Legal Knowledge and Information Systems

JURIX 2018: *The Thirty-first Annual Conference*

Editor:
Monica Palmirani

JURIX 2018

Foundation for Legal Knowledge Systems
JURIX

Legal Knowledge and Information Systems

Foundation for Legal Knowledge Systems

JURIX

JURIX 2018:
The Thirty-first Annual Conference

Editor:
Monica Palmirani

Artificial intelligence as applied to the legal domain has gained momentum thanks to the large, annotated corporate legal and case-law collections, human chats, and social media information now available in open data. Often represented in XML or other Semantic Web technologies, these now make it possible to use the AI theory developed by the JURIX community in over thirty years of research. Innovative machine and deep-learning techniques with which to classify legal texts and detect terms, principles, concepts, evidence, named entities, and rules are also emerging, and the last five years have seen a gradual increase in their practical application.

This book presents papers from the 31st International Conference on Legal Knowledge and Information Systems (JURIX 2018), held in Groningen, the Netherlands, in December 2018. The support of the Dutch Foundation for Legal Knowledge Based Systems for the JURIX conference has transformed a domestic workshop into an international event, with theoretical contributions, applied work, demo prototypes, a hackathon, and a doctoral consortium. Of the 72 submissions received, 17 full papers and 11 short papers were selected for publication, representing an acceptance rate of approximately 38%.

Machine learning for the legal domain prevails in the JURIX 2018 program, with traditional research mainstreams concerning legal reasoning and argumentation, natural-language processing, legal-text retrieval, and legal semantic modelling. An emerging topic is blockchain, which has graduated from the workshop area to the main program. The book offers an overview of the ways in which innovative information technologies are merging with legal theory, argumentation, and practice.

JURIX 2018

ISBN 978-1-61499-934-8

9 781614 999348

ISBN 978-1-61499-934-8 (print)
ISBN 978-1-61499-935-5 (online)
ISSN 0922-6389 (print)
ISSN 1879-8314 (online)

LEGAL KNOWLEDGE AND INFORMATION SYSTEMS

Frontiers in Artificial Intelligence and Applications

The book series Frontiers in Artificial Intelligence and Applications (FAIA) covers all aspects of theoretical and applied Artificial Intelligence research in the form of monographs, selected doctoral dissertations, handbooks and proceedings volumes. The FAIA series contains several sub-series, including 'Information Modelling and Knowledge Bases' and 'Knowledge-Based Intelligent Engineering Systems'. It also includes the biennial European Conference on Artificial Intelligence (ECAI) proceedings volumes, and other EurAI (European Association for Artificial Intelligence, formerly ECCAI) sponsored publications. The series has become a highly visible platform for the publication and dissemination of original research in this field. Volumes are selected for inclusion by an international editorial board of well-known scholars in the field of AI. All contributions to the volumes in the series have been peer reviewed.
The FAIA series is indexed in ACM Digital Library; DBLP; EI Compendex; Google Scholar; Scopus; Web of Science: Conference Proceedings Citation Index – Science (CPCI-S) and Book Citation Index – Science (BKCI-S); Zentralblatt MATH.

Volume 313

Recently published in this series

Vol. 312. T. Endrjukaite, A. Dudko, H. Jaakkola, B. Thalheim, Y. Kiyoki and N. Yoshida (Eds.), Information Modelling and Knowledge Bases XXX
Vol. 311. M. Coeckelbergh, J. Loh, M. Funk, J. Seibt and M. Nørskov (Eds.), Envisioning Robots in Society – Power, Politics, and Public Space – Proceedings of Robophilosophy 2018 / TRANSOR 2018
Vol. 310. N. Petkov, N. Strisciuglio and C. Travieso-González (Eds.), Applications of Intelligent Systems – Proceedings of the 1st International APPIS Conference 2018
Vol. 309. A.J. Tallón-Ballesteros and K. Li (Eds.), Fuzzy Systems and Data Mining IV – Proceedings of FSDM 2018
Vol. 308. Z. Falomir, K. Gibert and E. Plaza (Eds.), Artificial Intelligence Research and Development – Current Challenges, New Trends and Applications
Vol. 307. K. Muischnek and K. Müürisep (Eds.), Human Language Technologies – The Baltic Perspective – Proceedings of the Eighth International Conference Baltic HLT 2018
Vol. 306. S. Borgo, P. Hitzler and O. Kutz (Eds.), Formal Ontology in Information Systems – Proceedings of the 10th International Conference (FOIS 2018)
Vol. 305. S. Modgil, K. Budzynska and J. Lawrence (Eds.), Computational Models of Argument – Proceedings of COMMA 2018

ISSN 0922-6389 (print)
ISSN 1879-8314 (online)

Legal Knowledge and Information Systems

JURIX 2018: The Thirty-first Annual Conference

Edited by

Monica Palmirani

University of Bologna, CIRSFID, Italy

IOS
Press

Amsterdam • Berlin • Washington, DC

ISBN 978-1-61499-934-8 (print)
ISBN 978-1-61499-935-5 (online)
Library of Congress Control Number: 2018963896

Publisher
IOS Press BV
Nieuwe Hemweg 6B
1013 BG Amsterdam
Netherlands
fax: +31 20 687 0019
e-mail: order@iospress.nl

For book sales in the USA and Canada:
IOS Press, Inc.
6751 Tepper Drive
Clifton, VA 20124
USA
Tel.: +1 703 830 6300
Fax: +1 703 830 2300
sales@iospress.com

Preface

We are delighted to present the proceedings volume of the 31st International Conference on Legal Knowledge and Information Systems (JURIX 2018). For more than three decades, JURIX has organized an annual international conference for academics and practitioners, recently also including demos and a hackathon. The intention is to create a virtuous exchange of knowledge between theoretical research and applications in concrete legal use-cases. JURIX is also a good community where different skills work together to advance research by way of cross-fertilisation between law and computer technologies.

The JURIX conferences have been held under the auspices of the Dutch Foundation for Legal Knowledge Based Systems (www.jurix.nl). It has been hosted in a variety of European locations, extending the borders of its action and becoming an international conference in virtue of the the various nationalities of its participants and attendees.

The 2018 edition of JURIX, which runs from December 12 to 14, is hosted by the Faculty of Law and the Department of Artificial Intelligence in the Bernoulli Institute of Mathematics, Computer Science and Artificial Intelligence of the Faculty of Science and Engineering of the University of Groningen. JURIX 2018 is organised in cooperation with the Dutch Research School for Information and Knowledge Systems (SIKS). Special thanks go to Jeanne Mifsud Bonnici, Henry Prakken, and Bart Verheij and to their team for inviting us, hosting the event, and making this conference possible (http://jurix2018.ai.rug.nl).

For this edition we have received 72 submissions by 221 authors from 26 countries; 17 of these submissions were selected for publication as full papers (10 pages each) and 11 as short papers (five pages each), for a total of 28 presentations. We were inclusive in making our selection, but the competition stiff and the submissions were put through a rigorous review process with an acceptance rate of 38.8%. Borderline submissions, including those that received widely divergent marks, were accepted as short papers only.

The accepted papers have been grouped under six headings: (i) Machine Learning for the Legal Domain, a session presenting different methodologies and theoretical models applied to legislative texts and case-law (seven full papers and two short papers); (ii) Legal Reasoning and Argumentation, a session that ranges from theoretical aspects and demonstrations (three full papers and four short papers); (iii) Legal Knowledge Extraction, a session that presents natural-language processing of text for detecting terms, principles, concepts, evidence, rules, and named entities, and also speech in chatbots (two full papers and two short papers); (iv) Legal Knowledge Retrieval, a session focused on the answer-and-query approach (two full papers); (v) Legal Knowledge Modelling and Visualization, devoted to Semantic Web techniques, such as legal thesauri and ontologies (three full papers and one short paper); and (vi) Legal Blockchain, a session that has been growing in significance for several years in the workshop area, and now gains entry into the main conference (two full papers, one short paper).

This year we are honoured to have Marie-Francine Moens, full professor in the Department of Computer Science at Katholieke Universiteit Leuven, director of the Language Intelligence and Information Retrieval (LIIR) Research Lab, member of the Human Computer Interaction Group, and head of the Informatics section. She is a well-known researcher who experiments with novel methods for automated content recognition in text and multimedia, using statistical machine learning and exploiting insights from linguistic and cognitive theory. She has also successfully applied these techniques to the legal domain.

Also noteworthy is that the ethical aspects are increasingly relevant in big data and AI applications. For this reason we have asked to Jeroen van den Hoven to provide us with an overview of the ethical issues that emerge in connection with the use of emerging technologies. He is full professor of Ethics and Technology at the Delft University of Technology and editor in chief of Ethics and Information Technology.

We are very grateful to them for having accepted our invitation and for their interesting and inspiring talks.

Since 2013, JURIX has also hosted the Doctoral Consortium, now in its sixth edition. This initiative aims to attract and promote Ph.D. researchers in the area of AI & Law so as to enrich the community with original and fresh contributions. Many thanks are owed to Pompeu Casanovas, Ugo Pagallo, and Giovanni Sartor for organising the consortium this year, helped by other senior scholars.

As the previous editions, also this year the conference is growing richer with six co-located workshops. With long-running workshop like AICOL and LDA, we are continuing the TeReCom event and are hosting four new initiatives: XAILA, Legal Design, ManyLaws, and Legal Data Analytic Hackathon.

The Workshop on Artificial Intelligence and the Complexity of Legal Systems (AICOL), now in its tenth edition, is a stable event whose aim is to cut across multiple disciplines so as to examine the complexity of legal systems. The LDA workshop on Legal Data Analysis of the Central European Institute of Legal Informatics (CEILI), at its fifth edition, is devoted to the representation and analysis of legal data and documents, and to reasoning on such data and documents, using corpora and information systems.

The second edition of the Workshop on Technologies for Regulatory Compliance provides a forum for discussion of research on technologies for regulatory compliance on the basis of Semantic Web and Artificial Intelligence techniques. This workshop is supported by the LYNX European project: Building the Legal Knowledge Graph for Smart Compliance Services in Multilingual Europe (http://lynx-project.eu/).

The first-ever EXplainable AI in Law (XAILA) workshop aims to investigate the intersection of law and AI in order to provide a conceptual framework for ethical concepts and values in AI systems.

The first-ever ManyLaws workshop focuses on the semantic annotation of Big Legal Open Data, easily searchable and exploitable, on the basis of text-mining tools and algorithms offered through proper visualization techniques.

In this regard the Legal Design workshop integrates the previous ones with interdisciplinary and human-centered design principles to prevent or solve legal problems.

This year JURIX is also hosting a new challenge with the the Legal Data Analytics Hackathon (LeDAH), aiming to create applications and concrete projects using data-analytics methods applied to legal documents and data, with a specific focus on bringing out cognitive biases and providing visualization to support the transparency of legal information.

The JURIX 2018 conference was supported by IOS Press, BNKVI (the Benelux Association for Artificial Intelligence), and the OASIS LegalXML Steering Committee: many thanks to them, whose help made it possible to organise this event, and whose technical support contributed to attracting many participants from around the world.

Finally, we want to thank the program committee and sub-reviewers for reviewing the submissions with a professional and scientific attitude, enriched by active discussions that have ensured a fair reviewing process; the 221 authors who have submitted papers; the workshop organisers, who have enhanced the JURIX conference with emerging topics; the hackathon organizers for the applicative approach; and, finally, the members of the JURIX Steering and Executive Committees for supporting the conference year after year.

Monica Palmirani

CIRSFID, University of Bologna, Italy

This page intentionally left blank

Conference Organisation

Program Chair
Monica Palmirani, University of Bologna, CIRSFID

Local Co-Chairs
Jeanne Mifsud Bonnici, University of Groningen
Henry Prakken, University of Utrecht, University of Groningen
Bart Verheij, University of Groningen

Doctoral Consortium Chair
Pompeu Casanovas, Universitat Autònoma de Barcelona and La Trobe University

Local Organisation Committee
Annet Onnes, University of Groningen
Jelmer van der Linde, University of Groningen

Contact and administration
Elina Sietsema, University of Groningen
Sarah van Wouwe, University of Groningen

Program Committee
Tommaso Agnoloni, ITTIG-CNR
Laura Alonso Alemany, Universidad Nacional de Córdoba
Grigoris Antoniou, University of Huddersfield
Michał Araszkiewicz, Jagiellonian University
Kevin Ashley, University of Pittsburgh
Katie Atkinson, University of Liverpool
Matteo Baldoni, Università di Torino
Cesare Bartolini, University of Luxembourg
Nick Bassiliades, Aristotle University of Thessaloniki
Trevor Bench-Capon, University of Liverpool
Floris Bex, Utrecht University
Guido Boella, Università di Torino
Alexander Boer, University of Amsterdam
Danièle Bourcier, University of Paris 2/CNRS
Karl Branting, The MITRE Corporation
Elena Cabrio, Université Côte d'Azur, CNRS, Inria
Pompeu Casanovas, Universitat Autònoma de Barcelona and La Trobe University
Marcello Ceci, University of Luxembourg
Francisco Cifuentes-Silva, Biblioteca del Congreso Nacional de Chile
Jack G. Conrad, Thomson Reuters
Giuseppe Contissa, University of Bologna
Matteo Cristani, University of Verona
Claudia d'Amato, University of Bari

Sarah Sutherland, Canadian Legal Information Institute
Leon van der Torre, University of Luxembourg
Tom van Engers, University of Amsterdam
Marc van Opijnen, KOOP
Anton Vedder, KU Leuven
Bart Verheij, University of Groningen
Serena Villata, INRIA Sophia Antipolis, CNRS
Fabio Vitali, University of Bologna
Radboud Winkels, University of Amsterdam
Adam Wyner, Swansea University
John Zeleznikow, Victoria University
Haozhen Zhao, Navigant
Tomasz Zurek, Marie-Curie University of Lublin

Extra Reviewers
Son Nguyen Truong

JURIX steering committee
Katie Atkinson, Pompeu Casanovas, Erich Schweighofer, Serena Villata

JURIX executive committee
Floris Bex, Tom van Engers, Bart Verheij

jurix.nl
jurix2018.ai.rug.nl

This page intentionally left blank

Contents

This page intentionally left blank

Legal Knowledge and Information Systems
M. Palmirani (Ed.)

doi:10.3233/978-1-61499-935-5-1

Named Entity Recognition, Linking and Generation for Greek Legislation

Iosif ANGELIDIS [a], Ilias CHALKIDIS [b] and Manolis KOUBARAKIS [a]
[a] *National and Kapodistrian University of Athens, Greece*
[b] *Athens University of Economics and Business, Greece*

Abstract. We investigate named entity recognition in Greek legislation using state-of-the-art deep neural network architectures. The recognized entities are used to enrich the Greek legislation knowledge graph with more detailed information about persons, organizations, geopolitical entities, legislation references, geographical landmarks and public document references. We also interlink the textual references of the recognized entities to the corresponding entities represented in other open public datasets and, in this way, we enable new sophisticated ways of querying Greek legislation. Relying on the results of the aforementioned methods we generate and publish a new dataset of geographical landmarks mentioned in Greek legislation. We make available publicly all datasets and other resources used in our study. Our work is the first of its kind for the Greek language in such an extended form and one of the few that examines legal text in a full spectrum, for both entity recognition and linking.

Keywords. Named Entity Recognition and Linking, Dataset Generation, Entity Reference Representation, Deep Learning

1. Introduction - Related Work

Recently, there has been an increased interest in the adaptation of Artificial Intelligence technologies to the legal domain including text processing, knowledge representation and reasoning. Legal text processing [1] is a growing research area, comprising of tasks such as legal question answering [2] and legal entity extraction [3,4]. The same applies to the area of legal knowledge representation, where new standards have been developed and started to be adopted based on semantic web technologies. Relevant contributions here are the European Legislation Identifier (ELI) [5] for legislation, the European Case Law Identifier (ECLI) [6] for case laws, as well as LKIF [7] and LegalRule ML [8] for the codification of advanced legal concepts, such as rules and norms. The research community aims to develop tools and applications to help legal professionals as well as ordinary citizens. Based on these principles, Chalkidis et al. [9] developed Nomothesia (`http://legislation.di.uoa.gr`), a platform which makes Greek legislation available on the Web as linked open data to aid its sophisticated querying using SPARQL and the development of relevant applications.

Deepening this effort in order to build a bridge, as a point of reference, between those relative research fields of artificial intelligence (natural language processing and semantic web), we developed a Named Entity Recognizer (NER) and Linker (NEL) for Greek legislation.

Our main contributions are listed below:

- We study the task of named entity recognition in Greek legislation (Section 2) by evaluating state-of-the-art neural architectures that have been applied in legal text for other tasks (contract elements extraction [3], recognition of requisite-effectuation parts [4]). In these experiments, we compare two alternative token shape encodings, which signify the importance of an expressive feature representation.
- We introduce a novel RDF vocabulary for the representation and linking of textual references to entities (Subsection 3.1). As Chalkidis et al. [9], we consider RDF as a single data model for representing both metadata of a legislative document and knowledge that is encoded in the text.
- We deploy Nomothesia NER, based on the best model BILSTM-BILSTM-LR (Subsection 2.2) with a macro-averaged F_1 of 0.88, in the Greek legislation dataset [9] and produce new data for entity references, that we describe using the new RDF vocabulary.
- We link the references with open public datasets (Greek administrative units and Greek politicians) using rule-based techniques and the Silk framework [10] (See Section 3).
- We make publicly available new benchmark datasets (Subsection 2.1.1) of 254 annotated pieces of legislation related to named entity recognition and linking. Pre-trained word embeddings specialized in Greek legal text, demonstration code in python and other supplementary material are also provided.
- We generate a new RDF dataset of Greek geographical landmarks based on the results of Nomothesia NER by applying heuristic rules (Subsection 3.5).
- Based on the above procedures, we augment the knowledge base and the querying capabilities of the Nomothesia platform in two significant ways: tracing legislation citation networks and searching using entity-based criteria (Subsection 3.6).

All methods and practices that are described throughout this work can be applied in any given language given the appropriate pre-trained word embedding and datasets.

2. Entity Recognition

In this paper, we focus on extracting 6 entity types, when present:

Person Any formal name of a person mentioned in the text (e.g., Greek government members, public administration officials, etc.).

Organization Any reference to a public or private organization, such as: international organizations (e.g., European Union, United Nations, etc.), Greek public organizations (e.g., Social Insurance Institution) or private ones (e.g., companies, NGOs, etc.).

Geopolitical Entity Any reference to a geopolitical entity (e.g., country, city, Greek administrative unit, etc.)

Geographical Landmark References to geographical entities such as local districts, roads, farms, beaches, which are mainly included in pieces of legislation related to topographical procedures and urban planning.

Legislation Reference Any reference to Greek or European legislation (e.g., Presidential Decrees, Laws, Decisions, EU Regulations and Directives, etc.)

Public Document Reference Any reference to documents or decisions that have been published by a public institution (organization) that are not considered a primary source of legislation (e.g., local decisions, announcements, memorandums, directives).

2.1. Datasets

2.1.1. Benchmark Datasets

The benchmark datasets[1] contain 254 daily issues for classes A and D of the Greek Government Gazette over the period 2000-2017. Every issue contains multiple legal acts. Class A issues concern primary legislation published by the Greek government (e.g., laws, presidential decrees, ministerial decisions, regulations, etc.). Class D issues concern decisions related to urban, rural and environmental planning (e.g., reforestations, declassifications, expropriations, etc.). We uniformly splitted the issues across training (162), validation (45), and test (47) in terms of publication year and class. Thus the possibility of overfitting due to specific linguistic idiosyncrasies in the language of a government or due to specific entities and policies has been minimized. Our group annotated all of the above documents for the 6 entity types that we examine. We also created datasets that contain pairs of entity references and the respective matching Universal Resource Identifiers (URIs) in other open public datasets.

2.1.2. Word Embeddings

The last few years, feature engineering in NLP, which results in sparse feature representations, has gradually been replaced by the use of dense word vectors, most notably word embeddings. Word embeddings are pre-trained using unsupervized algorithms [11,12] over large corpora based on the linguistic observation that similar words tend to co-occur in similar contexts (phrases). Thus, word embeddings capture both semantic and syntactic information as well as correlations between words. In our work, we applied WOR2VEC (skip-gram model) [11] to an unlabelled corpus, which contains: 150,000 issues of the Greek Government Gazette in the period of 1990-2017; all publicly available pieces of legislation from the foundation of the Greek Nation in 1821 until 1990, which sum up to 50,000; 1,500 case laws published online by Greek Courts; all EU Treaties, Regulations and Decisions that have been translated in Greek and published in EUR-Lex; and the Greek part of the European Parliament Proceedings Parallel Corpus.

We produced 100-dimensional word embeddings for a vocabulary of 428,963 words (types), based on 615 millions of tokens (words), included in the unlabelled corpus. We used Gensim's implementation of WORD2VEC (http://radimrehurek.com/gensim/). Out of vocabulary words were mapped to a single 'UNK' embedding. To generalize across numbers with similar patterns and tokens that differ in letter-case formats, the unlabeled corpus was pre-processed to be upper-cased, de-accented and underwent replacement of all digits by 'D' for all tokens. We opted to do such transformations, in contrast to the usual lower-casing, based on the fact that the modern Greek alphabet consists of 24 upper-case letter cases and 25 lower-case ones, which also in many circumstances can be accented with 2 different accents in their lower-case formats. So, we normalize every word in an upper-case non-accented form (i.e.,'Νόμος', '1ή' encoded as 'ΝΟΜΟΣ', '1H'). English words were also mapped to a single word named 'ENGLISH_WORD'.

Moreover, we experimented with generic pre-trained 200-dimensional word embeddings (publicly available), trained with FASTTEXT [12] (https://fasttext.cc) and

[1] Datasets and Supplementary Material are published in https:/http://legislation.di.uoa.gr/publications under a non-commercial license. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/4.0/.

based on a much larger corpus with Greek Wikipedia articles. The experimental results were worse, possibly because legal expressions are under-represented (or do not exist) in generic corpora (e.g., wikipedia or news articles) that were used, while also the preprocessing seems really poor based on our observation.

We also experimented with two different formats of token shape embeddings [13]. Chalkidis et al. [3] proposed 5-dimensional token shape embeddings that represent the following seven possible shapes of tokens: token consisting of alphabetic upper-case characters, possibly including periods and hyphens (e.g., 'ΠΡΟΕΔΡΟΣ', 'Π.Δ.', 'ΠΔ/ΤΟΣ'); token consisting of alphabetic lower-case characters, possibly including periods and hyphens (e.g., 'νόμος', 'ν.', 'υπερ-φόρτωση'); token with at least two characters, consisting of an alphabetic upper-case first character, followed by alphabetic lower-case characters, possibly including periods and hyphens (e.g., 'Δήμος', 'Αναπλ.'); token consisting of digits, possibly including periods and commas (e.g., '2009', '12,000', '1.1'); line break; any token containing only non-alphanumeric characters (e.g., '.', '€'); and any other token (e.g., '1o', 'OIK/88/4522', 'EU').

In addition and for comparison, we used 25-dimensional token shape embeddings by generating 1578 shapes for tokens by replacing each alphabetic lower-case and upper-case character with 'c' and 'C' respectively, each digit with 'd' and punctuation characters kept from the original token (i.e. 'Π.δ./τος', 'Αναπλ.', '1963' encoded as 'C.c./ccc', 'Ccccc.' 'dddd'). If the same character is encountered more than 4 times in a token in the row, it is limited to 4 times (e.g., '123456' is mapped into 'dddd'). Intuitively, having such representations for tokens offers more information regarding the context. In general, the shape (form) of its token relies on the existence and relative position of alphabetic characters, digits and punctuations.

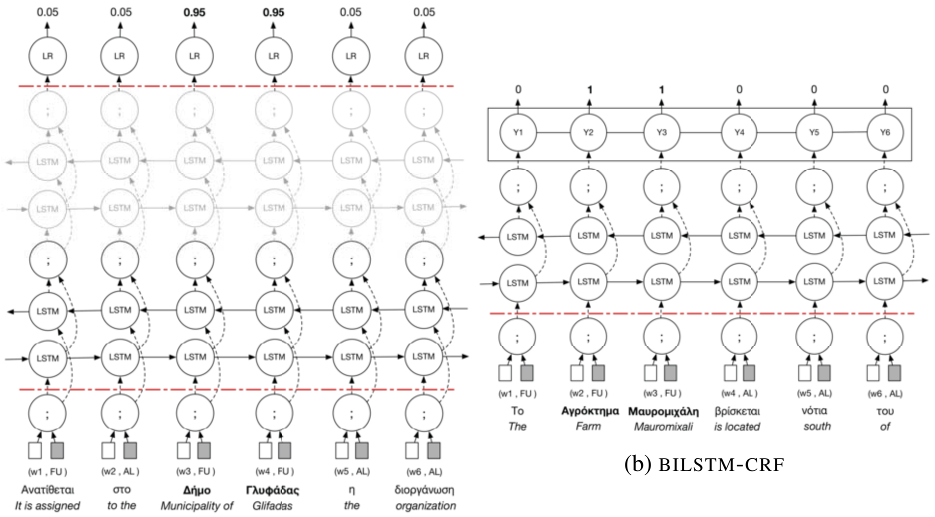
We were unable to embed the part-of-speech tag of each token due to the fact that so far there is no reliable POS tagger for the Greek language. We verified this by experimenting with the NLTK (http://www.nltk.org) POS tagger and, as also the one provided by CLTK (http://cltk.org), but both of them had a vast amount of wrong predictions, a fact that is even more profound in legal text.

2.2. *LSTM-based methods*

Until the recent advances in deep learning [14,15], NLP techniques were dominated by machine-learning methods that used linear models such as support vector machines or logistic regression. Currently, the focus of the community has switched from such linear models to multi-layer neural-network models. In this work, we experiment with Recurrent Neural Networks (RNNs), more specifically LSTM-based models that produce state-of-the-art results for language modeling [16,17], as well as for named entity recognition and part-of-speech tagging [18], all of which are sequence tagging tasks.

In this section, we describe the three LSTM-based methods we experimented with[2]. The LSTM (Long Short-Term Memory) [19] units are a more sophisticated form of the Simple Recurrent (RNN) units, tailored to resolve memory issues on sequential data. Each unit contains a self-connected memory cell and three multiplicative units: the input, output and forget gates, which provide continuous analogues of write, read and reset operations for the cells in each timestep (word). Further on, the LSTM layers are bidirectional in the sense that the input sequence is being processed from left to right and from right

[2] The methods were implemented using KERAS (https://keras.io/).



(a) BILSTM(-BILSTM)-LR

(b) BILSTM-CRF

Figure 1. LSTM-based methods

to left, conditioning information from the previous and next timesteps (tokens), respectively.

The first LSTM-based method that we have used, called BILSTM-LR (Figure 1a), uses a bidirectional LSTM (BILSTM) layer to convert the concatenated word and token shape embeddings of each token (lower ; nodes of Figure 1a) in each sentence to context-aware token embeddings (upper ; nodes), which better describe the semantics of each token given the specific task. Each context-aware token embedding is then passed on to the logistic regression (LR) layer (LR nodes of Figure 1a including the SOFTMAX activation) to estimate the probability that the corresponding token belongs to each of the examined categories (e.g., person, organization, etc.).

Based on experimentation, the pre-trained word embeddings are not updated during training on the labeled dataset, while in contrast the token shape embeddings are not pre-trained. The corresponding shape vectors are being learned during the actual training. We used Glorot initialization [20], binary cross-entropy loss, and the Adam optimizer [21] to train the BILSTM-LR recognizer with early stopping by examining the validation loss. Hyper-parameters were tuned by grid-searching the following sets, and selecting the values with the best validation loss: LSTM hidden units {100, 150}, batch size {16, 24, 32}, DROPOUT rate {0.4, 0.5}.

The second LSTM-based method, called BILSTM-BILSTM-LR, has an additional BILSTM layer (transparent upper LSTM nodes of Figure 1a) between the context-aware token embeddings (lower ; nodes) of the lower BILSTM layer, and the logistic regression (LR) layer (LR nodes). Stacking LSTM (or BILSTM) layers has been reported to improve efficacy in several natural language processing tasks [22] at the expense of increased computational cost. Our experimental results (presented in Section 2.3 below) show that the additional BILSTM chain of BILSTM-BILSTM-LR also leads to significant improvements.

In the third LSTM-based method, called BILSTM-CRF, we replace the logistic regression layer of the BILSTM-LR method with a linear-chain of Conditional Random Fields

(CRFs), as illustrated in the Figure 1b. CRFs [23] have been widely used in NLP sequence labeling tasks (e.g., POS tagging, named entity recognition). They have also shown exceptional results on top of BILSTMs in sequence labeling [18]. Hyper-parameter tuning and training are performed as in the previous LSTM-based methods.

2.3. Experiments and Evaluation

For each of the three methods we measured the performance on *precision*, *recall*, and F_1 scores based on the MUC guidelines [24][3]. Table 1 lists the results of this group of experiments based on the average of five individual runs for each method.

ENTITY	BILSTM-LR				BILSTM-CRF				BILSTM-BILSTM-LR			
TYPE	P	R	F1	F1'%	P	R	F1	F1'%	P	R	F1	F1'%
Person	0.95	0.86	0.90	-1%	0.94	0.91	0.92	-2%	0.94	0.92	**0.93**	-3%
Organization	0.90	0.71	0.79	-4%	0.87	0.72	0.79	-6%	0.92	0.76	**0.83**	-5%
GPE	0.90	0.80	0.85	-	0.90	0.76	0.83	+2%	0.92	0.85	**0.88**	-1%
GeoLandmark	0.88	0.77	0.83	-11%	0.83	0.80	0.81	-11%	0.93	0.86	**0.89**	-5%
Legislation Ref.	0.95	0.79	0.86	-3%	0.94	0.79	0.86	-7%	0.96	0.84	**0.90**	-5%
Public Document	0.82	0.69	0.75	+3%	0.76	0.69	0.73	-1%	0.88	0.76	**0.82**	-
Macro AVG	0.89	0.79	0.84	-1%	0.87	0.80	0.83	-2%	0.91	0.85	**0.88**	-2%

Table 1. Precision (P), Recall (R), and F_1 score. Best F_1 per entity type shown in bold font. The F_1 score show the performance using the 25-dimensional shapes, while $F_1'\%$ show the perfomance reduction using the 5-dimensional shapes for the seven predefined categories.

The results are highly competitive for all the examined methods. The best results, based on the macro-averaged F_1, are coming from BILSTM-BILSTM-LR (0.88), which indicates that the extra BILSTM layer, which deepens the model, expands its capacity by a significant margin, compared to BILSTM-LR (0.84) and BILSTM-CRF (0.83). Considering the generic FASTTEXT pre-trained embeddings, instead of our domain-specific ones, leads to a macro-averaged F_1 of 0.81 for the best reported method BILSTM-BILSTM-LR, especially in the latter four categories, in which domain knowledge matters the most (e.g., geographical aspects and codification of documents). Considering more expressive token shapes also seems to improve the performance of the examined model (by a factor of 2% in the case of the BILSTM-BILSTM-LR, 0.88 vs 0.86). Further on, we are going to rely on the BILSTM-BILSTM-LR classifier based on the fact that it outperforms the other models in every entity-type.

3. Entity Linking

As we already mentioned, the complexity of legal text and the particularities of the Greek language itself provide an additional challenge in our goal to link the identified references. Based on heuristic rules, we were able to segment and normalize entity references and proceed to the task of entity linking[4].

[3] MUC guidelines consider partial token overlaps between the gold annotations and the predicted entities (sequences of consecutive tokens that belong in the same class), given the correct (gold) class.

[4] Technical details have been stated in the assisting Supplementary Material document of the datasets.

3.1. A vocabulary for textual entity references

The first step towards linking entity references extracted by Nomothesia NER with the entities described in public open datasets is to represent those references using the RDF specification. The *legal text* of a document contains subdivisions (passagges of individual laws) that are defined as *LegalResourceSubdivisions* based on the Greek legislation ontology. Since some of those contain text, it is also possible to contain (*has_reference* to) a *Reference* to an entity (e.g., a law passage referring to a specific law that it modifies, or an organization). This reference is realized in an interval of characters. In other words, it *begins* and *ends* on specific sequential characters inside the text of the subdivision. This *Reference* most likely refers to (or, in another sense, is *relevant_for*) an *Entity*, which is probably described in open public datasets. Therefore, in our case, a *LegalResource-Subdivision* contains references to persons, administrative units and legal resources (e.g., laws, decisions etc.). The former description is depicted in Figure 2.

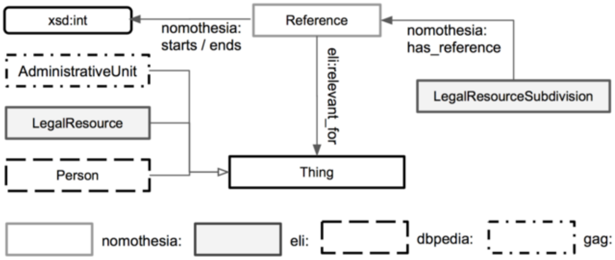


Figure 2. Textual Reference RDF Vocabulary

3.2. Linking entity references

We linked legal references with legal documents provided by the Greek legislation dataset[5]. We based on heuristic rules to directly interpret the relevant URI by capturing the *type*, *year* of publication and the *serial number*. We provide performance evaluation in Table 2.

3.3. Linking Greek Politicians and Greek Administrative Units using Silk

We linked person references with Greek politicians retrieved from the Greek DBpedia (`http://el.dbpedia.org/`) dataset and geopolitical entity references with the Greek administrative units, as they are described in the Greek Administrative Geography (GAG) dataset (published in `http://linkedopendata.gr/dataset`).

For both entity types, we proceed in interlinking the corresponding datasets using the Silk framework [10]. We experimented with two different textual linking operators: Levenshtein and substring distance [25] over the `rdfs:label` values provided by each dataset. For the case of the Greek Administrative Units, we also provided the *type* of the administrative units (e.g., local community, municipality, region, etc.) based on the naming conventions that we identified in the validation part of the labeled dataset.

[5] Published in `http://legislation.di.uoa.gr/legislation.n3`

3.4. Evaluation

For each interlinking method that we tried, we examine the performance of the interlinking in terms of *precision*, *recall*, and F_1 score *measured per entity pair* on the test part of our labeled dataset. Here, true positives (*TP*) are references correctly paired with an entity of each set, *false positives* (*FP*) are references incorrectly paired with entities, and *false negatives* (*FN*) are references incorrectly not paired with the relative entities of the examined sets. The acceptance threshold for both linking operators was tuned on the validation part of our datasets, while the entity pairs provided are those presented in the test part. Table 2 lists the results for this group of experiments.

METRICS	LINKING TECHNIQUE								
ENTITY	RULES			LEVENSHTEIN			SUBSTRING		
TYPE	P	R	F1	P	R	F1	P	R	F1
Person	-	-	-	0.99	0.55	0.71	0.90	0.68	**0.77**
GPE	-	-	-	0.99	0.79	0.88	0.95	0.92	**0.94**
Legislation Ref	0.99	0.97	**0.98**	-	-	-	-	-	-

Table 2. Precision (P), Recall (R), and F_1 score, *measured per entity pair*

Linking persons to Greek politicians was a great challenge, mainly because legislators tend to refer to a person's first name by its initials (e.g., 'A. Tsipras'), thus a fair amount of person references have been misclassified (precision: 0.71) for persons with the same surname. We successfully linked the geopolitical entities with the Greek administrative units (F_1: 0.92). Minor issues are related to the segmentation of compound references of multiple administrative units. The results for legislation references are robust (F_1: 0.98), while a short margin of documents are mis-linked due to the fact that ministerial decisions do not have a standard codification (nor a standard reference pattern), which varies from one ministry to another.

3.5. Greek geographical landmarks dataset generation

Greek geographical landmarks are a major asset for our legal recognizer since they are related to planning and architectural interests. However, there is no such public dataset to interlink between the references and the actual entities. We proceed in generating a new dataset by applying linguistic heuristics in order to form the entities and classify their type in 4 different abstract categories (classes):

Local District such as villages and small local communities.

Area sub-classified into *agricultural*, *forest*, *coastal* and *marine* areas.

Road sub-classified into *highway*, *local*, *bypass* roads.

Point of Interest such as a *farm*, an *islet*, or a *peninsula* which are commonly referred to urban planning legislation.

Further on, we interlink the new dataset with the Greek administrative units when there is a connection between them (*belongs_to*), indicated in terms of text (e.g., 'Beach Kavouri at Municipality of Varis-Voulas-Vouliagmenis').

3.6. Querying the augmented Greek legislation datasets

In this section we demonstrate new forms of querying the augmented Greek legislation dataset. The linking process expanded the Greek legislation dataset from approximately 2,9M triples to 4,4M triples in order to describe knowledge for 194,102 references of the supported entity types. Based on the above, we have the ability to pose queries against the resulting RDF graph using SPARQL (Table 3). Nomothesia platform also provides a SPARQL endpoint (http://legislation.di.uoa.gr/endpoint), as also a designated page for searching named entities (http://legislation.di.uoa.gr/entities).

<table>
<tr><td colspan="3"><b>Q1:</b> <i>Retrieve any legal act that refer to municipalities, which belong to the regional unit of Florinas.</i></td></tr>
<tr><td>SELECT DISTINCT ?municipality_name (group_concat(?act;separator=", ") as ?act_ids)
WHERE {
?act eli:id_local ?act_id. ?act eli:has_part+ ?part.
?part lego:has_reference ?reference. ?reference eli:relevant_for ?gpe.
?gpe owl:sameAs ?municipality. ?municipality rdfs:label ?local_district_name.
?municipality a gag:Municipality. ?municipality lego:belongs_to ?reg_unit.
?reg_unit rdfs:label "REGIONAL UNIT OF FLORINA"@en.
} GROUP BY ?municipality_name LIMIT 5</td><td><b>Municipality</b>
“DIMOS PRESPON”
“DIMOS AMINTAIOU”
“DIMOS FLORINAS”</td><td><b>Act ID</b>
“leg:pd/1998/310, ...”
“leg:law/1997/2539, ...”
“leg:law/2013/4109, ...”</td></tr>
<tr><td colspan="3"><b>Q2:</b> <i>Retrieve any legal acts that contain references to persons that have been born in Athens.</i></td></tr>
<tr><td>SELECT DISTINCT ?person_name
(group_concat(?act_id;separator=", ") as ?act_ids)
WHERE {
?act eli:id_local ?act_id. ?act lego:published_by ?signer.
?signer lego:relevant_for ?person. ?person a lego:Person.
?person owl:sameAs ?per. ?per rdfs:label ?person_name.
?per dbpedia-owl:birthPlace <http://el.dbpedia.org/resource/>.
} GROUP BY ?person_name LIMIT 5</td><td><b>Signer</b>
“ST. LABRINIDIS”
“G. GENNIMATAS”
“P. PIKRAMENOS”
“G. PAPAKONSTANTINOY”
“D. AVRAMOPOULOS”</td><td><b>Act ID</b>
“leg:dec/.3440_48844, ...”
“leg:law/1990/1874, ...”
“leg:la/1_14.06.2012, ...”
“leg:la/1_31.12.2009, ...”
“leg:dec/0546_.6551_55, ...”</td></tr>
<tr><td colspan="3"><b>Q3:</b> <i>Retrieve any legal act and its written code that is being referred to in Law 2014/4261.</i></td></tr>
<tr><td>SELECT DISTINCT ?ref_act ?ref_code
WHERE {
<http://legislation.di.uoa.gr/eli/law/2014/4261> eli:has_part+ ?part.
?part lego:has_reference ?ref. ?ref lego:relevant_for ?ref_act.
{?ref_act a lego:PresidentialDecree} UNION {?ref_act a lego:Law}.
?ref lego:has_original_label ?ref_label
} LIMIT 5</td><td><b>Act URI</b>
leg:law/2017/3356
leg:law/2007/3606
leg:law/2008/3691
leg:law/1920/2190
leg:law/2006/3455</td><td><b>Act Code</b>
“ν. 3556/2007”
“ν. 3606/2007”
“ν. 3691/2008”
“κ.ν. 2190/1920”
“ν. 3455/2006”</td></tr>
</table>

Table 3. Entity-based queries

Conclusion and Future Work

We evaluated LSTM-based methods for named entity recognition in Greek legislation, demonstrating the effectiveness of such techniques in language-diverse environments. We introduced and applied a novel vocabulary for the representation of textual references in RDF. Further on, we evaluated entity-linking between textual references and entities from third-party datasets, while we generated a new dataset for Greek geo-landmarks.

Our future plans include the investigation of deeper and/or CNN-based architectures and also the use of ELMo embeddings [17], replacing the current feature representation. In relation to the entity linking, we will conduct experiments with character-level neural approaches as alternative textual linking operators. We also endeavour to extract more geospatial information to augment the newly-published Greek geo-landmarks dataset.

Acknowledgement

This research has been cofinanced by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH-CREATE-INNOVATE (project code: T1EDK-01000, project name: GeoREgC).

References

[1] K. Ashley, *Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age*. Cambridge University Press, 2017.

[2] M. Kim and R. Goebel, “Two-step cascaded textual entailment for legal bar exam question answering,” in *Competition on Legal Inf. Extraction/Entailment*, (London, UK), 2017.

[3] I. Chalkidis and I. Androutsopoulos, “A deep learning approach to contract element extraction,” in *Int. Conf. on Legal Knowledge and Inf. Systems*, (Luxembourg), pp. 155–164, 2017.

[4] S. N. Truong, N. L. Minh, K. Satoh, T. Satoshi, and A. Shimazu, “Single and multiple layer BI-LSTM-CRF for recognizing requisite and effectuation parts in legal texts,” in *Int. Conf. on Artificial Intelligence and Law*, (London, UK), pp. 159–168, 2017.

[5] J. Dann, “European Legislation Identifier "ELI",” tech. rep., European Commission, 2014.

[6] M. V. Opijnen, “European Case Law Identifier: Indispensable Asset for Legal Inf. Retrieval,” in *From Inf. to Knowledge* (M. A. Biasiotti and S. Faro, eds.), vol. 236 of *Frontiers in Artificial Intelligence and Applications*, pp. 91–103, IOS Press, 2011.

[7] R. Hoekstra, J. Breuker, M. Di Bello, and A. Boer, “Lkif core: Principled ontology development for the legal domain,” in *Conf. on Law, Ontologies and the Semantic Web: Channelling the Legal Inf. Flood*, (Amsterdam, The Netherlands, The Netherlands), pp. 21–52, IOS Press, 2009.

[8] T. Athan, G. Governatori, M. Palmirani, A. Paschke, and A. Z. Wyner, “Legalruleml: Design principles and foundations,” in *Reasoning Web. Web Logic Rules*, 2015.

[9] I. Chalkidis, C. Nikolaou, P. Soursos, and M. Koubarakis, “Modeling and querying greek legislation using semantic web technologies,” in *European Semantic Web Conf.*, (Portorož, Slovenia), pp. 591–606, 2017.

[10] C. B. Anja Jentzsch, Robert Isele, “Silk - generating rdf links while publishing or consuming linked data,” in *Int. Semantic Web Conf.*, (Shanghai, China), 2010.

[11] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed Representations of Words and Phrases and their Compositionality,” in *Int. Conf. on Neural Inf. Proc. Systems*, (Stateline, NV), 2013.

[12] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword inf.,” *arXiv preprint arXiv:1607.04606*, 2016.

[13] D. Jurafsky and J. H. Martin, *Speech and Language Proc. :An Introduction to Natural Language Proc., Computational Linguistics, and Speech Recognition (3rd ed. draft)*. 2018.

[14] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[15] Y. Goldberg, *Neural Network Methods in Natural Language Processing*. Morgan and Claypool Publishers, 2017.

[16] S. Merity, N. S. Keskar, and R. Socher, “Regularizing and optimizing LSTM language models,” *CoRR*, vol. abs/1708.02182, 2017.

[17] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, “Deep contextualized word representations,” in *Conf. of NAn Chapter of the Association for Computational Linguistics*, (New Orleans, Louisiana, USA), 2018.

[18] X. Ma and E. Hovy, “End-to-end sequence labeling via bi-directional LSTM-cNNs-CRF,” in *Annual Meeting of the Association for Computational Linguistics*, (Berlin, Germany), pp. 1064–1074, 2016.

[19] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[20] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” in *Int. Conf. on Artificial Intelligence and Statistics*, (Sardinia, Italy), pp. 249–256, 2010.

[21] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Int. Conf. on Learning Representations*, (San Diego, CA), 2015.

[22] O. Irsoy and C. Cardie, “Deep recursive neural networks for compositionality in language,” in *Int. Conf. on Neural Inf. Proc. Systems*, (Montreal, Canada), pp. 2096–2104, 2014.

[23] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” in *Int. Conf. on Machine Learning*, (Williamstown, MA), pp. 282–289, 2001.

[24] D. Nadeau and S. Sekine, “A survey of named entity recognition and classification,” *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 3–26, 2007.

[25] G. Stoilos, G. Stamou, and S. Kollias, “A string metric for ontology alignment,” in *Int. Semantic Web Conf.*, (Galway, Ireland), pp. 624–637, 2005.

Legal Knowledge and Information Systems
M. Palmirani (Ed.)

doi:10.3233/978-1-61499-935-5-11

Lessons from Implementing Factors with Magnitude

Trevor BENCH-CAPON, Katie ATKINSON
Department of Computer Science, The University of Liverpool, UK

Abstract. We discuss the lessons learned from implementing a CATO style system using factors with magnitude. In particular we identify that giving factors magnitudes enables a diversity of reasoning styles and arguments. We distinguish a variety of ways in which factors combine to determine abstract factors. We discuss several different roles for values. Finally we identify the additional value related information required to produce a working program: thresholds and weights as well as a simple preference ordering.

Keywords. legal case based reasoning, dimensions, factors, values.

1. Introduction

Reasoning with legal cases has always been a central concern of AI and Law. Much of the investigation of this topic has been based on the pioneering work of Rissland and Ashley's HYPO system [23], [6]. Subsequent development of HYPO's ideas is described in [8]. While HYPO used *dimensions*, aspects of a case which were described using a range with one end favouring the plaintiff and the other end favouring the defendant, most subsequent work has used the simpler notion of *factors* to represent cases. Factors, as introduced in CATO [5], are stereotypical fact patterns which are legally significant and which can be seen in Boolean terms, as either present or absent [8]. For a discussion of the differences between dimensions and factors, see [24]. Using the simplification enabled by representing cases as a set of Booleans, a good understanding of this kind of reasoning has been developed, as formalised in [18] and [21]. It remains a simplification, however, and many researchers have often felt that it would become necessary to return to dimensions in order to give a full account of reasoning with legal cases (e.g. [11]). Both HYPO and CATO supported argumentation in US Trade Secret law and took as their starting point section 757 of the *Restatement of Torts*:

> "Some factors to be considered in determining whether given information is one's trade secret are: the *extent* to which the information is known outside of his business; the *extent* to which it is known by employees and others involved in his business; the *extent* of measures taken by him to guard the secrecy of the information; the *value* of the information to him and to his competitors; the

> *amount* of effort or money expended by him in developing the information; and the *ease or difficulty* with which the information could be properly acquired or duplicated by others"[1].

In CATO, we find several factors such as *InfoObtainableElsewhere*, *InfoKnownToCompetitors*, *SecurityMeasures*, *CompetitiveAdvantage* and *InfoReverseEngineerable*, all of which clearly originate in this section of the *Restatement*, but which are represented as Booleans, despite the emphasis on the *extent* to which these notions are satisfied in the *Restatement*. From this it is clear that concepts which are treated as Boolean factors in CATO, should not really be Boolean but have *magnitudes* (extents, amounts, degrees of difficulty and the like), and it should be for the court (rather than the analyst as in CATO) to decide whether the extent is sufficient, given the particular facts of the case. The simplification from dimensions to factors has proved useful, greatly facilitating understanding of several aspects of legal CBR, but the time has now come to return to the original notions of dimensions and factors with magnitude and much attention in recent AI and Law research has focussed on how this can best be done.

The persistent need for dimensions was argued in [11] and related to argumentation schemes for legal CBR in [20]. Formal representations for exploring a logic of precedent for dimensions and factors with magnitude have recently been proposed by Horty [17] and Rigoni [22]. In [7] a method for representing case law using factors with magnitude based on the methodology of [2] was described and an implementation of Aleven's CATO analysis using this representation as formalised in [1] was demonstrated at the COMMA 2018 conference [10]. In this paper we draw out the lessons learned from that implementation which can be used to guide future research on the topic.

Section 2 describes the representation used. Section 3 discusses the various ways in which factors are combined. Section 4 relates the implementation to the statement types of [3]. Section 5 discusses the additional types of reasoning enabled by using magnitudes and gives a hypothetical example case and variations to illustrate these different kinds of reasoning. Section 6 offers some concluding remarks.

2. Representation

The knowledge in the implemented program is represented following the ANGELIC methodology [2], specifically using the 2-regular structure of [1], in which statements (issues, intermediate concepts and factors) are represented as nodes and non-leaf nodes have exactly two children (see Table 1). Essentially this is used as the design documentation on which the implementation of [10] is based.

Each non-leaf node is associated with acceptance conditions, expressed in terms of its two children. Like [9] the statements have *degrees of acceptance*, in range [0,1]: 0 representing total rejection and 1 total acceptance. Some statements are genuinely Boolean and they use only 0 and 1. The acceptance conditions are then a set of conditions for attributing particular degrees of acceptance, together

[1] http://www.lrdc.pitt.edu/ashley/RESTATEM.HTM *italics ours.*

Table 1. 2-Regular ADF for CATO.

Parent	Child 1	Child 2
TradeSecretMisappropriation	SecretMisappropriated	EmployeeSoleDeveloper
SecretMisappropriated	InfoMiasappropriated	Info Trade Secret
Info Miasappropriated	BreachOfConfidence	ImproperMeans
Info Trade Secret	InfoValuable	EffortstoMaintainSecrecy
InfoValuable	InfoUseful	KnownOrAvailable
EffortstoMaintainSecrecy	AdequateEfforts	SecurityFailures
InfoUseful	CompetitiveAdvantage	UniqueProduct
KnownOrAvailable	Known	Available
Known	KnownOutside	Limitations
InfoAvailableElsewhere	InfoReverseEngineerable	InfoObtainableElsewhere
KnownOutside	InfoKnownToCompetitors	DisclosureInPublicForum
Limitations	UniqueProduct	MaintainSecrecyOutsiders
MaintainSecrecyOutsiders	SecretsDisclosedOutsiders	OutsiderDisclosuresRestricted
AdequateEfforts	SecurityMeasures	MaintainSecrecyDefendant
SecurityFailures	Reckless	WaiverOfConfidentiality
MaintainSecrecyDefendant	AgreedNotToDisclose	DisclosureInNegotiations
Reckless	NoSecurityMeasures	SecretsDisclosedOutsiders
Disclosed	DisclosureInPublicForum	MaintainSecrecyOutsiders
BreachOfConfidence	InfoUsed	ConfidentialRelationship
ImproperMeans	QuestionableMeans*	LegitimatelyObtainable
ConfidentialRelationship	NoticeofConfidentiality	ConfidentialityAgreement
ConfidentialityAgreement	AgreedNotToDisclose	WaiverOfConfidentiality
InfoUsed	GaveHelp	InfoIndependentlyGenerated
NoticeofConfidentiality	ValidAgreement	AwareConfidential
QuestionableMeans	IllegalMethods	DEfOKMethods
LegitimatelyObtainable	InfoKnownorAvailable	InfoAvailableElsewhere
GaveHelp	IdenticalProducts	GaveAdvantage
GaveAdvantage	BroughtTools	CompetitiveAdvantage
IllegalMethods	Criminal	Dubious
DefOKMethods	DisclosureInNegotiation	DefendantDiscovered
Criminal	BribeEmployee	InvasiveTechniques
Dubious	RestrictedMaterialsUsed	Deception
DefendantDiscovered	InfoIndependentlyGenerated	InfoReverseEngineered
AwareConfidential	RestrictedMaterialsUsed	KnewInfoConfidential
ValidAgreement	AgreementMade	AgreementInForce
AgreementMade	AgreedNotToDisclose	OutsiderDisclosuresRestricted
AgreementInForce	AgreementNotSpecific	WaiverOfConfidentiality
DisclosureInNegotiations		
BribeEmployee		
EmployeeSoleDeveloper		
AgreedNotToDisclose		
AgreementNotSpecific		
SecurityMeasures	Magnitude	
BroughtTools	Magnitude	
CompetitiveAdvantage		
SecretsDisclosedOutsiders	Magnitude	
VerticalKnowledge		
OutsiderDisclosuresRestricted		
NoncompetitionAgreement		
RestrictedMaterialsUsed		
UniqueProduct		
InfoReverseEngineerable		
InfoIndependentlyGenerated	Magnitude	
IdenticalProducts	Magnitude	
NoSecurityMeasures	Magnitude	
InfoKnownToCompetitors		
KnewInfoConfidential		
InvasiveTechniques		
WaiverOfConfidentiality		
InfoObtainableElsewhere	Magnitude	
InfoReverseEngineered		
Deception		
DisclosureInPublicForum		

with a default degree. Not all factors need to be ascribed magnitudes. Of the 26 factors in [5], 17 turned out to be adequately modelled as Boolean, but 9 were modelled with magnitudes. Even in HYPO, 10 of the 13 dimensions can adequately be modelled as Booleans [8], and Rigoni [22] recognises that there are these two types of factor and accordingly represents cases with a set of Booleans factors, as well as a set of factors with magnitude. Our program was implemented using SWI Prolog (Windows version).

It may be for some nodes (especially those with disjunctive children) that the acceptance conditions do not give an answer. For example there may be no evidence of competitive advantage offered, and hence there are no relevant facts for either side. Equally, there may be no mention of any disclosures. In such cases a default must be used. This is chosen to reflect the burden of proof. For example, it is generally uncontested that there was some competitive advantage from using the information, and this is usually simply accepted: thus it defaults to the plaintiff. Here the onus is on the defendant to show that the information provided no competitive advantage. In contrast, the existence of a confidential relationship must be demonstrated by the plaintiff, and so the relevant factors default to the defendant. In this way, the defaults can be used to assign the burden of proof, as revealed in the precedent cases.

3. Combining Factors

Examination of the structure used to represent the domain knowledge reveals that although each non-leaf node has two children, these children play a variety of roles for the different nodes. Consider for example the top of the tree, which represents the issues identified in the *Restatement*, which form the "logical model" used in [14]. At this level, we are dealing mainly with issues, such as *was the information misappropriated?* and *was the information a Trade Secret?*. These are Boolean (as discussed in [3]), found either for the plaintiff or the defendant, and the children are there either because the parent requires the satisfaction of multiple conditions (e.g. *SecretMisappropriated*), or because there are several conditions which enable the parent to be satisfied (e.g. *InfoMiasappropriated*). The very top level, however, represents a general rule with an exception: *TradeSecretMisappropriation* is normally satisfied if *SecretMisappropriated* but has an exception: if the case has the factor *EmployeeSoleDeveloper*, then *TradeSecretMisappropriation* is not satisfied. This represents a preference: the presence of factor *EmployeeSoleDeveloper* casts doubt on the ownership of the information and is taken to outweigh the considerations leading to *TradeSecretMisappropriation*: if the courts had decided otherwise in the relevant precedents, the exception would not appear.

As well as such preference based exceptions, we find exceptions which are motivated not by preferences and precedents, but by the very meanings of the terms. Consider *MaintainSecrecyOutsiders*. This is normally not satisfied, as one would expect, if we have *SecretsDisclosedOutsiders*. But there is an exception if we also have *OutsiderDisclosuresRestricted*. This is not a matter of preference: by making any disclosures to outsiders subject to restrictions, the plaintiff must be regarded as having made efforts to maintain secrecy with respect to outsiders, as a consequence of the meanings of the words involved.

Another kind of node is where a balance is struck between two competing factors [19]. This is illustrated by the node *ImproperMeans*, which requires that a balance be struck between the information having been obtained by *QuestionableMeans* and being *LegitimatelyObtainable*. It is, of course, possible that information which was obtainable legitimately was in fact obtained using questionable means. Thus the extent to which questionable means were used has to be weighed against the ease with which the information could have been obtained legitimately. This involves weighting the values concerned: do we want to encourage enterprise or strict adherence to ethical principles? In practice there has been a preference given to finding for the plaintiff if questionable means have been used, but this cannot be seen as absolute. If the information had been readily available on, say, Wikipedia, then some mild deception in obtaining the secret might be overlooked. In order to represent this relationship we ascribe weights to the values of the children involved. The values are taken from [15], which have also been used in later work such as [1]. An alternative set of values and ways of handling balances and trade offs using Boolean factors, can be found in [16].

Finally we have nodes at which conversion from children with magnitude to Boolean parents takes place. Here we are considering whether the extent of satisfaction represented by the children is *sufficient* to allow the parent to be *considered* satisfied. To achieve this, each value is associated with a threshold. Note that the thresholds can be set independently, since they are only required to be used in this way, not for comparison with one another. These thresholds receive a good deal of emphasis in the formal approaches of both [17] and [22]. The former divides the range into two factors, one a pro-plaintiff and one pro-defendant, with the point of division determined by precedents. Rigoni divides the range into a number of factors with a precedent determined "switching point" where the factors cease to be pro-plaintiff and become pro-defendant. These points correspond to our thresholds.

Acceptance conditions can thus take a variety of forms, depending on whether the children have magnitude, and how they are combined. Our program [10] uses the following acceptance condition types:

- **Conjunctions**: Used to provide multiple conditions all of which must be satisfied for the acceptability of a node. One or both conditions may have magnitude (represent degrees of acceptability). As in Fuzzy Logic [25], the minimum of the individual conjuncts is ascribed to the conjunction as a whole, enabling uniform treatment of Booleans and factors with magnitude.
- **Disjunctions**: Used to provide multiple conditions at least one of which must be satisfied for the acceptability of a node. One or both conditions may have magnitude (represent degrees of acceptability). As in Fuzzy Logic, the maximum of the individual disjuncts is ascribed to the disjunction.
- **Preference Based Exceptions**: Used to represent exceptions based on precedents expressing a preference between values. Preferences are based on values, taken from [15]. Should the value preference be changed, the exception disappears, since it is no longer sufficient to defeat the general case.
- **Definition Based Exceptions**: Used to represent exceptions based on precedents relying on word meanings. The exception depends on the meanings of the terms involved: no preference is required.

- **Comparison with a threshold**: This converts factors with magnitudes to Booleans. The result of conjoining (or disjoining) the children is compared to a threshold, and 1 or 0 is returned accordingly.
- **Balancing two factors using weights**. This is used to strike a balance between two features. It uses weights based on the values associated with the children and the extents to which the children are satisfied. This is similar to the method used in [15] to handle comparison between factors which may promote values to different extents.

In the program there were 37 nodes, of which 19 were disjunctions and 4 conjunctions. We had 4 preference based exceptions, 2 definitional exceptions and 7 nodes employing thresholds. 1 node expressed a trade off.

3.1. Values and Their Roles

What is particularly interesting about the above characterisation of the ways in which factors can combine, is the clarification it gives to the roles of values. Originally, as in [12], factors all promoted some value, and conflicts were resolved according to a value structure applicable to the whole domain. This uniformity was, however, shown to be inappropriate in [26], where it was shown that values could motivate not only whole rules (and the preferences between them), but also the inclusion of particular terms in the antecedents of rules, so as to enable the value to be given proper consideration. We now see that the picture is a little more complicated. The role of values is certainly to ensure that the various concerns they represent are properly considered. But their use to establish preferences is limited to the relatively few rules in which a balance must be struck to express a trade off, or a preference requires an exception to a general rule. In the first case we use weights established by precedents in the manner of [15], whereas in the second the preference is in the existence of the exception. A third role is to establish thresholds for determining whether a factor is satisfied to a sufficient extent. These thresholds do not use a preference ordering, since they are considered independently: they are, however, justified in terms of precedents[2] and may be set higher or lower to reflect the switching points manifest in the precedents. Thus, not only do values play several roles, but preferences between them are limited to a few nodes, and need not be consistent across the whole structure. This makes the reasoning more akin to Branting's reasoning with portions of precedent [13] than the holistic view taken by CATO and [12].

4. Statement Types

In [3] a number of different statement types were identified. Obviously the verdict must be Boolean: the court must decide for one party or the other. As we saw from the Restatement of Torts, the base level factors can have magnitude. The

[2] Some factors, like *SecretsDisclosedOutsiders* have a natural mapping to numbers, while others do not. For the latter, however an ordering on fact situations can be established (as with SecurityMeasures in [6]) and this order reflected in the magnitudes assigned.

implementation starts with the base level factors as input, so at some point the factors with magnitude have to be transformed to a Boolean. This must be done at, or before, the issues are reached.

The extents ascribed to the base level factors are decided by the court, and courts may disagree. Thus we find parts of the opinions discussing whether the lower court had correctly determined, for example, the extent to which the information was reverse engineerable, or was available elsewhere[3]. As we propagate the magnitudes assigned to base level factors up the tree, at some point the parent factor will need to become a Boolean, since issues are held to favour either the plaintiff, or the defendant. This means that if we have a branch ending in one or more base level factors with magnitude, at some point a threshold will need to be applied. This transition will not always occur at the border between abstract factors and issues, since some of the abstract factors, like some of the base level factors, may themselves be Boolean. However, the thresholds must be used to transform the factors with magnitude to Booleans when, or before, the issue level is reached. The use of thresholds is illustrated in the example given in section 5.

Note that there are two distinct points for the court to address: the *extent* to which the base level factor is satisfied, and the *threshold* that must be reached if the abstract factor is to be satisfied. If we do not allow factors to have magnitudes, these two points are conflated and both are decided by the analyst representing the cases: whereas they should be transparent and subject to explicit argumentation. They should kept distinct and both aspects decided by the court. In particular we often find minority opinions expressing disagreement as to the extent to which a factor is satisfied and, separately, as to whether the degree to which a factor is satisfied is sufficient to resolve the issue for a given party.

Transforming factors with magnitudes to Booleans is the primary role of thresholds, and it is essential. Examination of our program [10] reveals the importance of thresholds: they determine whether a factor is satisfied to a sufficient extent to enable an exception to be implied: without magnitudes any extent whatsoever could be considered sufficient. In the current implementation the threshold is dependent only on the value to which it relates, and the same threshold is used for all such tests relating to a given value. It remains possible, however, that different thresholds should be used to determine which party is favoured by an issue, and whether an exception should be applied. We leave this for future investigation, which will involve a careful analysis of actual opinions.

5. Reasoning with Factors with Magnitude

With the Boolean factors of CATO we can challenge a decision only by adding or removing a factor, or by changing a preference. With magnitudes we have the further options of:

[3] Of course, courts do not assign numbers to these extents, but they do order them, and this ordering can be mapped to numbers for computation. For example, Mr Justice Stevens says in *California v Carney* “It is perfectly obvious that the citizen has a much greater expectation of privacy concerning the interior of a mobile home than of a piece of luggage such as a footlocker”. It is, however, clear that the majority disagreed with Stevens when the mobile home was in use as a vehicle. This can be reflected in the magnitudes assigned in the different opinions.

- Increasing the degree of presence of a factor
- Decreasing the degree of presence of a factor
- Raising the threshold for factors relating to a particular value
- Lowering the threshold for factors relating to a particular value
- Adjusting the relative weights of factors

This provides a number of different ways in which decisions can be contested, whereas CATO allowed only for different preferences between the factors involved, and approaches such as [12] reduced all disputes to disagreements about value preferences.

We illustrate these options with the following hypothetical Trade Secrets case. Plaintiff in US produced a widget. This was a distinctive product, but something similar was manufactured in China (uniqueProduct, but less than 1). Drawings of the widget were kept in an unlocked drawer in the plaintiff's office (securityMeasures but less than 1). The defendant was in the plaintiff's office and was left alone. He searched the desk and looked at the drawings and photographed them on his phone (invasiveTechniques). Defendant claimed that the drawings confirmed his view that the product was reverse engineerable (infoReverseEngineerable but less than 1). Defendant also claimed that the information could have been obtained from the Chinese company (infoObtainableElsewhere but less than 1). Suppose we start with some initial parameters and facts (only non-zero base level factors are given). Relevant values are legitimate means (lm), questionable means (qm) and material worth (mw). Note that two values used in [15] (reasonable efforts and confidential agreement) do not need weights or thresholds since they appear only in exceptions:

```
weight(qm,1).  % Questionable methods and
weight(lm,1).  % Legitimate methods have equal weights
% Thresholds are all neutral between parties
threshold(lm,0.5). threshold(re,0.5). threshold(mw,0.5).
invasiveTechniques(1).
securityMeasures(0.6). % Security measures somewhat lax,
                       % but intended to keep info secret
uniqueProduct(0.8).    % Similar product, but not in US
infoReverseEngineerable(0.3). % Product considered rather
                              % hard to reverse  engineer
infoObtainableElsewhere(0.4). % Obtainable from Chinese firm
```

These facts will find for the plaintiff. We can now suggest arguments for modifying them:

- Perhaps the reverse engineerability should be considered more straightforward (after all the information had been independently developed in China). Increasing the magnitude to 0.6 (or greater: precision is not crucial) will enable a finding for the defendant
- Perhaps the threshold for reverse engineerability is too low: if the information is really readily ascertainable, why steal it? Raising the threshold to 0.8 will find again for the plaintiff.

- The security measures were rather poor: the drawer was not even locked. Decreasing the magnitude to 0.4 (or lower) will find for the defendant.
- But perhaps inventors should be given more protection: the defendant should not be rifling through the plaintiff's drawers. Lowering the threshold to 0.3 will restore the case for the plaintiff.
- Perhaps the balance between legitimate and questionable means is wrong: perhaps "all's fair in love, war and business". Raising the weight given to legitimate methods to 3 will find for the defendant.

6. Concluding Remarks

In this paper we have consolidated our current understanding of reasoning with legal cases using factors with magnitudes, by drawing out lessons we have learned by implementing the approach in the standard domain for factor based systems, US Trade Secrets as modelled in HYPO and CATO. In particular we have learned:

- That magnitudes are important to reflect the reasoning found in opinions. In particular they enable a variety of arguments which go beyond simply preferring one set of factors or values to another. These additional types of arguments have been identified in section 5. This requires that each value is given a threshold to represent the switching points of [22].
- We have confirmed the importance of the "switching points" as identified by the formal approaches of [17] and [22]. However, we have also identified another kind of argument not catered for in the papers: namely trade offs as discussed in [7] and [16]. This requires that values also be assigned weights to be used to strike an appropriate balance in the trade offs. Thus we require two sets of value related parameters: thresholds and weights.
- The key reason for values is to enable due consideration of various aspects the law is meant to address. We have identified three separate roles for values: establishing thresholds, motivating weights and justifying preference based exceptions. However, we have shown that values affect relatively few nodes (about one third in our particular application).
- We have shown that children in the factor hierarchy (ADF) do not contribute homogeneously to the acceptability of their parents: we have identified six different ways in which children are combined.

In sum, we have identified a diversity of types of reasoning, and arguments, required when considering legal cases, moving beyond the rather uniform reasoning found in previous factor and value based approaches. It shows the different ways in which factors combine, the roles played by values, and various argumentative options not available in a Boolean setting. It shows what is (e.g. thresholds), and what is not (e.g. trade offs), captured in the formalisations of [17] and [22]. We intend to use magnitudes in our on-going work with legal firms [4] and to use these lessons to inform and improve the planned future application of our approach to further domains.

References

[1] L Al-Abdulkarim, K Atkinson, and T Bench-Capon. Factors, issues and values: Revisiting reasoning with cases. In *Proceedings of the 15th ICAIL*, pages 3–12. ACM, 2015.

[2] L Al-Abdulkarim, K Atkinson, and T Bench-Capon. A methodology for designing systems to reason with legal cases using ADFs. *AI and Law*, 24(1):1–49, 2016.

[3] L Al-Abdulkarim, K Atkinson, and T Bench-Capon. Statement types in legal argument. In *Proceedings of JURIX 2016*, pages 3–12. IOS Press, 2016.

[4] L Al-Abdulkarim, K Atkinson, T Bench-Capon, S Whittle, R Williams, and C Wolfenden. Noise Induced Hearing Loss: An application of the Angelic methodology. In *Proceedings of JURIX 2017*, pages 79–88, 2017.

[5] V. Aleven. *Teaching case-based argumentation through a model and examples.* PhD thesis, University of Pittsburgh, 1997.

[6] K Ashley. *Modeling legal arguments: Reasoning with cases and hypotheticals.* MIT press, Cambridge, Mass., 1990.

[7] K Atkinson and T Bench-Capon. Dimensions and values for reasoning with legal cases. *Tech Report ULCS 17-004, Department of Computer Science, U of Liverpool*, 2017.

[8] T Bench-Capon. HYPO's legacy: introduction to the virtual special issue. *Artificial Intelligence and Law*, 25(2):205–250, 2017.

[9] T Bench-Capon and K Atkinson. Dimensions and values for legal CBR. In *Proceedings of JURIX 2017*, pages 27–32, 2017.

[10] T Bench-Capon and K Atkinson. Implementing factors with magnitude. In *Proceedings of COMMA 2018*, pages 449–450, 2018.

[11] T Bench-Capon and E Rissland. Back to the future: Dimensions revisited. In *Proceedings of JURIX 2001*, pages 41–52. IOS Press, 2001.

[12] T Bench-Capon and G Sartor. A model of legal reasoning with cases incorporating theories and values. *Artificial Intelligence*, 150(1-2):97–143, 2003.

[13] L Karl Branting. Reasoning with portions of precedents. In *Proceedings of the 3rd ICAIL*, pages 145–154. ACM, 1991.

[14] S Bruninghaus and K Ashley. Predicting outcomes of case based legal arguments. In *Proceedings of the 9th ICAIL*, pages 233–242. ACM, 2003.

[15] A Chorley and T Bench-Capon. An empirical investigation of reasoning with legal cases through theory construction and application. *AI and Law*, 13(3):323–371, 2005.

[16] Matthias Grabmair. *Modeling Purposive Legal Argumentation and Case Outcome Prediction using Argument Schemes in the Value Judgment Formalism.* PhD thesis, University of Pittsburgh, 2016.

[17] J Horty. Reasoning with dimensions and magnitudes. In *Proceedings of the 16th ICAIL*, pages 109–118. ACM, 2017.

[18] J Horty and T Bench-Capon. A factor-based definition of precedential constraint. *AI and Law*, 20(2):181–214, 2012.

[19] M Lauritsen. On balance. *Artificial Intelligence and Law*, 23(1):23–42, 2015.

[20] H Prakken, A Wyner, T Bench-Capon, and K Atkinson. A formalization of argumentation schemes for legal case-based reasoning in ASPIC+. *Journal of Logic and Computation*, 25(5):1141–1166, 2015.

[21] A Rigoni. An improved factor based approach to precedential constraint. *Artificial Intelligence and Law*, 23(2):133–160, 2015.

[22] Adam Rigoni. Representing dimensions within the reason model of precedent. *Artificial Intelligence and Law*, 26(1):1–22, 2018.

[23] E Rissland and K Ashley. A case-based system for trade secrets law. In *Proceedings of the 1st ICAIL*, pages 60–66. ACM, 1987.

[24] E Rissland and K Ashley. A note on dimensions and factors. *Artificial Intelligence and Law*, 10(1-3):65–77, 2002.

[25] L Zadeh. Fuzzy logic and approximate reasoning. *Synthese*, 30(3-4):407–428, 1975.

[26] T Zurek and M Araszkiewicz. Modeling teleological interpretation. In *Proceedings of the 14th ICAIL*, pages 160–168. ACM, 2013.

Legal Knowledge and Information Systems
M. Palmirani (Ed.)

doi:10.3233/978-1-61499-935-5-21

On Observing Contracts: Deontic Contracts Meet Smart Contracts

Shaun AZZOPARDI [a,1] and Gordon J. PACE [a,b] Fernando SCHAPACHNIK [c,2]
[a] *Department of Computer Science, University of Malta, Msida, Malta*
[b] *Centre for Distributed Ledger Technologies, University of Malta, Msida, Malta*
[c] *Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales, Departamento de Computación and ICC UBA-CONICET, Buenos Aires, Argentina*

Abstract. Smart contracts have been proposed as executable implementations enforcing real-life contracts. Unfortunately, the semantic gap between these allows for the smart contract to diverge from its intended deontic behaviour. In this paper we show how a deontic contract can be used for real-time monitoring of smart contracts specifically and request-based interactive systems in general, allowing for the identification of any violations. The deontic logic of actions we present takes into account the possibility of action failure (which we can observe in smart contracts), allowing us to consider novel monitorable semantics for deontic norms. For example, taking a rights-based view of permissions allows us to detect the violation of a permission when a permitted action is not allowed to succeed. A case study is presented showing this approach in action for Ethereum smart contracts.

Keywords. blockchain, smart contracts, contracts, deontic logic, monitoring

1. Introduction

Regulating the behaviour of interactive systems[3] has long been studied in a formal manner — from the formal semantics of contracts describing the expected modalities of behaviour to analysis techniques to enable automated verification of compliance of a system with such contracts. Deontic logics have been shown to be particularly effective in capturing the concepts behind such contracts, enabling them to be described as first-class logical objects which one can analyse, manipulate and transform. The deontic modalities of obligation and prohibition lend themselves easily to verification techniques, since through the observation of actions or the state of the system regulated by such a contract, one can easily ascertain whether or not an obligation or a prohibition has been violated. On the other hand, the notion of permission, if we take a rights-based view of it, is more problematic when it comes to verification, because an action that is permitted but (illegitimately) denied might be mistaken for one than never occurred. If a user is allowed to close an account but the bank does not allow it, the account remains open. From a (oversimplistic but common) point of view, there is no trace of the violation.

[1] Corresponding Author: Shaun Azzopardi, Department of Computer Science, University of Malta, Msida, Malta; E-mail: shaun.azzopardi@um.edu.mt.

[2] Partially supported by PICT-201-0112 and UBACyT 20020170100172BA.

[3] By interactive systems we refer to systems made up of different components interacting in such a manner that parts of their behaviour can be enabled or blocked by the other components.

1. *This contract is between ⟨buyer-name⟩, henceforth referred to as 'the buyer' and ⟨seller-name⟩, henceforth referred to as 'the seller'. The contract will hold until either party requests its termination.*
2. *The buyer is obliged to order at least ⟨minimum-items⟩, but no more than ⟨maximum-items⟩ items for a fixed price ⟨price⟩ before the termination of this contract.*
3. *Notwithstanding clause 1, no request for termination will be accepted before ⟨contract-end-date⟩. Furthermore, the seller may not terminate the contract as long as there are pending orders.*
4. *Upon enactment of this contract, the buyer is obliged to place the cost of the minimum number of items to be ordered in escrow.*
5. *Upon accepting this contract, the seller is obliged to place the amount of ⟨performance-guarantee⟩ in escrow, otherwise, if only a partial amount is placed, the seller is obliged to place the rest by a time period at the buyer's discretion.*
6. *While the contract has not been terminated, the buyer has the right to place an order for an amount of items and a specified time-frame as long as (i) the running number of items ordered does not exceed the maximum stipulated in clause 2; and (ii) the time-frame must be of at least 24 hours, but may not extend beyond the contract end date specified in clause 2.*
7. *Upon placing an order, the buyer is obliged to ensure that there is enough money in escrow to cover payment of all pending orders.*
8. *Before termination of the contract, upon delivery the seller must receive payment of the order.*
9. *Upon termination of the contract, if either any orders were undelivered or more than 25% of the orders were delivered late, the buyer has the right to receive the performance guarantee placed in escrow according to clause 5.*

Figure 1. A legal contract regulating a procurement process.

In open systems (i.e. systems whose internal structure is not fully known), such as ones where one of the intervening parties is a human, verification is problematic and the best one can hope for is to observe or monitor the interaction between the parties and deduce whether or not it is compliant with the contract. Even through simple monitoring, obligations and prohibitions are easily analysed for compliance or violation. Observed behaviour, however, does not always give sufficient information to enable a decision of whether or not there was a violation of a permission.

In this paper we limit ourselves to request-based interactive systems — systems in which a number of parties may initiate interaction (hence *request-based*), but for each interaction, one of the parties has the power to allow or reject such a request (hence *interactive*). In such systems, we observe that as long as the *intention* to initiate an action, and its success or otherwise can be observed, we can also monitor for violation of permissions i.e. if the user has permission to press a button and attempts to do so, but the backend rejects the request, then we can flag that the backend has violated the user's permission.

In particular, we focus on smart contracts as an instance of such a request-based interactive system, and show how a deontic logic with permission can be fully monitored on smart contracts on a platform such as Ethereum [19]. Since smart contracts have been proposed as executable implementations enforcing real-life contracts, the ability to ensure that an implementation on a distributed ledger technology (or blockchain), matches the description of the contract is a highly desirable feature. In order to illustrate our approach, we consider a smart contract regulated by a procurement contract, shown in Figure 1, as adapted from [1]. The contract regulates the behaviour of a buyer and a seller, with the contract setting minimum and maximum order counts, a contract end date, and ensures that there is enough money in escrow to ensure at least the minimum amount of orders is financially covered, along with a monetary performance guarantee from the seller to the buyer.

The paper starts off by describing similar and related approaches in the literature in Section 2. We then go on to define the semantics of an action-based deontic logic for request-based interactive systems in Section 3 and show how smart contracts written for the Ethereum platform in Solidity can implement sound and complete monitors of such contracts in Section 4. The approach is illustrated on a procurement smart contract in Section 5. We discuss the viability of our approach and conclude in Section 6.

2. Related Work

The interaction of deontic logic and smart contracts is an area of recent interest. Idelberger et al. [9] analyse pros and cons of logic-based smart contracts (as opposed to the more common procedural-code based ones). They discuss the ability to monitor logic-based contracts as an advantage but do not explore possible implementations, which is the topic of this article. With the blooming of smart contracts questions have begun to arise around the issue of how to be sure that the smart contract is correct (eg, [11,1]).

Indeed deontic-based contract monitoring is a well known topic, with proposals for using deontic languages to monitor business processes [15], building monitors out of logic contract specifications [4], or for using certain monitoring architectures [12], even using blockchain as base technology [18][4], amongst others.

However, these proposals takes a lightweight approach to handling permissions. For instance Modgil et al. [12] consider the monitoring of permissions, but only flag when they are exercised and not when denied by another party, as we do (see Section 3).

In [2] we presented the notion of a contract between interacting parties where permissions on one party could be denied by the other by not cooperating. That work, however, flagged violations when, at a given state, one party did not provide the needed synchronisation for a shared action, independently of whether the permission-bearer party attempted the action or not — if *A* is permitted to borrow a book from *B*, and *B*'s specification does not include the book-lending action, a violation is still flagged. However, one may argue that perhaps *A* never intended to exercise his or her right to borrow a book at the time — should that still count as a violation of the permission? Moreover that work requires *a priori* a full specification of the possible party actions, which is not available in the kind of interactive systems we consider here.

What was missing there, and this work addresses is the notion of *attempts*. It has been long established that a logic that only predicates over actions may be insufficient for certain deontic applications that refer to other concepts such as *intention* (e.g. [8]), much in the same sense as BDI agents. Later on, Lorini and Herzig [10] took a step further by proposing a logic that includes (and differentiates among) *intention* and *attempt*. While an *intention* is a mental state that has no direct observable effect, an *attempt* is an observable action, which may or may not succeed. Attempts permit to separate the state that the action is intended to bring about from whether the agent took some transition with a visible output. As an example, Bob might try to sell his car, yet failed to do so. The end state is Bob keeping the car and not having extra money, yet he engaged in the action of trying to sell it.

This work was built on prior philosophical traditions on what it means for an agent to intend to do an action. For space reasons we concentrate only on the two that are more

[4] Note that [18] discusses the use of blockchain as a decentralized way to monitor business processes but not the monitoring of the smart contract itself as we do.

relevant to our work: Sellars [17] poses that *"A tries to do α"* if and only if *A exercises actions to bring about the state α without being sure whether they will succeed or not.* Schroeder [16] however, reverses the word *"trying"* to the observation of the action being done by another agent: B perceives A trying an action a if B is not certain that action a will succeed. The labelling of something as an attempt is done by an observer, not by the exerciser of the action. Others, such as [3], have presented an alternative notion of attempt having to do with the probability p of bringing about a certain state α by certain action a. Interestingly, as the chance of $\neg\alpha$ is $1-p$, then a becomes both an attempt at α and an attempt at $\neg\alpha$ at the same time.

Our work follows the tradition of Lorini and Herzig (and others) of having actions that might succeed or not, but in our case the failure is an observation of an action attempted and denied by the system (which in turn, might be a permission violation).

3. An Operational Semantics for Deontic Contracts

In this section we present a simple action-based deontic logic for request-based interactive systems based on approaches from the literature such as [7,2], which will be used to synthesise smart contract violation monitors. By *request-based interactive systems* we mean systems in which only one party initiates interaction (hence *request-based*), but the other party may allow or reject any such request (hence *interactive*). Although we could have adopted other existing deontic logics which deal with interaction, we choose to define our own in order to obtain a clean translation process, which can then be adapted to other existing logics by adding the features as required.

Well-formed formulae in the logic will be used to denote a contract specifying the expected behaviour of the user interacting with a service. The interaction with the system will take the form of actions which are initiated by the user and accepted or rejected by the service. Formally, the logic is defined on an alphabet of event names Σ — we will use variables a, b to range over Σ. Given event name a, we will write a^Y to denote a successful attempt to perform event a, and a^N to denote that event a was attempted but was rejected. We will write $\mathbb{\Sigma}$ to denote the set of all possible observations over event names Σ: $\mathbb{\Sigma} \overset{df}{=} \{a^Y, a^N \mid a \in \Sigma\}$. We use variables x, y to range over $\mathbb{\Sigma}$.

For example, in the procurement contract, actions may include *login*, *requestExtension*, etc. Consider a contract where the user is allowed to request an extension only once. Then the first time the user would call *requestExtension*, she or he would expect it to succeed, but it may fail thereafter (i.e. any further requests are rejected).

The logic itself will allow the expression of obligations, prohibitions and permissions over these actions from the point of view of the user. It is worth noting that in a request-based interactive system, the obligations and prohibitions of invoking actions are the responsibility of the user (the caller), whereas satisfying the permissions which the user has is the responsibility of the underlying service. The syntax of the logic is the following:

$$\mathscr{C} ::= \top \mid \bot \mid \mathscr{O}(a) \mid \mathscr{F}(a) \mid \mathscr{P}(a) \mid [\psi]\mathscr{C} \mid \mathscr{C} \,\&\, \mathscr{C} \mid \mathscr{C};\mathscr{C} \mid \mathscr{C} \rhd \mathscr{C} \mid \mathrm{rec}\, X.\mathscr{C} \mid X$$

The logic contains the trivially satisfied ($\top$) and violated ($\bot$) contracts and ways of expressing obligation to perform an action ($\mathscr{O}(a)$), prohibition ($\mathscr{F}(a)$) and permission ($\mathscr{P}(a)$). Contracts can also be guarded by a condition on the actions which are performed ($[\psi]C$) — with the contract trivially holding if the condition does not hold. Contracts can be combined via conjunction ($C_1 \,\&\, C_2$), sequential composition ($C_1;C_2$) and reparation

($C_1 \rhd C_2$), with the last operator indicating that if the first contract is violated, the second comes into force. Finally, one can use recursion (rec $X.C$) to express repeating contracts. We will use variables C, C_1, C_2 to range over contracts. In the rest of the paper, we will assume that all recursion is well-formed in that it is guarded by a condition or a deontic modality e.g. rec $X.\mathscr{O}(a);X \,\&\, [hasNotPaid]X$ is fine, but not rec $X.(X \,\&\, \mathscr{P}(a));\mathscr{F}(b)$.

Consider a contract clause in the procurement example, which states that the user is permitted to terminate the contract once a delivery has been made. This can be expressed as: rec $X.[noDelivery]X \,\&\, [madeDelivery]\mathscr{P}(endContract)$.

We define a syntactic equivalence relation over contract formulae:

$$\begin{array}{llll} C \,\&\, \top \equiv C & \bot \,\&\, C \equiv \bot & \top \;;\, C \equiv C & \text{rec}\, X.C \equiv C[X\backslash \text{rec}\, X.C] \\ \top \,\&\, C \equiv C & C \,\&\, \bot \equiv \bot & \bot \rhd C \equiv C & \end{array}$$

As typically done in operational semantics, this relation will be used to simplify the presentation of the semantics, and will be applied in between the steps defined by the operational semantics. The only equivalence worth commenting on is the one for recursion, which states that free instances of the recursion variable may be replaced by the recursive formula itself. If unguarded recursion were allowed, this rule can be applied infinitely often. However, with guarded recursion, we note that if the equivalence relation is applied from left to right, we can prove that the process of applying the relation until no further reductions are available acts as a total and deterministic function. If, starting from C, the result obtained is C', we will write $C \hookrightarrow C'$. We say that a contract C is normalised if $C \hookrightarrow C$.

Note that the logic allows for multiple obligations to be in place at the same time, in order to enable conformance to such concurrent obligations, we give a semantics in which a set of concurrent actions (an *action set*) is observed at once. This approach follows other similar work in the literature (e.g. [14,2]). We will use variables A, B to range over action sets (i.e. in 2^{Σ})[5].

We can now define the semantics of the logic as a relation such that $C \xrightarrow{A} C'$ if contract $C \in \mathscr{C}$ evolves to contract $C' \in \mathscr{C}$ if an action set $A \subseteq \Sigma$ is observed. The full semantics is illustrated in Figure 2.

It is worth noting a number of features of the semantics of this logic: (i) obligations can be satisfied upon an attempt by the user to perform the action, irrespective whether or not it was approved, and dually, prohibition is violated by an attempt by the user to perform the action even if it is rejected; (ii) permission[6] is violated if an attempt to perform the action occurs but is rejected by the system; (iii) conditions are assumed to be predicates over the set of actions which the user has performed. However, one may adopt alternative choices (e.g. one may choose that an attempt to perform a forbidden action which is rejected not to be a violation) without changing the results in the rest of the paper. Finally, note that although at first sight some interactions may appear missing (e.g. semantics of recursion and the triggering of a reparation) they are handled by the equivalence relation presented earlier.

Given a trace of action-sets $T \in (2^{\Sigma})^*$, we write $C \overset{T}{\Rightarrow} C'$ to indicate that contract C evolves to contract C' following trace T, and applying the syntactic equivalences after

[5] In this article we use standard notation of 2^X to denote the power set of X, and X^* for finite lists over set X.

[6] We use the term permission to define a modality that other authors in the Hohfeldian tradition call a *right*, in the sense that what is permitted to a party posses a burden of compliance on the other party (or the system in our case).

$$\frac{}{\top \xrightarrow{A} \top} \qquad \frac{}{\bot \xrightarrow{A} \bot}$$

$$\frac{}{\mathscr{O}(a) \xrightarrow{A} \top}\ \{a^Y, a^N\} \cap A \neq \emptyset \qquad \frac{}{\mathscr{O}(a) \xrightarrow{A} \bot}\ \{a^Y, a^N\} \cap A = \emptyset$$

$$\frac{}{\mathscr{F}(a) \xrightarrow{A} \top}\ \{a^Y, a^N\} \cap A = \emptyset \qquad \frac{}{\mathscr{F}(a) \xrightarrow{A} \bot}\ \{a^Y, a^N\} \cap A \neq \emptyset$$

$$\frac{}{\mathscr{P}(a) \xrightarrow{A} \top}\ a^N \notin A \qquad \frac{}{\mathscr{P}(a) \xrightarrow{A} \bot}\ a^N \in A$$

$$\frac{C_1 \xrightarrow{A} C_1' \qquad C_2 \xrightarrow{A} C_2'}{C_1 \,\&\, C_2 \xrightarrow{A} C_1' \,\&\, C_2'} \qquad \frac{C_1 \xrightarrow{A} C_1'}{C_1 ; C_2 \xrightarrow{A} C_1' ; C_2} \qquad \frac{C_1 \xrightarrow{A} C_1'}{C_1 \rhd C_2 \xrightarrow{A} C_1' \rhd C_2}$$

$$\frac{C \xrightarrow{A} C'}{[\psi]C \xrightarrow{A} C'}\ \psi(A) \qquad \frac{}{[\psi]C \xrightarrow{A} \top}\ \neg\psi(A)$$

Figure 2. Semantics of deontic contract logic.

each step, defined as the smallest relation such that[7]: (i) if $C \hookrightarrow C'$, then $C \overset{\langle\rangle}{\Rightarrow} C'$; and (ii) if $C \hookrightarrow C' \xrightarrow{A} C'' \overset{T}{\Rightarrow} C'''$, then $C \overset{A:T}{\Longrightarrow} C'''$.

Although the contract semantics allows transitions over action sets, most services are accessed as a sequence of individual events. We choose to resolve concurrent norms by a special event that marks a time unit, using it to group single events into action sets. This may correspond to a temporal period (e.g. a day, with an event marking the occurrence of midnight) or otherwise (e.g. the time unit may be a login session, and the logging out event marks the end of the time unit). We will now show how such a small-step semantics can be defined using the action set semantics in order to be able to deal with systems such as smart contracts in which events happen individually.

The small-step semantics will process one action at a time, with the special event *tick* (not in Σ) to denote the end of the current time unit. We will write Σ_{tick} to denote the observable events Σ augmented by *tick*: $\Sigma_{tick} \overset{df}{=} \{tick\} \cup \Sigma$, and we will use variables α, β to range over this set. The small step semantics will be of the form $C_A \overset{\alpha}{\rightsquigarrow} C'_{A'}$ to denote that upon receiving event α, if $A \subseteq \Sigma$ were the events observed since the last tick event, then contract $C \in \mathscr{C}$ will evolve to contract $C' \in \mathscr{C}$ with the accumulated observed events now being $A' \subseteq \Sigma$:

$$\frac{}{C_A \overset{x}{\rightsquigarrow} C_{\{x\} \cup A}} \qquad \frac{C \xrightarrow{A} C' \qquad C' \hookrightarrow C''}{C_A \overset{tick}{\rightsquigarrow} C''_{\emptyset}}$$

As we did before with action set semantics, for a trace $t \in \Sigma^*_{tick}$ we define $C_A \overset{t}{\rightsquigarrow\!\!\!\!\rightsquigarrow} C'_{A'}$ to be the transitive closure of the small step semantics.

Another reasonable choice here is to immediately flag a violation. Note also how a tick event is essential, since at some point the actions performed need to be evaluated — an obligation to do an action without an explicit or implicit time limit is not an obligation.

[7] In the rest of the paper we use standard notation for traces: $\langle\rangle$ denotes the empty list, $x : xs$ denotes the list consisting of item x followed by list xs, $xs +\!\!+\, ys$ denotes the concatenation of lists xs and ys and $\mathrm{items}(xs)$ denotes the set of elements appearing in list xs.

We also define a relation $\flat \in (2^{\Sigma})^* \leftrightarrow \Sigma^*_{tick}$, such that action-set trace $T \in (2^{\Sigma})^*$ is related with singleton-action trace $t \in \Sigma^*_{tick}$ if the actions in t split by tick actions are the same as the actions sets appearing in T. This relation is defined as the least relation satisfying the following:

$$\begin{aligned}
\langle\rangle \flat \langle\rangle &\overset{df}{=} \mathit{true} \\
(A : T) \flat (t_1 \mathbin{+\!\!+} \langle \mathit{tick} \rangle \mathbin{+\!\!+} t_2) &\overset{df}{=} A = \mathrm{items}(t_1) \wedge \mathit{tick} \notin A \wedge T \flat t_2
\end{aligned}$$

Using this relation, we can formulate and prove the correctness of the small step semantics with respect to the action set semantics.

Theorem 1. Given a normalised contract $C \in \mathscr{C}$, an action-set sequence leads to a violation if and only if equivalent singleton-action traces also lead to violations:

$$\forall T \in (2^{\Sigma})^*,\, t \in \Sigma^*_{tick} \cdot T \flat t \implies (C \overset{T}{\Rightarrow} \bot) \Leftrightarrow (C_{\emptyset} \overset{t}{\rightsquigarrow} \bot_{\emptyset})$$

4. Monitoring Deontic Contracts in a Blockchain

Smart contracts are concrete instances of request-based interactive systems, with the intended purpose of serving as executable implementations of real-life contracts. By being executed on blockchains there is a certain degree of transparency and dependability. Although their out-of-the box immutable transaction record can be used for analysing past behaviour, such an approach does not allow for online monitoring with real-time and on-chain reactions to violations.

In smart contracts, actions are requested by the user (typically by calling an entrypoint or function of the smart contract's interface), which triggers specific business logic. In this manner, since a smart contract is intended to serve as a regulated environment within which the parties interact, the smart contract ensures that the action is allowed (according to the counterpart real-life contract) and if so carries out the expected behaviour. For example, if a buyer attempts to place an order when there is not enough money in escrow to cover it, then the smart contract should not let the order to be placed successfully (see Clause 7 of the procurement contract).

A smart contract thus enforces a deontic contract if it only allows compliant sequences of actions. This is not always possible — for instance when certain blockchain actions should signal an off-blockchain event which the smart contract should faithfully record (e.g. that a delivery was made). However, within the domain of what is observable on the blockchain, the challenge in monitoring that the smart contract behaviour matches that of the deontic contract lies in the semantic's rules which deal with attempts, particularly the rules for permission (where an attempt which is not allowed to go through results in a violation) and prohibition (where an attempt to perform a forbidden action is sufficient to trigger a violation). In other words, being able to monitor not just successful actions (i.e. a^Y) but also failed attempts (i.e. a^N). In this section we will look at how monitoring for attempts (as opposed to actual performance of the action) can be done on the Ethereum blockchain, and we discuss how we can instrument smart contracts written using the Solidity language [6] to monitor deontic contracts.

In Solidity smart contracts, the code defines an interface which parties may invoke in order to trigger particular behaviour specified in the code itself. The code itself may trigger a failure which results in the whole invocation to be dismissed (including any effects performed before the failure). For example, the code below shows a function used to terminate a contract which uses the `require` command to fail and abort execution if invoked by anyone other than the buyer or the seller:

```
function terminateContract() public {
  require(msg.sender == seller || msg.sender == buyer); // Can only be done by the seller or buyer
  ...
}
```

The fact that such failure triggers a revert of the smart contract state, as though the invocation never happened, is problematic for monitoring, since this means that if one monitors for a failure in an online manner, failure will also obliterate the observation itself. However, Ethereum provides different modalities for function calls, allowing for encapsulating function calls within a `call`, capturing such failures and signalling the outcome using a boolean value indicating success or failure of the invocation (instead of the failure). One can thus pre-process a smart contract to package its functions in order to detect both success and failure of invocations, and which are used to keep a record of events between `tick` events:

```
function terminateContractWithFailure() public {
    if (this.call(bytes4(keccak256("terminateContract()")))) {
        addSuccessfulAction(Action.TerminateContract);
    } else {
        addFailedAction(Action.TerminateContract);
    }
}
```

This approach allows us to listen to smart contract events, and record the ones that are relevant to the deontic contract to be used according to the small-step semantics of the contract language. The gathering of events corresponds to the first rule of the small-step semantics, while by encoding the deontic contract's semantics e.g. using a finite-state automaton or a symbolic automaton which we can easily instrument a smart contract with [5], we can then trigger a transition as per the second rule of the small-step semantics. That is, when `tick` is called we attempt to transition from the current state with the events that happened since the last tick:

```
function tick() public {
    ...
    transition();
    if (isViolating(currentState))
        doSomethingUponViolation();
    resetActions();
}
```

When a *tick* is invoked, the big step semantics are thus invoked and if a mismatch between the actions and the deontic contract is identified, appropriate action takes place[8]. How triggering of the `tick` function takes place depends on the contract itself, as discussed before. Instead of using the *tick* event to mark explicit time periods, we can also let it be triggered in other ways, for instance by agreement between parties (e.g. acknowledging that both will not perform any more actions in the current round), or by just one of the parties after a certain amount of time has elapsed since the last tick (e.g. as a signal that the party is ready to move to the next stage).

5. Case Study

Let us reconsider the procurement contract from Figure 1. In Figure 3, we show how the contract can be formalised in our logic. Here events do not just reflect actions of a party,

[8] How to deal with malfunctioning smart contracts is a challenge in itself, and is largely still an open problem, although an initial attempt at addressing this can be found in [13].

C1. $\mathscr{P}(\mathit{TerminateContractUnlessOtherwiseForbidden})$

C2. $\mathscr{F}(\mathit{TeminateContractWithItemsNotBetweenMinAndMaxItems})$

C3. $\mathscr{F}(\mathit{TerminateContractBeforeEndTimestamp})\,\&\,\mathscr{F}(\mathit{TerminateContractBySellerAndWithPendingOrders})$

C4. $\mathscr{F}(\mathit{EnactmentWithLessThanCostOfMinimumItems})$

C5. $\mathscr{F}(\mathit{AcceptContractWithLessThanGuarantee}) \rhd \mathscr{O}(\mathit{SendRestOfGuarantee})$

C6. $[\mathit{ContractNotTerminated}]\mathscr{P}(\mathit{OrderWithLessThanMaxItemsAndDeliveryTimeLessThan24HrsAndBeforeEnd})$

C7. $\mathscr{F}(\mathit{OrderWithLessThanEnoughMoneyForPendingOrders})$

C8. $[\mathit{ContractNotTerminated}]\mathscr{F}(\mathit{DeliveryWithPaymentLessThanCost})$

C9. $[\mathit{TerminateContractWithPendingOrdersOr25PercentLateOrders}]\mathscr{O}(\mathit{SendGuaranteeToBuyer})\,\&$
$[\mathit{TerminateContractWithoutPendingOrdersAnd25PercentLateOrders}]\mathscr{O}(\mathit{SendGuaranteeToSeller})$

$$\mathit{ProcurementContract} \overset{df}{=} \mathrm{rec}\ X.[\neg\psi]((C4\,\&\,C5);X)\ \ \&\ \ [\psi](\mathrm{rec}\ Y.(C1\,\&\,C2\,\&\,C3\,\&\,C6\,\&\,C7\,\&\,C8\,\&\,C9);Y)$$
$$\text{where } \psi \overset{df}{=} \mathit{EnactmentAndSellerAcceptanceWithEnoughInEscrow}$$

Figure 3. Formalisation of each of the clauses in Figure 1.

but rather actions happening in some smart contract state. It is worth noting here that the choice of meaning of the tick action is crucial, and one has to ensure that no party is able to unilaterally trigger this action to gain advantage, e.g. quickly invoking a tick so that the other party violates a pending obligation. For the sake of our case study, we implement a third party activated tick action, but there are other solutions which could be implemented, such as a unilateral time-bounded tick, in which either party can activate a tick action if at least a certain amount of time has passed since the last tick.

Both the contract and its monitoring based on the techniques outlined in Section 4 have been implemented on Ethereum in Solidity[9], thus achieving (i) a smart contract which should implement the logic of the legal contract; and (ii) a monitor which observes whether the behaviour of the parties and the logic of the smart contract actually complies with the legal contract as written in the formal deontic logic.

It should be noted that monitoring for compliance does not come for free. Execution of code in smart contracts on platforms like Ethereum require payment for so-called *gas*. The monitoring logic we add to a contract adds to the gas required when calling the original functions in such a contract. This increase (as a percentage of the original cost) is highly dependant on (i) the original smart contract; and (ii) the number of actions received between successive ticks. However, to get an idea of the magnitude of the extra cost we have done some previous studies at evaluating overheads in instrumenting monitors on Solidity smart contracts in [1]. It is worth mentioning that (i) given the critical nature of many smart contracts, the additional cost (in many cases being in the order of 10–20% of the original cost) is typically worth paying for; and (ii) given prices of ether at the time of writing the gas added is negligible in actual cost.

6. Conclusions

In computer science circles, one frequently finds conflation of contracts with specifications, mainly because obligations and prohibitions readily translate into standard specification languages. In contrast, permission has no direct correspondence in many specification languages, particularly linear time logics. In this paper, we have argued that in request-based interactive systems, respecting permission (or at least one form of permission) corresponds to respecting attempts to perform the permitted action or achieve the

[9] See `https://github.com/shaunazzopardi/deontic-monitoring-solidity/`

permitted state. Furthermore, in certain systems, attempts and their success (or otherwise) are observable events, which enables the definition of a trace semantics encompassing permission, thus allowing runtime monitoring of deontic contracts. This monitoring approach is applied to smart contracts (in our case study, to a procurement contract) in order to enable flagging when the prescribed behaviour and the behaviour of parties diverge.

One issue we have not addressed in our approach is directed modalities e.g. which party is obliged to perform an action, which would allow us to assign blame for violations. The current semantics can easily be extended to tag the modalities with parties and handle violations in a manner similar to the one we used earlier in [2].

References

[1] Shaun Azzopardi, Joshua Ellul, and Gordon J. Pace. Monitoring smart contracts: CONTRACTLARVA and open challenges beyond. In *the 18th International Conference on Runtime Verification*, 2018.

[2] Shaun Azzopardi, Gordon J. Pace, Fernando Schapachnik, and Gerardo Schneider. Contract automata - an operational view of contracts between interactive parties. *Artif. Intell. Law*, 24(3):203–243, 2016.

[3] Jan M. Broersen. Modeling attempt and action failure in probabilistic STIT logic. In Toby Walsh, editor, *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Spain, 2011*, pages 792–797. IJCAI/AAAI, 2011.

[4] María-Emilia Cambronero, Luis Llana, and Gordon J. Pace. A calculus supporting contract reasoning and monitoring. *IEEE Access*, 5:6735–6745, 2017.

[5] Joshua Ellul and Gordon J. Pace. Runtime verification of ethereum smart contracts. In *Proceedings of the First International Workshop on Blockchain Dependability*, 2018.

[6] Ethereum. Solidity v0.4.24 Language Documentation. `https://solidity.readthedocs.io/en/v0.4.24/`, 2018. [Online; accessed 09-September-2008].

[7] Stephen Fenech, Gordon J. Pace, and Gerardo Schneider. Clan: A tool for contract analysis and conflict discovery. In Zhiming Liu and Anders P. Ravn, editors, *ATVA*, volume 5799 of *Lecture Notes in Computer Science*, pages 90–96. Springer, 2009.

[8] Guido Governatori, Vineet Padmanabhan, Antonino Rotolo, and Abdul Sattar. A defeasible logic for modelling policy-based intentions and motivational attitudes. *Logic Journal of the IGPL*, 17(3), 2009.

[9] Florian Idelberger, Guido Governatori, Régis Riveret, and Giovanni Sartor. Evaluation of logic-based smart contracts for blockchain systems. In *Rule Technologies. Research, Tools, and Applications - 10th International Symposium, RuleML 2016, NY, USA, 2016. Proceedings*, pages 167–183, 2016.

[10] Emiliano Lorini and Andreas Herzig. A logic of intention and attempt. *Synthese*, 163(1):45–77, 2008.

[11] Daniele Magazzeni, Peter McBurney, and William Nash. Validation and verification of smart contracts: A research agenda. *IEEE Computer*, 50(9):50–57, 2017.

[12] Sanjay Modgil, Nir Oren, Noura Faci, Felipe Meneguzzi, Simon Miles, and Michael Luck. Monitoring compliance with e-contracts and norms. *Artif. Intell. Law*, 23(2):161–196, 2015.

[13] Gordon J. Pace, Joshua Ellul, and Christian Colombo. Contracts over smart contracts: Recovering from violations dynamically. In *8th International Symposium on Leveraging Applications of Formal Methods, Verification and Validation (ISoLA 2018)*, 2018.

[14] Cristian Prisacariu and Gerardo Schneider. A formal language for electronic contracts. In *FMOODS*, volume 4468 of *Lecture Notes in Computer Science*, pages 174–189. Springer, 2007.

[15] Shazia Wasim Sadiq, Guido Governatori, and Kioumars Namiri. Modeling control objectives for business process compliance. In *Business Process Management, 5th International Conference, BPM 2007, Brisbane, Australia, September 24-28, 2007, Proceedings*, pages 149–164, 2007.

[16] Severin Schroeder. The concept of trying. *Philosophical Investigations*, 24(3):213–227, 2001.

[17] Wilfrid Sellars. Science and metaphysics: Variations on kantian themes. 1968.

[18] Ingo Weber, Xiwei Xu, Régis Riveret, Guido Governatori, Alexander Ponomarev, and Jan Mendling. Untrusted business process monitoring and execution using blockchain. In *Business Process Management - 14th International Conference, (BPM) 2016, Rio de Janeiro, Brazil, September 18-22, 2016. Proceedings*, pages 329–347, 2016.

[19] Gavin Wood. Ethereum: A secure decentralised generalised transaction ledger. *Ethereum Project Yellow Paper*, 151:1–32, 2014.

Legal Knowledge and Information Systems
M. Palmirani (Ed.)

doi:10.3233/978-1-61499-935-5-31

Legal Representation and Reasoning in Practice: A Critical Comparison

Sotiris BATSAKIS [c,a,1], George BARYANNIS [a], Guido GOVERNATORI [b],
Ilias TACHMAZIDIS [a] and Grigoris ANTONIOU [a]
[a] *School of Computing and Engineering, University of Huddersfield, UK*
[b] *Data61, CSIRO, Brisbane, Australia*
[c] *Technical University of Crete, Greece*

Abstract. Representation and reasoning over legal rules is an important application domain and a number of related approaches have been developed. In this work, we investigate legal reasoning in practice based on three use cases of increasing complexity. We consider three representation and reasoning approaches: (a) Answer Set Programming, (b) Argumentation and (c) Defeasible Logic. Representation and reasoning approaches are evaluated with respect to semantics, expressiveness, efficiency, complexity and support.

Keywords. Legal knowledge representation, Legal reasoning, Normative reasoning

1. Introduction

Representation of legal rules and reasoning over them is a critical application area since laws and regulations are used in almost all human activities. Furthermore, corpora pertaining to laws and regulations are typically complex and enormous in size, thus experts are usually needed in order to apply legal reasoning in everyday life. Even for experts, this is a laborious, time-consuming, error-prone and costly process, which is why legal reasoning was one of the first application domains of artificial intelligence since its emergence, and more so after the proliferation of expert systems during the 1980s.

The logic-based structure of laws and regulations and the considerable benefits of automating the representation and reasoning processes has led to extensive research towards this direction. The importance and visibility of such research in recent years can be exemplified with the emergence of the field of Regulatory Technology (RegTech [1]). However, the complexity of the legal domain has made progress difficult. In civil law legal systems, which are the primary focus of this paper, several issues have been raised, such as: (a) the ambiguity of natural language used to express laws, which leads to different interpretations; (b) the existence of different conflicting rules that are applicable on the same case; and (c) the need to do reasoning in light of new information.

In this work, we identify and compare three distinct approaches that are capable of addressing some of these issues through features such as non-monotonic reasoning and

[1] Corresponding Author: Sotiris Batsakis, School of Computing and Engineering, University of Huddersfield, UK; E-mail: s.batsakis@hud.ac.uk.

conflict resolution. These are: Answer Set Programming, Argumentation and Defeasible Logic. Three use cases of increasing complexity are used for the comparison: a scenario on the presumption of innocence, licensing contractual clauses for the evaluation of a product and the reporting regulations on new drugs imposed by the US Food and Drug Administration (FDA). The main contributions of this work are the investigation of practical considerations of legal representation and reasoning through the aforementioned real-world use cases and a comparative evaluation of existing approaches and related tools used for these tasks. These, in turn, can assist in identifying strengths and weaknesses of various approaches, allowing for more informed choices on the appropriate solutions to different legal representation and reasoning problems.

This paper is organised as follows. A concise summary of efforts related to legal representation and reasoning is presented in Section 2. A description of the three use cases follows in Section 3. Representation and reasoning results using the three selected approaches are presented in Section 4 and they are critically evaluated in Section 5. Section 6 concludes and discusses directions for future research.

2. Background and Related Work

Early attempts at realising legal reasoning involved representing legislation in the form of Horn logic programs, subsequently extended with negation as failure, as in Sergot et al.'s seminal work on the British Nationality Act [2]. However, the treatment of double negation and counterfactual conditionals (e.g. “if it didn't rain”) proved problematic. Moreover, exceptions in legislation are modeled explicitly by negative conditions in the rules [3], which is more suitable for self-contained and stable legislation but may require some level of rewriting whenever previously unknown exceptions (or chains of exceptions) are introduced.

Following the advent of the Semantic Web and the introduction of the OWL family of languages, several research efforts focused on examining whether description logics are a suitable candidate for representing and reasoning about legislation. A prime example is HARNESS [4], which shows that well-established sound and decidable description logic reasoners such as Pellet can be exploited for legal reasoning, if, however, a significant compromise in terms of expressiveness is made.

A common issue that arises when using classical or description logics in legal representation and reasoning is the fact that they are monotonic: logical consequences cannot be retracted, once entailed. This is in contrast to the nature of law, where legal consequences have to adapt in light of new evidence and conflicts between different regulations must be accounted for and resolved. Therefore, it is natural to employ non-monotonic logic for the purposes of legal reasoning. The Defeasible Logic framework [5] has been applied in a legal reasoning setting due to its simplicity and flexibility and the fact that several efficient implementations exist. In this framework, rules can either behave in the classical sense (*strict*), they can be defeated by contrary evidence (*defeasible*), or they can be used only to prevent conclusions (*defeaters*).

The notions of permission and obligation are inherent in normative reasoning but are not explicitly defined in traditional logic systems; deontic logic was introduced to serve this purpose. Permission and obligation are represented by modal operators and are connected to each other through axioms and inference rules. While there has been some philosophical criticism on deontic logic due to its admission of several paradoxes (e.g. the gentle murderer), deontic modalities have been introduced to various logics to make them more suitable for normative reasoning. For instance, Governatori et al. [6] show how the aforementioned Defeasible Logic framework can be extended to model beliefs, intentions, obligations and permissions.

Legal reasoning, at its core, is a process of argumentation, with opposing sides attempting to justify their own interpretation, with appeals to precedent, principle, policy and purpose, as well as the construction of and attack on arguments [7]. AI and law research has addressed this with models that are based on Dung's [8] influential work, such as ASPIC+ [9], which offers means of producing argumentation frameworks tailored to different needs in terms of the structure of arguments, the nature of attacks and the use of preferences.

As argued by Brewka et al. [10], the high complexity of argumentation processes, including legal reasoning, makes them a prime application target for Answer Set Programming, a non-monotonic logic programming formalism, specifically aimed towards search problems of NP or higher complexity.

3. Use Cases

3.1. *Use case 1: Presumption of Innocence*

The first use case demonstrates the importance of the semantics of default inference in legal representation and reasoning. Presumption of innocence is the principle that one is considered innocent until proven guilty. Consider the following scenario in a legal case:

- Evidence A suggests that the defendant is not responsible.
- Evidence B suggests that the defendant is responsible.
- Sources for evidence A and B are equally reliable.
- According to the underlying legal system a defendant is presumed innocent (i.e. not guilty) unless responsibility has been proven (without any reasonable doubt).
- If the defendant is found to be innocent, they are entitled to compensation.

Given both evidences A and B, ambiguity exists as it is unclear whether the defendant should hold any responsibility. Thus, under this situation and with the underlying legal system, we should conclude that the defendant is not guilty and is entitled to compensation. However, if we allow the ambiguity with regard to responsibility to propagate, then the defendant's guiltiness should also be considered ambiguous; hence an undisputed conclusion cannot be drawn.

3.2. *Use case 2: Smart Contracts in Blockchain Systems*

The second use case, extracted from [11], is a typical example of legal reasoning related to contracts on blockchain systems and illustrates the intricacies inherent in the representation of the notions of permission and obligation. The following licensing contractual clauses are assumed for the evaluation of a product:

- Article 1. The Licensor grants the Licensee a licence to evaluate the Product.

- Article 2. The Licensee must not publish the results of the evaluation of the Product without the approval of the Licensor; the approval must be obtained before the publication. If the Licensee publishes results of the evaluation of the Product without approval from the Licensor, the Licensee has 24 hours to remove the material.
- Article 3. The Licensee must not publish comments on the evaluation of the Product, unless the Licensee is permitted to publish the results of the evaluation.
- Article 4. If the Licensee is commissioned to perform an independent evaluation of the Product, then the Licensee has the obligation to publish the evaluation results.
- Article 5. This licence will terminate automatically if Licensee breaches this Agreement.

Article 2 is of particular interest since it contains a reparation clause. Suppose that the licensee publishes the evaluation results without authorisation, then removes them within 24 hours. Then, according to Article 2, the licence to use the product still holds.

3.3. Use case 3: US FDA New Drugs Records and Reports Regulations

The third use case is much larger than the previous two; it includes legislation from the United States Electronic Code of Federal Regulations, specifically Title 21 - Food and Drugs, Part 310 - New Drugs, Subpart D - Records and Report. Specifically, the regulations define: (a) the reporting requirements for manufacturers, packers and distributors and (b) information on various life-threatening, serious or unexpected adverse drug experiences that have been reported in case safety reports. The legislation contains 8 paragraphs, with some of them containing as many as 12 requirements and can be found in [12]. For example, paragraph (c)(1) of the legislation requires that manufacturers, packers or distributors must report to FDA each adverse drug experience no later than 15 calendar days from initial receipt of complaint; however, this requirement is lifted if there is no reasonable possibility that the drug caused the adverse experience.

4. Representation and Reasoning

In this section, we present rule-based representations for the three use cases in the previous section, as well as reasoning results using these representations. The decision on candidate approaches was based on two fundamental criteria for practical legal reasoning: support for preferences over rules and availability of reasoning tools. Given these criteria, we selected the following: (1) Answer Set Programming (ASP), specifically Disjunctive Logic Programs with Inheritance [13]; (2) Argumentation, specifically structured argumentation within the ASPIC+ framework [9]; and (3) Defeasible Logic, specifically the extended version that supports deontic modalities [6].

4.1. Representation and Reasoning using ASP

ASP is an expressive form of Logic Programming based on the stable model semantics and supporting disjunction, among others. An extension introduced in [13] allows the expression of rule priorities by organising rules in inheritance networks and is implemented by the DLV system [14]. Given two objects (rules or rule sets) o_1 and o_2, $o_2 : o_1$ denotes that o_2 is more specific than o_1 and overrides it, in case of conflict. Use case 1 is encoded as follows:

```
r1 {responsible:- evidenceA.}       r2 {-responsible:- evidenceB.}
r4 {-guilty:- 1=1.}                 r3:r4 {quilty:-responsible.}
r6 {-compensation :- 1=1.}          r5:r6 {compensation:- -guilty.}
evidenceA.                          evidenceB.
```

Tautology 1=1 is required because objects in inheritance networks have to comprise rules and not facts. Running the DLV system using this encoding as an input yields no results (no stable models) because of the contradiction caused by rules r1 and r2; this is left unresolved and, as a result, no undisputed conclusions can be derived.

For use case 2, since deontic modalities are not supported natively by any ASP representation, we can only represent obligation explicitly within predicate names, while permission on a literal is encoded as a negated obligation on the literal's negation ($Oa \equiv \neg P \neg a$). The legislation of use case 2 is encoded as follows:

```
:- obl_publish, obl_not_publish.
art10 {obl_not_use :- 1=1.}
art11:art10 {-obl_not_use :- hasLicence.}
art21 {obl_not_publish :- 1=1.
       obl_remove :- publish, obl_not_publish.}
art22:art21 {-obl_not_publish :- hasLicense, hasApproval.}
art31 {obl_not_comment :- 1=1.}
art32:art31 {-obl_not_comment :- -obl_not_publish.}
art40:art21{obl_publish :- hasLicence,isCommissioned.
      -obl_not_publish :- hasLicence,isCommissioned.}
art51:art11 {obl_not_use :- violation.}
art52:art40,art22 {obl_not_publish :- violation.
                   -obl_publish :- violation.}
hasLicence.    publish.
```

The last two facts represent the case where the licensee has published results without having approval or being commissioned. Performing reasoning results in the stable model {hasLicence, publish, -obl_not_use, obl_not_publish, obl_remove, obl_not_comment}, which includes the permission to use, the obligations not to publish or comment and the obligation to remove published results. For use case 3, only the encoding for paragraph (c)(1) is presented due to space limitations. The full encoding contains 96 rules and can be found at https://github.com/gmparg/JURIX2018, along with all other encodings in this paper.

```
r5{obl_MPD_report_electronically_adverse_drug_exper_in_15_days
:- MPD_report_15_day_Postmarketing_Alert_reports.}
r6:r5{-obl_MPD_report_electronically_adverse_drug_exper_in_15_days
:-MPD_report_15_day_Postmarketing_Alert_reports_nocause.}
```

Note that if we are limited to the standard ASP format (ASP-Core-2), used by systems such as clingo [15] and DLV2 [16], then to compensate for the lack of support for preferences over rules, extra negation-as-failure literals need to be included in rule bodies; for instance, to express in use case 2 that permission to use applies to a licensee unless they have violated the agreement, we use the rule `-obl_not_use :- hasLicence, not violation.` Reformulated encodings using negation as failure can also be found at https://github.com/gmparg/JURIX2018.

4.2. Representation and Reasoning using Argumentation

In the case of argumentation, using an abstract framework would offer only a coarse-grained level of representation which is not enough for fully representing a legal document. Structured argumentation, on the other hand, as is the case of argumentation according to ASPIC+, is capable of representing legislation since it supports encoding knowledge bases with axioms and premises. We used TOAST [17], an online ASPIC+ implementation. Use case 1 is encoded as follows (presented using TOAST's argumentation theory structure and syntax):

```
Premises: evidenceA; evidenceB;
Assumptions: default;
Rules:
[r1] evidenceA => responsible;      [r2] evidenceB => ~responsible;
[r3] responsible => guilty;         [r4] default=> ~guilty;
[r5] ~guilty => compensation;       [r6] default => ~compensation;
Rule Preferences: [r4] < [r3]; [r6] < [r5];
```

Note that rules in TOAST must have bodies, hence the inclusion of the default predicate. The computed extension contains only asserted facts *evidenceA* and *evidenceB* since arguments from r1 and r2 defeat each other, which is equivalent to the reasoning outcome in ASP. For use case 2, deontic modalities again need to be encoded explicitly, since deontic extensions like the one in [18] have not been implemented yet:

```
Premises: hasLicence; publish;
Assumptions: default;
Preferences: default < hasLicence; hasLicence < violation;
Rules:
[r1] default => obl_not_use;
[r2] hasLicence => ~obl_not_use;
[r3] default => obl_not_publish;
[r4] publish, obl_not_publish => obl_remove;
[r5] hasLicence, hasApproval => ~obl_not_publish;
[r6] default => obl_not_comment;
[r7] ~obl_not_publish => ~obl_not_comment;
[r8] hasLicence, isCommissioned => obl_publish;
[r9] hasLicence, isCommissioned => ~obl_not_publish;
[r10] violation => obl_not_use;
[r11] violation => obl_not_publish;
Rule Preferences:
[r1] < [r2]; [r3] < [r5]; [r6] < [r7]; [r3] < [r8]; [r3] < [r9];
[r2] < [r10]; [r5] < [r11]; [r8] < [r11]; [r9] < [r11];
```

```
Contrariness:
obl_publish-obl_not_publish;
```

The added preferences prevent implicit attacks caused by rules with contradictory heads; also, a contrariness relation is needed for every pair of arguments that are contradictory. Similarly to ASP, the computed extension includes the arguments for permission to use, obligation not to publish or comment and obligation to remove published results. The encoding of use case 3 contains the same amount of rules as the DLV encoding.

4.3. Representation and Reasoning using Defeasible Logic

For Defeasible Logic, we opted to use SPINdle [19], since it is the only reasoner that natively supports deontic extensions, which are necessary for use case 2. The encoding for use case 1 is as follows:

```
r1: evidenceA => responsible         r2: evidenceB => -responsible
r3: responsible => guilty            r4:  => -guilty
r5: -guilty => compensation          r6:  => -compensation
>> evidenceA                         >> evidenceB
r3 > r4                              r5 > r6
```

Note that Defeasible Logic allows expressing defeasible rules without bodies (presumptions). The reasoning output includes the following (with +/-D meaning definitely provable/not provable and +/-d meaning defeasibly provable/not provable: -D responsible(X), -D -responsible(X), -d responsible(X), -d -responsible(X), -D guilty(X), -D -guilty(X), +d -guilty(X), -d guilty(X), -D compensation(X), -D -compensation(X), +d compensation(X), -d -compensation(X). This means that, in contrast to DLV and TOAST, SPINdle infers the defeasible conclusion that the defendant is not guilty. This is due to the default ambiguity blocking behaviour in SPINdle; if we switch to ambiguity propagation, then the result will contain no derived conclusions, as in DLV and TOAST.

For use case 2, we use deontic modalities (Permission/Obligation) both on literals and rules, by adding [P] or [O] before the literal or at the end of a rule label. Syntax `label[O]: A=>B` is equivalent to `label: A=>[O]B`. The encoding is as follows:

```
Art10[O]: => -use(X)
Art11[P]: hasLicence(X) => use(X)
Art21[O]: => -publish(X)
Art21rep[O]: [O]-publish(X), publish(X) => remove(X)
Art22[P]: hasLicence(X), hasApproval(X) => publish(X)
Art31[O]: => -comment(X)
Art32[P]: [P]publish(X) => comment(X)
Art40[O]: hasLicence(X), isCommissioned(X) => publish(X)
Art40p[P]: hasLicence(X), isCommissioned(X) => publish(X)
Art51[O]: violation(X) => -use(X)
Art52[O]: violation(X) => -publish(X)
Art11 > Art10          Art22 > Art21          Art32 > Art31
Art40 > Art21          Art51 > Art11
Art52 > Art40          Art52 > Art40p         Art52 > Art22
>> hasLicence(X)       >> publish(X)
```

Table 1. Comparison of the presented approaches

	ASP	Argumentation	Defeasible Logic
Expressiveness	Negation As Failure, Disjunction	Negation As Failure	Presumptions, Deontic Modalities
Inconsistency Handling	Propositional	Direct/Indirect [9]	Paraconsistent [20]
Support (Reasoning)	DLV	TOAST	SPINdle
Complexity	Σ_2^P, Π_2^P [13]	P [21]	Linear [6]

Rule Art21rep implements the reparation clause of Article 2, while Art40p explicitly derives permission from obligation. The output includes the following defeasible derivations: +d [O]-comment(X), -d [P]comment(X), +d [O]-publish(X), -d [O]publish(X), -d [P]publish(X), +d [P]use(X), -d [O]-use(X), +d [O]remove(X). These correspond to the same obligations and permissions that are derived from the DLV and TOAST encodings. The encoding of use case 3 contains the same amount of rules as the previous encodings.

5. Critical Evaluation

In this section, we present a comparative evaluation of the three approaches based on the experience gained by representing the three use cases and reasoning with the represented knowledge. We begin with a summary of common characteristics. All three approaches operate under the semantics of closed-world assumption. In terms of expressiveness, they all support non-monotonic reasoning, classical negation and preferences between rules. All approaches are efficient is terms of reasoning time. Even in use case 3 all reasoners return results within two or three tenths of a second. Finally, in all three approaches, while a number of reasoning tools have been developed, there is currently no tool support for representation, to the best of our knowledge. This means that any legislation has to be manually encoded in the respective languages supported by the reasoners.

Table 1 summarises the differences among ASP, argumentation and Defeasible Logic, in what concerns legal knowledge representation and reasoning. In terms of expressiveness, Defeasible Deontic Logic is clearly the most efficient approach, supporting presumptions, permissions and obligations, which make representation an easier, faster and less error-prone process. In contrast, the other approaches have to compensate by expressing modalities explicitly within predicate names and by using default predicates or tautological rule bodies to model presumptions. In cases where legislation changes relatively often, adapting ASP and argumentation encodings would require comparatively more time and effort. It should be noted that while disjunction (supported by ASP) and negation as failure (supported by ASP and TOAST), do not feature in any of the use cases we examined, they may still be useful in encoding different types of legislation.

An important difference that manifests in use case 1 is the way the three approaches handle inconsistency. The pair of conflicting rules on responsibility lead to unsatisfiability in ASP, because DLV (and other reasoners) assume the standard propositional definition of consistency. TOAST behaves in the same way, with the computed grounded extension containing only the asserted facts, due to ASPIC+'s definition of direct consistency. SPINdle, on the other hand, infers the defeasible conclusion of not guilty, due to the paraconsistent nature of Defeasible Logic; this can be argued to be closer to common sense reasoning. Note that Defeasible Logic offers the flexibility to support both cases:

the aforementioned derivation is for the default ambiguity blocking semantics, while the alternative ambiguity propagation would not derive any undisputed conclusion.

Differences in inconsistency handling appear also in the case of contradicting facts in the knowledge base. ASP reasoners conclude that there is no stable model, while both argumentation and Defeasible Logic approaches allow the contradicting facts to be used in further inferences, ensuring however that no form of the principle of explosion is possible. Both of these ways to address inconsistent input can be useful, depending on whether the point is to detect inconsistencies or to derive results despite them. However, it should be noted that both TOAST and SPINdle lack some form of acknowledgment of the existence of inconsistency (e.g. a warning message), which would be fundamental in addressing cases where inconsistency is not desired, but is the result of mistaken input.

In terms of reasoning support, there is at least one stable reasoner available for each approach. SPINdle is the most recent, released in 2014, while TOAST and DLV were released in 2012. Though there are more recent ASP reasoners (e.g. clingo and DLV2), they do not support preferences over rules. There are also several other alternatives for argumentation (e.g. the Tweety libraries at http://tweetyproject.org/). In what concerns complexity, all algorithms implemented by the aforementioned reasoners are of polynomial time complexity, albeit at different points in the polynomial hierarchy: Defeasible Logic is at the lowest level (linear), followed by argumentation (P) and then ASP (NP). A more detailed comparison of ASPIC+ and Defeasible Logic can be found in [22].

6. Conclusions and Future Research Directions

We have presented a practical demonstration of how existing approaches and tools can be used for legal representation and reasoning. In terms of representation, all approaches are capable of encoding standard elements of legislation, though with different levels of efficiency depending on their expressive power. Reasoning results can vary, especially when conflicting rules without preferences or conflicting facts are included; in such cases particular attention is needed in order to employ the most appropriate approach (and provide the proper encoding) that corresponds to the intended behaviour.

Interesting research directions that can be derived from the presented critical comparison include: (a) improving existing reasoners in terms of reporting inconsistencies and support for deontic modalities; (b) exploring much larger use cases that may require some form of large-scale reasoning, as discussed in [23]; and (c) providing tool support for representation on top of the existing reasoners by exploiting research in legal natural language processing.

Acknowledgements

The authors wish to thank Robert A. Kowalski for his insightful comments on legal reasoning using logic programming. This work has been supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 690974 for the project MIREL: MIning and REasoning with Legal texts.

References

[1] P. Mann, RegTech: The Emergence of the Next Big Disruptor, *International Banker* (2017). https://internationalbanker.com/finance/regtech-emergence-next-big-disruptor.

[2] M.J. Sergot, F. Sadri, R.A. Kowalski, F. Kriwaczek, P. Hammond and H.T. Cory, The British Nationality Act as a Logic Program., *Commun. ACM* **29**(5) (1986), 370–386.

[3] R. Kowalski and A. Burton, WUENIC - A Case Study in Rule-Based Knowledge Representation and Reasoning., in: *JSAI-isAI Workshops*, M. Okumura, D. Bekki and K. Satoh, eds, Lecture Notes in Computer Science, Vol. 7258, Springer, 2011, pp. 112–125.

[4] S. Van de Ven, J. Breuker, R. Hoekstra and L. Wortel, Automated Legal Assessment in OWL 2., in: *JURIX*, E. Francesconi, G. Sartor and D. Tiscornia, eds, Frontiers in Artificial Intelligence and Applications, Vol. 189, IOS Press, 2008, pp. 170–175.

[5] G. Antoniou, D. Billington, G. Governatori and M.J. Maher, A Flexible Framework for Defeasible Logics., in: *AAAI/IAAI*, H.A. Kautz and B.W. Porter, eds, AAAI Press / The MIT Press, 2000, pp. 405–410.

[6] G. Governatori, F. Olivieri, A. Rotolo and S. Scannapieco, Computing Strong and Weak Permissions in Defeasible Logic., *J. Philosophical Logic* **42**(6) (2013), 799–829.

[7] H. Prakken and G. Sartor, Law and logic: A review from an argumentation perspective., *Artif. Intell.* **227** (2015), 214–245.

[8] P.M. Dung, On the Acceptability of Arguments and Its Fundamental Role in Nonmonotonic Reasoning, Logic Programming, and n-Person Games, *Artificial Intelligence* **77**(2) (1995), 321–357.

[9] S. Modgil and H. Prakken, The ASPIC+ framework for structured argumentation: a tutorial, *Argument & Computation* **5**(1) (2014), 31–62. doi:10.1080/19462166.2013.869766.

[10] G. Brewka, M. Diller, G. Heissenberger, T. Linsbichler and S. Woltran, Solving Advanced Argumentation Problems with Answer-Set Programming, in: *AAAI*, 2017.

[11] G. Governatori, F. Idelberger, Z. Milosevic, R. Riveret, G. Sartor and X. Xu, On legal contracts, imperative and declarative smart contracts, and blockchain systems, *Artificial Intelligence and Law* (2018).

[12] US Government, Records and reports concerning adverse drug experiences on marketed prescription drugs for human use without approved new drug application, 2014. http://bit.ly/eCFR310_305.

[13] F. Buccafurri, W. Faber and N. Leone, Disjunctive Logic Programs with Inheritance, *Theory Pract. Log. Program.* **2**(3) (2002), 293–321.

[14] N. Leone, G. Pfeifer, W. Faber, T. Eiter, G. Gottlob, S. Perri and F. Scarcello, The DLV System for Knowledge Representation and Reasoning, *ACM Trans. Comput. Logic* **7**(3) (2006), 499–562.

[15] M. Gebser, R. Kaminski, B. Kaufmann, M. Ostrowski, T. Schaub and P. Wanko, Theory Solving Made Easy with Clingo 5, in: *Technical Communications of the 32nd International Conference on Logic Programming (ICLP'16)*, M. Carro and A. King, eds, OASIcs, Vol. 52, Schloss Dagstuhl, 2016, pp. 1–15.

[16] M. Alviano, F. Calimeri, C. Dodaro, D. Fuscà, N. Leone, S. Perri, F. Ricca, P. Veltri and J. Zangari, The ASP System DLV2, in: *Logic Programming and Nonmonotonic Reasoning*, M. Balduccini and T. Janhunen, eds, Springer International Publishing, 2017, pp. 215–221.

[17] M. Snaith and C. Reed, TOAST: Online ASPIC+ implementation, in: *Proc. of the 4th International Conference on Computational Models of Argument (COMMA 2012)*, B. Verheij, S. Szeider and S. Woltran, eds, Frontiers in Artificial Intelligence and Applications, Vol. 245, IOS Press, 2012.

[18] L.W.N. van der Torre and S. Villata, An ASPIC-based legal argumentation framework for deontic reasoning., in: *COMMA*, S. Parsons, N. Oren, C. Reed and F. Cerutti, eds, Frontiers in Artificial Intelligence and Applications, Vol. 266, IOS Press, 2014, pp. 421–432. ISBN 978-1-61499-436-7.

[19] H.-P. Lam and G. Governatori, The Making of SPINdle., in: *RuleML*, G. Governatori, J. Hall and A. Paschke, eds, Lecture Notes in Computer Science, Vol. 5858, Springer, 2009, pp. 315–322.

[20] M.J. Maher, A Model-Theoretic Semantics for Defeasible Logic., in: *Paraconsistent Computational Logic*, H. Decker, J. Villadsen and T. Waragai, eds, Datalogiske Skrifter, Vol. 95, 2002, pp. 67–80.

[21] P.E. Dunne and M. Wooldridge, Complexity of Abstract Argumentation, in: *Argumentation in Artificial Intelligence*, I. Rahwan and G.R. Simari, eds, Springer-Verlag, 2009, pp. 85–104.

[22] H.-P. Lam, G. Governatori and R. Riveret, On ASPIC+ and Defeasible Logic., in: *COMMA*, P. Baroni, T.F. Gordon, T. Scheffler and M. Stede, eds, Frontiers in Artificial Intelligence and Applications, Vol. 287, IOS Press, 2016, pp. 359–370. ISBN 978-1-61499-686-6.

[23] G. Antoniou, G. Baryannis, S. Batsakis, G. Governatori, L. Robaldo, G. Siragusa and I. Tachmazidis, Legal Reasoning and Big Data: Opportunities and Challenges, in: *MIREL 2018 Workshop*, 2018. https://doi.org/10.29007/tkmv.

Legal Knowledge and Information Systems
M. Palmirani (Ed.)

doi:10.3233/978-1-61499-935-5-41

A Question Answering System on Regulatory Documents

Diego COLLARANA [a,c,1], Timm HEUSS [b], Jens LEHMANN [a,c], Ioanna LYTRA [a,c], Gaurav MAHESHWARI [a,c], Rostislav NEDELCHEV [a,c] Thorsten SCHMIDT [b] and Priyansh TRIVEDI [a,c]

[a] *Enterprise Information Systems, Fraunhofer IAIS, Germany*
[b] *PricewaterhouseCoopers GmbH, Germany*
[c] *Smart Data Analytics Group, University of Bonn, Germany*

Abstract. In this work, we outline an approach for question answering over regulatory documents. In contrast to traditional means to access information in the domain, the proposed system attempts to deliver an accurate and precise answer to user queries. This is accomplished by a two-step approach which first selects relevant paragraphs given a question; and then compares the selected paragraph with user query to predict a span in the paragraph as the answer. We employ neural network based solutions for each step, and compare them with existing, and alternate baselines. We perform our evaluations with a gold-standard benchmark comprising over 600 questions on the MaRisk regulatory document. In our experiments, we observe that our proposed system outperforms other baselines.

Keywords. Question Answering; Reading Comprehension; Regulatory Domain

1. Introduction

Users of legal or regulatory documents usually access information through a keyword based search over a collection of related documents. Information Retrieval (IR) techniques can improve the search results by making use of synonyms, ontologies, and word embeddings, which allows an intelligent lookup rather than a simple keyword-based search. Still this approach requires the commitment of the legal expert in examining the retrieved documents and/or excerpts in order to find the right information providing answers to her queries. There is, hence, an increasing interest in developing systems able to provide accurate and precise answers to a user query, enabling a more intuitive, fast, and accurate access to information. One way for achieving this is the use of a question answering (QA) systems on such documents which can provide direct answers to users' questions by highlighting text spans in the target documents.

Question Answering is an important topic of research in the natural language processing field and and has seen major advances in the last years. Research in this area dates back to the 1960s with recent popular milestones being the development of IBM

[1] Corresponding Author: Enterprise Information Systems, Fraunhofer IAIS, Schloss Birlinghoven 53757 Sankt Augustin Germany; E-mail: diego.collarana.vargas@iais.fraunhofer.de.

Watson [4] and various conversational agents such as Alexa, Google Assistant, Siri, etc. Recent reading comprehension datasets like [12,10] further accelerated the growth of the field, by providing largescale general-domain datasets, enabling the use of complex neural network based approaches [16,7]. However, porting these techniques to domain-specific applications is non-trivial, and only few solutions [11,2] have been proposed for QA over legal documents. Amongst them, to the best of our knowledge, no comparable approach or system exists for the regulatory domain. This is mainly due to the fact that the specific domain – the regulatory domain – poses several challenges which need to be overcome in order to design and implement an effective QA system: (1) *Existence of long documents*: documents in the regulatory domain are usually long, consisting of tens or hundreds of pages, increasing the search space for potential answers; (2) *Structure of documents*: generally, regulatory documents have a rather complex structure including sections, subsections, paragraphs; (3) *Use of domain-specific language*: regulatory documents use a domain-specific vocabulary and a complex language which is difficult to be interpreted by machines. Further, the meaning of the words and the contexts often differs from the general domain or even across different regulatory documents; (4) *Lack of training data in the domain*: although several benchmarks for the general domain are available (e.g., SQuAD [12], MS MARCO [10], etc.), domain-specific datasets are very rare. Therefore, the key to a robust QA system over regulatory texts is a system that can comprehend both coarse and fine-grained information (across long documents) and adapt to the domain-specific vocabulary and complex language used in such documents.

In this paper, we propose a QA system over the regulatory domain which retrieves the paragraph and the precise span of the document as the answers to a user query. It consists of two major modules, namely *paragraph selection module*, which finds the relevant paragraph in the document given a question, and *answer selection module* which given a paragraph points to the exact span in the document. We show empirically, that our proposed approach can be effective for retrieving the answer span and outperforms traditional IR techniques. Further, we employ transfer learning techniques to increase the model performance, given the lack of training data in our specific domain. The proposed approach has been evaluated with question-answer pairs for the English translation of the MaRisk regulations document[2]. We tested different models and configurations of the proposed approach which allowed us to compare alternative QA pipelines.

The remainder of the paper is structured as follows. We introduce the difference between a traditional IR approach and a QA system based on reading comprehension through a motivating example in Section 2. The proposed approach is described in detail in Section 3 and the evaluation results are discussed in Section 4. Section 5 discusses the related work and, finally, Section 6 summarizes our key conclusions.

2. Motivating Example

In this section, we introduce a example of a question and answers based on the MaRisk regulatory document which defines Minimum Requirements for Risk Management for banks, insurances, and companies financially trading in Germany. The 62-page long MaRisk document in English consists of 2 main parts (General Part and Special Part),

[2] https://www.bafin.de/SharedDocs/Veroeffentlichungen/EN/Meldung/2014/meldung_140815_marisk_uebersetzung_en.html, last accessed on 2018-09-18

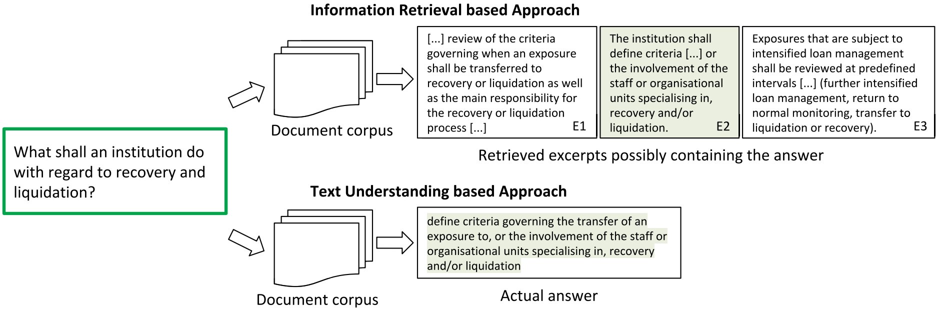

Information Retrieval based Approach
What shall an institution do with regard to recovery and liquidation?
Document corpus
[...] review of the criteria governing when an exposure shall be transferred to recovery or liquidation as well as the main responsibility for the recovery or liquidation process [...] E1
The institution shall define criteria [...] or the involvement of the staff or organisational units specialising in, recovery and/or liquidation. E2
Exposures that are subject to intensified loan management shall be reviewed at predefined intervals [...] (further intensified loan management, return to normal monitoring, transfer to liquidation or recovery). E3
Retrieved excerpts possibly containing the answer
Text Understanding based Approach
Document corpus
define criteria governing the transfer of an exposure to, or the involvement of the staff or organisational units specialising in, recovery and/or liquidation
Actual answer

Figure 1. Comparison of IR approach for retrieving answers to natural language questions to the QA approach based on reading comprehension. While with the first, we retrieve text passages which appear to be "similar" to the question, in the second approach the text excerpt is semantically mapped to the question.

64 sections and subsections having several paragraphs among which many are supported by annotations. It has approximately 24,000 words with many of them in German. A few representative questions are the following: "Where shall it be ensured that a trader can enter trades only under his/her own trader ID?", "Why are reverse stress tests performed?", and "What shall an institution do with regard to recovery and liquidation?". The spectrum of questions which can be asked is very broad (what-, how-, why-questions, etc.) and the answers may vary from a few words to a sentence, several sentences, or a paragraph. Figure 1 visualizes two alternative approaches for retrieving an answer from a document corpus given a question expressed in natural language: a traditional IR approach and a QA approach based on reading comprehension which is being investigated in the current work. While the first will return possible excerpts containing answer(s) to the user's question, the latter will find the actual answer to the initial question. For instance, given the exemplary question *"What shall an institution do with regard to recovery and liquidation?"*, based on the similarity of the question to several paragraphs in the MaRisk document, several excerpts with high confidence are being retrieved, with some of them (E1 and E3) not being related to the intent of the question. However, a reading comprehension based method is expected to interpret the question and find the actual answer, based on a deeper, albeit implicit understanding of the document.

3. Proposed Approach

To answer questions on long regulatory text documents we propose a two-step approach: 1. *Paragraph selection (PS)*: The purpose of this step is to find the most relevant paragraphs of a long document so that (costly) state-of-the-art reading comprehension techniques can be efficiently applied on the extracted paragraph(s). 2. *Answer Selection (AS)*: Given a paragraph retrieved by Step 1, the aim of this step is to find the exact word span which provides the answer to the question. The following subsections describe the details of paragraph and answer selection.

3.1. Paragraph Selection

The first phase of the pipeline involves selecting the relevant paragraph given a question. To do so, we employ an approach described in [17]. The proposed model, StarSpace

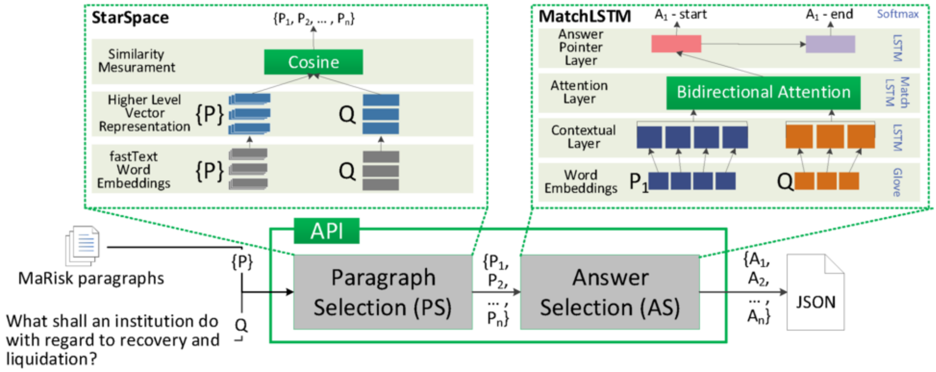


Figure 2. A two-step question answering approach for regulatory documents comprising: (1) Paragraph Selection and (2) Answer Selection.

(i) uses negative sampling to minimize margins between related entities, and maximizes margins between queries and irrelevant text, and (ii) is able to compare unrelated entities. This is accomplished by using a hierarchical structure of features to describe its entities, allowing to “embed all the things”.

In our implementation, words in both, queries, and documents are numerically represented by embeddings provided by fastText [8]. To generate positive pairs, we broke each of the documents into words. Then we generated n-grams of different sizes (n=3,4,5,6) that we used as a “pseudo-query” for each document. For a wider understanding of difficulty of the task, we also employ a Lucene index in parallel to select the paragraphs. The performance of both approaches can be found in Section 4.

3.2. *Answer Selection*

In order to find the exact span of the answer in the paragraph, retrieved by the Paragraph Selection module, we use the architecture proposed by [16]. The architecture consists of two major layers, a Match-LSTM layer and an Answer-Pointer layer. The Match-LSTM layer employs a word-by-word cross-attention to form a new weighted version of the paragraph representation based on the input question. This new representation indicates the degree of matching of each word in the paragraph to that of the question. This new weighted representation is then combined with the original question vector to form the final representation of the question which acts as an input to the Answer-Pointer layer. The Answer-Pointer layer employs a self-attention based mechanism to point to the exact span in the paragraph. The primary reason of using this architecture is that it points back to the span, rather than generating the answer on its own. This makes the architecture more robust to domain specific words, as it does not need to generate them from scratch.

Due to its complexity, and the inherent difficulty of the task, the model needs a large amount of data to train on. However, in our domain, supervised data is sparse and difficult to generate. To offset this general lack of data, we first train our model on SQuAD [12], and then fine-tune it over the domain dataset. Fine tuning is necessary as the underlying characteristics of one dataset is usually different from another. For example, while SQuAD has on average 11.31 words in the question and 2.47 words in the answer, our benchmark consists of questions including one more word on average and significantly longer answers (more than 10 words).

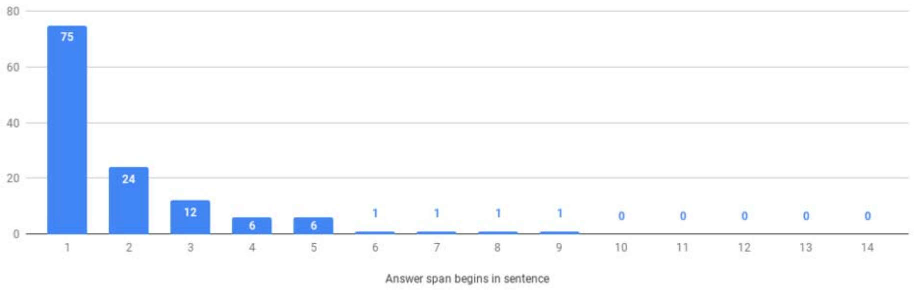


Figure 3. Distribution of the *start* of answers in MaRisk across paragraph sentences.

3.3. Question Generation

Contemporary machine learning models require copious amounts of data to achieve optimal results. And even when trained properly, their performance is highly representative of the inherent characters of the dataset. In our use case, we found out that the dataset is too small and disproportionately focuses on the first sentence of paragraphs. Also, models trained on this dataset perform poorly and do not generalize well.

To alleviate these issues, apart from using transfer learning, as described above, we also synthetically expanded the training data. We made use of a neural network question generation model proposed by [3]. The model is first trained on SQuAD [12]. We then apply the proposed method over the regulatory documents to generate a question for each of its sentences. In this manner, we overcome a major challenge hindering the use of statistical models for any subtask in the pipeline.

4. Evaluation Results

Dataset The ground truth consists of 631 question-answer pairs based on the MaRisk document. As mentioned above, a big majority of the questions have their answers in the beginning of the paragraph (first sentence), as visualized in Figure 3. On average, questions and answers have 12 and 13 words respectively. This is unlike SQuAD [12], whose answer spans are only three words long, on average. This distinction is noteworthy, as it makes transferring trained models from SQuAD to our dataset more challenging.

Metrics We report four metrics to evaluate the proposed approach and compare its effectiveness to alternative approaches: *Precision (P)*: ratio of the correct part of the predicted answer to the length of the predicted answer. *Recall (R)*: ratio of the the correct part of the predicted answer to the length of the true answer. *Partial Overlap (F1)*: Harmonic mean of P and R. *Exact Match (EM)*: This metric measures the percentage of predictions that exactly match the ground truth.

4.1. Answer Selection Evaluation

The answer selection module is trained in three different configurations, namely (i) on SQuAD without fine-tuning on the MaRisk data, (ii) on the MaRisk data (iii) pre-trained

	F1	EM
Trained on SQuAD	0.36	0.06
Trained on In-Domain data	0.20	0.10
Trained on SQuAD + In-Domain data	**0.59**	**0.34**

Table 1. Comparison of performance of the Answer Selection module trained in three different configurations: (i) only on the SQuAD dataset, (ii) on in-domain data, and (iii) on SQuAD and fine-tuned on in-domain data. In the third case, we are able to achieve significantly better results.

on SQuAD and then fine-tuned on the MaRisk data. The performance of the answer selection module trained in these configurations is summarized in Table 1.

We observe that training the model on SQuAD without any fine-tuning on the target dataset leads to very low performance. We attribute this to the difference in the inherent dataset characteristics. Moreover, while SQuAD consists of general domain text, the target task has a specialized vocabulary, which might impart specific, and sometimes completely different meaning to its constituent words. Also, the model performs almost equally worse when just trained on the domain dataset. This is as expected, as due to the limited size of the dataset the model fails to generalize on unseen data. However, when trained on SQuAD and afterwards fine-tuned in the regulatory domain, the results improve significantly. This suggests that transfer learning can, in general, improve the performance of our models, thereby helping to overcome the problem of the low number of training data in the specific domain.

4.2. Overall System Evaluation

The results of the QA system based on the two-step approach (Paragraph Selection and Answer Selection modules) are presented in Table 2. In particular, we compare four different pipelines composed by the combination of two variants for paragraph selection and two for answer selection. For paragraph selection, we compare StarSpace based model to a simple Lucene index (hereafter referred to as IR approach). For answer selection, we compare the MatchLSTM model, to a sentence selection based QA model, as proposed in [1].

In our experiments, we conclude that the IR based paragraph retrieval, and MatchLSTM based answer retrieval pipeline delivered the best results, across all the reported metrics. However, using StarSpace for paragraph retrieval instead of the IR approach also delivers comparable results. In addition, we observe that the sentence selection based approach for answer span selection performs worse. This is as expected since the answer spans in MaRisk do not follow sentence boundaries.

Instead of selecting the best paragraph in the first step, we also experiment with predicting more than one probable candidate paragraphs. We observe that the pipeline StarSpace (PS) + MatchLSTM (AS) performs the best amongst the baselines when k=3. However, when k=5, IR (PS) + MatchLSTM (AS) outperforms it. In general, the two paragraph selection algorithms can be used interchangeably since they deliver very similar results. In fact, the IR based approach achieves F1-measure and exact match only by 0.045 and 0.065 respectively higher compared to StarSpace when k=1.

Table 2. Evaluation of the proposed models. We report the Precision (P), Recall (R), F1-measure (F1) and Exact Match (EM). The metrics report accuracy over the final task of selecting the correct answer span.

QA Pipeline	P	R	F1	EM
IR (PS) + Choi et al. (SS)	0.182	0.479	0.263	0.000
IR (PS) + MatchLSTM (AS)	**0.667**	**0.574**	**0.617**	**0.343**
StarSpace (PS) + Choi et al. (SS)	0.182	0.446	0.259	0.000
StarSpace (PS) + MatchLSTM (AS)	0.618	0.532	0.572	0.278

A. Performance of baselines and the proposed model. Best performances are in **bold font**.

QA Pipeline	Top-k	P	R	F1	EM
IR (PS) + MatchLSTM (AS)	Top-1	0.667	0.574	0.617	0.343
	Top-3	0.702	0.617	0.657	0.343
	Top-5	**0.734**	**0.646**	**0.688**	**0.352**
StarSpace (PS) + MatchLSTM (AS)	Top-1	0.618	0.532	0.572	0.278
	Top-3	0.713	0.616	0.661	0.333
	Top-5	**0.74**	**0.641**	**0.687**	**0.343**

B. Two best performing models when considering top-1, top-3 and top-5 paragraphs for the paragraph selection step. Best performances are in **bold font**.

4.3. Discussion

The evaluation results suggest the following conclusions:

- Apart from the size of the dataset, its balance (in our case, position of the answer span) is pivotal for robust generalizations. In our approach, as mentioned in Section 3.3, we attempt to balance the dataset by synthetically generating questions from all the sentences in the document. We find out that synthetic data increases the model performance by 11%.
- The proposed approach is to a certain extent able to address the challenges of the regulatory domain question answering system. In particular, we address the fact that regulatory documents are usually long by introducing a two-step approach (paragraph and answer selection), also taking advantage of the structure of the documents which usually includes several sections, subsections, and paragraphs.
- From the performed evaluation, the combination of the IR approach for paragraph selection and the MatchLSTM model for answer selection delivers the best results in terms of F1-measure and exact match. In case we consider instead of the top-1 paragraph the top-3 selected paragraphs, the neural network based approach for paragraph selection performs slightly better. In general, both approaches for paragraph selection combined with the answer selection approach deliver comparable results.
- Transfer learning has impact on the domain only if domain-specific training data are taken into consideration. That means: (1) Models trained on other datasets in

order to address the same task do not generalize well on regulatory data. This was observed when SQuAD was used for training and the trained model was tested with in-domain data; (2) However, with careful fine-tuning, the models can substantially increase the performance in the regulatory domain.

5. Related Work

Question Answering over Legal Documents A significant amount of research work has been done in the direction of IR over legal documents. Some examples include the approach introduced by Landhaler et al. [9] which enhances full text search in document collections related to rights and obligations in contracts to find exact and semantically related matches using word2vec. Other approaches combine document embeddings and semantic word measures and natural language processing techniques for retrieving legal case documents [13]. For regulatory documents the efficiency of such techniques has not been tested yet. Also, QA over legal documents is not a new field of research. A QA system for retrieval of Portuguese juridical documents has been proposed by Quaresma et al. [11], able to answer factual questions related to criminal processes. This performs information extraction to build a structured knowledge base from the facts and transforms natural language questions into queries against this knowledge base. However, since only part of the document content can be semantically identified and described many questions remain unanswered. In a different approach, Ranking SVM and Convolutional Neural Networks are employed for building a question answering system able to retrieve answers included in legal articles at paragraph-level [2]. In fact, each article is split into single paragraphs before retrieving answers corresponding to these paragraphs. In our case, we employ an even finer granularity – at text span level. The main disadvantage of the aforementioned systems is that they retrieve relevant documents which may contain the answer rather than the actual answer to a user's question, as we are able to achieve with the current proposed approach.

Reading Comprehension Reading comprehension is a new field of research related to tasks like question answering and machine translation and its goal is "teaching" computers read and understand documents as humans do. To address this problem many deep learning models have been proposed. Wang and Jiang [16] have developed an end-to-end neural architecture for retrieving an answer given a paragraph; this approach is based on match-LSTM model and a Pointer Network [15], a sequence-to-sequence model which allows the generation of answers consisting of multiple tokens (i.e., text span) coming from the original input (i.e., paragraph). The implemented model in the current proposal for answer selection is based on the aforementioned approach. Several other models have been proposed addressing the problem of reading comprehension. One common approach is to use recurrent neural networks (RNNs) in order to predict or generate the answers together with the use of attention mechanisms for matching the words of the question to the corresponding text span in a given passage [5]. Memory Networks [14] have also been applied to reading comprehension. These methods are often slow for long documents because the model needs to be executed sequentially over possibly thousands of tokens. To address the problem of long documents Choi et al. [1] propose to combine a coarse, fast model for selecting relevant sentences and a more expensive RNN for retrieving the answer from those sentences. Similarly, we are following a two-step approach by first selecting the related paragraph and then the corresponding text span.

QA Benchmarks Most of the available datasets for training reading comprehension and QA related models are focusing on the general domain. We have shown in this paper how such datasets can be used for transfer learning to specific domains, in this case the regulatory domain. The Stanford Question Answering Dataset (SQuAD) [12] which was used to train the models in the proposed approach contains more than 100,000 question-answer pairs for documents taken from approximately 500 Wikipedia articles which have been annotated through crowdsourcing. Another large-scale dataset, MS MARCO, consists of 100,000 questions, 1 million passages, and links to over 200,000 Web documents; compared to SQuAD it includes several unanswerable queries and provides human generated answers rather than text spans. SQuAD has been proved more appropriate for the type of problem we are solving, therefore it has been selected for training our models. Apart from SQuAD and MS MARCO, a few other datasets exist addressing the tasks of reading comprehension and open-domain QA like the corpus of cloze style questions from CNN/Daily News summaries [5] and the Children's Book Test [6]. Apart from being too small for deep learning approaches the aforementioned datasets do not map natural questions to text spans directly, thus, they are inappropriate for our approach.

6. Conclusions

In this article we propose a two-step QA system for regulatory documents. We evaluate its various variants and compare it with contemporary IR techniques. Through our experiments, we conclude that the combination of the IR approach for paragraph selection and the MatchLSTM model for answer selection delivers the best results in terms of F1-measure and exact match. Furthermore we find that both transfer learning and data generation mechanisms can significantly improve the performance of the system.

In the future, we aim to further improve the system performance by generating and training our models on larger datasets. We look forward to extensions of this approach over other documents in the regulatory domain, and the generalization of the current system to work across them. Additionally, an interesting direction for further research would be to investigate the model's uncertainity quantification in its predictions. We hope that this would further bolster the adoption of these techniques in the community.

References

[1] E. Choi, D. Hewlett, J. Uszkoreit, I. Polosukhin, A. Lacoste, and J. Berant. Coarse-to-Fine Question Answering for Long Documents. In *Proc. of the 55th Ann. Meeting of the Association for Computational Linguistics, ACL 2017, Vol. 1: Long Papers*, pages 209–220, 2017.

[2] P. Do, H. Nguyen, C. Tran, M. Nguyen, and M. Nguyen. Legal Question Answering using Ranking SVM and Deep Convolutional Neural Network. *CoRR*, abs/1703.05320, 2017.

[3] X. Du, J. Shao, and C. Cardie. Learning to Ask: Neural Question Generation for Reading Comprehension. In *Proc. of the 55th Ann. Meeting of the Association for Computational Linguistics, ACL 2017, Vol. 1: Long Papers*, pages 1342–1352, 2017.

[4] D. A. Ferrucci. Introduction to "This is Watson". *IBM Journal of Research and Development*, 56(3):1, 2012.

[5] K. M. Hermann, T. Kociský, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom. Teaching Machines to Read and Comprehend. In *Advances in Neural Information Processing Systems 28: Ann. Conf. on Neural Information Processing Systems 2015*, pages 1693–1701, 2015.

[6] F. Hill, A. Bordes, S. Chopra, and J. Weston. The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations. *CoRR*, abs/1511.02301, 2015.

[7] M. Hu, Y. Peng, Z. Huang, N. Yang, M. Zhou, et al. Read+ verify: Machine reading comprehension with unanswerable questions. *arXiv preprint arXiv:1808.05759*, 2018.

[8] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*, 2016.

[9] J. Landthaler, B. Waltl, P. Holl, and F. Matthes. Extending Full Text Search for Legal Document Collections Using Word Embeddings. In *Legal Knowledge and Information Systems - JURIX 2016: The Twenty-Ninth Ann. Conf.*, pages 73–82, 2016.

[10] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. In *Proc. of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Ann. Conf. on Neural Information Processing Systems (NIPS 2016).*, 2016.

[11] P. Quaresma and I. P. Rodrigues. A Question Answer System for Legal Information Retrieval. In *Legal Knowledge and Information Systems - JURIX 2005: The Eighteenth Ann. Conf. on Legal Knowledge and Information Systems.*, pages 91–100, 2005.

[12] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. SQuAD: 100, 000+ Questions for Machine Comprehension of Text. In *Proc. of the 2016 Conf. on Empirical Methods in Natural Language Processing, EMNLP 2016.*, pages 2383–2392, 2016.

[13] K. Sugathadasa, B. Ayesha, N. de Silva, A. S. Perera, V. Jayawardana, D. Lakmal, and M. Perera. Legal Document Retrieval using Document Vector Embeddings and Deep Learning. *CoRR*, abs/1805.10685, 2018.

[14] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus. End-To-End Memory Networks. In *Advances in Neural Information Processing Systems 28: Ann. Conf. on Neural Information Processing Systems 2015*, pages 2440–2448, 2015.

[15] O. Vinyals, M. Fortunato, and N. Jaitly. Pointer Networks. In *Advances in Neural Information Processing Systems 28: Ann. Conf. on Neural Information Processing Systems 2015*, pages 2692–2700, 2015.

[16] S. Wang and J. Jiang. Machine Comprehension Using Match-LSTM and Answer Pointer. *CoRR*, abs/1608.07905, 2016.

[17] L. Y. Wu, A. Fisch, S. Chopra, K. Adams, A. Bordes, and J. Weston. StarSpace: Embed All The Things! In *Proc. of the Thirty-Second AAAI Conf. on Artificial Intelligence, 2018*, 2018.

Legal Knowledge and Information Systems
M. Palmirani (Ed.)

doi:10.3233/978-1-61499-935-5-51

Automated Processing of Privacy Policies Under the EU General Data Protection Regulation

Giuseppe CONTISSA [a], Koen DOCTER [b], Francesca LAGIOIA [b],
Marco LIPPI [c], Hans-Wolfgang MICKLITZ [b], Przemyslaw PALKA [d],
Giovanni SARTOR [a] and Paolo TORRONI [e]
[a] *CIRSFID, Alma Mater – Università di Bologna, Italy*
[b] *Law Department, European University Institute, Florence, Italy*
[c] *DISMI – Università di Modena e Reggio Emilia, Italy*
[d] *Center for Private Law, Yale Law School, New Haven, United States*
[e] *DISI, Alma Mater – Università di Bologna, Italy*

Abstract. Two years after its entry into force, the EU General Data Protection Regulation became applicable on the 25th May 2018. Despite the long time for preparation, privacy policies of online platforms and services still often fail to comply with information duties and the standard of lawfulness of data processing. In this paper we present a new methodology for processing privacy policies under GDPR's provisions, and a novel annotated corpus, to be used by machine learning systems to automatically check the compliance and adequacy of privacy policies. Preliminary results confirm the potential of the methodology.

1. Introduction: the legal and technological context

In Europe the processing of online personal data falls under the the General Data Protection Regulation (GDPR), which aims at making all data processing (from collection, to usage to transfers) lawful, fair and transparent. The enforcement of GDPR is based on two complementary approaches: (1) the administrative control by independent supervisory authorities and (2) the exercise of private rights by data subjects and/or civil society. The supervisory authority can either act on its own motion, or as a result of a complaint by a data subject or an NGO. To ensure transparency and enable the effective exercise of data subjects' rights, the GDPR requires controllers to provide the data subject with the information enlisted in Art. 13 and 14. Art. 12 stipulates that all this information must be given "in a concise, transparent, intelligible and easily accessible form, using clear and plain language". The document containing this information, namely the privacy policy, fails to be GDPR compliant if it foresees unlawful processings, if it does not contain required information, or if it uses unclear language. Our research indicates that many privacy policies fail to meet the requirements of the GDPR (see 4).

This undesirable state of affairs is due to the fact that even though data subjects, civil society and public authorities are legally empowered to conduct the control, they lack factual capabilities to do so, given the large amount of privacy policies to be checked and their complexity. Recent research has shown that tools for legal text analytics can be used to assess the completeness of privacy policies [2] and to automatically extract, categorize, and summarize information from privacy documents [3, 7, 6].

This work builds on previous research [4] where we have used machine learning methods to address the automated detection of potentially unfair clauses in online contracts. Its purpose is threefold: (1) to define the standard for a correctly designed privacy policy, in form and in content, under the currently existing standards put forward by the GDPR; (2) to analyse the privacy policies of 14 relevant online platforms and services in accordance with that standard; and (3) to verify to what extent such analysis can be automated, in order to empower consumers as a response to the technological supremacy of companies and business. If we succeed at (even partially) automating the analysis of privacy policies, this can pave the way for the development of tools that increase the efficiency and quality of the work of supervisory authorities and NGOs, and/or empower data subjects themselves. Section 2 provides an overview of the document corpus and describes the methodology adopted for evaluating privacy policies. In particular, it describes all the legal requirements that a properly designed privacy policy should meet. Further, we provide an overview of the document annotation procedures. Section 3 explains the machine learning methodology employed in the system, and some preliminary results. Section 4 concludes with a look at future research.

2. Classification of clauses and annotation guidelines

In this section, we shall provide the methodology adopted for evaluating privacy policies, and an overview of the annotated corpus.

According to the GDPR, privacy policies should be comprehensive, regarding the information they provide; comprehensible, regarding the form of expression; and substantively compliant, regarding GDPRs rules and principles (see 1). Thus, we defined a **Golden Standard** including the following three top-level dimensions: (1) Comprehensiveness of information: the policy should include all the information that is required by articles 13 and 14 of the GDPR; (2) Substantive compliance: the policy should only allow for the types of processing of personal data that are compliant with the GDPR; and (3) Clarity of expression: the policy should be framed in an understandable and precise language. With regard to these three dimensions we distinguished optimal and suboptimal achievement. In the first case the privacy policy clearly meets the GDPR requirements along the dimension at issue; while in case of suboptimal achievement the privacy policy apparently fails to reach the threshold required. In some cases we have distinguished two levels of suboptimal achievement: (a) questionable achievement: it may be reasonably doubted that the suboptimal policy reaches the threshold required (the clause could have been better framed, but still there is the possibility that the competent authorities view it as being good enough, i.e., that its improvement is only supererogatory); and (b) Insufficient (or no) achievement: the suboptimal policy clearly fails to reach the threshold. For each dimension, we have distinguished different aspects relatively to which the clause could be assessed. Each of these aspects of profile was denoted by a tag to be used, together with a number indicating the level of achievement, in the annotation of our corpus, as we shall see in the next section.

The corpus for our exploratory inquiry consisted of 14 relevant online privacy policies, i.e. Google, Facebook, Amazon, Apple, Microsoft, WhatsApp, Twitter, Uber, AirBnB, Booking, Skyscanner, Netflix, Steam and Epic Games. These privacy policies were selected among those provided by the main online platforms, taking into account their significance in terms of number of users and global relevance. Due to the limitation of our resources and time constraints we had to focus on such a limited number of documents (thanks to additional resources, we will be able to expand it substantially in the future), but the corpus still has a significant size. In fact, it contains overall 3,658 sentences (80,398 words), 401 sentences (11.0%) of which were manually marked as containing unclear language; 1,240 (33.9%) were marked as potentially problematic or as providing insufficient information. We used XML as a mark-up language. In cases where a single clause fell into multiple categories according to our classification, we applied to it multiple tags. If a clause span included multiple sentences, we tagged all such sentences. Readers can review full privacy policies annotated here: http://www.claudette.eu/gdpr/.

In the following subsection, we shall introduce, for each dimension of our golden standard, the different aspects of it that were distinguished in our annotation, being denoted by different tags.

2.1. Comprehensiveness of Information

The dimension of comprehensiveness of information concerns whether a privacy policy meets all the information requirements of Articles 13 and 14 of the GDPR, or fails to do so, either by not providing at all the required item of information, or by providing it insufficiently or imprecisely. We identified 12 types of required information clauses, for which we defined corresponding XML tags, as specified below. For each type of required information, we classified the corresponding clause either as optimal, i.e., fully informative (all the required information is present and well specified); or as suboptimal, i.e., insufficiently informative (information is hinted at, but non-comprehensive), appending to each XML tag respectively number 1 or 2. As noted above, a single clause in some cases may fall in different categories and consequently may have multiple taggings. In the following we present each category.

Identity of the controller and, where applicable, of the controller's representative (label:`<id>`). According to the GDPR this information must be provided both when personal data are collected from the data subject (Article 13(1)(a)) and when they have not been obtained from the data subject (Article 14(1)(a)). As an example of a suboptimal clause in this regard, thus labelled as `<id2>`, consider the following example taken from the Airbnb privacy policy (last updated on 16 April 2018):

```
<id2>If you change your Country of Residence, the Data Controller and/or Payments
Data Controller will be determined by your new Country of Residence as specified
above, from the date on which your Country of Residence changes.</id2>
```

Contact details of the controller and, where applicable, of the controller's representative (label:`<contact>`). Contact details should allow for different forms of communication with the data controller (See Article 29 Working Party Guideline on Transparency under Regulation 2016/679 (WP260), hereinafter "Transparency Guidelines", p. 26). In order to facilitate the exercise of data subjects rights, the data controller should "also provide means for requests to be made electronically, especially where personal data are processed by electronic means" (See GDPR, Recital 59). We labelled a clause on contact details as `<contact2>` when it did not allow for different forms of communication with the data controller (e.g. phone number, email, postal address etc) or it failed to provide adequate specifications.

Contact details of the data protection officer (label:`<dpo>`). They must be published and communicated by the controller or processor to the relevant supervisory authorities (Article 37(7)). This information should allow data subjects to contact the DPO easily and directly, without having to contact another part of the organisation. Article 37(7) does not require that the published contact details should include the name of the DPO. Whilst it may be a good practice to do so. We labelled a clause as `<dpo2>` when it only reached a low standard for the clarity and accessibility of the information, e.g. when it only provided a dedicated email address, omitting both the name of the DPO and a postal address.

Purposes of the processing (label:`<purp>`). Purpose specification must be provided by the controller (Articles 13(1)(c) and 14(1)(c)), to ensure a degree of user control and transparency for the data subject (See GDPR Art. 6(1)(a); Article 29 WP Guidelines on consent under Regulation 2016/679 (WP259 rev.01) p. 13; and Recital 42 GDPR). We labelled as `<purp2>` those cases where, for instance, it was unclear (i) what type of data would be processed; and (ii) what the correlation was between the collected information and the specific purposes, since a number of different purposes were listed one after the other.

Legal basis for the processing (label:`<basis>`). The legal basis of the processing (Articles. 13(1)(c) and 14(1)(c)) must be specified with regard to both personal data (Article 6(1)) and special categories of personal data (Article 9). For instance, we labelled as `<basis2>` those clauses that specified the purpose with reference to broad marketing practices, or that mixed multiple unrelated purposes.

Categories of personal data concerned (label:`<cat>`). This specification must be provided where personal data have not been obtained from the data subject (Article 14(1)(d)), as well as whenever the data subject consent constitutes the legal basis for the processing (Articles 6 and 9). A kind of clauses that we labelled as `<cat2>` were those that only provided a non-exhaustive list of examples of the collected data.

Recipients or categories of recipients of the personal data (label:`<recep>`) (Article 13(1)(e)). In accordance with the principle of fairness, controllers must provide information on the recipients that is meaningful for data subjects, i.e. the named recipients, so that data subjects know exactly who has their personal data, or the categories of recipients, by indicating the type of recipient (i.e. the activities it carries out), the industry, sector and sub-sector and its location. We labelled clauses as `<recep2>`, when they did not clearly indicate the type, the industry, the sector and the location of the mentioned recipients.

The period for which the personal data will be stored, or if that is not possible, the criteria used to determine that period (label:`<ret>`). 'We labelled as `<ret2>` those clauses that fail to specify the time for which the personal data will be stored, or at least the criteria used to determine that period, such as those clauses generically stating that personal data will be kept as long as necessary.

The rights to access; rectification; erasure; restriction on processing; objection to processing and data portability (label:`<correct>`) (Articles 13.2(b) and 14.2(c)). This information should be specific to the processing scenario and include a summary of what the right involves and how the data subject can take steps to exercise it, as well as any limitations on the right. We classified as `<correct2>` those clauses that failed to specify under what condition data subjects could exercise their rights and what steps were needed to exercise them.

The right to lodge a complaint with a supervisory authority (label:`<complain>`) (Articles 13.2(b) and 14.2(c)). We labelled as `<complain2>` the clauses that failed to specify in which State a complain should be presented, or that provided a wrong or misleading specification.

Information about source from which the personal data originate, and if applicable, whether it came from publicly accessible sources (label:`<source>`). Information on the source must be provided if personal data are not coming directly from the data subject (Article 14(2)(f)). We labelled as `<source2>` those clauses failing to specify the nature of the sources (i.e. publicly/ privately held sources; the types of organisation/ industry/ sector; and where the information was held (EU or non-EU) etc.).

The existence of automated decision-making, including profiling (label:`<auto>`). The policy must also include meaningful information about the logic involved, the significance and the envisaged consequences of such processing for the data subject (Articles 13.2(f) and 14.2(g))[1]. We labelled as `<auto2>` those clauses that failed to provide meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject; or that did not inform the data subject about the right not to be subject to a decision based solely on automated decision making, including profiling; and that the decisions referred would not be based on sensitive data.

2.2. Substantive compliance

This dimension concerns whether the types of processing stipulated are themselves GDPR compliant. We identified 10 categories of clauses and for each category, we defined a corresponding XML tag, as specified below. We assumed that each category could be classified either as a *fair processing*, clause; a *problematic processing* clause; and as an *unfair processing clause*. To this end, we appended a numeric value to each XML tag, with 1 meaning fair; 2 problematic; and 3 unfair.

[1] See Article 29 Working Party Guideline on Automated Individual Decision-making and Profiling for the Purposes of Regulation 2016/679 (WP251rev.01).

Processing of special categories of personal data (label:<sens>). Processing of sensitive data (e.g. data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, concerning health or a natural person's sex life or sexual orientation, etc.) are prohibited unless an exception applies (Articles 9 and 9(2)). We labelled as <sens2>, those clauses that allowed for the processing of sensitive data without providing full information e.g. clauses stating that the explicit consent is required for processing certain sensitive data, but failing to indicate the purpose of such processing. We labelled as <sens3> the clauses allowing for the processing of sensitive data outside the conditions specified in Article 9. As an example of a <sens2>-labelled clause, consider the following fragment taken from the Facebook privacy policy (last updated on 19 April 2018):

```
<sens2>To create personalized products that are unique and relevant to you, we
use your connections, preferences, interests and activities based on the data
we collect and learn from you and others (including any data with special
protections you choose to provide where you have given your explicit consent);
how you use and interact with our Products; and the people, places, or things
you're connected to and interested in on and off our Products.</sens2>
```

Consent by using (label:<cuse>). Consent should be given "by a statement or by a clear affirmative action" (art 4(11)). Thus we labelled as <cuse3>, among others, those clauses stating that by simply using the service, the user consents to the terms of the privacy policy. For instance, consider the following example taken from the Epic games privacy policy (last updated on 24 May 2018):

```
<cuse3> when you use our websites, games, game engines, and applications, you
agree to our collection, use, disclosure, and transfer of information as
described in this policy, so please review it carefully.</cuse3>
```

Take or leave it approach (label:<tol). "When assessing whether consent is freely given, utmost account shall be taken of whether, inter alia, the performance of a contract, including the provision of a service, is conditional on consent to the processing of personal data that is not necessary for the performance of that contract."(Article 7(4)). The situation of "bundling" consent with acceptance of terms or conditions, or "tying" the provision of a contract or a service to a request for consent to process personal data that are not necessary for the performance of that contract or service, is considered highly undesirable". We always labelled all clauses implementing a "take it or leave it approach" as <tol2>.

Transfers to third parties (label:<tp>). The data subjects should be informed on whether their information will be transferred to third parties, on the identity of the transferee, and on the legal basis for the transfer (such as consent by the party, necessity for the execution of a contract, or legitimate interest). We labelled a clause as <tp2> when the purpose of the transfer or identity of the transferee were not specified, but (a) the transfer presupposed consent of the data subject, which was not necessary to access the service, or (b) the transfer was needed to perform the contract. We labelled a clause as <tp3> when the purpose of the transfer or the identity of the transferee were not specified, but consent was necessary to access the service and the transfer was not needed to perform the contract.

Policy change (label:<pch>). The controller should adhere to the transparency principle when communicating both the initial privacy statement/ notice and any subsequent substantive or material changes to this statement/ notice, and consider several factors in assessing what is a substantive or material change. We labelled a clause as <pch2> when stating that a notice of the change would be provided, but would not require new consent or a confirmation of reading; and as <pch3> when even the commitment to a fresh notice was rejected (e.g. clauses stating that it is a responsibility of the data subject to check for the last updated version of the privacy policy).

Transfer to third country (label:<cross>). Articles 13(1)(f) and 14(1)(f) require the controller to inform the data subject about (i) whether he/she intends to transfer personal data to third countries and (ii) the existence or absence of adequacy decision by the Commission or appropriate safeguards. Besides, Chapter V (art. 44–49) governs the transfer of personal data to third countries. We labelled a clause as <cross2> when it only mentioned one of the transfer mechanisms listed by Articles 44–49 and did not provide any specific information allowing the data subject to be effectively informed. We labelled as <cross3> clauses failing to provide any information on the transfer mechanism requirements.

Processing of children's data (label:<child>). According to Article 8(1) The GDPR requires parental consent for the processing of data concerning children below 16 years (Article 8), and recommends a cautious and proportionate approach for all children (individuals under 18). We labelled as <child2> those clauses failing to mention and/or specify the types of efforts made to verify that the consent is given/authorised by the holder of parental responsibility. We labelled as <child3>, for instance, those clauses failing to specify what efforts would be taken to verify parental authorisation.

Advertising (label:<ad>). Whenever profiling for marketing purposes involves the use of personal data that were originally collected for something else, these marketing purposes must be compatible with the original purposes for which the data were collected, and moreover, the data subject has a right to opt-out (Articles 21(2), 21(3), and Recital 70). Consequently, we labelled a clause as <ad2> when the consent was not required, but the opt-out was possible, and as <ad3> when the consent was not requested, and the opt-out was not possible.

Any other type of consent (label:<c>). Since consent cannot be given through the general acceptance of a privacy policy or terms of use (Articles 4(11) and 7(2)), we labelled as <ad2> all clauses where "hidden" consents were present.

Any other type of clause we find "outstandingly problematic" (label:<out>). We labelled as <out2> clauses stating anything else that did not fall under the scope of the above-mentioned clauses, and yet could be considered as problematic for different reasons. For example, this included clauses that were obscure in their meaning, or that claimed the responsibility of the data subject regarding the processing of the data of third parties by the data controller.

2.3. Clarity of Expression

Article 5(1)(a) requires that personal data must be processed lawfully, fairly and in a transparent manner. Further, Article 12(1) requires that information must be provided in a concise, transparent intelligible and easily accessible form, using clear and plain language. Therefore, complex sentence and language structures should be avoided. Besides, the information should not be phrased in abstract or ambivalent terms or leave room for different interpretations (see Guideline on Transparency, pp. 9-10). The use of the language qualifiers such as “may”, “might”, “some”, “often”, and “possible” should be avoided (see Guideline on Transparency, p. 9), as well as “including” and “such as” when they are present within a list, for example of the category of data collected. We identified two types of clauses: clauses expressed in a clear language (not tagged) and clauses expressed in an unclear language, for which we defined the following XML tag: `<vag>`. As an example that fails to meet optimality concerning the *clarity* of expression, consider the following example taken from the Apple privacy policy (last updated on 22 May 2018):

```
<vag>Apple and its affiliates may share this personal information with each
other and use it consistent with this Privacy Policy. They may also combine it
with other information to provide and improve our products, services, content,
and advertising.</vag>
```

3. Machine Learning Methods and Experiments

Although this work focuses on offering a novel methodology for the labeling of privacy policies, with the purpose of automating the evaluation of such documents, we also present some preliminary results of the machine learning experiments conducted on the corpus.

Given the complexity of the problem, there are several ways in which the automatic system could be developed. For example, regarding problematic clauses, one could first detect all problematic clauses (in general) and then distinguish each category (data regarding children, advertising, etc.). Or the other way around, one could first detect all sentences regarding a certain category (e.g. advertising) and then decide whether they are problematic or not. The same holds for required clauses. In contrast, vague clauses do not have sub-categories.

As for the adopted algorithms, in our current approach we started experimenting with the technologies that have been successfully employed in the detection of potentially unfair clauses in online Terms of Service [5], including support vector machines (SVM) and deep networks. In some cases, a solution based on manually defined rules and patterns could also be used to detect some specific categories of problematic or required clauses, as often done in data mining.

We employed a standard leave-one-document-out (LOO) procedure, where training is repeated N times (N being the number of documents), with a different document of the corpus used as the test set, and the remaining $N - 1$ forming the training set. Performance is measured with standard metrics: precision, as the percentage of predicted positive sentences that are indeed positive in the corpus; recall, as the percentage of positive sentences that are indeed classified as positive; F_1, as the harmonic mean between precision and recall (the set of positive clauses depending on the task).

In a first experiment, we considered unclear language clauses only. Here, a simple grammar that detects whether some keyword (or combination of keywords) is present in each sentence is capable of recognizing 89% of vague clauses, yet with a low precision of 25%. A machine learning classifier based on Support Vector Machines and bag-of-words, instead, detects 72% of vague clauses, yet with a low precision of 30%. A combination of grammar and machine learning achieves 81%/32% recall/precision. Yet, a detailed analysis of the false positives (sentences detected as unclear, which actually were not tagged as such) shows that most of them are indeed problematic clauses. This observation made us argue that probably a machine learning classifier could take advantage of observing the combination of both problematic and unclear clauses. Therefore, we repeated the same experiments, this time considering the positive class (to be detected) as the union of problematic and vague clauses. Following the same approach described above, a hand-crafted grammar correctly detects 92% of positive clauses, yet with 31% precision. A pure machine learning classifier achieves instead 70% recall and 50% precision. A combination of the two approaches reaches a 75% recall with 47% precision, with an overall 57% F_1. Although these numbers could seem unimpressive to a lay observer, we shall remark that as preliminary results they are not bad at all. Indeed, they are comparable to the results obtained with the analysis of Terms of Service with a corpus of 20 documents [4], where we had initially obtained a 72%/62% recall/precision, which further increased to 80%/83% when the corpus was extended to include 50 documents. We could imagine a similar trend also for privacy policies.

Once problematic clauses have been detected, automatic categorization of such sentences into unlawfulness classes is a much simpler task: SVMs are capable of identifying the correct category with precision/recall usually around 80%/75%.

Some required information clauses can also be easily detected with grammars and regular expressions. For example, the category of automatic decision making can be identified with 95% precision and 83% recall, and the required information about complaints can be identified with 94% precision and 91% recall. Similarly, the data protection officer clause can be detected with 78% precision and 85% recall. Other tags are much more heterogeneous, and thus difficult to detect with hand-crafted rules (e.g., the purposes and the legal basis of data processing): in these cases, machine learning achieves performance comparable to the detection of problematic clauses.

4. Conclusions

This paper presented a first experimental study that used machine learning to evaluate privacy policy under the GDPR. Our inquiry was based on the identification of a golden standard for privacy policies, and on the definition of a methodology for assessing the extent to which policies get closer to such standard. From the legal perspective, our analysis of 14 privacy policies of online platforms and services suggests that there is still a significant margin for improvement. None of the analysed privacy policies gets close to meeting the standards put forward by the GDPR. Unsatisfactory treatment of the information requirements (e.g. with regard to contact details of the DPO; we could not retrieve an example of a fully informative clause from the policies we analysed); large amounts of sentences employing vague language; and an alarming number of “problematic” clauses cannot be deemed satisfactory. The results we obtained in our machine learning experiments, though still limited and provisional – mostly due to the limited extension of our corpus – show that there are promising prospects for developing automated tools to support the detection of unlawful clauses in privacy policies. Providing such tools will be an important contribution to the protection of individuals, in particular consumers or internet users. In the era of big data and automated decision-making, there is indeed a strong connection between data protection and consumer protection, as we are tracked, profiled, and directed/manipulated especially in our role as potential consumers. Thus there is a strong synergy between the purpose of our first Claudette project (empowering consumers and their associations to assess the legality and the fairness of online contracts) and the purpose of the project here presented (empowering data-subjects to assess privacy policies ([1])). Addressing privacy policies has been more difficult than examining consumer contracts: policies are less modular than contracts, they use a more variable and open-textured language, may have multiple layers, etc. Moreover, privacy policies may be defective not only for including wrong clauses, but also for omitting required information: thus to assess policies, negative tests are also needed. In the future, we plan to develop our research in different directions: addressing multilingualism in privacy policies, providing argument-based explanations for the assessment of clauses, detecting inconsistencies in policies, assessing the overall quality of privacy documents. Besides, as data protection regulation is rapidly evolving we will need to consider ways to minimize the effort involved in adapting our tool to new regulations and decisions by competent authorities.

References

[1] G. Contissa, F. Lagioia, M. Lippi, P. Palka, H.-W. Micklitz, G. Sartor, and P. Torroni. Towards consumer-empowering artificial intelligence. In *Proceedings of IJCAI Conference*, pages 5150–7. 2018.

[2] E. Costante, Y. Sun, M. Petković, and J. den Hartog. A machine learning solution to assess privacy policy completeness:(short paper). In *ACM workshop on Privacy in the electronic society*, pages 91–96. ACM, 2012.

[3] K. P. Joshi, A. Gupta, S. Mittal, C. Pearce, A. Joshi, and T. Finin. Alda: Cognitive assistant for legal document analytics. *AAAI Fall symposium*, 2016.

[4] M. Lippi, P. Palka, G. Contissa, F. Lagioia, H.-W. Micklitz, Y. Panagis, G. Sartor, and P. Torroni. Automated detection of unfair clauses in online consumer contracts. *Legal Knowledge and Information Systems*, page 145, 2017.

[5] M. Lippi, P. Palka, G. Contissa, F. Lagioia, H. Micklitz, G. Sartor, and P. Torroni. CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service. *CoRR*, abs/1805.01217, 2018. URL `http://arxiv.org/abs/1805.01217`.

[6] N. Tomuro, S. Lytinen, and K. Hornsburg. Automatic summarization of privacy policies using ensemble learning. In *Proceedings of the Sixth ACM Conference on Data and Application Security and Privacy*, pages 133–135, 2016.

[7] R. N. Zaeem, R. L. German, and K. S. Barber. Privacycheck: Automatic summarization of privacy policies using data mining. *ACM Transactions on Internet Technology (TOIT)*, 18(4):53, 2018.

Legal Knowledge and Information Systems
M. Palmirani (Ed.)

doi:10.3233/978-1-61499-935-5-61

Classifying Semantic Types of Legal Sentences: Portability of Machine Learning Models

Ingo GLASER [a], Elena SCEPANKOVA [a], and Florian MATTHES [a]
[a] *Software Engineering for Business Information Systems, Department of Informatics, Technical University of Munich, Germany*

Abstract. Legal contract analysis is an important research area. The classification of clauses or sentences enables valuable insights such as the extraction of rights and obligations. However, datasets consisting of contracts are quite rare, particularly regarding German language.

Therefore this paper experiments the portability of machine learning (ML) models with regard to different document types. We trained different ML classifiers on the tenancy law of the German Civil Code (BGB) to apply the resulting models on a set of rental agreements afterwards. The performance of our models varies on the contract set. Some models perform significantly worse, while certain settings reveal a portability. Additionally, we trained and evaluated the same classifiers on a dataset consisting solely of contracts, to be able to observe a reference performance. We could show that the performance of ML models may depend on the document type used for training, while certain setups result in portable models.

Keywords. legal sentence classification, portability of machine learning models, natural language processing, text mining

1. Introduction

Nowadays, many sectors face the obstacle called digitalization. So, does the legal domain as well. The rising of legal technology is highlighted by the increasing number of digitized legal documents, in particular legal contracts [1]. Due to the vast progress in research with regard to natural language processing (NLP), text mining is becoming more powerful in terms of its accuracy and performance. The tools and use cases for text mining in the legal field that are relevant for legal experts or practitioners, e.g., scientists, lawyers, judges, courts, etc., are diverse [2].

The computer-aided analysis of legal contracts is an important research are. Companies, law firms, government agencies, but also private individuals need to monitor contracts for a wide range of tasks [3]. For example, law firms and legal departments need to process large numbers of contracts to monitor the compliance. In terms of B2C, individuals are always involved as a contractual party. Therefore they need to understand their rights and obligations within that business relationship. However, the complex legal language hampers this understanding [4]. Thus we consider the task of extracting different legal concepts out of contracts.

In the legal domain however, the issue of data scarcity exists. This applies particularly to the German legal domain with regard to contracts, due to the nature of such documents, as they carry big privacy concerns. In this paper we want to investigate on the portability of machine learning (ML) models. Usually it is differentiated between two types of portability. Domain portability describes the capability of a ML model to perform its designated task on a different domain than it was originally trained on. In contrast, it can be also distinguished between different kind of documents. This paper focuses on the latter type of portability in order to overcome the lack of legal training data.

The reminder of the paper is structured as follows: Section 2 provides a short overview of the related work, Section 3 describes the semantic types leveraged, the experimental setup along with the used datasets are discussed in Section 4, finally the approaches and its performance is evaluated in Section 5, before Section 6 closes with a conclusion and outlook.

2. Related Work

The computer-assisted analysis of text phrases in legal documents with regard to their semantic type is highly relevant and has attracted researchers for quite some time. However, hardly any attempt has been made in the German contract domain. Waltl et al. [5] introduced a semantic type taxonomy for the German civil law, which was used to classify norms of German statutory texts. They used rule-based approaches as well as ML for the classification. In a previous work, Waltl et al. [6] incorporated active learning (AL) into that process in order to overcome the problem of data scarcity.

Approaches to classify legal norms with ML however exist in different jurisdictions. An early contribution to the classification of norms in legislative texts using ML approaches was made by Biagioli et al. [7] in 2005. The authors distinguish between 11 different semantic types, which are assigned on a norm level, achieving an average F_1 measure of 0.80. The same 11 functional classes were used by Francesconi and Passerin [8] to evaluate a multinomial naive bayes (MNB) classifier as well as a support vector machine (SVM). Maat et al. [9] classified legal norms within the Dutch legislation, using a taxonomy of 13 different classes. Utilizing SVMs, they could achieve an accuracy of more than 90%. O'Neill et al. [4] achieved an accuracy of 82% on the task of classifying sentences within different financial regulations leveraging deep learning (DL).

Research on classifying text phrases within contracts has been made in different domains. Indukuri and Krishna [10] applied a SVM on contract clauses to detect whether a clause is concerned with payment terms or not. Chalkidis et al. [11] tried to extract obligations and prohibitions out of Englisch contracts using hierarchical recurrent neural networks (RNN). While there is other work existing on the application of NLP for the information extraction (IE) of legal contracts [12,13,14,15], hardly any other work on classifying contract clauses with regard to their semantic types exists.

To the best of our knowledge no attempt to classify sentences for German legal contracts using ML approaches has ever been made before. Furthermore, we are not aware of any work related to the portability of ML models between statutory texts and legal contracts, even though portability of ML models as such, often referred to as transfer learning, is an examined research area.

3. Semantic Types of German Legal Sentences

The classification of legal sentences can be addressed from different perspectives, such as legal theoretical, philosophical, or a constructive one. In order to capture the semantics of sentences in legal documents, a functional classification approach seems to be most suitable. For the legal domain, different functional type systems were already introduced in other works as well as by us, e.g. [5,6,9]. We looked at existing classifications but also tried to leverage them, to come up with other possible taxonomies.

Waltl et al. [6] came up with a legal theoretically funded taxonomy of legal norms for German statutes. This classification system distinguishes at the first granularity between normative, auxiliary, legal-technical, and legal-mechanism statements (a deeper description of the taxonomy can be found in [6]). Due to the fact that this work investigates in the document type portability of ML models, a classification system appropriate to the legal domain, rather than special to a document type such as statutes, is required. Furthermore, this taxonomy constitutes 21 types at the finest fidelity. Section 4.2 introduces the datasets used during this research and reveals limited datasets in terms of size. Hence, a classification into 21 different types would cause a very low support for various classes, which is not a good setting for a supervised ML approach. Therefore we came up with three different taxonomies. The first system distinguishes between rights and obligations, including an additional fall-back class. Secondly, a taxonomy consisting of obligations, rights, references, definitions, legal consequences, and objections is proposed. A third taxonomy is shown in Table 2. We evaluated the different taxonomies from a legally theoretical perspective as well as from a technical one. For the technical analysis we annotated a dataset constituting the tenancy law of the German Civil Code (BGB) sentence-by-sentence with each of the taxonomies. We then trained two linear classifiers (SVM and logistic regression (LR)), as well as a decision tree (extra tree classifier (ETC)) using term-frequency (TF) on each dataset. Table 1 shows the results.

Classifier	F_1		
	3 Classes	6 Classes	9 Classes
SVM + TF	0.796	0.875	0.828
LR + TF	0.787	0.848	0.807
ETC + TF	0.817	0.874	0.831

Table 1. Technical analysis of the three taxonomies

The taxonomy consisting of three classes performed the worst. Moreover, the differentiation assumed by it is not sufficient for our purposes. While the taxonomy by Waltl et al. [5] seems to be more robust from a legal perspective, the six classes are from a technical point of view superior to the former. However, when applying NLP to the legal domain, we already learned, that domain knowledge is crucial and should not be neglected by technical assumptions. Hence the taxonomy described in Table 2 is used for this work. A detailed description of the thoughts behind this taxonomy can be found in [5].

4. Experimental Setup

4.1. Objective

Classifying sentences in legal contracts is, due to several reasons, attractive for the field of legal informatics. In the first place, it allows a more elaborate differentiation of a sen-

Semantic Type	Description
Duty	The primary function of a duty is to stipulate actions, inactions or states
Indemnity	The primary function of an indemnity is to clarify that, resp. under which conditions a duty does not exist
Permission	The primary function of a permission is to authorize actions, inactions or states
Prohibition	The primary function of a prohibition is to forbid or disallow actions, inactions or states
Objection	The primary function of an objection is to define that, resp. under which circumstances an existent claim may not be asserted
Continuation	The primary function of a continuation is to extend or limit the scope of application of a precedent legal statement
Consequence	The primary function of a consequence is to stipulate legal effects, without ordering or allowing character as far as the legal consequence part is concerned
Definition	The primary function of a definition is to describe and clarify the meaning of a term within the law
Reference	The primary function of a reference is to cite another norm with the aim of total or partial application transfer or non-application

Table 2. The used taxonomy created by us in an earlier work[5]

tence's meaning and thus enables subsequent contract analysis. Secondly, it is beneficial for information retrieval (IR) tasks in legal information databases and consequently supports the efficiency of e-discovery concerning legal documents. Last but not least, it helps determining dependencies and references between contracts and statutory documents.

Due to the lack of (annotated) data in this field, we are investigating on the portability of ML models. More precisely, the goal is to leverage the existing amount of statutory texts in order to create ML models which can be used to classify contract clauses or sentences.

4.2. Data

In order to prepare a proper setup for the legal sentence classification experiment, three different datasets were used. The first dataset compromises 601 sentences which constitute the tenancy law of the German Civil Code (§535-§597) in its consolidated version, effective from 21st of February 2017. Secondly, a dataset consisting of 169 sentences from rental agreements was required in order to test the trained ML models. Furthermore, this dataset was extended to 312 sentences, so that it could be used for training as well.

In a first preprocessing step, the raw text of these articles and clauses was segmented into sentences. Sentence segmentation in the legal domain can be a challenging task and results in a performance, yet inferior to other more common domains [16]. For this work a straight-forward rule-based approach by Waltl et al. [5] was chosen.

NLP typically involves various further pre-processing steps. Due to the fact that this work is based on a sentence classification problem and subject to be solved via supervised ML, the sentence segmentation was performed even before the actual NLP pipelines take place. Different normalization steps were incorporated into the varying pipelines and are discussed in Section 4.3.

Finally, the 913 sentences were manually classified by two human legal experts, according to the taxonomy described in Section 3. The annotations were performed using Gloss, the web-based annotation environment developed by Jaromir Savelka at the University of Pittsburgh. In this process, a third legal expert acted as the editor in order to decide on an annotation in the case of disagreement between the first two experts. The distribution of the different semantic types is revealed in Table 3. While some types have a very low occurrence, e.g. *Indemnity*, *Prohibition*, or *Definition*, some occur regularly, e.g. *Duty*, *Permission*, or *Consequence*.

Semantic type	BGB (n=601)		Rental agreements (n=169)		Rental agreements (n=312)	
	#	Relative (%)	#	Relative (%)	#	Relative (%)
Duty	117	19.5	52	30.8	105	33.7
Indemnity	8	1.3	0	0.0	1	0.3
Permission	148	24.6	35	20.7	75	24.0
Prohibition	18	3.0	2	1.2	3	1.0
Objection	98	16.3	8	4.7	4	4.5
Continuation	21	3.5	7	4.1	12	3.8
Consequence	117	19.5	64	37.9	101	32.4
Definition	18	3.0	1	0.6	1	0.3
Reference	56	9.3	0	0	0	0
	Σ 601	100	Σ 169	100	Σ 312	100

Table 3. Occurrence of the different semantic types in each of the datasets

4.3. Experiment

In order to measure the portability of ML models between document types, our experimental setting constituted four steps:

1. *Original Training:* We trained various classifiers on the BGB dataset and evaluated them using 10-fold cross-validation with 20% for testing.
2. *Portability Testing:* The resulting models were applied on the small set of rental agreements (n=169).
3. *Contract Training:* We extended the small set of rental agreements to 312 sentences and used it to train again various classifiers. The resulting models were evaluated on the same dataset using 10-fold cross-validation on 20% of the data.
4. *Portability Evaluation:* We applied the models from the original training on the new contract dataset (n=312). The results were compared with the performance from the contract training in order to assess the true portability.

Hereby, the classification of legal sentences using supervised ML was implemented following a basic workflow consisting of the following steps:

Data Acquisition: The data described in Section 4.2 was used[1].

Pre-Processing: We used three different pre-processing procedures: (1) The normalization (PRE) consisted of the removal of line breaks as well as duplicated whitespaces, replacing German umlauts, spelling numbers, and removing punctuation. (2) Stop word removal (SWR) was performed according to the spaCy[2] stop word list. (3) A lemmatization (Lemma) was conducted leveraging spaCy[2]. These three procedures were incorporated into pipelines in different combinations. Section 5.1 discusses the different variations.

[1] Available at: https://github.com/sebischair/Legal-Sentence-Classification-Datasets-and-Models

[2] https://spacy.io/usage/linguistic-features

Feature Extraction: Two different feature representations were used: (1) A bag-of-words approach was used to represent features. We used simple word count vectors and term frequency-inverse document frequency (TFIDF) transformer on these vectors. Where indicated, part-of-speech (POS) tags have been created and used as well. In order to keep the bag-of-words approach in this case as well, each token was combined with the respective POS tag using a dash. (2) The second feature representation leveraged word embeddings. We trained word2vec models on different legal corpora as well as used pre-trained models. These models were used to calculate the mean embedding of a sentence.

Training of Machine Learning Model: Six different classifiers were applied on the task of predicting the semantic types of legal sentences. We used MNB, LR, SVMs, multilayer perceptrons (P), random forests (RF), and a ETC. The models were trained using a 10-fold cross-validation on 80% of the dataset each iteration.

Evaluation and Error Analysis: Weighted variants of precision, recall, and F_1 were used to evaluate the performance of the trained models.

5. Evaluation and Error Analysis

5.1. Evaluating the portability

The objective of this experiment was to evaluate the portability of ML models with regard to different document types.

To achieve this, different classifiers were incorporated into various pipeline settings to train models on the BGB dataset. Waltl et al. [5] showed already that for simple pipelines, SVMs perform best on this dataset. For that reason, we initially trained the six classifiers by just relying on a simple count vectorizer (CV) as well as on TFIDF. Table 4 shows the performance of the models.

Classifier	Features	Precision	Recall	F_1
ETC	CV	0.814	0.836	**0.815**
	TFIDF	0.788	0.803	0.783
LR	CV	0.810	0.823	0.808
	TFIDF	0.724	0.749	0.719
MNB	CV	0.688	0.710	0.680
	TFIDF	0.699	0.646	0.616
P	CV	0.777	0.762	0.757
	TFIDF	0.798	0.780	0.777
RF	CV	0.728	0.718	0.705
	TFIDF	0.710	0.733	0.710
SVM	CV	0.838	0.839	**0.828**
	TFIDF	0.829	0.825	0.815

Table 4. Performance of the six classifiers on the BGB dataset

Two major observations can be made from this: (1) ETC and SVM perform the best, while (2) TFIDF creates worse results than simple TF. As a result, we tried several variations of combining different pipeline stages with these two classifiers based on TF. Furthermore we used sentence mean vectors by leveraging two pre-trained general word2vec models[3,4] as well as two manually trained word2vec models[5]. The results of the training are shown in Table 5.

ML classifier	Pipeline	Precision	Recall	F_1
ETC	CV	0.814	0.836	0.815
	PRE + CV	0.816	0.838	0.818
	PRE + SWR + CV	0.731	0.735	0.720
	PRE + SWR + Lemma + CV	0.701	0.698	0.688
	PRE + Lemma + CV	0.815	0.841	0.818
	Lemma + CV	0.810	0.831	0.809
	POS + CV	0.824	0.846	**0.827**
	POS + Lemma + CV	0.826	0.843	**0.825**
	SWR + POS + Lemma + CV	0.728	0.736	0.721
	word2vec JRCAquis	0.642	0.641	0.612
	word2vec Datev	0.661	0.652	0.623
	word2vec Google news	0.588	0.584	0.557
	word2vec Wikipedia	0.645	0.649	0.625
SVM	CV	0.838	0.839	**0.828**
	PRE + CV	0.838	0.834	**0.826**
	PRE + SWR + CV	0.748	0.743	0.735
	PRE + SWR + Lemma + CV	0.722	0.713	0.707
	PRE + Lemma + CV	0.830	0.823	0.816
	Lemma + CV	0.850	0.850	**0.839**
	POS + CV	0.831	0.831	0.823
	POS + Lemma + CV	0.839	0.839	**0.830**
	SWR + POS + Lemma + CV	0.703	0.702	0.694
	word2vec JRCAquis	0.687	0.716	0.691
	word2vec Datev	0.696	0.725	0.701
	word2vec Google news	0.622	0.636	0.614
	word2vec Wikipedia	0.680	0.666	0.658

Table 5. Performance of the best two classifier on the BGB dataset

At a first glance it is already obvious, that the bag-of-words approach outperforms the word embeddings by far. However, the word2vec models based on German legal corpora result in a greater F_1 than the pre-trained models. Our word2vec models were trained on two different corpora with default configuration, except dimensions is set to 300, window size to five and iterations to 10: (1) The JRCAcquis [17] with 33.686.085 token, and (2) a corpus consisting of judgments from the fiscal law constituting 114.091.840 token. The sizes are still pretty small for that matter. Therefore, further research is necessary to investigate in the suitability of word embeddings for such a classification task.

The highest F_1 was achieved by the following pipelines using a count vectorizer for feature extraction: (1) SVM using Lemma for pre-processing, (2) SVM using Lemma for pre-processing and the combination of the original POS tag along with the lemmatized token as features, (3) SVM without any pre-processing, (4) ETC without any pre-processing, but the combination of the token along with its POS tag as features, (5) SVM with PRE as pre-processing, and (6) ETC using Lemma as pre-processing and the combination of the lemmatized token along with its original POS tag as features. The F_1 varied from 0.839 to 0.825.

In the next step, the best resulting models were applied on the small contract dataset. Furthermore, the extended rental agreement dataset was used to train the six pipelines again, using a 10-fold cross-validation with 80% of the data. Afterwards, we took the

[3] https://devmount.github.io/GermanWordEmbeddings/

[4] https://code.google.com/archive/p/word2vec/

[5] Available at: https://github.com/sebischair/Legal-Sentence-Classification-Datasets-and-Models

six models from the original training and applied them on the bigger contract dataset as well. Table 6 shows the resulting F_1 measures.

Classifier	Pipeline	F_1			
		Train BGB	Test Rental (n=169)	Test Rental (n=312)	Train Rental (n=312)
ETC	POS + CV	**0.827**	**0.825**	0.720	0.713
	POS + Lemma + CV	0.825	0.789	0.709	0.718
SVM	CV	0.828	0.728	0.670	0.707
	PRE+ CV	0.826	0.705	0.667	0.682
	Lemma + CV	0.839	0.727	0.680	0.685
	POS + Lemma + CV	0.830	0.694	0.667	0.707

Table 6. Comparison of the performance on the BGB and contract dataset from the best classifiers

The column *Train BGB* captures the existing results from the first phase (see Table 5). The next column includes the results when applying the models from the BGB to the smaller contract dataset. A varying delta in F_1 from almost equal to zero (ETC with POS and CV) to 0.136 (SVM with POS + Lemma + CV) could be observed. This is already a first clue for a limited portability of ML models between document types in the legal domain. However, it is also obvious, that certain feature representations may be feasible for a portable model. Nonetheless, the actual performance of a classifier trained on contracts needs to be heeded yet. The last column (*Train Rental (n=312))* serves for this purpose. The models resulting in the training on the bigger contract dataset cease in an even worse performance between a F_1 of 0.718 and 0.682. These results were quite surprising. As a consequence, we conducted the original experiment phase (using all possible combinations of pre-processing, features and the six classifiers) again on the contract dataset. A F_1 measure of 0.734 however was the maximum, which is still significantly below the performance of the BGB models. Hence, we also applied the BGB models on the bigger contract dataset. The column *Test Rental (n=312))* reveals the respective results. The delta between the two models varies. The BGB model created with ETC and POS + CV performs better than the contract model. This model has already revealed a well suited portability in the previous step. For the remaining models however, the contract models outperformed the BGB model.

Our results reveal an evidence, that certain settings allow a portability between document types, even though portable models across-the-board are not given.

5.2. Error Analysis

To be able to better understand the classification process, but in particular the portability between document types, the worst portable configuration (SVM with POS + Lemma + CV) is examined in greater detail. Table 7 shows the performance of each class for the evaluation on the statutory dataset and the small contract dataset.

The resulting performance measures differ from the results shown in Table 6. This is because of different train test splits applied. While Table 6 is generated from a 10-fold cross-validation method, Table 7 is based on a static test split. The weighted mean mitigates the positive impact of small classes such as *Definition*, where only one instance is present in the test set, or even no instance as for *Reference*.

The results indicate that the portability issue of the models may be caused by the imbalance between the datasets. Particularly the types with a very low occurrence in

Type	BGB				Rental agreements (n=169)			
	Precision	Recall	F_1	#	Precision	Recall	F_1	#
Duty	0.783	0.857	0.818	21	0.693	0.736	0.684	53
Indemnity	1.000	1.000	1.000	1	0.250	1.000	0.400	1
Permission	0.919	0.810	0.861	42	0.944	0.829	0.883	41
Prohibition	0.333	0.500	0.400	2	0.000	0.000	0.000	1
Objection	0.800	0.923	0.857	13	0.454	0.714	0.556	7
Continuation	1.000	0.833	0.909	6	0.667	0.667	0.667	6
Consequence	0.731	0.950	0.826	20	0.660	0.525	0.585	59
Definition	1.000	0.769	0.870	3	0.000	0.000	0.000	1
Reference	1.000	0.769	0.870	13	0.000	0.000	0.000	0
Arith. mean (weigh.)	0.857	0.835	0.835	∑ 120	0.704	0.675	0.682	∑ 169

Table 7. Comparison of the performance by each type for the worst portable model

the contract dataset can have a huge impact on the results. Even though the language may vary between statutory texts and contracts, the characteristics of the different semantic types remain the same. In order to provide evidence for this hypotheses, we looked into the existing model (SVM with POS + Lemma + CV) and inspected the coefficients of each feature. The most important features for the class *Permission* are: (1) *können_verb*, (2) *berechtigen_verb*, and (3) *dürfen_verb*. For *Duties* the features with the highest weights are: (1) *muss_adj*, (2) *zu_part*, and (3) *verpflichten_verb*. Now it becomes also clear why SWR worsens the performance. Typical stop words such as *zu* depict important features for our models. This actually makes sense, since *zu* indicates the infinitive form and thus is crucial for determining the type of a sentence. Table 8 provides two examples per class from each corpora.

Type	Corpus	Sentence
Duty	BGB	Er hat die auf der Mietsache ruhenden Lasten **zu tragen**.
Duty	Contract	Insgesamt **zu zahlen** sind 2600 Euro.
Permission	BGB	Setzt der Mieter einen vertragswidrigen Gebrauch der Mietsache trotz einer Abmahnung des Vermieters fort, so **kann** dieser auf Unterlassung klagen.
Permission	Contract	Gegen Erstattung angemessener Kopier- und Portokosten **kann** der Mieter verlangen, dass ihm Kopien der Berechnungsunterlagen zugesandt werden

Table 8. Examples of Duties and Permission from both corpora

As one can see in Table 8, the most important features of the models are present in the examples. As a consequence, the models can properly represent these instances and thus classify them correctly. Looking at the most important features of the models, it seems obvious that using symbolic classification methods utilizing grammars or regular expressions may be more promising. However, Waltl et al. [5] examined such an approach already and could show the superiority of ML-based approaches.

6. Conclusion & Outlook

This work examined the portability of ML models with regard to different document types for the legal domain. Various classifiers were trained on the tenancy law of the German Civil Code and applied on a rental agreement dataset afterwards. Furthermore, the same settings were used to train models directly on the contractual dataset. We could show that ML models can be portable up to a certain degree in terms of document types.

Nonetheless, this research includes some limitations. The rental agreement dataset was pretty small for a supervised ML approach and differed in size in comparison to the statutory dataset. Furthermore, the class distribution varied between the two datasets. As a consequence, future research needs to define an even more suitable setting in terms of data distribution and size in order to provide more evidence on the portability of ML models. Yet, this work builds a solid base for future research in this area.

Another promising approach may be the incorporation of word2vec. Even though we have used word2vec features, it was not our focus and thus we have not investigated in greater detail into the vast amount of options concerning the training of word2vec models as well as the feature representations utilizing word2vec. Due to the nature of word2vec, capturing the semantics of words, it may be feasible for such a semantic classification task.

Lust but not least, this work did not look into domain portability of ML models, which is indeed another interesting and potentially helpful research field.

References

[1] M. Saravanan, B. Ravindran, and S. Raman, "Improving legal information retrieval using an ontological framework," *Artificial Intelligence and Law*, vol. 17, no. 2, pp. 101–124, Jun. 2009.

[2] K. D. Ashley, *Artificial intelligence and legal analytics: new tools for law practice in the digital age*. Cambridge University Press, 2017.

[3] Z. Milosevic, S. Gibson, P. F. Linington, J. Cole, and S. Kulkarni, "On design and implementation of a contract monitoring facility," in *First IEEE International Workshop on on Electronic Contracts*. IEEE, 2004, pp. 62–70.

[4] J. O. Neill, P. Buitelaar, C. Robin, and L. O. Brien, "Classifying sentential modality in legal language: a use case in financial regulations, acts and directives," in *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*. ACM, 2017, pp. 159–168.

[5] B. Waltl, G. Bonczek, E. Scepankova, and F. Matthes, "Semantic types of legal norms in german laws: classification and analysis using local linear explanations," *Artificial Intelligence and Law*, Jul 2018.

[6] B. Waltl, J. Muhr, I. Glaser, G. Bonczek, E. Scepankova, and F. Matthes, "Classifying legal norms with active machine learning," *Legal Knowledge and Information Systems*, p. 11, 2017.

[7] C. Biagioli, E. Francesconi, A. Passerini, S. Montemagni, and C. Soria, "Automatic semantics extraction in law documents," in *Proceedings of the 10th international conference on Artificial intelligence and law*. ACM, 2005, pp. 133–140.

[8] E. Francesconi and A. Passerini, "Automatic classification of provisions in legislative texts," *Artificial Intelligence and Law*, vol. 15, no. 1, pp. 1–17, 2007.

[9] E. de Maat, K. Krabben, and R. Winkels, "Machine learning versus knowledge based classification of legal texts." in *JURIX*, 2010, pp. 87–96.

[10] K. V. Indukuri and P. R. Krishna, "Mining e-contract documents to classify clauses," in *Proceedings of the Third Annual ACM Bangalore Conference*. ACM, 2010, p. 7.

[11] I. Chalkidis, I. Androutsopoulos, and A. Michos, "Obligation and prohibition extraction using hierarchical rnns," *arXiv preprint arXiv:1805.03871*, 2018.

[12] I. Chalkidis, l. Androutsopoulos, and A. Michos, "Extracting contract elements," in *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*. ACM, 2017, pp. 19–28.

[13] I. Chalkidis and I. Androutsopoulos, "A deep learning approach to contract element extraction." in *JURIX*, 2017, pp. 155–164.

[14] M. Curtotti and E. Mccreath, "Corpus based classification of text in australian contracts," in *Proceedings of the Australasian Language Technology Association Workshop*, 2010.

[15] I. Glaser, B. Waltl, and F. Matthes, "Named entity recognition, extraction, and linking in german legal contracts," in *Internationales Rechtsinformatik Symposium*, 2018.

[16] J. Savelka, V. R. Walker, M. Grabmair, and K. D. Ashley, "Sentence boundary detection in adjudicatory decisions in the united states," *Traitement automatique des langues*, vol. 58, no. 2, pp. 21–45, 2017.

[17] R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufis, and D. Varga, "The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages," *arXiv preprint cs/0609058*, 2006.

Legal Knowledge and Information Systems
M. Palmirani (Ed.)

doi:10.3233/978-1-61499-935-5-71

E-Science and the Law. Three Experimental Platforms for Legal Analytics

Nicola LETTIERI [a,1], Alfonso GUARINO [b] and Delfina MALANDRINO [b]
[a] *National Institute for Public Policy Analysis (INAPP), 00198, Rome, Italy*
[b] *Department of Computer Science, University of Salerno, 84084, Fisciano, Italy*

Abstract. The paper presents three experimental platforms for legal analytics, online environments integrating heterogeneous computational heuristics, information processing, and visualization techniques to extract actionable knowledge from legal data. Our goal is to explore innovative approaches to issues spanning from information retrieval to the quantitative analysis of legal corpora or to the study of criminal organizations for research and investigative purposes. After a brief introduction to the e-science paradigm and to the role played in it by research platforms, we focus on visual analytics as a viable way to interact with legal data. We then present the tools, their main features and the results so far obtained. The paper ends up with some considerations about the computational turn of science and its role in promoting a much needed interdisciplinary and empirical evolution of legal research.

Keywords. e-science, analytical platforms, legal visual analytics, network analysis

1. Introduction

Last few years have been marked by a growing integration of traditional research methods and data-driven computational heuristics. The term *E-science* [1] today embraces in one definition the features of an emerging research paradigm in which every stage of the scientific endeavor, from the formulation of research questions to the distribution of findings, is somehow ”enhanced” by digital information-processing, computational technologies and distributed collaboration infrastructures. The paradigm is spreading not only in empirical research [2]- mainly through big data analytics and machine learning - but also in theory-making - mostly by means of computer simulation model building [3]. In this scenario, analytical platforms - software environments integrating different tools of the e-science pipeline - are becoming the cornerstone of a change that challenges established epistemologies in all the areas of science with a peculiar impact in the social sciences where computational heuristics have spread at a slower pace. Against this backdrop, a major challenge for legal scholars is to start drawing up new tools capable of exploiting the data streams and the computing power today available to answer existing

[1] Corresponding Author: Nicola Lettieri; E-mail: n.lettieri@inapp.org

research questions or come up with new ones. This work presents three experimental analytical platforms combining data mining, visualization and machine learning to extract knowledge from the analysis of legal materials. Projects presented explore new ways to address the needs of legal science and practice. In the background, the belief that the computational turn of science is providing precious food for thought for a much needed empirical, quantitative and interdisciplinary evolution of legal research.

2. Research platforms and Visual Analytics

The evolution of ICT has greatly pushed forward the computational science paradigm: today almost every area of science is shifting towards the model of a research powered by machines [4] and computing intruments. We are witnessing the spread of data and computation-driven research platforms supporting scientists in different ways (for an overview see: [5]): allowing to more easily explore the ever growing amount of scientific papers today available (e.g. *PubChase*); supporting the analysis and the management of large sets of data and programming code (e.g. *GitHub*); facilitating the collaboration between colleagues and the handling of online experiments (e.g. *LabGuru, Asana*); simplifying the publication of papers and the analysis of their impact(e.g. *eLife, GigaScience*).

Tools above mentioned are all somehow triggering innovation in the practice of science. Particularly interesting is, in this evolving scenario, the possibility of using research platforms to combine the insights offered by computational heuristics, with intuitive visualizations allowing to better make sense of the interaction with huge and often obscure amounts of data. This is even more true in those fields of social sciences that, like law, are less familiar with quantitative approaches and advanced computational methods.

A promising frontier in this regard is represented by the adoption of methods and technical solutions coming from *Visual Analytics*(VA) [6], a fledgling research field aiming to provide scientists with innovative ways to turn data into knowledge while also enabling them to act on their findings in real-time. As highlighted in [7], VA explores new ways to: i) synthesize information and derive insights from massive, dynamic, ambiguous, and often conflicting data; ii) detect the expected and discover the unexpected; iii) provide timely, defensible, and understandable assessments; iv) communicate these assessments effectively for action.

3. Visualization, analytics and law

As a matter of fact, the idea of exploiting visual metaphors to ease the management and the understanding of legal information has repeatedly made its appearance in the history of law. The use of charts and maps dates back to the Middle Ages when the so called *"arbor"* (*"arbor"* is the Latin word standing for *"tree"*) diagrams [8,9] were used to graphically exemplify legal concepts like the impediments to marriage, or to depict the stages of procedure in Roman Law. Legal metaphors appear again centuries later, when Henry Wigmore [10] proposed the use of diagrams - the *"Wigmore charts"* - to support the analysis of ambiguous evidences and facilitate reasoning required to confirm or rebut hypotheses presented in court. Over the years, the interest in graphical methods led not only to the implementation of Wigmore diagrams through computational tools, but also to other visual representations of legal matters like Bayesian networks that support probabilistic inference in forensic science [11,12,13,14,15,16]. In more recent times, the availability of increasingly powerful technologies (e.g. user-friendly visualization tools and data mining libraries) and of insightful computational heuristics has prompted a growing interest in the development of advanced tools for the analysis of legal information. This is witnessed by a number of experiences [17,18] at the boundaries between visualization, analytics and law, in an area we could define as *Visual Legal Analytics* (VLA).

Some examples can be useful to get an idea of trends emerging in the field. *Ravel Law* is a commercial web platform for computer-assisted legal research that integrates machine learning, natural language processing, and visualization to help lawyers and legal scholars in retrieving and analyzing US case law. While traditional legal search engines use textual interfaces offering a poorer user experience, *Ravel Law* exploits graph visualization not only to allow the access to case law full texts, but also to convey information about cases like the relevance of precedents or the connections between judgments. Another experience worthy of attention is *Lexmex*, an online system visually representing the relations between the French Civil Code and other pieces of legislation. The tool generates a graph transforming laws in nodes and citations in edges: the size of nodes depends on the number of connections between the nodes, while the colors allow to identify groups of highly connected norms. Interaction with data is enabled by essential navigation solutions such as zooming, node selection (to show contextual information) and search by keywords. A more recent and interesting work [19] presents an open source software for the analysis and the visualization of citation network of Dutch case law. The goal is to support legal research questions, including the identification of relevant judgments, the comparison of precedents with those identified in the literature, and the determination of clusters of related cases. In a similar direction, again, we can cite a project using visualization to depict and explore the history of Swiss Federal Law [20].

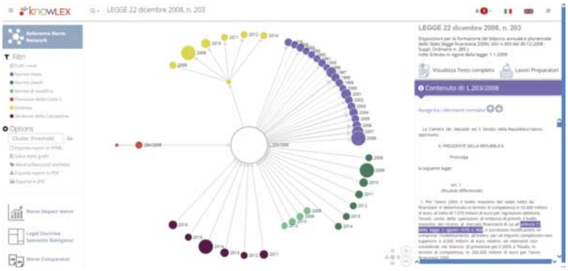
4. Three experimental platforms for legal analytics

In this section we present *KnowLex*, *EUCaseNet* and *CrimeMiner*, three experimental research platforms [21,22,23] exploiting e-science methods to meet in new ways the needs of legal science and practice. The description of the features of each tool is accompanied by a brief presentation of the results obtained so far.

4.1. Mapping and exploring norms' "neighborhood": KnowLex

KnowLex is a web application (https://bit.ly/2PfmqPu) designed for visualization, exploration, and analysis of legal documents coming from different Italian sources and connected to a given piece of legislation. Understanding the legal framework relating to a given issue often requires the analysis of complex legal corpora: when legal professionals or citizens try to understand how a given phenomenon is disciplined, their attention cannot be limited to a single source of law, but has to be directed on the bigger picture resulting from all the sources related to the theme under investigation. *KnowLex* exploits data visualization and quantitative analysis to support this activity by means of interactive maps making sense out of heterogeneous documents (norms, case law, legal literature, etc.) and their properties. The tool results from an analysis conducted with legal professionals and students and has already undergone a preliminary evaluation study aiming at evaluating the effectiveness of visualization (compared with that of textual interfaces of traditional database), the usability of the proposed system, and the overall user satisfaction [22].

Features. *KnowLex* is made up of several modules offering different insights and ways of interacting with available data. The "*Reference Norm Network*"(RNN) uses interactive graphs to represent the set of materials connected to a given piece of legislation. *KnowLex* gathers documents (amendments, Supreme Court judgments, constitutional judgments, preparatory works, legal literature) from different datasets and websites starting from a norm chosen by the user (the"Root" norm", Legge 22 Dicembre 2008, n. 203", in our example), and builds a map connecting all of them. The graph (see Fig.1) not only offers an overall view of documents properties and relations but also allows user to access text and information simply by interacting with the nodes. The *Semantic Doctrine Navigator* module (Figure 2) shows a treemap depicting the tags used to classify the corpus of scientific articles published with reference to the *Root norm* exploiting the classification by subject used by the Italian bibliographic database DOGI. The map allows to: i) visually explore the law's impact on the different areas of the legal system (e.g., if 70% of the articles related to a certain law is tagged administrative law, it is likely that this is the field on which the law has had the most impact); ii) understand how the doctrine has evolved and what themes have drawn the attention over time (areas have a color that varies according to the date of the articles with a given tag); iii) retrieve papers abstracts and bibliographic data by clicking on the tiles. The *Norm Impact Meter* module (Figure 3) combines graphs relating to different categories of documents linked to the *Root Norm* (amendments, repeals, citations contained in other laws, judgments of different authorities that apply the norm, and reviews of constitutionality). The visualization allows to extract a coarse-grained quantitative image of the laws impact on the legal system. The *Norm Comparator* module (see Figure 4) exploits data relating to the papers tags from the *Semantic Doctrine Navigator* to compare two laws in semantic terms. Different views allow the user to understand: *i)* whether two given laws have dealt with the same topics (which is represented with an histogram), and *ii)* how much the two laws are similar to each other (which is done by calculating Euclidean distance and the Pearson



LEGGE 22 dicembre 2008, n. 203

Contenuto di: L.203/2008

IL PRESIDENTE DELLA REPUBBLICA

Promulga

la seguente legge:

Art. 1

Figure 1. KnowLex: the Norm Graph Navigator.

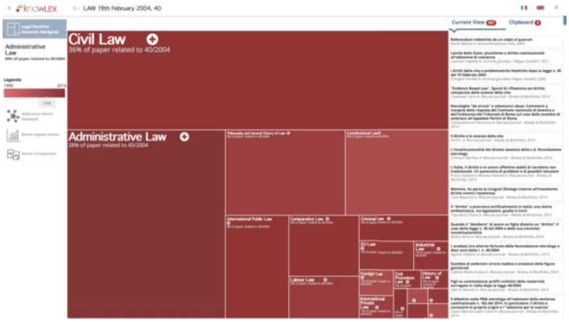


Figure 2. KnowLex: Semantic Doctrine Navigator.

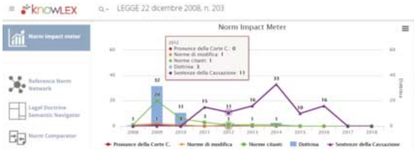


Figure 3. KnowLex: Norm Impact Meter.

correlation coefficient and is represented through gauge meters). The feature becomes interesting when used to compare two norms that have the same function like the Finance acts. A different "semantic fingerprint" of the two laws suggests that the legislator has focused his attention on different priorities (e.g., public education rather than health-care) in two budget years.

Technologies. From the technological point of view, KnowLex is built on a three-layer architecture: on the client-side, it has been developed using JavaScript open-source libraries (i.e., *Sigma.js*, *Linkurious.js* and *D3.js*). On the server-side, data are gathered through HTTP requests using cURL, while PHP wrappers parse different external sources and produce structured data in JSON format. Data are stored in a MySQL database.

4.2. Analyzing the features of EU case law: EUCaseNet

EUCaseNet is an online laboratory allowing legal scholars to explore the EU case law corpus in real time using computational heuristics and visualization techniques. The tool is freely accessible (https://bit.ly/2yVTFNr) and has already allowed [23] a series of interesting experiments about the advantages potentially deriving from network-based inferences in the discovery of features characterizing both the EU case law as a whole, and single judgments (e.g. the relevance of precedents)

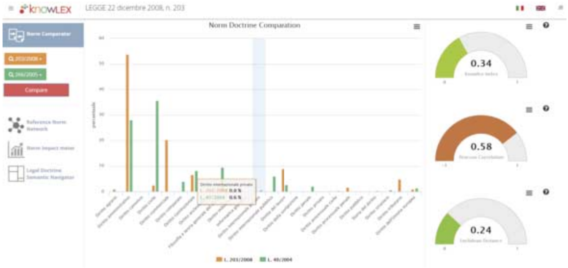


Figure 4. KnowLex: Norm Comparator.

Features.*EUCaseNet* offers basically three functionalities. *Citation Analysis*. EUCaseNet allows the interactive application of NA measures (centrality, Page Rank, Community detection etc.) to the network of the citations connecting all the ECJ judgments. The size and the color of the nodes can be varied based on the value of specific NA metric (e.g. betweenness centrality) allowing users to visually explore the potential correspondence between the results of computational analysis and the features of judicial decisions like judges' behavior (e.g., the propensity to cite other cases) or the nature and the features of the most cited precedents (Figure 5). *Topics and trends analysis*.A heatmap (Figure 6) allows to visually compare the topics most covered by the entire EU case law (a sort of trending topics dened by the system using ofcial ECJ data). The visualization allows to intuitively understand if a given ruling (e.g. a judgment recognized as particularly important by legal literature) deals with issues to which the EU case law has already reserved particular attention in the past. The measurement, moreover, suggests objective indices of the emergence of new trends, offering insights for further investigations. *Evolution of case law topics* A linegraph (Figure 7) depicts, in diachronic terms, some aspects of the evolution of the EU case law allowing to see how the number

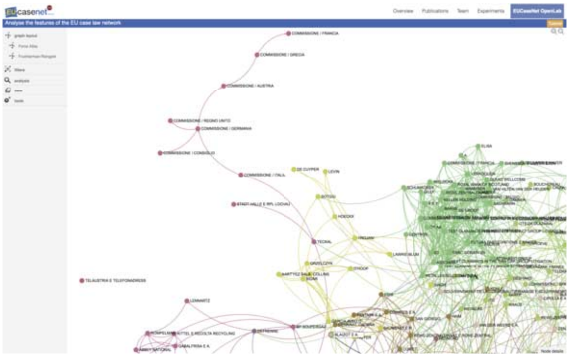


Figure 5. EUCaseNet: citations network graph of the first 350 judgments in terms of in-degree and their citation network.

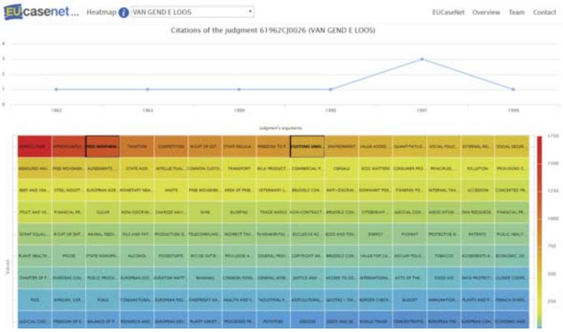


Figure 6. EUCaseNet: heatmap.

of judgments dealing with a given topic (e.g. free movement of goods in EU) evolved over time. The possibility of overlaying several lines related to different topics, allows legal scholars (e.g., historians of law) to make useful comparisons and to identify notable correlations in the changes of most frequent topics dealt with by ECJ.

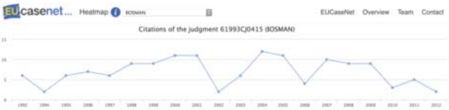


Figure 7. EUCaseNet: linegraph.

Technologies. *EUCaseNet* is built upon a three-tier architecture, implemented by following a typical Model-View-Controller layer architecture. The Data Persistence and Business layers are implemented server-side, through Java Servlet components, within Apache Tomcat. The User Interface Layer is implemented with commonly used JavaScript libraries (i.e., Sigma.js).

4.3. A social-legal analysis of organized crime: CrimeMiner

CrimeMiner is a platform aiming to explore how the combination of data mining, SNA, machine learning and data visualization can contribute to a deeper understanding of structural and functional features of criminal organizations starting from the analysis of even simple relational and investigative data. The tool, aimed at both scientific and investigative purposes, has been developed in collaboration with public prosecutors of the Italian Direzione Investigativa Antimafia (the Italian Investigative Directorate anti-mafia) and has been validated within a case study based on data coming from real criminal investigations [21]. Data currently handled by *CrimeMiner* consist in people records/charges and telephone/environmental tappings and they are visualized as a graph G=(V, E) where V=people and E=telephone/environmental tapping shaping the relationship among individuals that can be analyzed using SNA metrics.

Features. *CrimeMiner* offers visualizations that illuminate different aspects of the criminal group under investigation. The *Wiretaps graph* (Figure 8) offers an intuitive view of the communications/social interactions network within the organization. User can apply different SNA metrics (e.g. community detection algorithms, PageRank) to retrieve insights about the social role of single individuals (leader, broker etc.) or the features of subcommunities (e.g. specialization in given criminal activities).

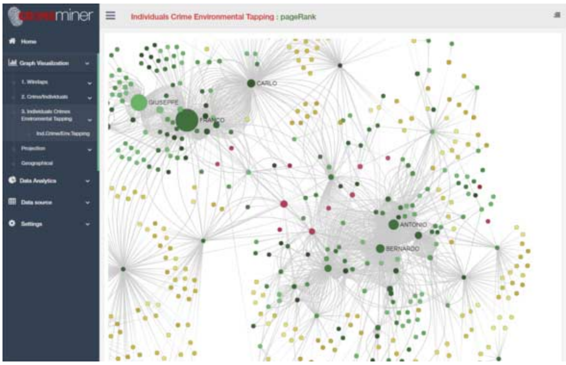


Figure 8. CrimeMiner: individual wiretaps graph

The *GIS map* displays on a map the geographical position (place of residence) of the people under investigation, preserving information about the properties (degree, betweenness, Page Rank values etc.) of single nodes. This allows to discover potentially invisible relationships between the features of the node and its geographical position.

Technologies. *CrimeMiner* is built upon the Java EE Spring Data *Neo4j* framework whose architecture is structured in four layers. In details: *i)* the *Storage layer* stores all data (including graphs) under examination, such as personal details of investigated people and tapping records; *ii)* the *Mapping layer* is responsible of the mapping of *Neo4j* relations and entities in Java classes; *iii)* the *Business layer* processes data mapped in the Mapping layer and provides developed services to the top layer (SNA metrics are defined in within this layer); *iv)* the *Presentation layer* includes the user interface allowing users to interact with *CrimeMiner* features (exploiting Linkurious.js and HighCharts.js).

5. Conclusions

The projects and the tools above described are still very preliminary attempts to explore potential intersections between the legal world and some of the emerging e-science methodologies, and will thus be subject of further developments. On the other hand, despite the publication of some works [24], the adoption of the e-science techniques in legal contexts is still in its infancy and in-depth reflections will be necessary to better exploit them. What is probably before us is in any case the opportunity of a significant methodological and scientific change, especially if we look at analytics and visualization not simply as ways to ease the access to legal materials, but as unprecedented solutions to delve into the complexity of legal world. Actually, collecting data and extracting knowledge from them is contributing to change the whole infrastructure of practices, technologies and perspectives in social sciences [25], and this is going to affect ever more legal research. From this viewpoint, our experience has resulted also into an opportunity to reflect in general terms on the future of the law both on a scientific and methodological standpoint. A first outcome of our reflection is a clearer idea of the gradual shift of law to what has been called the "machine science" paradigm [4], a research model that promises unprecedented scientific insights thanks to a wise combination of theories, data and code. In the perspective of an instrument-enabled future of law - similar arguments can be made for the evolution of legal enforcement towards techno-regulation [26,27,28] - legal scholars will ever more be involved in the challenge of designing tools, including platforms, with implications unfolding also on the theoretical level. It will be up to lawyers to gradually learn new skills and languages (from the computational to the technical ones), so to rethink their research questions, conceptual categories, methods of study and relationships with other sciences. A second worthy result of the work so far done, is the hunch of considering e-science methods as the entry point, in the legal field, of what has been called "computational empiricism" [29,30], the new perspective of an empirical research mainly rooted in the power of data and computational heuristics. After all, the project of what we call "legal computational empiricism", is consistent with some of the emerging trends in the debate on the future of legal science. As witnessed by the flourishing of fields like Empirical Legal Studies [31] or New Legal Realism [32], last few years have been marked by growing attention for the application to legal issues of empirical methods developed in the social sciences. Thanks to their capacity to integrate increasingly advanced ways of tracking and measuring reality, analytical platforms will probably be part of the empirical (experimental, quantitative) evolution of the legal science with an impact on the way scholars and practitioners think about their goals and methods. What is needed is not a sloppy juxtaposition of methods and scientific perspectives but a serious work on the theoretical and experimental level. To use the words of Franz Leeuw [33], "*the more empirical legal research is a "growth industry", the more important it is to understand and discuss epistemological problems of this field of study*" tackling with fundamental issues including "*how to operationalise legal concepts, where to find data (stored, but also Big Data)*" and, "*how to bring empirical evidence to the fore, in such a way that it can be understood and used by lawyers, legislators and regulators*". Research at the borders between e-science and law is not the only solution, but will be for sure part of the effort.

References

[1] T. Hey and A. Trefethen, "The data deluge: An e-science perspective," *Grid computing: Making the global infrastructure a reality*, 2003.

[2] R. Kitchin, "Big data, new epistemologies and paradigm shifts," *Big Data & Society*, 2014.

[3] E. Winsberg, *Science in the age of computer simulation.* University of Chicago Press, 2010.

[4] J. Evans and A. Rzhetsky, "Machine science," *Science*, 2010.

[5] N. Lettieri, A. Altamura, R. Giugno, A. Guarino, D. Malandrino, A. Pulvirenti, F. Vicidomini, and R. Zaccagnino, "Ex machina. Analytical platforms, Law and the challenges of Computational legal science," *Future Internet*.

[6] J. Kohlhammer, D. Keim, M. Pohl, G. Santucci, and G. Andrienko, "Solving problems with visual analytics," *Procedia Computer Science*, 2011.

[7] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon, "Visual analytics: Definition, process, and challenges," in *Information visualization*, Springer, 2008.

[8] A. Errera, *Arbor actionum: genere letterario e forma di classificazione delle azioni nella dottrina dei glossatori*. Monduzzi Ed., 1995.

[9] C. Radding and A. Ciaralli, *The Corpus iuris civilis in the Middle Ages: Manuscripts and transmission from the sixth century to the juristic revival*. Brill, 2006.

[10] J. H. Wigmore, "Problem of proof," *Ill. LR*, 1913.

[11] P. Tillers, "Picturing factual inference in legal settings," 2005.

[12] A. Biedermann and F. Taroni, "Bayesian networks and probabilistic reasoning about scientific evidence when there is a lack of data," *Forensic science international*, 2006.

[13] F. Taroni, C. G. Aitken, P. Garbolino, and A. Biedermann, *Bayesian networks and probabilistic inference in forensic science*. Wiley Chichester, 2006.

[14] A. B. Hepler, A. P. Dawid, and V. Leucari, "Object-oriented graphical representations of complex patterns of evidence," *Law, Probability & Risk*, 2007.

[15] T. F. Gordon, "Visualizing carneades argument graphs," *Law, Probability and Risk*, 2007.

[16] B. Verheij, "Argumentation support software: boxes-and-arrows and beyond," *Law, Probability and Risk*, 2007.

[17] R. Winkels, N. Lettieri, S. Faro, *et al.*, *Network analysis in law*. Napoli Edizioni Scientifiche Italiane9788849527698, 2014.

[18] R. Whalen, "Legal networks: The promises and challenges of legal network analysis," *Mich. St. L. Rev.*, 2016.

[19] D. V. KUPPEVELT and G. V. DIJCK, "Answering legal research questions about dutch case law with network analysis and visualization," in *Legal Knowledge and Information Systems: JURIX 2017: The Thirtieth Annual Conference*, IOS Press, 2017.

[20] S. N. André Ourednik, Peter Fleer, "A Visual Approach to the History of Swiss Federal Law," in *DHd 2016: Modelling - Networking - Visualization*, 2016.

[21] N. Lettieri, D. Malandrino, and L. Vicidomini, "By investigation, i mean computation," *Trends in Organized Crime*, 2017.

[22] N. Lettieri, A. Altamura, and D. Malandrino, "The legal macroscope: Experimenting with visual legal analytics," *Information Visualization*, 2017.

[23] N. Lettieri, A. Altamura, A. Faggiano, and D. Malandrino, "A computational approach for the experimental study of EU case law: analysis and implementation," *Social Netw. Analys. Mining*, 2016.

[24] E. Gomez-Nieto, W. Casaca, I. Hartmann, and L. G. Nonato, "Understanding large legal datasets through visual analytics," in *Proceedings of 6th Workshop on Visual Analytics, Information Visualization and Scientific Visualization (WVIS) in SIBGRAPI*, 2015.

[25] J.-C. Plantin, C. Lagoze, and P. N. Edwards, "Re-integrating scholarly infrastructure: The ambiguous role of data sharing platforms," *Big Data & Society*, 2018.

[26] R. Leenes, "Framing techno-regulation: An exploration of state and non-state regulation by technology," *Legisprudence*, 2011.

[27] R. Brownsword, "In the year 2061: from law to technological management," *Law, Innovation and Technology*, 2015.

[28] P. De Filippi and S. Hassan, "Blockchain technology as a regulatory technology: From code is law to law is code," *arXiv preprint arXiv:1801.02507*, 2018.

[29] P. Humphreys, *Extending ourselves: Computational science, empiricism, and scientific method*. Oxford University Press, 2004.

[30] P. Humphreys, "Computational empiricism," in *Topics in the Foundation of Statistics*, Springer, 1997.

[31] P. Cane and H. Kritzer, *The Oxford handbook of empirical legal research*. OUP Oxford, 2010.

[32] T. J. Miles and C. R. Sunstein, "The new legal realism," *U. Chi. L. Rev.*, 2008.

[33] F. L. Leeuw, "Empirical legal research the gap between facts and values and legal academic training," *Utrecht L. Rev.*, 2015.

Legal Knowledge and Information Systems
M. Palmirani (Ed.)

doi:10.3233/978-1-61499-935-5-81

Chronicle of a Clash Foretold: Blockchains and the GDPR's Right to Erasure

Ugo PAGALLO[a,1], Eleonora BASSI[b], Marco CREPALDI[a], and Massimo DURANTE[a]
[a] *University of Turin, Law School*
[b] *Politecnico of Turin, Nexa Center*

Abstract. GDPR abiding blockchain systems are feasible. Jurists, programmers, and other experts are increasingly working on this aim nowadays. Still, manifold blockchain networks functioning out there suggest a new generation of data protection issues brought about by this technology. Some of these issues will likely concern the right to erasure set up by Art. 17 of the EU data protection regulation ('GDPR'). These cases will soon be discussed before national authorities and courts, and will likely test the technical solutions explored in this paper, such as hashing-out methods, keys destruction, chameleon hash functions, and more. By taking into account matters of design and the complex architecture of blockchains, we shall distinguish between blockchains that have been thought about to expressly meet the requirements of the EU regulation, and blockchains that, for one reason or another, e.g. ante GDPR designed blockchains, trigger some sort of clash with the legal order, that is, (i) matters of principle on e.g. political decentralization; (ii) standards on security and data protection; (iii) a mix of them; and, (iv) social clash. It is still unclear how the interplay of legal regulation, technological constraints, social norms, and market interests, will end up in this context. Rulings and court orders will be instructive. It is a clash foretold, after all.

Keywords. Blockchain, chameleons hash functions, data protection, encryption, hashing-out methods, right to erasure.

1. Introduction

There is a multi-faceted parallel between current blockchain technologies and peer-to-peer ('P2P') systems, i.e. the massively distributed platforms for information storage and retrieval, which became popular in the late 1990s with the Napster case [1]. Among their features, such as tamper-proof and append-only properties, blockchains are a kind of P2P network [2]. As occurred with the spread of such distributed and decentralized systems twenty years ago [3], many claim that blockchains can strengthen social interaction to such an extent, that the creation of a libertarian cyberspace, a direct online democracy, or even a digital form of communism could be at hand [4].

There is the other side of the coin, though. Similar to the first wave of P2P networks in the 1990s and the 2000s, blockchains raise several legal issues. The DAO case, namely, a short-lived experiment that aimed to create a decentralized, directly managed crowdfunding and investment vehicle to back development projects on the Ethereum blockchain, illustrates this point with a theft of millions of dollars [5]. One of

[1] Ugo Pagallo, Law School, University of Turin, Lungo Dora Siena 100 - 10135 Torino, Italy; E-mail: ugo.pagallo@unito.it.

our main contentions will be that such legal issues of the blockchain are systemic, that is, they concern the very architecture of blockchain, as occurred with P2P technology two decades ago. What, then, was the main legal issue of these latter systems, which can help us understand the most relevant challenges triggered by the functioning and design of blockchains in the legal domain today?

After the legal misadventures of Napster—the first popular file-sharing system on the internet, which bankrupted in September 2002 after a copyright lawsuit filed by the US record industry association—the main legal issue of P2P systems revolved around whether or not this technology is "incapable of non-infringing uses." Three years later, in the 2005 *Grokster* case, the US Supreme Court's decision had to clarify to what extent technologies promoting the ease of infringing on copyrights have to be condemned, so that producers of P2P software, like Grokster and Steamcast, could be sued for "inducing copyright infringement committed by their users." In connection with the figures of the Wikipedia entry, according to which "90% of files shared on Grokster were downloaded illegally," the point of the claimants was clear: plaintiffs claimed that infringing uses of P2P technology constituted the primary aim of such systems. Although Justices in Washington did not buy this argument, they unanimously held that companies could be sued for inducing copyright infringements for acts taken in the course of marketing file sharing software. A more distributed and decentralized generation of P2P systems followed as a result of the legal arguments of the Court. The "servent" (server/client) architecture of these networks was adapted to meet the legal constraints of copyright [1, 6]. All in all, we reckon that something similar will occur in the EU law with most of today's blockchain architectures in the field of privacy and data protection.

Next, Section 2 illustrates which specific features of blockchain technologies and hence, what current blockchain networks are under scrutiny in this paper, e.g. closed or open blockchains, rather than, say, Turing-complete or incomplete blockchains. Since some of the data stored in these blockchains are personal, Section 3 takes into account current EU regulation on data protection, i.e. the "GDPR," and more particularly Article 17 on the right to erasure. Certain provisions of Art. 17 appear simply at odds with some properties of blockchains, such as their "immutability." Section 4 considers some solutions for either abiding by the rules of the GDPR, or preventing blockchains from processing personal data. Each solution, so far, has its own drawbacks, as shown by methods of "hashing-out," of keys destruction, of chameleon hash functions, and more. This leads to a paradox. Should we admit that no one-size-fits-all solution exists in this context, except from banning the technology [7]? Should Western countries, e.g. the EU under the GDPR, follow Chinese suit, and throw out the baby (blockchain) with the dirty water (e.g. data illegally stored that have to be erased)?

The conclusion of the paper brings us back to the parallel between the legal fate of P2P systems and current debate on blockchains. As illustrated by the former's cases in the early 2000s, an outright ban of blockchain systems that do not provide mechanisms for erasing data illegally stored is for real. In the jargon of the US Supreme Court, we shall distinguish between technologies capable, or "incapable of non-infringing uses." The paper examines feasible technical solutions, much as the role that social norms and the forces of the market may play in this context. The forecast is twofold: on the one hand, it is likely that the "Grokster mechanism" will reappear in the EU under the GDPR. There will be a new generation of blockchain networks, whose design and architecture intend to abide by the EU provisions on data protection. Yet, on the other hand, we should consider that some current blockchains, such as Bitcoin, operate in more than 100 different jurisdictions with lack of formal governance, thus suggesting further problems of enforcement. The final outcome of this interplay between law and technology is of course far from clear and, nevertheless, the clash between legal and technological regulations, economic interests, or social values, e.g. trust, seems inevitable. In homage to Gabriel García Márquez 1981 novel, let us start our "chronicle of a clash foretold."

2. Blockchains

There are many definitions of blockchain out there. Some present blockchains as P2P, append-only, tamper-proof, ever-growing distributed and decentralized networks that function as records for transactions [2]. Others insist on the properties of the network that allow the nodes to agree on the order, validity, existence, and authenticity of all the transactions ever occurred within the system [8]. In a nutshell, blockchains link chunks of data together in blocks by including the hash of the previous block, i.e. the function used to map data of arbitrary size to data of a fixed size. By defining the height of a block as the number of blocks in the chain between it and the genesis block, whose height is 0, it follows that a block with height *x* includes the hash of the block with height *x-1*, the block with height *x-1* includes the hash of the block with height *x-2* and so on. This is crucial. Any modification to the data stored in a block necessarily entails having to re-compute all the blocks that came after it, otherwise the system rejects the modified block. We return to this issue later in the next section, where the provisions of the GDPR's right to erasure are illustrated with their pros and cons.

Still, there are different kinds of blockchains, e.g. Turing-complete, or incomplete systems [4]. In this context, dealing with Art. 17 of the GDPR, the distinction between open and closed blockchains is particularly relevant. Open blockchains, such as Bitcoin and Ethereum, are "permissionless"; closed blockchains, such as Corda and Hyperledger Fabric, are vice versa "permissioned," i.e. participants in the system are known. Contrary to politically decentralized blockchains, in which nobody controls the network, permissioned blockchains are "politically centralized" [9]. Therefore, notions of "data processors" and "data controllers" do not seem particularly problematic [7, 10]. They regard at least the validators of the network in light of Articles 24 and 28 of the GDPR.

Things appear legally more complex in the case of some open blockchains. Their design aims to make impossible to tamper the information, namely, once the information is appended on a blockchain according to the protocol rules, it is impossible to alter it without compromising the entire network. In addition, such information should persist through time, because the blockchain has to record all the history of the previous states of the system without ever deleting any part of it. Some argue that blockchains are not immutable but rather, hard to change from the users' side [11]. The aforementioned DAO case, or the Ethereum improvement proposal n. 999, shed light on why some blockchains are better understood as immutable for users and non-users of the technology, while simply hard-to-change for the gatekeepers of the system. Therefore, as occurs with validators of closed blockchains, notions of "data processors" and "data controllers" do not appear particularly complex in this second scenario [7, 10]. Responsibilities related to such notions regard at least the gatekeepers of the network.

But, how about "politically decentralized" blockchains [9]? Should we represent all the nodes in the network as data processors, or even data controllers [7, 10]?

Data processed by blockchains concern two different kinds: system-dependent data and arbitrary data. The former is necessary for the functioning of the system: public network addresses, cryptographic primitives, input and outputs of transactions, or private keys, are examples of system-dependent data. On the other hand, most open blockchains allow users to store arbitrary data, namely, any kind of data. Whilst some system-dependent data might be considered personal data under the GDPR [11], others have found that Bitcoin's blockchain already stores more than 1600 arbitrary files [12], some of which likely contain personal data. There are in addition several methods that allow users to store arbitrary data on Bitcoin's blockchain [13]. Correspondingly, we may suspect that sooner, rather than later, matters of data protection will concern the current functioning of such blockchain networks. Next section explores what kind of obligations data protection and data controllers have in the case of the right to erasure, and how the enforcement of this right may work with different kinds of blockchain in the EU legal system.

3. GDPR's Art. 17

The provisions of Art. 17 are divided into three sections, which correspond to (i) the substantial grounds of the right; (ii) its mechanisms; and, (iii) restrictions, rather than exceptions, of the right to erasure. As to the grounds of the right, data subjects can exercise it—i.e. to obtain from the data controller the erasure of their personal data without undue delay—under six circumstances. According to Art.17(1), they regard (a) cases in which such data are no longer necessary in relation to the purposes for which they were collected, or otherwise processed; (b) withdrawal of consent pursuant to Art. 6(1)(a) and Art. 9(2)(a); (c) objection to the processing pursuant to Art. 21(1); (d) unlawful processing; (e) legal obligations to which the controller is subject; and, (f) personal data that have been collected by information society-services when the data subject was a 'child.'

As to the mechanism set up by Art. 17(2), the data controller has not only to erasure the data. It "shall take reasonable steps, including technical measures, to inform controllers which are processing the personal data that the data subject has requested the erasure by such controllers of any links to, or copy and replication of, those personal data." This duty of information by the first data controller shall be related to the "available technology" and "cost of implementation."

As to the limits of the right, Art. 17(3) establishes five cases in which the right to erasure does not apply. They concern (a) the right of freedom of expression and information; (b) compliance with certain legal obligations, to which the controller is subject, e.g. the performance of a task carried out in the public interest; (c) reasons of public interests pursuant to Art. 9(2) and (3); (d) archiving purposes in the public interest, scientific or historical research purposes, or statistical purposes; and, (e) the establishment, exercise or defense of legal claims.

Regardless of the technology or online services under scrutiny, e.g. search engines' liability, the critical parts of this new set of provisions can be summed up in accordance with three issues. First, it is up to the data controller to strike a balance between the request of the data subject—pursuant to Art. 17(1)—and the restrictions set up by Art. 17(3). This means the data controller has to evaluate whether the protection of further rights and interests, such as freedom of speech, should prevail in a certain case. The mechanism, at work in the field of data protection in Europe since the 2014 ruling of the EU Court of Justice ('EUCoJ') on the right to be forgotten, is not new. The EU version of the 'notice-and-takedown' procedure is well known for example in the field of intellectual property and liability of online service providers [14]. In both cases, it is still an open issue whether this mechanism of 'notice-and-takedown' is consistent with due process provisions in Europe, e.g. ECHR's Article 6 on the 'equality of arms,' which is mentioned by art. 6(2) of the EU Treaty as a source of fundamental rights in the European Union [15]. Although data subjects can always challenge the decisions of data controllers before their national authorities and courts, it remains controversial whether this two step-procedure can guarantee both rights of data subjects and the public interest. For example, after the ruling of the EUCoJ, Google has rejected so far about 56% of delisting requests. Out of this considerable amount of rejected requests, only a negligible percentage appealed to a National Authority, with very few requests to a national Court. In the case of the right to be forgotten, it may be argued, "the European Union has not only ordered Google to comply with European law; it has essentially handed off enforcement of the right in the first instance to Google" [16]. Could such scenario ever apply to the data controllers of some blockchain networks?

The second problem concerns the duty to inform pursuant to Art. 17(2). The latter can be understood either as a toothless mechanism or as a powerful means to safeguard the data subjects. The wording of the regulation on "reasonable steps," "available technology," or "cost of implementation," allows either to be true. At their best possible light, these rather vague provisions may represent a wise mechanism of legal flexibility with which to address the challenges posed by the astonishing advancements in technology and prevent over-frequent revisions to tackle such progress [17]. Yet, it is hard to envisage how the entire chain of controllers of Art. 17(2)—triggered by the request of the data subject pursuant to Art. 17(1)—will end up erasing "any links to, or copy and replication of, those personal data." By taking into account the extra-territorial effects of permissionless blockchains operating in multiple jurisdictions, how shall the EU legislators attain their purpose?

The third set of problems regards the restrictions provided by Art. 17(3)(d) on archiving purposes in the public interest, scientific or historical research purposes, or statistical purposes. These provisions make of course sense. Still, pursuant to Art. 89(2) and (3), safeguards and conditions for the processing of personal data are delegated in such cases back to national legal systems, e.g. the divergent provisions of Art. 35 of the German Law and Art. 16 of the Luxembourg Law implementing the norms of the GDPR on "erasure." This delegation of power may either entail a European form of experimental federalism, or trigger threats and risks of fragmentation [17]. For instance, according to certain scholars, this mechanism "will certainly reduce the impact of the regulation as a harmonizing force of data protection regulation in Europe in the context of Big Data, and it will make life harder for companies and organizations operating not just in one but multiple Member States in Europe" [18]. Articles 60, 61, 75(4) and 97(2)(b) of the GDPR provide a number of ways to cope with the centrifugal forces of the system, i.e. methods of coordination between multiple jurisdictions of national supervisory authorities. Still, it seems fair to concede that the first data controller triggered by the data subject's request pursuant to Art. 17(1), may have a hard time in striking a fair balance between this request and the restrictions that shall be often evaluated on a national basis, in accordance with Art. 17(3)(d).

Scholars have extensively discussed the impact of the GDPR on the functioning and design of blockchains [7, 10, etc.]. The specificity of the technology suggests three possible scenarios. First, we can imagine a bunch of GDPR abiding blockchain systems that lawfully process and store personal data: think again about the validators of a permissioned blockchain network, or the gatekeepers of some open blockchains. There is no reason why such networks cannot legally process personal data in accordance with obligations of data processors and data controllers established by the GDPR. Moreover, this compliance represents the bread and butter of several EU projects, e.g. DECODE, in which the aim of lawyers, developers, and other experts, is to design win-win solutions for blockchain-driven projects that process personal data [19].

Yet, something can go wrong. Consider a Court's order to delete personal data on a blockchain pursuant to either Art. 17(1)(d) of the GDPR on unlawful processing, or Art. 17(1)(b) on withdrawal of consent, in accordance with Art. 6(1)(a), or Art. 9(2)(a). Once a Court identifies the data processors and controllers of the blockchain with, say, the gatekeepers of the network, what should the next legal step be? Since irreversibly encrypting personal data on blockchains seems unfeasible [20], should the law force the entire blockchain network to re-compute all the blocks that follow the block in which the network stored the personal data to be erased? But, how about the transactions that occurred after the target block? Are they not at risk? Shouldn't the network halt, until all the subsequent blocks are mined? In more general terms, how should we address the retroactive effect of such measures?

The third scenario is even more worrying. Here, personal data have to be removed from a blockchain pursuant to Art. 17 of the GDPR, and still it is hard to find out who should remove such data; who is, in other words, responsible. After all, one of the mantras in today's debate on blockchains conceives them as a sort of "distributed autonomous organizations" [4]. The intricacy of the interaction between humans and computers can make it extremely difficult to ascertain what is, or should be, the information content of the natural or artificial entity, as foundational to determining the responsibility of individuals. Such cases of distributed responsibility that hinge on multiple accumulated actions of humans and computers may lead to cases of impunity that already have recommended some legal systems to adopt new forms of criminal accountability [21]. In the case of highly distributed networks, such as DAOs, what should the response of the law thus be? After the legal misadventures of P2P systems in the 2000s, will history repeat itself with the ban of some kinds of blockchain networks, e.g. illegal non-modifiable blocks of data and information?

Next section further explores these scenarios with some technical solutions.

4. The Clash

The troubles of blockchains with the GDPR can be examined with some solutions that experts and scholars have proposed over the past years. Most efforts comprehensibly revolve around how to make personal data anonymous on the blockchain. Methods for generating a new public key pair for each transaction have been developed for Bitcoin and Ethereum, whilst cryptographic techniques, such as "ring confidential transactions" [22], and zero-knowledge proofs [23], have been implemented into the Monero and the Z-Cash cryptocurrencies, respectively.

In the case of the right to erasure of personal data, three main approaches have emerged. The first is "hashing-out," that is, storing personal data off-chain in a database under the control of an identifiable data controller. The blockchain maintains a hash that can be used as a link to the database in which personal data are stored. Once the right to erasure of Art. 17 is triggered, that which should be erased is only the off-chain data that are employed to identify any subject linked to the in-chain hash. The price to pay for this solution, however, seems too high to many: "this [solution] may be considered a betrayal to the decentralization principle of blockchains, as a certain degree of control of data remains in the hands of a single centralized party" [10]. Here, the clash appears as a matter of principle that involves both the design of technology, its architecture, and the distribution of power among the nodes of the network.

The second approach concerns "key destruction." The use of sufficiently strong encryption for data stored on blockchains suggests that data can be "erased" by destroying the encryption key. The information is no longer accessible because it is impossible to decrypt the data. Yet, it is an open question whether the provisions of Art. 17 can be interpreted in such a way, that data should not be really erased from the chain but only accessible by the data subject, or not accessible at all. According to the 2014 Opinion of the Art. 29 Working Party, "state-of-the-art encryption... does not necessarily result in anonymisation" [20]. Correspondingly, the destruction of data or keys does not eliminate the possibility to re-identify individuals. Furthermore, in the Opinion of the EU authorities, hypotheses of brute force attacks and the evolution of technology have also to be taken into account. The clash appears here more as a matter of security standards, than a matter of principle.

The third approach has to do with chameleon hashes. Rather than hashing-out data, or making such data inaccessible with the destruction of the keys, the aim is to devise "redactable blockchains" through the use of hash functions that involve a trapdoor [24]. The knowledge of such trapdoor allows re-writing the blocks under specific constraints, e.g. transparency and accountability. The redaction occurs either through a trusted third party that knows the trapdoor to open the block, or by adding the hash function as a primitive of the blockchain's protocol. In the first case, some of the critiques to the hashing-out approach, i.e. "betrayal to the political decentralization principle," reappear. In the second case, we may wonder how to properly address the data protection issues of most blockchains, since blockchains need to include chameleon hash functions from their own inception in order to be redactable [10]. Moreover, according to some others, "chameleon hashes can't eliminate old copes of the blockchain that will still contain the redacted information and miners also have the discretion as to whether to accept the changes or not" [7, 24]. Here, the reasons for a legal clash appear as a mix of principles on political decentralization and rules for data protection.

Further approaches do exist, e.g. µchains [25]. However, it seems fair to admit that all the solutions illustrated in this section have some problems of their own and, all in all, they appear unfit to deal with the erasure of personal data *already present* in some blockchains today. By further considering multiple types of blockchain, according to their architecture, uses, services, or functions, it is then hard to say what solution to the right to erasure-problem could prevail in the next future. In any event, we should not overlook another reason of clash between today's legal frameworks and most current blockchains. Going back to the final scenario of the previous section on highly distributed networks, such as DAOs—in which it can be tricky to determine who is the data processor, and who is the data controller—the clash concerns regulatory systems that compete and even render the claim of the other regulatory system superfluous.

From the viewpoint of technology as a regulatory system [26], several examples illustrate how the legal intent to regulate the process of technological innovation may fail. The EU e-money directive 46 from 2000 is instructive: soon after its implementation, further forms of online payments, such as PayPal, forced the Bruxelles legislators to intervene, finally amending themselves with the new directive 110 from 2009. From the viewpoint of the law, however, several counter-examples stress the multiple ways in which legal systems may affect—and even hinder—technological innovation. As shown by the aforementioned 2005 ruling of the US Supreme Court in the Grokster case, the legal parable of P2P systems is instructive. It draws our attention to (i) whether or not a technology should be banned; (ii) whether designers and producers of such technology are accountable; and, (iii) whether users of P2P, and now blockchains, can be held liable as well. In the case of users of highly distributed blockchain networks with no obvious data controller, some argue, "a large amount of nodes would need to be contacted and compelled to comply... this may lead to forcing all nodes to stop running the blockchain software where GDPR rights cannot be achieved through alternative means" [7].

Others claim that even in the case of individuals directly interacting with a permissionless blockchain, a solution can be found through voluntary clauses embedded into the system via smart rules [19], or conditions and terms of use [10]. These latter approaches could indeed strengthen GDPR requirements for lawful data processing, by either prohibiting the processing of certain types of personal data, or requiring users to have consent or another legal basis for processing. According to Art. 25 of the GDPR on the so-called principle of privacy by default and by design, designers and developers of such blockchains could also be forced to embed some of the techniques mentioned in this section, e.g. zero-knowledge proofs, into the design of the blockchain. The problem with this line of argument, however, is that issues are not always 'technical,' but 'social.' Social issues may concern matters of principle on the design of the network and its degrees of 'political decentralization' [9]; namely, the distribution of power among the nodes of a blockchain. In addition, social clash may regard the extra-territorial effects of such distributed blockchain networks [26], their terms of use and enforcement [6], or the consensus algorithm on whose proof-of-work the process of block computing relies in most blockchains [4]. In light of P2P legal misadventures, we may thus expect in the short term either courts potentially targeting the whole bunch of users interacting with a permissionless blockchain, or blocking some blockchains with court orders. This is what already occurred with the Peppermint case on P2P systems in Italy, back to 2008 [27]. Will history repeat itself?

5. Conclusions

One of our main contentions in the paper has been that GDPR abiding blockchain systems are feasible. Jurists, programmers, and other experts are increasingly working on this nowadays. Some argue that blockchain solutions could even strengthen rights and obligations enshrined in the GDPR [7, 10, 19]. Still, manifold blockchain networks functioning out there suggest a new generation of data protection issues brought about by this technology. Some of these issues will likely concern the right to erasure set up by Art. 17. These cases will soon be discussed before national authorities and courts, and will likely test all the solutions that have been illustrated in the previous sections. In the case of permissioned, or closed blockchains, it is arguable that "hashing-out" strategies can properly address Art. 17 related issues. The "betrayal" of the decentralization principle seems however the price to be paid for such solution [9, 10].

In the case of permissionless blockchains, the legal test under the GDPR will often be about "key destruction" techniques, chameleon hashes, or µchains. Yet, we may suspect that some highly distributed blockchain networks will find a harder time with the implementation of Art. 17. Contrary to the legal misadventures of P2P systems, e.g. the 2008 Peppermint case in Italy [27], we think it is unlikely that national authorities and courts will target the bunch of individuals directly interacting with a permissionless blockchain, so as to enforce a data subject's right to erasure. Rather, it is probable that the target will be the blockchain network as such. This was the legal output of another famous P2P case, i.e. the 2005 Grokster case of the US Supreme Court, and this has also been the recent EU policy on other crucial issues of internet governance, such as liability of search engine services as controllers of personal data processing. It is thus likely that this trend will go on with some blockchain cases, "forcing all nodes to stop running the blockchain software" [7].

The parallel with the legal issues of P2P systems sheds light on two further aspects of today's debate on legal blockchains. First, in the phrasing of the US Supreme Court, there is no doubt that blockchain networks are "capable of non-infringing uses." Contrary to many 2000s, early 2010s opponents of P2P systems, there is no Western advocate of the ban of blockchain technologies up today. This is not to say that current blockchains are business as usual in the legal domain, and should not be rather re-conceptualized, and accordingly designed. The paper illustrated some solutions, e.g. chameleon hash functions and zero-knowledge proofs, which should be embedded into the design of the blockchain since its inception, in order to prevent data protection-related issues. The reference point of the GDPR was here Article 25 on privacy by design, and by default. This brings us to the second facet of our parallel. In the case of P2P networks, the 2005 decision of the US Supreme Court represented a threshold for the design of these systems: after the ruling of Justices in Washington, the design solution for the most relevant legal issue of many P2P systems, e.g. copyright infringement, was a more decentralized and distributed architecture for such file-sharing networks [1]. In the case of blockchains, it is likely that we will similarly refer soon to their architecture, by distinguishing between blockchains designed before or after the GDPR. The threshold indicates the set of blockchains that have been thought about to expressly meet the requirements of the EU regulation through e.g. privacy by design techniques, and blockchains that, for one reason or another, e.g. ante GDPR designed blockchains, trigger some sort of clash with the legal order. The paper has sorted out four different kinds of clash, i.e. on principles, security standards, rules for data protection, and 'social clashes.' It is unclear how the interplay between legal regulation, technological constraints, social norms, and market interests, will end up in this context; still, we should be ready to address, or even prevent, such different kinds of clashes. Rulings and court orders will be instructive over the next years.

References

[1] A. Glorioso, G. Ruffo, and U. Pagallo, The Social Impact of P2P Systems, in X. Shen, H. Yu, J. Buford and M. Akon, *Handbook of Peer-to-Peer Networking*, pp. 47-70, Springer Heidelberg, 2010.

[2] R. Wattenhofer, *The science of the blockchain*, CreateSpace Independent Publishing Platform, 2016.

[3] M. Bauwens, *P2P and Human Evolution: Placing Peer to Peer Theory in an Integral Framework*, at http://integralvisioning.org/article.php?story=p2ptheory1, 2005 (last accessed April 20th, 2018).

[4] M. Crepaldi, The path toward an ethics of Distributed Autonomous Organizations (DAOs), *Ethicomp* Proceedings, September 2018.

[5] M. Campbell-Verduyn (ed.), *Bitcoin and Beyond Cryptocurrencies, Blockchains, and Global Governance*, Routledge New York, 2018.

[6] U. Pagallo and M. Durante, Three roads to P2P systems and their impact on business ethics, *Journal of Business Ethics* **90** (2009), 551-564.

[7] M. Finck, Blockchain and Data Protection in the European Union. *Max Planck Institute for Innovation & Competition, Research Paper No. 18-01*, 2017 (Available at SSRN https://ssrn.com/abstract=3080322 or http://dx.doi.org/10.2139/ssrn.3080322).

[8] F. Glaser, Pervasive decentralization of digital infrastructures: a framework for blockchain enabled system and use case analysis, 2017.

[9] V. Buterin, The Meaning of Decentralization, available at https://medium.com/@VitalikButerin/the-meaning-of-decentralization-a0c92b76a274, February 2017.

[10] L-D. Ibáñez, K. O'Hara and E. Simperi, On Blockchains and the General Data Protection Regulation, University of Southampton, June 2018.

[11] A. Walch, A. The path of the blockchain lexicon (and the law), 2017.

[12] R. Matzutt, J. Hiller, M. Henze, J.H. Ziegeldorf, D. Müllmann, O. Hohlfeld, and K. Wehrle, *A Quantitative Analysis of the Impact of Arbitrary Blockchain Content on Bitcoin.* Paper presented at the Proceedings of the 22nd International Conference on Financial Cryptography and Data Security (FC). Springer, 2018.

[13] A. Sward, I. Vecna, and F. Stonedahl, F., Data Insertion in Bitcoin's Blockchain. *2018, 3*. doi:10.5195/ledger.2018.10,1

[14] P. Van Eecke, Online service providers and liability: A plea for a balanced approach, *Common Market Law Review* **48**(5) (2011), 1455–1502.

[15] U. Pagallo and M. Durante, Legal Memories and the Right to Be Forgotten, in L. Floridi (ed.), *Protection of Information and the Right to Privacy. A New Equilibrium?,* Law, Governance and Technology Series 17, Springer, Dordrecht, 2014, 17-30.

[16] J.M. Balkin, Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation (September 9, 2017). *Yale Law School, Public Law Research Paper No. 615*. Available at SSRN: https://ssrn.com/abstract=3038939.

[17] U. Pagallo, The Legal Challenges of Big Data: Putting Secondary Rules First in the Field of EU Data Protection **3**, 1 (2017), 34-46.

[18] V. M. Schönberger and Y. Padova, Regime change? Enabling big data through Europe's new data protection regulation, *Columbia Science and Technology Law Review* **17** (2016), 315-335.

[19] E. Bassi, M. Ciurcina, J.C. De Martin, S. Fenoglietto, G. Rocchi, O. Sagarra Pascua, and F. Bria, D1.8 Legal Framework for Digital commons DECODE OS Legal Guidelines, Decode Project, 2017., available at https://www.decodeproject.eu/publications/legal-frameworks-digital-commons-decode-os-and-legal-guidelines.

[20] Art. 29 Working Party, Opinion on Anonymization Techniques, n. 05/2014.

[21] U. Pagallo, AI and Bad Robots: The Criminology of Automation, in M.R. McGuire and Th. J. Holt (eds.), *The Routledge Handbook of Technology, Crime and Justice*, 643-653. London & New York, Routledge, 2017.

[22] S. Noether, A. Mackenzie, and the Monero Research Lab, Ring Confidential Transactions, *Ledger* **1**(1), December 2016.

[23] E. Ben-Sasson, A. Chiesa, E. Tromer and M. Virza, Succinct Non-Interactive Zero Knowledge for a von Neumann Architecture. Technical Report 879, 2013.

[24] G. Ateniese, B. Magri, D. Venturi, E. Andrade, Redactable Blockchain–or–rewriting History in Bitcoin and Friends. *Security and Privacy (EuroS&P)*, IEEE European Symposium, 2017.

[25] I. Puddu, A. Dmitrienko, S. Capkun,. μchain: How to Forget without Hard Forks. *IACR Cryptology ePrint Archive*, December 2017.

[26] U. Pagallo, The Realignment of the Sources of the Law and their Meaning in an Information Society, Philosophy & Technology **28**, 1 (2015), 57-73.

[27] U. Pagallo, Let Them Be Peers: The Future of P2P Systems and Their Impact on Contemporary Legal Networks, in M. Fernandez-Barrera, N. Nuno Gomes de Andrade, P. de Filippi, M. Viola de Azevedo Cunha, G. Sartor e P. Casanovas (eds.), *Law and Technology: Looking into the Future*, pp. 323-338. European Press Academic Publishing, Florence, 2009.

Legal Knowledge and Information Systems
M. Palmirani (Ed.)

doi:10.3233/978-1-61499-935-5-91

Legal Ontology for Modelling GDPR Concepts and Norms

Monica PALMIRANI*, Michele MARTONI*, Arianna ROSSI*, Cesare BARTOLINI**, Livio ROBALDO**

**CIRSFID, University of Bologna.*
{monica.palmirani, michele.martoni, arianna.rossi}@unibo.it

*** SnT - Interdisciplinary Centre for Security, Reliability and Trust,*
Université du Luxembourg
JFK Building, 29, Avenue J.F. Kennedy, L-1855 Luxembourg
{cesare.bartolini, livio.robaldo}@uni.lu

Abstract. This paper introduces PrOnto, the privacy ontology that models the GDPR main conceptual cores: data types and documents, agents and roles, processing purposes, legal bases, processing operations, and deontic operations for modelling rights and duties. The explicit goal of PrOnto is to support legal reasoning and compliance checking by employing defeasible logic theory (i.e., the LegalRuleML standard and the SPINDle engine).

Keywords. Semantic web, legal reasoning, legal ontology, GDPR.

1. Introduction

The GDPR (General Data Protection Regulation) is the new common framework for data protection that applies to the whole European Union and harmonizes the legal principles of its Member States that can thus be more effectively applied in the Digital Single Market. The Regulation places upon entities involved in the processing of personal data a number of obligations, among which is the obligation to assess the risks they could encounter and adapt their duties on the basis of the impact assessment (Article 35, GDPR), whereas specific measures for the safeguard of data subjects' human dignity and fundamental rights are introduced. Instruments such as audits and compliance checking are intended to ensure the application of the principles of *data protection by design* (Article 25, GDPR) during software development (ex-ante phase), but also a punctual detection of violations (ex-post phase) when they occur. Since public administrations, enterprises and non-profit organizations alike will need to observe these newly-introduced, demanding duties, semantic web and legal reasoning techniques can offer a valuable support and ease compliance.

A legal ontology that formalizes data protection norms is therefore needed and timely. This paper introduces PrOnto [21], the privacy ontology that models the GDPR main conceptual cores: data types and documents, agents and roles, processing purposes, legal bases ex Article 6 GDPR, processing operations, and deontic operations for modelling rights (Chapter 3- articles 12-23) and duties (Chapter 4- articles 24-43). This ontology considers the GDPR as a starting point, however it is meant to be extended to the concepts and relative relations of other legal frameworks (such as

Member State laws). The explicit goal of PrOnto is to support legal reasoning and compliance checking by employing defeasible logic theory (i.e., the LegalRuleML standard [5] and the SPINDle engine [12]), as opposed to exclusively execute information retrieval. This article focuses on the analysis of deontic operators in order to manage the checking of compliance with the GDPR obligations. We use the Right to Data Portability (Art. 20) for illustrating the PrOnto benefits.

2. MeLOn Methodology

PrOnto was developed through an interdisciplinary approach called MeLOn (Methodology for building Legal Ontology), which has been successfully used to develop several legal ontologies by legal experts[1]. MeLOn is explicitly designed for legal ontologies and the related difficulties encountered by the legal operators during the definition of a model of reality through ontological techniques, such as Protégé, or patterns design method or the foundational approach.

The MeLOn methodology iterates over ten steps: 1) Describe the goal of the ontology; 2) Evaluation indicators. PrOnto's criteria, based on the existing state of the art, are [6]: (i) coherence, (ii) completeness, (iii) efficiency, (iv) effectiveness, (v) usability, (vi) agreement; 3) State of the art survey: PrOnto reuses existing ontologies, ontology patterns [13][14], and other existing domain vocabularies; 4) List the whole relevant terminology, extracted from legal sources, in particular legal definitions; 5) Use usable tools (such as tables, UML diagrams and the Graffoo tool); 6) Refine and optimize: an ontology expert manually adds the axioms; 7) Test the output in terms of completeness, effectiveness and usability; 8) Evaluate the ontology: OntoClean method and SPARQL queries; 9) Publish the document with the LODE tool [20]; 10) Collect feedbacks from the community in order to reach the agreement criteria. The MeLOn methodology allows to successfully work within interdisciplinary group that include engineers, lawyers, linguists, logicians and ontologists, and to model the legal knowledge rapidly and accurately while integrating the contributions of different disciplines.

3. The Right to Data Portability

Chapter 3 of GDPR lists all the rights of the data subject (right to access, right to be forgotten, right to portability, right to erasure, etc.) and Chapter 4 the duties for the controller and processor (such as the obligation to notify the data breach to the data subject). PrOnto aims at modelling the deontic operators (right, obligation, prohibition, permission) but also customize the obligations and rights for the GDPR. In particular we consider the Right to Data Portability (Art. 20). The Right to Data Portability is a complex right composed basically of two obligations:

"*1. The data subject shall have the right to receive the personal data concerning him or her, which he or she has provided to a controller, in a structured, commonly used and machine-readable format and have the right to transmit those data to another controller without hindrance from the controller to which the personal data have been provided, where:*

(a) the processing is based on consent pursuant to point (a) of Article 6(1) or point (a) of Article 9(2) or on a contract pursuant to point (b) of Article 6(1); and

(b) the processing is carried out by automated means."

[1] http://amsdottorato.unibo.it/8215/; http://amsdottorato.unibo.it/7804/; http://amsdottorato.unibo.it/7261/.

The Right to Data Portability involves the Controller that has the obligation to take some Steps (see Fig.1). A Step is executed by Actions and each Step commits LegalRules, in our case *ObligationOfPortability* The actions involved are "Provide" personal data to the data subject, and "Transmit" the same data also to other controllers. The data type is determined by the action performed by the Controller. This means that all the personal data provided by the data subject or observed are involved in this obligation. The personal data inferred, derived or stored by the Controller are not a matter of this Right[2].

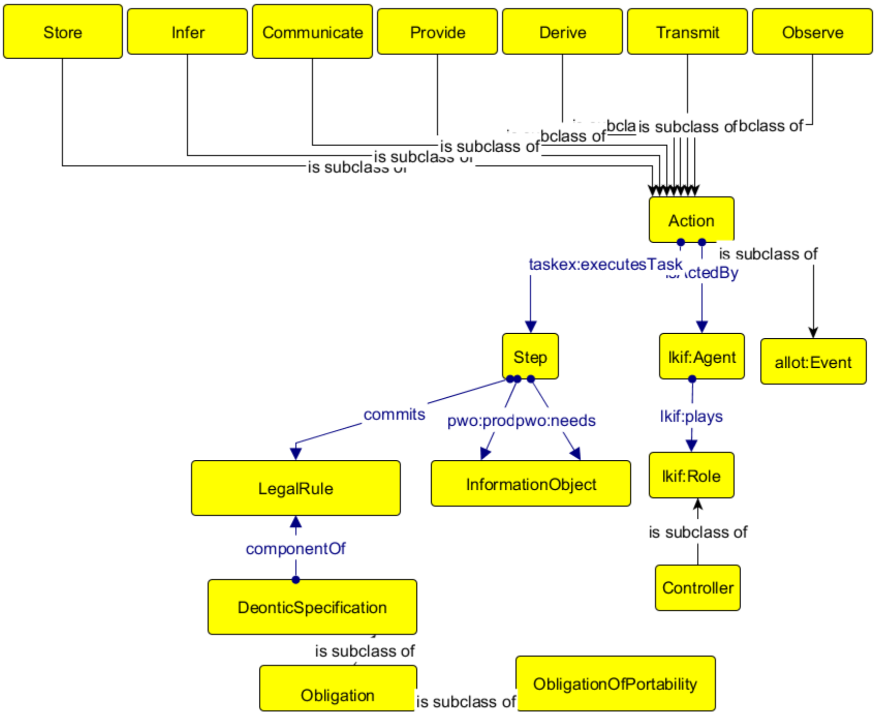


Figure 1 - Fragment of PrOnto concerning the ObligationOfPortability

4. PrOnto Modules

PrOnto is composed by modules, following the main structure of the GDPR legal principles: i) data and documents, ii) agents and roles, iii) processing purposes and legal bases; iv) data processing and workflow, risk management, and v) legal rules and deontic operators. Some documents and data refer to the data subject, which is a role of an agent (natural person). Data is processed following a given workflow, i.e., a plan of actions. When it is executed, each action assumes specific temporal parameters (e.g., interval of time of the processing), context (e.g., jurisdiction where the data processing

[2] Article 29 WP 242 rev.01 "In contrast, inferred data and derived data are created by the data controller on the basis of the data "provided by the data subject". For example, the outcome of an assessment regarding the health of a user or the profile created in the context of risk management and financial regulations (e.g. to assign a credit score or comply with anti-money laundering rules) cannot in themselves be considered as "provided by" the data subject."

is carried out), and value (e.g., place where the data processing is performed). The processing is lawful only if a legal basis is provided. Each processing activity involves a number of stakeholders: controller, processor, and other actors, and each has obligations or rights (for instance, data subjects have data protection rights). Such rights and obligations are linked to documents where the provisions appear, such as terms of use, information, privacy policies, consent forms.

4.1. Data and Document

In the context of data protection, personal data (Article 4.1(1)) is the object of the Regulation and the target of its protection, but data are also the information source that regulates the relations among different agents (e.g., controller, processor, etc.) using privacy policies, informed consent, contracts, codes of conduct, law, case-law, and any other legal document. Since data and documents are documental sources, the FRBR[3] ontology is employed: their representation over time can thus be modelled by following a robust design pattern that has been adopted for the publication process. Data is organized in the categories defined in the GDPR: personal data (Article 4.1(1)), non-personal data, anonymized data, pseudonymised data (Article 4.1(5)). The duties and rights depend on the type of data. For instance the DPO (Data Protection Officer) is mandatory for processing "a large scale of special categories of data" (Art. 9). This is why a specific version of data can be detected by using the time when the event occurred (e.g., a data breach event) and the dynamic versioning of the FRBR model is applied also to the class data.

4.2. Agent and Role

Agents and roles are frequently mistaken in legal ontologies. PrOnto, on the contrary, distinguishes the two classes. An agent might play multiple roles in different processing operations or contexts (e.g., a controller could act as processor or third party in relation to different data processing activities). Not only physical persons and organizations are included in the agents' class, but also IT organizations, artificial intelligence and software, or robots. Each role is fixed in a given time period, which is linked to the time version of the dataset and the duration of the data processing. This implies that there is an event that assigns the role to an agent (e.g., designation of the processor by the controller ex Article 28, GDPR). Concerning the different roles, we have the Controller that *isRepresentedBy* a Representative in the European Union (Art. 27), *designates* a DPO and *nominates* a Processor.

4.3. Purposes and Legal Basis

Under the GDPR, personal data processing (Article 4.1(2)) is lawful only if motivated by a purpose that must be supported by a legal basis (see Article 6, GDPR, on the lawfulness of processing). This is why a lawfulness status was needed and was thus added as a Boolean data property of the PersonalDataProcessing class, whilst each personal data processing is based on a Purpose.

[3] FRBR— Functional Requirements for Bibliographic Records https://www.ifla.org/publications/functional-requirements-for-bibliographic-records

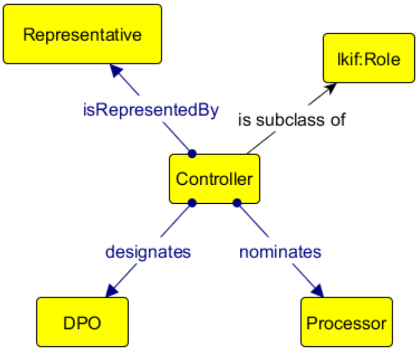


Figure 2 - Controller class and its proprieties

By modelling the knowledge in this manner, a rule engine that, for instance, is based on a rule-based language such as LegalRuleML is able to return this value after the rule reasoning process. The legal basis is also involved in the Right to Data Portability, considering that the Right is applicable only in the presence of a consent or a contract.

4.4. Data Processing

Human activities can be modelled through a workflow, i.e., a sequence of steps that takes some resources in input and produces certain outcomes. However, a workflow is composed of two parts: first a plan to do something is laid out (e.g., workflow), then the concrete sequence of actions is actually performed (e.g., execution of the workflow). This distinction is of utmost relevance in the GDPR framework: the plan (e.g., Impact Assessment Plan made by steps) is different from the real execution (e.g., the countermeasures acted in the event of a data breach), which is made up of a set of actions. Compliance checking presumes both a plan in line with the law, and countermeasures in the event of violations during the actual execution (e.g., remedies). For this goal, the Publishing Workflow Ontology (PWO) proved perfectly suitable as a basis to model the data processing ontology module because it includes both a workflow and an executed workflow. The workflow execution is composed by actions. An action [1] is a kind of event that is described by temporal parameters (e.g., interval) and contextual values (Time-indexed Value in Context - TVC). One of the values it can take is the place where the event occurs (e.g., within the EU borders) and the relevant jurisdiction (e.g., Regional competence). Other values and statuses can also be included to enrich the context description.

4.5. Deontic Operators

In order to model legal norms, deontic operators such as right, obligation, permission and prohibition are fundamental. Under the GDPR, it is also important to include violation/compliance as the status in which an obligation or a prohibition is violated or maintained. The deontic operators have temporal parameters and refer to a jurisdiction to consider those rights that are only effective in a specific domestic regulation. For all these reasons, this section of PrOnto allows to model those predicates that are necessary to implement legal rules and is an extension of the LegalRuleML metamodel, which allows the synchronization of the legal rule language modelling with the ontology.

This module also defines the relationships among deontic rules, actors' rights and obligations, obligations and permissions, and violation/compliance. This modelling allows the population of the ontology, or the creation of RDF triples, in order to perform queries such as "give me all the processing activities that have been violated by some actors in a given time". This knowledge is processed by the rule engine, but it is also transformed into individuals (e.g., materialization) of the ontology (or RDF triples) without the need to perform a query on the rule engine each time.

It is also worth noticing that, within the project DAPRECO [8], PrOnto is used to formalize GDPR in reified I/O logic [22], and in the Cloud4Eu is used with defeasible logic. In both cases formulae are connected to the concepts in PrOnto via the LegalRuleML constructs. This means that PrOnto is neutral respect the type of logic adopted.

4.6. Rights and Obligations relationships

For each obligation there is also a right connected with the data subject (Bearer). Fig. 3 shows how the ObligationOfPortability is connected to the RightToPortability using DeonticSpecification super class.

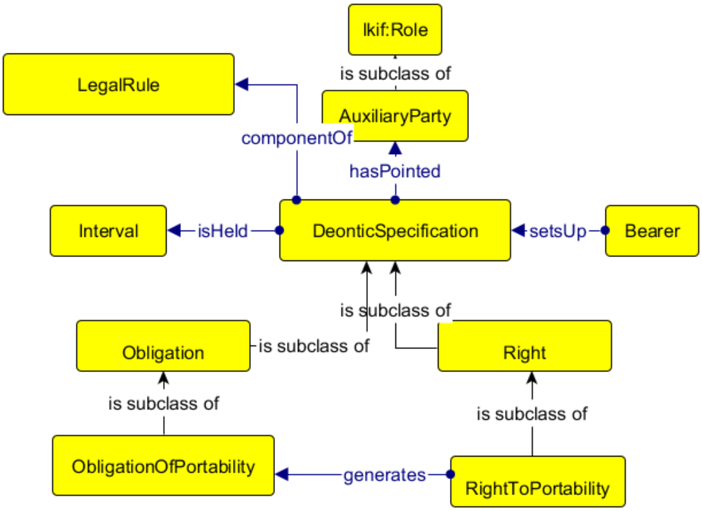


Figure 3: Right to Portability and Obligation of Portability

For implementing this fragment of ontology we have extended the LegalRuleML meta-model with several axioms. We report here some of axioms that intend to connect the right with the obligation and prohibition, and the right with the permission.

LegalRule	<owl:ObjectProperty rdf:about="https://w3id.org/ontology/pronto#componentOf"> <rdfs:domain rdf:resource="http://docs.oasis-open.org/legalruleml/ns/v1.0/metamodel#DeonticSpecification"/> <rdfs:range rdf:resource="https://w3id.org/ontology/pronto#LegalRule"/>

	</owl:ObjectProperty>
isViolatedBy it defines the relationship between obligation/prohibition and violation	<owl:ObjectProperty rdf:about="https://w3id.org/ontology/pronto#isViolatedBy"> <rdfs:domain rdf:resource="http://docs.oasis-open.org/legalruleml/ns/v1.0/metamodel#Obligation"/> <rdfs:domain rdf:resource="http://docs.oasis-open.org/legalruleml/ns/v1.0/metamodel#Prohibition"/> <rdfs:range rdf:resource="http://docs.oasis-open.org/legalruleml/ns/v1.0/metamodel#Violation"/> </owl:ObjectProperty>
isFulfilledBy	<owl:ObjectProperty rdf:about="https://w3id.org/ontology/pronto#isFulfilledBy"> <rdfs:domain rdf:resource="http://docs.oasis-open.org/legalruleml/ns/v1.0/metamodel#Obligation"/> <rdfs:domain rdf:resource="http://docs.oasis-open.org/legalruleml/ns/v1.0/metamodel#Prohibition"/> <rdfs:range rdf:resource="http://docs.oasis-open.org/legalruleml/ns/v1.0/metamodel#Compliance"/> </owl:ObjectProperty>
Relationship between obligation and permission	<owl:ObjectProperty rdf:about="https://w3id.org/ontology/pronto#implies"> <rdfs:domain rdf:resource="http://docs.oasis-open.org/legalruleml/ns/v1.0/metamodel#Obligation"/> <rdfs:range rdf:resource="http://docs.oasis-open.org/legalruleml/ns/v1.0/metamodel#Permission"/> </owl:ObjectProperty>
Relationship between prohibition and obligation	<owl:ObjectProperty rdf:about="https://w3id.org/ontology/pronto#isAKindOf"> <rdfs:domain rdf:resource="http://docs.oasis-open.org/legalruleml/ns/v1.0/metamodel#Prohibition"/> <rdfs:range rdf:resource="http://docs.oasis-open.org/legalruleml/ns/v1.0/metamodel#Obligation"/> </owl:ObjectProperty>
Right as a subclass of Permission	<owl:Class rdf:about="http://docs.oasis-open.org/legalruleml/ns/v1.0/metamodel#Right"> <rdfs:subClassOf rdf:resource="http://docs.oasis-open.org/legalruleml/ns/v1.0/metamodel#Permission"/> </owl:Class>

Table 1 – Some axioms of the extension of the LegalRuleML ontology

4.7. Duties and Violation

Finally, the Steps are connected with LegalRuleML that is the deontic part of the ontology, capable to model and perform reasoning with right, obligation, permission, prohibition. The violation is connected with the obligation/prohibition that is violated and the compliance states when the obligation is complied with. In this way, we are able to detect the steps that create violations of some obligations and the connected risks, along with the related measures (see Fig. 4).

5. Evaluation

The evaluation is carried out inside the Cloud4EU European project PCP, that intends to provide legal compliance checking systems for eGovernment services that are delivered across the cloud. We are currently in the phase of testing PrOnto on three different scenarios related to school services. PrOnto is also used inside the MIREL European project and the DAPRECO Luxembourgish project.

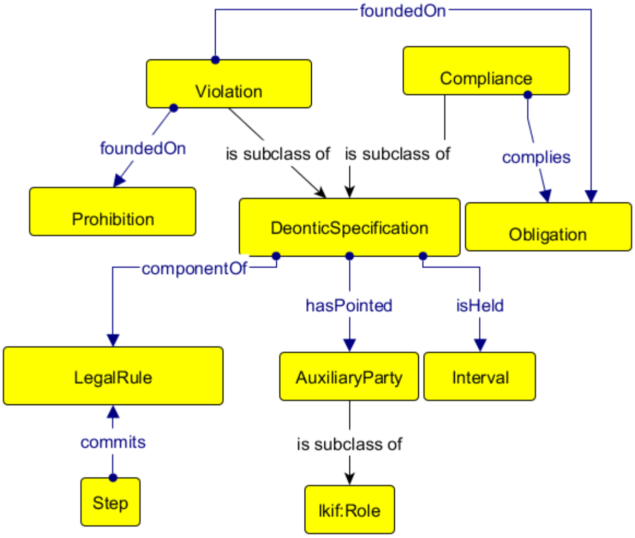


Figure 4: Violation, Compliance and Obligation, Prohibition

6. Related Work

A few privacy ontologies with specific goals [4]; [11]; have been designed, for instance the HL7 privacy ontology [13] for electronic health records. Other ontologies were created to ensure secure messaging among Internet of Things devices, whilst others are meant to manage data flows in the linked open data environment or on the blockchain. UsablePrivacy and PrivOnto [16] are more oriented to provide linguistic instruments in order to define glossary and taxonomy for the privacy domain, basically starting from the bottom-up annotation of the privacy policies (crowdsourcing annotation). GDPRtEXT [18] provides a list of concepts present in the GDPR text without really entering the modelling of the norms and the legal axioms (e.g., the actions performed by the processor, the obligations of the controller and the rights of the data subject). Morover GDPRtEXT does not foster FRBR information for managing versioning of the legal text over the time and consequently the changes of the legal concepts due to modifications in the legal system. GDPRov aims to describe the provenance of the consent and data lifecycle in the light of the Linked Open Data principles such as Fairness and Trust [17]. The SPECIAL Project[4] aims to provide tools for checking compliance in privacy domain. However, no ontology with foundational concepts, patterns, deontic operators and privacy principles has been designed to support legal

[4] https://www.specialprivacy.eu/

reasoning and check compliance yet. ODRL provides predicates and classes for managing obligations, permission, prohibitions, but several parts of the deontic logic are missing (e.g., right and penalty classes). ODRL is good for modelling simple policies capable to be searchable in SPARQL, but it is quite limited to manage the complex organization of the legal rules (e.g., exception in the constitutive rules or in the prescriptive rule). PrOnto is more exhaustive in this field. In order to do so, rights and obligations must be modelled through deontic operators. Moreover, actors and processing operations described in the normative prescriptions must also be included.
This is why PrOnto considers and reuses existing ontologies and follows ontology design patterns [21]: ALLOT [7], FRBR [15], LKIF we use in particular lkif:Agent to model lkif:Organization, lkif:Person and lkif:Role [9], the Publishing Workflow Ontology (PWO) [10]; Time-indexed Value in Context (TVC) and Time Interval [19].

7. Conclusions and Future Work

The existing privacy ontologies presented in the state of the art (e.g., HL7 for eHealth, PPO for Linked Open Data, ODRL for modelling rights, etc.) do not integrate deontic logic models that can be used for legal reasoning. PrOnto aims at the integration of different levels of semantic representation for multiple goals: 1) document and data modelling can support information retrieval in the Semantic Web, in particular with Linked Open Data (e.g., SPARQL queries); 2) workflow and processing modelling can represent helpful tools to plan privacy policies, but also BPMN modelling can be useful in system design (e.g., privacy-by-design); rights and obligations are necessary modules to enable automated legal reasoning that employ rule languages (e.g., LegalRuleML and compliance checking); 3) and finally, human-centered approaches can allow the visualization and the presentation of data protection principles and concepts in different contexts and directed to different audiences.
The research described in these pages has a long-term goal. Our intention is that of continuing the modelling and optimization of the formal model of the ontology, but also to evaluate it with a number of use-cases. In the meantime, we deem fundamental a discussion about the ontology within a large community, in order to establish consensus and to place such results in a standardization body for future governance (e.g., OASIS, W3C). In the future, it will also become necessary to develop specific profiles, one for each specific national law or even by thematic domain (e.g., Privacy in IoT, Privacy in AI, etc.).

Acknowledgements

This work was partially supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 690974 “MIREL: MIning and REasoning with Legal texts”, and by Luxembourg National Research Fund (FNR) CORE project C16/IS/11333956 “DAPRECO: DAta Protection REgulation COmpliance”.

References

[1] Abrams, M., 2014. The Origins of Personal Data and its Implications for Governance. SSRN Electron. J. https://doi.org/10.2139/ssrn.2510927

[2] Article 29 Working Party, 2018. Guidelines on Personal data breach notification under Regulation 2016/679 (No. wp250rev.01).

[3] Article 29 Working Party, 2017. Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679 (No. wp248rev.01).

[4] Ashley, K.D., 2017. Artificial intelligence and legal analytics: new tools for law practice in the digital age. Cambridge Univ Press, Cambridge New York Melbourne Delhi Singapore.

[5] Athan, T., Governatori, G., Palmirani, M., Paschke, A., Wyner, A., 2015. LegalRuleML: Design Principles and Foundations, in: Faber, W., Paschke, A. (Eds.), Reasoning Web. Web Logic Rules. Springer International Publishing, Cham, pp. 151–188. https://doi.org/10.1007/978-3-319-21768-0_6

[6] Bandeira, J., Bittencourt, I.I., Espinheira, P., Isotani, S., 2016. FOCA: A Methodology for Ontology Evaluation. Eprint ArXiv.

[7] Barabucci, G., Cervone, L., Di Iorio, A., Palmirani, M., Peroni, S., Vitali, F., 2010. Managing semantics in XML vocabularies: an experience in the legal and legislative domain. Balisage Ser. Markup Technol. 5. https://doi.org/10.4242/balisagevol5.barabucci01

[8] Bartolini, Cesare, Andra Giurgiu, Gabriele Lenzini, and Livio Robaldo. 2016. Towards legal compliance by correlating standards and laws with a semi-automated methodology. In BNCAI, volume 765 of Communications in Computer and Information Science, pages 47{62. Springer.

[9] Breuker, J., Hoekstra, R., Boer, A., van den Berg, K., Sartor, G., Rubino, R., Wyner, A., Bench-Capon, T., Palmirani, M., 2007. OWL Ontology of Basic Legal Concepts (LKIF-Core) (Deliverable No. 1.4). IST-2004-027655 ESTRELLA: European project for Standardised Transparent Representations in order to Extend Legal Accessibility.

[10] Gangemi, A., Peroni, S., Shotton, D., Vitali, F., 2017. The Publishing Workflow Ontology (PWO). Semantic Web 8, 703–718. https://doi.org/10.3233/SW-160230

[11] Gharib, M., Giorgini, P., Mylopoulos, J., 2017. Towards an Ontology for Privacy Requirements via a Systematic Literature Review, in: Mayr, H.C., Guizzardi, G., Ma, H., Pastor, O. (Eds.), Conceptual Modeling. Springer International Publishing, Cham, pp. 193–208. https://doi.org/10.1007/978-3-319-69904-2_16

[12] Governatori, G., Hashmi, M., Lam, H.-P., Villata, S., Palmirani, M., 2016. Semantic Business Process Regulatory Compliance Checking Using LegalRuleML, in: Blomqvist, E., Ciancarini, P., Poggi, F., Vitali, F. (Eds.), Knowledge Engineering and Knowledge Management. Springer International Publishing, Cham, pp. 746–761. https://doi.org/10.1007/978-3-319-49004-5_48

[13] Health Level Seven International, 2015. HL7 Specification: Clinical Quality Common Metadata Conceptual Model, Release 1 (HL7 Informative Document).

[14] Hitzler, P., Gangemi, A., Janowicz, K., Krisnadhi, A., Presutti, V. (Eds.), 2016. Ontology engineering with ontology design patterns: foundations and applications, Studies on the semantic web. IOS Press, Amsterdam Berlin.

[15] IFLA Study Group on the Functional Requirements for Bibliographic Records, 1996. Functional Requirements for Bibliographic Records, IFLA Series on Bibliographic Control. De Gruyter Saur.

[16] Oltramari, A., Piraviperumal, D., Schaub, F., Wilson, S., Cherivirala, S., Norton, T.B., Russell, N.C., Story, P., Reidenberg, J., Sadeh, N., 2016. Privonto: A semantic framework for the analysis of privacy policies. Semantic Web (1-19).

[17] Pandit H.J., Lewis D., 2017. Modelling Provenance for GDPR Compliance using Linked Open Data Vocabularies, Proceedings of the 5th Workshop on Society, Privacy and the Semantic Web - Policy and Technology (PrivOn2017) co-located with the 16th International Semantic Web Conference (ISWC 2017), http://ceur-ws.org/Vol-1951/PrivOn2017_paper_6.pdf.

[18] Pandit H.J., Fatema K., O'Sullivan D., Lewis D., 2018. GDPRtEXT - GDPR as a Linked Data Resource. In: Gangemi A. et al. (eds) The Semantic Web. ESWC 2018. Lecture Notes in Computer Science, vol 10843. Springer, Cham.

[19] Peroni, S., Palmirani, M., Vitali, F., 2017. UNDO: The United Nations System Document Ontology, in: d'Amato, C., Fernandez, M., Tamma, V., Lecue, F., Cudré-Mauroux, P., Sequeda, J., Lange, C., Heflin, J. (Eds.), The Semantic Web – ISWC 2017. Springer International Publishing, Cham, pp. 175–183. https://doi.org/10.1007/978-3-319-68204-4_18

[20] Peroni, S., Shotton, D., Vitali, F., 2012. The Live OWL Documentation Environment: A Tool for the Automatic Generation of Ontology Documentation, in: ten Teije, A., Völker, J., Handschuh, S., Stuckenschmidt, H., d'Acquin, M., Nikolov, A., Aussenac-Gilles, N., Hernandez, N. (Eds.), Knowledge Engineering and Knowledge Management.

[21] Monica Palmirani, Michele Martoni, Arianna Rossi, Cesare Bartolini, Livio Robaldo: PrOnto: Privacy Ontology for Legal Reasoning. EGOVIS2018, 7th International Conference, EGOVIS 2018, Regensburg, Germany, September 3-5, 2018, Proceedings. LNCS 11032, Springer, pp. 139-152 (2018)

[22] Robaldo, L. and X. Sun. 2017. Reifed input/output logic: Combining input/output logic and reification to represent norms coming from existing legislation. The Journal of Logic and Computation, Vol. 7

Legal Knowledge and Information Systems
M. Palmirani (Ed.)

doi:10.3233/978-1-61499-935-5-101

Modelling Legal Knowledge for GDPR Compliance Checking

Monica PALMIRANI, Guido GOVERNATORI

[1] *CIRSFID, University of Bologna.*
monica.palmirani@unibo.it
[2] *Data61, CSIRO, Australia.*
guido.governatori@data61.csiro.au

Abstract. In the last fifteen years, Semantic Web technologies have been successfully applied to the legal domain. By composing all those techniques and theoretical methods, we propose an integrated framework for modelling legal documents and legal knowledge to support legal reasoning, in particular checking compliance. This paper presents a proof-of-concept applied to the GDPR domain, with the aim to detect infringements of privacy compulsory norms or to prevent possible violations using BPMN and Regorous engine.

Keywords. LegalRuleML, Akoma Ntoso, Legal Ontology, Legal Compliance Checking, GDPR Rule Modelling

1. Introduction

Over the last fifteen years, Semantic Web technologies have been successfully applied to the legal domain by defining unique identifier naming conventions for legal resources (e.g., ELI, ECLI, URN:LEX)[1], legal document vocabularies for the representation of sources (e.g., Metalex/CEN, Akoma Ntoso[2][11][15]), legal ontologies for modelling legal concepts (e.g., LKIF ontology,[3] PrOnto for GDPR[4][14]), and legal rule-based languages for modelling norms (e.g., LegalRuleML[5][1][2]). However, such components, and the related research communities, are not integrated enough to produce a robust and scientific framework that can be usable in real applications and that takes the needs of end users into account. By composing all these techniques and theoretical methods, we propose an integrated framework for modelling legal documents and legal knowledge to support legal reasoning, and in particular to check for compliance. This paper presents a proof-of-concept of this framework carried out in the Cloud for Europe (C4E) European project, where these techniques (Akoma Ntoso, PrOnto, LegalRuleML) have been applied to the GDPR domain with the aim of

[1] ELI—European Legislation Identifier https://eur-lex.europa.eu/eli-register/about.html; ECLI—European Case Law Identifier https://e-justice.europa.eu/content_european_case_law_identifier_ecli-175-en.do; URN:LEX—https://datatracker.ietf.org/doc/draft-spinosa-urn-lex/

[2] Metalex/CEN—http://www.metalex.eu/; Akoma Ntoso—Architecture for Knowledge-Oriented Management of Any Normative Texts using Open Standards and Ontologies https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=legaldocml

[3] LKIF—Legal Knowledge Interchange Format, https://github.com/RinkeHoekstra/lkif-core

[4] GDPR–General Data Protection Regulation (EU) 2016/679.

[5] LegalRuleML—https://www.oasis-open.org/committees/legalruleml/

detecting infringements of privacy rules (*ex-post analysis*) or of preventing possible violations (*ex-ante analysis*) using BPMN[6] [7] and the Regorous engine [9].

2. The Cloud for Europe Project

The Cloud for Europe project (C4E)[7] aims to design a cloud computing platform for eGovernment sevices compliant with GDPR rules. Cloud computing gives rise to several legal issues related to data protection rules: jurisdiction defines which legal system is applicable (e.g., in Germany, Section 130 of the Criminal Code bans Nazi symbolism, but no such ban exists in the USA); the geographic location of servers defines special rules (e.g., Cross-Border Data Transfer, Arts. 44–50 GDPR); security attacks in a cloud computing environment could cause multiple data breaches affecting different servers and consequently different data processors.

The solution we propose is an innovative architecture for managing legal compliance checking for public-sector cloud-computing network services. The GDPR includes several provisions that have a significant impact on this domain:

i) It defines constraints that must be included *by design* as part of information-system specifications and implementation (e.g., obligations, rights, permissions, prohibitions, penalties, remedies).

ii) It shapes policies (e.g., security, privacy) and business processes (e.g., the nodes of the brokers admitted for the transmission of data out of EU borders) that affect processing workflow.

iii) It changes over time, and this produces dynamic situations where a prompt reaction is fundamental (e.g., data breach, notification). A classic static rule engine is not enough to achieve a feasible legal-compliance-checking environment: it is necessary to include a defeasible, deontic, and temporal logic model connected with the original legal texts if (evidence based) reports are to be produced that can justify the outcomes of compliance-checking activities.

Our integrated framework prototypes (see Fig. 1) are capable of managing four main crucial functional requirements in legal compliance checking:

1. **Managing changes** made to the legal document over time, especially as this applies to acts, regulations, and contracts, which by nature are variable and subject to frequent change, significantly affecting coordination between the text and the rules that need to be remodeled. Our framework uses two different specialized Web editors integrated in a dashboard with a harmonized design interface: one for marking up the legal text in Akoma Ntoso, the other for modelling legal norms in LegalRuleML. The legal rules, the texts, and the legal ontologies are connected with one another other via FRBR (Functional Requirements for Bibliographic Records). Native NoSQL and XML databases store the legal sources marked up in Akoma Ntoso and LegalRuleML. The application level (server and client side) is able to maintain the legislative repositories updated (*point-in-time*) and to discover new pertinent law available in the Web. A legal temporal model is implemented in an application and data model based on three main parameters: a norm's time of entry into force, its time of efficacy, and its time of application to a specific case. This temporal model is extensively used in a coordinated manner in Akoma Ntoso and in LegalRuleML XML standards. It ensures that all legal

[6] BPMN— Business Process Model and Notation, http://www.bpmn.org/

[7] https://www.fokus.fraunhofer.de/en/dps/projects/cloudforeurope

rules affected by a legislative change are detected. In particular, if a legislative text changes, we can also detect the business processes that have been affected or are missing, so as to make it possible to promptly update the entire system.

2. **Modelling Legal concepts** using the PrOnto ontology for the data protection legal domain, and in particular for the GDPR.

3. The **legal reasoning** component uses the legal sources previously marked up using the Web editors. The legal reasoning engine is based on defeasible and temporal logic specific to the pertinent legal domain, and it is also scalable and computable with the relevant volume of rules [9]. It ensures legal compliance checking by means of a specific algorithm making it possible to answer queries submitted by cloud service providers or by the national service brokerage infrastructure. When a fact or a service is required, the cloud computing infrastructure asks the legal reasoning engine to verify the legality of the operation using, among the other resources, the contract's general provisions, the relevant case law, and soft-law policies. The result of legal reasoning is a report detailing violation [3], remedies, and possible alternative solutions that need to be interpreted by decision-makers (cloud actors).

4. **Business process** integration with legal reasoning is necessary in order to guarantee the correct application of technical operations, events, and processes connected with cloud computing services. To that end we have a special editor for modelling business processes using BPMN 2.0. This module is invoked in each legal action to determine whether the legal rules are also consistent with the real applicative scenario. Legal reasoning is also invoked when a law changes and the need arises to check whether business processes are still compliant with the new modified legislative scenario. If not, an alert is sent to the business process designer to update the workflow component that works with the cloud computing service platform and with the content management system.

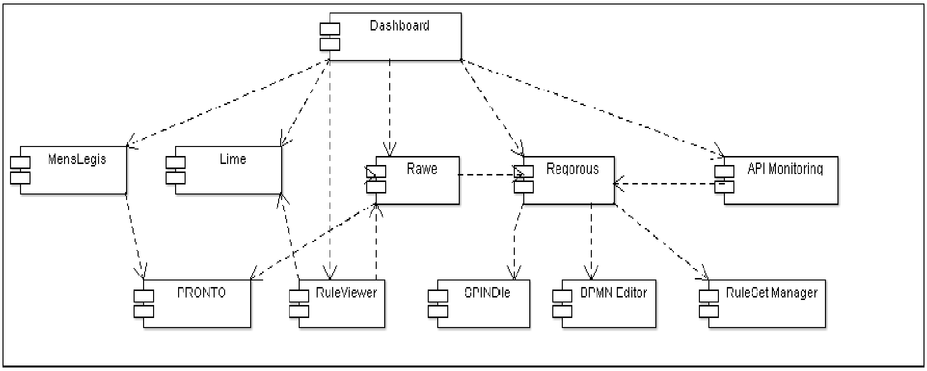


Figure 1 - Legal Tools Architecture

The architecture (Fig. 1)[8] was implemented and a prototype was tested under the Cloud 4 Europe project. Three legal experts manually modelled the rules and checked the BPMN. In our framework, presented below, we find LIME and RAWE [12], which are two web editors (JavaScript) capable of semi-automatically marking up the text in Akoma Ntoso and the manually formalized norms in LegalRuleML. PrOnto is a legal ontology for modelling GDPR concepts and axioms. It feeds concepts and predicates to the legal rule-modelling layer in order to make the formalization consistent and

[8] http://sinatra.cirsfid.unibo.it/c4eu-dashboard/

harmonized. Regorous [9] is a tool (written in Java) that makes it possible to design BPMN 2.0 and to connect each step of the process with the legal rules. Regorous provides an API to SPINdle [10], a defeasible legal reasoning engine. Regorous presents at the end the results of compliance checking in a user interface for the end user.

3. Use-Case Scenario

A student wants to access an online service provided by a public school platform in cloud computing. The platform provides an online environment where student and parents can access grades, information, administrative communications, and courseware. Students can also upload their material connected with training activities so as to share it with other students and teachers. Additionally, the platform includes a chatline with specialized school staff (e.g., psychology counselling, health service). Art. 8 GDPR reads as follows:

"Article 8 - Conditions applicable to child's consent in relation to information society services

1. Where point (a) of Article 6(1) applies, in relation to the offer of information society services directly to a child, the processing of the personal data of a child shall be lawful where the child is at least 16 years old. Where the child is below the age of 16 years, such processing shall be lawful only if and to the extent that consent is given or authorised by the holder of parental responsibility over the child. Member States may provide by law for a lower age for those purposes provided that such lower age is not below 13 years."

In order to access to the service/platform the student must (i) agree to the general service conditions for authentication and (ii) provide the consent for the controller's processing of personal data (Art. 4 GDPR), including sensitive data (Art. 6 GDPR).
The BPMN modelling of the process above is illustrated in Fig. 2 below.

4. Legal Knowledge Modelling Framework

The lifecycle of legal-knowledge management starts by modelling Art. 8 GDPR in Akoma Ntoso so as to describe the structure of the provisions, the normative references, and the legal concepts using the PrOnto ontology and also the temporal parameters (e.g., entry into force). After this step the rules are modelled and connected with the BPMN. Finally, the Regorous engine provides the result of compliance checking.

4.1. LegalRuleML Metamodel Extension

The current LegaRuleML metamodel is very elementary and intended to design LegalRuleML constructs. However, it is a good starting point for developing extensions suitable for other goals, like compliance checking. Table 1 shows an extension of the LegalRueML metamodel included in the PrOnto ontology. It includes relationships between deontic operators, disjoint-class axioms, and better modelling of remedies, violations, and penalties.

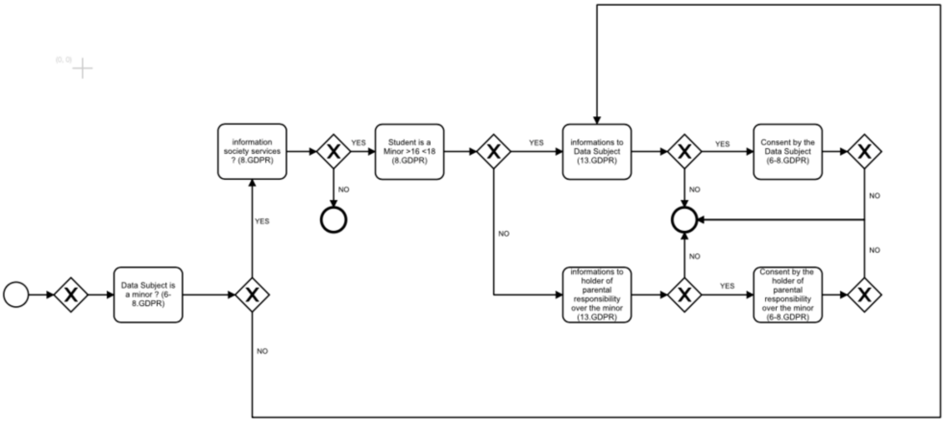

(0, 0)
Data Subject is a minor ? (6-8.GDPR)
YES
NO
information society services ? (8.GDPR)
YES
NO
Student is a Minor >16 <18 (8.GDPR)
YES
NO
informations to Data Subject (13.GDPR)
informations to holder of parental responsibility over the minor (13.GDPR)
YES
NO
NO
YES
Consent by the Data Subject (6-8.GDPR)
Consent by the holder of parental responsibility over the minor (6-8.GDPR)
NO
NO

Figure 2 – BPMN modelling of an eGovernment service

In particular, some predicates and axioms are included in PrOnto for modelling deontic relationship like `repairs`, which connects a PenaltyStatement with a LogicalFormulaStatement. Another example is the restriction applied to the `generates` property in order to model the Obligation generated by Right for the AuxiliaryParty. In LegalRuleML we have Reparation and PenaltyStatement, where Remedy links a PenaltyStatement with a PrescriptiveStatement, and the PenaltyStatement is a Deontic Specification.

Table 1. Some axioms in the extension of the LegalRuleML ontology

Reparation <owl:ObjectProperty rdf:about="https://w3id.org/ontology/pronto#repairs"> <rdfs:domain rdf:resource="http://docs.oasis-open.org/legalruleml/ns/v1.0/metamodel#PenaltyStatement"/> <rdfs:range rdf:resource="http://docs.oasis-open.org/legalruleml/ns/v1.0/metamodel#PrescriptiveStatement"/> </owl:ObjectProperty>
Restriction: Obligation *hasHeld* CounterParty generated by a Right <owl:ObjectProperty rdf:about="https://w3id.org/ontology/pronto#generates"> <rdfs:domain rdf:resource="http://docs.oasis-open.org/legalruleml/ns/v1.0/metamodel#Right"/> <rdfs:range> <owl:intersectionOf rdf:parseType="Collection"> <owl:Class rdf:about="http://docs.oasis-open.org/legalruleml/ns/v1.0/metamodel#Obligation"/> <owl:Restriction> <owl:onProperty rdf:resource="https://w3id.org/ontology/pronto#hasHeld"/> <owl:someValuesFrom rdf:resource="http://docs.oasis-open.org/legalruleml/ns/v1.0/metamodel #AuxiliaryParty"/> </owl:Restriction> </owl:intersectionOf> </rdfs:range> </owl:ObjectProperty>

In the scenario of Art. 8 GDPR, we have the following OWL-DL:

Table 2. Legal Axioms

Legal concepts OWL-DL Axioms	Individual(PrOnto:child type(PrOnto:person))
	Individual(PrOnto:child type(PrOnto:person))
	SubClassOf(PrOnto:data_controller PrOnto:data)
	SubClassOf(PrOnto: personal_data_processing PrOnto: process)
	Individual(PrOnto: information_society_service PrOnto:process)
	Individual(PrOnto: obligation_to_obtain_consent PrOnto:obligation)
	ObjectProperty(PrOnto:has_at_least16years domain(PrOnto:child) range(PRONTO:status))

4.2. Legal Rule Modelling: LegalRuleML Formalization

Art. 8 is modelled using the RAWE graphic tool implemented using Scratch diagrams[9] (see Fig. 3) to help legal experts approach logic formalization. The idea is that the following logic rule is directly modelled using visual diagrams even in order to properly connect the PrOnto legal ontology:

Table 3. Logic rule modelling

Logic rule modelling	IF
	personal_data_processing(d,x) ∧ *child*(x) ∧ *at_least16years*(x) ∧ *information_society_service*(s,d) ∧ *data_controller*(y,s)
	THEN
	obligation_to_obtain_consent(y,x,s)

Thanks to the official legal text in a single window, we can also model the legal rules in LegalRuleML connected to the ontology terms previously marked up in the text (see Fig. 4). Secondly, the XML id attribute connects the rules with the original legal official text. This helps to detect the rules that need to be updated when a legal text changes. LegalRuleML manages temporal defeasible logic to detect the correct set of rules point-in-time. The Art. 8 GDPR admits being trumped by domestic regulation.[10] At present in Europe different age limitations are in place (e.g, age 13 in Spain; 14 in Italy;[11] 15 in France).[12] LegalRuleML makes it possible to use defeasible operators (e.g., <lrml:appliesStrength iri="lrmlv:Defeasible"/>), implementing hierarchies between rules, jointly with metadata that tracks jurisdictions [2] (e.g., <lrml:appliesJurisdiction keyref="jurisdictions:it"/>).

4.3 Checking for Compliance

The LegalRuleML representation of norms is imported in Regorous via an API. Thus, the RAWE LegalRuleML is also imported in the Regorous editor with the corresponding BPMN previously designed by the person in charge of the eGov service. A legal expert annotates the BPMN tasks with the terms present in the LegalRuleML

[9] https://scratch.mit.edu/

[10] "Member States may provide by law for a lower age for those purposes provided that such lower age is not below 13 years."

[11] Legislative Decree n. 101 of 10 August 2018.
http://www.gazzettaufficiale.it/atto/serie_generale/caricaDettaglioAtto/originario?atto.dataPubblicazioneGazzetta=2018-09-04&atto.codiceRedazionale=18G00129&elenco30giorni=false

[12] https://www.betterinternetforkids.eu/web/portal/practice/awareness/detail?articleId=3017751

rules. The ex-ante checking is guaranteed on the basis of the BPMN. The live monitoring is guaranteed using the flow of data coming from the eGov service. Using these log files as facts, we can check the operations' compliance with the rules in Art. 8 GDPR.

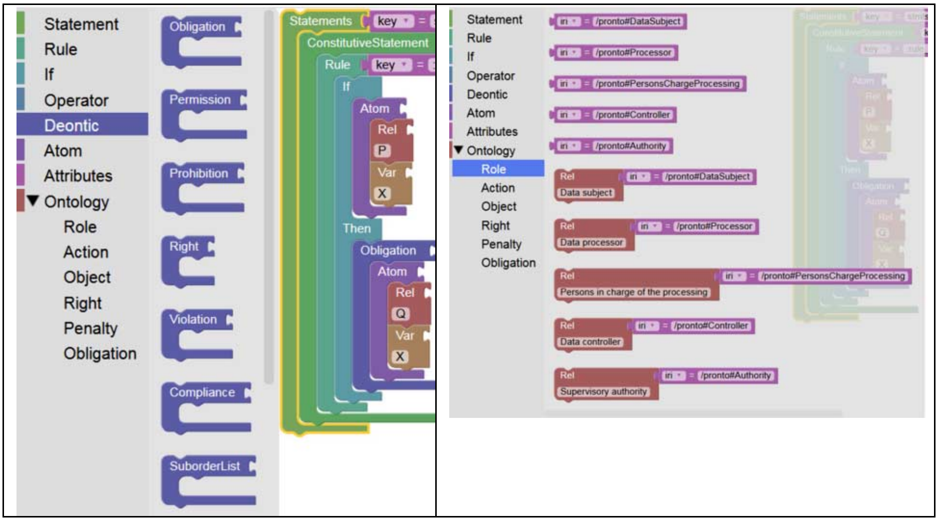
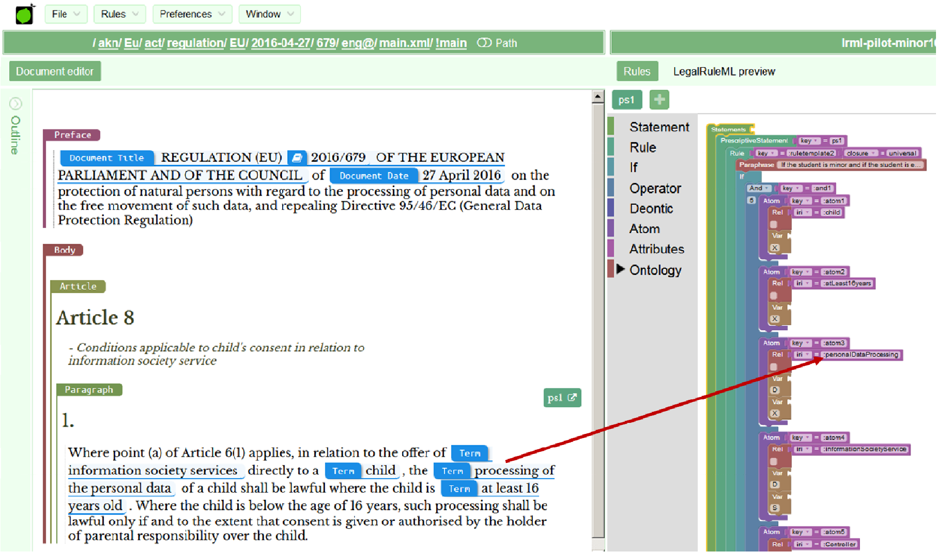


Figure 3 – RAWE web editor for modelling legal rules using graphic diagrams



Preface

Document Title REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of Document Date 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

Body

Article

Article 8

- *Conditions applicable to child's consent in relation to information society service*

Paragraph

1.

Where point (a) of Article 6(1) applies, in relation to the offer of Term information society services directly to a Term child , the Term processing of the personal data of a child shall be lawful where the child is Term at least 16 years old . Where the child is below the age of 16 years, such processing shall be lawful only if and to the extent that consent is given or authorised by the holder of parental responsibility over the child.

Figure 4 – RAWE web editor for modelling the connection with text, ontology concepts, and rules

The Regorous [9] mechanism for managing compliance checking is as follows: for ex-ante (design time) compliance checking, Regorous dynamically generates the traces in the process (corresponding to the possible ways in which the process can be executed), and for every step it computes the state of the process after executing the corresponding task, so as to then make a call to SPINdle [10] to determine the obligations, prohibitions, and permissions in force at that particular step. It then compares the state with the legal requirements in force to determine whether they have been fulfilled or violated, or whether the status is still pending. It repeats the procedure for all tasks in the trace, and then for the full set of traces. At the end it reports the compliance status according to the following conditions:

- A trace is compliant if no task in it results in a violation.
- A trace is weakly compliant if every violation is compensated for.
- A process is (weakly) compliant if, and only if, all its execution traces are (at least weakly) compliant.
- A process is partly compliant if, and only if, there is at least one compliant trace.

The live monitor (run-time compliance checking) uses the log files generated by the underlying process-execution workflow, and it first extracts the information from it and passes them as facts for each executed task. At this stage, we can use the same mechanism as the design-time procedure, noting that the log for an instance corresponds to a single trace in the process.

5. Related Work

Over the last decade, the problem of modelling legal knowledge has been addressed by different scholars [15], but unfortunately in a patchy manner, without any integrated vision that combines different technologies (e.g., Semantic Web, AI and Law, NLP) into a single usable framework. Several different standards for modelling text and rules arise (RuleML, SWRL, RIF, LKIF, ODRL, etc.), but they are not specific to the task of modelling the complexity of different legal contexts. Unlike any of these projects, our approach (i) connects text with rules for tracking changes over time, (ii) uses a legal reasoning level on top of the ontological layer of the Semantic Web stack, (iii) permits multiple alternative interpretations, and (iv) connects BPMN with the legal reasoning layer. None of Semantic Web previously mentioned languages and technologies are compliant with the guidelines established in [5] [13] for representing legal knowledge and legal reasoning. In addition, these approaches have severe limitations when it comes to modelling legal reasoning, since they do not provide a conceptually sound model of legal reasoning [6]. A good project is [4], but it is not totally integrated with LegalRuleML.

6. Conclusions

We have presented an integrated framework for checking compliance with legal rules, focusing in particular on the GDPR. We have used Akoma Ntoso to model the text, PrOnto to model legal concepts, LegalRuleML to model norms, and Regorous to combine BPMN and facts with the rules expressed in LegalRuleML and to provide the final report. We have provided an integrated user interface and dashboard based on diagrams to help legal engineers align with the legal text as it evolves over time. This framework makes it possible to track the changes a legal text goes through over time, and hence detect the legal rules that need to be updated. From the text it imports constraints on the metadata level: these constraints include temporal information and jurisdiction. Finally, the legal ontology connected with the text is imported into the legal rules, without redundancy or errors, so as to maintain a coherent taxonomy of predicates in the rule base.

Acknowledgement

This work was supported by Cloud for Europe project no. 610650 FP7-ICT, and was also partly supported by the European Union’s Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement no. 690974, “MIREL: MIning and REasoning with Legal Texts.”

Annex A

X is a “child,” D is “personal data, S is the “information society service” and Y is the “controller.”

```
<lrml:LegalRuleML xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns="http://docs.oasis-open.org/legalruleml/examples/compactified/ex9-alternatives-compact#"
xmlns:lrml="http://docs.oasis-open.org/legalruleml/ns/v1.0/" xmlns:ruleml="http://ruleml.org/spec"
xmlns:rulemlmm="http://ruleml.org/1.0/metamodel#" xml:base="http://docs.oasis-
open.org/legalruleml/examples/compactified/ex9-alternatives-compact"
xsi:schemaLocation="http://docs.oasis-open.org/legalruleml/ns/v1.0/ ./xsd-schema/compact/lrml-
compact.xsd">
omissis
    <lrml:Statements>
        <lrml:PrescriptiveStatement key="ps1">
            <ruleml:Rule key=":ruletemplate2" closure="universal">
                <lrml:Paraphrase> If the student is minor and if the student is emancipated, in any case,
he/she can provide autonomous consent, if it is considered an action of ordinary
administration</lrml:Paraphrase>
                <ruleml:if>
                    <ruleml:And key=":and1">
                        <ruleml:Atom key=":atom1">
                            <ruleml:Rel iri=":child"/>
                            <ruleml:Var>X</ruleml:Var>
                        </ruleml:Atom>
                        <ruleml:Atom key=":atom2">
                            <ruleml:Rel iri=":atLeast16years"/>
                            <ruleml:Var>X</ruleml:Var>
                        </ruleml:Atom>
                        <ruleml:Atom key=":atom3">
                            <ruleml:Rel iri=":personalDataProcessing"/>
                            <ruleml:Var>D</ruleml:Var>
                            <ruleml:Var>X</ruleml:Var>
                        </ruleml:Atom>
                        <ruleml:Atom key=":atom4">
                            <ruleml:Rel iri=":informationSocietyService"/>
                            <ruleml:Var>D</ruleml:Var>
                            <ruleml:Var>S</ruleml:Var>
                        </ruleml:Atom>
                        <ruleml:Atom key=":atom5">
                            <ruleml:Rel iri=":Controller"/>
                            <ruleml:Var>Y</ruleml:Var>
                            <ruleml:Var>S</ruleml:Var>
                        </ruleml:Atom>
                    </ruleml:And>
                </ruleml:if>
                <ruleml:then>
                    <lrml:Obligation iri=":obligation">
                        <ruleml:Atom key=":atom6">
                            <ruleml:Rel iri=":ObtainConsent"/>
```

```
                                <ruleml:Var>X</ruleml:Var>
                                <ruleml:Var>Y</ruleml:Var>
                                <ruleml:Var>S</ruleml:Var>
                        </ruleml:Atom>
                    </lrml:Obligation>
                </ruleml:then>
            </ruleml:Rule>
        </lrml:PrescriptiveStatement>
    </lrml:Statements>
</lrml:LegalRuleML>
```

```
<lrml:hasQualification>
    <lrml:Overrides over="#ps2" under="#ps1"/>
</lrml:hasQualification>
```

References

[1] Athan, T., Boley, H., Governatori, G., Palmirani, M., Paschke, A., Wyner, A.: OASIS LegalRuleML. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law, pp. 3–12, New York (2013)

[2] Athan, T.; Governatori, G.; Palmirani, M.; Paschke, A.; Wyner, A., LegalRuleML: Design Principles and Foundations, in: Reasoning Web. Web Logic Rules 2015, LNCS, vol. 9203, pp. 151–188, Springer, Heidleberg (2015)

[3] Elgammal, A., Türetken, O., van den Heuvel, W.J.: Using patterns for the analysis and resolution of compliance violations. Int. J. Coop. Inf. Syst. 21(1), 31–54 (2012)

[4] Gandon, T.F., Governatori G., Villata S., Normative Requirements as Linked Data, JURIX2017, Amsterdam, IOS, (2017)

[5] Gordon, T.F., Governatori, G., Rotolo, A.: Rules and norms: Requirements for rule interchange languages in the legal domain. In: Governatori, G., Hall, J., Paschke, A. (eds.) RuleML 2009. LNCS, vol. 5858, pp. 282–296. Springer, Heidelberg (2009)

[6] Governatori, G., Hashmi, M.: No time for compliance. In: Hallé, S., Mayer, W. (eds.) EDOC 2015, pp. 9–18. IEEE Press (2015)

[7] Governatori, G., Mustafa M., Lam H.-P., Villata S. and Palmirani, M.: Semantic Business Process Regulatory Compliance Checking using LegalRuleML, In Proceedings of the 20th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2016), November, 2016, Bologna, Italy (2016)

[8] Governatori, G.: On the relationship between Carneades and defeasible logic. In: van Engers, T. (ed.) Proceedings of the 13th International Conference on Artificial Intelligence and Law (ICAIL 2011), pp. 31–40. ACM Press (2011)

[9] Governatori, G.: The Regorous approach to process compliance. In 2015 IEEE 19th International Enterprise Distributed Object Computing Workshop, pages 33–40. IEEE Press (2015)

[10] Lam, H.-P., Governatori, G.: The making of SPINdle. In: Governatori, G., Hall, J., Paschke, A. (eds.) RuleML 2009. LNCS, vol. 5858, pp. 315–322. Springer, Heidelberg (2009)

[11] Palmirani M, Vitali F, Akoma-Ntoso for Legal Documents, in: Legislative XML for the Semantic Web. Principles, Models, Standards for Document Management, BERLIN, Springer Verlag, 2011, pp. 75 – 100.

[12] Palmirani, M., Cervone, L., Bujor, O., Chiappetta, M.: RAWE: an editor for rule markup of legal texts. In: Fodor, P., Roman, D., Anicic, D., Wyner, A., Palmirani, M., Sottara, D., Lévy, F. (eds.) CEUR Workshop Proceedings, Seattle, USA, 11–13 July 2013, vol. 1004, CEUR-WS.org (2013)

[13] Palmirani, M., Governatori, G., Contissa, G.: Temporal dimensions in rules modelling. In: Winkels, R. (ed.) JURIX. Frontiers in Artificial Intelligence and Applications, vol. 223, pp. 159–162. IOS Press, Amsterdam (2010)

[14] Palmirani, M., Martoni, M., Rossi, A., Bartolini, C., Robaldo, L.: PrOnto: Privacy Ontology for Legal Reasoning. EGOVIS2018, 7th International Conference, EGOVIS 2018, Regensburg, Germany, September 3-5, 2018, Proceedings. LNCS 11032, Springer, pp. 139-152 (2018)

[15] Sartor, G., Palmirani, M., Francesconi, E., Biasiotti, M.: Legislative XML on Semantic web. Springer, Heidelberg (2011)

Legal Knowledge and Information Systems
M. Palmirani (Ed.)

doi:10.3233/978-1-61499-935-5-111

Segmenting U.S. Court Decisions into Functional and Issue Specific Parts

Jaromír ŠAVELKA [a,b,1] and Kevin D. ASHLEY [a,b,c]
[a] *Intelligent Systems Program, University of Pittsburgh, USA*
[b] *Learning Research and Development Center, University of Pittsburgh, USA*
[c] *School of Law, University of Pittsburgh, USA*

Abstract. In common law jurisdictions, legal research often involves an analysis of relevant case law. Court opinions comprise several high-level parts with different functions. A statement's membership in one of the parts is a key factor influencing how the statement should be understood. In this paper we present a number of experiments in automatically segmenting court opinions into the functional and the issue specific parts. We defined a set of seven types including Background, Analysis, and Conclusions. We used the types to annotate a sizable corpus of US trade secret and cyber crime decisions. We used the data set to investigate the feasibility of recognizing the parts automatically. The proposed framework based on conditional random fields proved to be very promising in this respect. To support research in automatic case law analysis we plan to release the data set to the public.

Keywords. case law, legal analysis, information retrieval, text segmentation, conditional random fields

1. Introduction

In this paper we examine an application of natural language processing (NLP) and machine learning (ML) to facilitate one of the initial steps in case law analysis. Court opinions consist of several high-level parts each of which has a different function. The main Analysis part often contains several sub-parts each of which is dedicated to a different issue. Distinguishing the functional as well as issue-specific parts is crucial for a lawyer to be able to focus attention on the pieces of the opinion that matter. We are interested if and how could NLP and ML techniques be helpful in recognition of the individual parts. We address this question by assessing the ability of the presented NLP/ML pipeline to perform this step in the analysis automatically.

2. Background and Motivation

We define the task as a two step process (diagrammed in Figure 1). In the first step an opinion is segmented into a varying number of consecutive non-overlapping parts. Each part is assigned one of the following types:

[1] Corresponding Author: Jaromír Šavelka, Learning Research and Development Center, 3939 O'Hara St, Office 520, Pittsburgh, PA 15260, USA; E-mail: jas438@pitt.edu.

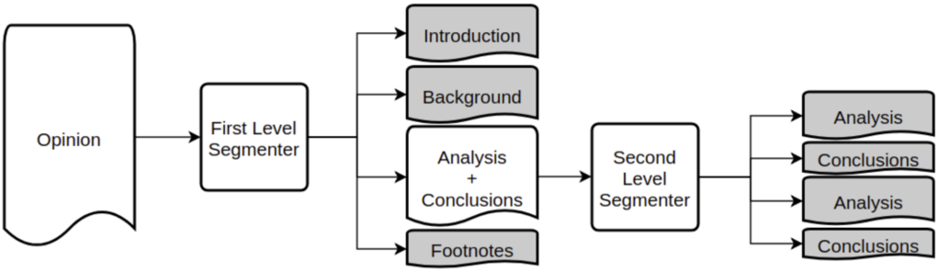


Figure 1. The diagram shows the two-step segmentation process and its interim and final outcomes.

1. **Introduction** – the opening part which typically consists of lines indicating the deciding court, judges, the case citation, parties, etc. It is not uncommon that the court would include a summary of the decision.
2. **Background** – the part where the court describes the procedural history of the case, the relevant facts, as well as what the parties are claiming. Its tone is usually descriptive, i.e., the court refrains from expressing its own opinions.
3. **Analysis** – the part where the court discusses and reasons about the issues of the case and states its outcome. Quite often the tone is deliberative, i.e., the court expresses opinions on the issues, arguments, or claims. The court may deal with a single issue or it may treat several issues separately.
4. **Concurrence or Dissent** – the part where opinions of concurring or dissenting judges are presented. There may be dedicated sections for each concurrence/dissent but there could also be just a single sentence informing about the list of concurring/dissenting judges.
5. **Footnotes** – a list the indices of which are references to different parts of a decision. Each item of the list provides additional information as to what is in the text at the place of reference.
6. **Appendix** – a separate document attached to a decision as supplemental material.

The type system is a variation on type systems presented by a number of authors in the past (see Section 3 on related work). It is specifically tailored for court decisions in the U.S. In our opinion it would generalize well to other jurisdictions. Note that the task is more complex than merely identifying sections or section headings. Most of the time the decisions are not explicitly segmented into sections that would map onto the scheme.

In the second step the Analysis part, that typically spans the larger portion of an opinion, is annotated with the one additional type:

7. **Conclusions** – the part where a court states the outcome of its analyses, i.e., its decision concerning each issue it addresses.

The annotations are then used to segment the Analysis part into the treatments of specific issues. Although, the situation is more complex than that, one could adopt an assumption that argumentation about a specific issue is finished with its outcome (Conclusions). The segments could then be obtained in a straightforward way.

This line of work has been recognized as "important but often neglected" in [25] where the author speaks about segmentation of a legal document into its structural elements such as, e.g., facts, arguments, and rulings. As explained in [15], "The ability to identify and partition a document into segments is important for many NLP tasks, including information retrieval, summarization, and text understanding." Mainly, this is because the segments provides clues to the meaning of their contents [1].

For instance, knowing in which high-level part a sentence appears will help to annotate sentences in terms of the roles they play in legal argument. Sentences that state a finding of fact or state a legal rule are more likely to be found in the Background or Analysis parts, respectively. Annotating Conclusions will help to indicate where the analysis of one issue ends and another begins. We plan to use document segmentation as the next step in our long-term project of automatically analyzing the texts of court decisions to support statutory interpretation [21,22,23].

The annotation task also may have important pedagogical potential. While it is easy for law student annotators to identify instances of Background, Concurrence or Dissent, and Footnotes, it is more difficult to identify a court's conclusions regarding the issues raised. Student annotators could benefit from this kind of practice while providing training instances for machine learning and legal text analytics.

3. Related and Prior Work

Unlike our project, much of the related work on segmenting texts into multi-paragraph passages focuses on dividing the segments by topic (e.g., [14] and [15]). Some domain general approaches to segmenting texts into multi-paragraph passages by topic are based on statistical similarity and lexical cohesion, the repetition of similar words in coherent segments and the tendency for vocabulary to change across segment boundaries (e.g., [4] and [12]). Segmenting legal texts into topics or, as in our project, into functional sections or parts, has required the application of more legal domain-specific knowledge [7,26]. For instance, one must first settle on the types of functional sections that are present in the legal texts of interest such as courts' legal decisions. In [5], based on discussions with a focus group of West editors and an exercise in which each editor tagged 2-3 cases, the authors developed a short list of segment types which could be linked to fixed issues or annotations: Issues / Contentions (Substantive and Procedural) Analysis, Facts/Evidence, and Conclusions (Abstract and Concrete).

One approach to segmentation has focused on automatically identifying the rhetorical roles of sentences. For instance, a case document to be summarized has been divided into parts for purposes of selecting the important sentences and organizing them into a summary based on a standard model of case structure [7,10,16]. Supervised ML has been applied to learn rhetorical role sentence classifiers based on a wide range of features [10,18]. Unsupervised ML was applied to select relevant portions of texts for summaries [16]. [16]

Other work has focused more specifically on linguistic analysis of sectional texts to identify features characteristic of section types. The authors of [9] employed verb tense and aspect in sentences stating legal background knowledge, case description, or a judge's opinions. [3] employed other linguistic markers with contextual dependencies to construct a thematic structuring rule base for contextual exploration. In [7] the authors employed linguistic markers to segment Canadian decisions into four units: Introduction, Context, Juridical Analysis, and Conclusion. The first three unit types appear to map onto our Introduction, Background, and Analysis. The Conclusion appears to correspond to the "Final decision of the court," whereas we treat it more broadly; where the court treats several issues separately there may be a conclusion at the end of the Analysis of each issue. A similar scheme was proposed in [11], including some additional types such as Dissent, Footnotes, or Party Claims. Identifying Conclusions in our work appears to be related to some aspects of the work presented in [27]. The authors identify typical language structures that are used in various types of Premises or Conclusions. These are then expressed in the form of Context Free Grammar for parsing legal arguments. Some of the types used in the pipeline presented in [8] and [2] appear to partially map to our Conclusions as well.

By contrast to the above work, however, we have not employed rhetorical roles or linguistic analysis in our project.

In [19] conditional random fields (CRF) were applied to segment legal documents into seven labeled components with each label representing a corresponding rhetorical role. CRF were applied to identify the rhetorical roles: identifying case (F_1 .853), establishing case facts (F_1 .824), arguing case (F_1 .805), case history (F_1 .851), arguments (F_1 .787), decision ratio (F_1 .888), final decision (F_1 .973). Features included key phrases, named entities recognized, proper names, location in layout, and legal vocabulary, neighboring sentence similarity, paragraph structure, and citation [18].

The authors of [26] applied other ML algorithms (naive Bayes, logistic regression, decision trees, support vector machines and neural networks) to identify sections but only of legal briefs and considering only section titles. The authors in [24] employed a naive Bayes multinomial classifier to automatically annotate legal principles in case texts based on features including deontic modalities of verbs such as must, may, or should.

We employ a corpus containing considerably more legal decisions than in the above work and covering a wider range of legal domains. We also use the sentence boundary detection system developed in [20] and [22]. This is a crucial component that allows us to use sentence-level segments and sentence-level features in the presented NLP/ML pipeline.

4. Experimental Design

4.1. Data Sets

We downloaded 316 court decisions from the online Court Listener[2] and Google Scholar[3] services. Of these 143 are from the area of cyber crime (cyber bullying, credit card frauds, possession of electronic child pornography), and 173 cases involve trade secrets (typically misappropriations). The trade secret part of the corpus is a slightly extended data set assembled in [6]. We use cases from the two different areas of law to gain a sense of how well the trained models generalize (see Section 6 for details).

We created guidelines for manual annotation of the decisions[4] with the types introduced in Section 2. Two human annotators (the authors) then annotated the decisions using Gloss, the web-based annotation environment developed by the authors. Each decision was annotated by one of the annotators. A subset (25 from each domain, i.e., 50

[2] `www.courtlistener.com`

[3] `scholar.google.com`

[4] Available at `https://github.com/jsavelka/us-dec-func-iss-sgm.git`.

	Cyber Crime			Trade Secrets			Total		
	Filtered	Original	Agree	Filtered	Original	Agree	Filtered	Original	Agree
Documents	139	143	–	158	173	–	297	316	–
Introduction	139	143	.965	158	173	.934	297	316	.947
Background	133	139	.758	152	172	.787	285	311	.774
Analysis	139	143	.932	158	173	.936	139	143	.935
Conclusions	398	429	.689	833	909	.763	398	1338	.732
Concurr/Diss	0	18	1.0	0	44	1.0	0	62	1.0
Footnotes	95	97	.983	103	113	.982	198	210	.982
Appendix	0	7	–	0	6	–	0	13	–

Table 1. The data sets before (Original) and after processing (Filtered) with inter-annotator agreement (Agree).

in total) was annotated by both authors to measure inter-annotator agreement (see Table 1). The inter-annotator agreement is computed using the following formula:

$$A = \sum_{a=0}^{1} \sum_{i=1}^{|S|} \frac{t(s_i,a) = T \wedge t(s_i,\neg a) = T}{t(s_i,a) = T}$$

In the formulas S stands for the set of all sentences in the corpus; T means a specific type; $t(s_i,0)$ stands for the type of the sentence s_i assigned by the first human annotator; $t(s_i,1)$ those assigned by the second human annotator; $\neg$ is a negation that can be applied only to 0 or 1 in order to reverse them into each other. We use the accuracy formula because it maps nicely to the F_1-measure that we use for the evaluation of the proposed framework. This allows us to assess how well the machine performs when compared to a human.

Table 1 provides detailed statistics of the created annotations. Certain pre-processing steps involving exclusion of some documents as well as parts of others were necessary. We report the statistics of the full data set in the "Original" columns of Table 1; statistics of the data set after pre-processing are reported in the "Filtered" columns. These were performed in order to keep the experiments reported in this paper focused on the main task. A small number of decisions have a non-standard structure where, for example, Background interleaves with Analysis and Conclusions. Since we do not have enough data in our data set to deal with these, we decided to exclude the documents from our experiments. In addition, it turned out that we have very little data for Appendix and Concurrence and Dissent types. We manually eliminated them out for the purposes of the experiments reported here.

4.2. Classification Pipeline

The classification pipeline is schematically depicted in Figure 2. Each document is first split into individual paragraphs. Since the opinions are relatively clean a simple regular expression was used to perform this step. Information about the document from which a sentence comes and its order in a sequence are retained. The paragraphs are then split into individual sentences using the sentence boundary detection system of [20].

The paragraphs and sentences are transformed into vectors of paragraph level features. The features include, e.g., lower-case tokens and POS-tagged lemmas that appear in a paragraph, the position of a paragraph within a document, its length as well as the average length of sentences it contains. The first and the last five tokens (paragraph boundaries) are described through more detailed features, such as a token's signature, length, and type (i.e., digit, case, white space).

The resulting feature vectors and human created annotations (as labels) are then used in the training of the Functional parts segmenter (the first step of the analysis). The segmenter consists of three CRF models. A CRF is a random field model that is globally conditioned on an observation sequence O. The states of the model correspond to event labels E. We use a first-order CRF in our experiments (observation O_i is associated with E_i). We use the CRFsuite[5] implementation of a first-order CRF [13,17].

The three models are trained in an iterative manner. First, the model for recognizing the Introduction type is trained on full texts. Then we train a model for separating the Background type from the rest in documents that were stripped of the Introduction parts. Finally, a model for finding the boundary between the Analysis and Footnotes is trained on the documents that were stripped of the Introduction and Background types.

The sentences from the Analysis part are transformed into vectors of sentence level features. The types of features are similar to those described above except they are applied at a different level (sentence as opposed to paragraph). The feature vectors and human-created annotations are then used in the training of the Conclusions recognizer (the second step of the analysis). The recognizer consist of the single CRF model.

The two resulting components—the Functional parts segmenter and the Conclusions recognizer—are then used in the automated annotation process on the unseen document. This process is similar to the one described above. Where the models required human-created annotations during training, the automatically generated annotations are used. This happens in the interim stages of the Functional parts segmenter and when feeding the (predicted) Analysis part to the Conclusions recognizer.

4.3. Evaluation

We use 10-fold cross validation as the evaluation method. This means that the data set is split into 10 folds of roughly equal sizes. The splitting is performed at the level of documents, i.e., all the sentences from a single document are in the same fold. We did not consider the size of a document for the purposes of splitting. Therefore, the number of sentences in each fold could vary significantly. We evaluate performance on each fold separately based on the model trained on the other 9 folds. The reported results are an aggregate from all the 10 folds.

We use precision (P), recall (R), and F$_1$-measure (F$_1$), i.e., the traditional information retrieval measures, to evaluate performance of the presented pipeline. The performance is evaluated at the level of sentences where the type of the reported measures is micro. Therefore, measures are computed as follows:

$$P = \sum_{i=1}^{|S|} \frac{t(s_i,g) = T \wedge t(s_i,p) = T}{t(s_i,p) = T} \qquad R = \sum_{i=1}^{|S|} \frac{t(s_i,g) = T \wedge t(s_i,p) = T}{t(s_i,g) = T} \qquad F_1 = \frac{2PR}{P+R}$$

[5] www.chokkan.org/software/crfsuite

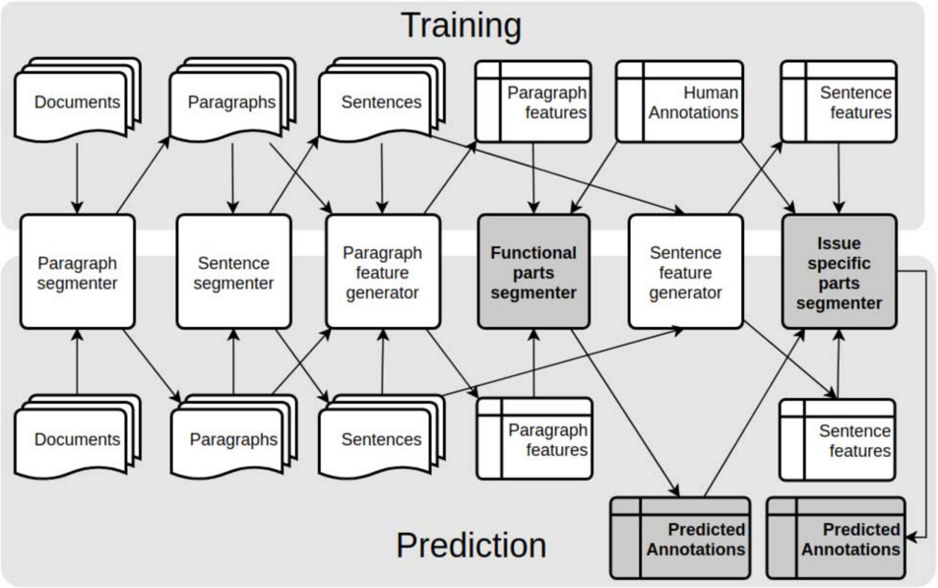


Figure 2. The diagram shows the automatic classification pipeline (training and prediction).

	Cyber Crime			Trade Secrets			Total		
	P	R	F_1	P	R	F_1	P	R	F_1
Introduction	.926	.945	.935	.907	.947	.926	.914	.947	.930
Background	.634	.752	.689	.640	.775	.701	.638	.767	.697
Analysis	.922	.879	.900	.948	.871	.908	.939	.873	.905
Footnotes	.874	.963	.916	.845	.981	.908	.855	.973	.910

Table 2. Results of the segmentation into the high-level functional parts.

In the formulas S stands for the set of all sentences in the corpus; T means a specific type; $t(s_i, g)$ stands for the type of the sentence s_i assigned by a human annotator; $t(s_i, p)$ that assigned by the automatic pipeline.

5. Results

Table 2 summarizes the results of the experiments evaluating the feasibility of the first step described in Section 2, i.e., the segmentation of an opinion into high-level functional parts. The performance of the models differs considerably across the types but it correlates well across the two different domains as well as with the inter-annotator agreement. The recognition of the Introduction and the Analysis types has a high success rate. It appears to be comparable to the human performance. The performance on the Background and the Footnotes types is slightly lower but still very close to the human performance measured through the inter-annotator agreement (see Table 1). Due to data sparsity we did not attempt to predict the Appendix and the Concurrence and Dissent types.

Table 3 presents the results of the experiments evaluating the feasibility of the second step described in Section 2, i.e., annotation of the Analysis part with the Conclusions

	Cyber Crime			Trade Secrets			Total		
	P	R	F_1	P	R	F_1	P	R	F_1
Conclusions	.521	.461	.489	.576	.493	.532	.552	.482	.515

Table 3. Results of the Conclusions recognition.

type. Although, the performance is promising the experiments confirm that this task is very challenging. From the inter-annotator agreement it appears that this task is the most challenging one. We further elaborate on the results in Section 6.

6. Discussion

We have detected a problem of data sparsity pertaining to certain types. Whereas types such as the Introduction or the Analysis are nearly guaranteed to appear in each single document, types such as the Appendix or the Concurrence and Dissent are present only occasionally. In order to solve this problem one would have to enrich the data set with many opinions that specifically contain these parts. However, this will necessarily lead to a biased data set. The result may be a model that is over-predicting the rare types. On the other hand, unbiased sampling could require a corpus that would be prohibitively expensive (in terms of annotation labor).

We performed a detailed error analysis on the Conclusions recognition step. The sentences that were identified by the human annotator as Conclusions but missed by the system were often quite short (e.g., "There is no error."). Typically, these sentences consist of words that are quite common and are not too informative for a system to pick up the signal. Some of the shorter length sentences were recognized correctly (e.g., "The judgment of the district court is affirmed.").

With respect to the sentences that the system erroneously predicted as the Conclusions it appears that attribution is a major challenge. Some sentences included verbs like "conclude" or "hold" which are likely very suggestive for a sentence to be classified as the Conclusions. However, the sentence was not attributed to the deciding court. Instead it was attributed to some other entity, such as a lower court or a party.

Words such as "therefore" or various forms of "find" (i.e., "finds" or "finding") may often indicate the Conclusion type. However, they also appear in many sentences that are not the Conclusions of issues, for instance, "Therefore, we now address appellants' claims," "finds that defendant's reliance [on certain cases] is misplaced," "hereby finds the following facts and state separately its conclusions of law," or "because there was no jury finding against her."

The error analysis also confirmed the task is very challenging even for humans. From the trade secrets case we sampled 40 sentences that were predicted as the Conclusions by the system, but not annotated as such by the human annotator. In case of 15 of those the human annotator would be willing to change the original decision. This is tightly connected to the problem of ambiguity as to what constitutes the conclusion of an issue. For example, when are findings of fact conclusions as to an issue? In addition, a court may break an issue down into multiple sub-issues like whether a rule applies (e.g., "We conclude that this rule applies equally to both blueprints and/or drawings and customer lists ...") or whether evidence supports a conclusion (e.g., "We find that X is further evidence of Y", "We credit that testimony.")

7. Future Work

The models trained in our experiments operate on low-level textual features. While examining the errors it became clear that while these could be sufficient for certain tasks (e.g., the segmentation into functional parts) they might be insufficient for others (e.g., recognizing Conclusions). We already pointed out that an attribution resolution would likely improve the Conclusions detection.

One of the problems that we detected is data sparsity in case of certain types. The challenge is how to obtain a significantly larger set of annotated data. We hypothesize that law students can annotate legal texts as a useful pedagogical exercise and that their annotations could then be used for purposes of ML and legal text analytics. In a small pilot study, ten students in Ashley's Intellectual Property course employed Gloss to annotate the parts of the full texts of four trade secret law cases before reading the edited versions in the case book. They received a 15-page set of "Instructions for Annotating Cases Using Gloss comprising a 2-page guide to access Gloss and use it to annotate the cases, the definitions of the seven parts as above, specific guidelines for annotating each part, and multiple samples illustrating the annotations and the borders between parts. Across the four cases, the students made nearly 1000 annotations.

Given this preliminary evidence that law students can use Gloss, in future work, we plan to arrange for students to annotate legal decisions in terms of key aspects of the reasoning in a case beyond high-level parts and conclusions, including stating a legal rule, expressing a judge's holding that a rule requirement is satisfied (or not), reporting a finding of fact, describing evidence, and substantive features of legal domains such as legal factors, patterns of fact that strengthen or weaken a sides position on a claim. We also plan to evaluate if and how much they learn by testing students' knowledge gains and by monitoring any increases in the extent to which their annotations agree with an instructor's and each other's.

8. Conclusions

In this paper we examined the possibility of automatically segmenting court opinions into high-level functional (step 1) and issue specific (step 2) parts. We have shown that segmentation into the functional parts could be done automatically in a quality that is not too far from human performance. Although, the model for segmenting the Analysis part into issue-specific segments via the Conclusions recognizer shows promise, there appears to be a gap between its performance and that of a human annotator. We hope that this work will stimulate further research in segmentation of court opinions. For this reason we will release the data set we employed in these experiments.

Acknowledgements

This work was supported in part by the National Institute of Justice Graduate Student Fellowship (Fellow: Jaromir Savelka) Award # 2016-R2-CX-0010, "Recommendation System for Statutory Interpretation in Cybercrime."

References

[1] Ashley, K. *Artificial intelligence and legal analytics: new tools for law practice in the digital age*. Cambridge University Press, 2017.

[2] A. Bansal, Z. Bu, B. Mishra, S. Wang, K. Ashley, and M. Grabmair. “Document Ranking with Citation Information and Oversampling Sentence Classification in the LUIMA Framework.” *JURIX*, v. 294, p. 33. IOS Press, 2016.

[3] Berrazega, I., and Faiz, R. “Automatic structuring of Arabic normative texts.” *Proceedings of the Computational Methods in Systems and Software* 238-249. Springer, 2017.

[4] Choi, F. Y., Wiemer-Hastings, P., and Moore, J. “Latent semantic analysis for text segmentation”. *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, 2001.

[5] Conrad, J. and Dabney, D. “A cognitive approach to judicial opinion structure: applying domain expertise to component analysis.” *Proc. 8th Int’l Conf. on Artificial intelligence and Law* 1-11. ACM, 2001.

[6] Falakmasir, M.H. and Ashley, K.D. “Utilizing Vector Space Models for Identifying Legal Factors from Text.” *JURIX* 183-192, 2017.

[7] A. Farzindar and G. Lapalme. “Letsum, an automatic legal text summarizing system.” *Legal knowledge and information systems, JURIX* 11-18, 2004.

[8] Grabmair, M., K. Ashley, R. Chen, P. Sureshkumar, C. Wang, E. Nyberg, and V. Walker. “Introducing LUIMA: an experiment in legal conceptual retrieval of vaccine injury decisions using a UIMA type system and tools.” In *Proc. 15th Int’l Conf. on Artificial Intelligence and Law*, pp. 69-78. ACM, 2015.

[9] Grover, C., Hachey, B., Hughson, I., and Korycinski, C. “Automatic summarisation of legal documents.” *Proc. 9th Int’l Conf. on Artificial intelligence and law* 243-251. ACM. 2003.

[10] B. Hachey and C. Grover. “Extractive summarisation of legal texts.” *Artificial Intelligence and Law* 14,4 305-345, 2006.

[11] Harasta, J., F. Kasl, J. Misek, and J. Savelka. “Segmentation of Czech Court Decisions into Subtopic Passages.” *CEILI Workshop on Legal Data Analysis* JURIX, 2017.

[12] Hearst, M. “TextTiling: Segmenting text into multi-paragraph subtopic passages.” *Computational Linguistics* 23.1, 33-64. 1997.

[13] Lafferty, J., A. McCallum, and F. Pereira. ”Conditional random fields: Probabilistic models for segmenting and labeling sequence data.” 2001.

[14] Le, T. T. N. “A study on Hierarchical Table of Indexes for Multi-documents.” *Ph.D. Diss., Japan Advanced Inst. of Science and Technology* 1-47. 1999.

[15] Lu, Q., Conrad, J., Al-Kofahi, K., and Keenan, W. “Legal document clustering with built-in topic segmentation.” *Proc. 20th ACM Int’l Conf. on Information and Knowledge Management* 383-392. ACM. 2011.

[16] M.-F. Moens. “Summarizing court decisions.” *Information Processing and Management* 43, 6, 1748-1764. 2007.

[17] Okazaki, N. “Crfsuite: a fast implementation of conditional random fields.” 2007.

[18] Saravanan, M., and B. Ravindran. “Identification of rhetorical roles for segmentation and summarization of a legal judgment.” *Artificial Intelligence and Law* 18.1 45-76. 2010.

[19] Saravanan, M., B. Ravindran, and S. Raman. “Improving legal document summarization using graphical models.” *JURIX* 2006.

[20] Savelka, J., V. Walker, M. Grabmair, and K. Ashley. “Sentence Boundary Detection in Adjudicatory Decisions in the United States.” *Traitement Automatique des Langues* 58.2 21-45. 2017.

[21] Savelka, J., and K. Ashley. “Detecting Agent Mentions in US Court Decisions.” *JURIX*. 2017.

[22] Savelka, J., and K. Ashley. “Using Conditional Random Fields to Detect Different Functional Types of Content in Decisions of United States Courts with Example Application to Sentence Boundary Detection.” *ASAIL 2017: 2nd Workshop on Automated Semantic Analysis of Information in Legal Texts*.

[23] Savelka, J., and K. Ashley. “Extracting case law sentences for argumentation about the meaning of statutory terms.” *Proc. of the Third Workshop on Argument Mining (ArgMining2016)*. 2016.

[24] Shulayeva, O., Siddharthan, A., and Wyner, A. “Recognizing cited facts and principles in legal judgements.” *Artificial Intelligence and Law* 25(1),107-126. 2017.

[25] Schweighofer, E. “The Role of AI and Law in Legal Data Science.” *JURIX* 191-192. 2018.

[26] Vanderbeck, S., Bockhorst, J., and Oldfather, C. “A Machine Learning Approach to Identifying Sections in Legal Briefs.” *MAICS* 16-22. 2011.

[27] Wyner, A., Mochales-Palau, R., Moens, M. F., & Milward, D. Approaches to text mining arguments from legal cases. In *Semantic processing of legal texts* (pp. 60–79). Springer, Berlin, Heidelberg. 2010.

Legal Knowledge and Information Systems
M. Palmirani (Ed.)

doi:10.3233/978-1-61499-935-5-121

Analysing the Impact of Legal Change Through Case Classification

Roos SLINGERLAND, Alexander BOER & Radboud WINKELS[1]
Leibniz Center for Law, University of Amsterdam, Netherlands

Abstract. In this paper an automated solution for finding cases for analysing the impact of legal change is proposed and the results are analysed with the help of a legal expert. It focuses on the automatic classification of 15,000 judgments within civil law. We investigated to what extent several machine learning algorithms were able to classify cases 'correctly'. This was done with accuracies around 0.85. However, the data were scarce and the initial labelling not perfect, so further research should focus on these aspects to improve the analysis of the impact of legal change.

Keyword. Automatic classification, case-law, law changes.

1. Introduction

A regulation that involves all Member States of the EU is the Brussels I Regulation, which is a set of rules about the jurisdiction, recognition and enforcement of judgments in civil and commercial matters involving individuals resident in different Member States of the European Union and the European Free Trade Association (EFTA). The EU Regulation (EC) 44/2001, as the regulation is officially called, was created by the Council of the European Union and came into force in March 2002. However, the Court of Justice of the European Union (CJEU) stated that Article 23 regarding application of jurisdiction was too concise and therefore suggested a recast. Hence, a recast was created by both the Council and the European Parliament. It was implemented in 2015 and is officially called Regulation (EU) No 1215/2012. The main difference between the old and new regulation is that in the recast the rules of Brussels I were extended to defendants not domiciled in a Member State of the EU.

Although this recast was supposed to solve the shortcomings of the 44/2001 regulation, Danov [1] states that the application of this recast has been largely overlooked by both policymakers and literature. Since the regulation is still new and opinions differ as to its usefulness, more insight in its usage would be of great value for legal professionals and the EU. This research is complicated because there are no agreements regarding the referencing of the regulation or the recast in case decisions. Therefore these cases are hard to find manually. The use of automated tools could help.

This paper describes research on text analysis of cases involving the Brussels I Regulation and will discuss to what extent it is possible to design a supervised

[1] Corresponding author, PO Box 1030, 1000 BA Amsterdam, Netherlands, winkels@uva.nl

classification system that uses judgments of civil law cases from the Dutch portal to distinguish:

1. cases about Brussels I Regulation from all the other civil law cases;
2. cases from the Brussels I Regulation Recast from the Brussels I Regulation.

Answers to these questions will indicate whether or not a system could be used to reliably distinguish cases involving the Regulation or Recast after which further (manual) research is possible. The method of binary classification could then be extended to other Member States or even to other regulations that were changed. The system could then be used by experts at the start of a new law to assess its effects and impact. It is expected that because the second classification question has to deal with a smaller set of data, those results will score lower on accuracy than the results of the first classification problem.

The rest of this paper is organized as follows: We will first give a short overview of text classification in the legal field and describe an earlier attempt at automatic classification of civil law cases. Next we will discuss our research method and describe the two classifiers we built and evaluated. We will end with a discussion, conclusions and future work.

2. Text Classification

Text classification is a problem that has been studied in many domains, including the legal domain. Bruninghaus and Ashley [1] explain this demand by the desire of attorneys to find the most relevant cases and argue that this information need caused the wide interest of classification in the legal domain. De Maat e.a. [3] for example describe a study about the classification of legal sentences and the comparison of machine learning techniques against knowledge based classification. Goncalves and Quaresma [4] applied multiple algorithms to European legal texts and stated that legal texts are very suitable for text classification because of the unstructured format of the data. Bag-of-words method was used, but also part of speech tagging and lemmatisation were applied. They note the shortcomings of the bag-of-words method - the method being too simplistic to obtain good results - which was also mentioned by [1] and [3]. Above that, the researchers state that the legal language has a unique style and that the vocabulary and word-distributions differ from 'regular' English. All authors also point at the need for a pre-tagged training set and the difficulty of obtaining one. It is tedious and hard work and legal experts are busy and expensive.

Besides picking the right algorithms, proper feature selection is of great importance. Not only is it necessary to make large problems computationally efficient, but it can improve the accuracy substantially [5]. This increase of accuracy could also mean that less data is needed to obtain good results, which is a big advantage for a system.

2.1 An earlier attempt

Zheng [6] made an analysis of a data set obtained from the Dutch portal rechtspraak.nl for cases until October 2016 and used the MAchine Learning for LanguagE Toolkit (Mallet). The Dutch portal for case law contains a small, but growing part of all judicial decisions in the Netherlands. Case citations in these decisions are sometimes explicitly marked in metadata (e.g. the first instance case); references to legislation only the main one(s) in recent cases. The texts are available in an XML format, basically divided in paragraphs, with a few metadata elements. The court decisions do not contain inline, explicit, machine readable links to cited legislation or other cases. So even when the metadata contain such references, we do not know in which paragraph the case or article was cited, nor how often.

First, an indication of the field of law for our purpose was made: 'civil law' is the most common in both the old (77%) and new (73%) regulation. Topic modelling was used to generate multiple topics and then multiple classifiers were trained and tested. For the old regulation an accuracy of 0.64 was obtained, and for the recast an accuracy of 0.78.

There are several differences between this earlier work and the research reported in this paper. First of all, we will not use topic modelling, but we will see the judgments as a bag-of-words, where pre-processing should be executed to decrease the number of features to improve efficiency and accuracy. Secondly, whereas Zheng retracted more than 2 million cases (also unpublished) about the entire field of law up to October 2016, we work on 15,000 published civil law cases until May 2017.

2.2 Tools

In this research, the following tools were used:

- KNIME: an open source platform focusing on data mining, manipulation, visualization and prediction. With its easy user interface, many machine learning applications can be used by building a workflow with different building blocks.[2] An example of the workflows built for this research can be seen below in **Figure 1**.
- MongoDB: an open source and free tool, focusing on storing data in JSON-like documents. It is possible to handle large amounts of data, to change data later on and to retrieve specific sets of data based on specific requirements. For example: retrieve all documents that are classified positively and contain 3 keywords. Each data point is stored as an object, where multiple instances can be added with different sizes.[3]

3. Classification Question 1

From rechtspraak.nl 15,000 cases were retrieved, starting with the newest from May 2017 and working 'down'. From these cases the XML was obtained, including all sorts of tags. For each case a new object was made in MongoDB with the XML as instance and the title as meta-data which had to be unique in the database. The tags were then stripped, the type of document was changed into txt-files, and these were added as new instances to each case in MongoDB. Next, these txt-instances were checked for the appearance of keywords referring to the Brussel I regulations. A legal expert, who also helped interpreting the results, made a list of words in multiple languages that indicate both the 44/2001 and 1215/2015 regulations. In Dutch the lists are as follows:

Brussel I: `Brussels I Regulation: EEX-Vo ; EG-Executieverordening; EEX-Verordening; Brussel I-Verordening; Brussel I; 44/2001`

[2] KNIME.COM AG. KNIME Open for Innovation. 2017. http://www.knime.org.

[3] MongoDB.COM AG. What is MongoDB. 2017. https://www.mongodb.com/.

Brussel I recast: `Brussels I Regulation recast: EEX-Vo II; Brussel Ibis; Brussel I-bis; EU-executieverordening; Brussel I bis-Verordening; EEXVerordening II; Brus-sel 1 bis-Vo; Brussel 1 bis; herschikte EEX-Vo; 1215/2012`

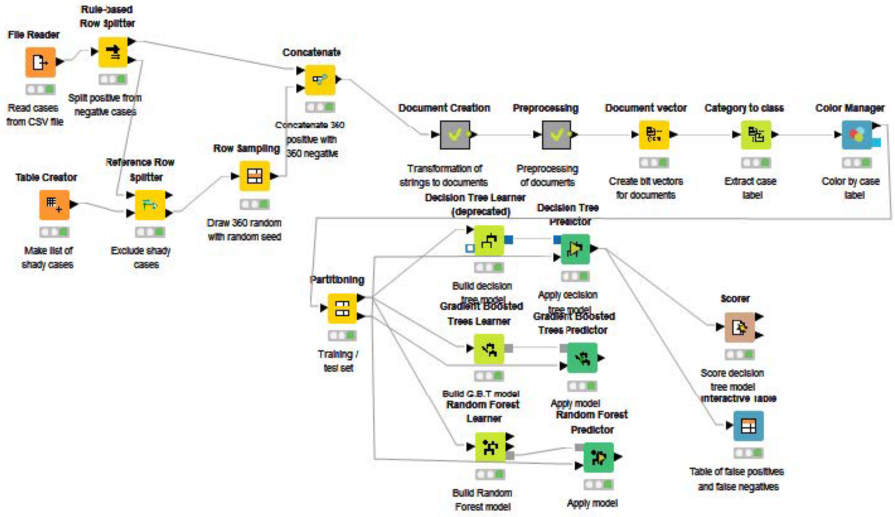

File Reader
Read cases from CSV file
Rule-based Row Splitter
Split positive from negative cases
Concatenate
Concatenate 360 positive with 360 negative
Table Creator
Make list of shady cases
Reference Row Splitter
Exclude shady cases
Row Sampling
Draw 360 random with random seed
Document Creation
Transformation of strings to documents
Preprocessing
Preprocessing of documents
Document vector
Create bit vectors for documents
Category to class
Extract case label
Color Manager
Color by case label
Decision Tree Learner (deprecated)
Build decision tree model
Decision Tree Predictor
Apply decision tree model
Partitioning
Training / test set
Gradient Boosted Trees Learner
Build G.B.T model
Gradient Boosted Trees Predictor
Apply model
Random Forest Learner
Build Random Forest model
Random Forest Predictor
Apply model
Scorer
Score decision tree model
Interactive Table
Table of false positives and false negatives

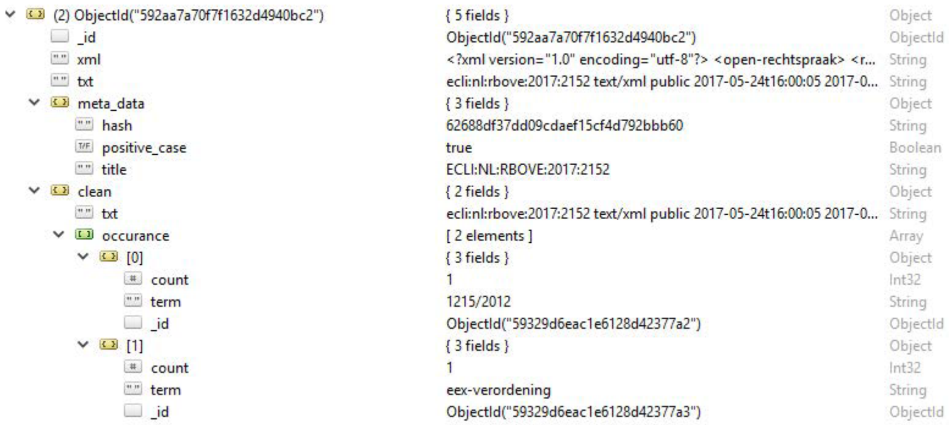
Figure 1: KNIME workflow used for the classification problems

The lists were combined and changed into one regular expression and a new instance ‘clean’ was created. In this instance the txt-file without the keywords was inserted. All the terms that matched the regular expression were listed as new instances and a new meta-data instance was set to ‘true’ if the case did contain a keyword, ‘false’ if it did not. See the example of **Figure 2** how all the instances are related. This way we created a labelled set of cases we can use for training and testing the classifiers.

It is important to understand that the labels ‘true’ and ‘false’ are used as ‘golden standard’. However, this does not mean that there are no cases involving the regulation that are labelled false. Since we are researching the reliability of such a classification system, the incorrectly classified cases are important as well. As stated in the introduction, there are no agreements regarding referencing this regulation in case decisions, so the number of cases involving the regulation is expected to be larger than the number of ‘true’ cases.

Once this was done, an analysis of the true and false cases could be made. From the 15,000 cases, there were 360 cases classified as ‘true’, and 14,640 classified as ‘false’ using the keywords. Most positive cases contained between 1-3 keywords, but two even contained 6 keywords, namely ECLI:NL:GHSHE:2017:1873 and ECLI:NL:GHSHE:2017:1874 and these two are related. The frequency distribution of the keywords over the documents can be seen in **Table 1**.

From all these instances, a CSV-file of 15,000 rows was created with the columns ‘title’, ‘document’ and ‘label’ containing the title of the case, the txt-file of judgment without keywords and the classification true or false respectively. This CSV-file was then ready to be handled by KNIME.



```
(2) ObjectId("592aa7a70f7f1632d4940bc2")   { 5 fields }                                                           Object
  _id                                      ObjectId("592aa7a70f7f1632d4940bc2")                                   ObjectId
  xml                                      <?xml version="1.0" encoding="utf-8"?> <open-rechtspraak> <r...        String
  txt                                      ecli:nl:rbove:2017:2152 text/xml public 2017-05-24t16:00:05 2017-0...  String
  meta_data                                { 3 fields }                                                           Object
    hash                                   62688df37dd09cdaef15cf4d792bbb60                                       String
    positive_case                          true                                                                   Boolean
    title                                  ECLI:NL:RBOVE:2017:2152                                                String
  clean                                    { 2 fields }                                                           Object
    txt                                    ecli:nl:rbove:2017:2152 text/xml public 2017-05-24t16:00:05 2017-0...  String
    occurance                              [ 2 elements ]                                                         Array
      [0]                                  { 3 fields }                                                           Object
        count                              1                                                                      Int32
        term                               1215/2012                                                              String
        _id                                ObjectId("59329d6eac1e6128d42377a2")                                   ObjectId
      [1]                                  { 3 fields }                                                           Object
        count                              1                                                                      Int32
        term                               eex-verordening                                                        String
        _id                                ObjectId("59329d6eac1e6128d42377a3")                                   ObjectId
```

Figure 2: Example of positive case in MongoDB with all instances.

Table 1: The number of documents each keyword appears in.

Keyword (in Dutch)	Frequency
1215/2012	103
44/2001	72
Brussel I	134
Brussel I bis	-
Brussel I Bis-Verordening[4]	8
Brussel I bis-Vo	-
Brussel Ibis	12
Brussel I-bis	3
Brussel I-Verordening	7
EEX-Verordening	123
EEX-Verordening II	7
EEX-Vo	71
EEX-Vo II	-
EG-Executieverordening	1
EU-Executieverordening	1
Herschikte EEX-Vo	23

3.1 Pre-processing of data

To be able to use algorithms on the data, a binary vector of unique terms for each case was needed. However, since the texts of the cases were sometimes very extensive and the number of positive examples scarce, proper pre-processing was important:

1. For each document delete:
   - Terms consisting of the characters: !#$%()*+,./:;¡¿=?@ˆ ‘— []
   - Terms consisting only of numbers
   - Terms consisting of less than 4 characters
   - Terms that occur in a stop list or occur in more than 95% of the documents

[4] ‘Verordening’ = Regulation; ‘executieverordening’ = implementing regulation.

- Terms that occur in less than 1% of the documents

2. For each term in each document:
   - Convert all characters to lowercase
   - Use Snowball Stemmer for Dutch language[5]

This resulted in around 6,000 unique terms that were then used as features. Of these 6,000 terms a few stood out, because of their similarity with the keywords. These terms were further investigated and also analysed by our domain expert. He concluded that most of these keywords have no relation to Brussel I, so to be certain not to train the model on wrongly labelled cases, the 34 cases containing one of these 'grey keywords' were excluded from the total set before selecting the test- and training data.

3.2 Experimental Setup

To create a baseline of 50%, 360 negative examples were drawn randomly next to the 360 positive examples. To enlarge reliability of this random sample, this was done with 10 different random seeds. Each experiment of 720 cases was then split in 504 cases to train on (70%) and 216 cases to test the classification on (30%).

Since earlier experiments already showed the poor results of the algorithms naive Bayes (accuracy of 0.51) and k-nearest neighbour (accuracy of 0.77), these were excluded in further experiments. We noted that algorithms based on trees resulted in the best accuracy, so we will use decision trees, gradient boosting trees and random forest.

Decision trees is a tree-structured algorithm, where each internal node presents a test on an attribute, each branch corresponds to an attribute value and each leaf node represents a class label. Decision trees can deal with noisy data and function well with disjunctive hypotheses [7]. It does not have any requirements about the distribution of the data (for example Naive Bayes requires independent variables), since it is a non-parametric technique.

Gradient Boosting trees is an algorithm that keeps improving its model by calculating the error and fitting new Decision Trees to the corresponding cost function, and by doing so increasing its complexity. Lawrence e.a. [7] state that in most cases it outperforms decision trees or at least performs equally. It is said to deal with overfitting better than decision trees.

Random forest is a machine learning algorithm that again uses decision trees, by learning multiple decision trees simultaneously. It then chooses the most common label of all the models. This has the advantage of decreasing the overfitting problem that decision trees tend to have. But Prasad e.a. [8] state the disadvantages of time and computational resources and the 'black-box' characteristic.

In the pre-processing all the cases were labelled true or false based on the occurrence of a few keywords. The research question is to what extent the classifier can classify cases based on their texts without these keywords. This can be measured by using the relative number of correctly classified cases, also known as the accuracy. Precision is the number of correctly classified positive examples divided by the number of examples labelled by the system as positive and recall is the number of correctly classified positive examples divided by the number of positive examples in the data. The F1-measure is the harmonic mean of both precision and recall. Since this study is mainly interested in cases that involve the regulation (positive cases), but are not

[5] http://snowball.tartarus.org/algorithms/dutch/stemmer.html

classified as true because of a lack of keywords or lack of agreement on referencing, recall is in this case more important than precision. A measure that weighs recall higher than precision is the F2-measure and for each experiment this value was calculated, its formula is as follows:

$$F_2 = 5 * \frac{precision * recall}{4 * precision + recall}$$

3.3 Results

Table 2 presents an overview of the various measures for the average over 10 samples for the three methods. In each sample we drew 360 negative cases together with all 360 positive ones. Random forest scores best on all performance measures. All methods do much better than the baseline of 50%.

Table 2: Overview of average of 10 samples of multiple evaluation methods using different algorithms. + means case involving the regulation and - means not.

Sample	Recall-	Recall+	Precision-	Precision+	F1-	F1+	F2-	F2+	Accuracy
Decision trees	0.91	0.90	0.91	0.91	0.91	0.90	0.85	0.88	0.91
Gradient boosting	0.90	0.90	0.91	0.90	0.90	0.90	0.90	0.90	0.91
Random forest	0.94	0.95	0.95	0.94	0,94	0.95	0.95	0.94	0.95

4. Classification Question 2

For the second classifier the goal was to reliably distinguish old cases from new cases. So all the negative examples were excluded and only the 360 positive examples were used. The keywords that were used for the first classifier, were used again, but this time they were not combined. New instances were then created in MongoDB, indicating whether cases were labelled 'old' or 'new' (or both). The labelling resulted in 310 new cases, 124 old cases, of which 74 were labelled both new and old. In Table 2 the frequency of each keyword can be found for both old and new cases.

Table 3: The number of documents each keyword appears in, divided in old cases (44/2001) and new cases (1215/2015) where overlap is possible.

Keyword	Frequency 'old'	Frequency 'new'
1215/2012	65	103
44/2001	72	18
Brussel I	134	18
Brussel I bis	-	-
Brussel I Bis-Verordening[6]	2	8
Brussel I bis-Vo	-	-
Brussel Ibis	6	12

[6] 'Verordening' = Regulation; 'executieverordening' = implementing regulation.

Brussel I-bis	2	3
Brussel I-Verordening	7	1
EEX-Verordening	123	37
EEX-Verordening II	2	7
EEX-Vo	71	29
EEX-Vo II	-	-
EG-Executieverordening	1	-
EU-Executieverordening	1	1
Herschikte EEX-Vo	12	23

4.1 Experimental Setup

The pre-processing steps were the same as with the first experiment. This classification problem was split into two parts: old against not old and new against not new. By doing so, the classification remained binary, just like in the first experiment. In the first classification problem there were 50% positive examples and 50% negative examples and the baseline could be set to 50. In this second classification problem that is different. For the old cases the baseline is 310/360 = 86% and for the new cases (360-124)/360 = 66%.

Again, the data was split into a train and test set using a 70/30 ratio. Because of the small number of new cases, we also used a 80/20 ratio for that part. The same machine learning algorithms were used as in the first classifier: decision trees, gradient boosting trees and random forest. For the algorithm gradient boosting trees, this time also multiple settings were performed for the number of maximum tree-depth. In the first classifier this hardly changed the accuracy, but now it did, as can be seen below.

4.2 Results

From **Table 4** below it can be concluded that the baseline of 86% for classifying old cases was not reached by any machine learning algorithm. The highest accuracy was obtained by using gradient boosting trees with a threshold of maximum tree-depth set on 4 (0.85). Also when looking at the F1-values this algorithm outperforms the rest.

Table 4: Results of classifying old cases (+) against not old cases (-).

Sample	**Recall-**	**Recall+**	**Precision-**	**Precision+**	**F1-**	**F1+**	**F2-**	**F2+**	**Accuracy**
Decision trees	0.27	0.98	0.75	0.84	0.40	0.90	0.31	0.95	0.84
Gradient boosting max 10	0.36	0.90	0.47	0.84	0.41	0.87	0.38	0.89	0.79
Gradient boosting max 4	0.27	1	1	0.85	0.43	0.92	0.32	0.97	0.85
Random forest	0.91	0.98	0.50	0.81	0.15	0.89	0.11	0.94	0.80

From **Table 5** and **Table 6** it can be concluded that for classifying new cases, the baseline of 66% is met by all algorithms. According to the highest accuracy, again gradient boosting trees with a threshold of maximum tree-depth set to 4 was best. Only for random forest the accuracy was lower when using 80/20 ratio instead of 70/30.

Although random forest (80/20) obtained the highest re-call for the negative examples, it got the lowest recall for the positive examples. When looking at the F1-values (80/20) it can be seen that gradient boosting with threshold 4 scores best on both positive and negative examples. This also holds for the F2-values and as stated before, for the accuracy.

Table 5: Results of classifying new cases (+) against not new cases (-) with 70/30 ratio for train-test set.

Sample	Recall-	Recall+	Precision-	Precision+	F1-	F1+	F2-	F2+	Accuracy
Decision trees	0.83	0.56	0.73	0.69	0.77	0.62	0.58	0.81	0.72
Gradient boosting max 10	0.80	0.62	0.75	0.68	0.77	0.65	0.63	0.79	0.73
Gradient boosting max 4	0.91	0.58	0.75	0.81	0.82	0.68	0.61	0.87	0.77
Random forest	0.91	0.44	0.70	0.77	0.79	0.57	0.49	0.86	0.72

Table 6: Results of classifying new cases (+) against not new cases (-) with 80/20 ratio for train-test set.

Sample	Recall-	Recall+	Precision-	Precision+	F1-	F1+	F2-	F2+	Accuracy
Decision trees	0.79	0.67	0.77	0.69	0.78	0.68	0.67	0.78	0.74
Gradient boosting max 4	0.86	0.67	0.78	0.77	0.82	0.71	0.69	0.84	0.78
Random forest	0.88	0.40	0.67	0.70	0.76	0.51	0.44	0.83	0.68

5. Conclusions and Discussion

The objective of the research was to answer the following question: To what extent is it possible to design a supervised classification system that uses judgments of civil law cases from rechtspraak.nl to distinguish:

1. cases about Brussels I Regulation from all the other civil law cases?
2. cases from the Brussels I Regulation Recast from the Brussels I Regulation?

Classifying Brussels I Regulation cases from other civil law cases can be done with an accuracy of 0.95, but with computational limitations. Classifying Brussels I Regulation from the Brussels I Regulation Recast is harder, due to the small amount of data.

With only 124 cases referring to the 1215/2015 regulation, the second classifier did not perform well on accuracy. The number of false positives and false negatives was high and further research is needed to find out what these incorrectly classified cases say about the data. Are they counter examples due to too few data or was the initial labelling incorrect?

Because of limitations in computer power, we decided to use only 360 negative cases in the first classifier. However, choosing 360 out of 14,640 can be done in many ways and even though this was done 10 different times this may have influenced the results.

To draw conclusions for the entire European Union, it is necessary to expend the research to other languages. We have keywords for several languages, but we may need access to legal experts from these jurisdictions to help analysing results. Extending research to other countries will also increase the number of positive cases.

Another option is using unpublished cases from courts, but we will have to see whether their format is similar to the cases from the official portal.

The performance of the classification systems was evaluated against the original labelling of the cases based on keyword matching. If the original labelling was not correct, this will of course influence the performance evaluation. We did some analysis of the original labelling and removed some cases which contained 'grey area' keywords, but a thorough analysis by human experts would be better. However, given the nature of the work and the scarcity of experts, this was not feasible at present. Our approach has the benefit of little human effort, but the disadvantage of possible mistakes in classification. Since the final analysis of the impact of legal change will be done by human experts anyway, we do not see this as a major problem.

Acknowledgements. Part of this research is co-funded by the Civil Justice Programme of the European Union in the Brussel I project under grant JUST/2014/JCOO/AG/CIVI/7754. We would like to thank our domain expert Michiel van Rooijen.

6. References

[1] Danov, M. "The Brussels I Regulation: Cross-Border Collective Redress Proceedings and Judgments". In: *Journal of Private International Law* 6.2 (2017), pp. 359–393.

[2] Bruninghaus, S. and Ashley, K.D. "Toward Adding Knowledge to Learning Algorithms for Indexing Legal Cases". In: *Proceedings of the 7th international conference on Artificial intelligence and law*. ACM, 1999, pp. 9–17.

[3] Maat, E. de, Krabben, K. and Winkels, R. "Machine Learning versus Knowledge Based Classification of Legal Texts". In: JURIX (2010), IOS press, pp. 87–96.

[4] Goncalves, T. and Quaresma, P. "Is linguistic information relevant for the classification of legal texts?" In: *Proceedings of the 10th international conference on Artificial intelligence and law*. ACM, 2005, pp. 168–176.

[5] Forman, G. "An extensive empirical study of feature selection metrics for text classification". In: *Journal of machine learning research* 3.Mar (2003), pp. 1289–1305.

[6] Zheng, K.H.. *Brussel I project*. Tech. rep. University of Amsterdam, 2016.

[7] Lawrence, R., Bunn, A., Powell, S. & Zambon, M. "Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis". In: *Remote sensing of environment 90:3* (2004), pp. 331–336.

[8] Prasad, A.M., Iverson, L.R., and Liaw, A. "Newer classification and regression tree techniques: bagging and random forests for ecological prediction". In: *Ecosystems 9:2* (2006), pp. 181–199.

Legal Knowledge and Information Systems
M. Palmirani (Ed.)

doi:10.3233/978-1-61499-935-5-131

Towards the Assessment of Gold-Standard Alignments Between Legal Thesauri

Armando STELLATO[a], Andrea TURBATI[a],
Manuel FIORELLI[a], Tiziano LORENZETTI[a],
Peter SCHMITZ[b], Enrico FRANCESCONI[b,c],
Najeh HAJLAOUI[b], Brahim BATOUCHE[b]
[a] *University of Rome Tor Vergata, Department of Enterprise Engineering, Rome, Italy*
[b] *Publications Office of the European Union, Luxembourg*
[c] *Institute of Legal Information Theory and Techniques of CNR (ITTIG-CNR), Italy*

Abstract. In this paper we report on the experience gathered in producing two gold-standard alignment datasets between the European Union thesaurus EuroVoc and two other notable resources adopted in legal environments: the thesaurus of the Italian Senate TESEO and the IATE European terminological resource. The realization of these two resources has been performed in the context of the PMKI project, an European Commission action aiming at creating a Public Multilingual Knowledge management Infrastructure to support e-commerce solutions in a multilingual environment. As of the numerous lexical and terminological resources involved in this project, ontology and thesaurus alignment and, as a consequence, the evaluation of automatically generated alignments, play a pivotal role for the success of the project.

Keywords. legal thesauri, gold-standard alignments, semantic interoperability.

1. Introduction

The Semantic Web [1] has offered a powerful stack of standard languages and protocols for modeling, sharing and reuse of knowledge on the Web. However, the advantages brought by metadata standards and infrastructures for shareability of information cannot avoid (but can support) the reconciliation work needed on the information content. Different, domain-overlapping, redundant to different extents, ontologies, thesauri, vocabularies, datasets etc. are expected to emerge on the Web in order to satisfy specific needs and exigencies and, at different points in time, are expected as well to be – somehow – "reconciled" out of their heterogeneities, for interoperability's sake.

This "reconciliation" takes the form of alignments, that is, sets of correspondences between the different entities that populate lexical and semantic resources on the Web. The expression "ontology alignment" is often used in a broader sense than the one that the first word of the term would suggest. "Ontology" is in this case a synecdoche for ontologies, thesauri, lexicons and any sorts of knowledge resources modeled according to core knowledge modeling languages for the Semantic Web, which shared and made available on the Web itself. The expression ontology alignment thus defines the task of discovering and assessing alignments between ontologies and other data models of the RDF family; alternative expressions are *ontology mapping* or *ontology matching* (as the produced alignments are also referred to as *matches*). In the RDF jargon, and following the terminology adopted in the VoID metadata vocabulary [2], a set of alignments is also called a *Linkset*.

The production of alignments is an intensive and error prone task; for this reason, several approaches for automating the task have been devised [3] since the early years of the Semantic Web. An Ontology Alignment Evaluation Initiative[1] [4] is also held every year since 2004 with the intent of evaluating available tools against benchmarks consisting in well-assessed alignments between notable semantic resources (mostly ontologies, and some thesauri). The task are also divided into T-Box/Schema matching, dealing – as the name suggests – with the alignment of ontology vocabularies, i.e. word models, including class and properties, and instance matching, involving the creation of links between domain objects represented in different datasets.

In this work, we report on the experience gathered in producing two gold-standard alignment datasets between EuroVoc[2] – EU's multilingual thesaurus covering the activities of the European Union – and two other notable resources adopted in legal environments: the thesaurus of the Italian Senate TESEO[3] (TEsauro SEnato per l'Organizzazione dei documenti parlamentari) and IATE[4] (InterActive Terminology for Europe), the EU's multilingual terms base. The realization of these resources has been performed in the context of the PMKI[5] project, a European Commission action launched to promote the Digital Single Market in the European Union (EU) by creating a Public Multilingual Knowledge management Infrastructure. PMKI aims at sharing maintainable and sustainable public multilingual knowledge systems making them interoperable in order to be reused by different services that support the creation of a Multilingual Digital Single Market; proposing a common data model for the representation of multilingual terminologies based on existing recognised standards and developing semantic capabilities to enable the automated alignment of data from different sources in order to further increase their interoperability. Such infrastructure will consist in a set of tools able to create interoperability between multilingual classification systems (like thesauri) and other language resources, so that they can be easily accessible through a Web dissemination platform and reusable by small and medium-sized enterprises (SMEs), as well as by public administrations. In this paper, the standards used to represent aligned resources and a methodology for the production of gold-standard alignments between the language resources mentioned above are presented.

2. Motivation

As far as producing mappings is an intensive and error prone task, it is also a very difficult one which cannot be easily automated. If alignments could be logically inferred, there would be processes generating them automatically with guaranteed precision. In fact, the necessity for alignments comes into play when there is no a-priori semantic agreement between two resources. The identification of proper alignments is thus a discovery process (some approaches [5] treat it as an Information Retrieval problem indeed) based on the analysis of the content. This content is mostly, on a first step, pure textual information, in order to gather the first anchors between the two resources, as common language expressions are indeed the last stand for creating hypothesis about alignments [6, 7], which is later followed by inference mechanisms based on the created hypotheses. This kind of analysis obviously benefits from background knowledge –

[1] http://oaei.ontologymatching.org/

[2] http://eurovoc.europa.eu/

[3] http://www.senato.it/3235?testo_generico=745

[4] http://iate.europa.eu

[5] https://ec.europa.eu/isa2/actions/overcoming-language-barriers

owned by the mapping agent – that goes beyond the one available in both datasets and from the capability to process natural language content.

As of the state of the art, the importance and main contribution of automatic alignment consists in providing in a short time a seed that would be hard to produce manually, in order to support the subsequent refinement performed by domain experts. Even though meant to be completed by means of human supervision, the quality of the automatically produced alignments is important: besides the trivial aspect that a good automatic process shortens the following validation, it is also true that, especially in alignments of large corpora, an unprecise seed will in any case provide a bias on the validator: a single "almost perfectly matched" concept would drift the attention of the validator away from some best representative that is still laying somewhere in the target dataset. Even more important, an incomplete mapping between two datasets (i.e. missing some important matches) can be validated in the precision of the produced matches, but missing matches would be no less hard to be discovered than in a point-blank situation. For this reason, even though precision is important, recall (as in every Information Retrieval process) is fundamental, as users may easily discard wrong results but their trust in the system depends on the feeling that no important information is left unretrieved.

A proper evaluation of alignment processes demands for gold-standard alignments, guaranteeing maximum precision (quality-produced matches) and recall (completeness of the alignment). Another possible exploitation of such resources is their use as training sets for mapping algorithms based on learning techniques. Also, to avoid the negative effects of in-vitro experimentation, alignments between real datasets should be produced.

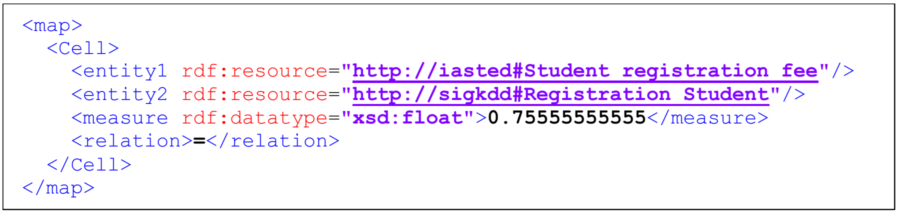
With these requirements in mind and with the legal domain as an important topic for the project that has been the context for this work; we chose to create two alignments for subsets of three important resources of the European Union: EuroVoc, TESEO and IATE. In order to create reliable gold-standard alignments, we selected a range of circa 300 concepts per each resource, so that the task of manual alignment could be performed entirely by humans without any bias nor errors induced by machine processing.

3. Standards for the Representation of Semantic Resources and their Alignments

PMKI considers resources that are modeled and published according to Semantic Web standards, thus adopting RDF (and the various vocabularies for creating ontologies, thesauri, etc.) and Linked Open Data best practices for publication. Different actions of the project also aim at providing services (transformation, hosting etc.) for bringing non-RDF resources into the PMKI platform, thus leveling all of them to the desired standards.

Within this context, the kind of resources that have been considered (at least for a first bootstrap of the project) are multilingual language resources, including thesauri and lexicons. W3C offers two core modeling vocabularies for dealing with these kind of resources: for thesauri there is the W3C recommendation SKOS [8] (and its extension SKOS-XL [9], for modeling reified labels) while a community group under the W3C umbrella has recently developed the OntoLex-Lemon (https://www.w3.org/2016/05/ontolex/) suite of vocabularies for modeling lexicons.

SKOS is a vocabulary defined in OWL introducing a set of terms characterizing thesauri: skos:Concepts are described as representing ideas or notions, units of thought. However, as the reference manual clarifies, what constitutes a unit of thought is subjective, and this definition is meant to be suggestive, rather than restrictive. In a few words, while OWL classes provided a formal classification mechanism for describing entities, adopting a first-level logic perspective made of objects and predicated over them,



```
<map>
  <Cell>
    <entity1 rdf:resource="http://iasted#Student_registration_fee"/>
    <entity2 rdf:resource="http://sigkdd#Registration_Student"/>
    <measure rdf:datatype="xsd:float">0.75555555555</measure>
    <relation>=</relation>
  </Cell>
</map>
```

Figure 1. an excerpt of the RDF/XML code for an alignment expressed for the Alignment API

SKOS concepts are meant to provide a simplified view, with shallow semantics, where entities can be generally described with values. Concepts are, in this sense, all objects with respect to an OWL perspective (they are indeed instances of the skos:Concept class), whether they are merely conceptual entities or concrete objects of a domain. A more detailed description of differences between SKOS and OWL (and what to model as a thesaurus or as an ontology) is provided in [10].

Another important aspect of SKOS lies in its terminological properties. Preferred labels (skos:prefLabel) describe the lexical expressions that are most recurring for referring to a certain concept. Alternative labels (skos:altLabel) provide alternative expressions, less common than the preferred ones. Hidden labels (skos:hiddenLabel) cover common misspells and other expressions that should not be explicitly shown while being at the same adopted for technical aspects such as indexing and retrieval. SKOS also provides several properties for expressing notes about the elements of a thesaurus. These notes can be in-domain or extra-domain, representing respectively the description (e.g. skos:definition) of the object of the domain represented by the SKOS concept or notes taken by the editors (e.g. skos:editorialNote) that concern the concept as an element of the thesaurus construct and not as its denoted domain object.

A third important aspect of SKOS lies in its mapping vocabulary: a set of properties describing relationships of similarity or, at least, connection, between concepts from two different thesauri. skos:exactMatch provides a shallow semantics (i.e. no logical entailment) approach to identity, expressing the fact that two SKOS concepts denote the same object of the world. skos:closeMatch implies a strong closeness between two SKOS concepts even though they are not considered to be exactly the same thing. skos:broad/narrowMatch express a “more specific”/“more generic” relationship between two concepts from different thesauri while skos:related connects two concepts denoting two things that, while being clearly different, are somehow related to each other.

The vocabulary that has been used for the alignments is the one created and adopted for the Alignment API [11], and it is a base for the more elaborated EDOAL mapping language. In the Alignment API ontology, mappings are reified into resources (see figure 1), in order to be decorated with further metadata: for instance, besides the type of relation (expressed by the property :relation) the vocabulary foresees a property :measure, representing the confidence of the mapping. However, other additional information can then be added as well (e.g. rdfs:comments). The vocabulary is also extensible, as new relations can be added in custom extensions of the vocabulary.

For instance, the basic vocabulary only foresees an equivalence relation, probably due to the fact that it is not expected for a machine to be so precise in distinguishing among subtle variations of a match, thus postponing this specification to later refinement by means of human supervision. However, in the context of the maintenance of the EuroVoc thesaurus, the Publications Office of the European Union adopted an extension of the alignment ontology where a property semantically close to skos:closeMatch has been added to the range of available ones, thus making the porting to SKOS mapping relations a simple 1-1 transformation.

4. Alignment Completeness and Minimality

As outlined in section 2, an important aspect when considering an alignment effort is, besides the (trivially clear, yet not easy to achieve) precision of the outcome, its completeness. Most alignment vocabularies consider at least three kind of relations: equivalence and the two inverse more specific/more general relations, plus usually a relation for non-matchable entities. In an alignment effort between two datasets D_A and D_B it would be theoretically possible to assert the cartesian product of $|D_A| \times |D_B|$ possible alignments. However, there are some aspects that, intuitively, should be taken into consideration:

- *Direction of the alignment*. An alignment has a direction, from a "source" dataset to a "target" dataset
- *"Nobleness" of the relationship*. If a concept $c_A \in D_A$ is aligned with a relation of a certain "nobleness" to a concept $c_B \in D_B$, then there is no need to establish other "*less noble*" relationships between c_A and any other concept in D_B. "nobleness" is a total order over the available relations in a given mapping vocabulary. E.g. in the SKOS mapping properties, the following order would hold: exactMatch, closeMatch, narrowMatch, broadMatch, related. We consider then two relevant sets of relations: *equivalence* relations and *tolerance* relations. *Equivalence* relations are reflexive, symmetric and transitive. Tolerance relations are reflexive, symmetric but not necessarily transitive. skos:exactMatch match is an example of a equivalence relation, while skos:closeMatch is a tolerance relation. Notably, in SKOS neither :exactMatch nor :closeMatch have been defined as reflexive. The absence of reflexivity is due to historical reasons (the SKOS model is older than OWL2, where reflexivity has been introduced).
- Note that more relationships of the same type (if allowed by the relation itself) would be admitted. For instance, a concept $c_A \in D_A$ could be declared to be in a skos:narrowMatch relation with two concepts $c_{1B,}c_{2B} \in D_B$. The rationale for this double assignment is that being no better (more noble) match for c_A in D_B, then as many concepts from D_B, which (together) provide a best representation for c_A, are linked to it.

We can thus provide the following definitions:

Definition 1 (Linkability and Reachability). *If a concept $c_A \in D_A$ is aligned with a tolerance relation to a concept $c_B \in D_B$,(expressed as $l(c_A, r, c_B)$) then $\forall\ c_i \in D_A$: $c_i < c_A$ it holds that $c_i \in reach(c_B)$.*
In other words, in absence of an explicit relationship for c_i, c_i can still be considered linked to c_B (thus being able to "reach" c_B) through a traversal of the hierarchy in D_A until concept c_A that is perfectly matching c_B is met.

Definition 2 (link to a Dataset). *Given two datasets D_A and D_B linked through linkset L, a set of alignment relations R, a concept $c_A \in D_A$* is *said to be linked to D_B iff at least one of the two following statements is true:*

- *1.* $\exists\ l(c_A, r, c_B)$ *:* $c_A \in D_A, c_B \in D_B$, $r \in R$

- *2.* $\exists c_B \in D_B: c_A \in reach(c_B)$

In particular, given R^T the set of tolerance relations:

- c_A *is said to be perfectly linked to D_B iff* c_A *is linked to D_B and statement 1 is false if it is not possible to find an explicit link* $l(c_A, r, c_B)$ [6]
- c_A *is said to be redundantly linked to D_B iff both statements 1 and 2 are true and the asserted alignment is not a match based on a tolerance relation*

Definition 3 (properly unlinked concepts). *Given two datasets D_A and D_B linked through linkset L, a concept* $c_A \in D_A$ is *said to be properly unlinked from D_B if there is no possible assignment of alignments* $\in L$ *so that c_A can be linked to D_B.*

So, the notion of *perfectly linked* concept guarantees that no concept is simply linked as the consequence of a tree-traversal when it could have been better linked through an equivalence link. This is necessary for defining the completeness of an alignment. Conversely, the notion of redundantly linked concept supports the definition of minimal and complete alignment.

Definition 4 (Mapping completeness). *A linkset L between two datasets D_A* and *D_B is said to be complete with respect to dataset D_A iff each concept in D_A is perfectly linked to or is properly unlinked from D_B.*

Definition 5 (Minimal complete mapping). *A linkset L between two datasets D_A* and *D_B is said to be complete and minimal with respect to dataset D_A iff each concept in D_A is either perfectly and non-redundantly linked to D_B or is properly unlinked from D_B.*

5. The Procedure for Creating the Datasets Corpus

As the objective for the golden-standard corpus is to reach a number of 300 concepts per dataset, a procedure for selecting the 300 concepts without de facto performing a mapping (which we wanted to perform on a second phase, after the cut for the corpus had been done) had to be established.

1. Having the two concept trees at hand or simply, by knowing the domains covered by the two thesauri, a domain is chosen so that the selected concepts will be generally expected to be connected with it.
2. In order to generate cuts of the thesauri that can be still considered as real resources, macro portions of the hierarchy should be selected, without performing too much pruning. The user performing the cut should thus select a concept relevant for the domain (and general enough to subsume other relevant concepts) and perform a bird's eye inspection (again, the real alignment has to be performed later) to verify that roughly a non-trivial number of concepts is relevant to the domain (and thus possibly linkable to concepts in the cut of the other thesaurus). If the number of concepts in the retrieved branch is too high, some cuts will be performed on it. Conversely, if it is too low, a new branch can be selected – with the same approach – from the original thesaurus in order to increment the number of chosen units.

[6] To understand this apparent tautology, we have to distinguish links which are asserted in a concrete linkset, and links that could possibly be found. So, a link is perfect either if it is explicitly asserted (and sound) or if it is expressed by a *reach*, having established that there is no possible explicit alignment. This *having established* is a truth value that is associated to the presence of an *oracle*, representing all that is know about the relations between two given datasets. The objective of realizing such an oracle is thus the scope of this work, and these properties describe how this oracle can be used to describe alignments computed by machines.

3. Once the two cuts have been produced, a further bird's eye over their content should confirm if there is an acceptable overlap between the two selected portions of thesauri.

An important aspect to clarify, when working with alignment corpora based on real thesauri cuts, is that a perfectly sound linkset which is also minimal and complete with respect to the aligned excerpts, cannot be considered as a subset of a linkset (with same properties of soundness and completeness) between the two original thesauri. The act of cutting both source and target thesauri implies that the search for alignments will incur in local minimums, dictated by the restricted search space.

6. The experience in aligning EuroVoc and TESEO

The alignment work has been organized as follows:

- The pivot language for the alignment has been the Italian language, as TESEO is only expressed in this idiom.
- A domain expert has manually developed a first version of the linkset by analyzing all concept combinations from the two thesauri. The search space is not trivial (300x300=90000 combinations). For this reason, a quick a-priori identification of palely exact-matching concepts has been performed in order to reduce its dimension.
- A Semantic Web expert has revised the work performed by the domain expert, by both validating/rejecting/modifying the alignments produced in the first step and by performing a further search (notably reduced in size thanks to the first step and the following validation), discovering new alignments.

The following table provides some figures for the outcome of the alignment process:

Table 1. Results of the alignment

alignments produced by the domain expert	74
alignments discarded by the SW expert	18
alignments modified (type of relation) by the SW expert	3
alignments added by the SW expert	45
total amount of alignments at the end of SW expert's review	101

As it is possible to observe, the Semantic Web expert discarded a non-trivial fraction (~25%) of the original alignments, while being able to discover many more that had been overlooked by the domain expert (though he later re-confirmed these additions).

Had we more resources, a further improvement we would have added to the procedure would have consisted in doubling the first step with two domain experts. This would have allowed us to report on standard measures, such as inter-annotator agreement [12], which are a valid instrument for evaluating the easiness of the task (e.g. if humans agree on a given percentage of the results after carefully performing the job, we cannot expect a machine to perform better than them). In retrospect, considering the numerous changes and improvements that the Semantic Web expert brought during the second step, the inter-annotator agreement would have been extremely low.

This probably reflects the inherent difficulty for humans in performing the alignment task (especially if their familiarity with the domain is not matched by an equal proficiency with thesauri and computer systems) more than the specific case study. In this case, the inter-annotator agreement could probably have not been considered a reliable upper bound for machine performance.

6.1. Report on observed phenomena

Besides the direct outcome of producing the linkset resource, aligning two thesauri is also an occasion to delve into the details of the examined resources and to discover many inconsistencies and issues (because of the deep attention in identifying the nature and identity of concepts) that would have been overlooked otherwise.

The experience in linking the two (cut) thesauri EuroVoc and TESEO has presented a series of results that are indeed not surprising, if some background knowledge about the origins, invested human resources and profile of the two datasets is known a priori. We list here a few considerations emerged during the alignment work:

Rigor. Compared to TESEO, the EuroVoc thesaurus is characterized by a stricter rigor and better precision in the representation and organization of concepts. Vice versa, despite multilinguality is one of its stronger characteristics, EuroVoc loses in precision when we consider labels for languages that are not English or French (respectively, the de facto lingua franca spoken in EU and the first language used in the offices of the European Commission in Luxembourg). This is probably due to the natural interpretation gap generated by different users dealing with the conceptual content and with its translation in different languages (whereas, in a monolingual thesaurus such as TESEO, the two roles, at least for the preferred labels selected at concept creation, converge into the same user). For instance, the domain expert initially expressed a match between “aviazione militare” from EuroVoc and “aeronautica militare” from TESEO. However, despite the two terms show an apparently almost-synonymic expression in Italian, “aviazione militare” in EuroVoc revealed to be a wrong translation for “military aircraft”, whereas the best representative in EuroVoc should have been “forze aree” (“air force”).

Different practices. While SKOS is already a very permissive standard, leaving much liberty in terms of modeling choices (which is also a limit in terms of universal understandability and thus shareability of content), a further degree of freedom is given by different practices at content level that characterize thesauri since the dawn of their time. For instance, in TESEO it is possible to find many cases in which, rather than modeling a domain concept and its more specific interpretations as different formal concepts in SKOS, they are all converged into a single SKOS concept where the preferred label represents the most general interpretation in the domain and alternative labels provide its specializations. In terms of mere representation this approach might be considered a severe mistake. However, if we consider thesauri merely as tools supporting retrieval of annotated resources (usually documents), this conglomeration seems to be justifiable in all cases in which there is no particular interest in exploring the details of the domain concept (thus considering one element as enough), while it is still considered important to broaden the range of its potential lexical expressions. For instance, in TESEO the concept of “conventional weapon” is represented by several labels including “artillery” that is indeed a lexical expressions denoting only one of the possible conventional weapons. In EuroVoc, there seems to be no use (at least, not as broad as in TESEO) of this conceptual collapse, probably due to the larger resources invested in developing the resource, and in the wider scope of its application. TESEO has indeed been explicitly developed for indexing the documents of the Italian Senate, while EuroVoc represents a conceptual hub for all member states, possibly reused in different contexts other than indexing of the documents published by the EU Publications Office.

Indeed, in cases such as this one, it is logically appropriate to consider the concept's extension as the one invoked by its most general lexical representations (e.g. "conventional weapon") and, consequently, to map it as such as, implicitly, a concept extension contains the extensions of its narrower concepts. It also true that the broader/narrower relation in SKOS is not formally defined as in OWL and this extension inclusion is not guaranteed. On the other side, the described phenomenon seems to consistently appear in the case of really inclusive representation of concepts.

Differences in the Domain. Aligning two thesauri does not require to deal only with different modeling and lexical choices. The domains of the two resources might not be perfectly overlapping, especially with respect to concrete objects and the classifying concepts that include them. For instance, the domain expert simply mapped two concepts, both labeled as "esercito", as an exact match. However, by looking at the structure of EuroVoc and other translations of the term, it appears that EuroVoc's "esercito" was actually a representation of "armed forces", whilst it is "esercito di terra" that most closely represents the "army", that is what truly the "Esercito" is in Italian. Until now, it could seem another mere translation issue. However, still, if "esercito di terra" has to be interpreted in its broader sense of "ground force" then, strictly speaking, different corps may include ground forces (e.g., US' navy seals belong to the navy while being a ground force, as much as Italian "battaglione San Marco" does). Simply, there is a blurred definition for these elements, as each country has its own military organization and it is not possible to make a 1-1 transposition. For this reason, the new alignment replacing the wrong one has in any case been represented as a close match. Analogous issues happened with all job-related concepts related to workers, rights, separation of legal and physical persons and their abilities. Another troubled area lies in government aspects, as the different perspectives given by the governance of the EU and of a specific government of a member state, brought clearly diverging bias affecting even the most general concepts. However, these phenomena mostly pose a threat to the quality of the alignment, offering several pitfalls to the annotators during the alignment process, and do not represent inherent representation issues.

7. A new challenge: EuroVoc and the IATE terminology

The alignment between EuroVoc and IATE is being produced at the time of writing. We mention here the main differences in terms of knowledge representation models being adopted by this third resource, IATE. IATE is a terminology, giving more emphasis on the lexical aspect rather than the conceptual one. Despite that, lexical entries in IATE are still organized around meanings, which are however, not structured into a hierarchy like a traditional thesaurus. IATE is thus being modeled as a OntoLex-lemon lexicon. The lexical entries are described in their details (e.g. term composition) thanks to the above standard, and the meanings are represented as ontolex:LexicalConcepts, a dedicated class of OntoLex (subclassing skos:Concept) expressly thought for representing units of thought in lexicons. The presence of (lexical) concepts allowed us to keep a "concept to concept" approach to alignment representation. Conversely, the lack of a conceptual hierarchical structure forced us to adopt a different strategy for the pre-selection of a set of concepts for the initial development of the corpus to be aligned. Luckily, IATE includes an explicit concept of domain called "subject field", thus we have selected the field "Civil Law" which includes 270 concepts (so we could take them all), and we could proceed with the already experimented process of branch-selection for EuroVoc.

8. Conclusions

In this paper, we have described our guidelines for the production of golden-standard alignments with a reasonably-limited amount of resources and reported on our application of the guidelines for the development of such alignments. We hope that this work can provide both a theoretical and practical foundation for similar initiatives.

Acknowledgements

This work has been supported by the PMKI project, under the 2016.16 action of the ISA^2 Programme (https://ec.europa.eu/isa2/). ISA^2 is a EUR 131 million programme of the European Commission, supporting the modernization of public administrations in Europe through the development of eGovernment solutions.
The produced alignments will be published, when ready, on the PMKI page.

References

[1] T. Berners-Lee, J. A. Hendler and O. Lassila, "The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities," *Scientific American,* vol. 279, no. 5, pp. 34-43, 2001.

[2] K. Alexander, R. Cyganiak, M. Hausenblas and J. Zhao, "Describing Linked Datasets with the VoID Vocabulary (W3C Interest Group Note)," World Wide Web Consortium, 3 March 2011. [Online]. Available: http://www.w3.org/TR/void/. [Accessed 16 May 2012].

[3] J. Euzenat and P. Shvaiko, Ontology Matching, 2nd ed. ed., Springer-Verlag Berlin Heidelberg, 2013.

[4] Z. Dragisic, K. Eckert, J. Euzenat, D. Faria, A. Ferrara, R. Granada, V. Ivanova, E. Jimenez-Ruiz, A. O. Kempf, P. Lambrix, S. Montanelli, H. Paulheim, D. Ritze, P. Shvaiko, A. Solimando, C. Trojahn, O. Zamazal e B. Cuenca Grau, «Results of the ontology alignment evaluation initiative 2014,» in *9th International Workshop on Ontology Matching, October 20, 2014*, Riva del Garda, Trentino, Italy, 2014.

[5] P. Schmitz, E. Francesconi, N. Hajlaoui e B. S. A. Batouche, «Semantic Interoperability of Multilingual Language Resources by Automatic Mapping,» in *Electronic Government and the Information Systems Perspective (Lecture Notes in Computer Science)*, vol. 11032, E. Francesconi, A cura di, Springer, Cham, 2018, pp. 153-163.

[6] M. T. Pazienza, S. Sguera and A. Stellato, "Let's talk about our "being": A linguistic-based ontology framework for coordinating agents," *Applied Ontology, special issue on Formal Ontologies for Communicating Agents,* vol. 2, no. 3-4, pp. 305-332, 26 December 2007.

[7] M. T. Pazienza, A. Stellato and A. Turbati, "Linguistic Watermark 3.0: an RDF framework and a software library for bridging language and ontologies in the Semantic Web," in *5th Workshop on Semantic Web Applications and Perspectives (SWAP2008), Rome, Italy, December 15-17, 2008, CEUR Workshop Proceedings*, FAO-UN, Rome, Italy, 2008.

[8] World Wide Web Consortium (W3C), "SKOS Simple Knowledge Organization System Reference," World Wide Web Consortium, 18 August 2009. [Online]. Available: http://www.w3.org/TR/skos-reference/. [Accessed 22 March 2011].

[9] World Wide Web Consortium (W3C), "SKOS Simple Knowledge Organization System eXtension for Labels (SKOS-XL)," World Wide Web Consortium, 18 August 2009. [Online]. Available: http://www.w3.org/TR/skos-reference/skos-xl.html. [Accessed 22 March 2011].

[10] A. Stellato, "Dictionary, Thesaurus or Ontology? Disentangling Our Choices in the Semantic Web Jungle," *Journal of Integrative Agriculture,* vol. 11, no. 5, pp. 710-719, May 2012.

[11] J. David, J. Euzenat, F. Scharffe and C. Trojahn dos Santos, "The Alignment API 4.0," *Semantic Web Journal,* vol. 2, no. 1, pp. 3-10, 2011.

[12] R. Artstein, "Inter-annotator Agreement," in *Handbook of Linguistic Annotation*, J. Pustejovsky, Ed., Springer, Dordrecht, 2017.

Legal Knowledge and Information Systems
M. Palmirani (Ed.)

doi:10.3233/978-1-61499-935-5-141

Using Agreement Statements to Identify Majority Opinion in UKHL Case Law

Josef VALVODA [a], Oliver RAY [a] and Ken SATOH [b]
[a] *University of Bristol, Bristol, UK,*
e-mail: {jv16618,csxor}@bristol.ac.uk
[b] *National Institute of Informatics, Tokyo, Japan,*
e-mail: ksatoh@nii.ac.jp

Abstract. This paper is concerned with the task of finding majority opinion (MO) in UK House of Lords (UKHL) case law by analysing agreement statements (AS) that explicitly express the appointed judges' acceptance of each other's reasoning. We introduce a corpus of 300 UKHL cases in which the relevant AS and MO have been annotated by three legal experts; and we introduce an AI system that automatically identifies this AS and MO with a performance comparable to humans.

Keywords. Agreement Statements, Majority Opinion, UK House of Lords

1. Introduction

The court of the *UK House of Lords* (UKHL) is the former judicial arm of the British parliament's upper house, which served as the country's highest appellate court until it became the *UK Supreme Court* (UKSC) in 2009. In this court, a *majority decision* (MD) is an outcome agreed by more than half of the participating judges, while a *majority opinion* (MO) is a line of reasoning accepted by more than half as legal grounds for that decision. The distinction is crucial because an MO sets a *binding* precedent in UK law, while a non-majority view is merely *persuasive* even if it supports an MD. Thus it is common for UK law lords to discuss their opinions in draft and explicitly state any agreements with each other in their judgments. Usually this is done through formulaic phrases, that we call *agreement statements* (AS), used specifically for this purpose.

The goal of our work is to develop a computational method for detecting AS in UKHL judgments and using them to identify cases with a binding MO. This is needed because legal research tools[1] currently offer little help in this respect: since, unlike other jurisdictions, the obvious instances of dissent which are often flagged up in case digests are generally insufficient to establish the presence or absence of MO in UKHL cases. This paper takes the natural first step towards a solution by looking for unqualified (in-full) AS that suffice (per-se) to imply a definite (non-contestable) MO.

Our first contribution is to introduce a new legal corpus, called ASMO, consisting of 300 UKHL cases in which relevant AS and implied MO have been annotated by

[1] See for example westlaw.co.uk , lexisnexis.com/uk/legal , justcite.com and bailii.org/uk/cases/ukhl/

three legal experts. We derive a consensus labeling for the corpus and determine how accurately humans perform this task in practice. Our second contribution is to introduce a novel AI system that uses *machine learning* (ML) and *natural language processing* (NLP) to identify relevant AS and MO with a performance comparable to humans. We argue this is a useful first step towards the development of practical tools to help lawyers identify legal precedents in UKHL judgments, but we also demonstrate why this task is far more complex than it might first seem.

2. Background

Legal scholars have long discussed [1,2] how the UKHL tradition of publishing *seriatim* opinions of individual judges, with no accompanying statement of official consensus on the underlying reasoning, can make it hard to distinguish a binding MO from a persuasive MD, even when judges use explicit AS to express their agreements with each other.

In a speech [3] on the first anniversary of the UKSC Lady Hale stated while *"there should never be any doubt about what has been decided and why [...] This may not always be achieved even when we think that we have"*. Citing several UKHL & UKSC cases, she singled out the ongoing failure to solve this as the *"low point"* of the year.

Our aim is to approach this problem from a computational perspective by automating the task of detecting unqualified AS in UKHL judgments and using them to identify cases where they are sufficient to establish a definite MO. This can be broken down into two key tasks which are explained in the following two sub-sections.

2.1. Agreement Statements (AS)

The first challenge is to identify those sentences in which the judges actually specify their agreements with each other. Although a stock of formulaic phrases have evolved for this purpose, subtle variations in the precise English wording allow judges to express a myriad range of full or partial agreements - as illustrated in Table 1 below:

1 I fully agree with my noble and learned friend Lord Woolf that this appeal should be dismissed for the reasons he gives.

2 I agree with it, and for the reasons which he gives I, too, would dismiss the appeal.

3 I have read the speeches of my noble and learned friends, Lord Hoffmann and Lord Slynn, and for the reasons given by Lord Hoffmann, I would dismiss this appeal.

4 Therefore, like Lord Hoffmann, I see no reason in principle why, today, prerogative legislation, too, should not be subject to judicial review on ordinary principles of legality, rationality and procedural impropriety.

5 I too would allow the appeal and make the orders my noble friend, Lord Millett, proposes.

6 I have had the advantage of reading in draft the speech of my noble and learned friend, Lord Hoffmann.

7 For these reasons I would allow the appeal and direct that the case be remitted to the County Court for trial.

8 For these reasons and also for those contained in the opinions of my noble and learned friends Lord Scott of Foscote and Lord Brown, I agree with the conclusion reached by Langley J and the Court of Appeal and would dismiss the appeal.

9 I am in full agreement with the reasons expressed in the House today by my noble and learned friends.

Table 1. *Example AS, taken from our corpus, representing Full Agreement (1,2,3,8), Partial Agreement (4), Order Agreement (5), Acknowledgement (3,6), Self Agreement (7,8) and Generic Agreement (9).*

As our aim is to find incontestable MO based on *full* agreements (where one judge accepts another's reasoning in its entirety) our first priority is to distinguish these from weaker *partial* agreements (with just some aspects of the other's reasoning) or *order* agreements (with only the outcomes or orders proposed by the other). For example, in Table 1, sentence 1 is a full agreement with the reasons for the outcome, but sentence 4 is a partial agreement with just a part of the reasoning (relating to prerogative legislation), and sentence 5 is an order agreement with only the outcome and orders (but not the reasoning).

When considering such agreements a few pitfalls must be avoided. For example, sentence 2 is a full agreement where the name of the judge being agreed with is not explicitly contained in the AS, but must be inferred from the relevant pronoun. This is why we must also consider *acknowledgment* statements, like sentence 6, which contain the actual names of the judges referred to. Sentence 3 is a full agreement with one judge combined with the acknowledgement of another. This shows one sentence may contain several AS, and not all the judges mentioned are necessarily being agreed with.

It's worth pointing out the notion of full *agreement* can be more precisely viewed as an *acceptance* that the opinions of some set of judges comprise the binding reasoning of a case. If a judge believes their own reasoning is indispensable, this can be seen as *self* agreement. Although self agreement is often left implicit, judges do frequently refer to their own reasons explicitly, as shown in sentence 7 - especially when they see them as a necessary addition to the reasoning of some other judges, as in sentence 8 (which also includes a partial agreement with a judge in a lower court for good measure).

Finally, it is not uncommon for judges express a full agreement with their *learned friends*, but without saying exactly who. As shown in sentence 9, we call these *generic* agreements. While this judge is likely agreeing with at least two of his peers, we cannot be certain which ones. Could it be all the (other) judges, or just those which have posited arguments of their own, or only those whose arguments are acknowledged by this judge, or something else? But, although generic agreements are ambiguous, if the other judges are more explicit, it may still be possible to find an MO, as explained below.

2.2. Majority Opinions (MO)

The second challenge is to determine if an MO is implied by the AS. For us, this means finding a *set* of judges whose reasoning is collectively agreed with by more than half the court. To this end, we find it convenient to depict the judges of a case as nodes in a graph, using arrows to show full agreements (with loops denoting self agreement and ellipses representing generic agreements) and bold circles to show any judges forming an MO. As most cases in our corpus have 5 judges, we depict them by the letters A-E. In this way, we can use the four hypothetical cases shown in Figure 1 below to illustrate some of the key challenges involved in identifying MO from the AS. In so doing, we will explain the sort of inferences that a lawyer would make - some of which are necessarily based on a familiarity with the way that judges actually express themselves in such cases.

In example (i), a 3-of-5 judge majority {B,C,E} establishes {D} as the MO - since the former all fully agree with the latter's reasoning. But, while the MO is clear, it is worth noting we would also be justified in including D in this majority because the fact he expresses no full agreement with any other judges implies he must be relying on his own arguments. Indeed, it happens in practice that one lord writes a stand-alone *lead*

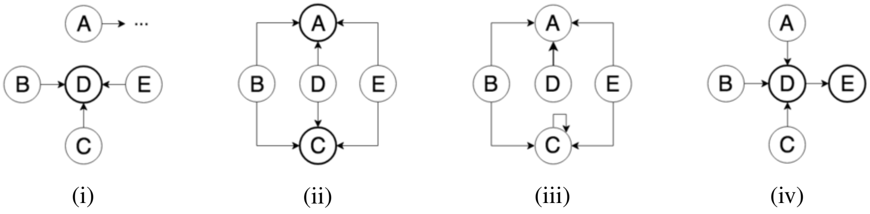


Figure 1. *Graphs showing three hypothetical cases; where letters denote judges, arrows denote full agreement, ellipses denote generic agreement, and bold circles show the judges, if any, whose reasoning forms a MO.*

judgment which all the other judges agree with. Of course, in this example, since A gives only a generic agreement, we cannot actually be sure if he shares the majority view (that the ratio of the case is contained fully in D's opinion) or if he instead believes some additional orbiter from B, C, E, or even A himself, are also indispensable. Fortunately, the MO is not affected by A's position, as D is already supported by a majority here.

In example (ii), a 3-of-5 judge majority {B,D,E} establishes {A,C} as the MO - since all the former agree with the reasoning of both the latter. While the MO is again clear, it is worth noting that neither A nor C are part of the majority electing them as the MO. For, while each certainly agrees with his own reasoning (as neither provides any other external grounds for their decision), we can't assume they agree with each other. It does happen in practice that two judges explicitly express their mutual agreement with each other, but when that occurs they will explicitly agree with each other. Even though that doesn't happen here, a majority nevertheless does still believe that the arguments in the opinions of A and B are both needed to adequately justify the decision.

Example (iii) is nearly identical to the previous one, but shows a case where no MO can be inferred from the AS. The problem is that B and E now form a minority (voting for both A and C), while C forms another minority (voting for C but not A), which means there is no majority view. Note that when a judge agrees with the reasoning of two or more judges, the agreement is with the combined reasoning as opposed to reasoning of one judge or the other. Therefore, the agreement with reasoning of judges A and C, is different from agreement with the reasoning of judge A or C. In our example (iii), B and E are not agreeing with A or C, they agree with both A and C. Crucially, this also shows the MO cannot be determined by simply finding nodes with three or more incoming edges. Example (iv) also demonstrates this point. Here a majority {A,B,C} all agree with D, who in-turn agrees with E. But, since the reasoning of D relies on E, it follows they both must be included in the MO which is therefore {D,E}. Note that, as we only consider full agreements here, A, B and C must implicitly agree with E - since if they did not then they would not be able to agree with D who clearly does. This is another illustration of how hard this task can be.

3. Manual Annotation Study

To better understand the practical significance of the challenges outlined in the previous section, we created a corpus of 300 UKHL judgments and asked three experts to identify the relevant AS and MO in each case. We then used an arbitration process to derive a consensus annotation for the entire corpus. The following subsections describe how our corpus, called ASMO[2], was constructed, annotated and arbitrated.

3.1. Creating the Corpus

At the outset, we decided a corpus of 300 cases should provide an adequate basis for reliably training and testing an AI system. But, sourcing such a large number of judgments is not trivial as it would violate the terms-of-service of the legal research tools that would ordinarily be used to obtain them. Fortunately, UKHL judgments between 1996 and 2009 are publicly available from the UK parliament website[3]. The only problem is that cases are split across multiple web pages whose HTML format differs markedly from year to year. Therefore we created a bespoke web scraper using the BeautifulSoup library[4] which downloads the body of each case, taking into account the yearly format changes, and combining the text from successive pages into a single file.

To facilitate the potential future integration of our work with prior work on rhetorical zoning, we chose to include 69 UKHL cases previously used in the HOLJ corpus of Hachey et al. [4] in our own corpus. We then randomly selected the remaining 231 cases from the 755 cases available on the UK parliament website. The resulting files were split into individual sentences, with each sentence beginning on a new line. We first used the NLTK toolkit [5] for sentence splitting, but the complexity of a typical sentence with legal citations and abbreviations led to poor results. We then used the StanfordNLP toolkit [6] to achieve much better results. Finally, any remaining errors were manually corrected. In this way we obtained a total of 134,953 sentences in the ASMO corpus.

3.2. Annotating the Corpus

We used three experts to each annotate all 300 cases with relevant AS and MO. We had two junior annotators (both reading law at university) and a senior annotator (working as a paralegal in the UK). We provided them with a set of formal guidelines and had them participate in joint training sessions which explained the two key tasks: to identify which sentences in a case contain AS representing full agreement or acknowledgement; and to identify which judges form a conclusive MO based on those AS.

Before engaging the annotators we experimented with the BRAT [7], GATE [8] and Tagtog [9], annotation tools. But these are optimised for highlighting sub-sentence structure and not for selecting entire sentences - especially when they span multiple screen lines (as is usually the case in our corpus where sentences are 29 words long on average). It soon became clear that just trying to highlight a sentence with these tools can take longer than working out the correct label. Thus we built our own web-based annotation tool, which allows users to quickly classify UKHL sentences as *full agreement*, *acknowledgement* or *neither* and select the names of any judges forming a MO. Each expert used our tool to annotate all 300 cases over a period of two months.

[2] The ASMO corpus can be accessed at http://www.holj.ml/asmo where the full text and consensus annotations of all 300 judgements can be seen by simply clicking on the case number (or by following the direct links embedded in text of this paper discussing specific examples).

[3] See https://publications.parliament.uk/pa/ld/ldjudgmt.htm

[4] See https://www.crummy.com/software/BeautifulSoup/

3.3. Arbitrating AS Annotations

In the first task, the annotators collectively labelled 2357 sentences as containing AS. They unanimously agreed on the label of 1671 (71%) of these. Of the other 686, at least two annotators agreed in all but two instances. A consensus labelling was therefore obtained by taking the majority view where it existed, or taking the senior annotator's view in the few instances where it didn't. We found three common sources of dispute.

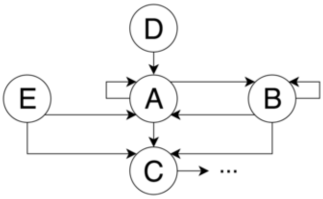
The first sort of disagreement concerns the distinction between agreement on the outcome and agreement on the reasons for the outcome. For example, one annotator mislabeled the sentence: *"In agreement with my noble and learned friend Lord Scott of Foscote, I too would allow the appeal."*, as a full agreement, when in actual fact it can only be used to infer agreement with the outcome (since, unlike sentence 1 in Table 1, it doesn't actually mention the *"reasons"*) and so it is only a partial agreement.

The second type of disagreement is on whether a sentence contains an agreement in-full with reasons leading to an outcome, or only an agreement with some subset of those reasons. For example in the sentence: *"On the basis of the wider approach to the problem of comparison which my noble and learned friend Lord Slynn of Hadley has adopted I am in full agreement with him that the rules of procedure for a claim under section 2(4) of the 1970 Act are not less favourable than those which would apply to a claim for breach of contract in the circumstances of the present cases."*, the agreement is limited to the interpretation of the 1970 Act. Hence this sentence is also a partial agreement.

The third issue concerns the ambiguous phrasing judges sometimes use to describe their agreement. For example the sentence: *"For these reasons, which really do no more than echo a part of the altogether fuller reasoning contained in the opinion of my noble and learned friend Lord Neuberger of Abbotsbury, which I have had the advantage of reading in draft and with all of which I am in complete agreement, I too would dismiss both appeals."* was understood by two experts as a full agreement with Lord Neuberger, and by one expert as a self agreement of Lady Hale together with a full agreement with Lord Neuberger. It is hard to say definitively who is correct in cases like this.

3.4. Arbitrating MO Annotations

In the second task, the experts unanimously agreed in 230 (77%) of the 300 cases. Of the other 70, at least two annotators agreed in all but six cases. Again, a consensus was obtained by taking the majority view where it existed, or the senior annotator's view in the few cases where it didn't. This time, the two main sources of dispute came from differing views about which judges some problematic AS were actually agreeing with, and from errors made in determining what MO a given set of agreements implied. A good example is shown by the graph below, depicting the full agreements in case 102:



In this case, one expert inferred {A,B,C} as the MO, presumably because (for the reasons previously explained in the last sentence of Section 2.2) it seems to be supported by all the judges. Now, if C meant to only agree with A and B, then that would indeed be true. But, given his generic agreement, we can't rule out the possibility, unlikely as it might be, that C also meant to agree with D or E - which would then force them into the MO as well! Since we can't infer C's intentions from the AS, we must conclude along with the other two experts that there is no unambiguous MO in this case.

4. Automated Annotation Study

This section describes our approach to automating the identification of MO. To avoid over-fitting our training data, we randomly split our corpus into three equal subsets of 100 cases, called the AS-set, MO-set and AI-set, that were used in the following three subsections, respectively . The AS-set was used for the training and testing our ML model for the sub-task of AS classification. The MO-set was used for development and testing of our rules for the sub-task of MO identification. The AI-set was used to evaluate the complete AI system obtained by combining our ML model for AS classification with our rules for MO identification.

4.1. Automating AS Identification

We approached the task of AS identification as a text classification problem. Three commonly used ML algorithms, Support Vector Machines (SVM), Logistic Regression (LR) and Naive Bayes (NB) were trained to classify sentences of our corpus as acknowledgement, full agreement or neither. Each algorithm was evaluated and the F1-score was used to select the best one. All our experiments were implemented using the scikit-learn library [10] using 10-fold cross validation.

As we are only interested in a few very specific sentences from a vast corpus, our classes are highly imbalanced by nature. To prevent our models simply favouring the most frequent category of *neither* (42 525 sentences), we down-sampled the AS-Set, by randomly selecting sentences to obtain a more balanced training set with 1 292 sentences in both the *none* and *full agreement* categories, and 374 sentences in the *acknowledgement* category.

Inspired by Hachey et al.'s research on rhetorical zoning [4] and Palau et al.'s research on argumentation mining [11] we take advantage of a combination of traditional features used for the task of legal text classification. These include unigrams, part of speech tags (POS), sentence lengths, sentence position and named entities (NE). We also employ custom designed cue phrase feature, inspired by Teufel et al.'s research on rhetorical zoning [12,13]. Our features are extracted using the NLTK library [5]. The cue phrases were manually selected by a human annotator based on commonly occurring phrases in the sentences of interest. They include phrases such as: *"for these reasons", "allow/dismiss the appeal" or "I have had the advantage"*. Some of the words contained in our cue phrases (e.g. *appeal*, *dismiss*, *reasons*) are already automatically identified by the ML algorithms as the most informative unigram features. However, the cue phrases also capture their word order, making them a valuable feature. Our POS and unigram features were normalized using term frequency-inverse document frequency (TFIDF).

The best individual feature are unigrams, followed by POS tags, cue phrases, NE, position and length. The high performance of cue phrases alone suggests AS are indeed

	SVM		LR		NB	
	Ind	Cum	Ind	Cum	Ind	Cum
Unigrams	0.938	**0.938**	0.938	0.938	0.922	0.922
POS tags	0.907	**0.938**	0.893	0.935	0.851	0.916
Cue phrases	0.800	0.935	0.802	0.939	0.800	**0.936**
NE	0.490	0.935	0.490	0.942	0.186	0.931
Position	0.342	0.937	0.342	**0.943**	0.186	0.931
Length	0.282	0.937	0.329	**0.943**	0.186	0.922

Table 2. *ML experiments reporting weighted-average F-scores for 10-fold cross-validation for individual (Ind) and cumulative (Cum) performance of features.*

formulaic. On the other hand, the drop of performance between unigrams and cue phrases point to the necessity of employing a ML model to achieve F-score of 0.90 and above.

As shown in Table 2 the LR model performs the best, achieving an F-score of 0.943. It is also the only model able to successfully integrate all of the features (except sentence length) to cumulatively improve its performance. Based on this we chose to use the the LR model trained on all features but length in our complete AI system.

4.2. Automating MO Identification

Resolving AS to obtain MO consists of two sub-tasks, parsing the AS to build a graph of the case relations and resolving such graph to identify MO. To build a graph of judge agreements we parse the AS based on our observations of different types of AS structures from Section 2.1 Table 1. First however, we remove all sentences with a number in them, since these can't be AS and are a result of our ML model misannotation. The number in a sentence is a proxy for a case citation, indicating an agreement with a past judgment, or specification of a legal point of another judge, indicating only a partial agreement. We remove all acknowledgements which were not followed by a full agreement, since they are not necessary for the purpose of anaphora resolution. We remove the parts of the sentences proposing orders, since orders sometimes contain names of the judges who the agreement isn't with. We remove the acknowledging sentences if the names of the judges in them don't match the names of judges in full agreement sentences. The remaining AS sentences of a judgment are merged together. We check an indication of a self-agreement expressed by phrases such as *"For these reasons"*. Finally, we extract the names of the judges. For each case we store the agreements in a dictionary, representing our graph.

To resolve our graph we follow three rules that we believe capture the common inference patters that we previously explained in Sections 2.2 and 3.4. First we take the transitive closure of the agreement graph, using the well-known Floyd-Warshall algorithm. Intuitively this progressively adds implicit agreements from a judge A to any judge C who is agreed with by any judge B that A already agrees with. Second we take a qualified reflexive closure of the resulting graph, which means adding self agreements to any judges that have no outgoing edges (i.e. who have not expressed a full agreement with any other judge). Finally we take each judge in turn and see what precise set of judges they agree with. If any such set has the support of more than half the judges then it is taken as the MO. Applied on our MO-set, our method finds (human-identified) MO from (human-classified) AS with 89% accuracy.

4.3. Complete AI System

We built our full AI system by combining the ML model of Section 4.1 with the rules of Section 4.2. When tested on the the independent AI-set, our system replicated the consensus MO with 81% accuracy. By contrast, the average expert agreement with the consensus MO is 91%, with the lowest being 85%. Hence we claim our system is close to achieving a human level of performance. It also significantly outperforms all obvious baselines we tested such as saying there isn't any MO (29%), choosing the single most cited judge (38%), choosing the judge most mentioned in AS (48%), choosing the set of judges with opinions longer than some optimal number of sentences (43%), or choosing the single judge with the most sentences (48%).

4.4. Error Analysis

There are three reasons why we don't achieve 100% agreement with human annotators. The first cause of error is imperfectly parsed AS caused by an unusual formulation of a sentence. For example in case 233, the wording *"in the manner Lord Hutton proposes"* is used to express an agreement with an order, instead of the traditional expression which explicitly contains the word *"order"*, thereby confusing our system. To resolve these instances in the future, a statistical approach trained on a larger data set could perhaps learn to correctly identify such unusual expressions of an agreement.

The second cause of error arises when annotators arguably misannotate a case. The two common sources of annotator error in ASMO corpus are exemplified in Figure 2. The first error arises in instances of complex transitive agreement (e.g. case 90). Our experts labelled this as having no MO. But we would argue this is incorrect as there are three judges {C,D,E} all in mutual agreement once we add in the implicit transitive agreement of E with himself. The second error arises where there are only two agreements with multiple judges. For example in case 300 the human consensus is that the MO is {B,E}, but we would say there is no MO, as only two judges {A,C} actually agree on this.

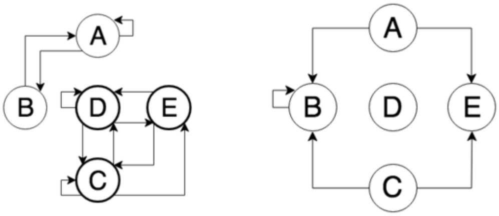


Figure 2. *Arguably mislabeled cases 90 (left diagram) and 300 (right diagram) from our corpus.*

Finally our ML model sometimes misclassifies AS. Since most of our cases have only 5 judges, a single erroneous agreement in our graph can often result in identifying the wrong MO. The low error rate we report for our ML model is therefore amplified on the task we employ it for. Particularly confusing for our model are partial agreements. For example the sentence *"I also agree with the supplementary observations made by Lord Walker and by Lord Neuberger in their opinions."* in case 91, contains the word stems *"agree"* and *"opinion"* which our ML model associates with full agreement and yet the the phrase *"supplementary observations"* gives away that this agreement is not with the full reasoning.

5. Conclusion

Our goal was to develop a computational method for identifying AS and MO in UKHL case law. To this end we demonstrated the considerable difficulties inherent in this task and we introduced a novel way of approaching these challenges from a graph-theoretic point of view. We also introduced an extensive expert-annotated ASMO corpus together with an AI system that is able to automatically identify explicit AS and implied MO with an accuracy approaching that of humans (81%).

Although we have only been concerned with the identification of MO supported by explicit inter-judge expressions of agreement, this is sufficient to establish the existence of legally binding MO in over two thirds (71%) of our corpus. Moreover, we believe the work we have done is a necessary stepping-stone towards the development of more sophisticated methods that might attempt to resolve the remaining cases through more nuanced partial agreements between judges and/or appealing to implicit semantic similarities in their legal arguments.

References

[1] J. Wilson. UKSC judgments: the case for a single, identifiable majority opinion UKSCBlog. http://ukscblog.com/uksc-judgments-the-case-for-a-single-identifiable-majority-opinion/.

[2] J. Lee. A defence of concurring speeches. *Public Law*, (Apr):305–331, 2009.

[3] B. Hale. Judgment Writing in the Supreme Court Brenda Hale UKSCBlog. http://ukscblog.com/judgment-writing-in-the-supreme-court-brenda-hale/.

[4] B. Hachey and C. Grover. Extractive summarisation of legal texts. *Artif. Intell. Law*, 14(4):305–345, December 2006.

[5] E. Loper and S. Bird. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, pages 63–70, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[6] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014.

[7] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii. Brat: A web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 102–107, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[8] H. Cunningham, V. Tablan, A. Roberts, and K. Bontcheva. Getting more out of biomedical documents with gate's full lifecycle open source text analytics. *PLOS Computational Biology*, 9(2):1–16, 02 2013.

[9] J. M. Cejuela, P. McQuilton, L. Ponting, S. J. Marygold, R. Stefancsik, G. H. Millburn, B. Rost, and the FlyBase Consortium. tagtog: interactive and text-mining-assisted annotation of gene mentions in plos full-text articles. *Database: The Journal of Biological Databases and Curation*, 2014:33, 2014.

[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[11] R. M. Palau and M. Moens. Argumentation mining: The detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, ICAIL '09, pages 98–107, New York, NY, USA, 2009. ACM.

[12] S. Teufel, A. Siddharthan, and D. Tidhar. Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 103–110, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[13] S. Teufel and M. Moens. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445, 2002.

Legal Knowledge and Information Systems
M. Palmirani (Ed.)

doi:10.3233/978-1-61499-935-5-151

Exploiting Causality in Constructing Bayesian Network Graphs from Legal Arguments

Remi WIETEN [a], Floris BEX [a,b], Henry PRAKKEN [a,c] and Silja RENOOIJ [a]
[a] *Information and Computing Sciences, Utrecht University, The Netherlands*
[b] *Institute for Law, Technology and Society, Tilburg University, The Netherlands*
[c] *Faculty of Law, University of Groningen, The Netherlands*

Abstract. In this paper, we propose a structured approach for transforming legal arguments to a Bayesian network (BN) graph. Our approach automatically constructs a fully specified BN graph by exploiting causality information present in legal arguments. Moreover, we demonstrate that causality information in addition provides for constraining some of the probabilities involved. We show that for undercutting attacks it is necessary to distinguish between causal and evidential attacked inferences, which extends on a previously proposed solution to modelling undercutting attacks in BNs. We illustrate our approach by applying it to part of an actual legal case, namely the Sacco and Vanzetti legal case.

Keywords. Bayesian networks, legal reasoning, argumentation, causality

1. Introduction

Bayesian networks (BNs) are probabilistic reasoning tools that are being applied in many complex domains where uncertainty plays a role, including forensics and law [1]. A BN consists of a graph, which captures the probabilistic independence relation among the modelled domain variables, and locally specified (conditional) probability distributions that collectively describe a joint probability distribution. BNs are well-suited for reasoning about the uncertain consequences that can be inferred from the evidence in a case. However, especially in data-poor domains, their construction needs to be done mostly manually, which is a difficult, time-consuming and error-prone task. Domain experts such as crime analysts and legal experts, therefore, typically resort to using qualitative reasoning tools, such as argument diagrams and mind maps [2] and Wigmore charts [1].

The benefits of BNs can be exploited if their construction can be facilitated by extracting information specified by experts using the tools and techniques they are familiar with. In previous research, Bex and Renooij [3] contributed to this idea by deriving constraints on a BN given structured arguments [4]. Their approach suffices for automatically constructing an undirected graph. However, for setting arc directions, as required for the directed BN graph, the authors resort to the standard approach used in BN construction: the BN modeller and the domain expert together specify the arc directions using the notion of causality as a guiding principle [5]. The resulting graph then has to be verified and refined in terms of the independence relation it represents.

From a legal perspective, it is, however, justified to assume that information regarding causality is present in the domain expert's original argument-based analysis [6, 7]. In this paper, we therefore propose to make causality information explicit in the argument-based analysis and to exploit this information in automatically constructing a completely directed BN graph. Moreover, we demonstrate that causality information in addition provides for constraining several conditional probability distributions. We illustrate our approach by applying it to the Sacco and Vanzetti legal case and compare our results to a previous BN modelling of the same case [1].

2. Preliminaries

2.1. Argumentation

Throughout this paper, we assume that the domain expert's analysis of a case is captured in an *argument graph*, in which claims are substantiated by chaining inferences from the observed case evidence; an example is depicted in Figure 1a. Argument graphs, which we define below, are closely related to Wigmore charts [1], mind maps and argument diagrams [2] with which many crime analysts and legal experts are familiar. We note that our formalism is an abstraction of existing argumentation formalisms (an overview is provided by [8]). Specifically, the evaluation of arguments is not taken into account, as it is not needed for our current purposes. By embedding our formalism in other existing formalisms, the dialectical status of arguments can be accounted for (cf. [9]).

Formally, an argument graph is a directed graph $AG = (\mathbf{P}, \mathbf{A})$, where $\mathbf{P}$ is a set of nodes representing propositions and $\mathbf{A}$ is a set of directed (hyper)arcs. Nodes $\mathbf{Ev} \subseteq \mathbf{P}$ corresponding to the (observed) *case evidence* are shaded root nodes in the argument graph. We assume that for every $p \in \mathbf{Ev}$, there is no node representing proposition p's negation $\neg p$ elsewhere in the graph. The set of directed (hyper)arcs $\mathbf{A}$ is comprised of three pairwise disjoint sets $\mathbf{S}$, $\mathbf{R}$ and $\mathbf{U}$, which are sets of support arcs, rebuttal arcs and undercutter arcs, respectively. A *support arc* is a solid (hyper)arc $s\colon p_1, \ldots, p_n \rightarrow p \in \mathbf{S}$, indicating an inference step from one or more propositions $p_1, \ldots, p_n \in \mathbf{P}$ (the *tails* of the arc) to a single proposition $p \in \mathbf{P}$ (the *head* of the arc). Two support arcs s_1 and s_2 form a *support chain* (s_1, s_2) in case the head of s_1 is a tail of s_2.

There are two types of attack arcs. A *rebuttal arc* $r \in \mathbf{R}$ is a bidirectional dashed arc in the argument graph that is drawn between two nodes in $\mathbf{P}$ iff these nodes represent a proposition p and its negation $\neg p$. An *undercutter arc* $u \in \mathbf{U}$ is a dashed hyperarc directed from a node $p \in \mathbf{P}$ (the undercutter of the inference) to a support arc $s \in \mathbf{S}$. Figure 3a depicts an example including undercutter arcs. Informally, a rebuttal is an attack on a proposition, while an undercutter attacks an inference by providing exceptional case circumstances under which the inference is not applicable.

In reasoning about evidence, a distinction can be made between causal and evidential inferences [6, 10]. Causal inferences are of the form '*c is a cause for e*' (e.g. fire causes smoke), whereas evidential inferences are of the form '*e is evidence for c*' (e.g. smoke is evidence for fire). We assume that every support arc $s \in \mathbf{S}$ is annotated with a causal 'c' or an evidential 'e' label; $\mathbf{S}$ then falls into two disjoint sets $\mathbf{S}^{\mathbf{C}}$ and $\mathbf{S}^{\mathbf{E}}$ of causal and evidential support arcs, respectively. We consider the tails of two causal arcs with the same head to be competing (but not necessarily mutually exclusive) causes for the common effect expressed by the head. Similarly, in case two evidential support arcs have distinct heads and a single, identical tail, we consider the heads to be competing causes for the common effect expressed by the tail. In case a causal or evidential support arc has multiple tails, we assume that the tails are not in competition, as the arc expresses that only the tails together allow us to infer the head.

As noted by Pearl [10], the chaining of a causal inference and an evidential inference can lead to undesirable results. Consider the example in which a causal inference states that a smoke machine causes smoke and an evidential inference states that smoke is evidence for fire. If we now know that there is a smoke machine, then this is evidence that there is a fire, which is clearly undesirable. To this end, we assume that an argument graph does not contain a support chain (s_1, s_2) where $s_1 \in \mathbf{S^C}$, $s_2 \in \mathbf{S^E}$.

2.2. Bayesian Networks

A BN [5] is a compact representation of a joint probability distribution $\Pr(\mathbf{V})$ over a finite set of discrete random variables $\mathbf{V}$. The variables are represented as nodes in a directed acyclic graph $G = (\mathbf{V}, \mathbf{E})$, where $\mathbf{E} \subseteq \mathbf{V} \times \mathbf{V}$ is a set of directed arcs $V_i \rightarrow V_j$ from parent V_i to child V_j. Each node describes a number of mutually exclusive and exhaustive values; in this paper, we assume all nodes to be Boolean. The BN further includes, for each node, a conditional probability table (CPT) specifying the probabilities of the values of the node conditioned on the possible joint value combinations of its parents. A node is called instantiated iff it is set to a specific value. Given a set of instantiations, or evidence, the probability distributions over the other nodes in the network can be updated using probability calculus [5]. An example of a BN graph and one of its CPTs is depicted in Figures 1b and 1c, where ovals denote nodes and instantiated nodes are shaded.

The BN graph G captures the independence relation among its variables. Let a chain be defined as a sequence of distinct nodes and arcs in the BN graph. A node V is called a *head-to-head node* on a chain c if it has two incoming arcs on c. A chain c is *blocked* iff it includes a node V such that (1) V is an uninstantiated head-to-head node on c without instantiated descendants; or (2) V is instantiated and has at most one incoming arc on c. A chain is inactive if it is blocked; otherwise it is called *active*. If no active chains exist between V_1 and V_2 given instantiations of nodes in $\mathbf{Z}$, then they are considered conditionally independent given $\mathbf{Z}$. After constructing an initial BN graph, it should be verified that this graph is acyclic and that the graph correctly captures the (conditional) independencies. If the graph does not yet exhibit these properties, arcs should be reversed, added or removed by the BN modeller in consultation with the domain expert. We will refer to this step as the *'initial validation step'*.

In case the two parents of a head-to-head node are seen as causes of a common effect, then instantiation of the head-to-head node or one of its descendants will induce an active chain between the causes. If one of the causes is now observed, then the probability of the other cause being present as well can either increase, decrease or stay the same upon updating, depending on the probabilities in the CPT for the head-to-head node. In case the probability of the other cause decreases, this is called the *'explaining away'* effect [11]. To achieve the explaining away effect between two causes $V_1 = \mathit{true}$ (denoted v_1) and $V_3 = \mathit{true}$ (v_3) of $V_2 = \mathit{true}$ (v_2), the CPT for V_2 needs to be constrained such that V_1 and V_3 exhibit a *negative product synergy wrt* v_2: $\Pr(v_2 \mid v_1, v_3) \cdot \Pr(v_2 \mid \neg v_1, \neg v_3) \leq \Pr(v_2 \mid v_1, \neg v_3) \cdot \Pr(v_2 \mid \neg v_1, v_3)$. A *zero product synergy* is a special case of a negative product synergy, where the inequality in the above equation is replaced by an equality. In this case, no intercausal reasoning can occur between causes V_1 and V_3 of v_2.

3. Exploiting Causality in Constructing BN Graphs from Legal Arguments

In this section, we first present the steps of our structured approach for automatically constructing BN graphs from argument graphs, and then illustrate and explain the steps with several examples. Let $node\colon \mathbf{P} \to \mathbf{V}$ be an operator which maps every proposition p or $\neg p \in \mathbf{P}$ to a variable $node(p) = node(\neg p) \in \mathbf{V}$ that describes values p and $\neg p$. For a given argument graph $AG = (\mathbf{P},\mathbf{A})$, a BN graph $G = (\mathbf{V},\mathbf{E})$ is constructed as follows:

1. For every p or $\neg p \in \mathbf{P}$, the BN graph includes $node(p)$. For every $p \in \mathbf{Ev} \subseteq \mathbf{P}$, the set of observed nodes $\mathbf{Ev} \subseteq \mathbf{V}$ includes $node(p)$.
2. For every support arc $s\colon p_1,\ldots,p_n \to p \in \mathbf{S^C}$, $\mathbf{E}$ includes arcs $node(p_1) \to node(p),\ldots,node(p_n) \to node(p)$.
3. For every support arc $s\colon p_1,\ldots,p_n \to p \in \mathbf{S^E}$, $\mathbf{E}$ includes arcs $node(p) \to node(p_1),\ldots,node(p) \to node(p_n)$.
4. For every undercutter arc $u \in \mathbf{U}$ directed from a $p \in \mathbf{P}$ to support arc $s\colon q_1,\ldots,q_n \to q \in \mathbf{S^E}$, $\mathbf{E}$ includes arcs $node(p) \to node(q_1),\ldots,node(p) \to node(q_n)$.
5. For every undercutter arc $u \in \mathbf{U}$ directed from a $p \in \mathbf{P}$ to support arc $s\colon q_1,\ldots,q_n \to q \in \mathbf{S^C}$, $\mathbf{E}$ includes arc $node(p) \to node(q)$.
6. Verify the properties of the BN graph by performing the initial validation step.

We propose the following constraints on the CPTs of the BN under construction:

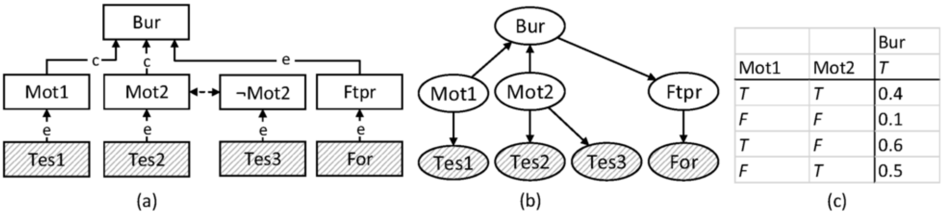
7a) For every pair of support arcs $s_1\colon p_1,\ldots,p_n \to p, s_2\colon q_1,\ldots,q_m \to p \in \mathbf{S^C}$, the CPT for $node(p)$ should be constrained such that $node(p_i)$ and $node(q_j)$ exhibit a negative product synergy wrt value p of $node(p)$ for $1 \leq i \leq n, 1 \leq j \leq m$.

7b) For every pair of support arcs $s_1\colon p \to p_1, s_2\colon p \to p_2 \in \mathbf{S^E}$, the CPT for $node(p)$ should be constrained such that $node(p_1)$ and $node(p_2)$ exhibit a negative product synergy wrt value p of $node(p)$.

7c) For every support arc $s\colon p_1,\ldots,p_n \to p \in \mathbf{S^C}$, the CPT for $node(p)$ should be constrained such that $node(p_i)$ and $node(p_j)$ exhibit a zero product synergy wrt value p of $node(p)$ for $1 \leq i \leq n$, $1 \leq j \leq n$, $i \neq j$.

7d) For every undercutter arc $u \in \mathbf{U}$ directed from a $p \in \mathbf{P}$ to support arc $s\colon q_1,\ldots,q_n \to q \in \mathbf{S^E}$, the CPT for $node(q_i)$ should be constrained such that $node(p)$ and $node(q)$ exhibit a negative product synergy wrt value q_i of $node(q_i)$ for $1 \leq i \leq n$. For every pair of undercutter arcs $u_1,u_2 \in \mathbf{U}$ directed from $p_1 \in \mathbf{P}$ respectively $p_2 \in \mathbf{P}$ to s, the CPT for $node(q_i)$ should be further constrained such that $node(p_1)$ and $node(p_2)$ exhibit a negative product synergy wrt value q_i of $node(q_i)$ for $1 \leq i \leq n$.

In Section 3.1, we illustrate that steps 1-3 of our approach suffice for constructing BN graphs from argument graphs without undercutter arcs, where the CPTs of the BN under construction should be constrained according to steps 7a-7c. We then illustrate in Section 3.2 that the BN under construction needs to be further constrained in case undercutter arcs are present in the argument graph; this is accounted for in steps 4, 5 and 7d of our approach. Finally, we note that, while our approach exploits the domain knowledge captured in the argument graph, the argument graph may not contain all the information needed to resolve conflicts such as cycles and unwarranted (in)dependencies in the obtained BN graph; hence, step 6 of our approach.



		Bur
Mot1	Mot2	*T*
T	*T*	0.4
F	*F*	0.1
T	*F*	0.6
F	*T*	0.5

Figure 1. An argument graph involving two competing causes Mot1 and Mot2 for Bur (a); the corresponding BN graph constructed by our approach (b); a possible CPT for node Bur (c).

3.1. Argument Graphs without Undercutter Arcs

We first describe the main idea behind our approach. As an argument graph describes the inferences that can be made between a set of proposition nodes and a BN graph describes an independence relation over a set of variables, a transformation step from propositions to variables needs to occur upon constructing a BN graph; this is accounted for in step 1 of our approach. By the same step, two propositions involved in a rebuttal are captured as two mutually exclusive values of the same node. In steps 2 and 3, arcs in the BN graph are then directed using the *notion of causality*, that is, for every causal support arc $s \in \mathbf{S^C}$, arcs in the BN graph are directed from the nodes corresponding to the tails of s to the node corresponding to the head of s, and vice versa for an evidential support arc. This formalises the approach typically taken in the manual construction of BN graphs [5].

In comparing the knowledge captured in a BN graph to the knowledge captured in the original argument graph, it is important to note that a BN graph necessarily represents more than an argument graph, as for every proposition $p \in \mathbf{P}$ a variable is created which describes both values p and $\neg p$. Furthermore, without instantiating variables, a BN in itself is a joint probability distribution which does not model directionality; only when instantiating variables do reasoning patterns arise in the form of induced active chains. As the notion of an active chain is a symmetrical concept (an active chain exists between variables A and B given $\mathbf{Ev}$ iff an active chain exists between B and A given $\mathbf{Ev}$), a BN graph will also capture reasoning patterns in the opposite direction from the support arcs present in the argument graph. Our current focus will be on verifying whether the (chains of) support arcs in the argument graph are present in the BN graph in the form of active chains given the relevant case evidence. As the observed case evidence $\mathbf{Ev} \subseteq \mathbf{P}$ is indicated in the argument graph, we instantiate only the corresponding variables $\mathbf{Ev} \subseteq \mathbf{V}$. Here, we note that in determining the influence of a variable $E_i \in \mathbf{Ev}$, active chains given the variables $\mathbf{Ev} \setminus \{E_i\}$ need to be considered.

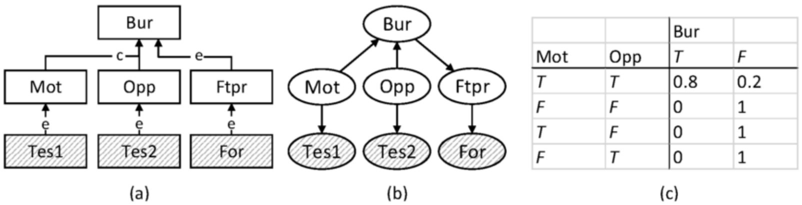
We illustrate our approach through the example depicted in Figure 1a. Suppose that a burglary has taken place and that we are interested in whether some suspect is in fact the burglar (Bur). Forensic analysis (For) shows there is a match between a pair of shoes owned by the suspect and footprints found near the crime scene, which provides us with evidence that the suspect left his footprints at the crime scene (Ftpr). According to witness testimony (Tes1), the suspect had a motive to commit this burglary, namely to commit theft (Mot1); however, according to another testimony (Tes2), the suspect had a different possible motive, namely to harm the owner (Mot2). These motives are expressed as competing causes for Bur in the argument graph. Finally, the suspect denied in his testimony (Tes3) that he had intentions to harm the owner (¬Mot2), which rebuts Mot2.

By following steps 1-3 of our approach, the BN graph of Figure 1b is constructed. For every support chain (s_1, s_2) in the argument graph, there exists an active chain between the nodes corresponding to the tails of s_1 and the head of s_2 in the BN graph given **Ev**. Specifically, given the case evidence $\mathbf{Ev} = \{$Tes1, Tes2, Tes3, For$\}$, there exist active chains between Tes1 and Bur, Tes2 and Bur, Tes3 and Mot2, and For and Bur in the BN graph. Generally, we note that for any given argument graph, no head-to-head node is formed in the node corresponding to the head of s_1 for any support chain (s_1, s_2) in the argument graph by steps 1-3 of our approach. Since propositions in **Ev** are root nodes in an argument graph, corresponding instantiated nodes are root nodes in the BN graph; chains in the BN graph corresponding to support chains in the argument graph are, therefore, always active given **Ev**. Manual verification of whether support chains are captured in the form of active chains in the BN graph can, therefore, be skipped.

Besides structural constraints, information regarding causality in the argument graph also allows us to derive constraints on the CPTs of the BN. In the argument graph of Figure 1a, Mot1 and Mot2 are competing causes for the common effect Bur in that presence of one of the causes makes the other cause less likely. We propose to link this type of intercausal interaction in argument graphs to the explaining away effect in BNs. Specifically, as proposed in step 7a of our approach, the CPT for Bur should be constrained such that Mot1 and Mot2 exhibit a negative product synergy wrt value Bur $=$ *true*. For example, the CPT for Bur can be chosen as in Figure 1c, as in this case $0.4 \cdot 0.1 \leq 0.6 \cdot 0.5$. Similarly, the heads of two evidential support arcs with the same tail are in competition for the effect expressed by the tail. Suppose that the two causal support arcs Mot1 $\rightarrow$ Bur and Mot2 $\rightarrow$ Bur in Figure 1a are replaced by two evidential support arcs Bur $\rightarrow$ Mot1 and Bur $\rightarrow$ Mot2. By following steps 1-3 of our approach, again the BN graph of Figure 1b is constructed. Competition between the causes can again be captured by constraining the CPT for Bur accordingly, as proposed in step 7b of our approach. We note that, in case two evidential support arcs have multiple (shared) tails, intercausal interactions will not in general be of some predetermined type that constrains the CPTs.

By following steps 2 and 3 of our approach, two competing causes automatically form a head-to-head connection in the node corresponding to the common effect for any given argument graph; interaction between two competing causes in an argument graph can, therefore, always be directly captured in the CPT for the node corresponding to the common effect. We note that the intended explaining away effect is only active in case the node corresponding to the common effect is instantiated or has an instantiated descendant. This is the case when in the corresponding argument graph the common effect is an element of **Ev** or in case there exists a chain of evidential support arcs for the common effect. For example, in Figure 1b the explaining away effect is active between Mot1 and Mot2, as Bur has instantiated descendant For; this is the case as there exists a chain of evidential support arcs from For to Bur in Figure 1a.

Figure 2a depicts an adjustment of the argument graph of Figure 1a, where the single causal support arc between Mot, Opp and Bur indicates that the presence of motive and opportunity together caused the suspect to commit the burglary. According to steps 1-3 of our approach, the BN graph of Figure 2b is constructed. Similar to the BN graph of Figure 1b, an active chain exists between Mot and Opp given **Ev**; however, as Mot and Opp are not in competition for Bur in this example, we need to assure that no intercausal reasoning can occur between Mot and Opp for value Bur $=$ *true*. This can be achieved by constraining the CPT for Bur such that Mot and Opp exhibit a zero product synergy wrt value Bur $=$ *true*, as proposed in step 7c of our approach. Note that the argument graph only informs us that there should be a zero product synergy between Mot and Opp



		Bur	
Mot	Opp	*T*	*F*
T	*T*	0.8	0.2
F	*F*	0	1
T	*F*	0	1
F	*T*	0	1

Figure 2. An argument graph involving two non-competing causes Mot and Opp for Bur (a); the corresponding BN graph constructed by our approach (b); a possible CPT for node Bur (c).

wrt value Bur $=$ *true*; it does not inform us whether this should also hold wrt value Bur $=$ *false*. Figure 2c depicts a possible CPT for Bur, where Mot and Opp exhibit a zero product synergy wrt value Bur $=$ *true* as $0.8 \cdot 0 = 0 \cdot 0$. However, as $0.2 \cdot 1 \leq 1 \cdot 1$, Mot and Opp also exhibit a negative product synergy wrt value Bur $=$ *false*. Care should be taken, therefore, in eliciting the involved probabilities.

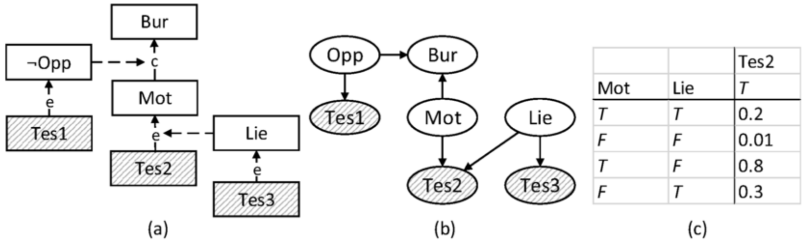
3.2. Argument Graphs including Undercutter Arcs

Next, undercutting attacks are considered. Bex and Renooij [3] interpret undercutting attack as explaining away and propose constraints on the BN graph and CPTs accordingly. We demonstrate that their solution only has the desired effect in case the undercut support arc is evidential. The solution of Bex and Renooij is captured by steps 4 and 7d of our approach; step 5 then accounts for undercut causal support arcs.

In Figure 3a, an example of an argument graph is depicted in which both an evidential and a causal support arc are undercut. The evidential support arc Tes2 $\rightarrow$ Mot is undercut by proposition Lie, which states that person x who testified to Tes2 had reason to lie when giving his testimony. Since Tes2 is either the result of the true motive or due to a lie, Mot and Lie can be seen as competing causes of x's testimony. Generally, undercutters of evidential support arcs can be considered competing causes for the common effects expressed by the tails of the support arc; for undercut evidential support arcs, we therefore follow the approach of [3]. The authors specify that the nodes corresponding to the propositions involved in an undercutting attack should form head-to-head nodes in the tails of the undercut support arc. By step 3 of our approach, the BN graph under construction includes arc Mot $\rightarrow$ Tes2. A head-to-head node can, therefore, be formed in node Tes2 by adding additional arc Lie $\rightarrow$ Tes2 to the BN graph; this is captured by step 4 of our approach. The explaining away effect can then be achieved by constraining the CPT for Tes2 such that Lie and Mot exhibit a negative product synergy wrt value Tes2 $=$ *true*; this is captured by step 7d of our approach. For example, the CPT for Tes2 can be chosen as in Figure 3c, as in this case $0.2 \cdot 0.01 \leq 0.8 \cdot 0.3$.

We note that, in case there are multiple undercutters of an evidential support arc, the CPTs for the nodes corresponding to the tails need to be further constrained such that the explaining away effect occurs between the nodes corresponding to each pair of undercutters of the undercut support arc, as each undercutter expresses an additional competing cause; this is also captured by step 7d of our approach.

In the argument graph of Figure 3a, the causal support arc Mot $\rightarrow$ Bur is undercut by proposition $\neg$Opp. In contrast with the undercut evidential support arc, this undercutter cannot be considered a competing cause for the tail of the undercut support arc; the absence of opportunity cannot be considered a cause for motive. Instead, it can be considered a causal explanation for the fact that the suspect is not the burglar ($\neg$Bur). For undercut causal support arcs, we therefore propose to form a head-to-head node in



		Tes2
Mot	Lie	*T*
T	*T*	0.2
F	*F*	0.01
T	*F*	0.8
F	*T*	0.3

Figure 3. An argument graph involving both an undercutter of an evidential support arc and a causal support arc (a); the corresponding BN graph constructed by our approach (b); a possible CPT for node Tes2 (c).

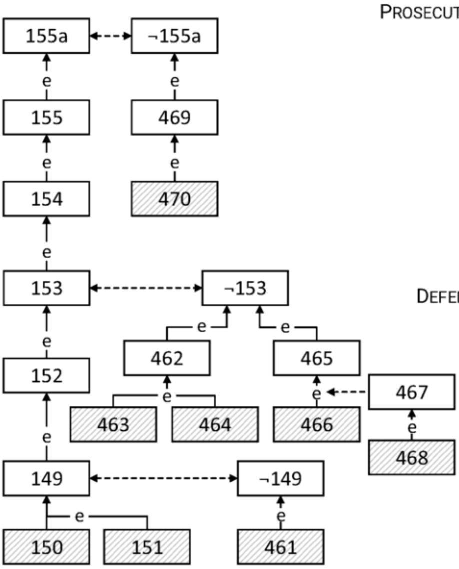
the node corresponding to the head of the support arc as opposed to in the node corresponding to the tail of the support arc as proposed by [3]. By step 2 of our approach, the BN graph under construction includes arc Mot → Bur. A head-to-head node can, therefore, be formed in Bur by adding additional arc Opp → Bur to the BN graph; this is captured by step 5 of our approach. As $node(\text{Bur})$ describes both values Bur and ¬Bur, possible interactions, if any, between Mot and ¬Opp can be captured in the CPT for this node. We note that, in this case, intercausal interactions will not in general be of some predetermined type that constrains the CPT.

4. Case Study: The Sacco and Vanzetti Case

In this section, we apply our approach to part of an actual legal case, namely the well-known Sacco and Vanzetti case. The case concerns Sacco and Vanzetti, who were convicted for shooting and killing a payroll guard during a robbery. Kadane and Schum [1] performed a probabilistic analysis of this case by first constructing Wigmore charts, which are a type of (evidential) argument graph. The authors then manually constructed corresponding BNs by assessing the modelled independence relation and eliciting the necessary (conditional) probabilities. The authors did not provide a systematic approach for constructing BN graphs from Wigmore charts. Specifically, they constructed BN graphs by compiling and aggregating the evidence present in multiple Wigmore charts, where the choice which evidence to aggregate and compile was decided upon a case-by-case basis. In this section, we illustrate for one of their Wigmore charts that our approach provides a more structured way of constructing BN graphs.

In this case study, we only consider the evidence regarding Sacco's consciousness of guilt. During their arrest, Sacco and Vanzetti were armed. According to the two arresting officers, Connolly and Spear, Sacco and Vanzetti made suspicious hand movements, from which the prosecution concluded that Sacco and Vanzetti intended to draw their concealed weapons in order to escape their arrest. This suggests that Sacco and Vanzetti were conscious of having committed a criminal act.

In Figure 4, an argument graph concerning Sacco's consciousness of guilt corresponding to the Wigmore chart from Kadane and Schum [1, pp. 330–331] is depicted, along with the corresponding key list which indicates for every number in the argument graph to which evidence it corresponds. For reasons of space, the original Wigmore chart from Kadane and Schum is omitted. In formalising the Wigmore chart from Kadane and Schum, we have followed the approach of Bex and colleagues [7]. As noted by Kadane and Schum [1][p. 74–76], the natural direction of reasoning in their Wigmore charts is from the evidence to the ultimate conclusion, as all the evidence is known to them. All support arcs in the argument graph of Figure 4 are, therefore, evidential.



PROSECUTION
149. Following his arrest, Sacco attempted to put his hand under his overcoat.
150. Connolly's testimony to 149.
151. Spear's testimony to 149.
152. Sacco intended to draw his concealed weapon.
153. Sacco intended to use his weapon on the arresting officers.
154. Sacco intended to escape from his arrest.
155. Sacco was conscious of having committed a criminal act.
155a. Sacco was conscious of having been involved in a robbery and shooting.

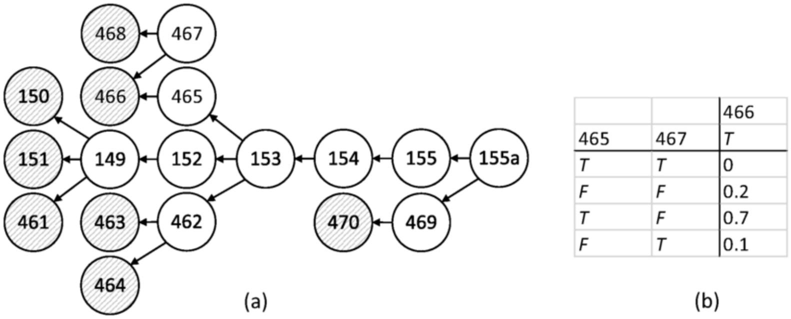
DEFENCE
461. Sacco's testimony to denying 149.
462. Sacco carried a weapon because he intended to shoot rabbits with it.
463. Sacco's testimony to 462.
464. Sacco's wife's testimony to 462.
465. Sacco carried a weapon because of his duties as a night watchman.
466. Sacco's testimony to 465.
467. Sacco was not a night watchman.
468. Sacco's admission on cross-examination.
469. Sacco believed he was being arrested because of his political beliefs.
470. Sacco's testimony to 469.

Figure 4. An argument graph corresponding to the Wigmore chart that Kadane and Schum constructed concerning Sacco's consciousness of guilt, along with the corresponding key list, adapted from [1, pp. 330–331].

By applying our construction approach, the BN graph of Figure 5a is obtained. Rebutting propositions are captured as two values of the same node by step 1 of our approach; arcs in the BN graph are then directed in the opposite direction of the evidential support arcs by step 3. Additional arc 467 $\rightarrow$ 466 is then added to the BN graph according to step 4. The obtained graph is largely identical to the BN graph that Kadane and Schum manually constructed for this part of the case [1, p. 232]. The only difference is that Kadane and Schum aggregated nodes 466, 467 and 468 into a single node; possible intercausal effects between 467 and 465 can, therefore, not be explicitly captured in their BN graph. While aggregation reduces the number of conditional probabilities to be elicited, we prefer to explicitly capture all elements of the argument graph in the corresponding BN graph to prevent loss of information. We note that, by step 7d of our approach, a constraint on the CPTs of the BN under construction is automatically obtained, which partially simplifies the elicitation procedure. By this step, nodes 465 and 467 should exhibit a negative product synergy wrt value 466 $=$ *true*. For example, the CPT for 466 can be chosen as in Figure 5b, as in this case $0 \cdot 0.2 \leq 0.7 \cdot 0.1$.

5. Conclusion

In this paper, we have studied how legal arguments, including causality information, can be used to inform the construction of BN graphs. We have proposed a structured approach that allows domain experts to automatically construct a BN graph corresponding to their initial argument-based analysis which captures similar reasoning patterns as present in the original argument graph. We have illustrated that for undercutting attacks it is necessary to distinguish between causal and evidential attacked inferences. As a result, we extend on Bex and Renooij's [3] previously proposed solution to modelling undercutting attacks in BNs. In addition, we have identified when constraints on the CPTs of the BN are required to capture the desired effect of (induced) dependencies. In related research, Grabmair and colleagues [12] claim (without proof) that the Carneades argument model



		466
465	467	*T*
T	*T*	0
F	*F*	0.2
T	*F*	0.7
F	*T*	0.1

Figure 5. The BN graph corresponding to the argument graph of Figure 4, as constructed by our approach (a); a possible CPT for node 466 (b).

can be given probabilistic semantics using BNs. However, the authors do not discuss issues related to forensics and causality.

In this paper, we have assumed that all support arcs are labelled with a causal or evidential label. In our future research, we intend to study the construction of BN graphs from partially labelled argument graphs. Furthermore, we have made no assumptions regarding inference strength; within our proposed constraints, the CPTs, therefore, need to be elicited manually. As arcs in the BN graph are directed from cause to effect, the necessary conditional probabilities can be elicited in the form of likelihood ratios, which are ratios of probabilities of observing the evidence under two mutually exclusive hypotheses. These can be directly entered in the CPT for each node, as is also done by e.g. Kadane and Schum [1]. In our future research, we will focus on deriving more probabilistic constraints, for example, by taking inference strength into account.

References

[1] J.B. Kadane and D.A. Schum. *A Probabilistic Analysis of the Sacco and Vanzetti Evidence.* John Wiley & Sons Inc., 1996.

[2] P.A. Kirschner, S.J. Buckingham Shum, and C.S. Carr, eds., *Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making*. Springer, 2003.

[3] F. Bex and S. Renooij. From arguments to constraints on a Bayesian network. In P. Baroni, T.F. Gordon, T. Scheffler, and M. Stede, eds., *Computational Models of Argument: Proceedings of COMMA 2016*, volume 287, pages 95–106. IOS press, 2016.

[4] H. Prakken. An abstract framework for argumentation with structured arguments. *Argument & Computation*, 1(2): 93–124, 2010.

[5] F.V. Jensen and T.D. Nielsen. *Bayesian Networks and Decision Graphs*. Springer, 2nd ed., 2007.

[6] F. Bex. An integrated theory of causal stories and evidential arguments. *Proceedings of the 15th International Conference on Artificial Intelligence and Law*, pages 13–22. ACM Press, 2015.

[7] F. Bex, H. Prakken, C.A. Reed, and D. Walton. Towards a formal account of reasoning about evidence: argumentation schemes and generalisations. *Artificial Intelligence and Law*, 11(2-3): 125–165, 2003.

[8] H. Prakken. Historical overview of formal argumentation. In P. Baroni, D. Gabbay, M. Giacomin, and L. van der Torre, eds., *Handbook of Formal Argumentation*, pages 73-141. College Publications, 2018.

[9] F. Bex, S. Modgil, H. Prakken, and C.A. Reed. On logical specifications of the argument interchange format. *Journal of Logic and Computation*, 23(5): 951–989, 2013.

[10] J. Pearl. Embracing causality in default reasoning. *Artificial Intelligence*, 35(2): 259–271, 1988.

[11] M.P. Wellman and M. Henrion. Explaining "explaining away". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(3): 287–292, 1993.

[12] M. Grabmair, T.F. Gordon, and D. Walton. Probabilistic semantics for the Carneades argument model using Bayesian networks. In P. Baroni, F. Cerutti, M. Giacomin, and G.R. Simari, eds., *Computational Models of Argument: Proceedings of COMMA 2010*, volume 216, pages 255–266. IOS Press, 2010.

Legal Knowledge and Information Systems
M. Palmirani (Ed.)

doi:10.3233/978-1-61499-935-5-161

Japanese Legal Term Correction Using Random Forests

Takahiro YAMAKOSHI [a], Takahiro KOMAMIZU [a,b], Yasuhiro OGAWA [a,b] and Katsuhiko TOYAMA [a,b]

[a] *Graduate School of Informatics, Nagoya University*
[b] *Information Technology Center, Nagoya University*
Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan

Abstract. We propose a method that assists legislation officers in finding inappropriate Japanese legal terms in Japanese statutory sentences and suggests corrections. In particular, we focus on sets of similar legal terms whose usages are defined in legislation drafting rules. Our method predicts suitable legal terms in statutory sentences using Random Forest classifiers, each of which is optimized for each set of similar legal terms. Our experiment shows that our method outperformed existing modern word prediction methods using neural language models.

Keywords. Japanese Legal Terms, Legal Term Correction, Random Forest

1. Introduction

Legislation drafting requires a lot of careful attention. The Japanese government deals with this task by means of thorough legislation drafting rules and final inspection by the Cabinet Legislation Bureau.

The drafting rules regulate document structures, orthography, and phraseology of statutes. These rules have been utilized for more than 100 years and are published as legislation manuals (e.g. [6]). Among the drafting rules, it is a noteworthy feature that they explicitly define distinct usage and meaning to many legal terms that look mutually similar. For example, the three Japanese words "者 $_{\text{(a)}}$," "物 $_{\text{(b)}}$," and "もの $_{\text{(c)}}$" are all pronounced *mono*. The Japanese legislation drafting rules prescribe that the term (a) only means a natural or juristic person, the term (b) only means a tangible object that is not a natural or juristic person, and the term (c) only means an abstract object or a complex of these objects. Phrases in Fig. 1 contain each legal term.

Using the drafting rules, legislative officers in the Cabinet Legislation Bureau strictly inspect legislative bills which are prudently written in the Cabinet Office or in each ministry, including the legal term usage. Therefore, any legal term defined in the rules must not appear vaguely or mistakenly in inspected bills. However, these inspections are still conducted mainly by human experts in legislation and that requires deep knowledge and an enormous amount of labor. Furthermore, according to Enami [4], this legislative work has become even tougher because of recent increased enactment of statutes.

Considering the above, we propose a method that assists legislation officers in finding inappropriate legal terms in a draft and offers correction ideas. By regarding a set of similar legal terms as a set of choices, we handle the legal term correction as a special

chosakubutsu	*wo*	*sosakusuru*	*mono*
著作物	を	創作する	者(a)
work	ACC	create	person

【a person who creates a work】

chikuonkiyoonban	,	*rokuontepu*	*sonotano*	*mono*
蓄音機用音盤	、	録音テープ	その他の	物(b)
phonograph disc	,	recording tape	such as	tangible object

【a material object such as a phonograph disc or recoding tape】

shiso	*matawa*	*kanjo*	*wo*	*sosakutekini*	*hyogenshita*	*mono*
思想	又は	感情	を	創作的に	表現した	もの(c)
thought	or	sentiment	ACC	creatively	expressed	abstract object

【a production in which thoughts or sentiments are creatively expressed】

Phrases are from the Copyright Act (Act No. 48 of 1970)

Figure 1. Phrases with a legal term (underlined)

case of the multiple-choice sentence completion test. Although language models are typically used for the general multiple-choice sentence completion test (e.g. [5,10,12,13,15]), we apply Random Forest classifiers [2] to our method. Each classifier in our method is trained and optimized for a single set of similar legal terms. We assume that this term-specializing approach brings better performance.

This paper contributes to the legal term correction task by formally defining its problem, proposing a Random Forest-based method for the problem, and showing the performance of our method compared with existing language models.

2. Japanese Legal Terms

As described in Section 1, the Japanese legislation drafting rules define a number of sets of similar legal terms and each usage. The list below displays some examples:

- "直ちに (d)" (*tadachini*), "速やかに (e)" (*sumiyakani*), and "遅滞なく (f)" (*chitainaku*)
  These are adverbs and share the concept of "speedily." In Japanese statutory sentences, these words express different degrees of speed: (d), (e), and (f) express most, moderately, and least speedy, respectively. No such strict difference among them exists in general Japanese sentences. According to the Standard Legal Terms Dictionary [14], (d), (e), and (f) should be translated to "immediately," "promptly," and "without delay," respectively.
- "前項 の 場合 に おいて (g)" (*zenko no baai ni oite*) and "前項 に 規定する 場合 に おいて (h)" (*zenko ni kiteisuru baai ni oite*)
  Both of these phrases behave as conjunctive and share the concept of "mentioning the preceding paragraph." In Japanese statutory sentences, (g) is used to mention the whole paragraph, while (h) is used to mention only the condition prescribed in the paragraph. According to the dictionary, (g) and (h) should be translated as "in the case referred to in the preceding paragraph," and "in the case prescribed in the preceding paragraph" respectively.

We note that legal terms have wide grammatical diversity: each legal term can be a noun, a verb, and so forth. Furthermore, some legal terms consist of multiple words.

3. Related Work

Since we regard legal term correction as a special case of the multiple-choice sentence completion test, we explain the test in Section 3.1. Then, we mention several studies on language models for solving the test in Section 3.2. In Section 3.3, we introduce Random Forest [2], which we utilize instead of language models in our problem.

3.1. Multiple-choice Sentence Completion Test

In the general multiple-choice sentence completion test, a sentence with a blank and choices to fill in the blank are given. The statement below represents a typical example of the test:

> He is ______ at the scoreboard.
> (A) look (B) looks (C) looking (D) looked

One must choose the best option for filling in the blank "______" (in this case, (C)). The combination of choices can vary diversely depending on the situation. For example, the sentence with a blank below can be associated with the following choices to examine verb usage.

> He is ______ at the scoreboard.
> (A) looking (B) watching (C) seeing

Therefore, a method for this problem has to cope with any combination of choices.

3.2. Language Models

In the previous situation, language models are useful because they predict a word from the whole vocabulary. To evaluate language models, Zweig and Burges [16] presented a dataset of the multiple-choice sentence completion test called the MSR Sentence Completion Challenge Data.

A variety of language models are evaluated by this dataset. First, Zweig and Burges [16] evaluated n-gram models by their dataset. Most powerful language models evaluated by this dataset have a neural network architecture, which overcomes the curse of dimension by treating each word and each sequence of words as vectors [1]. For instance, Mikolov et al. [10] proposed two neural language models: Continuous Bag-of-Words Model (CBOW) and Continuous Skip-gram Model (Skipgram). Mnih and Kavukcuoglu [12] proposed the vector Log-bilinear model (vLBL) and ivLBL. Mori et al. [13] proposed vLBL(c) and vLBL+vLBL(c), which are improved models of vLBL so that they are sensible of relative positions of words adjacent to the target word. Mirowski and Vlachos [11] proposed a recurrent neural network (RNN) [3,7] language model by incorporating the syntactic dependencies of a sentence.

While most studies propose neural language models, some propose non-neural language models. Gubbins and Vlachos [5] proposed an n-gram-like language model that handles dependency trees. Woods [15] proposed a novel method based on pointwise mutual information.

Algorithm 1 Algorithm that solves our problem

```
Input: $W, T$
Output: Suggests
  Suggests $\leftarrow \emptyset$
  for all $(i, j)$ such that $w_i\ w_{i+1} \cdots w_j = t \in T$ do
    $W_l \leftarrow w_1\ w_2 \cdots w_{i-1}$
    $W_r \leftarrow w_{j+1}\ w_{j+2} \cdots w_{|W|}$
    $t_{\text{best}} \leftarrow \arg\max_{t' \in T} \text{score}(W_l, t', W_r)$
    if $t \neq t_{\text{best}}$ then
      Suggests $\leftarrow$ Suggests $\cup$ { a suggestion that $t$ in position $(i, j)$ should be replaced into $t_{\text{best}}$}
    end if
  end for
```

3.3. Random Forest

Random Forest [2] is a kind of machine-learning algorithm for classification.

It learns the training data by building a set of decision trees, which is also called a random forest.[1] A decision tree is conceptually a suite of if-then rules. Then, a random forest predicts the class of the given data by taking a vote of each decision tree. Here, each decision tree is constructed by randomly selected data records and features. Therefore, even if a single decision tree makes an unsophisticated decision, the ensemble of decision trees predicts unseen data better.

4. Proposed Method

In this section, we describe our proposed method for legal term correction. In Section 4.1, we formally define the legal term correction problem and a general algorithm for the problem. In Section 4.2, we regard our problem as a special case of the sentence completion test and compare it with the general one. In Section 4.3, we state the way to use Random Forest for our problem and the advantages of using it.

4.1. Definition

Our method inspects legal terms found in given statutory sentences, and outputs correction ideas for some legal terms that seem to be mistakenly used. We define this task as a problem below:

- A statutory sentence $W = w_1\ w_2 \cdots w_{|W|}$ and a set of legal terms $T \subseteq V^+$ are given, where V^+ is the Kleene plus of the vocabulary V, that is, a legal term $t \in T$ can be either a word or multiple words;
- One judges whether each legal term t found in W is adequate;
- If another legal term $t_{\text{best}} \in T$ $(t_{\text{best}} \neq t)$ seems more adequate in the context, one suggests an idea that t_{best} should be placed instead of t.

We define a general algorithm for this problem in Algorithm 1, where $\text{score}(W_l, t, W_r)$ is any scoring function that calculates the likelihood of the term t when two word sequences W_l and W_r are adjacent to the left and right of t, respectively.

For example, let the statutory sentence W and the legal term set T be as follows:

[1] From here on, we call the algorithm "Random Forest" and a classifier "a random forest."

$$W = \begin{array}{cccccc} \textit{chosakubutsu} & \textit{wo} & \textit{sosakusuru} & \textit{mono} & \textit{no} & \textit{hogo} \\ \text{著作物} & \text{を} & \text{創作する} & \text{もの}_{(c)} & \text{の} & \text{保護}, \\ \text{work} & \text{ACC} & \text{create} & \text{abstract object} & \text{of} & \text{protection} \end{array} \tag{1}$$

$$T = \{\ \text{者}_{(a)}, \text{物}_{(b)}, \text{もの}_{(c)}\}. \tag{2}$$

Here, T is the legal term set mentioned in Section 1. In this case, the algorithm finds (c) from W, which is a legal term in T. Then, the algorithm processes two word sequences $W_l =$ 著作物を創作する (*chosakubutu wo sosakusuru*; creating a work) and $W_r =$ の保護 (*no hogo*; protection of). Using W_l and W_r, the algorithm calculates scores of each legal term by the following equations:

$$\text{score}\left(\begin{array}{cccccccc} \textit{chosakubutsu} & \textit{wo} & \textit{sosakusuru} & & \textit{mono} & & \textit{no} & \textit{hogo} \\ \text{著作物} & \text{を} & \text{創作する} & & \text{者}_{(a)} & & \text{の} & \text{保護} \\ \text{work} & \text{ACC} & \text{create} & , & \text{person} & , & \text{of} & \text{protection} \end{array}\right), \tag{3}$$

$$\text{score}\left(\begin{array}{cccccccc} \textit{chosakubutsu} & \textit{wo} & \textit{sosakusuru} & & \textit{mono} & & \textit{no} & \textit{hogo} \\ \text{著作物} & \text{を} & \text{創作する} & & \text{物}_{(b)} & & \text{の} & \text{保護} \\ \text{work} & \text{ACC} & \text{create} & , & \text{tangible object} & , & \text{of} & \text{protection} \end{array}\right), \tag{4}$$

$$\text{score}\left(\begin{array}{cccccccc} \textit{chosakubutsu} & \textit{wo} & \textit{sosakusuru} & & \textit{mono} & & \textit{no} & \textit{hogo} \\ \text{著作物} & \text{を} & \text{創作する} & & \text{もの}_{(c)} & & \text{の} & \text{保護} \\ \text{work} & \text{ACC} & \text{create} & , & \text{abstract object} & , & \text{of} & \text{protection} \end{array}\right). \tag{5}$$

Each calculates the likelihood of “著作物 を 創作する 者$_{(a)}$ の 保護” (Protection of a person that creates a work), “著作物 を 創作する 物$_{(b)}$ の 保護” (Protection of a tangible object that creates a work), and “著作物 を 創作する もの$_{(c)}$ の 保護” (Protection of an abstract object that creates a work), respectively. The algorithm is highly expected to choose the first option and to output a suggestion that (c) in W should be replaced into (a).

4.2. Characteristics of Problem

We regard the problem defined in Section 4.1 as a kind of sentence completion test by introducing the following ideas:

- W_l ___ W_r is the sentence with a blank, where ___ is a blank, and W_l and W_r are as defined in Algorithm 1.
- T is the choices, one of which adequately fills the blank in the sentence.

Our problem is different from the general multiple-choice sentence completion test in two ways. First, a set of choices (i.e. a legal term set) relates to many sentences with a blank. For example, each term of the legal term set $T = \{\ \text{者}_{(a)}, \text{物}_{(b)}, \text{もの}_{(c)}\}$ appears tens of thousands times in a statutory sentence corpus of nearly 29 million words that is compiled from almost four thousand acts and cabinet orders in effect in Japan. This means that we can make several hundred thousand questions for the legal term set. On the other hand, we cannot assume that such a large number of sentences relate to a set of choices in the general multiple-choice sentence completion test, since we usually consider that each sentence with a blank has a different set of choices.

Second, we can consider only meaningful legal term sets that Japanese legislation manuals mention. On the other hand, we may consider any combination of choices in the general multiple-choice sentence completion test, since there is no restriction of them.

4.3. Using Random Forests

Because of the characteristics described in the previous section, we apply Random Forest [2] to our problem. We utilize it as the scoring function score(W_l, t, W_r), which is calculated by the following equation:

$$\text{score}(W_l,t,W_r) = \sum_{d \in D} P_d(t|w_l^{|W_l|-N+1},\dots,w_l^{|W_l|-1},w_l^{|W_l|},w_r^1,w_r^2,\dots,w_r^N), \tag{6}$$

where D is a set of decision trees, d is a decision tree, and $P_d(t|w_1,w_2,\dots,w_N)$ is the probability (actually 0 or 1) that d chooses t based on features $w_1,w_2,\dots,w_N$. w_l^i and w_r^i are i-th word of W_l and W_r, respectively. N is the window size (the number of left or right adjacent words focused on).

For example, we calculate the scoring function in Equation (3) by the following equation when $N = 2$:

$$\begin{aligned} &\text{score}\left(\begin{array}{cccccc} \textit{chosakubutsu} & \textit{wo} & \textit{sosakusuru} & \textit{mono} & \textit{no} & \textit{hogo} \\ \text{著作物} & \text{を} & \text{創作する} & \text{者}_{\text{(a)}} & \text{の} & \text{保護} \\ \text{work} & \text{ACC} & \text{create} & \text{, person} & \text{, of} & \text{protection} \end{array}\right) \\ &= \sum_{d \in D} P_d \left(\begin{array}{c|cccc} \textit{mono} & \textit{wo} & \textit{sosakusuru} & \textit{no} & \textit{hogo} \\ \text{者}_{\text{(a)}} & \text{を} & \text{創作する} & \text{の} & \text{保護} \\ \text{person} & \text{ACC ,} & \text{create} & \text{, of ,} & \text{protection} \end{array}\right). \end{aligned} \tag{7}$$

Our method treats each legal term as a class. Thus, it builds a random forest for each set of legal terms.

We use random forests for three reasons. First, Random Forest classifiers for each legal term set learn from different datasets, and thus they can optimize their parameters for each set. In particular, it is useful to adjust the window size per legal term set because the distance between a legal term and its clue word can vary per legal term. On the other hand, language models learn from a single integrated dataset, and thus they use the same parameters throughout the legal terms. Second, Random forests can predict multiple-word legal terms because they equally handle any legal term as a single class. On the other hand, language models predict only a single word by given words. Therefore, we need some technique like word concatenation to predict multiple-word legal terms using a language model. Third, decision trees with naive if-then rules seem to be sufficient to predict legal terms because statutory sentences are rather more formal than general sentences since their orthography and phraseology are thoroughly regulated by the legislation drafting rules.

5. Experiment

To evaluate the effectiveness of our method, we conducted an experiment on predicting legal terms in Japanese statutory sentences.

5.1. Outline of Experiment

We compiled a statutory sentence corpus from e-Gov Statute Search[2] provided by the Ministry of Internal Affairs and Communications, Japan. We acquired 3,983 existing

[2] http://elaws.e-gov.go.jp/

Japanese acts and cabinet orders on May 18, 2018. Next, we tokenized each statutory sentence in the corpus by MeCab (v.0.996), a Japanese morphological analyzer. Statistics of the corpus are as follows: the total number of sentences is 622,527, the total number of tokens is 28,816,368, and the total number of different words is 23,236.

We defined 26 legal term sets by referencing the Japanese legislation manual [6]. Table 1 shows some examples of legal term sets. English translations in this table are taken from the Standard Legal Terms Dictionary (March 2018 edition) [14] provided by the Ministry of Justice, Japan, except for items with an asterisk.

We compared the following models with Random Forest [2]: CBOW [10], Skipgram [10], vLBL [12], vLBL(c) [13], vLBL+vLBL(c) [13], and n-gram. As for neural language models (CBOW, Skipgram, vLBL, vLBL(c), and vLBL+vLBL(c)), we set the window size to 5 in accordance with their papers. Other parameters are as follows: dimension of vectors is 200, number of epochs is 5, minibatch size is 512, number of negatively sampled words is 10 (only in Skipgram and the vLBL family), optimization function is Adam [9]. We implemented, trained and tested the models by Chainer (v.1.7.0). As for the n-gram model, we used Katz's backoff trigram and 4-gram [8] in reference to Zweig and Burges [16].

We prepared two experiment designs for Random Forest: (1) setting the window size to 5—the same as the neural methods and (2) using a variable number $\{2, 5, 10, 15\}$ of window size that is suitable for each legal term set. The window size in the latter is determined by five-fold cross validation. From here on, we call the former method "Random Forest (fixed)" and the latter method "Random Forest (variable)." In both methods, we used the Gini coefficient to build decision trees, and optimized the number of decision trees $\{10, 50, 100, 500\}$, the maximum depth of each tree $\{10, 100, 1000, \text{unlimited}\}$, and the window size (in case of (2)) by five-fold cross validation. Implementation, training, and testing are done by Scikit-learn (v.0.19.1).

Since neural language models and n-gram models are designed to predict a single word, we combined all legal terms with multiple words into single words by the longest match principle. After this operation, the total number of tokens in the corpus became 27,718,637. Also, we changed words that appear less than five times in the corpus into unknown words to reduce computational cost. In training and predicting words, we utilized an end-of-sentence token to pad short word sequences.

We divided the 3,983 acts and cabinet orders in the corpus into training data and test data. The training data has 3,784 documents, where there are 598,522 sentences and 26,707,937 tokens in total. The test data has 199 documents with 24,005 sentences and 1,010,700 tokens in total. There are 166,959 legal terms appearing in the test data.

In the evaluation, we measured accuracy of predicting legal terms in two averages: micro average acc_{micro} and macro average by legal term set acc_{macro}.

5.2. Experimental Results

Table 2 shows the experimental results of each model. As a baseline, we calculated the micro and macro averages of accuracy in maximum likelihood estimation (MLE), in which the most frequent legal terms in the train data are always selected.

Random Forest-based methods achieved the best accuracy in both the micro and macro averages. On the other hand, Skipgram was inferior to MLE in the two averages. Both Random Forest-based methods have less than 1% of gap between the two averages, while any other method has more than 2.6% of gap between the two. This means that Random Forest predicts any legal term set with high accuracy.

Table 1. Examples of legal term sets

Legal Term	Pronunciation	Meaning	Count
者 (a)	*mono*	natural or juristic person*	286,762
物 (b)	*mono*	tangible object*	24,622
もの (c)	*mono*	abstract object*	159,496
に 係る	*ni kakaru*	pertaining to*	102,924
に 関する	*ni kansuru*	regarding*	76,405
に 関係する	*ni kankeisuru*	regarding*	80
直ちに (d)	*tadachini*	immediately	2,293
速やかに (e)	*sumiyakani*	promptly	1,947
遅滞なく (f)	*chitainaku*	without delay	6,054
する こと が できる	*suru koto ga dekiru*	may	26,337
しなければ ならない	*shinakereba naranai*	must, shall	38,864
する もの と する	*suru mono to suru*	is to	8,976

6. Discussion

In this section, we investigate the experimental results in more detail to reveal the characteristics and effectiveness of our Random Forest-based methods.

First, we decompose the experimental results per part-of-speech (POS) in order to determine whether our method is good at predicting any POS of legal terms. Table 3 shows micro averages of accuracy per POS: noun, modifier, verb, and conjunction. In Table 3, Random Forest (variable) achieved the highest accuracy in nominal, verbal and conjunctional legal terms, and Random Forest (fixed) achieved the best in nominal. On the other hand, vLBL+vLBL(c) achieved the best in modifier legal terms.

It is an interesting tendency that while vLBL(c) and vLBL+vLBL(c) compare favorably with Random Forest in accuracy of nominal and modifier legal terms, they achieved worse accuracy for verbs and conjunctions. Specifically, vLBL(c) achieved 5.0% and 8.0% worse accuracy for verbs and conjunctions than Random Forest (variable), respectively, and vLBL+vLBL(c) did 6.1% and 10.8% for the same groups.

To reveal a cause for these results, we look into the accuracy per legal term. Table 4 shows the accuracy of conjunctional legal terms in vLBL+vLBL(c), Random Forest (fixed), and Random Forest (variable). Each method is denoted by “vLBL+,” “RF

Table 2. Experimental results

Method	acc_{micro}	acc_{macro}
Random Forest (fixed) [proposed]	91.8%	**91.0%**
Random Forest (variable) [proposed]	**91.9%**	**91.0%**
CBOW	86.9%	82.6%
Skipgram	72.5%	65.0%
vLBL	78.8%	76.1%
vLBL(c)	90.2%	83.0%
vLBL+vLBL(c)	90.1%	85.1%
Backoff trigram	84.9%	83.8%
Backoff 4-gram	86.6%	85.5%
MLE (baseline)	76.5%	66.4%

Table 3. Accuracy per POS

Method	Nominal	Modifier	Verbal	Conjunctional
Random Forest (fixed)	**92.7%**	92.0%	94.3%	87.9%
Random Forest (variable)	**92.7%**	92.1%	**94.7%**	**88.2%**
CBOW	87.4%	87.9%	82.9%	83.7%
Skipgram	77.0%	75.3%	51.5%	64.3%
vLBL	82.1%	78.4%	80.9%	73.6%
vLBL(c)	91.4%	92.2%	89.7%	80.2%
vLBL+vLBL(c)	92.1%	**92.4%**	88.6%	77.4%
Backoff trigram	83.3%	86.1%	89.1%	81.1%
Backoff 4-gram	86.6%	87.2%	91.3%	81.9%
MLE (baseline)	63.3%	82.5%	70.1%	78.5%

Table 4. Accuracy of conjunctional legal terms

No.	Legal term (Pronunciation; meaning)	Count	vLBL+	RF (F)	RF (V)	WS
1	又は (*matawa*; or)	9,116	74.2%	97.4%	**98.9%**	10
	若しくは (*moshikuwa*; or)	2,425	**72.9%**	38.7%	33.9%	
	Total (micro average)	11,541	73.9%	85.0%	**85.2%**	
2	及び (*oyobi*; and)	6,523	79.2%	98.9%	**99.0%**	5
	並びに (*narabini*; and)	921	**87.4%**	46.3%	45.1%	
	Total (micro average)	7,444	80.2%	**92.3%**	**92.3%**	
3	その他の (*sonotano*; other)	1,299	85.1%	91.4%	**91.6%**	5
	その他 (*sonota*; other)	1,036	80.9%	**81.9%**	81.3%	
	Total (micro average)	2,335	83.0%	**87.2%**	87.0%	
4	前項 の 場合 に おいて (*zenko no baai ni oite*; in the case referred to in the preceding paragraph)	87	63.2%	**100.0%**	98.9%	15
	前項 に 規定する 場合 に おいて (*zenko ni kiteisuru baai ni oite*; in the case prescribed in the preceding paragraph)	7	**57.1%**	28.6%	42.9%	
	Total (micro average)	94	62.8%	**94.7%**	**94.7%**	
5	ただし (*tadashi*; provided, however, that ⋯)	725	82.1%	91.4%	**96.4%**	15
	この 場合 に おいて (*kono baai ni oite*; in this case)	466	**85.8%**	82.8%	84.8%	
	Total (micro average)	1,191	83.5%	89.2%	**91.9%**	

(F),” and “RF (V)” in Table 4. In this table, “Count” means the number of the legal terms appearing in the test data, and “WS” means the optimized window size.

According to the table, Random Forest (variable) outperformed Random Forest (fixed) and vLBL+vLBL(c) in legal term set 5. Here, Random Forest (variable) set the window size to 15 for the legal term set. From this fact, we assume that Random Forest (fixed) and vLBL+vLBL(c) utilized insufficient context for the legal term set, while Random Forest (variable) could choose optimal context length. However, Random Forest has a tendency to choose frequent legal terms. For example, according to the table, it prefers to choose major legal terms “又は” (*matawa*; or) and “及び” (*oyobi*; and) from legal term sets 1 and 2, respectively.

7. Summary

In this paper, we proposed a legal term correction method in Japanese statutory sentences, focusing on sets of similar legal terms whose usages are defined in the legislation drafting rules. We regarded this legal term correction as a special case of the sentence completion test with multiple choices, considering a set of similar legal terms as the choices. Our method uses Random Forest classifiers [2], each of which is optimized for a set of similar legal terms. Our experiment has shown that our method outperformed existing modern methods for word prediction using neural language models.

In future work, we aim to improve performance by resolving the problem of frequent term preference by introducing techniques for biased datasets.

Acknowledgments

This work was partly supported by JSPS KAKENHI Grant Numbers 18H03492 and 17K00460.

References

[1] Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C.: A neural probabilistic language model. Journal of Machine Learning Research 3, 1137–1155 (2003)

[2] Breiman, L.: Random forests. Machine Learning 45, 5–32 (2001)

[3] Elman, J.L.: Finding structure in time. Cognitive Science 14(2), 179–211 (2003)

[4] Enami, T.: Rippobakuhatsu to opun gabamento ni kansuru kenkyu — horeibunsyo niokeru "opun kodingu" no teian —. Tech. rep., Fujitsu Research Institute (2015)

[5] Gubbins, J., Vlachos, A.: Dependency language models for sentence completion. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. pp. 1405–1410 (2013)

[6] Hoseishitsumu-kenkyukai: Shintei wakubukku hoseishitsumu (2nd edition). Gyosei (2018) (In Japanese)

[7] Jordan, M.I.: Serial order: a parallel distributed processing approach. Tech. Rep. ICS Report 8604, Institute for Cognitive Science, University of California. 39 pages (1986)

[8] Katz, S.M.: Estimation of probabilities from sparse data for the language model component of a speech recognizer. IEEE Transactions on Acoustics, Speech, and Signal Processing 35(3), pp. 400–401 (1987)

[9] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations. 15 pages (2015)

[10] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: International Conference on Learning Representations. 12 pages (2013)

[11] Mirowski, P., Vlachos, A.: Dependency recurrent neural language models for sentence completion. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics. pp. 511–517 (2015)

[12] Mnih, A., Kavukcuoglu, K.: Learning word embeddings efficiently with noise-contrastive estimation. In: Proceedings of the Advances in Neural Information Processing Systems 26. pp. 2265–2273 (2013)

[13] Mori, K., Miwa, M., Sasaki, Y.: Sentence completion by neural language models using word order and co-occurrences. In: Proceedings of the 21st Annual Meeting of the Association for Natural Language Processing. pp. 760–763 (2015)

[14] The Japanese Law Translation Council: Standard Legal Terms Dictionary (March 2018 Edition) (2018), `http://www.japaneselawtranslation.go.jp/dict/download`

[15] Woods, A.M.: Exploiting linguistic features for sentence completion. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. pp. 438–442 (2016)

[16] Zweig, G., Burges, C.J.: The Microsoft Research sentence completion challenge. Tech. rep., Microsoft Research (2011)

Legal Knowledge and Information Systems
M. Palmirani (Ed.)

doi:10.3233/978-1-61499-935-5-171

Efficient and Effective Case Reject-Accept Filtering: A Study Using Machine Learning

Robert BEVAN, Alessandro TORRISI, Katie ATKINSON, Danushka BOLLEGALA and Frans COENEN
Department of Computer Science, University of Liverpool, Liverpool, L69 3BX, UK
e-mail: {robert.bevan, alessandro.torrisi, k.m.atkinson, danushka.bollegala, coenen}@liverpool.ac.uk

Abstract. The decision whether to accept or reject a new case is a well established task undertaken in legal work. This task frequently necessitates domain knowledge and is consequently resource expensive. In this paper it is proposed that early rejection/acceptance of at least a proportion of new cases can be effectively achieved without requiring significant human intervention. The paper proposes, and evaluates, five different AI techniques whereby early case reject-accept can be achieved. The results suggest it is possible for at least a proportion of cases to be processed in this way.

Keywords. Early Case Accept-Reject, Heuristics, Machine Learning

1. Introduction

A common challenge for any commercial law firm is deciding whether to take on a case or not. There are many issues of significance here; however, in most cases the primary concerns are the resource required to process a case and the potential economic gains that may result. To arrive at a decision, lawyers use their experience accompanied by established heuristics. The nature of these heuristics is frequently dependent on the nature of the area in which a particular legal concern operates. One example is to reject cases where some statutory limitation is imminent, another is to reject cases which have an international aspect. Whatever the case, the established approach is time consuming, subjective and requires reference to domain experts. However, most legal enterprises have substantial repositories of previous cases. These repositories can be effectively used to determine the potential outcome of a case. There has been a substantial amount of previous work directed at the use of machine learning to build classification models that can be used to predict the legal outcome of new cases [1,2,3,4,5,6]. However, these approaches either require substantial analysis of the new case so as to identify the key features, or operate on highly structured text composed by domain experts.

In this paper it is suggested that a more efficient approach is to use a classification model founded on features extracted from information that can be readily ascertained by an administrator on first contact with a potential client. In this way, an early "definite reject" or "definite accept" decision can be made with respect to at least a proportion of cases, with the remaining cases set aside for further consideration. It is conjectured that this approach will be of particular benefit in the context of legal firms that operate in well defined domains, such as road traffic or household insurance litigation, where many of the cases received fall into a limited number of categories. More specifically, this paper considers five different techniques whereby this early reject/accept decision can be implemented. The first technique considered is a simple rule-based approach, using established heuristics, to filter incoming cases into three classes: *definite reject*, *definite accept* and *further consideration*. The second is to use an existing case repository to build a classification model, directed at the above three classes, using established machine learning algorithms [7]. The third is founded on the idea of first clustering the cases held in the repository into a small number of clusters and then building individual classification models for each cluster. The fourth is then a combination of the first and second approaches, whilst the fifth is a combination of the first and third.

2. Rule-Based Early Reject-Accept

The rule-based approach is the simplest and is founded on the observation that lawyers frequently apply heuristics to decide whether to accept or reject a case. These heuristics are designed to capture potential risks/rewards associated with a case. For example, if a case has an international element, pursuing the case may incur additional costs, reducing the potential reward. Conversely, there may exist some domain-specific green flags that indicate a high chance of success.

The business benefits derived from each heuristic depend on how frequently they are satisfied and how strongly they indicate whether a case should be accepted or rejected. With this in mind, two heuristics were chosen with respect to the experiments presented in Section 6: reject any cases with an international element, and reject any cases in which a statutory limitation will be reached within a specified time period. The international element heuristic was deemed to be satisfied if the initial contact indicated an overseas context. The limitation heuristic was implemented as follows: every date mentioned in the claimant's interview statement was extracted and ordered chronologically and the most recent date used to provide a conservative estimate of whether some limitation was imminent.

3. Single Classifier Based Early Reject-Accept

The idea underpinning the single classifier-based technique is to use the case repository, that most legal firms maintain, to build a classification model. In this work, classification features included in a telephone interview were converted into a bag-of-words representation, composed of unigrams and person/organisation entities. A large number of algorithms can be used to build such classification models. The authors experimented with a number of these and found there was little to choose between them. For the evaluation presented in Section 6, a regularized logistic regression model was used.

4. Multiple Classifier Based Early Reject-Accept

The idea underpinning the multiple classifier-based technique was that many of the cases typically received by legal firms can be categorised as being of a certain type and this categorisation is frequently an indicator of whether a case should be accepted or rejected. With respect to the third technique, the idea was thus to first cluster the cases in a given repository into a small number of clusters and to then construct individual classification models with respect to the cases in each cluster. The intuition was that this might produce more accurate predictions.

In order to effectively cluster the dataset, cases were initially converted into topic vectors, with the topics constructed in a semi-supervised fashion [8]. The topic-word distributions were then inspected, and all but the four most coherent topics were discarded (i.e. those that best represented particular case types). The reduced topic representations were then partitioned into five clusters using the k-means algorithm [9]. After clustering, cluster-specific classification models were trained using the same features as described in Section 3. Again, a regularized logistic regression model was used for the evaluation presented below.

5. Evaluation Data

The dataset used for experimentation comprised 40,000 accident claim records collected over a period of four years. Each record contained two parts: a free-text statement transcribed from a telephone interview, and a second, shorter text summarising the claim. Each record was accompanied by a label indicating whether or not the case was accepted. Note that the dataset classes were imbalanced, with the majority of cases rejected. The first technique, the rule-based technique, was applied to the entire dataset. For the classification based techniques, techniques two and three, both training and test sets were required. Thus evaluation results in these two cases were generated using Ten-Fold Cross-Validation. As the labels included in the dataset indicate only whether a case was accepted or rejected, some extra work was required in order to classify cases into the *definite reject*, *definite accept* and *further consideration* classes. For this purpose the test set class membership probabilities were sorted, and the top 10% most confident predictions for the accept and reject classes were selected as the *definite accept* and *definite reject* class predictions, the remaining 80% were assigned to the *further consideration* class. For the combination techniques, techniques four and five, heuristic filtering was applied to the test set in addition to invoking a classification model, and the heuristic filtering class predictions took precedence over the classification model predictions.

6. Results

The results obtained from the comparative evaluation are given in Table 1. The results presented are with respect to the *definite reject* and *definite accept* classes. From the table it can be seen that the rule-based approach achieved the highest precision but also produced the lowest recall by a large margin. This method failed to make any *definite accept* predictions as the chosen heuristics target the *definite reject* class only. Both machine learning approaches increased the recall at the expense of precision. No significant performance difference was observed between the general classifier and pre-clustering approaches. This may be a result of the clustering approach assigning the majority of examples to a single broad cluster, meaning the majority of test examples were evaluated using this cluster's classification model, which is nearly equivalent to the general classification model. Pre-filtering cases using the rule-based method before applying either machine learning approach improved performance. Note the disparity in performance across classes can be partially explained by the class imbalance in the dataset.

Table 1. Comparative Evaluation Results. Metrics computed separately for the *definite accept* (+) and *definite reject* (-) classes. Note the final column (MCC) refers to the Mathews correlation coefficient.

Technique	Prec (+)	Prec (-)	Rec (+)	Rec (-)	F1 (+)	F1 (-)	MCC
1. Rule-based	-	**0.946**	-	0.041	-	0.078	**-**
2. Single Classifier-Based	0.445	0.936	**0.875**	0.628	**0.59**	0.752	0.438
3. Multiple Classifier-Based	0.439	0.934	0.869	0.625	0.584	0.749	0.42
4. Rule & Single Classifier-Based	**0.449**	0.937	0.844	**0.693**	0.586	**0.797**	0.455
5. Rule & Multiple Classifier-Based	0.445	0.934	0.836	0.691	0.581	0.794	**0.457**

7. Conclusions

In this paper five alternative techniques have been proposed directed at the task, frequently undertaken by legal firms, of deciding whether to accept or reject a new case. The techniques were evaluated using an accident claims dataset of 40,000 records. The evaluation indicated it is easily possible to implement a case reject-accept system that can automatically process at least a proportion of cases. In practice, the volume of cases an automated early-rejection/acceptance system is able to process will depend on the domain and business strategy of the operator. For future work the authors intend to investigate a confidence scoring mechanism that can be attached to a particular classification and an explanation facility.

References

[1] D. Arditi and T. Pulket, "Predicting the outcome of construction litigation using an integrated artificial intelligence model," *Journal of Computing in Civil Engineering*, vol. 24, no. 1, pp. 73–80, 2009.

[2] S. Brüninghaus and K. D. Ashley, "Combining case-based and model-based reasoning for predicting the outcome of legal cases," in *Proc. International Conference on Case-Based Reasoning (ICCBR '03)*, pp. 65–79, Springer LNAI, 2003.

[3] K. Ashley and S. Brüninghaus, "Automatically classifying case texts and predicting outcomes," vol. 17, pp. 125–165, 06 2009.

[4] M. Dunn, L. Sagun, H. Şirin, and D. Chen, "Early predictability of asylum court decisions," in *Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law*, ICAIL '17, (New York, NY, USA), pp. 233–236, ACM, 2017.

[5] D. L. Chen and J. Eagel, "Can machine learning help predict the outcome of asylum adjudications?," in *Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law*, ICAIL '17, (New York, NY, USA), pp. 237–240, ACM, 2017.

[6] N. Aletras, D. Tsarapatsanis, D. Preotiuc-Pietro, and V. Lampos, “Predicting judicial decisions of the european court of human rights: a natural language processing perspective,” *PeerJ Computer Science*, vol. 2, p. e93, 2016.

[7] M. Bramer, *Principles of Data Mining*. Springer, 3 ed., 2016.

[8] Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa, “Incorporating lexical priors into topic models,” pp. 204–213, Association for Computational Linguistics, 2012.

[9] J. B. MacQueen, “Some methods for classification and analysis of multivariate observations,” p. 281297, University of California Press, 1967.

Legal Knowledge and Information Systems
M. Palmirani (Ed.)

doi:10.3233/978-1-61499-935-5-176

Dealing with Qualitative and Quantitative Features in Legal Domains

Maximiliano C. D. Budán [a,b], María Laura Cobo [d], Diego I. Martínez [a], and Antonino Rotolo [c]

[a] *Inst. for Computer Science and Eng. (UNS-CONICET), Dept. of Computer Science and Engineering, Universidad Nacional del Sur, Argentina.*
[b] *Dept. of Mathematics, Universidad Nacional de Santiago del Estero, Argentina.*
[c] *Dept. of Mathematics, University of Bologna, Italy.*
[d] *Dept. of Computer Science and Eng., Universidad Nacional del Sur, Argentina.*

Abstract. In this work we enrich a formalism for argumentation by including a formal characterization of features related to the knowledge, in order to capture proper reasoning in legal domains. We add meta-data information to the arguments in the form of labels representing quantitative and qualitative data about them. These labels are propagated through an argumentative graph according to the relations of support, conflict, and aggregation between arguments.

Keywords. Commonsense Reasoning, Argumentation Process, Qualitative and Quantitative Feature, Legal Domain

1. Introduction

Argumentation theory specializes in modeling the process of human reasoning to determine which conclusions are acceptable in a context of disagreement [1,5]. The basic idea is to identify arguments in favor and against a claim, then select the acceptable ones through a proper analysis, determining whether the original statement can be accepted or not. In general, arguments are self-contained logical constructions of defeasible deduction, and dialectical procedures of justification are defined on these constructions. There are, however, some important knowledge related to the use, nature and role of arguments that must be considered when analysing a particular case. In legal domains, there exist different non-logical features that represent different uncertain dimension . The aim of this work is to propose a legal-inspired enrichment of argumentation frameworks by using decorations to arguments in the form of *labels*. These labels are integrated and transformed through the argumentation analysis according to its intended meaning.

Consider the following legal case about Medically Assisted Reproduction where an agent must decide whether it is appropriate this procedure for a specific couple.

$\mathbb{A}$ *The couple can generate children, then they are not sterile. Furthermore, there exists reasons to think that the couple can grants a reasonably expectancy of life for the child. In conclusion, the couple can not perform the medically assisted reproduction.*

$\mathbb{B}$ *The couple is affected by a serious genetic disease. However, they enjoy psychological well-being. Also, they have access to medically assisted reproduction techniques due to the fact that the woman do not have any physical impediment. Thus, there is a legal solution for the reproduction problems.*

$\mathbb{C}$ *The couple has a stable economic position to provide health and education for the future child. Thus, there is a legal solution for the reproduction problems.*

The facts of the case are: the couple is able to conceive and generate children, but they are both carriers of a severe genetic disease, which does not allow children to live for more than a few years. This couple has a good economic position an they enjoys psychological well-being.

In order to reach a conclusion for a particular decision, certain features of the knowledge can be taken into account. For instance, the *relevance* of the information representing that some pieces of knowledge may be more pertinent than others. Also the *intuition* behind the knowledge is important. Note that these two features are of different nature: one of them (relevance) is a measure usually expressed with real numbers, while the other (intuition) is just a label denoting the implicit or explicit intuition associated with the argument. Hence, the former is a quantitive feature about the knowledge while the latter is a qualitative one. Our intuition behind the combination of qualitative and quantitative features associated with arguments is that, in most cases, these features are complementary or with a strong dependency between each other and they are extremely important for legal decision making. Thus, by combining these kind of information in the argumentation process, an essential part of legal debates is captured.

In this work we propose an extension of *Labeled Argumentation Frameworks* (LAF) [2,3] in the legal domain that allows the representation of qualitative and quantitative features associated to the arguments involved in a legal dispute. In particular, labels are combined and propagated through an argumentation graph according to the manner in which the interactions between arguments are defined: support, conflict, and aggregation. Once the propagation process is complete, with the *definitive* argumentation labels, we establish the acceptability status of arguments by using the information on these labels.

2. Labeled Argumentation Framework

Labels represent quantitative and qualitative domain-dependant information about the arguments. We define an *Algebra of Argumentation Labels* as an abstract algebraic structure that contains the operations related to manipulation of arguments. The effect of aggregation, support, and conflict of arguments will be reflected in their labels, evidencing how the arguments have been affected by their interaction.

Definition 1 *An algebra of argumentation labels is a tuple* $\mathsf{A} = \langle A, \leq, \odot, \oplus, \ominus, \top, \bot \rangle$, *where* A *is a set of labels called the* domain of labels, $\leq$ *is a partial order over* A *with* $\top$ *and* $\bot$ *two distinguished elements of* A $\odot : A \times A \to A$ *is a* support operation, $\oplus : A \times A \to A$ *is an* aggregation operation, *and* $\ominus : A \times A \to A$ *is a* conflict operation.

A natural way of representing *quantitative* information is to use a numeric scale. We will consider fuzzy valuations ranging between two distinguished elements: $\top$ and $\bot$, where $\bot$ represents the less possible degree, while $\top$ is the maximum degree. Regarding *qualitative* features, we are interested here in being as general as possible using *order theory*, as a mean for describing statements such as "*this is less than that*" or "*this precedes that*". The appropriate behavior in the argumentation domain is defined by properties associated to the operators.

We extend a formalism called *Labeled Argumentation Framework* (LAF) that combines the knowledge representation features provided by the *Argument Interchange For-*

mat with the processing of meta information using the *Algebra of Argumentation Labels*. This framework allows a good representation of arguments, by taking into account their internal structure, the argument interactions, and special features of the arguments. Labels then are propagated and combined through the argumentation process using the algebra operators. The final label attached to each argument is obtained, and this information is used to establish the status of acceptance of these arguments. In LAF, we use the AIF ontology as the underlying knowledge representation model for the internal structure of the arguments and its relations [4,3].

Definition 2 *A Labeled Argumentation Framework* (LAF) *is a tuple of the form* $\Phi = \langle \mathscr{L}, \mathscr{R}, \mathscr{K}, \mathscr{A}, \mathscr{F}_{\mathscr{K}} \rangle$ *where:* $\mathscr{L}$ *is a logical language for knowledge representation;* $\mathscr{R}$ *is a set of inference rules* $R_1, R_2, \ldots, R_n$ *defined in terms of* $\mathscr{L}$*;* $\mathscr{K}$ *is the knowledge base, a set of formulas of* $\mathscr{L}$ *describing the knowledge about a specific domain of discourse;* $\mathscr{A}$ *is a set of algebras of argumentation labels* $\mathsf{A}_1, \mathsf{A}_2, \ldots, \mathsf{A}_n$*, one for each feature that will be represented by the labels; and* $\mathscr{F}_{\mathscr{K}}$ *is a function that assigns to each element of* $\mathscr{K}$ *a n-tuple of elements in the algebras* $\mathsf{A}_i, i = 1, \ldots, n$.

Example 1 *The running example about assisted fertility can be formalized by the following LAF:*

– $\mathscr{L}$ *is a language defined in terms of two disjoint sets: a set of presumptions and a set of defeasible rules, where a presumption is a ground atom X or a negated ground atom* $\sim$*X, a defeasible rule is an ordered pair, denoted* $\mathtt{C} \mathrel{-\!\!\prec} \mathtt{P}_1, \ldots, \mathtt{P}_n$*, where* $\mathtt{C}$ *is a ground atom (conclusion) and* $\mathtt{P}_1, \ldots, \mathtt{P}_n$ *is a finite non-empty set of ground atoms (premises).*

– $\mathscr{R}$= $\{$ *dMP*$\}$*, denoting Defeasible Modus Ponens:* *dMP:* $\frac{\mathtt{P}_1,\ldots,\mathtt{P}_n \quad \mathtt{C} \mathrel{-\!\!\prec} \mathtt{P}_1,\ldots,\mathtt{P}_n}{\mathtt{C}}$ *–*$\mathscr{A}$ =

$\{\mathsf{A}, \mathsf{B}\}$ *the set of algebras of argumentation labels where:*

A *represents the trust degree attached to arguments. The domain of labels A is the real interval* $[0,1]$ *representing a normalized relevance valuation of the information.*

$\alpha \odot \beta = \alpha\beta$*, where the relevance of a conclusion is based on the conjunction of the relevances corresponding to its premises*
$\alpha \oplus \beta = \alpha + \beta - \alpha\beta$*, where if there is more than one argument for a conclusion, its relevance valuation is the sum of the valuations of the arguments supporting it, with a penalty term.*
$\alpha \ominus \beta = max(\alpha - \beta, 0)$*, reflects that the relevance valuation of a conclusion is weakened by the relevance in its contrary.*

B *is an algebra of argumentation labels representing the intuition attached to arguments. The domain of labels B is the set* $\{$ *PL (Preserve Life), NG (No Eugenesis), FCH (Preserve the Future of the Child), PCH (Preserve the Couple Health),* $\emptyset$ *(without intuition)*$\}$ *representing the intuition behind an argument, where the order of these elements is* $PL > NG > FCH > PCH > \bot$.

$\alpha \odot \beta = min(\alpha \cup \beta)$*, models that the governing intuition for an argument is based on the less influential intuition associated to the arguments that support it (conditional instantiation).*
$\alpha \oplus \beta = \alpha \cup \beta$*, where if there is more than one argument for a conclusion, its governing intuitions are the composition of the predominant intuition proposed by the arguments supporting it.*
$\alpha \ominus \beta = \alpha \setminus \beta$*, reflects that only the unquestionable intuition are preserved.*

- $\mathscr{K}$ *is composed by the set of norms representing the interpretation of the provisions of the legal systems involved in the case, and the intuition behind it. We display below the set of formulas of* $\mathscr{L}$ *forming* $\mathscr{K}$*, where labels (between brackets), denote relevance and intuition behind the formula:*

$$\left\{ \begin{array}{l} \mathtt{r}_1 : \mathtt{Med\_Repr(X)} \mathrel{-\!\!\prec} \sim\!\mathtt{Physical\_Imp(X)} : \{\mathtt{07,PL}\} \\ \mathtt{r}_2 : \sim\!\mathtt{Steril(X)} \mathrel{-\!\!\prec} \mathtt{Genere\_Child(X)} : \{\mathtt{0.7,NE}\} \end{array} \right\}$$

$$\left\{\begin{array}{l} \mathtt{n}_1 : \mathtt{Sol\_Rep\_Prob(X)} \mathrel{-\!\!\prec} \mathtt{Genetic\_Dis(X),Med\_Repr(X)} : \{\mathtt{1,PL}\} \\ \mathtt{n}_2 : \sim\!\mathtt{Med\_Repr(X)} \mathrel{-\!\!\prec} \mathtt{Rsn\_Exp\_Life(X)},\sim\!\mathtt{Steril(X)} : \{\mathtt{0.8,NE}\} \\ \mathtt{n}_3 : \mathtt{Sol\_Rep\_Prob(X)} \mathrel{-\!\!\prec} \mathtt{Edu\_Health\_Child(X)} : \{\mathtt{0.5,FCH}\} \end{array}\right\}$$

$$\left\{\begin{array}{ll} \sim\!\mathtt{Physical\_Imp(CP)} : \{\mathtt{0.8,PL}\} & \mathtt{Genere\_Child(CP)} : \{\mathtt{0.8,NE}\} \\ \mathtt{Rsn\_Exp\_Life(CP)} : \{\mathtt{0.5,PL}\} & \mathtt{Genetic\_Dis(CP)} : \{\mathtt{1,PL}\} \\ \mathtt{Edu\_Health\_Child(CP)} : \{\mathtt{0.6,FCH}\} & \end{array}\right\}$$

We use *argumentation graphs* to represent the argumentative analysis derived from an *LAF*. We assume that nodes are named with different sentences of $\mathscr{L}$, so we will use the naming sentence to refer to the I-node in the graph.

Definition 3 *Given an LAF Φ, its associated argumentation graph is the digraph $G = (N,E)$, where $N \neq \emptyset$ is the set of nodes and E is the set of the edges, constructed as follows: (i) each element $X \in \mathscr{K}$ or derived from $\mathscr{K}$ through $\mathscr{R}$, is represented by an I-node $X \in N$. (ii) for each application of an inference rule defined in Φ, there exists an RA-node $R \in N$ such that: (a) the inputs are all I-nodes $P_1,\ldots,P_m \in N$ representing the premises necessary for the application of the rule R; and (b) the output is an I-node $Q \in N$ representing the conclusion. (iii) if X and $\overline{X}$ are in N, then there exists a CA-node with edges to and from both I-nodes X and $\overline{X}$. (iv) for all $X \in N$ there does not exist a path from X to X in G_Φ that does not pass through a CA-node.*

Once the argumentation graph is obtained, we proceed to attach a label to each I-node, representing the features referring to extra information that we want to represent.

Definition 4 *Let Φ be an LAF, and G be its corresponding argumentation graph. Let A_i be one of the algebras in $\mathscr{A}$. A labeled argumentation graph is an assignment of two valuations from each of the algebras to all I-nodes of the graph, denoted with μ_i^X and δ_i^X, where μ_i^X accounts for the aggregation of the reasons supporting the claim X, while δ_i^X displays the state of the claim after taking conflict into account, such that μ_i^X, $\delta_i^X \in A_i$. If X is an I-node, its valuations are determined as follows: (i) If X has no inputs, then its accrued valuation is given by function $\mathscr{F}$; thus, $\mu_i^X = \mathscr{F}_i(X)$. (ii) If X has an input from a CA-node representing conflict with $\overline{X}$, then: $\delta_i^X = \mu_i^X \ominus \mu_i^{\overline{X}}$. If there is no input from a CA-node then: $\delta_i^X = \mu_i^X$. (iii) If X is an element of $\mathscr{K}$ with inputs from RA-nodes $R_1,\ldots,R_k$, where each R_s has premises $X_1^{R_s},\ldots,X_{n_s}^{R_s}$, then: $\mu_i^X = \mathscr{F}_i(X) \oplus [\oplus_{s=1}^{k}(\odot_{t=1}^{n_s}\delta_i^{X_t^{R_s}})]$. If X is not an element of $\mathscr{K}$ and has inputs from RA-nodes $R_1,\ldots,R_k$, where each R_s has premises $X_1^{R_s},\ldots,X_{n_s}^{R_s}$, then: $\mu_i^X = \oplus_{s=1}^{k}(\odot_{t=1}^{n_s}\delta_i^{X_t^{R_s}})$.*

This procedure determines the system of equations representing the constraints that all features must fulfill. Once the I-nodes of the argumentative graph are labeled, we can determine the acceptability status associated with an argument.

Definition 5 *Let Φ be an LAF. For each of the algebras A_i in $\mathscr{A}$, representing a feature to be associated with each I-node X. Then, X has assigned one of four possible acceptability statuses with respect to A_i:* `Assured` *if and only if $\delta_i^X = \top_i$ or $\top_i \in \delta_i^X$;* `Unchallenged` *if and only if $\mu_i^X = \delta_i^X \neq \bot_i$;* `Weakened` *if and only if $\bot_i < \delta_i^X < \mu_i^X$; and* `Rejected` *if and only if $\delta_i^X = \bot_i$.*

Finally, for each claim, we form a vector with the acceptability of that claim with respect to each of the attributes, and take the least degree of those that appear in the vector as the acceptability degree for the claim as a whole.

Example 2 *We show in Figure 1 the solution to EQS, which represent the model of the instantiated LAF. For reasons of readability and space, we do not include the equations that determine the features of the graph.*

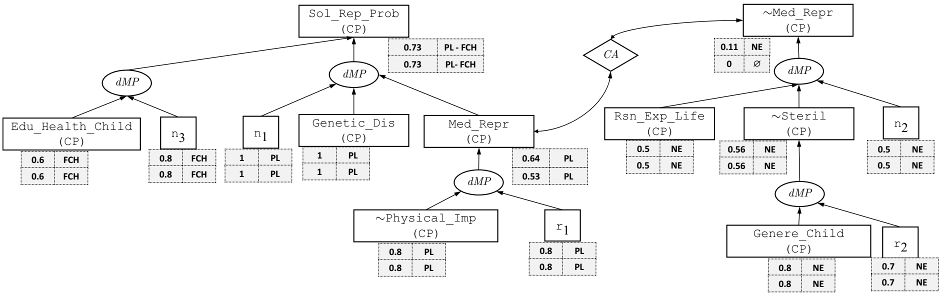


Figure 1. Labeled for an Argumentation graph

We calculate the acceptability status for each claim, considering the valid labeling for the argumentation graph that describes the discussion about medically assisted reproduction. Thus, we have that $S_g^A = \{\texttt{Genetic\_Dis(CP)}, \texttt{n}_1\}$*,* $S_g^U = \{\sim\texttt{Physical\_Imp(CP)}, \texttt{Edu\_Health\_Child(CP)}, \texttt{Genere\_Child(CP)}, \texttt{Rsn\_Exp\_Life(CP)}, \sim\texttt{Steril(CP)}, \texttt{n}_2, \texttt{n}_3, \texttt{r}_1, \texttt{r}_2\}$*,* $S_g^W = \{\texttt{Med\_Repr(CP)}\}$*, and* $S_g^R = \{\sim\texttt{Med\_Repr(CP)}\}$*. Finally, the medically assisted reproduction is legally accepted with a relevance of 0.72/), and a intuition of preserve the life and the future of the child.*

3. Conclusions

In this work we present an extension of *Labeled Argumentation Framework* that combines the KR capabilities provided by the *Argument Interchange Format* (*AIF*), together with the management of labels through properly defined algebras. We have associated operations in an algebra of argumentation labels to three different types of argument interactions, allowing to propagate qualitative and quantitative information on the argumentation graph. We defined a process to determine the status of acceptance of arguments, by integrating the information shown in labels. The conflict operation defined in the algebra allows the weakening of arguments, which contributes to a better representation of application domains. In our running example, labels are used to characterize relevance and the intuition associated with the arguments supporting decision making. These qualitative and quantitative features are the key to provide a properly, founded decision by providing awareness of different aspects of the knowledge exposed.

References

[1] Philippe Besnard and Anthony Hunter. *Elements of Argumentation*. MIT Press, 2008.

[2] Maximiliano CD Budán, Mauro Gómez Lucero, Carlos Chesñevar, and Guillermo R Simari. Modeling time and valuation in structured argumentation frameworks. *Information Sciences*, 290:22–44, 2015.

[3] Maximiliano Celmo Budán, Gerardo I. Simari, Ignacio Darío Viglizzo, and Guillermo Ricardo Simari. An approach to characterize graded entailment of arguments through a label-based framework. *Int. J. Approx. Reasoning*, 82:242–269, 2017.

[4] Carlos Iván Chesñevar, Jarred McGinnis, Sanjay Modgil, Iyad Rahwan, Chris Reed, Guillermo R. Simari, Matthew South, Gerard Vreeswijk, and Steven Willmott. Towards an argument interchange format. *Knowledge Eng. Review*, 21(4):293–316, 2006.

[5] Iyad Rahwan and Guillermo R. Simari. *Argumentation in Artificial Intelligence*. Springer Verlag, 2009.

Legal Knowledge and Information Systems
M. Palmirani (Ed.)

doi:10.3233/978-1-61499-935-5-181

Linguistic Legal Concept Extraction in Portuguese

Alessandra Cid [a] Alexandre Rademaker [b] Bruno Cuconato [c] Valeria de Paiva [d]

[a] *FGV/Direito Rio and FGV/EMAp*
[b] *IBM Research and FGV/EMAp*
[c] *FGV/EMAp*
[d] *Nuance Communications*

Abstract. This work investigates legal concepts and their expression in Portuguese, concentrating on the “Order of Attorneys of Brazil” Bar exam. Using a corpus formed by a collection of multiple-choice questions, three norms related to the Ethics part of the OAB exam, language resources (Princeton WordNet and OpenWordNet-PT) and tools (AntConc and Freeling), we began to investigate the concepts and words missing from our repertory of concepts and words in Portuguese, the knowledge base OpenWordNet-PT. We add these concepts and words to OpenWordNet-PT and hence obtain a representation of these texts that is mostly “contained” in the lexical knowledge base.

Keywords. wordnet, law, legal informatics, lexical resources

1. Introduction

The “Order of Attorneys of Brazil” (in Portuguese ‘Ordem dos Advogados do Brasil’ or OAB), the Brazilian Bar association, administers a bar examination nationwide three times a year. The exam is divided in two stages. The first consists of 80 multiple choice questions covering several disciplines. The candidate must score at least 40 questions correctly to proceed to the second part of the exam. Success in the examination allows one to practice in any court or jurisdiction of the country.

We would like to use Natural Language Processing (NLP) tools to develop a computer system capable of providing question-answering facilities, based on Brazilian laws and regulations. An ideal legal system would take a question `Q` in natural language and a corpus of all legal documents in a given jurisdiction `LawCorpus`, and would return both a correct answer (easier if using multiple choice) and its legal foundation. However, this is too broad and too hard: we hope to provide a sample corpus (a subset of `LawCorpus`) with a single processed law, to see how far we can get the processing done.

Previous work [6] on a corpus constructed from multiple choice questions, attests to the suitability of the data obtained from the OAB Bar questions. The data from OAB’s previous exams and their answer keys were cleaned and prepared for processing, and a simple question answering system, targeting the exams, based on shallow NLP methods was described. The work in [7] improved the system by incorporating wordnet [1] data to its analysis, and started a preliminary effort to expand OpenWordnet-PT (OpenWN-PT or simply OWN-PT), our basic lexical resource, to the legal domain.

The expansion of a wordnet with legal terms was also investigated by [15] where legal vocabulary was added to the Italian Wordnet (ItalWordNet). Unfortunately, we could not get access to the final resource.

This work follows [7]. It is clear from inspection that the legal domain has many concepts and words that are only used within the legal profession. These concepts and words need to be added to OWN-PT, described below if this is to be used to reason about the law.[2]

2. OpenWordNet-PT

OpenWordnet-PT [5] is an open access wordnet for Portuguese, originally developed as a syntactic projection of Universal WordNet (UWN) [3]. OWN-PT has been constantly improved through linguistically motivated additions, manual and semi-automatic. Most of this work has been based on grammatical functions: we improved the verb lexicon [4], we provided nominalizations and their links to verbs [10], and we increased demonyms and gentilics [13], which was meant to break down the huge class of adjectives into smaller subsets. Regarding specialized domains, we did a preliminary study of Geological Eras [11], and here we are tackling another sophisticated and specialized field, the Law domain.

In order to deal with legal texts we need to expand OWN-PT with legal terms and *multiword expressions* (MWEs), that describe the field, but this is known to be a hard problem in linguistics [14]. Some of these are in Latin, such as *habeas corpus* or *data venia*. But most others are simply common Portuguese words, used in fixed expressions, which have more specific meanings. For example the expression *defensor público* could be used for someone who defends the public or someone who defends something in public, but it is mostly used to describe the attorney, appointed by the Estate to defend the interests of poor citizens, who are not able to pay for a lawyer. Some recent work, especially on English noun compounds [8], makes the point that MWEs can be compositional or non-compositional, conventionalized and not conventionalized.

It is clear that specific domains like Law require a big set of MWEs, both compositional (or not) and conventionalized (or not). Briefly we can say that *semantic non-compositionality* is the property of a compound whose meaning can not be readily interpreted from the meanings of its components. *Conventionalization* refers to the situation where a sequence of words that refer to a particular concept is commonly accepted in such a way that its constituents cannot be easily substituted for near-synonyms, because of some cultural or historical conventions. A large fraction of compounds are to some extent conventionalized, however we are interested in clear and well-known conventionalizations, which [8] refer to as “marked conventionalization”. We assume that non-compositional compounds are by definition conventionalized, hence it only makes sense to consider conventionalization (or not) of compositional compounds.

[1] A wordnet is a lexical database that groups nouns, verbs, adjectives and adverbs into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations.

[2] A longer version of this article was deposited in the `arXiv.org` [2].

3. Experiments

In order to identify relevant legal terms and MWEs and to analyze how this legal vocabulary can be incorporated to the OWN-PT, we describe three small experiments.

In our first experiment, we investigated the English terms in the Princeton WordNet (PWN) [9] synsets that were 'classified by' the synset {`08441203-n`: *jurisprudence, law*} in PWN and OWN-PT.[3] Our hypothesis was that, by translating the English terms that were already classified as legal vocabulary, we would incorporate important legal terms in Portuguese to the OWN-PT. We analyzed the terms that were on the topic, verifying the quality of the translations and seeing if we could translate the ones that were not. We reached the conclusion that the synsets related to {`08441203-n`: *jurisprudence, law*} were very specific to American Law and that by adding their translations to the OWN-PT we were not expanding it with relevant words for legal vocabulary in Portuguese. For example, we found several expressions for specific types of laws in English, such as "Gag Law" or "Blue sky law", that are not used in Portuguese. Given this situation, we moved on to our second experiment.

Our second experiment deals with a wholesale construction of a glossary of legal terms extracted from the OAB questions and from the three norms that are the base of the Ethics questions of the OAB exam. These three norms are: the Law 8906 of July of 1994, the 'Código de Ética da OAB' (the Ethics Code of the OAB) and the 'Regulamento Geral da OAB' (OAB's General Regulation). We analyzed these documents using AntConc [1], a corpus analysis toolkit for concordancing and text analysis. Using AntConc, we obtained a list of 6,890 bi-grams and tri-grams on the texts that occur more than 9 times. AntConc works over the raw text, without any linguistic annotation, and we have to filter n-grams that were clearly not MWEs. Two annotators filtered the list independently and we combined the results ending up with 430 MWEs candidates.

Instead of deciding which n-grams are true MWEs as opposed to simple collections of words that occur together, we used a simple test to classify each candidate as compositional or non-compositional and conventional or non-conventional. Is the meaning of this expression explained by the meanings of its parts? If not, then we think we have a non-compositional MWE. If the meaning of the expression is compositional, is it a title of an article in the Portuguese Wikipedia?[4] If yes, we reckon this is sufficient evidence to characterize a conventional MWE. If it is not a Wikipedia title, it may be that Wikipedia should have one such page and is missing it. Therefore, our process is an oversimplification that could be improved in the future. Finally, we identified the head words from expressions and added them to the proper synsets in OWN-PT, when they exist. If a head word suggest a concept that does not exist, we create a new synset in OWN-PT, placing it in the right position of the network of concepts, and assign the word to it. In both cases, the expression is finally added to a new synset, hyponym of the synset where its head word was added.[5]

[3] PWN contains many semantic relations between synsets, besides the most well-known `hyponym`, `hypernym`, and `antonym`, we also have the relation `classifiedByTopic` for grouping synsets into domains. All OWN-PT synsets have 1-1 mappings to Princeton WordNet synsets. Data is available at `http://wnpt.brlcloud.com/wn/`.

[4] We obtained a list of titles of all Portuguese pages from Wikipedia at `https://dumps.wikimedia.org/other/`.

[5] The list of MWEs and all data from the experiments will be made available at `http://github.com/own-pt/`.

	total	unique	no sense
Nouns	2629	727	190
Adjectives	634	234	60
Verbs	1167	330	16
Adverbs	268	77	32

Table 1. Analysis of Law 8906 by Freeling

In our third experiment, we investigated the lexical units of the Law 8906, one of the norms used in the second experiment. Since the Ethics part of the Bar examination is one of the most straightforward sections of the exam, it makes sense for us to make sure that the whole law is processed correctly and that all the required vocabulary is in place, before trying to relate the OAB ethics questions to their answers and justifications.

The experiment was carried out using Freeling [12], a well-known NLP library to analyze Brazilian Portuguese. We processed the Law 8906, investigating the results of the tokenization, lemmatization, part-of-speech (PoS) tagging and word sense disambiguation. We checked if all the content words are assigned to OWN-PT senses in the context of the articles of the law. This allowed us to evaluate how Freeling's modules could be adapted to process the law more accurately and enabled us to measure how many words belonging to the legal vocabulary were already on OWN-PT or needed to be added.

Some of Freeling's results after processing the law were expected. Since OWN-PT, just as PWN, does not cater for pronouns, determiners or prepositions, it did not have a meaning assignment for these cases. Freeling's lemmatization and PoS tagging modules are driven by a dictionary of word forms. The words that are not in Freeling's dictionary must have the lemma and part-of-speech tag guessed, which introduces some errors. For example, the Portuguese word *juizado* (court) was not in the dictionary, so the lemmatization of *juizados*, was wrongly ascribed as *juizados*. This was evidence that FreeLing's dictionary did not have it and we simply added it. The multiword expressions identified and added to OWN-PT must also be added to the Freeling locutions file. Other bugs are still under investigation, but the results obtained so far are summarized in Table 1, where we present basic statistics of Freeling's analysis of Law 8906. To obtain the unique totals we considered pairs (lemma, PoS tag), and we only considered that a word was missing a sense if it was tagged as the right PoS tag. Law 8906 comprise 87 articles summing up to 231 sentences and 10,242 tokens (1,508 unique types/words). Table 1 shows in the last column that we are still missing some words in OWN-PT.

4. Conclusion

This preliminary work investigates legal concepts and their expression in Portuguese. Using the corpus formed by the collection of multiple-choice questions in the exams, three ethics norms, language resources and NLP tools we began to investigate the concepts missing from our repertory of concepts and words in Portuguese, the knowledge base OWN-PT.

As for future work, we need to complete the expansion of OWN-PT that we started constructing. When the mappings are consistently investigated, we need to establish a process to make sure that newer changes do not undermine the previous work, i.e. we need to establish test suites and regression tests. Finally we would like also to design and implement our own system for computing "entailment and contradiction detection" between the OAB examination questions and their answers and justifications (segments of text, or spans, in the laws that justify the answer of the question).

References

[1] Laurence Anthony. Antconc (version 3.5.7) [computer software], 2018. Available from http://www.laurenceanthony.net/software.

[2] A. Cid, A. Rademaker, B. Cuconato, and V. de Paiva. Linguistic Legal Concept Extraction in Portuguese. *ArXiv e-prints*, October 2018. https://arxiv.org/abs/1810.09379.

[3] Gerard De Melo and Gerhard Weikum. Towards a Universal WordNet by learning from combined evidence. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 513–522. ACM, 2009.

[4] Valeria de Paiva, Fabricio Chalub, Livy Real, and Alexandre Rademaker. Making virtue of necessity: a verb lexicon. In *PROPOR – International Conference on the Computational Processing of Portuguese*, Tomar, Portugal, 2016.

[5] Valeria de Paiva, Alexandre Rademaker, and Gerard de Melo. Openwordnet-pt: An open Brazilian Wordnet for reasoning. In *Proceedings of COLING 2012: Demonstration Papers*, pages 353–360, Mumbai, India, December 2012. The COLING 2012 Organizing Committee. Published also as Techreport, http://hdl.handle.net/10438/10274.

[6] Pedro Delfino, Bruno Cuconato, Edward Hermann Haeusler, and Alexandre Rademaker. Passing the Brazilian OAB Exam: Data preparation and some experiments. In Adam Wyner and Giovanni Casini, editors, *Legal Knowledge and Information Systems*, volume 302 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, 2017. 30th International Conference on Legal Knowledge and Information Systems (JURIX 2017). Expanded version at https://arxiv.org/abs/1712.05128.

[7] Pedro Delfino, Bruno Cuconato, Guilherme Paulino Passos, Gerson Zaverucha, and Alexandre Rademaker. Using OpenWordnet-PT for Question Answering on Legal Domain. In *Global Wordnet Conference 2018*, Singapore, January 2018. to appear.

[8] Meghdad Farahmand, Aaron Smith, and Joakim Nivre. A multiword expression data set: Annotating non-compositionality and conventionalization for english noun compounds. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 29–33, 2015.

[9] Christiane Fellbaum. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, 1998.

[10] Cláudia Freitas, Valeria de Paiva, Alexandre Rademaker, Gerard de Melo, Livy Real, and Anne de Araujo Correia da Silva. Extending a lexicon of portuguese nominalizations with data from corpora. In Jorge Baptista, Nuno Mamede, Sara Candeias, Ivandré Paraboni, Thiago A. S. Pardo, and Maria das Graças Volpe Nunes, editors, *Computational Processing of the Portuguese Language, 11th International Conference, PROPOR 2014*, São Carlos, Brazil, October 2014. Springer.

[11] Henrique Muniz, Fabricio Chalub, Alexandre Rademaker, and Valeria de Paiva. Extending wordnet to geological times. In *Global Wordnet Conference 2018*, Singapore, January 2018. to appear.

[12] Lluís Padró and Evgeny Stanilovsky. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May 2012. ELRA.

[13] Livy Real, Valeria de Paiva, Fabricio Chalub, and Alexandre Rademaker. Gentle with gentilics. In *Joint Second Workshop on Language and Ontologies (LangOnto2) and Terminology and Knowledge Structures (TermiKS) (co-located with LREC 2016)*, Slovenia, May 2016.

[14] Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. Multiword Expressions: a pain in the neck for NLP. In *Conference on Intelligent Text Processing and Computational Linguistics*, pages 1–15, Heidelberg, 2002. Springer Berlin.

[15] Maria Teresa Sagri, Daniela Tiscornia, and Francesca Bertagna. Jur-wordnet. In *Proceedings of the 2nd International Global Wordnet Conference*, pages 305–310. Citeseer, 2004.

Legal Knowledge and Information Systems
M. Palmirani (Ed.)

doi:10.3233/978-1-61499-935-5-186

A Tool to Highlight Weaknesses and Strengthen Cases: CISpaces.org

Federico CERUTTI [a], Timothy J. NORMAN [b] and Alice TONIOLO [c]
[a] *Cardiff University, UK*
[b] *University of Southampton, UK*
[c] *University of St Andrews, UK*

Abstract. We demonstrate CISpaces.org, a tool to support situational understanding in intelligence analysis that complements but not replaces human expertise, for the first time applied to a judicial context. The system combines argumentation-based reasoning and natural language generation to support the creation of analysis and summary reports, and to record the process of forming hypotheses from relationships among information.

Keywords. argumentation, case analysis, report generation

1. Introduction

We demonstrate the application of CISpaces.org [15,14,2], Collaborative Intelligence Spaces Online, in judicial context. CISpaces.org is a suite of tools and algorithms for the support of sense-making of complex situations, complementing human expertise, and for the dissemination of natural language reports. This web-based tool builds on top of argumentation-based systems, combining a structured graphical representation of the reasoning process with efficient algorithms for the automated identification of plausible hypotheses.

We demonstrate how CISpaces.org supports the data-to-decision process, from hypotheses formation, to report generation that can be used for briefings to inform legal practitioners, or even judges. CISpaces.org facilitates sensemaking in a declarative format. Differently from existing tools [7,6], CISpaces.org provides a method to record and support the process of forming hypotheses from the relationships among information which enables the analyst to highlight information or assumptions that may lead to interrelated as well as alternative hypotheses. CISpaces.org makes this core process of reasoning explicit, providing further support for structuring reasoning and mitigating biases. The reasoning mechanism identifies what evidence and claims together constitute a plausible interpretation of an analysis.

CISpaces.org is freely available for being used at `http://tiny.cc/CISpaces` (username: demo, password: demo), and it can be downloaded at GitHub, `https://github.com/CISpaces`, with MIT licence.

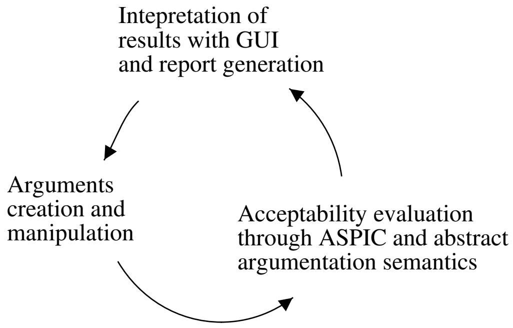


Figure 1. Overview of CISpaces.org functionalities.

2. Components

CISpaces.org builds on top of (1) argumentation schemes; (2) formal argumentation; and (3) natural language generation techniques (cf. Figure 1).

2.1. Argumentation Schemes

Argumentation schemes [16] are abstract reasoning patterns. Schemes have been derived from empirical studies of human argument and debate, and further adapted in this work from literature and experts [15]. Each scheme has a set of critical questions that represents standard ways of critically probing into an argument to find aspects of it that are open to criticism. For instance, the following is the argumentation scheme for argument from cause *A* to effect *B*:

Major Premise: Generally, if A occurs, then B might occur.
Minor Premise: In this case A might have occurred.
Conclusion: Therefore, in this case B might have occurred.

Critical questions are:

CQ1: Is there evidence for C to occur?
CQ2: Is there a general rule for C causing E?
CQ3: Is the relationship between A and B causal?
CQ4: Is there any exception to the causal rule that prevents B from occurring?
CQ5: Has A happened before B?
CQ6: Is there any other cause A' that might have caused B?

The purpose of schemes in CISpaces.org is to guide analysts in drawing inferences, critical questions are available to analysts as a means to reflect on potential issues during the formation of hypotheses. Instantiated schemes can be mapped to the overall ASPIC+ framework following the approach proposed in [11].

2.2. ASPIC+ [10]

ASPIC+ [10] is a formal framework able to transform logical statements and logical rules into arguments. Due to space constraints, we refer an interested reader to the relevant literature.

ASPIC+ uses Dung's abstract argumentation framework [5] to compute the acceptability status of arguments and, consequently logical statements. CISpaces.org uses state-of-the-art algorithms for computing such acceptability statuses [4].

2.3. Natural Language Generation

The use of graphical models to represent arguments is the most common approach used in the formal argumentation community to capture argument structures [3,1]. This requires a significant level of training that cannot be assumed for the recipients of intelligence analysis, viz. decision makers such as group commanders. To this reason, CISpaces.org has been equipped with a Natural Language Generation system. A Natural Language Generation (NLG) system requires [12]: a knowledge source to be used; a communicative goal to be achieved; a user model; and a discourse history.

In CISpaces.org we followed a rather pragmatic approach. Indeed, as our main audience are legal practitioners and judges, we strictly obey to the principle of providing them with the important pieces of information in the most concise way. We implemented: (1) a template-based NLG system; (2) a greedy, heuristics-based approach for chaining together premises and conclusion of arguments; (3) an assert-justify writing style suitable for speed reading.

3. Applications to Intelligence and Legal Analysis

CISpaces.org is the result of a collaboration with the US Army research Lab in the NIS ITA programme and with the UK Defence Science and Technology Laboratory in both the NIS ITA programme and follow-on Defence and Security Accelerator (DASA) programme. CISpaces.org is available for use by professional analysts in both the US (Army Research Laboratory) and the UK (Joint Forces Intelligence Group). The first version of CISpaces [15] was one of three key research highlights in the NIS ITA programme [13]. The refinement of the CISpaces software to take it to Technology Readiness Level 4 (characterised as "validation in a laboratory environment") was informed by evaluation conducted with professional analysts in the US and the UK as part of the NIS ITA programme, and enabled by the DASA programme. Development work funded by the DASA programme led to CISpaces being made available as an open-source project under a permissive (MIT) licence: `https://github.com/CISpaces`.

CISpaces.org supported an analysis to determine whether Karadžić possessed *mens rea*[1] for genocide in relation to the Srebrenica mass killing. The results of our analysis were submitted to the Mechanism for International Criminal Tribunals as an *amicus curiæ*[2] brief [9] pursuant to Rule 83 of the MICT Rules of Procedure and Evidence. We based our analysis only on the judgement of Prosecutor v. Radovan Karadžić [8].

We have continued collaboration with the UKs NCA National Cyber Crime Unit, where there is considerable interest in the technologies underpinning CISpaces.org and Open Source Intelligence extraction.

4. Conclusion

CISpaces.org complements human expertise in sense-making activities. The system and its underpinning technologies have attracted positive interest by the UK Joint Forces In-

[1] *Mens rea*: the intention or knowledge of wrongdoing that constitutes part of a crime.

[2] *Amicus curiæ*: a non-party in a lawsuit who argues or presents information relevant to the lawsuit.

telligence Group as well as by the UK National Crime Agencies Dark Web Intelligence Unit and National Cybercrime Unit, in addition to triggering interest in the legal community due to its use for analysing whether Karadžić possessed *mens rea* for genocide in relation to the Srebrenica mass killing.

Acknowledgements

This project was awarded funding through the Defence and Security Accelerator themed competition Revolutionise the human information relationship for defence under Contract DSTLX1000113927. This research was partially sponsored by the U.S. Army Research Laboratory and the U.K. Ministry of Defence under Agreement Number W911NF-16-3-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

References

[1] Cerutti, F.: On scientific enquiry and computational argumentation. In: Proceedings of the 18th Workshop on Computational Model of Natural Argument (2018)

[2] Cerutti, F., Norman, T.J., Toniolo, A., Middleton, S.E.: Cispaces.org: from fact extraction to report generation. In: COMMA 2018. pp. 269–280 (2018)

[3] Cerutti, F., Toniolo, A., Norman, T.J.: On natural language generation of formal argumentation. https://arxiv.org/abs/1706.04033 (2017)

[4] Cerutti, F., Vallati, M., Giacomin, M.: An Efficient Java-Based Solver for Abstract Argumentation Frameworks: jArgSemSAT. International Journal on Artificial Intelligence Tools 26(02) (2017)

[5] Dung, P.M.: On the Acceptability of Arguments and Its Fundamental Role in Nonmonotonic Reasoning, Logic Programming, and n-Person Games. Artificial Intelligence 77(2), 321–357 (1995)

[6] Heuer, R.: Psychology of intelligence analysis. US Government Printing Office (1999)

[7] IBM: i2 Analyst's Notebook. http://www-03.ibm.com/software/products/en/analysts-notebook/, [Accessed: Feb. 2018]

[8] ICTY: Prosecutor v. Karadžić. http://www.icty.org/x/cases/karadzic/tjug/en/160324_judgement.pdf (2016)

[9] McDermott Rees, Y., Cerutti, F.: Request for leave to make submissions as amicus curiae. http://jrad.unmict.org/webdrawer/webdrawer.dll/webdrawer/rec/240941/view/ (2018)

[10] Modgil, S., Prakken, H.: A general account of argumentation with preferences. Artificial Intelligence 195, 361–397 (2013)

[11] Prakken, H.: An abstract framework for argumentation with structured arguments. Argument & Computation 1(2), 93–124 (2010)

[12] Reiter, E., Dale, R.: Building Natural Language Generation Systems. Cambridge (2006)

[13] The Network and Information Science (NIS) International Technology Alliance (ITA): Online legacy - capstone event. http://nis-ita.org/capstone (2016)

[14] Toniolo, A., Braines, D., Preece, A.D., Webberley, W., Norman, T.J., Sullivan, P., Dropps, T.: Conversational intelligence analysis. In: Proceedings of the First International Workshop on Understanding Situations Through Multimodal Sensing. pp. 42:1–42:6 (2016)

[15] Toniolo, A., Norman, T.J., Etuk, A., Cerutti, F., Ouyang, R.W., Srivastava, M., Oren, N., Dropps, T., Allen, J.A., Sullivan, P.: Supporting Reasoning with Different Types of Evidence in Intelligence Analysis. In: AAMAS 2015. pp. 781–789 (2015)

[16] Walton, D., Reed, C., Macagno, F.: Argumentation schemes. Cambridge University Press, NY (2008)

Legal Knowledge and Information Systems
M. Palmirani (Ed.)

doi:10.3233/978-1-61499-935-5-190

Coding Suspicion

Arthur Crivella[a], Wesley M. Oliver[b,1] and Morgan Gray[c]
[a]*Partner, Crivella Technologies, Ltd., Pittsburgh, PA*
[b]*Professor of Law, Duquesne University*
[c]*Third Year Law Student, Duquesne University*

Abstract. Legal standards for suspicion involve seemingly limitless possible factors, leaving them vague and subject to concerns of illegitimate biases by decision makers. Beginning with the relatively small number of factors present in drug interdiction stops, a model can be developed that not only predicts judicial behavior but the odds of discovering drugs. This technology will require legislatures or judges to begin the process of determining what numerical threshold of suspicion justifies investigatory detentions and searches.

Keywords. probable cause, reasonable suspicion, end user vernacular, vector regression

Crivella Technologies, in conjunction with researchers at Duquesne University School of Law, is developing a prototype for assessing whether officers performing drug interdiction stops have adequate suspicion for a search or prolonged detention. This technology holds the potential to improve the accuracy – and decrease implicit biases – in the on-the-scene decisions officers must make daily.

1. Unpredictable and Unreliable Legal Standards for Assessing Suspicion

Assessing suspicion is essential to law enforcement. American law requires police officers to have probable cause to conduct a search for evidence or arrest a suspect. With a lower quantum of suspicion – something the law calls reasonable suspicion – an officer may briefly detain an individual, and the car he is driving, for certain types of investigations. An officer, with this lesser level of concern, may question the detained suspect or have a drug dog sniff a motorist's car. [1]

The seemingly limitless number of factors upon which an officer may rely – and the low threshold requirements for permitting a search – provide little objective guidance for officers, or judges reviewing their decisions to search or arrest. Oliver Wendell Holmes, Jr., famously said that the law is a prediction of what a judge will do. [2] Multi-factored legal standards like probable cause or reasonable suspicion lack meaningful predictability, except when applied to circumstances identical to those judges have previously ruled upon.

Further, and perhaps more importantly, the conclusions of judges that particular sets of facts are sufficiently suspicious to justify an investigatory detention or search are never supported by empirical evidence. While police officers are forbidden to use "hunches," in deciding when to search, the limits of human ability prevent judicial

[1] Corresponding Author: Wesley M. Oliver, Duquesne Law School, 900 Locust Street, Pittsburgh, PA 15282, E-mail: oliverw@duq.edu.

findings of probable cause from being anything more than hunches with footnotes. Advanced cognitive computing, however, can be used to determine whether judges are attributing the appropriate weight to any particular suspicious factor, or combination of suspicious factors.

Modern technology further permits a mean of applying our standards for assessing suspicion in a way that minimizes the implicit racial bias human actors are unaware they possess. This is not to say that we predict that a bias-proof system can be achieved through our system. We do, however, contend we are building a device that considers the likelihood of drug possession on the bases of suspicion officers articulate, not their unstated hunches, a decided improvement over decisions that rely on prejudices that human actors unwittingly possesses.

2. Reasonable Suspicion in Drug Interdiction Stops as an Ideal Model for Machine Learning of Legal Tests

The development of technology that aids in the assessment of suspicion must begin with a manageable subcategory of cases, one that contains a discrete number of possible suspicious factors. Drug interdiction stops provide such a starting point.

A large number of these cases exist. Officers must decide daily whether they will search motorists stopped for an ordinary traffic offense. Specialty units within police departments essentially do nothing else.

A necessarily small number of factors could give an officer a basis for further investigation in this context. During the brief time it takes an officer to write a traffic citation, and wait for a dispatcher to inform him whether the motorist has any outstanding warrants, he or she must determine what, if any, level of suspicion exists. Short of actually seeing or smelling illegal drugs themselves, the short list of bases of suspicion includes: observing a driver’s nervousness, smelling air fresheners or talcum powder designed to disguise other smells, learning that the motorist reports he is traveling to a location different than one reported by his passenger or identified on a car rental agreement.

These factors are not binary and a meaningful model for assessing reasonable suspicion must account for this complexity. Each will exist in degrees and be described using very different terms. A smell can be strong or faint. Nervousness could be identified by sweating, stuttering, or a much more vague claim of anxiousness – and these symptoms could be explained by weather, physical impediments, or the manner of the officer in dealing with the motorist. A driver traveling north of Manhattan reporting that he is traveling to Maine offers a story somewhat inconsistent with his passenger who claims they are traveling to Vermont, but quite inconsistent with a rental agreement identifying Miami as the destination, and potentially consistent with a claimed destination of Boston.

There is, nevertheless, a limited universe of the circumstances that could justify suspicion of drug trafficking in an ordinary traffic stop. There are cases in which officers have reported rather unique circumstances that they claim raised their concern – playing loud gospel music for instance – but most efforts to justify either a search, or continued detention awaiting a drug dog, fit into a fairly small group of categories.

Modern computers have substantial advantages over humans in applying legal standards of suspicion. Computers obviously have faster reading capacity and better memories than any person. Machines can provide answers more consistent with existing case law than human judges. More importantly, machines given the proper databases can determine the actual likelihood that drugs will be discovered in a car when it is provided with the officer's observations.

Further, machines can overcome prejudices much more quickly than humans. Unlike humans, they can be easily coded to ignore race. An officer may deem the driver's Hispanic origin relevant in deciding whether to search the car, but computers can be programmed to ignore at this explicit consideration of race. Machines, however, are obviously not a panacea for discriminatory policing. Officers may ask for the device's assistance disproportionately with minority motorists. The existence of such a device though would improve even this concern. Such a machine would record racial patterns of the requests made of them, providing a valuable tool in identifying and rooting out biases in the training of police and unrealized biases in the algorithm evaluating suspicion.

Using computers and metadata to evaluate probable cause and reasonable suspicion standards does not displace humanity from judging in a way that would be of concern in other multi-variable tests. Many multi-variable standards in constitutional law, for instance, similarly lack predictability. Suspicion, however, is not a matter of values or philosophy – it is truly a question of odds. There may be legitimate disagreements about the degree of certainty required to justify a search. A law-and-order judge may comfortable with a search based on a 15 percent chance of finding drugs, while a civil libertarian may insist only a 40 percent chance of success. At its core, though, probable cause is stated in terms of the likelihood of success. Thus far court have rejected efforts put actual numbers to the test. That reluctance, however, has doubtlessly been driven by the absence of any meaningful way to actually identify the odds of a search's success.[4] Machine learning thus holds the potential not only to make suspicion assessments more reliable and less biased, but to change the question courts ask in determining whether the legal standard has been satisfied. In the context of drug interdiction stops, artificial intelligence provides an opportunity to accurately assess the prediction the law requires officers to make.

3. Developing a Prototype to Assess Suspicion

The current prototype still under development by Crivella Technologies, a Probable Cause Advisor, uses existing case law to identify salient features of suspicion, or stated another way, categories of suspicious facts. Examples in a drug interdiction stop include, for instance, nervousness, masking odors, and inconsistent travel plans. Artificial intelligence applications and language analysis is then used to identify various ways courts and officers describe the presence, and degree, of these salient features. A smell of talcum powder can be faint or strong; the inconsistencies of the travel plans described by the driver and passenger can be minor or extreme. And these salient features can be described in any number of ways.

Language used by a court or officer is identified as a salient factor by comparing text obtained from officers in the field against more formal language used in court proceedings and orders. Measuring semantic text similarity and the use of artificial intelligence for key decision support has been a very active area of research and development to support litigators and experts in complex mass tort pharmaceutical and medical device litigation. The prototype builds upon the methods developed in these contexts and research related to Twitter tweet searches [5] and paraphrase recognition[6] to establish language translation between end user vernaculars.

Our general approach has been a vector regression model to combine a large number of textual and metadata general and domain specific features. The prototypes have been completed using proprietary marker sets tested against large litigation language corpora, in accordance with Content of Interest seed set sampling and evaluation and a powerful semantic word similarity model based on latent semantic analysis. [7]

At present, we have assembled two major corpora of pertinent documents and language. The first, the Corpus of Judicial Probable Cause Opinions, contains over one-hundred thousand decisions assessing whether reasonable suspicion exists in drug interdiction stops. A language corpus, comprising the complete set of written words pertaining to these courts opinions, has been developed and made statistically analyzable. Focused derivative sub-corpora have been further developed to aid in advice algorithm design, testing and training.

The second major corpora is the Language of Active Interdiction. As the name implies this corpus contains the language of police officer real time recording of interdiction observations. This corpora will identify the language officers use in the field to identify the concepts relevant to courts. In some jurisdictions, internal department requirements or judicial decrees require officers to explain their decisions to stop, detain, and search any automobile, indicating whether or not drugs were discovered. For databases of police reports identifying circumstances that led both to the discovery of drugs as well as fruitless searches, training can begin to start developing algorithms for the presence of drugs.

Databases including false-positives are somewhat obviously less available than judicial decision admitting or excluding evidence from drug interdiction stops. Every federal and state court is required to determine whether the evidence it considers in a criminal case was obtained legally. By contrast, only some jurisdictions are retaining records of the suspicion that led an officer to conduct a fruitless search. Additionally, there is often no indication in these records when officers recorded their suspicions before or after the search. Post-search statements may be tainted by facts discovered in the search. Nevertheless, the limited records that exist provide training material so the system can learn the various ways salient factors are described and, more substantially, begins the development of an algorithm for predicting whether drugs are actually present – as opposed to merely the odds that a court will find the search justified. The system will obviously be able to predict judicial behavior with far greater accuracy than the presence of drugs in its early phase.

The trial implementation phase of the prototype will, however, begin the process of producing the best version of a database of suspicion. There will be substantial law enforcement benefits to using even early pilot versions of the program that predict only the judicial perspective on the officer's observations. Officers who type the facts they observe into the program prior to a search, or continued detention, will discover the odds that a court would find a detention or search justified, enhancing their efforts to justify the search to a judge hearing a motion to suppress if drugs are discovered. In entering the data prior to search, the officers will be building a system that increasingly is able to conclude just how probable the officer's cause is.

References

[1] *Warrantless Searches and Seizure,* Geo. L. J. 91 (2003), 36-54.

[2] O.W. Holmes, Jr., *The Path of the Law*, 10 Harv. L. Rev. (1897) 457-478.

[3] D.H. Kaye, *The Double Helix and the Law of Evidence* (2010).

[4] B. Alarie, A. Niblett & A.H. Yoon*, How Artificial Intelligence Will Affect the Practice of Law*, Univ. of Toronto L. J. 68 (2018) 106-24.

[5] B. Siriam, D. Fuhry, E. Demir, H. Ferhatosmanoglu, M. Demirbas, *Short Text Classification in Twitter to Improve Information Filtering*, Proceedings of the 33rd Int'l AEM SIGR Conference in Research Development in Information Retrieval, 841-42.

[6] B. Dolan, C. Quirk, C. Brockett, *Unsupervised Construction of Paraphrase Corpora: Exploiting Massively Parallel News Sources*, Proceedings of the 20th Int'l Conference on Computational Linguistics, Article No. 350 (2004).

[7] The patented technology and extensive analytic software of Crivella Technologies Limited (see International Publication Number WO 2007/146809 A2 and International Publication Number WO 03/012697 A2 for additional information) has been utilized for data analysis and the development of machine learning and analytic decision framework development.

Legal Knowledge and Information Systems
M. Palmirani (Ed.)

doi:10.3233/978-1-61499-935-5-195

Is Blockchain Hashing an Effective Method for Electronic Governance?

Oleksii Konashevych,[a,b,1] Marta Poblet[b,2]
[a]*Erasmus Mundus Joint International Doctoral Fellow in Law, Science, and Technology*
[b] *RMIT University, Graduate School of Business and Law*

Abstract. Governments across the world are testing different uses of the blockchain for the delivery of their public services. Blockchain hashing--or the insertion of data in the blockchain--is one of the potential applications of the blockchain in this space. With this method, users can apply special scripts to add their data to blockchain transactions, ensuring both immutability and publicity. Blockchain hashing also secures the integrity of the original data stored on central governmental databases. The paper starts by analysing possible scenarios of hashing on the blockchain and assesses in which cases it may work and in which it is less likely to add value to a public administration. Second, the paper also compares this method with traditional digital signatures using PKI (Public Key Infrastructure) and discusses standardisation in each domain. Third, it also addresses issues related with concepts such as "distributed ledger technology" and "permissioned blockchains." Finally, it raises the question of whether blockchain hashing is an effective solution for electronic governance, and concludes that its value is controversial, even if it is improved by PKI and other security measures. In this regard, we claim that governments need first to identify pain points in governance, and then consider the trade-offs of the blockchain as a potential solution versus other alternatives.

Keywords. Blockchain, hashing, e-governance, digital signatures, PKI

1. Introduction

Governments across the world are testing potential uses of the blockchain in their public sectors and, in particular, for property registries and notarisation of deeds. In 2016 Estonia launched a "notarisation on the blockchain" project with BitNation [1], although depriving such notarised acts of legal force. Some other countries have also started pilots: Honduras announced a blockchain-based real estate registry, but the project was eventually discontinued [2]; Chromaway—a Swedish start-up—announced in 2016 promising plans to upgrade the Swedish real estate registry by conducting records of deeds on the blockchain [3]. Yet, two years later the third phase of the pilot has been concluded but no results are yet available [4]. In the USA, the project Velox.re in Cook County (Chicago) tested a transaction outside of the real registry to imitate the use of the blockchain for deeds with real estate [5]. The clerk's office of the county issued a report [5] but the project never went ahead. Ubitquity.io announced the project Bitland to

[1] Oleksii Konashevych (a.konashevych@gmail.com) is Erasmus Mundus Joint International Doctoral Fellow in Law, Science and Technology, Last-JD.

[2] Marta Poblet (marta.pobletbalcell@rmit.edu.au) is an Associate Professor at RMIT University.

implement a blockchain-based real estate registry in Ghana with no further continuity either [6].

Other pilots are currently work in progress. Ukraine and the Republic of Georgia announced their cooperation with Bitfury [7] to apply distributed ledger technology and blockchain to their cadastral registries.

Different organisations in the EU and UK have recently released reports on the use of the blockchain [8] [9] [10]. These reports generally express positive views about the impact of the technology and its value in the development of the informational society. Yet, they are much less specific about the design of blockchain-based e-government systems and how to implement blockchains in particular areas.

In this paper we raise some caveats about the use of the blockchain for public—state-owned—registries. We contend that the present lack of standards and regulatory frameworks makes its adoption contentious, as in some cases it could undermine e-government services and public interest in general. In this context, we analyse the potential of blockchain hashing in the e-government space and compare this method with the more traditional ways of digital signatures using standardised PKIs (Public Key Infrastructures) such as eIDAS in the European Union [11]. We conclude that governments should identify and consider pain points in the administrative processes before making decisions that involve blockchain adoption.

2. How does hashing on the blockchain work?

Broadly, hashing on the blockchain refers to inserting a hash sum in a blockchain transaction. Insertion implies that data is "published in the ledger and cannot be censored or retracted and will be permanently available to the world" [12]. The paper [12] offers a comprehensive analysis of different methods of data insertion in Bitcoin.

The method of hashing is recognised as a way of securing public data on the blockchain. Each entry in the central database of the public registry is hashed and casted to the blockchain. This concept is considered by some governments as a way of the use of the blockchain. An example of this is the project by Bitfury in Ukraine, developed in partnership with Transparency International (TI) and the Ukrainian government. The project, launched in 2017, applies Bitfury's DLT "Exonum" to hash records of the geocadastral registry [7]. Authorised nodes that are controlled by the government cast hashes to the Exonum-based ledger. The hash of the current state of the ledger is periodically anchored on the public blockchain. Initially, this public blockchain was Bitcoin, now it is Emercoin [13]. TI keeps another node, which plays the role of the observer and has permission to read the ledger only. As we see, Exonum is an example of a private permissioned DLT. Technical details can be found at [13].

3. Digital signatures vs hashing on the blockchain

Asymmetric cryptography consists of a pair of keys: public and private. The public key is a code string that uniquely identifies a certain individual or company. While the public key can be shared with anyone, the private key must be kept secure and undisclosed [14].

In essence, digital signatures and blockchain hashing do the same thing. Both methods leverage cryptographic algorithms to detect the forgery of an original entry. If any bit of information is altered, the hashes will not match. Since the signing of a blockchain transaction is based on asymmetric cryptography [15] the cryptographic signing of data is not different from signing blockchain transactions. The main difference between the two methods is that hashing on the blockchain adds another layer of data. With digital signatures, users insert their data as an INPUT to the cryptographic function; with hashing on the blockchain, users insert payment data along with the required data.

Digital signatures have not been used for public purposes in this basic form. For this to happen, the system needs to be supported by a Public Key Infrastructure (PKI) which includes a set of standards. PKI consists of technologies, procedures and actors that enable deployment of public-key cryptography-based security services [16]. Within PKI, one provider—known as Certificate Authority (CA) or Trust Service Provider (TSP)—is responsible for the provision of identity services, while other providers—Timestamp Authorities (TSA)—authenticate time and date information. Blockchain hashing, in contrast, does not require a centralised TSA since all blocks are chronologically stored. Timestamps are embedded in the blocks and remain immutable.

At present, PKI benefits from a complete infrastructure with regulations, standards, and procedures. For example, a roadmap of standards in the EU is described in [17]. In contrast, there is no standardised protocol to manage access in permissioned blockchains. There are no procedures for authorising nodes and operators (clerks) either. Clerks use private keys, but there is no standardised protocol if these keys are compromised. More generally, there is no standard for DLTs.

DLT and the blockchain, in summary, are not fully equipped yet. Without regulations, standards, procedures, and certifications, hashing on the blockchain can only work in limited environments, under supervision and, presumably, for research purposes. This is definitely not a scalable approach for e-government at this point in time.

4. How hashing on the blockchain should work?

There are at least three issues that must be addressed to leverage hashing and improve the security of centralised public registries: (i) identification; (ii) bi-directional relations of entries in the database and hashes on the blockchain; and (iii) standardisation.

4.1. Identification

The blockchain was initially conceived to preserve the anonymity (or, at least, the pseudonymity) of both nodes owners (“miners”) and owners of blockchain addresses (“users”). Therefore, blockchains and DLTs must be supplemented with overlaid solutions for identification, authorisation and authentication, with the component of trust services (similar to eIDAS) where necessary.

In a permissioned DLT, the administrator will keep the key to the system and grant permissions (therefore, standards and procedures for managing keys and accesses must be applied). However, if hashing takes place on a public blockchain no one will know who did that—an authorised officer or an attacker—and therefore addresses will need to be identified, or the inserted data will have to contain an identifiable digital signature. Moreover, if the private key to the blockchain address is compromised or stolen, a procedure for a stop list where such an address is added to the special database and any further action from this address is considered invalid will be required. Again, we come back to the need for PKI standards based on proved principles.

4.2. Bi-directional relations between databases

The probable attack scenario here is that the record on the central database is changed or replaced with a new one. If there is no reverse relation with the hash stored on the ledger, a new hash can be created for the corrupted record and published in the distributed ledger, and so presented as a correct one. Apparently, such use of the blockchain does not add any value in terms of the security.

The issue is that hashing does not protect the data itself from being changed and deleted. It just helps to reveal the forgery if the user still keeps in their hands the original record. If the database is closed and centrally controlled—which is how databases work in the public administration—such manipulation is possible.

Traditionally, in government registries this issue is addressed by deploying a multilayered system of security measures, logs, and access procedures. One of the possible ways to use the blockchain is to build a central registry in the style of "chain of blocks" (chain of records, actually) similar to the blockchain, either public or private, or just to move the registry to the blockchain. Therefore, the use of the blockchain for public administration implies both the transfer of the central base to the blockchain—and not just hashing—and the use of PKI for identification.

4.3. Permissioned blockchains, standardisation, and public policy

The current hype about the blockchain may lead to some confusion when it comes to adoption by governments. The alleged achievements in the blockchainisation of e-government services tend to refer to "permissioned blockchains." Yet, permissioned blockchains are architecturally centralised and have a hierarchical system of nodes. In this context, using the notion of "blockchain" or "permissioned blockchain" to refer to a centralised system can be misleading. Arguably, what qualifies as a blockchain is still under discussion: Original Bitcoin-like networks only? Decentralised systems based on other types of consensus mechanisms? A mix of them? What properties does a network have to exhibit to qualify as a blockchain? Ultimately, governments need to rely on shared conceptual frameworks and standards to make the appropriate technology choices. Otherwise, we may end up with scenarios where different government departments use different types of conflicting blockchains, or apply blockchain protocols with only a few nodes (unable to maintain the required level of security of the network), or use DLTs that are not blockchains at all.

Another risk associated with the lack of standardisation is that governments must keep the list of public blockchains whose technologies are proven trustworthy. Yet, such lists of "trusted" blockchains can lead to discrimination, arbitrarily excluding potentially appropriate networks and technologies. Standardisation is a better way to ensure fair competition and stable development.

Public policy and clear roadmaps are needed to avoid voluntarism and data breaches. Yet, a consistent blockchain public policy should also openly acknowledge the role of cryptocurrencies. Cryptocurrencies are the blood of the system, a key mechanism and incentive to maintain large networks sustainable. The alternative to cryptocurrencies are governments creating and maintaining decentralised network infrastructures. Ultimately, this leads to architectural decentralisation by political centralisation. A difficult balance to strike.

5. Conclusion

The standardisation of the blockchain domain is still in the early stages [18]. For this reason, hashing on the blockchain may be premature for public registries. Rather, the alternative option of PKI supported with standards and regulation is a more plausible solution at the present stage. Why, and what for, should we use the blockchain in e-government? Does the first generation of blockchains improve the security standards of existing centralised databases? These questions remain open.

Arguably, governments will need to rethink the nature of public databases and perhaps look in the direction of the new generation of smart contracts and decentralised applications (DApps) (Blockchain 2.0). Smart contracts [19] and DApps can leverage many more functionalities than blockchain hashing, which is primitive and does not unleash the full potential of the blockchain. They could also help to reduce human-generated mistakes, bureaucracy, and costs for the public administration. Yet, smart contracts and DApps will equally require not just consistent regulatory frameworks, but broader organizational change in public administration and governance.

References

1. Estonian Government and Bitnation Begin Cooperation - e-Estonia, https://goo.gl/88pBui.
2. Jun, M.: Blockchain government-a next form of infrastructure for the twenty-first century. J. Open Innov. Technol. Mark. Complex. 4, 7 (2018).
3. Blockchain and Future House Purchases, https://chromaway.com/landregistry/.
4. Kim, C.: Sweden's Land Registry Demos Live Transaction on a Blockchain, https://www.coindesk.com/sweden-demos-live-land-registry-transaction-on-a-blockchain/.
5. Mirkovic, J.: Blockchain Pilot Program. Final Report. (2017).
6. Real Estate Land Title Registration in Ghana Bitland, http://bitlandglobal.com/.
7. Chavez-Dreyfuss, G.: Ukraine launches big blockchain deal with tech firm Bitfury, http://www.reuters.com/article/us-ukraine-bitfury-blockchain-idUSKBN17F0N2.
8. Boucher, P. (Scientific F.U.E.P.: How Blockchain Technology Could Change Our Lives. (2017).
9. ENISA: Security Guidelines on the Appropriate Use of Qualified Electronic Signatures. Guidance for Users. European Union Agency for Network Information Security (2016).
10. Walport, M.: Distributed ledger technology: Beyond block chain. A report by the UK Government Chief Scientific Adviser. (2015).
11. Thomas Fillis: Electronic Registered Delivery Service (ERDS) and the eIDAS Regulation, https://goo.gl/w1TEvC, (2016).
12. Sward, A., Vecna, I., Stonedahl, F.: Data Insertion in Bitcoin's Blockchain. Ledger. 3, 1–23 (2018).
13. Consensus Algorithm Specification, https://exonum.com/doc/advanced/consensus/specification/.
14. eIDAS for Dummies.
15. Nakamoto, S.: Bitcoin: A Peer-to-Peer Electronic Cash System, https://bitcoin.org/bitcoin.pdf.
16. Trček, D.: Managing information systems security and privacy. (2006).
17. Electronic Signatures and Infrastructures (ESI); The framework for standardization of signatures : overview. (2015).
18. Cuschieri, H., Chen, Y.: ISO/TC 307 - Blockchain and distributed ledger technologies, https://www.iso.org/committee/6266604.html.
19. Szabo, N.: Formalizing and Securing Relationships on Public Networks. First Monday. 2, (1997).

Legal Knowledge and Information Systems
M. Palmirani (Ed.)

doi:10.3233/978-1-61499-935-5-200

Towards Explainable Semantic Text Matching

Jörg LANDTHALER, Ingo GLASER and Florian MATTHES
Software Engineering for Business Information Systems, Department of Informatics, Technical University of Munich, Germany

Abstract. The growing amount of textual data in the legal domain leads to a demand for better text analysis tools adapted to legal domain specific use cases. Semantic Text Matching (STM) is the general problem of linking text fragments of one or more document types. The STM problem is present in many legal document analysis tasks, such as argumentation mining. A common solution approach to the STM problem is to use text similarity measures to identify matching text fragments. In this paper, we recapitulate the STM problem and a use case in German tenancy law, where we match tenancy contract clauses and legal comment chapters. We propose an approach similar to local interpretable model-agnostic explanations (LIME) to better understand the behavior of text similarity measures like TFIDF and word embeddings. We call this approach eXplainable Semantic Text Matching (XSTM).

Keywords. Semantic Text Matching, Explainable AI, Word Embeddings, TFIDF, Text Similarity Measure, German Tenancy Law

1. Introduction

The amount of textual data relevant in the legal domain is continuously growing. This leads to a demand for text analysis tools that capture more and more of the semantics of the textual data. Automated semantic processing of texts requires an adequate representation of texts. Many scientific applications of NLP for legal information systems leverage word embeddings, for example, question answering by Adebayo et. al [1], information extraction by Chalkidis et. al [2] or argumentation mining by Rinott et. al [3]. It is unclear when TFIDF or word embeddings is the superior technology. While word embeddings' characteristics are intriguing, to date it is not understood why certain structures occur in the embedding spaces nor efficient and effective quality measures for word embeddings are available.

Explainable artificial intelligence (XAI) is a research area that aims to better understand the behavior of algorithms. Waltl and Vogl [4] elaborate on the importance of XAI approaches for the legal domain. Waltl et al. [5] investigated the application of the particular XAI method LIME [6] to explain the behavior of supervised machine learning algorithms on classification tasks. In this paper we focus on text similarity measures to solve Semantic Text Matching (STM) problems. STM is the general problem identifying implicit semantic or logical relationships among text fragments. We propose an explanation approach similar to LIME for an unsupervised machine learning pipeline that we call eXplainable semantic text matching (XSTM). XSTM performs a sensitivity analysis for the words that are part of a text similarity measurement. This allows to investigate the contribution of the individual words to the text similarity measurement. We show preliminary results of XSTM on a German tenancy law use case, where text fragments of tenancy contracts and legal comments are matched.

The remainder of this paper is organized as follows: Section 2 recapitulates STM and briefly summarizes results of our previous research. We introduce our XSTM approach in Section 3. In Section 4 we set our work into the context of related work. Section 5 concludes the paper with a short summary and an outlook to future work.

2. Semantic Text Matching (STM)

STM has been introduced in [7]. STM is the general problem of identifying implicit links among text fragments where text fragments stem from documents of one, two or more different document types. For example, in argumentation mining premises need to be matched against claims. Problems of this kind are often tackled with text similarity methods. A high text similarity indicates a high probability for a link. A benefit of text similarity measures based on TFIDF or word embeddings is the unsupervised nature, i.e. no labeled data is required. STM can be seen as generalization of several problems that can be solved using text similarity measure technologies. In contrast to explicit citation networks, where references to other text fragments are present in the text, STM is an approach to identify implicit semantic or logical references among text fragments. STM is related to general information retrieval. The more different document types are involved, the more STM approaches a general search problem. A restriction to one or two document types leads to a problem easier to solve than general search. This enables a deeper investigation of the involved text similarity measures with less side effects.

Our use case is an envisioned support tool for lawyers that analyze or edit contracts. For our novel human-computer interaction method lawyers interactively

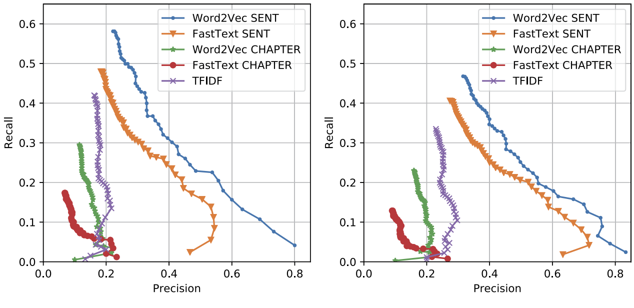


Figure 1. Precision/Recall curves for our STM use case that matches tenancy contract clauses and legal comment chapters. The SENT approach significantly performs better than TFIDF and CHAPTER approaches. word2vec performs better than FastText. TFIDF performs better than word embeddings with the CHAPTER approach. Results shown for the *narrow* tagging in the left column and for the *broad* tagging in the right column.

select a text fragment of interest and see relevant results of a suitable corpus. We implemented this as a web application. The dataset encompasses six tenancy contracts (37 cosmetic repairs related clauses) and three legal comments (1800 chapters). The identification of related legal comment chapters can be seen as a STM problem. We compared three approaches to recommend users relevant legal comment chapters:

- **TFIDF:** We encode the corpus of all contracts and legal comment chapters using the traditional TFIDF[1] representation and calculate the cosine similarity to rank all chapters.
- **CHAPTER:** Contract clauses and legal comment chapters are represented with vectors where a text fragment is the sum of all word embedding vectors representing the contained words[8]. We rank results again with cosine similarity.
- **SENT:** Equal to the CHAPTER approach except that the legal comment chapters are further segmented into sentences and the chapters with the most similar sentences are retrieved.

From the dataset described in [7] we only show the results for the cosmetic repair contract clauses where our ground truth for evaluation contains 223 links for the broad tagging and 127 links for the narrow tagging among contract clauses and legal comments. The differences among the broad and narrow tagging are explained in [7]. The word embeddings have been trained with word2vec[2][9] and FastText[3][10] with standard parameters except size is set to 300, iterations to 100 and min-count is set to zero. We use the CBOW model. Fig. 1 shows that our SENT approach performs best. However, for our use case the user needs some training to use our system effectively and the ground truth can only capture a subset of potential queries.

3. eXplainable Semantic Text Matching (XSTM)

We propose the XSTM approach to investigate the effect of individual words in text similarity applications. XSTM draws from ideas of LIME [6]. The idea is to perform a sensitivity analysis with the text fragments words as input features and the text similarity score among text fragments as output. XSTM can be applied to any text similarity application with different text similarity technologies, for example word embeddings. In order to assess the impact of an individual word we remove the word from one text fragment and re-calculate the text similarity between two text fragments. The difference among the original text-similarity and the newly calculated text similarity can be seen as the contribution for this particular word for the similarity among the two involved text fragments. This can be extended to all words of two text fragments by subsequently removing all words one after the other. Fig. 2 illustrates the contributions for all words among two text fragments of our tenancy law use case. XSTM can be further extended

[1] https://radimrehurek.com/gensim/, version 3.1.0, last accessed September 2018
[2] https://github.com/kzhai/word2vec, version 0.1c (for OSX), last accessed September 2018
[3] https://fasttext.cc/, version 0.1.0, last accessed September 2018

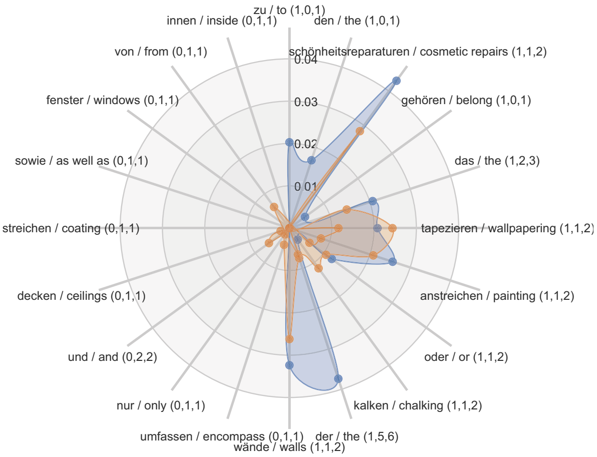


Figure 2. The contribution of individual words to a single match can be visualized with a radar chart for both: the query (contract clause) and the matching sentence (from a legal comment chapter). The query (orange) is *Zu den Schönheitsreparaturen gehören das Tapezieren, Anstreichen oder Kalken der Wände.* (To the cosmetic repairs belongs the wallpapering, painting or chalking of the walls.). The matching sentence (blue) using our SENT approach is *Schönheitsreparaturen umfassen nur das Tapezieren, das Anstreichen der Wände und Decken sowie das Streichen der Fenster von innen.* (The cosmetic repairs encompass only the wallpapering, painting of walls and ceilings as well as the coating of windows from the inside.). The axes display the contribution for the participating words (in brackets: German, English translation, frequency in query, frequency in matching sentence, occurrence frequency in both, query and matching sentence). A simplified example was chosen to facilitate visualization and it is not representative for the dataset. An interesting observation is that nouns seem to contribute most to the similarity among query and matching sentence. A structured evaluation of several matches will be necessary.

to assess the contribution of words among text fragments that are part of several links.

4. Related Work

LIME, proposed by Ribeiro et. al [6], is a XAI approach that performs a sensitivity analysis on black box machine learning classifiers. The effect of input variations on the output can serve as ‘explanation’ of the importance of different features to a specific classification result. In contrast to that, we focus on an unsupervised application. Waltl et. al [5] compare a rule-based approach and machine learning classifiers to classify sentences of the tenancy law part of the German Civil Code. On one concrete classification result they showed, using LIME, that the machine learning classifiers most significant features are similar to the features of the manually crafted rules. Semantic Text Matching [7] is the general problem of identifying implicit semantic or logical links among text fragments and is related to or present in several legal domain applications, for example, information retrieval and argumentation mining. A subtask of argumentation mining is to identify premises that support a claim. Rinott et. al [3] use TFIDF and word embeddings to identify evidence for claims in debates. However, it is rarely investigated why one text similarity measure surpasses another. Qureshi and Greene [11] present an unsupervised explainable word embeddings technique (EVE) that modifies the training of word embeddings in way so that individual dimensions of the word embeddings are clamped to specific concepts of a knowledge base such as Wikipedia. EVE is a constructive approach to build explainable embedding models by nature. In contrast to that, our approach is an attempt to investigate the characteristics of native embedding methods like word2vec or FastText from the outside.

5. Conclusion

We recapitulated STM as a general problem that also occurs in legal domain specific applications such as argumentation mining. We compare TFIDF and word embeddings as text similarity measures to solve a particular STM in German tenancy law. We propose XSTM as an approach to assess the impact of individual features (words) when used in a text similarity application. We hope that XSTM will enable us to deeper investigate the behavior of the different text representation and text similarity methods TFIDF and word embeddings.

References

[1] G. B. Adebayo Kolawole John, Luigi Di Caro and C. Bartolini, “- an approach to information retrieval and question answering in the legal domain,” in *Proceedings of the 10th International Workshop on Juris-informatics (JURISIN 2016)*, 2016.

[2] I. Chalkidis and I. Androutsopoulos, “A deep learning approach to contract element extraction,” in *Legal Knowledge and Information Systems - JURIX 2017: The Thirtieth Annual Conference, Luxembourg, 13-15 December 2017.*, 2017, pp. 155–164.

[3] R. Rinott, L. Dankin, C. Alzate, M. M. Khapra, E. Aharoni, and N. Slonim, “Show me your evidence–an automatic method for context dependent evidence detection,” in *Proceedings of the 2015 Conference on Empirical Methods in NLP (EMNLP), Lisbon, Portugal*, 2015, pp. 17–21.

[4] B. Waltl and R. Vogel, “Explainable artificial intelligence the new frontier in legal informatics,” *Jusletter IT 22*, 2018.

[5] B. Waltl, G. Bonczek, E. Scepankova, and F. Matthes, “Semantic types of legal norms in german laws: Classification and analysis using local linear explanations,” *Artificial Intelligence and Law*, 2018.

[6] M. T. Ribeiro, S. Singh, and C. Guestrin, “”why should I trust you?”: Explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, 2016, pp. 1135–1144.

[7] J. Landthaler, I. Glaser, E. Scepankova, and F. Matthes, “Semantic text matching of contract clauses and legal comments in tenancy law,” in *Tagunsband IRIS: Internationales Rechtsinformatik Symposium*, 2018.

[8] Q. V. Le and T. Mikolov, “Distributed representations of sentences and documents.” in *ICML*, vol. 14, 2014, pp. 1188–1196.

[9] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *CoRR*, vol. abs/1301.3781, 2013.

[10] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[11] M. A. Qureshi and D. Greene, “Eve: explainable vector based embedding technique using wikipedia,” *Journal of Intelligent Information Systems*, Jun 2018.

Legal Knowledge and Information Systems
M. Palmirani (Ed.)

doi:10.3233/978-1-61499-935-5-205

A Chatbot Framework for the Children's Legal Centre

Jay MORGAN [a], Adeline PAIEMENT [a,c], Monika SEISENBERGER [a], Jane WILLIAMS [b], Adam WYNER [a,b],

[a] *Swansea University, Department of Computer Science*
[b] *Swansea University, School of Law*
[c] *Université de Toulon, Laboratoire LIS*

Abstract. This paper presents a novel method to address legal rights for children through a chatbot framework by integrating machine learning, a dialogue graph, and information extraction. The method addresses a significant problem: we cannot presume that children have common knowledge about their rights or express themselves as an adult might. In our framework, a chatbot user begins a conversation, where based on the circumstance described, a neural network predicts both speech acts, relating to a dialogue graph, and legal types. Information is extracted in order to create a case for a legal advisor. In collaboration with the Children's Legal Centre Wales, who advocate for the improvement of legal rights in Wales, a corpus has been constructed and a prototype chatbot developed. The framework has been evaluated with classification measures and a user study.

Keywords. Children's Legal Rights, Chatbot, Natural Language Processing, Machine Learning, Recurrent Neural Networks

1. Introduction

Chatbots are computer programs that allow for interaction with systems through natural language [3]. They can be used to make legal processes more accessible by reducing the burden of legal knowledge. In this paper, we present a new usage of chatbots to improve children's access to their legal rights, through making it easier for them to contact and get help from a (human) advisor. The goals of this chatbot are: 1) identify the legal circumstances; 2) identify the involved parties; and 3) from information in 1 and 2, create a case for an advisor. To validate this approach, a prototype and a new corpus dataset have been created and evaluated in collaboration with the Children's Legal Centre Wales (CLC)[1]. The centre provides consultations and information about laws that affect children in Wales. They are developing a Virtual Legal Practice to manage legal cases between children and practising lawyers. This presents a use-case to demonstrate the chatbot.

[1] https://childrenslegalcentre.wales/

Speech Act	**Artificial**	**Real**	**Example Statement**
Greetings	153	36	Hello, can you help me?
Statement	467	134	Is it legal for my dad to hit me?
Positive Response	151	24	yeah please
Negative Response	143	17	no thanks
Legal Type	**Artificial**	**Real**	**Example Statement**
Abuse	187	46	My boyfriend hit me what can I do?
Cyber-Crime	105	15	Someone online is bullying me
Hate-Crime	78	19	Is it illegal to make fun of other religions?
Under-age Sex	97	54	My gf is under 16, can we have sex?

Table 1. Breakdown of the number of artificial and real statements in the corpus.

There have been successful chatbots that provide legal services. Most notable are *DONOTPAY*[2] that assists motorists in appealing parking tickets, and *Visabot* that aims to help with immigration issues. Both chatbots ask questions, gather data, and draft documents for the user to proceed with the case themselves. As they are proprietary tools, they are unavailable for academic development that would alleviate problems such as: being limited in their dialogue interactions; a presumption of an adult's level of comprehension. We are not aware of any chatbots designed for children's rights, where we must relate the language of children to legal concepts, as a child may describe a problem in everyday rather than legal terms.

In Section 2, we discuss the corpus. Section 3 presents the methodology, and Section 4 evaluates the framework with classification measures and user studies. This paper concludes with a discussion and future work in Section 5.

2. Corpus Dataset

While there are legal corpora on which machine learning methods may be trained, e.g., the British Law Report Corpus (BLaRC) [6], some of which bears on family law, the terminology is not such as we might expect children to use and the format is not that of chat logs. We are unaware of any corpus of chats by children about legal matters or any corpus modelling children's language. Gathering a corpus of children's language about sensitive legal matters is intrinsically problematic. Therefore, we created a novel corpus of messages. The corpus (Table 1) is comprised of "artificial" statements, approximating the language of a child, and "real" statements extracted from a user study in which adults modelled children's language. The statements are classified in terms of *speech act* and *legal type*.

3. Methodology

When the user accesses the chatbot's interface, a dialogue graph (Fig. 1) is created to track turns with the user. To interact with the user, the chatbot must perform

[2] https://www.donotpay.com/

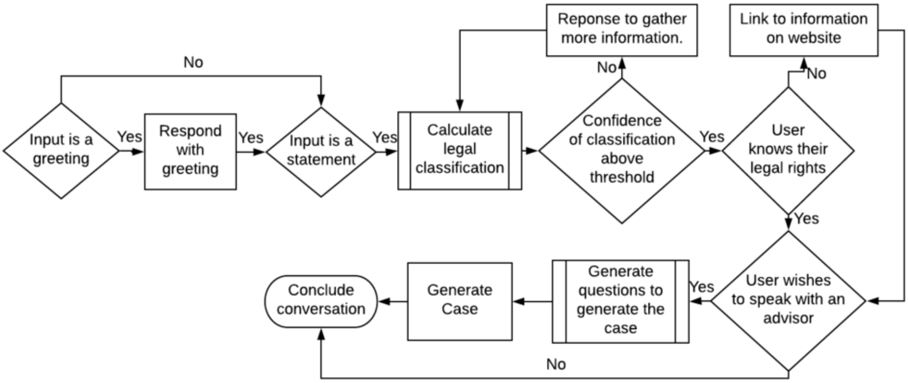

Input is a greeting
Yes
No
Respond with greeting
Yes
Input is a statement
Yes
Calculate legal classification
Reponse to gather more information.
No
Confidence of classification above threshold
Yes
Link to information on website
No
User knows their legal rights
Yes
User wishes to speak with an advisor
Yes
Generate questions to generate the case
Generate Case
Conclude conversation
No

Figure 1. Dialogue graph used by the chatbot.

Legal Type	**Required Information**
General Information	Contact name, time of event, contact information, and contact time.
Abuse	Event location. Who is the abuser and abused.
Cyber-Crime	Which platform the event occurred. The reason behind the case.
Hate-Crime	Who committed the act. What act was committed.
Underage-Sex	The age of parties involved. The reason behind the request.

Table 2. Information that the chatbot must acquire for the advisor's case.

two tasks. First, it reasons as to the role of each user input (its speech act) as well as to the legal type. These classification tasks are performed by a neural network. For a current position in the conversation a predefined response that suits the identified situation will be returned. Any user statements that are not recognised lead to a default response being returned to keep the conversation going. Secondly, during this conversation the chatbot recognises named entities (name, location, time). At the end of the conversation, a report is generated for an advisor to take over.

Classification of the Message's Functions and Contents To allow the neural network (Fig. 2) to classify input messages as a speech act and legal type, words are tokenized and converted to word vectors of 200 dimensions, aiming to capture the semantic similarities between words [1]. These word vectors go through two LSTM layers to encode the impact of word order in the sentence's meaning. The output of these recurrent layers are further transformed by a dense layer with a ReLU activation function, and a dropout rate of 20% to reduce overfitting. The sentence is classified by two parallel dense layers with a softmax activation.

Named Entity Recognition Class The system recognises and extracts named entities in the user's statements to be used later without having to ask for the information. Regular expressions are used for well-formatted inputs, e.g. email addresses. Otherwise, a neural network is used [2].

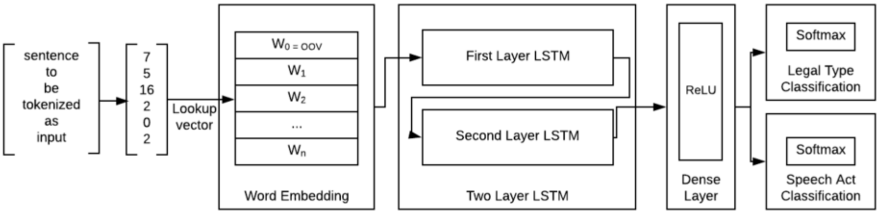


Figure 2. Neural Network Architecture

Network Type	**Speech Act F_1 Score**	**Legal Type F_1 Score**	**Avg F_1 score**
Dense Neural Network	97.36% (+/- 0.57%)	93.93% (+/- 1.21%)	95.65%
Two Layer RNN	95.16% (+/- 1.73%)	93.23 (+/- 1.24%)	94.20%
Pre-trained Embedding	98.41% (+/- 0.90%)	98.06% (+/- 0.35%)	98.24%

Table 3. Comparison between Dense and Recurrent Neural Network classification scores.

4. Results and Evaluation

4.1. Evaluation of the Classification

We perform a comparative evaluation of our neural network against a baseline, using the standard cross-validation procedure, beginning with a simple dense neural network with an untrained embedding layer. As we replace the dense layers with LSTM layers (Fig. 2) the classification scores for both speech act and legal type drop. A possible cause could be the LSTM layers not learning accurate representations. Using a pre-trained embedding [5], the proceeding layers are provided with descriptive vectors, resulting in the best score of 98.24%.

4.2. User Studies

We invited 14 participants to select 3 situations from a total of 5, then to converse with the chatbot and complete a questionnaire of 11 questions. Responses are ranked 1 (strongly disagree / no) to 5 (strongly agree / yes). Two questions for each measure were accumulated, and the average score for all participants were taken (Table 4).

1. How easy was the chatbot to use?
2. How easy was it to create a case with an advisor?
3. How well do you feel that the chatbot understood you?
4. The pace of the interactions were suitable.
5. If the chatbot did not understand, was it easy to reformulate the response?
6. How friendly was the chatbot?
7. Were the questions the chatbot was asking you clear?
8. The conversation felt natural.
9. Would you use the system again?
10. Overall, do you find yourself satisfied with the experience?
11. Free form feedback.

User Study Measure	Minimum	Maximum	Average
Ease of Use (Q1, Q2)	6	10	7.42
Interaction Performance (Q4, Q8)	6	9	6.71
Politeness & Responses (Q6, Q7)	7	10	7.57
Perceived Understanding (Q3, Q5)	3	10	5.00
Future Use (Q9, Q10)	4	10	8.29

Table 4. User study questionnaire responses.

Ease of use shows the degree of difficulty in using the chatbot to create a case – an essential purpose of our chatbot. *Interaction performance* and *Politeness* are determined by the dialogue graph and templating of responses. With this prototype, participants found the pace of the conversation to be suitable.

Perceived understanding shows the participant's belief that they have been understood by the chatbot. From the free-form feedback in the questionnaire, we see this measure drop due to the usage of templated responses. Indeed, the chatbot does not indicate that it had understood, rather, it would move on without acknowledgement. This may be addressed without the need for a complex text generator, by rephrasing the user input to create an echo effect [4].

5. Conclusion and Future Work

We presented a chatbot framework to improve children's access to a legal advisor and their legal rights. Our method uses machine learning to perform joint predictions of the speech act and the legal type being described by the user, in addition to extracting named entity extraction for the case creation.

This approach was evaluated through classification tests, and a user study in which participants interacted with the system to describe a legal situation to create a case for an advisor. Our framework may now be expanded to more legal case types and population groups in future works.

References

[1] M. Faruqui and C. Dyer. Improving Vector Space Word Representations Using Multilingual Correlation. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471, 2014.

[2] M. Honnibal and M. Johnson. An Improved Non-monotonic Transition System for Dependency Parsing. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, 2015.

[3] A. Kerly, P. Hall, and S. Bull. Bringing chatbots into education: Towards natural language negotiation of open learner models. *Knowledge-Based Systems*, 20(2):177–185, 2007.

[4] W. Kulesza, D. Dolinski, A. Huisman, and R. Majewski. The Echo Effect: The Power of Verbal Mimicry to Influence Prosocial Behavior. *Journal of Language and Social Psychology*, 33(2):183–201, 2014.

[5] J. Pennington, R. Socher, and C. Manning. Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, 2014.

[6] M. J. M. Pérez and C. R. Rizzo. Design and compilation of a legal English corpus based on UK law reports: the process of making decisions. In *Las Tic: Presente y Futuro en el analisis de corpus*, pages 101–110. 1996.

Legal Knowledge and Information Systems
M. Palmirani (Ed.)

doi:10.3233/978-1-61499-935-5-210

Query Generation for Patent Retrieval with Keyword Extraction Based on Syntactic Features

Julien ROSSI [a,1,2], Evangelos KANOULAS [a]

[a] *Informatics Institute & Amsterdam Business School, University of Amsterdam, Netherlands*

Abstract. This paper describes a new method to extract relevant keywords from patent claims, as part of the task of retrieving other patents with similar claims (search for prior art). The method combines a qualitative analysis of the writing style of the claims with NLP methods to parse text, in order to represent a legal text as a specialization arborescence of terms. In this setting, the set of extracted keywords are yielding better search results than keywords extracted with traditional method such as tf-idf.

Keywords. Patent, Claims, Legal Text, NLP, Information Retrieval

1. Introduction

This work focuses on improving the effectiveness of the search for prior art. It is a high recall task, where the target is to retrieve from databases a list of documents that relate to the invention described in a patent and demonstrate whether it is indeed novel or not.

A number of methods have been proposed in the literature with the purpose of searching for prior art without the need to manually construct a Boolean Query, based on automated keyword-based query generation. Many of the previous works were based on keyword extraction out of the description or the claims, we refer to Konishi [3], Golestan et al. [1], Mase et al. [8], Lopez and Romary [4,5] and Verberne and dHondt [10].

In this work we also follow the methodology of keyword-based query generation, but different from all the previous work as it does not rely on term frequency to identify semantic salience (TF-IDF, BM25). Instead we depend on creating trees of words out of the claim section of the patent and based on these trees we attempt to identify novelty terms. Suzuki and Takatsuka [9] tackled the problem of identifying novelty-related keywords based on claims in an Information Extraction task. Our more generic method based on a constituency parser and chunking that relies on grammar to judge word salience. We show that this method can generate queries yielding to better search results, and we evaluate improved recall but also improved ranking.

[1] Corresponding Author: Julien ROSSI, e-mail: *j.rossi@uva.nl*

[2] This work is a part of a paid internship offered by the European Patent Office. Opinions expressed in this paper are authors only, and do not reflect the opinion of the European Patent Office.

To summarize, in this work we attempt to answer the following research question: *Can we improve the retrieval of relevant document by selecting keywords based on syntactical signals of semantic importance, rather than term-frequency?*

2. Methodology

In a nutshell our method works as follows: (1) Based on the complete claim set of a patent filing we generate a claim tree; (2) we then generate a specialization tree for each claim, and (3) score words based on their appearances in tree nodes; (4) we then submit the selected top-n keywords into a search engine.

2.1. Generating the Claim Tree

Each claim can establish itself as a refinement of one or more other claims. We used regular expressions to identify the dependencies and create the claim trees. [3]

Each claim is parsed by the Stanford Core NLP Constituency Parser [7], which provides the word tokenization, the POS tagging and the constituency parsing itself. Because of the unusual language of the claims, it is typical that words like "said" or "claim" are misclassified as verbs where they are instead used as relative adjectives. This incorrect POS tagging has repercussions in the chunking. We created a new annotator in the Stanford Core NLP Server to correct these tags, we refer to Hu et al. [2] for the correction.

2.2. Generating Specialization Trees

The constituency parsing tree is traversed depth-first, creating a string of tags, both POS and Chunks tags. Our set of regular expression identifies two types of patterns that can be expressed as head-to-child relation within a tree:

- Composition : a system comprising this and that, the constituency parsing allows for a lot of flexibility in the actual wording, as the chunk structure stays stable in that situation. Head node contains a system, attached to two child nodes this and that
- Specialization : a system made of this, which is on top of this and that, again the chunk structure is very stable and resistant to the diversity of the wording. Head node contains a system, attached to one child node this, itself attached to 1 child node this and that

We observe that the chunking produced by the Stanford Parser is very stable over the actual phrasing and choice of verbs, words and delimiters. It efficiently reduces lexical and morphological variations of the concepts of composition and specialization to a few chunk patterns. We leverage the chunking stability to then fold sentences into trees.

The specialization tree is the representation of one claim as a tree based on the relations of composition and specialization between chunks of the text. We identify words w as belonging to specialization tree nodes n_i: $\forall w, N(w) = \{n_i, where\, w \in n_i\}$, $\forall w, P(w) = \{(nd(n_i), nh(n_i), cd(n_i)), n_i \in N(w)\}$

[3] Strictly speaking, these are not trees, since each claim has an identified list of parent nodes.

nd, *nh* are the depth and height of the node within the specialization tree, *cd* is the depth of this claim within the claim tree of that patent.

2.3. Scoring keywords

Words are then grouped by stem, using the Porter Stemmer. For a specific stem, we record which one of all the words with the same stem had the highest number of occurrences: $\forall stem\, s,\, P(s) = \bigcup_{stem(w)=s} P(w)$

The scoring method has to favor words that are located deep within the specialization tree, as they relate to finer details of the invention, which is where we expect an invention to stand out of other similar inventions. We also want the scoring to favor words that are within claims that are deep into the claim tree, for the same reason as above, as a claim discloses finer details about the claims it depends on.

We devised two scoring methods: $CLST05(s) = \sum_{P(s)} e^{\alpha_{05} * \frac{nd}{nd+nh-1} + \beta_{05} * cd}$, and $CLST06(s) = \sum_{P(s)} e^{\alpha_{06} * \max(nd) + \beta_{06} * \max(cd)}$

The hyperparameters $\alpha_{05}, \alpha_{06}, \beta_{05}, \beta_{06}$ are determined by experimenting and keeping the values that generate the highest metrics.

The top-n stems with the highest scores are selected to construct a query, which is the concatenation of the words associated with these stems.

3. Experimental Setup

Dataset. We used the CLEF-IP 2011 Topic Collection as a basic dataset. This collection contains patent documents with qrels to identify the definitive list of relevant documents for each case. The search database is the complete historical worldwide repository of patents. The setting is to search for relevant documents based on the claims from the seed document, searching through the claims of the documents in the corpus.

Search Engine. We used an instance of Lucene search engine.

Baseline. The baseline is a system developed at the EPO, known under the MLT acronym, configured to act within the same parameters, using the claims as source for query, and the claims within the search database as corpus. We use it as it is representative of related work that uses TF-IDF term weighting to extract keywords.

System Configuration. Our methods are called CLST-05 and CLST-06, and each method has three variants: "As is", "BOOST" where the word scores are used as boost factors for the search engine, and "NO-RETAG" where the POS tagging is not corrected. The boost increases the scoring of a document that contains the boosted terms. We used the Lucene instance in place at the EPO. This search engine receives our query and returns a list of ranked search results.

Evaluation Metrics. We evaluate the performance of the different keyword-based query generation methods on the basis of Recall@100, and PRES@100, similar to the evaluation performed under CLEF-IP. PRES is a metric introduced by Magdy and Jones [6]. [4]

[4] We had to correct the formula presented in the original paper as it was producing results out of the range $[0,1]$. We used $\sum r_i = \sum_{i=1}^{nR} r_i + \sum_{i=nR+1}^{n} (N_{max} + n - (i - nR - 1))$

4. Results and Analysis

In the first place, we can compare the average PRES@100 and Recall@100 to the baseline. We developed 6 different systems that we can evaluate. For each system we can also select how many words we extract as keywords, from 50 to 100 by increment of 10. The performance of all those systems is given in Table 1.

Table 1. PRES@100 and Recall@100

	PRES@100					**Recall@100**					**Summary**	
	Number of Keywords					**Number of Keywords**					**R@100**	**PRES@100**
System Name	60	70	80	90	100	60	70	80	90	100		
CLST-05	0.18	0.19	0.19	0.19	0.19	0.24	0.24	0.25	0.25	0.25	**0.2479** (***)	**0.1923** (***)
CLST-05B	0.14	0.15	0.15	0.15	0.15	0.19	0.19	0.20	0.20	0.20	0.2019 (***)	0.1525 (***)
CLST-06	0.18	0.18	0.19	0.19	0.19	0.23	0.24	0.24	0.24	0.25	0.2463 (***)	0.1918 (***)
CLST-06B	0.12	0.12	0.12	0.13	0.13	0.16	0.17	0.17	0.17	0.18	0.1749	0.1275
CLST-06NR	0.18	0.18	0.19	0.19	0.19	0.23	0.24	0.25	0.25	0.25	0.2467 (***)	**0.1925** (***)
CLST-06NRB	0.12	0.12	0.13	0.13	0.13	0.16	0.17	0.17	0.18	0.18	0.1756	0.1294
MLT		0.13					0.17				0.1742	0.1325

The summary clarifies which results are a statistically significant improvements over the MLT baseline (randomization test, *** means $p < 0.001$)
The results show that BOOST is significantly decreasing performance. This can be interpreted as the scoring system having a good effect on selecting salient keywords over more general words, but not being able to catch the variations in relative importance of words in a way that is numerically in line with the boost factors of the search engine.

The correction of the POS tagging, which was tried only on the system CLST-06, does not generate a statistically significant improvement over the vanilla version. We analyze that the distortions on the parsing occur at different places than those where a specialization or combination occurs, which makes the system oblivious to this correction to a certain extent. Nonetheless, well keep this correction in mind for future work, especially when additional features get extracted from the dependency parser.

The evaluation metrics keep increasing with the number of keywords, the difference being statistically non-significant between 80, 90 and 100 keywords. Our system overperforms the existing system, with a statistically significant improvement with 30 keywords. We keep the results based on 100 keywords.

The significant result is that both CLST-05 and CLST-06 largely outperform the TFIDF-based baseline. Results show significant improvement in this setting of Query Generation, although the setting mixes the performance of the keyword extraction and the tweaking of the underlying search engine. Nonetheless, the approach of going away from term frequency methods to identify salient words in presence of an enforced writing style is proven to make sense. The term frequency allows for identification in absence of other information on how the text is written, while we can leverage the additional information that authors are restricted to deliver information in a way that is reflected in grammar, thus enabling us to work at the semantic level by working at the grammatical level.

5. Conclusion

In this work we used the sentence morphological features to identify keywords within patent claims, and used these keywords as query terms to retrieve other relevant patents. In this setting we establish a significant improvement over the existing baseline, based on term-frequency weighting methods.

In the future we plan to apply and expand this work on other text sources with constrained writing style. We also see potential in adapting NLP tools that were designed or trained on conventional literature.

References

[1] Far, G., Mona, Sanner, Scott, Bouadjenek, Reda, M., Ferraro, Gabriela, David, H.: On term selection techniques for patent prior art search. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 803–806. SIGIR '15, ACM, New York, NY, USA (2015). https://doi.org/10.1145/2766462.2767801, http://doi.acm.org/10.1145/2766462.2767801

[2] Hu, M., Cinciruk, D., Walsh, J.M.: Improving automated patent claim parsing: Dataset, system, and experiments. CoRR **abs/1605.01744** (2016), http://arxiv.org/abs/1605.01744

[3] Konishi, K.: Query terms extraction from patent document for invalidity search. In: Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access, NTCIR-5, National Center of Sciences, Tokyo, Japan, December 6-9, 2005 (2005)

[4] Lopez, P., Romary, L.: Multiple Retrieval Models and Regression Models for Prior Art Search. In: CLEF 2009 Workshop. p. 18p. Corfu, Greece (Sep 2009), https://hal.archives-ouvertes.fr/hal-00411835

[5] Lopez, P., Romary, L.: Experiments with citation mining and key-term extraction for Prior Art Search. In: CLEF 2010 - Conference on Multilingual and Multimodal Information Access Evaluation. Padua, Italy (Sep 2010), https://hal.inria.fr/inria-00510267

[6] Magdy, W., Jones, G.: Pres: A score metric for evaluating recall-oriented information retrieval applications. In: SIGIR 2010 Proceedings - 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 611–618 (9 2010). https://doi.org/10.1145/1835449.1835551

[7] Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., McClosky, D.: The stanford corenlp natural language processing toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 55–60. Association for Computational Linguistics (2014). https://doi.org/10.3115/v1/P14-5010, http://www.aclweb.org/anthology/P14-5010

[8] Mase, H., Matsubayashi, T., Ogawa, Y., Iwayama, M., Oshio, T.: Proposal of two-stage patent retrieval method considering the claim structure. ACM Transactions on Asian Language Information Processing **4**(2), 190–206 (Jun 2005). https://doi.org/10.1145/1105696.1105702, http://doi.acm.org/10.1145/1105696.1105702

[9] Suzuki, S., Takatsuka, H.: Extraction of keywords of novelties from patent claims. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. pp. 1192–1200. The COLING 2016 Organizing Committee (2016), http://www.aclweb.org/anthology/C16-1113

[10] Verberne, S., d'Hondt, E.: Prior art retrieval using the claims section as a bag of words. In: et al., C.P. (ed.) CLEF 2009 Workshop, Part I, LNCS 6241. pp. 498–502. Springer-Verlag Berlin Heidelberg (2010)

Legal Knowledge and Information Systems
M. Palmirani (Ed.)

doi:10.3233/978-1-61499-935-5-215

K-Means Clustering for Controversial Issues Merging in Chinese Legal Texts

Xin Tian [a], Yin Fang [a], Yang Weng [a], Yawen Luo [b], Huifang Cheng [c] and Zhu Wang [b,1]
[a] *College of Mathematics, Sichuan University*
[b] *Law School, Sichuan University*
[c] *China Justice Big Data Institute*

Abstract. In the fact of growing number of cases, Chinese courts have gradually formed a trial mode to improve the efficiency of trials by conducting trials around the controversial issues. However, identifying the controversy issue in specific cases is not only affected by the uncertainty of facts and laws, but also by the discretion of the judges and extra-case factors, and cannot be expressed as a standard format, which lead to the controversial issues based case retrieval a challenge problem. In this paper, we propose a controversial issues merging algorithm based on K-means clustering for Chinese legal texts. The proposed algorithm can determine the number of clusters of the given cause of action automatically and merge the controversial issues semantically, which makes the case information retrieval more accurate and effective.

Keywords. information retrieval, K-means clustering, controversial issues

1. Introduction

Over the past 20 years, the number of cases accepted by the people's courts has increased rapidly. From 2013 to 2017, more than 89 million cases were accepted by Chinese people's courts at all levels. In order to cope with the high-speed growth of cases, Chinese courts have gradually formed a trial mode of launching trials around controversial issues to improve the efficiency of trials, and as embodied in judgments, it is mainly to develop reasoning around controversial issues in the reasoning part of the courts [1]. In Chinese adjudicative documents, controversial issues mainly exist in civil and commercial cases, and are also common in administrative cases, but seldom exist in criminal cases.

According to the content of disputes involved, controversial issue can be generally divided into factual controversial issue and legal controversial issue. In terms of trial organization, factual controversial issue can help to focus on fact investigation, while legal controversial issue can help to organize court debate. Both play a role in improving the efficiency of court trial. The controversial issues sorted out, organized to investigate, and debated by the courts during trials will be embodied in the judgment, and become the contents which could mostly restore the scene of court trials and the judgment thoughts of the judges in the judgment.

[1] Corresponding Author: wangzhu@scu.edu.cn.

It should also be noted that judges' summary about controversial issues is unformatted. As a verbal interaction aiming at multiple parties, the summary of controversial issues reflects the judges' skill in using laws and court trial rules to ascertain the facts of a case. The determination of controversial issues in individual cases is influenced not only by the uncertainty of facts and laws, but also by the discretion of administrative judges and extrajudicial factors, so it is impossible to express controversial issues in a precise format.

Due to the limited access to cases by judges, it is very difficult for the judges to draw lessons from the experience of other judges in summarizing and discussing controversial issues, except from the cases tried or discussed by themselves, and this has greatly hindered the accumulation of legal knowledge and the dissemination of judges' experience. It is difficult to retrieve unformatted controversial issues with key words, so the combination of homogeneous controversial issues has become the basis for judges to retrieve similar controversial issues [2].

As a matter of fact, the controversial issues are finite for a case in same cause of action, but it is difficult to distinguish the similar controversial issues due to huge corpus and different expressions. Therefore, we need machine learning algorithm for identifying the similar controversial issues in a large legal corpus. Most legal information is expressed in text, such as facts of the case, laws and rules, etc. Firstly, we should transform this semantic information into vector space. Several approaches are utilized for the semantic vectorization, such as *term frequency-inverse document frequency* (TF-IDF) [3], *Latent semantic analysis* (LSA) [4], Word2Vec and Doc2Vec. Unsupervised learning is one of the machine learning task of inferring a function that describes the structure of unlabeled data, and the clustering is a form of unsupervised learning that classifies data into different classes or clusters automatically.

In this paper, we extract controversial issues from the Chinese adjudicative documents of personality right disputes in 2017 and introduce four classes of controversial issues including repeated cause of action, general procedure law, general substantive laws and non-general substantive laws controversial issue or factual controversial issue group. The first three controversial issues were extracted by regular expression, then the last one was extracted by machine learning. Experiments show that semantic-based methods capture the semantic information in the text, whose clustering precision is higher than others.

2. Controversial Issues Overview

The courts divide controversial issues into factual controversial issues and legal controversial issues during court trials, mainly aiming to ascertain the facts first, and then proceed to legal reasoning. But from the perspective of referential property to other judges, some legal controversial issues may not have reference significance, and they are of limited types and general-purpose, so they may be sorted out first artificially. However, the combination of controversial issues based on machine learning mainly aims at the non-general substantive laws controversial issues which are not sorted out in advance and the factual controversial issues. We divide controversial issues in judgments into four types:

First type: Controversial issue group of repeated cause of action (G1). Such controversial issues are featured by that, upon the request of the parties concerned, the judges consider that the issues with the nature of controversial issue are actually the causes of action involved in the cases. For example, in the disputes over portrait right, the expression Does the defendant infringe upon the plaintiffs portrait right? is clearly directed at the cause of action. Similar controversial issues may be combined directly.

Second type: Controversial issue group of general procedure law (G2). Such controversial issues are featured by that, in different causes of action, there will probably be similar procedural controversial issue group. For example, for such controversial issues of Is the plaintiff a competent subject?, there must be clear regulations in the Civil Procedure Law, so it is available to sort out the controversial issues by type first before technical combination.

Third type: Controversial issue group of general substantive laws (G3). Such controversial issues are featured by that, the judges make value judgment on whether minor premise (facts of a case) meets major premise (legal provisions) according to the clearly expressed provisions of law. For example, for such cause of action, like Is the amount of compensation claimed by the plaintiff reasonable?, there must be clear regulations in substantive laws, which widely exist in different causes of action, so it is worth sorting out the causes of action before technical combination.

Fourth type: non-general substantive laws controversial issue and factual controversial issue group (G4). Different causes of action involve different non-general substantive laws controversial issue and factual controversial issue groups. Wherein, the level-four causes of actions under the same level-three cause of action are possibly repeated with the factual controversial issue group of the level-three cause of action, and have relatively great retrieval value for similar cases. Non-general substantive laws controversial issues have relatively great reference significance. Such causes of action are mainly sorted out by relying on machine learning.

3. Feature Extraction and K-means

Feature extraction aims to transfer text to a vector which represents the feature of the text. There are many methods which can capture different features of text, such as frequency-based (TF-IDF, LSA) and network-based (Word2Vec, Doc2Vec). In fact, the network-based methods are more appropriate because the combination of controversial issues are based on semantics and the network-based methods can capture semantic feature exactly. For instance, Word2Vec expresses the semantic information of words by learning huge corpus and makes the embedding vectors of similar words more compact as well [5]. Doc2Vec is an extension of Word2Vec that is designed to get an embedding vector of documents [6]. The difference between them is that Doc2Vec adds the identification information of the document, and it can be regarded as the topic of the document.

K-means is the most important hard clustering algorithm [7]. It aims to assign the data set into K clusters, where the value of K is already known. As we know, K-means is a simple and efficient clustering method, but a big disadvantage of this method is that it needs to determine the number of clusters first. In reality, often K is nothing more than a good guess based on experience or domain knowledge. It is similar in merging controversial issues. In cases of small data sets, we can give an ideal K with the help of some legal experts, but we can not do so in huge data sets. In this part, we will focus on a method for determining an appropriate K.

Schtze et al. introduced a heuristic method in the "*Introduction of Information Retrieval*", which can capture some possible values of K [8]. In more details, we first perform i (e.g. $i = 10$) clusters with a fixed K (Note that you should initialize each one differently) and compute the objective functions of each cluster. Then the minimum of these objective function values can be denoted by $J_{min}(K)$. Now, we implement this process and compute $J_{min}(K)$ with the increase of K. Finally we can capture a series of $J_{min}(K)$ with different K, and find the "knee" in the curve - the point where the successive decreases in J_{min} become noticeably smaller.

4. Experiments

In this part, the experiments are implemented on the adjudicative documents of personality right disputes in 2017 to demonstrate the performance of our approach. We attempt to prove that the feature model can capture semantic information, and meanwhile the clustering algorithm can also assign controversial issues according to our expectations, especially, an appropriate K can be given according to the heuristic method.

Throughout the experiments, we utilize two indicators to measure the degree of conformity between ground truth clusters and our evaluation from algorithm outputs: *adjusted mutual information* (AMI) and *V-measure* [9]. Another measure is V-measure [10]. V-measure is an entropy-based measure which explicitly measures how successfully the criteria of homogeneity and completeness have been satisfied.

The proposed method is implemented on the cause of action of portrait right disputes belongs to disputes over personality right. There are 87 controversial issues in the cause of action of portrait right disputes that were extracted by regular expression and then clustered. In more details, the component we set will be retained at 50 in LSA, which means the dimension of the text feature vector from LSA is 50. We utilized huge corpus to train Word2Vec vector embedding including two parts: Controversial issues of disputes over private lending, disputes over traffic accident liability for motor vehicles and disputes over personality right, on the other hand, adjudicative documents of personality right disputes which includes the facts affirmed by the court, the reasoning of the courts and the result of judgement.

We set the number of K separately according to "*manual*" and "*heuristic*". When deciding about the "*manual*" number of K, we rely on expert opinion. The ideal value of K we choose is 33, which equals to groundtruth. The "*heuristic*" number of K comes from former method are 31, 29, 27, 40 and 36 respectively. The experimental results are shown in Table 1, we see that, as expected, the Word2vec and Doc2vec based on semantic feature get better results, meanwhile the "*heuristic*" results are comparable with the "*manual*". It further proves that the semantic-based method we used can capture the semantic information inside the document, and the heuristic search of K in K-means is applicable to this problem.

5. Conclusions and Future Works

In this paper, the controversial issues in judgments are divided into four types. The non-general substantive laws controversial issues and the factual controversial issues are

Table 1. Comparison of several methods

Method	AMI	$V - measure$
LSA_manual	0.4263	0.7759
LSA_heuristic	0.4137	0.7689
Word2Vec (dim=50)_manual	0.4318	0.7926
Word2Vec (dim=50)_heuristic	0.4627	0.7864
Word2Vec (dim=100)_manual	0.4603	0.8097
Word2Vec (dim=100)_heuristic	0.4334	0.7727
Doc2Vec (dim=50)_manual	**0.4936**	**0.8229**
Doc2Vec (dim=50)_heuristic	0.4720	**0.8390**
Doc2Vec (dim=100)_manual	**0.5300**	0.8202
Doc2Vec (dim=100)_heuristic	0.4902	0.8184

merged by the proposed machine learning algorithm. More precisely, we utilize network-based word embedding models such as Word2Vec and Doc2Vec to extract text feature and cluster data with K-means. Finally, the results of experiments demonstrate that the proposed algorithm is more effective than baseline. However, we found that there is a hierarchical structure between different cause of action. Therefore, it will be better to introduce some hierarchical clustering algorithms for the intrinsic hierarchical structure of controversial issues in Chinese legal texts.

Acknowledgements

This work was supported by National Key R&D Program of China (No. 2018YFC0830300).

References

[1] CHEN Gui-ming. (2004). Issues Concerning Several Relations in the Design of Pretrial Preliminary Procedure. *The Political Science and Law Tribure*, 23(4), 10-15.

[2] Hu Ya-qiu. (2012). Arrangement Procedure of Controversial Points in the Evolution of Civil Litigation System. *Journal of Soochow University* , 12(3), 58-67.

[3] Salton, G., & McGill, M. J. (1986). Introduction to modern information retrieval.

[4] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6), 391-407.

[5] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781.*

[6] Le, Q., & Mikolov, T. (2014, January). Distributed representations of sentences and documents. *In International Conference on Machine Learning* (pp. 1188-1196).

[7] Lloyd, S. (1982). Least squares quantization in PCM. *IEEE transactions on information theory*, 28(2), 129-137.

[8] Schtze, H., Manning, C. D., & Raghavan, P. (2008). *Introduction to information retrieval* (Vol. 39). Cambridge University Press.

[9] Vinh, N. X., Epps, J., & Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(Oct), 2837-2854.

[10] Rosenberg, A., & Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. *In Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL).*

Legal Knowledge and Information Systems
M. Palmirani (Ed.)

doi:10.3233/978-1-61499-935-5-220

Checking the Validity of Rule-Based Arguments Grounded in Cases: A Computational Approach

Heng ZHENG [a,1], Minghui XIONG [b] and Bart VERHEIJ [a]
[a] *Artificial Intelligence, University of Groningen, The Netherlands*
[b] *Institute of Logic and Cognition, Sun Yat-sen University, Guangzhou, China*

Abstract. One puzzle studied in AI & Law is how arguments, rules and cases are formally connected. Recently a formal theory was proposed formalizing how the validity of arguments based on rules can be grounded in cases. Three kinds of argument validity were distinguished: coherence, presumptive validity and conclusiveness. In this paper the theory is implemented in a Prolog program, used to evaluate a previously developed model of Dutch tort law. We also test the theory and its implementation with a new case study modeling Chinese copyright infringement law. In this way we illustrate that by the use of the implementation the process of modeling becomes more efficient and less error-prone.

Keywords. Artificial Intelligence and Law, Rule-based Reasoning, Case-based Reasoning, Argumentation Modeling, Prolog

1. Introduction

The recent case model formalism [1] is a hybrid theory showing connections between cases, rules and arguments [2]. The formalism defines different ways in which rule-based arguments can be valid in cases: arguments can be coherent, conclusive or presumptive. The formalism has been applied to model Dutch tort law, showing how a rule-based legal domain can be grounded in legal cases. In this way, a formal connection is established between the civil law tradition focusing on rules and the common law tradition focusing on cases.

The present paper provides a computational version of the case model formalism. A Prolog program is presented that can computationally check whether a case model is correct, whether rule-based arguments are valid (in the three kinds of validity coherence, conclusiveness and presumptiveness), and whether defeating circumstances are rebutting, undercutting or undermining. The computational tool can be used to support the manual modeling of a complex legal domain, making that more manageable. As an example, we provide a new domain model, namely Chinese copyright infringement law, both formally (as a case model) and computationally (in Prolog).

[1] Corresponding Author: zhengh48@mail2.sysu.edu.cn.

```
case(model_num(1),case_num(101),[not(dmg)]).
...
case(model_num(1),case_num(104),[not(dut),dmg,unl,imp,not(cau)]).
case(model_num(1),case_num(105),[dut,dmg,unl,imp,cau,vrt,not(vst),not(vun),ift
     ,not(ila),not(ico),not(jus),prp]).
...
case(model_num(1),case_num(114),[not(dut),dmg,not(unl),vrt,not(vst),jus]).
...
case_order(model_num(1),[case_num(101),case_num(102),case_num(103),case_num
     (104),[case_num(105),case_num(106),...,case_num(113)],...]).
```

Listing 1: Definition of the Dutch tort law case model in Prolog

2. The implementation of case model formalism in Prolog with case studies

We have implemented the case model formalism in Prolog. We use the previously developed model of Dutch tort law [2] as an illustration. Cases are represented as Prolog lists, of which the elements consist of strings and their negations (represented using not/1). Case models are represented as lists of cases and their ordering, where case models and cases are referred to using identifiers.

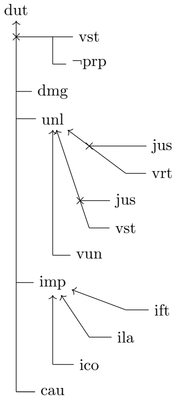
Listing 1 provides a part of the representation of the case model for Dutch tort law (model_num(1)). The full model consists of the 16 cases discussed in [2]. The ordering is represented as a list of cases and lists of cases, each representing an equivalence class of the total preorder, in decreasing level of preference. Hence the first element of this list represents the case or cases that are maximal in the total preorder. Here there is one maximal case case_num(101): not(dmg), representing that there are no damages. When an equivalence class consists of several cases, such as [case_num(105),...,case_num(113)], it is represented as a list of cases (actually: of case identifiers).

The predicate case_model_valid/1 (with a case identifier as single argument) checks whether all cases are consistent, incompatible and different. Checking whether the ordering of cases is total and transitive is not directly represented since we use an explicit representation of the total preorder as a list of equivalence classes. In this representation another validity check is helpful, namely whether each case of which an identifier appears in the ordering is defined and whether each case identifier of a defined case appears in the ordering exactly once. This check has been implemented using the predicate ordering_valid/1.

Arguments are represented by a list of premises and a list of conclusions. The three kinds of validity of arguments can be checked using the predicates coherent/1, conclusive/1 and presumptively_valid/1, each taking an argument (represented using argument/2) as input. Coherence is checked by first determining the case made by an argument, found by appending the list of premises to the list of conclusions, and then checking whether there is a case in the case model that contains all elements of the case made by the argument.

The conclusiveness of an argument is checked by first checking whether the argument is coherent and then checking, if all cases in the case model that contain the argument's premises also contain its conclusions.

An argument's presumptive validity is checked by determining a maximally preferred case that witnesses the argument's coherence (i.e., finding a case in



```
presumptively_valid([dmg,unl,imp,cau],[dut]).
    rebutting_attack(argument([dmg,unl,imp,cau],[dut]),[vst,not(prp)]).

    rebutting_attack(argument([vrt],[unl]),[jus]).
presumptively_valid(argument([vrt],[unl])).
    rebutting_attack(argument([vst],[unl]),[jus]).
presumptively_valid(argument([vst],[unl])).
conclusive(argument([vun],[unl])).

conclusive(argument([ift],[imp])).
conclusive(argument([ila],[imp])).
conclusive(argument([ico],[imp])).
```

Figure 1. The Dutch tort law model: argument structure (left); in Prolog (right)

which both the premises and conclusions of the argument hold and that is maximal in the ordering with this property) and then checking for each case in which the argument's premises hold whether that case is of equal or lower ordering.

Attack of arguments has been implemented using the predicates successful_attack/2, rebutting_attack/2, undercutting_attack/2 and undermining_attack/2. The predicate successful_attack/2 takes an argument and defeating circumstances (as a list) as input. The predicate checks whether the argument is presumptively valid and whether the argument to the same conclusions but from the premises with the defeating circumstances appended is not presumptively valid.

The three kinds of attack—rebutting, undercutting and undermining—are defined in terms of successful attack as follows. Rebutting attack requires a successful attack of an argument such that also the argument from the premises to the opposite of the argument's conclusion is presumptively valid. Note that this only makes sense for arguments with a single element as conclusion as we use no Prolog expression for the negation of a series of conclusions. Successful attacks that are not rebutting attacks are undercutting. Undermining attack is a special kind of successful attack, namely an attack of an argument with a tautology as premise. In the program, a tautology is represented as an empty Prolog list [ ].

The Prolog program can be used to validate hand-made domain models, such as the case model for Dutch tort law of [2]. In Figure 1, we show the argument structure that is valid in the hand-made formal case model of that paper (left). On the right, we show Prolog clauses that all evaluate to true given the Prolog version of that case model (partially shown in Listing 1). In other words, the model is computationally validated.

Another case study is about Chinese copyright infringement. The article of Copyright Infringement in Chinese Criminal Law [3] is below:

> **Article 217** Whoever, for the purpose of making profits, commits any of the following acts of infringement on copyright shall, if the amount of illegal gains is relatively large, or if there are other serious circumstances, be sentenced to fixed-term imprisonment of not more than three years or criminal detention and shall also, or shall only, be fined; if the amount of illegal gains is huge or if there are other especially serious circumstances, he shall be

> sentenced to fixed-term imprisonment of not less than three years but not more than seven years and shall also be fined:
> (1) reproducing and distributing a written work, musical work, motion picture, television programme or other visual works, computer software or other works without permission of the copyright owner;
> (2) publishing a book of which the exclusive right of publication is enjoyed by another person;
> (3) reproducing and distributing an audio or video recording produced by another person without permission of the producer; or
> (4) producing or selling a work of fine art with forged signature of another painter.

According to the articles related to Art. 217 in Chinese criminal law [3], and relevant official judicial interpretations [4], there is a defeating circumstance: the action was not belong to “without permission of the copyright owner”.

In the light of Art. 217 and the judicial interpretations related to it, a case model based on Verheij's case model formalism with a similar modeling approach to the Dutch tort law model in [2] can be built. We use the elementary propositions in Table 1, shown with their formal abbreviations. The full model has 30 cases. A selection of the cases is shown in its case list version in Table 2. The model has identifier model_num(2). In the text below, cases are numbered 1, 2, 3, ... corresponding to cases 201, 202, 203, ... in the Prolog version.

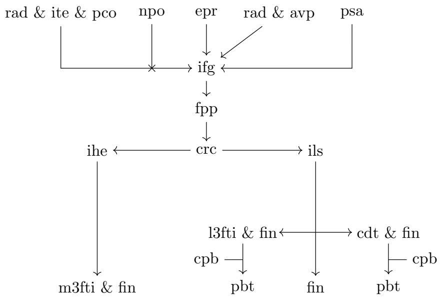
From Chinese copyright infringement, we can analyze the argument structure as in the diagram in Figure 2 (left). This argument structure shows multiple

Table 1. Elementary propositions in the copyright infringement model with abbreviations

ifg	there is a copyright infringement
fpp	the act was for the purpose of making profits
rad	the act was reproducing and distributing something
ite	the act concerned the items in Art. 217:1
pco	the act was without permission of the copyright owner
npo	the act was not belong to ”without permission of the copyright owner”
epr	a book is published of which the exclusive right of publication is enjoyed by another person
avp	a audio or video recording is produced by another person without permission of the producer
psa	a work of fine art with forged signature of another painter is produced or sold
ils	the amount of illegal gains is large or other serious circumstances
ihe	the amount of illegal gains is huge or other especially serious circumstances
crc	the person commits the crime of copyright infringement
l3fti	the person shall be sentenced to fixed-term imprisonment of at most three years
cdt	the person shall be sentenced to criminal detention
fin	the person shall be fined
m3fti	the person shall be sentenced to fixed-term imprisonment of not less than three years but not more than seven years
cpb	the defendant satisfies the conditions of probation
pbt	the defendant will be put on probation

Table 2. The Chinese copyright infringement case model (selection)

Case 1	¬rad,¬ite,¬pco,¬epr,¬avp,¬psa,¬ifg
Case 2	rad,ite,pco,npo,¬ifg
Case 3	rad,ite,pco,¬npo,¬epr,¬avp,¬psa,ifg,¬fpp
Case 4	rad,ite,pco,¬epr,¬avp,¬psa,ifg,fpp,ihe,¬ils,crc,m3fti,¬l3fti,¬cdt,fin
Case 5	rad,ite,pco,¬epr,¬avp,¬psa,ifg,fpp,¬ihe,ils,crc,¬m3fti,¬l3fti,¬cdt,fin
Case 6	rad,ite,pco,¬epr,¬avp,¬psa,ifg,fpp,¬ihe,ils,crc,¬m3fti,l3fti,¬cdt,fin,cpb,pbt
Case 7	rad,ite,pco,¬epr,¬avp,¬psa,ifg,fpp,¬ihe,ils,crc,¬m3fti,¬l3fti,cdt,fin,cpb,pbt
Case 8	rad,ite,pco,¬epr,¬avp,¬psa,ifg,fpp,¬ihe,ils,crc,¬m3fti,l3fti,¬cdt,fin,¬cpb,¬pbt
Case 9	rad,ite,pco,¬epr,¬avp,¬psa,ifg,fpp,¬ihe,ils,crc,¬m3fti,¬l3fti,cdt,fin,¬cpb,¬pbt
...	...
Case order	Case 1 > Case 2 = Case 3 > Case 4 = Case 5 = Case 6 = Case 7 = Case 8 = Case 9



```
presumptively_valid(argument([rad,ite,pco],[ifg])).
presumptively_valid(argument([crc],[ils])).
presumptively_valid(argument([crc,ils],[l3fti,fin])).

rebutting_attack(argument([rad,ite,pco],[ifg]),[npo]).

conclusive(argument([crc,ihe],[m3fti,fin])).
conclusive(argument([ifg,fpp],[crc])).
conclusive(argument([crc,ils,l3fti,fin,cpb],[pbt])).
conclusive(argument([crc,ils,cdt,fin,cpb],[pbt])).
```

Figure 2. The Chinese copyright infringement model: argument structure (left); in Prolog (right)

rule-based steps and an exception-based attack. The structure is valid in the case model we built. Following the definitions of the case model formalism, the arguments in Figure 2 (right) can be extracted in the model. The Prolog program confirms the validity of these arguments. These Prolog queries are evaluated as true, which means the results of the Prolog program correspond to our analysis of the Chinese copyright infringement model.

3. Conclusion

The results of this paper show that an implementation of the case model formalism can be used to support the modeling of a legal domain with a complex argument structure involving combined support and attack[2]. In this way, we have shown a computational connection between cases, rules and arguments, applied to the civil law system of the Netherlands and to criminal law in the Chinese legal system. AI and legal reasoning technology needs to combine rule-based reasoning, case-based reasoning and argumentation together, paving the way for argumentation technology that bridges cases and rules, as it is common in the law.

References

[1] B. Verheij. Correct Grounded Reasoning with Presumptive Arguments. In L. Michael and A. Kakas, editors, *15th European Conference on Logics in Artificial Intelligence, JELIA 2016. Larnaca, Cyprus, November 9–11, 2016. Proceedings (LNAI 10021)*. Springer, Berlin, 2016.

[2] B. Verheij. Formalizing Arguments, Rules and Cases. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law (ICAIL 2017)*, 2017.

[3] The National People's Congress of the People's Republic of China. Criminal Law of the People's Republic of China. http://www.npc.gov.cn/englishnpc/Law/2007-12/13/content_1384075.htm, 2011.

[4] The Supreme People's Court of The People's Republic of China, The Supreme People's Procuratorate of the People's Republic of China, and The Ministry of Public Security of the People's Republic of China. Judicial interpretation on the application of law in handling criminal cases involving infringement of intellectual property rights (in Chinese). http://www.court.gov.cn/fabu-xiangqing-2903.html, 2011.

[2] The full program code and the Chinese copyright infringement model are available at https://github.com/Zhe333/Appendix.git

Subject Index

Legal Knowledge and Information Systems
M. Palmirani (Ed.)

Author Index