Legal data extraction and possible applications

IOP Conf. Series: Materials Science and Engineering **618** (2019) 012037 doi:10.1088/1757-899X/618/1/012037

Legal data extraction and possible applications

K Stoykov[1] and S Chelebieva

[1]Technical university of Sofia, 8 "Kl. Ohridski" Street, Sofia, Bulgaria

[1]k.stoykov@hotmail.com

[2]stefka.chelebieva1202@gmail.com

Abstract. The functioning of the modern society provokes creation of significant number of documents, most of them in text form. Many of them are legally relevant and complex in nature, which makes their understanding difficult. They are written in nature language for people, which is hard for computer processing and hence hard for further analysis. Documents represent an abundance of data in unstructured form. They can be handled using NLP methodology. Different models, trained for specific use cases are used, depending on a type of legal document. Automation of number of processes can save time and speed up work on document processing. This, of course, does not rule out the need for highly qualified lawyers. The goal is to ease their work in terms of processing large amount of documents. Increased productivity should be beneficial for both natural and legal persons when working with textual and legal issues.

1. Introduction

The presence of many data leads to a major delay in processing by humans. Automated processing is many times faster, but the data must be extremely well-structured, that is, computer easily understands SQL, but a much more difficult the context of a text [1]. The innovations in hardware in recent years has led to a boom in machine learning algorithms, including an understanding of natural language. These approaches allow a much faster analysis of texts, including legal ones. Legal documents are considered to be difficult to understand and apply [2]. However, different approaches are used to solve legal problems in specific areas [3]. The goal of the article is to show the connection between legal processes and how their analysis in large amounts can be useful and saves time [3]. The processing of natural language is directly related to the preliminary preparation of models in the given language [4], which is not very well covered in Bulgarian [5]. In order to achieve the goal of the article, different types of documents will be considered, what algorithms and approaches can be used to retrieve data. Also, tools that will be involved in processing [5, 6, 7] will also be considered.

2. Legal context

The law is divided into different branches - the common and the classic ones are: civil law, criminal law and administrative law, which in turn are divided into branches. According to the public relations governed by public law, the bond, the copyright, the hereditary, the family are branches of civil law, [8] according to the structural division of the positive law can be distinguished - material and procedural, private and public, domestic and international law. Each of them has its peculiarities and specifics, but there is a common, the normative basis of which they are subordinate, the documents and in general all the texts of a legal nature are in natural language. Large amounts of data are generated from and in the

form of various documents - contracts, powers of attorney, declarations, orders, applications, requests, decisions, letters, notices, receipts, notifications, etc. Though each one of them to have an abstract structure, they are treated as unstructured information from the point of view of processing them as computer data. Different documents, depending on their type, have elements that have a specific location. The personal data of the parties to the contracts are always mentioned at the beginning of the documents, not as an annex or between the paragraphs itself, this is an example of an abstract structure.

The availability of a large set of data which must be processed is a good sign to involve AI methods for task solving and process acceleration. Below, the article lists various tools that can be used to solve different tasks, such as the use of NLP to understand text. The introduction of such a methodology is not only because it is modern or because of some major change over the past one or two years, but rather because the legal sphere is ready to accept software solutions that use technology to understand natural language and machine learning [3]. These adaptation processes are not the same throughout the world, but have begun and cover more and more countries.

3. Applications

The article will focus on examining three main areas where natural language understanding can be applied [3]. The first is dealing with contracts. They can be a variety of types - contract for purchase, rent, order, donation, replacement, etc. NLP methods could retrieve contract information and verify that it meets the requirements [9]. In some cases, this can helpful to finalize the contract more quickly and avoid mistakes. The other strand that will be considered is the legal study - analyzing law cases (judgments, definitions and orders), thus predictive systems can be trained. The last direction that will be addressed in the article is intelligent interfaces. We can imagine them as web-based platforms that are the kind of question-response. They can be used for both lawyer and lawyer training.

3.1. Contracts

The term "contract" is used with different meanings, it is legally embedded in the Law on Obligations and Contracts (the Law on Obligations and Contracts), in foreign literature it is considered as a legal relationship and as a document. [10]. The contract is a phenomenon common to the different branches of law - real estate, family (marriage), labor (individual and collective labor contracts). Administrative law is familiar with the administrative contract [11]. One of the sources of international law is the international treaty, which can be bilateral or multilateral. Many companies, institutions, governments, and so on. must handle contracts for a wide range of tasks. An example of this may be a process that verifies whether a clause in the contract is contrary to another or to a mandatory rule of law. As another example, it is possible to examine whether the legal document includes clauses that are binding on the company. Both cases only illustrate the current volume of work. Lawyers have to check many contracts many times until a final consensus is reached on both sides. Some of these views can be automated. This involves understanding and interpreting text [12]. Depending on the type of document, it must contain certain elements and have a specific structure. Processes of classification and verification of elements and structure can be automated. Another key task is to specify what the clauses are (required - price, goods, property, behavior, etc.) of the parties to the contract and what would happen and what would be due in case of non-execution of the contract – penalties, interest, etc. This can also be delegated to stand-alone systems that check the understanding document.

Similar contract processing involves reading with understanding using NLP and similar approaches like word-embeddings. For the training of such a system, a certain amount of contracts will be needed. Also other set of contacts is necessary to verify the correctness of the model obtained. The task of such a system is to define the elements of a text, to extract the structure and the parts of the contract. Including to emphasize how much the document differs from the standard of this kind.

An important part of the processing of such information is the need to anonymize personal data. We do not want our model to make decisions on a name-based basis and, above all, such data should be well protected and not publicly available.

3.2. Legal study

Apart from the analysis of a document, AI can also be used with a wider range [3]. This range includes knowledge and prediction systems.

Knowledge-based systems are software based on law, time-based regulatory changes, judgments and motivations in it, opinion of experts in the field, lawyers' opinions, scientific developments, etc. Such systems can analyze a problem very well thanks to the understanding of natural language and provide a response to a case. Although such systems seem like a search engine in a large database, their great advantage is the understanding of the language and the results they provide are not just based on keywords but also on the meaning itself.

Such systems work in a given language in most cases English is used. Though it is not part of the current article, a system that combines good practices across different legal system could be a very good research field. Such combination will significantly increase training data set and most of all, will help to transfer good practices from one legal system to another if local law permits it. This study remains for future development.

The systems that "predict" include an analysis of court judgments that have already entered into force and the production of a result for a new case. We are far from the moment when such systems can replace people in the decision-making process, but they can speed up the whole process of dealing with a case. This approach is extremely useful when we have analogous case studies and can quickly show us the relevant case law. This trained model will have an extremely large set of cases to work on and the result obtained can be as accurate as possible. The main purpose is to reduce the manual search of different judgments, motives, opinions, etc., which process takes a very long time.

3.3. Intelligent interface

In addition to the direct application for document processing and casework, the possibility of such software being used for both student and non-lawyer training in answering legal questions should not be underestimated. The training process involves examining and analyzing good practice and case studies. The typical process involves searching in documents and textbooks to find the right text. An interface that responds to user queries will be very useful in the learning process and will speed it up and, above all, increase quality because the information that will be offered will be as accurate, up to date, comprehensive as possible in more understandable language. By the same logic, such an interface would also be useful to non-jurists in solving trivial problems, which would again save a huge amount of time globally. This type of software will resemble an intelligent search engine that has to work very well with a legal context.

An example of such an interface in the field of learning is one that could answer the question: "What kind of documents are need during car police check?" “Could a pedestrian be guilty in case of an accident on a pedestrian crossing? "," What documents are needed in particular case - when issuing new identity documents - ID card or passport?". Such a platform would save a great amount of time and bring confidence and security to citizens, whether lowers or not.

4. Data extraction

The extraction of knowledge can be defined in many different ways. Even the name can be defined as the discovery of knowledge in the data [13]. It is so vast discipline that none of the definitions describe the entire process that takes place during the acquisition of knowledge. In general, the process involves data processing in order to extract useful information. It can be represented as 7 consecutive steps that are performed until the desired results are found [13].

1. Clear data - data must be cleared from noise before it can be processed [13].
2. Data Integration - Checks whether multiple sources can be combined to achieve as much knowledge as possible [13].
3. Selection of data - check how much data is correct for the task we are trying to solve [13].
4. Transformation of the data - the data shall be prepared in a suitable type for processing [13].

5. Knowledge extraction - a major part of the process where intelligent processing systems are involved. This part is fundamental to the current article, because the NLP method is just right for it [13].
6. Model Evaluation - Identify the really interesting models (templates) found in the data [13].
7. Presenting knowledge in a man-readable form [13].

5. NLP

Natural language processing (NLP) is a set of approaches that are used to process text data. The entire processing text being worked on is called a corpus. The use cases of such data processing are many - check grammar and spelling, knowledge of the next word in speech, retrieve user-based query information (intelligent interfaces, including the legal ones that were discussed in the article), categorization text, summary, etc. [14].

Data mining is a basic method that will also be used to process legal documents, and ongoing work will show several specific approaches and algorithms that can be used [15].

A word bag [15] - one of the simplest yet effective methods for presenting machine learning data. This approach ignores the order of words in sentences, the sentences and structures themselves. It loses part of the information but gives a general idea of the content of the text. Three things are taken into account:

1. Tokenization - splitting the document into words
2. Define the words and number them
3. Count how many times a word occurs.

Not all words have the same weight and a list of words that are not involved in the processing should be made - pronouns, pronouns, etc. There are similar lists of words, mainly in English, but it is not a problem to do so in other languages as well.

This method can be improved by not losing the word order. In order to do this, the word bag is not built with just one word, but with a set of consecutive words. The approach is called n-gram model, n being the number of words used to compose the word bag [15].

TF-IDF is another, very powerful method for extracting a common knowledge of text where, apart from counting words, this number is also compared to all other words in the corpus. [16]

The above methods give a general idea of the content of the text, but the legal documents will need a deeper understanding. To achieve such an understanding, methods need to be improved. One such improvement can be made in the first step of the bag with words - tokenization. In the texts there are always words that are in different forms (verb tenses, singular or plural, congressional) which do not contribute much to the ultimate understanding of the text, but rather saturate the pattern. During processing, these words can be replaced with their roots, thus reducing the number of words being worked on [15].

Also, to improve tokenization, not only the roots of the words, but also the synonyms can be used, i.e. how close the words to each other are in terms of meaning. This can be achieved by using another approach in advance - presenting the words as vectors and searching for the proximity of the vectors [4].

Another very good method of text processing is the use of recurrent neural networks. Unlike other methods that analyze a test, neural networks can generate text that makes them very convenient for use in intelligent interfaces [15].

6. Tools

To achieve the above tasks, NLP is used. As it has become clear, this is an approach to understanding text written in natural language. Based on it, various tools have been developed to facilitate the development of client-end software that can solve a given legal task with understanding.

Standford-parser [4] – one of the most commonly used NLP tools. Provides tools for solving subtasks from the above-described - basic parts of words, parts of speech, normalized teachings, numbers, etc. This way you can find how the different words are connected to one another. The library is based on the statistical model and works fast. It is mainly used for English.

Apache OpenNLP [6] – a very good text-processing library in natural language. It is very useful for solving some of the tasks in the 3 main directions that are discussed in the article. Maintains basic algorithms such as tokenization, segmentation of sentences, determination of the parts of the speech. These techniques are needed to create more sophisticated text-processing systems.

TensorFlow [17] – an open-source Google library for creating and training neural networks. Everything, working with this library, is presented as tensors of different rows.

SpaCy [18] – a statistical tool for extracting data from large arrays. Works fast with large text processing as input. As part of the functionalities it supports, it is syntax-based sentence separation, good integration with machine learning systems that includes TensorFlow.

Parse-IT – part of the digital law platform [19]. The tool adopts normative text as input and finds logical structures. Provides a fine tuning capability.

The tools mentioned here are fundamental to text processing in natural language, most commonly used is the first tool. Using only once of them could not give a direct answer and does not solve directly the tasks listed above in the article. Rather, these tools are the alphabet used to write a book, a book that solves legal problems with AI and machine learning. Instruments are mainly designed to work in English, which opens the next obstacle to their broad application, namely how easily they can be adapted to the local language of the countries. Which will be more appropriate - whether reworking the tools and re-engineering the models in a given language or automatically translating a text into English and vice versa. The answer to this question is beyond the scope of the article and is rather a reason to develop a study in this area.

7. Summary

Automated approaches, based on natural language processing, combined with machine learning can be used to process the legal documents. They cannot replace experienced lawyers, but they can do a lot about initial analysis - whether the contract is correct and includes the necessary clauses, what kind of document we have and whether it is structured correctly, automatic structure correction, intelligence interfaces and platforms, which take decision what lay is applicable in particular case. Even if it is not 100% correct, this type of systems could be an “initial” light when solving a problem and thus would save time for lawyers themselves, as well as for citizens, that is, processes can happen more quickly, more efficiently, and more accurately.

8. Conclusion

The article far cannot cover all aspects of the application of intelligent systems in law. The legal system will change over the coming years, and that should be for good. In some countries this has already begun, and in others it is about to start [5]. New companies engaged in AI and machine learning in law appear constantly [3], each solving different problems. The sphere is extremely difficult and complex, but it should start from somewhere and move forward.

The growing acceptance of AI in world law is before us. In the technological world, the new and modern today will be old tomorrow. We are not at a time when we have to think if technology can be useful, but rather try to accept it as quickly as possible and improve it so that we can make the most of it.

References

[1] Christian H¨anig, Martin Schierle, Daniel Trabold, 2010, Comparison of Structured vs. Unstructured Datafor Industrial Quality Analysis
[2] Risto Hiltunen, 2012, The Grammar And Structure Of Legal Texts
[3] Richard Tromans, 2016, Legal AI
[4] Douwe Osinga, Deep learning cookbook. 2018
[5] Neural Reasoning For Legal Text Understanding (Guido Boella Adebayo Kolawole John, Luigi Di Caro), In Proceedings of the 29th International Conference on Legal Knowledge and Information Systems (JURIX2016), 2016

[6] Andor, Daniel, et al. "Globally normalized transition-based neural networks." arXiv preprint arXiv:1603.06042 (2016)
[7] Milagro Teruel, Cristian Cardellino, Fernando Cardellino, Laura Alonso Alemany, Serena Villata, 2017, Legal text processing within the MIREL project
[8] Pavlova, M., Civil law - common part, 2nd edition, Sofi-R (Павлова, М. Гражданско право - обща част - Второ преработено и допълнено издание - Софи-Р), 2002, 38-44.
[9] Ilias Chalkidis, Achilleas Michos, Ion Androutsopoulos, 2017, Extracting Contract Elements
[10] Calamari, J.,J. Perillo. Contracts.Saint Paul, Minn.: West Publishing Co.,1987. p.3.
[11] Ruschev, I., Administrative and Private Law, Market and law (Русчев, И. Административният и частноправният договор. - Пазар и право), 2000, №1, 25-32.
[12] Adrien Bibal, Benoît Frénay, 2016, Interpretability of Machine Learning Models and Representations: an Introduction
[13] Jiawei Han, Micheline Kamber, Jian Pei, Data mining: Concepts and techniques, 2012
[14] Dr. Mariana Neves, Natural Language processing, SoSe 2016
[15] Andreas C. Müller and Sarah Guido, Introductin to machine learming with python, 2017
[16] Stoykov, K., Knowledge Extraction and Machine Learning, International Conference "Automatics and Informatics'2018", 4-6 Oct. 2018, Sofia, pp. 117 – 120, ISSN 1313-1869.
[17] Floris Bex, Joeri Peters and Bas Testerink: A.I. for Online Criminal Complaints: From Natural Dialogues to Structured Scenarios, in prof. of Artificial Intelligente for Justice workshop, collocated at the 22nd European Conference on Artificial Intelligence
[18] C. Bartolini, G. Lenzini and L. Robaldo: Towards legal compliance by correlating Standards and Laws with a semi-automated methodology, in proc. of the 28th Annual Benelux Conference on Artificial Intelligence. Amsterdam, 2016.
[19] Deerwester, Scott, et al. "Indexing by latent semantic analysis." Journal of the American society for information science 41.6 (1990): 391.