Combining shallow and deep learning approaches against data scarcity in legal domains

SyntagmTuner: Combining Shallow and Deep Learning Approaches against Data Scarcity in Legal Domains

ARTICLE INFO

Keywords:
Data Scarcity
Deep Learning
TF-IDF
Syntagmatic Relations
Law

ABSTRACT

We are recently witnessing a radical shift towards digitisation in many aspects of our daily life, including law, public administration and governance. This has sometimes been done with the aim of reducing costs and human errors by improving data analysis and management, but not without raising major technological challenges. One of these challenges is certainly the need to cope with relatively small amounts of data, without sacrificing performance. Indeed, cutting-edge approaches to (natural) language processing and understanding are often *data-hungry*, especially those based on deep learning. With this paper we seek to address the problem of data scarcity in automatic Legalese (or legal English) processing and understanding. What we propose is an ensemble of shallow and deep learning techniques called SyntagmTuner, designed to combine the accuracy of deep learning with the ability of shallow learning to work with little data. Our contribution is based on the assumption that Legalese differs from its spoken language in the way the meaning is encoded by the structure of the text and the co-occurrence of words. As result, we show with SyntagmTuner how we can perform important tasks for e-governance, as multi-label classification of the United Nations General Assembly (UNGA) Resolutions or legal question answering, with data-sets of roughly 100 samples or even less.

1. Introduction

The new emphasis on digitisation of data applies to much more than private documents and services. More and more aspects of public life and of the services provided by public administrations and governments come into play. Indeed, the data-centred paradigm of businesses and services on our societies is impressive, raising new significant issues on how feeble is the border between truth and convenience, between democracy and regimes (Farkas and Schou, 2019; Hinds, Williams and Joinson, 2020). In this context, many advanced countries are trying to automate several aspects of governance (Jaeger and Thompson, 2003) in order to reduce costs and minimise human errors, enforcing in many cases important principles such as transparency, lawfulness and fairness, as stated for instance in the General Data Protection Regulation (GDPR) and many other official documents (Basu, 2004; Jaeger and Bertot, 2010). An example among many is the European Commission actively supporting e-government both at the national level and at its own supranational level (EU-Commission, 2018).

This need for digitisation has grown into ad-hoc interdisciplinary efforts in law, computer science, engineering, resulting in the birth of institutions like the Ministry of Digital Governance of Greece, the Australian Digital Transition Agency, and long-term plans such as the European Digital Transition Action Plan and many others. In many ways, e-governance aims to put the benefits of citizens at the centre, through a more effective and efficient social organisation, while also ensuring fair and lasting public access to government information. (Jaeger and Bertot, 2010).

At the same time, data-centred Artificial Intelligence can be considered a fundamental paradigm for effective intelligent interaction between machines and humans, and this kind of interaction is probably the biggest part of what e-governance is about. Data-centred AI is capable of learning from data and is designed to be able to capture huge amounts of data and improve its results in proportion to the quantity and quality of the data acquired. Deep Learning technologies such as Deep Neural Networks (DNNs) can now perform tasks previously approachable only by biological intelligence, pursuing long-term objectives through cunning strategies. Nonetheless, this impressive level of performance seems to be offset by a lack of transparency, as trained DNNs are defined by (gigabytes of) opaque numbers and data with no apparent symbolic interpretation usable for human understanding.

Given the copious amount of textual documents in governance, some of the main applications of Artificial Intelligence to e-governance encompass text classification, information extraction, information retrieval and content generation. Applications span from data analysis and enrichment of natural language documents, to assistants for content generation and understanding, to tools for litigation mitigation, etc. Natural language processing/understanding is of

ORCID(s):

utmost importance in this domain, and many works have focused on general-purpose state-of-the-art trained language models over government documents of various kinds. These language models are then specialised and applied in ad-hoc applications for text classification, question answering, etc. (e.g., see (Shao, Mao, Liu, Ma, Satoh, Zhang and Ma, 2020; Condevaux, Harispe, Mussard and Zambrano, 2019; Vink, Netten, Bargh, van den Braak and Choenni, 2020))

Natural language processing for legal texts has recently raised a lot of interest, highlighting "the need to create a bridge between conceptual questions, such as the role of legal interpretation in mining and reasoning, as well as computational and engineering challenges, such as the handling of big legal data and the complexity of regulatory compliance" (Robaldo, Villata, Wyner and Grabmair, 2019).

There are several recent examples of this interest. For example, in 2018, Bommarito II, Katz and Detterman published a framework for natural language processing and information extraction for legal and regulatory texts. In 2019, Chalkidis and Kampas proposed one of the first models for legal word embeddings, and the Incorporated Council of Law Reporting for England and Wales (ICLR) published Blackstone (ICLR, 2019), a library meant to allow researchers and engineers to automatically extract information from long, unstructured legal texts (such as judgements, skeleton arguments, scholarly articles, Law Commission reports, pleadings, etc.

But, do we always have huge data collections to work with? What if our datasets are smaller, not sufficient for pre-trained general models? This situation is not uncommon: we can think of at least one situation where data scarcity is unavoidable. That is when legal English (or *Legalese*) or other peculiar and specialised variants of natural languages are involved in tasks requiring automated processing or understanding. Specialised language variants have a clear similarity to their corresponding base languages. Thus, fine-tuning a general pre-trained model can certainly handle the parts of Legalese that do not differ much from plain language. Yet, it is hard to believe that fine-tuning general models is sufficient, or even beneficial when the specialised parts of the corpora are just few, and furthermore if they are not consistent across documents and legislations (Chalkidis and Kampas, 2019).

In these language variants it is not uncommon to find out that the minimal training-set that needs to be annotated manually for adequate deep learning tasks ends up being more or less of the same size as the whole corpus, since labels or their definitions change widely across the available data. An example of this could be the corpus of the UNGA Resolutions, written over several years by different authors and with an incredible variety of language constructs and vocabulary choices.

In this work we address the problem of data scarcity in processing and understanding texts written using a variety of Legalese constructs, for e-governance. We base our analysis on the following hypothesis, that the Legalese variant is more or less similar to its base language, and the meaning of a Legalese text, generally, does not deviate much from plain language, but for certain constructions of words. In fact, Legalese is not repetitive, it is often canonical and tends to avoid terms with multiple meanings and rather adopts vocabulary that is used in a punctual way in very specific contexts, as if the sentences they form are governed by very formal rules. For example, in Legalese, highly meaningful fragments tend to never be ambiguous, to have associated definitions, and to make use of combinations of specific nouns and verbs. The application of these formal rules impacts directly on local meanings by constraining the relationships (also known as *syntagmatic* relations) that words have with others when co-occurring in the sequence of writing[1]. In other words, what we hypothesise is that Legalese constructs play with syntagmatic relations in a very unique way and that this fact can be exploited to tackle the data scarcity problem.

If our hypothesis is correct, we can improve the performance of pre-trained general-purpose deep language models on processing and understanding (relatively) small-size document sets by simply combining them with ad-hoc models for capturing the patterns of syntagmatic relations across texts, without the need to re-train the deep language models. The point is that such syntagmatic relations can be identified even with little data, e.g. by techniques such as Term Frequency–Inverse Document Frequency (TF-IDF) (Rajaraman and Ullman, 2011) or Latent Semantic Analysis (Dumais, 2004). Anyway, these simple tools (also known as *shallow embedders*, because they try to embed the shallow meaning of a text in a numeric vector) can be quickly trained in an unsupervised manner on the available data and they can be used for capturing a part of the meaning that in Legalese is encoded into syntagmatic information. On the other hand, more sophisticated tools (e.g., *deep embedders*) (pre-)trained on generic natural language can be used to capture parts of meaning that are not peculiar to the Legalese variant of the base language.

To prove our assumption, we carried out a few experiments that showed that a linear combination (named *SyntagmTuner*) of these two types of embedders may be an appropriate solution to *approximate* effectively and efficiently a solution to many problems typical of e-governance and in which data scarcity is inevitable. More in detail, we focused

[1] Syntagmatic associations indicate compatible combinations of words (i.e. the word "rotten" combined with "apple"), excluding others (i.e. the syntagm "curdled apple").

on the tasks of multi-label text classification of UNGA Resolutions, and on the problem of automatically answering questions on International Private Law, studying how syntagmatic relations impact on performance. The results partly confirmed our hypothesis that the meaning encoded in Legalese by syntagmatic relations is not always captured by general-purpose language models and that techniques like SyntagmTuner may be of great help in this context.

This paper is structured as follows: in Section 2 we give the background information to understand the pipeline presented in Section 3, trying to show enough insights on the Distributional Hypothesis and the application of Natural Language Processing (NLP) and Natural Language Understanding (NLU) in the Legal Domain. In Section 4 we provide a few experiments in support of our main hypothesis, showing the benefits of our contribution. In Section 5 we conclude by discussing the results and pointing to future work.

2. Background: Shallow vs Deep Learning

Nowadays, state-of-the-art NLP has started to rely more and more on statistical semantics, finding its roots in what is called "Distributional Hypothesis", a fairly important discovery that the meaning of words statistically depends on their context and their co-occurrences (Harris, 1954). The hypothesis can be rephrased as following: "a word is characterised by the company it keeps" (Firth, 1957). The use of statistical semantics over text documents has led to the discovery of impressive technologies capable of encoding the meaning of words and documents as numerical representations (*embeddings*), i.e., mathematical objects that can be represented as multi-dimensional points in an Euclidean space, so that classical mathematical operations can be operated on them. For example, by computing the distance (or the cosine similarity) between two of these embeddings it is possible to quantitatively estimate the degree of similarity between the meaning of their corresponding words or document fragments. Distributional hypothesis is arguably one of the fundamental gears behind the astonishing performances of the most recent deep language models. Its impressive compatibility with deep learning technologies is probably the reason why the distributional hypothesis, that originated in Linguistics, it is now receiving attention also in Cognitive Science (McDonald and Ramscar, 2001).

Several techniques exist for learning numerical representations of texts from their occurrence information, some of them are specialised on words while others are on longer snippets of text such as sentences or even whole documents. According to the leveraged distributional information, existing models could be broadly grouped into two categories (Sahlgren, 2008). The first category leverages more on the *syntagmatic relations* between words, which relate to words that co-occur within the same text region (Sun, Guo, Lan, Xu and Cheng, 2015). While the second leverages more on the *paradigmatic relations*, which relate to words that occur within similar contexts but may not co-occur anywhere in the text.

One of the most basic techniques for text embedding is probably Bag of Words (BoW) (Harris, 1954), where a text snippet is represented as a non-ordered set of its words with representation of individual occurrences (i.e., a bag of words), thereby disregarding grammar and even word order but paying attention only to frequency. An example of BoW embedding for the sentence "This sentence is cool even if a sentence" could be:

```
this = 1;
sentence = 2;
is = 1;
and = 0;
gibberish = 0;
cool = 1;
even = 1;
if = 1;
a = 1.
```

As we can see, BoW tokenizes the documents, it counts the occurrences of tokens, and then it returns them as a (sparse) matrix.

Another important technique for text embeddings is TF-IDF (Jones, 1972). TF-IDF is computed as the product of two statistics: *Term Frequency* (TF) and *Inverse Document Frequency* (IDF). Term Frequency is basically the output of a BoW model. For a specific document, TF determines how important a word is by looking at how frequently it appears in the document. The Inverse document Frequency statistic (IDF), on the other hand,is based on the idea that important words for a specific document (also called signature words) appear frequently inside this document but likely not as often inside other documents. Thus, the frequency of signature words is usually low in different documents,

and this its Inverse Document Frequency must be high. Thus similarity between TF-IDF embeddings is said to be *syntagmatic* (Sahlgren, 2008; Sun et al., 2015), since it concerns words that co-occur within the same text region (e.g., the same sentence, paragraph, or document).

One of the main issues with TF-IDF and BoW is that they usually generate very sparse embedding matrices, depending on the size of the context snippets (Picard, 1999), which are hard to manage effectively for large scale datasets. To this end, techniques such as Latent Semantic Analysis (LSA) (Deerwester, Dumais, Furnas, Landauer and Harshman, 1990) can be used to reduce the sparsity of the embeddings.

In addition, TF-IDF performs poorly on capturing the paradigmatic meaning of words. For word-level embeddings, or *word embeddings*, other (more paradigmatic) techniques are generally used. The term "word embedding" Bengio, Ducharme, Vincent and Jauvin (2003) refers to a type of mapping that provide similar numerical representations to words with similar meaning. For instance, Word2Vec (Mikolov, Sutskever, Chen, Corrado and Dean, 2013), GloVe (Pennington, Socher and Manning, 2014) and fastText (Bojanowski, Grave, Joulin and Mikolov, 2017) are approaches of (unsupervised) learning algorithms for word embeddings, based on Artificial Neural Networks (ANNs), and they consist in an ANN usually trained to optimally predict a word given its context and vice-versa.

An important aspect of the resulting embeddings is that they can be used to find word analogies (e.g., analogies of the form "A is to B what C is to D") by using simple arithmetic. For example, in Word2Vec, we might see that the following word-embeddings equations are valid: "Paris - France + Germany = Berlin", "King - Man + Woman = Queen". This same idea of paradigmatic word embeddings can be extended to sentences or any other bigger-than-word snippet of text.

A simple approach to build document embeddings might be averaging the word embeddings of a document, a technique called Average Word Embedding (AWE). More sophisticated combinations, in addition to simple average, have been proposed in literature to compute weighted averages, some of them involving TF-IDF (Le and Mikolov, 2014; Arroyo-Fernández, Méndez-Cruz, Sierra, Torres-Moreno and Sidorov, 2019).

One of the disadvantages of AWE, just like TF-IDF, is that it is not sensible to words order. More sophisticated approaches for document embeddings where the order becomes meaningful are represented by the encoder-decoder models based on Artificial Neural Networks (ANNs). Some famous examples include the *paragraph vectors* by Le and Mikolov (2014) or the *skip-thought* vectors by Kiros, Zhu, Salakhutdinov, Zemel, Torralba, Urtasun and Fidler (2015). More recently, together with the understanding that deeper ANNs create better embeddings, researchers started to devise more complex and performing document embedding techniques, such as the Transformers (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser and Polosukhin, 2017), based on Deep Neural Networks (DNNs) and specialised in NLP; see also BERT by Devlin, Chang, Lee and Toutanova (2018), Universal Sentence Encoder (USE) by Cer, Yang, Kong, Hua, Limtiaco, John, Constant, Guajardo-Cespedes, Yuan, Tar et al. (2018), and T5 by Raffel, Shazeer, Roberts, Lee, Narang, Matena, Zhou, Li and Liu (2019).

Two variants of the USE have been proposed by Cer et al. (2018): the Transformer-based and the DAN-based approaches. The Transformer-based variant, from the homonym Deep Neural Network(Vaswani et al., 2017) has high accuracy, but quadratic complexity with respect to the input size. The DAN-based variant takes its name from the Deep Averaging Network (Cer et al., 2018), and has linear complexity, but apparently reaches a lower accuracy. Differently from AWE, USE learns to embed the whole sentence directly in an end-to-end manner, providing higher quality results. In both cases the similarity between these embeddings is more *paradigmatic*.

Overall, on the one hand we find shallow syntagmatic techniques for text embedding, such as TF-IDF. Models based on them are very easy to obtain, and in fact a 2015 survey (Beel, Gipp, Langer and Breitinger, 2016) showed that 83% of text-based recommender systems in digital libraries use it. On the other hand we find deeper and more paradigmatic techniques such as BERT or USE, based on deep learning techniques, are very powerful yet much harder to obtain, requiring huge amounts of data, specialised hardware and many hours or days of learning for achieving reasonable performance.

Despite this, recent advancements in deep learning and fine-tuning, together with the advent of new and powerful frameworks (Wolf, Debut, Sanh, Chaumond, Delangue, Moi, Cistac, Rault, Louf, Funtowicz et al., 2019), have made techniques such as BERT or USE very easy to apply, adapt and scale also to simple applications.

However, the application of deep learning in the context of legal information technology is having some difficulties due to the scarce amount of available data. When applying deep learning to legal documents, the main problem is that the legal language (Legalese) differs from the spoken language. Therefore a deep learning algorithm trained on a data-set composed only of spoken language texts often does not work correctly on Legalese. The reason is that legal language is a slightly different language, and training the same deep learning algorithm on a data-set composed only

by Legalese texts often does not give the same results. Examples of legal informatics problems addressable using AI could be: monitor changes to tax laws and regulations, judging small claims, predict court cases, legal question answering, drafts generations, etc..

3. Proposed Solution

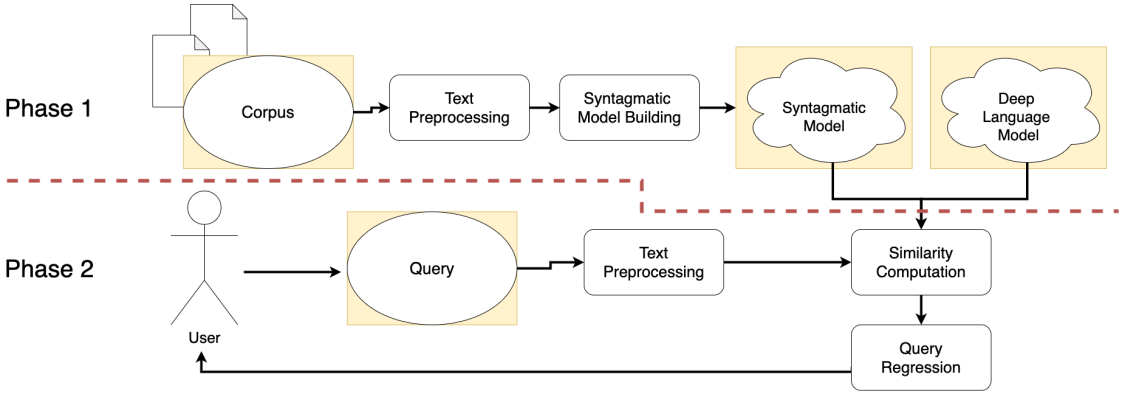


Figure 1: SyntagmTuner: Sketch of the pipeline used in our experiments for both text classification and question answering. Phase 1 is performed only once, when instantiating SyntagmTuner, while phase 2 is performed any time the user queries the system.

In this work we address the problem of data scarcity in Legalese by hypothesising that the way syntagmatic relations encode meaning in Legalese is different from that of its corresponding base language. If this hypothesis is correct, we can improve the performance of general-purpose deep language models on tasks of Legalese processing/understanding affected by data scarcity. We can do it by simply combining them with other ad-hoc models for capturing the syntagmatic relations across texts, without changing the deep language models. In fact, these syntagmatic relations can be partially[2] captured even with little data, e.g. by shallow text embedders such as TF-IDF.

For instance, we could use TF-IDF to model domain-specific information in combination with USE (or any other state-of-the-art deep language model) to model generic information (e.g. semantic relationships among non-domain-specific words). Interestingly, this idea of an ensemble of TF-IDF and ANN-based language models is not completely new. For example, Kowsari, Heidarysafa, Brown, Meimandi and Barnes (2018) and Du, Chen, Peng, Xiang, Tao and Lu (2018) proposed to exploit TF-IDF for improving the training of an ANN. While Zhu, Zhang, Li, He and Zhang (2016), Le and Mikolov (2014) or Arroyo-Fernández et al. (2019) proposed to use TF-IDF for computing a weighted AWEs.

Differently, what we propose is to combine similarities instead of embeddings and to exploit pre-trained paradigmatic language models without re-training or fine-tuning them. To do so we perform case-based reasoning, exploiting the similarities between embeddings for tasks as classification, regression or question answering over texts. For example, for the classification task we define the classes by means of a cluster of representative texts (i.e. a few paragraphs) and then we classify any new fragment of text by looking at how distant (or similar, or distant) are the clusters to it. So, if such distance is above a given threshold then we say that the query belongs to the cluster's class. The same can be seemingly done also for regression or question answering, without needing any threshold. In fact, for question answering we simply look for the closest (or most relevant) answer to a question, without bothering to define any classes or clusters.

More in detail, the distance between snippets of texts (used to understand how close is a text to a question or a class) is computed as the combined similarity of: the embeddings obtained with existing general purpose language models (lacking of domain-specificity), and the embeddings of other more shallow models (i.e. TF-IDF) capable of generating syntagmatic-like similarities. So that the TF-IDF document similarity would be a sort of *topical* similarity extracted by populating the embeddings with information about "which text regions the linguistic items occur in". While the deep

[2]We say "partially" because these techniques do not take into account relative word orderings (that partly constitute syntagmatic relations), but they rather focus on estimating the informative significance of words given by their co-occurrence matrix (Aizawa, 2003).

language model's similarity would be more a sort of *paradigmatic* similarity extracted by populating the embeddings with information about "which other linguistic items the items co-occur with". In other words, the idea behind our proposed ensemble is to combine the unique and different properties of the aforementioned similarities, in order to get a new *paradigmatic* similarity potentially able to express *topical* similarity in a domain on which the pre-trained language models have not been trained on.

To achieve this we designed SyntagmTuner, a pipeline of AI techniques, shown in Figure 1, consisting of two main phases. During the first phase the syntagmatic model (i.e. TF-IDF) is built (after pre-processing, cleaning, the corpus of documents) and combined with a deep language model (i.e. USE). While, during the second phase the queries performed by the user are converted into embeddings by the combined model produced with phase 1. These embeddings are then compared with the embeddings of the corpus. After that, the elements of the corpus with the most similar embeddings to the query's are selected and returned to the user during a sub-step called *Query Regression*.

More in detail, this pipeline can be summarised by the following sub-steps:

Corpus pre-processing : For building the TF-IDF model we need properly formatted data. This sub-step is responsible for normalising the input texts, by manipulating syntagms according to task-specific heuristics (e.g. stop-words filtering) and then performing tokenisation, lemmatization and/or stemming.

TF-IDF model building : This sub-step consists in building a TF-IDF model, learned from the corpus of documents relevant to the task, so to identify the *signature words*[3] contained within it. The TF-IDF model is built only once (unless the corpus changes in time) and before any query is given to the system.

Query pre-processing : This sub-step is identical to the first sub-step (corpus pre-processing), but it is applied to the queries instead of the corpus.

Query similarity computation : This sub-step is responsible for combining the similarities coming from the different models and then used during the next step. The combination is performed linearly, by summing the TF-IDF similarity to the similarity of the deep language model. Both similarities are weighted by task-dependent weights. For example, query q is embedded by TF-IDF into vector V_q and by the deep language model into vector $\bar{V}_q$. On the other hand, the corpus paragraph p is embedded into vectors V_p and $\bar{V}_p$. Let s be a similarity function (i.e. cosine similarity), while w_S is a predefined weight for the syntagmatic model and w_P for the deep language model. The "syntagmatic" similarity $s(V_q, V_p)$ of q to p is S while the similarity $s(\bar{V}_q, \bar{V}_p)$ is P. So, the final combined similarity of p to q is given by the formula $w_S \cdot S + w_P \cdot P$.

Query regression : This sub-step selects the most similar snippets of text of the corpus and it returns them to the user.

For more technical details about the pipeline, please read Sovrano, Palmirani and Vitali (2020a).

4. Experiments and Results

Given all the premises stated in Sections 1 and 2, we designed a few experiments to better understand the role of syntagmatic relations in Legalese, so to figure how to exploit existing state-of-the-art general-purpose natural language models for effectively addressing the data scarcity issue on real-case scenarios. The experiments we performed are two. The first one is about multi-label text classification of UNGA Resolutions, as published in Sovrano et al. (2020a). The second one is an extension of the work presented by Sovrano, Palmirani and Vitali (2020b) and it is about automated question answering on International Private Law (PIL).

During both the experiments we conduct an ablation study to see whether SyntagmTuner is better than using only general-purpose deep language models. In fact, if the hypothesis is true, we expect the performance of any general-purpose deep language model to be considerably exceeded by that of SyntagmTuner on corpora such as the UNGA Resolutions or PIL regulations.

[3] Words that are typical of the topics depicted by the corpus.

4.1. Multi-Label Text Classification of UNGA Resolutions

In the Thirty-Third Session of the High Level Committee on Management (HLCM) held in Budapest, 30-31 March 2017, the United Nations adopted the *Akoma Ntoso XML standard for the United Nations System (AKN4UN)* as well as the *United Nations System Document Ontology (UNDO)* to provide a formal representation of the fundamental entities of UN documents and of their relationships.

HLCM adopted the AKN4UN Guidelines for the markup of UN normative and parliamentary documents, and UNDO as the main reference model for the implementation of UNSIF, the United Nations Semantic Interoperability Framework, in order to identify the structural parts and the semantic aspects of the sentences according to the specific goals of each UN Agency or Department. One common task is to qualify the UN documents following the Sustainable Development Goal (SDG) in order to monitor the progresses at the world level to fight poverty, discrimination, climatic changes. In 2015, 17 SDGs (e.g. no poverty, no hunger, good health and well-being, quality education, etc..) and 169 targets were adopted by the world leaders and in 2016 this list officially came into force. Those goals define the Agenda for Sustainable Development till 2030 and the intention is to universally apply them to eradicate poverty, fight inequalities, tackle climatic change, support inclusion. Progresses are monitored using 232 unique indicators and open data are used for such information.

An interesting and worthy task within this framework is therefore the classification of UN documents, as well as other kinds of documents, according to the above-mentioned SDGs and targets, so as to detect trends and indicators and to produce open data-sets (UN, 2015c) useful for statistics and predictors, and consequently to better inform political strategies of the UN and participating countries to reach such goals. In this scenario, manually annotating a minimal training-set (for a deep learning based classifier) would be the same of annotating the whole corpus (every time the SDGs or their definitions change).

UN Resolutions are the texts of the formal expressions of the opinion or will of United Nations organs. All the UN Resolutions are publicly available on the UN website (UN, 2015a). A UN resolution has a regular structure composed of a preface with the title and the identification information (date and number), a preamble with justificatory and introductory paragraphs, and the full body with the actual norms. The multi-label text classification of a UN Resolution can be performed at different granularity of the document, for example at the document level or approaching each paragraph separately. Sovrano et al. (2020a) decided to work at the paragraph level, because:

i) A paragraph is smaller than the whole document and thus, intuitively, it is easier to classify correctly.

ii) There are 609 UN Resolutions with a grand-total of 26784 paragraphs. Intuitively, it is harder to reliably associate a multi-label classification to 609 texts, while it becomes more appropriate with 26784 texts.

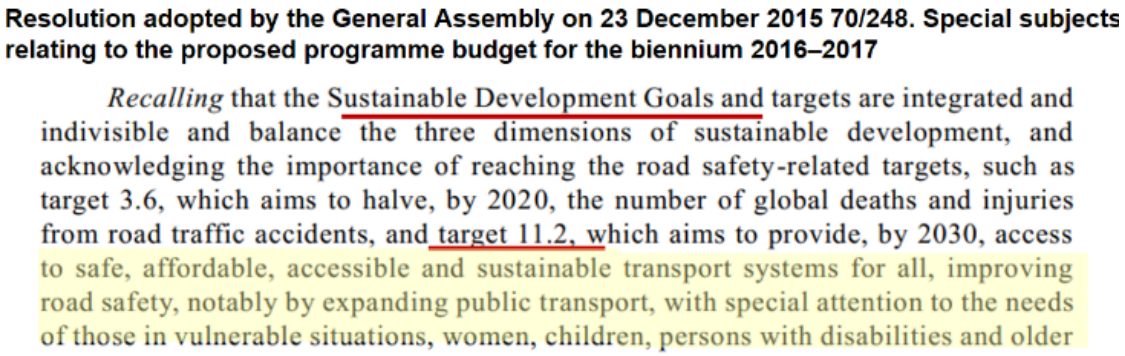
In this particular setting, every SDG states a well-defined and different concept. Therefore classifying a resolution paragraph according to its most related SDGs is equivalent to understand whether the concepts expressed in the paragraph are similar enough to one or more SDGs. The task requires to identify whether a paragraph of an official English resolution of the UN is related (if any) to one or more SDGs. Every SDG may have different targets that may change in the near future. In fact, some of the targets have a short- or mid-term deadline such as 2020, 2030.

Mentions to a target or a goal, within a paragraph, can be both explicit or implicit, as shown in Figure 2. Intuitively, identifying the implicit references is harder. As consequence, a few reasonable requirements for a multi-label text classifier in this context would be:

1. The algorithm should be able to decide whether a given paragraph is related or not to a SDG.

2. The (learning) algorithm should require almost no annotated training set for properly working, and should allow us to easily change the SDG definitions without incurring significantly slower or more error-prone pre-processing (eg. a slow model-training phase).

Every SDG has an official English description publicly available at UN (2015b). But these descriptions alone seem to be not enough for properly training an ANN-based model from scratch, nor for traditional transfer learning or fine-tuning. This why in order to tackle this problem we need an approach like the one presented in Section 3. More in detail, let A (the *query*; a paragraph) and B (a *corpus* document describing a SDG) be two distinct documents, we want to compute the similarity between A and B. In order to do that, we combine the cosine similarity of the TF-IDF embeddings of A and B with the cosine similarity of the USE embeddings weighted by the cosine similarity of some Average GloVe embeddings. The pre-trained models we are going to use are:


Resolution adopted by the General Assembly on 23 December 2015 70/248. Special subjects relating to the proposed programme budget for the biennium 2016–2017

Recalling that the Sustainable Development Goals and targets are integrated and indivisible and balance the three dimensions of sustainable development, and acknowledging the importance of reaching the road safety-related targets, such as target 3.6, which aims to halve, by 2020, the number of global deaths and injuries from road traffic accidents, and target 11.2, which aims to provide, by 2030, access to safe, affordable, accessible and sustainable transport systems for all, improving road safety, notably by expanding public transport, with special attention to the needs of those in vulnerable situations, women, children, persons with disabilities and older

Figure 2: An UNGA paragraph containing some examples of both explicit and implicit goals/targets. The explicit goals/-targets are underlined in red, while the implicit ones are highlighted in yellow.

- The GloVe model coming from Spacy (Honnibal, 2016) and pre-trained on data from Common Crawl (Crawl, 2011).
- The USE model for document embedding coming from TensorFlow Hub (Cer et al., 2018).

Therefore, summarising Sovrano et al. (2020a), the pipeline presented in Section 3 has been adapted as follows. We have that the corpus C is made of 34 different documents, two for every SDG. These documents have been extracted from the official English descriptions of the SDGs, publicly available at UN (2015b). Both the "Corpus and Query Pre-Processing" steps have been modified so to remove stop-words and punctuation, and to better identify relevant uncommon syntagms (such as the ids of the SDGs). The *Query Similarity Computation* step has been slightly changed in order to combine more than two language models (GloVe and USE), while both the syntagmatic w_S and the paradigmatic w_P weights were set to 0.5. The *Query Regression* step produces as output a regression represented by a vector of real numbers, one for each different class (the SDGs). After that a threshold T is used to convert the regression into a classification.

Ablation Study

Considering the immediate compatibility of the work presented by Sovrano et al. (2020a) with our experimental setting, we picked the same evaluations done in Sovrano et al. (2020a). In fact, those evaluations are conducted so to have the ablation study we need in order to test our hypothesis. The pipeline presented in Section 3 is tested on 3 different data-sets, built in different ways by different annotators with different expertise:

- **Dev-Set**: used to tune the algorithm during development. This data-set has 36 annotated elements by A. These elements do not appear in the other data-sets.
- **Test-Set A**: it contains 121 paragraphs manually annotated by A.
- **Test-Set B**: it contains 105 paragraphs manually annotated by B. This set shares 50 paragraphs with set A with possibly different annotations.

We decided to skip both the test-sets CB and CL presented in Sovrano et al. (2020a) because they contain a lot of false negatives, as pointed out also in the paper. As shown in Table 1, all the aforementioned data-sets are imbalanced, in

	No SDG	SDG 16	SDG 17	Remaining SDGs
Dev	28,8%	11,1%	13,3%	46,8%
A	28,5%	25,7%	8,5%	37,3%
B	42,8%	19,6%	1,7%	35,9%

Table 1: SDGs distribution across datasets Dev, A and B.

fact most of the annotated labels are of type 0 (no SDG), or 16 ("Promote just, peaceful and inclusive societies") or 17

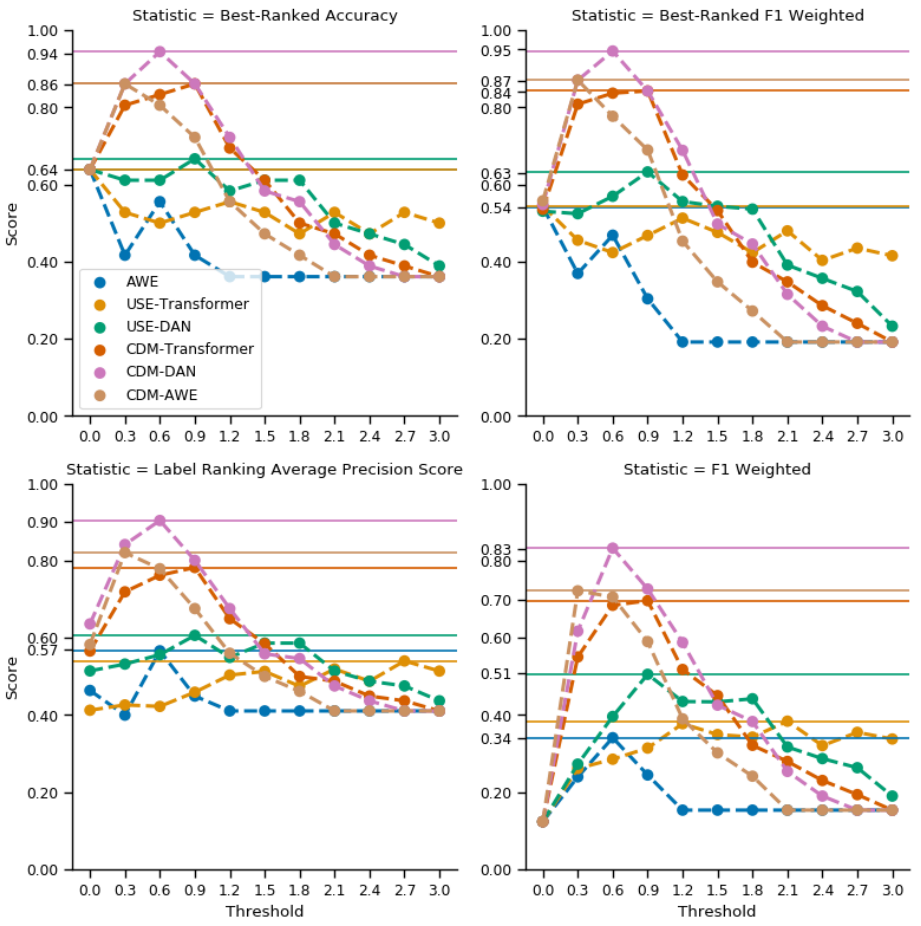


Figure 3: Dev-Set - Score comparison of the statistics of different algorithms when changing the threshold, on the Dev-Set. The baselines are labelled as AWE or USE, while SyntagmTuner is labelled as CDM.

("Revitalize the global partnership for sustainable development"). This imbalance would be normally an issue for any deep learning classifier trained from scratch or fine-tuned, but thanks to the approach we are going to defend imbalance is no more a significant issue.

Considering that all those data-sets were unbalanced and that the task involves multi-label text classification, the adopted evaluation metrics were:

- Label Ranking Average Precision,
- Weighted F1,
- Best-Ranked (BR) Accuracy,
- BR Weighted F1.

Where the *Weighted F1* is a variant of the Macro F1, weighted by support (the number of true instances for each label) in order to address the data-set imbalance. The Best-Ranked statistics are the statistics of the best ranked label in the intersection of *true labels* with the *predicted labels*. If the aforementioned intersection is empty, then a random *true label* is taken. BR statistics seem reasonable due to the fact that the average labels per point (paragraph) is very low: between 1 and 1.5 depending on the test-set. Here, the BR Accuracy is equivalent to the BR Micro F1.

The ablation study is performed by comparing the proposed algorithm (using a linear combination of syntagmatic and paradigmatic similarities) with the baselines (using only state-of-the-art paradigmatic similarities), while changing

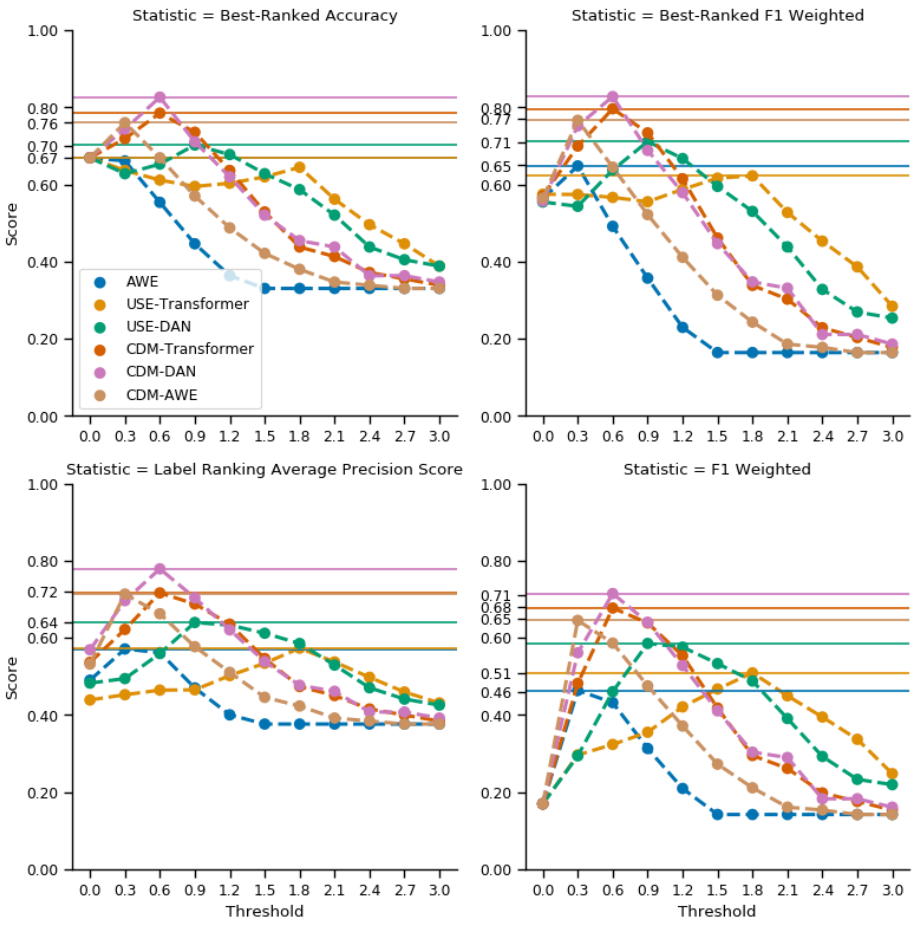


Figure 4: Test-Set A - Score comparison of the statistics of different algorithms when changing the threshold, on the Test-Set A. The baselines are labelled as AWE or USE, while SyntagmTuner is labelled as CDM.

the value of the classification threshold T. In Figures 3, 4 are shown the statistics obtained respectively on the dev-set and on test-set A

In this experiment, three different baselines (using only paradigmatic similarities) are compared against two other contenders (exploiting both syntagmatic and paradigmatic similarities). The baselines are:

- AWE: based solely on GloVe AWE similarity.
- USE-DAN: relying solely on the DAN-based USE similarity.
- USE-Transformer: relying solely on the Transformer-based USE similarity.

The contenders are:

- CDM-DAN: our SyntagmTuner, relying on both the TF-IDF similarity the DAN-based USE similarity.
- CDM-Transformer: our SyntagmTuner, relying on both the TF-IDF similarity the Transformer-based USE similarity.

As we can see also in Table 2, the results of the experiments are against all the baselines, suggesting that a naive linear combination of syntagmatic and paradigmatic similarities can boost a performance improvement that is on the order of +30% on the Dev-Set, of +12% on Test-Set A and of +8% on Test-Set B. This gives us strong evidence in favour of our main hypothesis.

	Dev	A	B
BR Accuracy	+27%	+12%	+8%
BR F1 Weight.	+32%	+12%	+8%
LRAP	+30%	+15%	+10%
F1 Weight.	+32%	+12%	+8%

Table 2: SyntagmTuner improvements over the best baseline, on the UNGA Resolutions classification task, for each metric, across the 3 data-sets: Dev, A and B. LRAP stands for Label Ranking Average Precision.

4.2. Answering Questions on International Private Law

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., criminal law vs. civil law). Scientific research on PIL reveals the need to create a bridge between European and national laws on this domain, accessing heterogeneous legal sources. In this context, legal experts have to access heterogeneous sources, being able to recall all the norms and to combine them using case-laws[4] 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. In fact, searching within thousands and thousands of pages of legal documents with different sources and legislations is surely a task requiring days and days of human effort by a workforce with highly specialised expertise. This is probably one of the reasons why researchers, governments and industry have long looked for a way to build "intelligent" machines capable of helping humans in detecting the relevant legal provisions over such complex corpora (Kratochwil, 1991).

In 2015 Kim et al. presented one of the very first algorithms based on DNNs for Legal Question Answering (reasoning) applied to a data-set of Boolean questions from Japanese legal bar exams, then followed up by Do, Nguyen, Tran, Nguyen and Nguyen (2017) and others (Ravichander, Black, Wilson, Norton and Sadeh, 2019; Holzenberger, Blair-Stanek and Van Durme, 2020).

At about the same time, Sovrano et al. (2020b) 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 improve the quality of the relationships inside the provisions and between them. More generally, automated reasoning over legal texts (not just the PIL's ones) is not a trivial task, due to the fact that Legalese is rarer and in many ways different from a commonly used natural language. In this context it is not uncommon to work with small data-sets that make very hard to effectively build a deep language model, especially when corpora are not consistent across different laws and Regulations Chalkidis and Kampas (2019). 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).

The baseline tool published by Sovrano et al. (2020b) is relying on the same technique presented in Section 3, for properly retrieving a set of answers from a Knowledge Graph (KG), given an input question. Sovrano et al. (2020b) extract KGs from legal documents by exploiting the grammatical dependencies of their content, through an automated dependency parser. The legal documents considered for the show-case are three European regulations: Rome I Regulation EC 593/2008, Rome II Regulation EC 864/2007, and Brussels I bis Regulation EU 1215/2012. The proposed pipeline for performing question answer retrieval can be summarised by the following phases: extracting, matching, selecting, answering. The question answering retrieval algorithm is as follows:

1. **Extracting**: extract the set of concepts C from a question Q.
2. **Matching**: find the most syntactically similar KG's concepts to C, and retrieve all their related template-triplets including those of the sub-classes of C.
3. **Selecting**: among the natural language representations of both the retrieved triplets and their respective subject-s/objects[5], select those snippets of text that are sufficiently likely to be an answer to Q.
4. **Answering**: return as set of answers the contexts (the source paragraphs) of the selected snippets of text (triples or simple concepts).

[4] http://www.interlexproject.eu/del/Deliverable2dot3.pdf

[5] Some questions can be succinctly answered through a single concept, while others require a more elaborated sentence (therefore a template-triplet).

More in detail, the *selection* phase is performed by means of a variation of Sovrano et al. (2020a), that combines TF-IDF with a version of the Universal Sentence Encoder (USE) for Question Answering (QA) (Yang, Cer, Ahmad, Guo, Law, Constant, Abrego, Yuan, Tar, Sung et al., 2019). Therefore, similarly to the work presented in Section 4.1, during the *selection* phase the pipeline presented in Section 3 is used to associate a given question to a set of answers.

The goal in this specific domain is to extract knowledge according to specific situations and to detect the useful legal sources capable to help the expert to interpret them and to find a solution. A user can interact with the KG through the Question Answering (QA) tool, posing natural language questions and expecting useful answers from the system.

Some frequently asked questions, in these cases, might be related to where a legal trial is celebrated (e.g., the pertinent jurisdiction and court), because there are many nuances and conflicting rules depending to the typology of actors, the country of residence (e.g., habitual residence), the country where the activity is performed (e.g.,country where the employee habitually carries out his work). To this end, Sovrano et al. (2020b) has published a set of questions on those 3 regulations, where each question is correlated with a set of answers that a domain expert would expect.

Ablation Study

Considering the immediate compatibility of the work presented by Sovrano et al. (2020b) with our experimental setting, we had to change very little of that in order to make it fit our needs. We took the baseline tool (SyntagmTuner) given by Sovrano et al. (2020b), without any modification. We found that SyntagmTuner is using $w_S = 0.5$ and $w_P = 0.5$. As contender (NoTune) we simply used SyntagmTuner with $w_S = 0$ and $w_P = 1$, hence with no syntagmatic similarities. With the given KG extractor we were able to extract a knowledge graph of roughly 9175 different grammatical clausal triplets. In order to test our main hypothesis, we compared the answers given by the two tools with the set of expected answers identified for each question in the data-set. Considering that we are not interested in the order answers are ranked, as metric for estimating the performance of the algorithms we chose the 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 top5-precision is a measure of how many selected answers in the top five are relevant. Knowing the top5-recall and the top5-precision we can also compute the top5-F1 score. The results, shown in Tables 3, 4 and 5, indicate that, on

	Top5-Recall	Top5-Precision	Top5-F1
NoTune	28.17%	24.70%	25.27%
SyntagmTuner	**37.58%**	**45.17%**	**38.05%**

Table 3: Average top5 recall, precision and F1 for both NoTune and SyntagmTuner, on the Question Answering task. Maximal scores are shown in bold. SyntagmTuner is achieving the best results on all the metrics.

the whole data-set, the legal tool using the syntagmatic relations modelled by TF-IDF is the best in all the identified metrics, therefore giving further support to our main hypothesis. In fact SyntagmTuner has a top5-recall of 37.58%, a top5-precision of 45.17% and a top5-F1 of 38.05%, beating NoTune by roughly +12.78% on the top5-F1 and by +20.47% on the top5-precision, this without any complicated training procedure. What is interesting to notice is that NoTune and SyntagmTuner have a similar recall, suggesting that syntagmatic relations are of most importance for reasoning with higher precision.

5. Conclusions

Modern state-of-the-art AI for language processing and understanding (i.e. deep learning) usually have performances that scale proportionally to the amount of available data. These AI are normally trained directly on large collections of texts, building a (deep) language model that can be used to solve downstream tasks as classification, regression, reasoning, etc.

In this paper we presented and discussed a new technique called SyntagmTuner, to tackle many problems of language processing that are typical of e-governance and in which data scarcity is unavoidable. SyntagmTuner is based on a specific hypothesis according to which Legalese (i.e. legal English) differs from its corresponding base language (i.e. common English) in the way it encodes meaning through syntagmatic relations. Starting from this hypothesis, we

Question	Expected Answers	SyntagmTuner's Top5	SyntagmTuner's Scores	NoTune's Top5	NoTune's Scores
Who determines disputes under a contract?	B Art. 7.1 B Art. 8.3 B Art. 8.4 B Art. 17	RI Rec.12 B Art.17.2 RI Rec.24	**R: 25%** **P: 33%** **F1: 28.44%**	RI Rec. 12 RI Art. 11.5 RI Art. 4.1 B Art.25.1 RII Art.11.1	R: 0% P: 0% F1: 0%
What factors should be taken into account for conferring the jurisdiction to determine disputes under a contract?	B Art. 7.1 B Art. 17 B Art. 20 B Art. 25	RI Rec. 12 B Art.25 B Art.25.5 B Rec.15 RI Rec. 21	R: 25% P: 40% F1: 30.76%	RI Rec. 12 B Art.25.4 B Rec. 12 B Art.25.1 B Art.25.5	R: 25% **P: 60%** **F1: 35.29%**
Which parties of a contract should be protected by conflict-of-law rules?	RI Rec. 23 RI Art. 6 RI Art. 8 RI Art. 13	RI Rec.23 B Rec.18 RI Rec.24 RI Art.25.1 RI Rec.27	**R: 25%** **P: 20%** **F1: 22.22%**	RII Art.4.3 RII Art.5.2 RII Art.11.1 RII Art.10.1 RI Rec.12	R: 0% P: 0% F1: 0%
In which case claims are so closely connected that it would be better to treat them together in order to avoid irreconcilable judgments?	B Art. 8 B Art. 30 B Art. 34	B Art. 8.1	R: 33% **P: 100%** **F1: 49.62%**	B Art.8.1 B Art.71.2 B Rec.22 B Rec.21 B Rec.28	R: 33% P: 20% F1: 24.90%
What kind of agreement between parties are regulated by these Regulations?	B Rec. 6 B Rec. 10 B Rec. 12 B Art. 1 RI Rec. 7 RI Art. 1	B Art.73.3 B Rec. 12 B Rec. 36 B Art.71.2 B Art. 71.1	**R: 20%** **P: 20%** **F1: 20%**	RI Rec.12 B Art.60 B Art.24 B Art.71.2 B Art.71.1	R: 0% P: 0% F1: 0%
In which court is celebrated the trial in case the employer is domiciled in a Member State?	B Art. 21 B Art. 22 B Art. 23	B Art. 21.1 B Art.22.1 B Art. 21.2 B Art. 20.1 B Art. 20.2	R: 66% **P: 60%** **F1: 62.85%**	B Art.22.1 B Art.18.1 B Art.15 B Art.15 B Art.21.1	R: 66% P: 40% F1: 49.81%
How should a contract be interpreted according to this regulation?	RI Rec. 22 RI Rec. 12 RI Rec. 26 RI Rec. 29 RI Art. 12	RI Art. 10.1 RI Rec.17	R: 0% P: 0% F1: 0%	RI Art.3.2 RI Art.9.1 RI Art.14.1 RI Art.10.1 RI Rec.28	R: 0% P: 0% F1: 0%
Which law is applicable to a non-contractual obligation?	RII Rec. 17 RII Rec. 18 RII Rec. 26 RII Rec. 27 RII Rec. 31 RII Art. 4-20	RI Art. 8.1 RII Art.15 RII Art.16 RII Art.8.1 RII Rec. 22	R: 60% P: 60% F1: 60%	RII Art.18 RII Art.6.2 RII Art.15 RII Art.12.1 RII Art.6.1	**R: 100%** **P: 100%** **F1: 100%**

Table 4: First block of questions, expected answers and given answers (ordered by the pertinence to the question estimated by the tool) of NoTune and SyntagmTuner. "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. The best top5 scores are in bold, unless they are the same across all the algorithms. In the "scores" columns: "P" stands for Precision and "R" stands for Recall.

Question	Expected Answers	SyntagmTuner's Top5	SyntagmTuner's Scores	NoTune's Top5	NoTune's Scores
Can the parties choose the applicable law in consumer contracts?	RI Rec. 11 RI Rec. 25 RI Rec. 27 RI Art. 6	B Art. 18.2 B Art. 18.1 RI Rec. 28 RI Art. 5.2 RI Art. 6.2	R: 25% P: 20% F1: 22.22%	RI Art. 5.2 RI Art. 3.1 RI Art. 6.2 B Art. 18.2 RI Art. 8.1	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?	B Rec. 18 B Art. 17 B Art. 18 B Art. 19 B Art. 26	RI Rec. 12 RI Rec. 24 B Art. 19 B Art. 17.1 B Art. 25.5	**R: 40%** **P: 40%** **F1: 40%**	RI Rec. 27 RI Rec. 24 RI Art. 6.3 RI Rec. 28 RI Art. 6.3 B Art. 17.1	R: 20% P: 20% F1: 20%
Can the parties choose a different applicable law for different parts of the contract?	RI Rec. 11 RI Art. 3.1	RI Art. 3.1 RI Art. 5.2 RI Art. 7.3 RII Art. 25.2 RI Art. 22.2	R: 50% P: 20% F1: 28.57%	RI Art.5.2 RI Art.3.1 RI Art.8.1 RI Art.7.2 RI Art.8.2	R: 50% P: 20% F1: 28.57%
What non-contractual obligations fall into the scope of Regulation Rome II?	RII Rec. 10 RII Rec. 11 RII Art. 1 RII Art. 2	RII Stat. 1 RI Rec. 7	R: 0% P: 0% F1: 0%	RII Art.15 RII Art.22.1 RII Art.1.2 RII Art.6.3 RII Art.12.1	**R: 25%** **P: 20%** **F1: 22.22%**
What is the applicable rule to protect the weaker party of a contract?	RI Rec. 23 B Rec. 18	RI Rec. 23 B Rec. 18	**R: 100%** **P: 100%** **F1: 100%**	RI Rec. 28 RII Art. 5.2 RII Art. 4.3 B Art. 19 RI Art. 13	R: 0% P: 0% F1: 0%
What is the applicable law to determine the validity of consent?	RI Art. 3.5 RI Art. 10 RI Art. 11 RI Art. 13	RI Art. 3.5 RI Art. 10.2 RI Art. 10.1 B Rec. 20	**R: 50%** **P: 75%** **F1: 60%**	RI Art. 7.3 B Rec. 20 B Art. 65.2 RI Art. 6.2 B Rec. 10	R: 0% P: 0% F1: 0%
When are two actions to be considered related according to the Regulation Brussels I Bis?	B Rec. 21 B Art. 30.3		R: 0% P: 0% F1: 0%	B Rec. 2 B Rec. 22 B Rec. 12 B Art. 30.2 B Art. 8.4	R: 0% P: 0% F1: 0%
What court has jurisdiction in case of a counter-claim?	B Art. 8.3 B Art. 14.2 B Art. 18.3 B Art. 22.2	B Art. 18.3 B Art. 14.2 B Art. 22.2 B Art. 8 B Art. 24	**R: 100%** **P: 80%** **F1: 88.88%**	B Rec. 11 B Art. 2 B Art. 18.3 B Art. 14.2 B Art. 22.2	R: 75% P: 60% F1: 66.6%
Where can an employee sue their employer?	B Rec. 14 B Rec. 18 B Art. 21.1 B Art. 22.1 B Art. 23	B Art. 21.1	R: 20% **P: 100%** F1: 33.33%	B Art. 21.1 B Art. 20.1 B Art. 22.1 B Art. 23.2 RII Rec. 27	**R: 60%** P: 60% **F1: 60%**

Table 5: Second block of questions, expected answers and given answers of SyntagmTuner and NoTune. See the caption of Table 4 for more details about how to read this table.

designed a pipeline of AI techniques specialised at capturing the syntagmatic relations of Legalese text, integrating it in general-purpose deep language models (pre-)trained on the base language.

The point is that the syntagmatic relations across texts can be partially captured even with little data (e.g. by shallow text embedders such as TF-IDF or Latent Semantic Analysis), avoiding us the need for large quantities of data. We verified this on two different tasks. The first task is a multi-label text classification of UNGA Resolutions, while the second one is automatic question answering on International Private Law.

The results we obtained evidently support the initial hypothesis. As shown in table 2 and in table 3, in both the considered tasks of classification and question-answering, as expected, SyntagmTuner significantly outperformed the general-purpose language models, by a large margin (i.e. +20%). Our expectations were based on the hypothesis that the meaning encoded in legalese by syntagmatic relations is not captured by general-purpose language models, and the results clearly confirmed it.

A consequence of these results is that, if the initial hypothesis is true, it is possible to specialise general-purpose AI on many e-governance tasks without the need for large amount of data, hence consistently reducing the costs of development and maintenance. In fact, building a large training-set is usually an expensive procedure requiring the effort of several experts over several days (if not months).

The theoretical implications of the initial hypothesis are manifold. First of all it might foster future research on deep learning technologies specifically tailored to handle Legalese texts, i.e. by creating specialised neural networks that can better capture the meaning encoded by syntagmatic relations. Furthermore, it could also give some insights about how to produce better formal representations of texts (i.e. for symbolic reasoning), so that the meaning of the Legalese text is better preserved in its formal representation. Finally, another implication is that it could help designing new mechanisms for fine-tuning pre-trained deep language models, as we did with SyntagmTuner.

Nonetheless, the technology behind our current implementation of SyntagmTuner is far from being perfect. It is important to remember that TF-IDF can model syntagmatic relations only partially, by not taking into account relative word orderings. Therefore further research could focus on how to better capture syntagmatic relations with SyntagmTuner, perhaps without sacrificing all the good qualities of TF-IDF. Another direction of research could also be to analyse the consequences that a combination of shallow and deep learning approaches can have on the overall explainability. Indeed, shallow techniques are generally known to be much more interpretable, whereas the deeper is the approach to learning, the less is explainability.

References

Aizawa, A., 2003. An information-theoretic perspective of tf–idf measures. Information Processing & Management 39, 45–65.

Arroyo-Fernández, I., Méndez-Cruz, C.F., Sierra, G., Torres-Moreno, J.M., Sidorov, G., 2019. Unsupervised sentence representations as word information series: Revisiting tf–idf. Computer Speech & Language 56, 107–129.

Basu, S., 2004. E-government and developing countries: an overview. International Review of Law, Computers & Technology 18, 109–132.

Beel, J., Gipp, B., Langer, S., Breitinger, C., 2016. Paper recommender systems: a literature survey. International Journal on Digital Libraries 17, 305–338.

Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C., 2003. A neural probabilistic language model. Journal of machine learning research 3, 1137–1155.

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T., 2017. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics 5, 135–146.

Bommarito II, M.J., Katz, D.M., Detterman, E.M., 2018. Lexnlp: Natural language processing and information extraction for legal and regulatory texts. arXiv preprint arXiv:1806.03688 .

Cer, D., Yang, Y., Kong, S.y., Hua, N., Limtiaco, N., John, R.S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., et al., 2018. Universal sentence encoder. arXiv preprint arXiv:1803.11175 .

Chalkidis, I., Kampas, D., 2019. Deep learning in law: early adaptation and legal word embeddings trained on large corpora. Artificial Intelligence and Law 27, 171–198.

Condevaux, C., Harispe, S., Mussard, S., Zambrano, G., 2019. Weakly supervised one-shot classification using recurrent neural networks with attention: Application to claim acceptance detection., in: JURIX, pp. 23–32.

Crawl, C., 2011. Common crawl. http://commoncrawl.org. Online; accessed 22-May-2019.

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, 391–407.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 .

Do, P.K., Nguyen, H.T., Tran, C.X., Nguyen, M.T., Nguyen, M.L., 2017. Legal question answering using ranking svm and deep convolutional neural network. arXiv preprint arXiv:1703.05320 .

Du, J., Chen, Q., Peng, Y., Xiang, Y., Tao, C., Lu, Z., 2018. Ml-net: multi-label classification of biomedical texts with deep neural networks. arXiv preprint arXiv:1811.05475 .

Dumais, S.T., 2004. Latent semantic analysis. Annual review of information science and technology 38, 188–230.

EU-Commission, 2018. Digital transition action plan. https://ec.europa.eu/futurium/en/system/files/ged/digital_transition_action_plan_for_dgum_300818_final.pdf. Online; accessed 09-Mar-2021.

Farkas, J., Schou, J., 2019. Post-truth, fake news and democracy: Mapping the politics of falsehood. Routledge.

Firth, J.R., 1957. A synopsis of linguistic theory, 1930-1955. Studies in linguistic analysis .

Harris, Z.S., 1954. Distributional structure. Word 10, 146–162.

Hinds, J., Williams, E.J., Joinson, A.N., 2020. "it wouldn't happen to me": Privacy concerns and perspectives following the cambridge analytica scandal. International Journal of Human-Computer Studies 143, 102498.

Holzenberger, N., Blair-Stanek, A., Van Durme, B., 2020. A dataset for statutory reasoning in tax law entailment and question answering. arXiv preprint arXiv:2005.05257 .

Honnibal, M., 2016. Spacy (version 1.3. 0). Explosion AI, Berlin, Germany, Tech. Rep .

ICLR, 2019. Blackstone. https://research.iclr.co.uk. Online; accessed 09-Mar-2021.

Jaeger, P.T., Bertot, J.C., 2010. Transparency and technological change: Ensuring equal and sustained public access to government information. Government Information Quarterly 27, 371–376.

Jaeger, P.T., Thompson, K.M., 2003. E-government around the world: Lessons, challenges, and future directions. Government information quarterly 20, 389–394.

Jones, K.S., 1972. A statistical interpretation of term specificity and its application in retrieval. Journal of documentation .

Kim, M.Y., Xu, Y., Goebel, R., 2015. A convolutional neural network in legal question answering, in: JURISIN Workshop.

Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R.S., Torralba, A., Urtasun, R., Fidler, S., 2015. Skip-thought vectors. arXiv preprint arXiv:1506.06726 .

Kowsari, K., Heidarysafa, M., Brown, D.E., Meimandi, K.J., Barnes, L.E., 2018. Rmdl: Random multimodel deep learning for classification, in: Proceedings of the 2nd International Conference on Information System and Data Mining, ACM. pp. 19–28.

Kratochwil, F.V., 1991. Rules, norms, and decisions: on the conditions of practical and legal reasoning in international relations and domestic affairs. 2, Cambridge University Press.

Le, Q., Mikolov, T., 2014. Distributed representations of sentences and documents, in: International conference on machine learning, PMLR. pp. 1188–1196.

McDonald, S., Ramscar, M., 2001. Testing the distributioanl hypothesis: The influence of context on judgements of semantic similarity, in: Proceedings of the Annual Meeting of the Cognitive Science Society.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013. Distributed representations of words and phrases and their compositionality, in: Advances in neural information processing systems, pp. 3111–3119.

Pennington, J., Socher, R., Manning, C., 2014. Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532–1543.

Picard, J., 1999. Finding content-bearing terms using term similarities, in: Ninth Conference of the European Chapter of the Association for Computational Linguistics, pp. 241–244.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J., 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683 .

Rajaraman, A., Ullman, J.D., 2011. Mining of massive datasets. Cambridge University Press.

Ravichander, A., Black, A.W., Wilson, S., Norton, T., Sadeh, N., 2019. Question answering for privacy policies: Combining computational and legal perspectives. arXiv preprint arXiv:1911.00841 .

Robaldo, L., Villata, S., Wyner, A., Grabmair, M., 2019. Introduction for artificial intelligence and law: special issue "natural language processing for legal texts".

Sahlgren, M., 2008. The distributional hypothesis. Italian Journal of Disability Studies 20, 33–53.

Shao, Y., Mao, J., Liu, Y., Ma, W., Satoh, K., Zhang, M., Ma, S., 2020. Bert-pli: Modeling paragraph-level interactions for legal case retrieval, in: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, pp. 3501–3507.

Sovrano, F., Palmirani, M., Vitali, F., 2020a. Deep learning based multi-label text classification of unga resolutions, in: Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance, pp. 686–695.

Sovrano, F., Palmirani, M., Vitali, F., 2020b. Legal knowledge extraction for knowledge graph based question-answering, in: Legal Knowledge and Information Systems: JURIX 2020. The Thirty-third Annual Conference, IOS Press. pp. 143–153.

Sun, F., Guo, J., Lan, Y., Xu, J., Cheng, X., 2015. Learning word representations by jointly modeling syntagmatic and paradigmatic relations, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 136–145.

UN, 2015a. General assembly resolutions. https://www.un.org/en/sections/documents/general-assembly-resolutions/. Online; accessed 22-May-2019.

UN, 2015b. Sustainable development goals. https://www.un.org/sustainabledevelopment/sustainable-development-goals/. Online; accessed 22-May-2019.

UN, 2015c. Sustainable development goals catalog. http://www.sdg.org/. Online; accessed 22-May-2019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need, in: Advances in neural information processing systems, pp. 5998–6008.

Vink, M., Netten, N., Bargh, M.S., van den Braak, S., Choenni, S., 2020. Mapping crime descriptions to law articles using deep learning, in: Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance, pp. 33–43.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al., 2019. Transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771 .

Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Abrego, G.H., Yuan, S., Tar, C., Sung, Y.H., et al., 2019. Multilingual universal sentence encoder for semantic retrieval. arXiv preprint arXiv:1907.04307 .

Zhu, W., Zhang, W., Li, G.Z., He, C., Zhang, L., 2016. A study of damp-heat syndrome classification using word2vec and tf-idf, in: 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE. pp. 1415–1420.