

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

A dataset for evaluating legal question answering on private international law

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:

Sovrano, F., Palmirani, M., Distefano, B., Sapienza, S., Vitali, F. (2021). A dataset for evaluating legal question answering on private international law. New York : Association for Computing Machinery [10.1145/3462757.3466094].

Availability:

This version is available at: <https://hdl.handle.net/11585/829241> since: 2021-09-27

Published:

DOI: <http://doi.org/10.1145/3462757.3466094>

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>).
When citing, please refer to the published version.

(Article begins on next page)

A Dataset for Evaluating Legal Question Answering on Private International Law

FRANCESCO SOVRANO, University of Bologna - DISI, Italy

MONICA PALMIRANI, University of Bologna - CIRSFID, Italy

BIAGIO DISTEFANO, Universität Wien, Austria

SALVATORE SAPIENZA, University of Bologna - CIRSFID, Italy

FABIO VITALI, University of Bologna - DISI, Italy

International Private Law (PIL) is a complex legal domain that presents frequent conflicting norms between the hierarchy of legal sources, legal domains, and the adopted procedures. Scientific research on PIL reveals the need to create a bridge between European and national laws. In this context, legal experts have to access heterogeneous sources, being able to recall all the norms and to combine them using case-laws and following the principles of interpretation theory. This clearly poses a daunting challenge to humans, whenever Regulations change frequently or are big-enough in size. Automated reasoning over legal texts is not a trivial task, because legal language is very specific and in many ways different from a commonly used natural language. When applying state-of-the-art language models to legalese understanding, one of the challenges is always to figure how to optimally use the available amount of data. This makes hard to apply state-of-the-art sub-symbolic question answering algorithms on legislative texts, especially the PIL ones, because of data scarcity. In this paper we try to expand previous works on legal question answering, publishing a larger and more curated dataset for the evaluation of automated question answering on PIL.

CCS Concepts: • **Applied computing** → **Law**; • **Computing methodologies** → **Reasoning about belief and knowledge**; **Information extraction**.

Additional Key Words and Phrases: Legal Question Answering, Private International Law, Knowledge Graph Extraction

ACM Reference Format:

Francesco Sovrano, Monica Palmirani, Biagio Distefano, Salvatore Sapienza, and Fabio Vitali. 2021. A Dataset for Evaluating Legal Question Answering on Private International Law. In *Proceedings of Eighteenth International Conference for Artificial Intelligence and Law (ICAIL '21)*. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3462757.3466094>

1 INTRODUCTION AND BACKGROUND

The International Private Law (PIL) is a complex legal domain that presents frequent conflicting norms between the hierarchy of legal sources (e.g., national vs. European level), between legal domains (e.g., consumer law vs. labour law), between the adopted procedures (e.g., alternative dispute resolution vs. litigation). Scientific research on PIL reveals the need to create a bridge between European and national laws on this domain by accessing heterogeneous legal sources. The European project Interlex¹ intended to investigate this domain and to use technology to fill the gap between different legal sources. This need to rely on technology is due to the complexity of the PIL domain. In fact,

¹<http://www.interlexproject.eu/index.html>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2021 Association for Computing Machinery.

Manuscript submitted to ACM

in this context, legal experts have to access heterogeneous sources, being able to recall all the norms and to analyse them using case-law² and following the principles of interpretation theory. This poses a daunting challenge to humans, whenever Regulations change frequently or are big-enough in size. In fact, searching within thousands and thousands of pages of legal documents from different sources and jurisdictions is undoubtedly a task requiring human effort and specialised expertise. This is probably one of the reasons why researchers, governments and industry have for long looked for a way to build “intelligent” machines capable of helping humans in detecting the relevant legal provisions over such complex corpora [11]. In literature we may find at least two distinct main approaches to reasoning and artificial intelligence. The first approach is more symbolic and formal, capable to model legal knowledge into a formal representation. For example, the legal ontology modelling method [4, 8] is a relevant instrument for defining the legal concepts and relationships included in legal texts (e.g., hard law, judgement, soft law, etc.) but it is extremely expensive, it depends on the hermeneutic approach adopted by each scholar or community (e.g., common law vs. civil law), it is influenced by a strong localisation due to the local jurisdiction (e.g., domestic regulation and local court action), by the cultural and social norms (e.g., concept of gender) and, furthermore, modifications in the legal framework (e.g., new legislation) require a refinement or (even worse) a whole extension of the ontology is required. The second approach is the most recent and in many ways the most versatile, but sub-symbolic and opaque. A sub-symbolic approach is said to be more data-oriented and it follows the recent success of Deep Neural Networks (DNNs) on natural language processing and understanding. Current state-of-the-art on natural language understanding is heavily based on this data-centred approach, and many models specifically applied to legalese have already been published. For example in 2018 [3] published a framework for natural language processing and information extraction for legal and regulatory texts. In 2019 [5] proposed one of the first models for legal word embeddings. While, in 2015 Kim et al. [10] presented one of the very first algorithms based on DNNs for Legal Question Answering (reasoning) applied to a dataset of Boolean questions from Japanese legal bar exams, then followed up by [7] and others [9, 12]. In 2020, Sovrano et al. [13] proposed a novel and hybrid approach for legal question answering on PIL, using a legal ontology based on Ontology Design Patterns (like agent, role, event, temporal parameter, action) in order to mirror the legal significance of the relationships within and among the provisions. More generally, automated reasoning over legal texts (not just the PIL's ones) is not a trivial task, due to the fact that the legal jargon (legalese) is less frequent and more ambiguous than commonly-used natural language. This is probably the reason why some works have decided to focus on corpora, such as privacy policies [12], with a legal language that would be more similar to its natural counterpart, or to focus on more argumentative texts (e.g. sentences, procedural documents, cross-examinations, parliamentary court reports) instead of legislative texts or contracts. Anyway, this challenge makes hard to apply state-of-the-art sub-symbolic question answering algorithms on legislative texts, especially the PIL ones, because of data scarcity or novel topics introduced for the first time in the legal system (e.g., no historical series).

With this work we are interested into advancing on automated answering to questions written in legalese and on PIL legislative texts. Our goal is to be able to properly evaluate canonical question answering techniques for PIL. This is why we try to expand the work presented by Sovrano et al. in [13], publishing a larger and more curated dataset extracted from Regulations such as: Rome I Regulation EC 593/2008; Rome II Regulation EC 864/2007; and Brussels I bis Regulation EU 1215/2012.

In Section 2 we describe our dataset, and the methodology we followed to design it. While in Section 3 we analyse the results obtained by re-running the experiment of [13] on the new dataset, pointing to future work in Section 4.

²<http://www.interlexproject.eu/del/Deliverable2dot3.pdf>

2 A DATASET FOR EVALUATING LEGAL QUESTION ANSWERING ON PIL

In this Section, we explain how we expanded the dataset presented in [13], doubling its size. We improved over [13], publishing a larger and more curated dataset for the evaluation of automated question answering on PIL.

Both the old and the new dataset were extracted from the following Regulations, in English:

- Rome I Regulation EC 593/2008;
- Rome II Regulation EC 864/2007;
- and Brussels I bis Regulation EU 1215/2012.

These regulations are, respectively, on the law applicable to contractual obligations; on the law applicable to non-contractual obligations; on jurisdiction and the recognition and enforcement of judgements in civil and commercial matters. These Regulations aim to provide a tool for identifying the applicable law and the jurisdiction in cases when two or more legal systems connect and generate complex relationships (e.g. a sale of goods contract between an Italian and a German citizen regarding commodities situated in Spain).

It is important to highlight the fact that, for the construction of the new dataset, we decided to inherit some methodological choices from [13], considering PIL as a subject simply from the point of view of these three EU Regulations, as a self-contained environment, i.e., excluding references to other international conventions and general principles. So that it is possible to evaluate Q&A techniques with respect to their ability to handle the *general principles* in the recitals, the *scope of application* in the initial articles, and the *specific cases* (e.g. exceptions) in the other articles. The methodological choices we kept raised some issues with regard to the formulation of the questions and their relevance. Conceptual questions (e.g. “What is a non-contractual obligation?”) cannot be fully answered by relying solely on these 3 Regulations, as the goal of this legislation - when considered atomistically - is limited to discipline conflict of law and conflict of jurisdiction cases. While the Regulations, as with any other piece of legislation, rely somewhat on external definitions and legal concepts, including those derived from jurisprudence and opinions from commentators, they also define intrinsically and specifically for their own purposes, key concepts (e.g. “judgement” in Art. 2 of Reg. Brussels I-bis). Therefore we decided to exclude any conceptual question but those involving key concepts defined within the Regulations.

The legal question answering tools we are interested in evaluating are meant to be used by practising lawyers, with reasonable - yet, not expert - knowledge of PIL to:

- explore the contents of the Regulations;
- get support in the reasoning concerning large Regulations.

The dataset for evaluating such tools shall comprise a set of questions for each of which there is also a set of expected answers in the form of Articles, Recitals or Commission Statements³. Recitals are considered beside Articles because the user persona could be interested in *prima facie* interpretive tools emerging from the text itself, let alone the debated bindingness of Recitals. The dataset published in [13], was designed following a methodology that is similar to the one we are going to use for this extension. For the selection of the questions and the identification of the expected answers we adapted to our case a specific methodology encoded by Ashley and others in their works [1, 2, 6] during the last years. This methodology is common to other works in the field and it is meant to validate the experiment also from a legal perspective. In our case, the questions were selected by two legal experts, while other two independent legal experts matching our intended user persona were responsible for identifying the expected answers by relying

³Rome II Regulation contains three Commission Statements meant to bind the EU Commission to publish studies on selected topics

solely on the verbatim information that can be found in the Regulations. Therefore, legal experts were instructed to prevent case-law, general principles or scholar opinions from influencing their answers, as well as requested to avoid interaction with each other. As stated above, the research wants to model only the neutral legislative information from the three Regulations without any interpretation other than the literal one. The inclusion of other knowledge will be left to further research. First, experts read the three Regulations and answered to the questions without any assistance from auxiliary sources, including tools and previous knowledge. Then, they were allowed to compare their answers with those provided by the tool for legal question answering, selecting tool-assisted correct answers and missing replies to be used to calculate performance scores in the later stages. Despite the efforts to draft interpretation-neutral questions, each independent expert has a certain margin of appreciation both when providing her/his answers and when assessing the correctness of the tool-provided answers. Therefore, another intervention was necessary when divergences in their evaluation occurred. When identifying the expected answers, the aggregation kept into account only theoretical replies that were common between the two independent experts. This aggregation was conducted by one legal expert who dispose of a higher level of expertise in comparison to the independent evaluators, yet relying on the same criterion, i.e. literal interpretation only.

At the end of the process we got the 9 new questions shown in Table 2.

The questions were chosen with the following criteria: they had to be sufficiently specific to find adequate answer in the Regulations (we avoided too broad or excessively conceptual questions); the questions needed not to be focused on specific cases but with a reasonable level of abstraction (e.g., instead of “Where can an employee that carries out their work in Spain sue an employee located in Spain, if they had not agreed on the jurisdiction?”, a question such as “Where can an employee sue their employer?”); the questions needed to be sufficiently different from one another (i.e., not asking repetitive questions such as “What is the applicable law in contracts of carriage?” and “What is the applicable law in insurance contracts?”). Some of the questions in the dataset are relatively similar to one another, with some of them being the a more correct specification of another, such as “Which parties of a contract should be protected by conflict-of-law rules?” vs “What is the applicable rule to protect the weaker party of a contract?”

Questions in the dataset are not speculative or *de iure condendo* and are agnostic to elements that are placed outside the Regulation (e.g. jurisprudence, general principles, etc.). As such, they are not meant to nudge towards forms of interpretations other than the literal one (e.g. analogy, principle-based reasoning, *lex specialis*, etc.)

Furthermore, in order to be able to further analyse the results of any evaluation based upon our dataset, we decided to pick an heuristic for classifying questions, that is the *context specificity* (Low, Normal, High), and we applied it also to the old dataset we extended. *Context specificity* is a subjective concept and it is highly dependant on each jurist. For this reason, we opted to use a criterion that would ensure an acceptable level of objectivity. Thus, specific questions whose answer is exactly in the domain of the Regulations were labelled as Highly specific (e.g., “Can the parties choose a different applicable law for different parts of the contract?”); questions whose answer falls in part within the scope of the Regulations, but somewhat relies on external concepts were labelled as Normally specific (e.g., “Which parties of a contract should be protected by conflict-of-law rules?”); finally, broad questions whose answer requires the significant use of external legal concepts and resources and whose answer is found through an articulate combinations of articles and recital were labelled as having Low specificity (e.g., “How should a contract be interpreted according to this regulation?”).

Of the 17 questions that compose the new extended dataset: 29.41% have a Low specificity; 35.29% have a Normal specificity; 35.29% have a High specificity.

Table 1. First block of answers (ordered by the pertinence to the question estimated by the tool) given by the baseline to the questions in [13]. “B” stands for Brussels, “RI” for Rome I and “RII” for Rome II. “Rec.” stands for Recital, “Art.” for Article, and “Stat.” for Commission Statement. For each answer, the top5 scores (precision, recall, F1) are shown. In the “scores” columns: “P” stands for Precision and “R” stands for Recall. In the “Specificity” column: “L” stands for Low, “N” stands for Normal and “H” stands for High.

Question	Specificity	Expected Answers	Baseline’s Top5	Baseline’s Scores
Who determines disputes under a contract?	L	<u>B Art. 7.1</u> , <u>B Art. 8.3</u> , <u>B Art. 8.4</u> , <u>B Art. 17</u>	<u>RI Rec. 12</u> , <u>B Art.17.2</u> , <u>RI Rec. 24</u>	R: 25% P: 33% F1: 28.44%
What factors should be taken into account for conferring the jurisdiction to determine disputes under a contract?	N	<u>B Art. 7.1</u> , <u>B Art. 17</u> , <u>B Art. 20</u> , <u>B Art. 25</u>	<u>RI Rec. 12</u> , <u>B Art.25</u> , <u>B Art.25.5</u> , <u>B Rec.15</u> , <u>RI Rec. 21</u>	R: 25% P: 40% F1: 30.76%
Which parties of a contract should be protected by conflict-of-law rules?	N	<u>RI Rec. 23</u> , <u>RI Art. 6</u> , <u>RI Art. 8</u> , <u>RI Art. 13</u>	<u>RI Rec. 23</u> , <u>B Rec.18</u> , <u>RI Rec. 24</u> , <u>RI Art.25.1</u> , <u>RI Rec. 27</u>	R: 25% P: 20% F1: 22.22%
In which case claims are so closely connected that it would be better to treat them together in order to avoid irreconcilable judgments?	H	<u>B Art. 8</u> , <u>B Art. 30</u> , <u>B Art. 34</u>	<u>B Art. 8.1</u>	R: 33% P: 100% F1: 49.62%
What kind of agreement between parties are regulated by these Regulations?	L	<u>B Rec. 6</u> , <u>B Rec. 10</u> , <u>B Rec. 12</u> , <u>B Art. 1</u> , <u>RI Rec. 7</u> , <u>RI Art. 1</u>	<u>B Art.73.3</u> , <u>B Rec. 12</u> , <u>B Rec. 36</u> , <u>B Art.71.2</u> , <u>B Art. 71.1</u>	R: 20% P: 20% F1: 20%
In which court is celebrated the trial in case the employer is domiciled in a Member State?	H	<u>B Art. 21</u> , <u>B Art. 22</u> , <u>B Art. 23</u>	<u>B Art. 21.1</u> , <u>B Art.22.1</u> , <u>B Art. 21.2</u> , <u>B Art. 20.1</u> , <u>B Art. 20.2</u>	R: 66% P: 60% F1: 62.85%
How should a contract be interpreted according to this regulation?	L	<u>RI Rec. 22</u> , <u>RI Rec. 12</u> , <u>RI Rec. 26</u> , <u>RI Rec. 29</u> , <u>RI Art. 12</u>	<u>RI Art. 10.1</u> , <u>RI Rec.17</u>	R: 0% P: 0% F1: 0%
Which law is applicable to a non-contractual obligation?	N	<u>RII Rec. 17</u> , <u>RII Rec. 18</u> , <u>RII Rec. 26</u> , <u>RII Rec. 27</u> , <u>RII Rec. 31</u> , <u>RII Art. 4-20</u>	<u>RI Art. 8.1</u> , <u>RII Art.15</u> , <u>RII Art.16</u> , <u>RII Art.8.1</u> , <u>RII Rec. 22</u>	R: 60% P: 60% F1: 60%

3 DATASET ANALYSIS

In order to understand the behaviour of existing question answering tools on the new dataset, we repeated on it the experiment described in [13], changing the metrics used for the evaluation. Considering that we are not interested in the order answers are ranked, as metric for estimating the performance of the algorithm we chose: top5-recall, top5-precision and top5-F1, defined as follows. Let m be the number of strictly-correct answers that are produced as output by the algorithm, let $|E|$ be the number of expected answers for a question, let $|A|$ be the number of given answers to a question, then the top5-recall is given by $\frac{m}{\min(|E|, 5)}$, while the top5-precision is given by $\frac{m}{\min(|A|, 5)}$. The top5-recall is a measure of how many relevant answers are selected by the algorithm in the top five answers, while the

Table 2. Second block of expected answers and answers given by the baseline. See the caption of Table 1 for more details about how to read this table.

Questions	Specificity	Expected Answers	Baseline's Top5	Baseline's Scores
Can the parties choose the applicable law in consumer contracts?	H	<u>RI Rec. 11</u> , <u>RI Rec. 25</u> , <u>RI Rec. 27</u> , <u>RI Art. 6</u>	<u>B Art. 18.2</u> , <u>B Art. 18.1</u> , <u>RI Rec. 28</u> , <u>RI Art. 5.2</u> , <u>RI Art. 6.2</u>	R: 25% P: 20% F1: 22.22%
What factors should be taken into account for conferring the jurisdiction to determine disputes under a consumer contract?	N	<u>B Rec. 18</u> , <u>B Art. 17</u> , <u>B Art. 18</u> , <u>B Art. 19</u> , <u>B Art. 26</u>	<u>RI Rec. 12</u> , <u>RI Rec. 24</u> , <u>B Art. 19</u> , <u>B Art. 17.1</u> , <u>B Art. 25.5</u>	R: 40% P: 40% F1: 40%
Can the parties choose a different applicable law for different parts of the contract?	L	<u>RI Rec. 11</u> , <u>RI Art. 3.1</u>	<u>RI Art. 3.1</u> , <u>RI Art. 5.2</u> , <u>RI Art. 7.3</u> , <u>RII Art. 25.2</u> , <u>RI Art. 22.2</u>	R: 50% P: 20% F1: 28.57%
What non-contractual obligations fall into the scope of Regulation Rome II?	H	<u>RII Rec. 10</u> , <u>RII Rec. 11</u> , <u>RII Art. 1</u> , <u>RII Art. 2</u>	<u>RII Stat. 1</u> , <u>RI Rec. 7</u>	R: 0% P: 0% F1: 0%
What is the applicable rule to protect the weaker party of a contract?	N	<u>RI Rec. 23</u> , <u>B Rec. 18</u>	<u>RI Rec. 23</u> , <u>B Rec. 18</u>	R: 100% P: 100% F1: 100%
What is the applicable law to determine the validity of consent?	L	<u>RI Art. 3.5</u> , <u>RI Art. 10</u> , <u>RI Art. 11</u> , <u>RI Art. 13</u>	<u>RI Art. 3.5</u> , <u>RI Art. 10.2</u> , <u>RI Art. 10.1</u> , <u>B Rec. 20</u>	R: 50% P: 75% F1: 60%
When are two actions to be considered related according to the Regulation Brussels I Bis?	N	<u>B Rec. 21</u> , <u>B Art. 30.3</u>		R: 0% P: 0% F1: 0%
What court has jurisdiction in case of a counter-claim?	N	<u>B Art. 8.3</u> , <u>B Art. 14.2</u> , <u>B Art. 18.3</u> , <u>B Art. 22.2</u>	<u>B Art. 18.3</u> , <u>B Art. 14.2</u> , <u>B Art. 22.2</u> , <u>B Art. 8</u> , <u>B Art. 24</u>	R: 100% P: 80% F1: 88.88%
Where can an employee sue their employer?	H	<u>B Rec. 14</u> , <u>B Rec. 18</u> , <u>B Art. 21.1</u> , <u>B Art. 22.1</u> , <u>B Art. 23</u>	<u>B Art. 21.1</u>	R: 20% P: 100% F1: 33.33%

top5-precision is a measure of how many selected answers in the top five are relevant. Knowing the top5-recall and the top5-precision, it is easy to compute the top5-F1 score by following the formula.

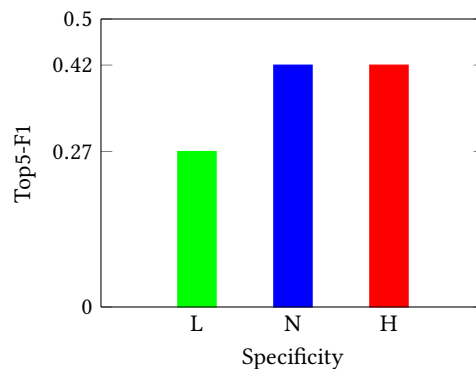
After running the experiment we computed the average top5-F1 for all the questions in the dataset presented in Section 2 (that is the old dataset of [13] plus our new extension). The results on the whole dataset are a Top5-Recall of 37.58%, a Top5-Precision of 45.17% and a Top5-F1 of 38.05%.

We also performed an error analysis taking under consideration how top5-F1 scores vary when the *context specificity* change, expecting that questions with low *context specificity* are harder to answer correctly.

Results partly confirmed our expectations. In fact, we can observe a trend where top5-F1 scores increase proportionally to the *context specificity*. Our expectations were based on the fact that:

- the specificity of a question is low when it asks something that is not closely related to the Regulations;
- multi-hop reasoning is usually required to answer questions with a low specificity, but the baseline is not equipped for that kind of reasoning (yet).

Fig. 1. Average top5-F1 scores for each class of *context specificity*: Low, Normal, High. Scores are respectively: 27.40%, 42.16%, 42.81%



For example, the question “How should a contract be interpreted according to this regulation?” has a very low specificity and it would probably require to pinpoint both recitals and articles for a proper answer, therefore more distinct and distant paragraphs. Probably, most of the speculative questions would require a broader view on the subject matter, having a low specificity to the Regulations, therefore requiring multi-hop reasoning.

4 CONCLUSIONS

With this paper we extended the work presented by Sovrano et al. in [13], proposing a larger and more curated dataset for the evaluation of automated question answering on PIL. In the future we will use these datasets for evaluating new algorithms for question answering, exploiting Akoma Ntoso XML⁴ models of the Regulations, for better capturing the relationships between different portions of the legal hierarchy (e.g. recitals connected via metadata to articles) and also for reusing as much as possible other legal metadata like: i) temporal legal information concerning modifications occurred over time, ii) life-cycle information concerning the history of the regulations, iii) normative references (citations). We also intend to make the question answering tool “aware” of the LegalRuleML ontology⁵ for better handling: obligations, permissions, exceptions, derogations, prohibitions.

ACKNOWLEDGEMENTS

This paper is was conducted with the contribution of CIRSFD-Alma AI and DISI University of Bologna (Interlex Project Grant Agreement Number 800839 and LAILA PRIN2017). The questions were selected by Biagio Distefano, Salvatore Sapienza, while Pier Giorgio Chiara and Noemi Conditì picked the expected answers, separately. All of them are PhD candidates at the “Law, Science and Technology” International PhD program.

REFERENCES

- [1] Kevin D Ashley. 2017. *Artificial intelligence and legal analytics: new tools for law practice in the digital age*. Cambridge University Press.
- [2] Trevor Bench-Capon, Michał Araszkiewicz, Kevin Ashley, Katie Atkinson, Floris Bex, Filipe Borges, Daniele Bourcier, Paul Bourguine, Jack G Conrad, Enrico Francesconi, et al. 2012. A history of AI and Law in 50 papers: 25 years of the international conference on AI and Law. *Artificial Intelligence and Law* 20, 3 (2012), 215–319.

⁴<http://docs.oasis-open.org/legaldocml/akn-core/v1.0/akn-core-v1.0-part1-vocabulary.html>

⁵<http://docs.oasis-open.org/legalruleml/legalruleml-core-spec/v1.0/cs02/rdfs/>

- [3] Michael J Bommarito II, Daniel Martin Katz, and Eric M Detterman. 2018. LexNLP: Natural language processing and information extraction for legal and regulatory texts. *arXiv preprint arXiv:1806.03688* (2018).
- [4] Pompeu Casanovas, Monica Palmirani, Silvio Peroni, Tom Van Engers, and Fabio Vitali. 2016. Semantic web for the legal domain: the next step. *Semantic Web* 7, 3 (2016), 213–227.
- [5] Ilias Chalkidis and Dimitrios Kampas. 2019. Deep learning in law: early adaptation and legal word embeddings trained on large corpora. *Artificial Intelligence and Law* 27, 2 (2019), 171–198.
- [6] Jack G Conrad and John Zeleznikow. 2013. The significance of evaluation in AI and law: a case study re-examining ICAIL proceedings. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law*. 186–191.
- [7] Phong-Khac Do, Huy-Tien Nguyen, Chien-Xuan Tran, Minh-Tien Nguyen, and Minh-Le Nguyen. 2017. Legal question answering using ranking SVM and deep convolutional neural network. *arXiv preprint arXiv:1703.05320* (2017).
- [8] Meritxell Fernández-Barrera and Giovanni Sartor. 2011. The legal theory perspective: doctrinal conceptual systems vs. computational ontologies. In *Approaches to Legal Ontologies*. Springer, 15–47.
- [9] Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2020. A Dataset for Statutory Reasoning in Tax Law Entailment and Question Answering. *arXiv preprint arXiv:2005.05257* (2020).
- [10] Mi-Young Kim, Ying Xu, and Randy Goebel. 2015. A convolutional neural network in legal question answering. In *JURISIN Workshop*.
- [11] Friedrich V Kratochwil. 1991. *Rules, norms, and decisions: on the conditions of practical and legal reasoning in international relations and domestic affairs*. Number 2. Cambridge University Press.
- [12] Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. 2019. Question answering for privacy policies: Combining computational and legal perspectives. *arXiv preprint arXiv:1911.00841* (2019).
- [13] Francesco Sovrano, Monica Palmirani, and Fabio Vitali. 2020. Legal Knowledge Extraction for Knowledge Graph Based Question-Answering. In *Legal Knowledge and Information Systems: JURIX 2020. The Thirty-third Annual Conference*, Vol. 334. IOS Press, 143–153.