Relation Extraction from Environmental Law Text Using Natural Language Understanding

Heiko Thimm [1] and Phil Schneider[2]

Abstract: In the last decades the highly active area of environmental legislation has produced a vast amount of text documents that contain laws and regulations enacted by various types of rule setters. This large body of legal text documents is still growing with an increasing speed. In order to assure compliance with the regulations, today, corporate specialists spend a lot of time with the reviewing and assessment of these documents. It seems that through the use of text processing assistance tools these important corporate environmental compliance management tasks can be completed in less time. To develop corresponding assistance tools has been the broader goal of this work in which initial text processing experiments with a common Natural Language Understanding pipeline are described. The obtained results confirm that in order to extract meaningful relations from text documents of the environmental legislation area, domain-specific processing techniques that are tailored to the specific language and format of legal text are required.

Keywords: environmental legislation, legal text, Natural Language Processing, Natural Language Understanding, Relation Extraction, text processing pipeline.

Addresses Sustainable Development Goal 9: Industry, innovation and infrastructure

1. Introduction

Due to recent technology advancements in numerous domains, text processing applications that are based on Natural Language Understanding (NLU) techniques are being increasingly used for a variety of tasks [AHN19]. These applications, for example, support human users in administrative tasks, information search and acquisition tasks, judgement tasks, and decision tasks. Obviously, at the forefront of the use of NLU techniques are domains were large amounts of text documents are at the center of the core business processes. This characteristic is particularly true for the legal domain in general. Not only the daily tasks of law firms require to deal with large amounts of legal text documents. Also, corporate business processes require various legal specialist to frequently (often even daily) review large amounts of text documents. This is in particular a duty of environmental compliance management specialists. As part of their common daily duties [Th15] they need to check environmental regulations described in text documents published by environmental rule setters of various levels (community level,

[1] Pforzheim University, School of Engineering, Tiefenbronner Str. 65, Pforzheim, 75175, heiko.thimm@hs-pforzheim.de, https://orcid.org/0000-0002-6200-9655
[2] Pforzheim University, School of Engineering, Tiefenbronner Str. 65, Pforzheim, 75175, Phil.Schneider@web.de

state level, country level, international level, supra-national level). The documents either describe a new environmental regulation (e.g. law, directive, mandatory standard) or a revision of an already existing regulation (or revision). According to a rough estimate of environmental compliance management practitioners, a typical mid-sized globally acting production company with multiple production sites in different world regions, needs to check on a monthly basis several douzens of new English and Non-English environmental regulations (or revisions). Many of them can be filtered out right away because eligibility criteria of the regulation are not fulfilled. For the set of remaining regulations more extensive investigations are necessary possibly requiring group decisions. The investigations typically yield about 2-3 regulations which require compliance enforcement measures such as information measures, training measures, equipment/infrasturcture measures, or product revision or production process revision measures. Note that also any revision of products or/and production processes may require to review text documents of environmental legislation in order to check and assure that the change is in compliance with the relevant environmental legislation [Th15].

Our long term research targets to develop tools that assist environmental compliance specialists in their duty to review and analyse legal text documents for judgement and decision tasks. In particular the tools are intended to enable companies to complete core environmental compliance management tasks in less time such as the relevance assessment task, the measure determination task, and the maintenance of a regulation registry [Th15]. Inspired by recent advancements in the area of LegalTech [DA19], [Ha19] and LegalAI [Zh20], in the initial phase of our research, we explore potential possibilities of NLU approaches. The results will be used to develop and test assistance systems for corporate compliance management tasks in particular the task to review an analyse legal text documents. Based on a review of NLU methods and Natural Language Processing (NLP) methods and techniques, a NLU pipeline was implemented which is able to extract relations from documents. The pipleline was tested with various text documents from the environmental legislation domain. In the further continuation of this ongoing research this initial NLU pipeline will be optimized and also other pipelines will be developed in order to test other NLU approaches including approaches that use machine learning techniques.

The following Section 2 gives a general overview of NLP and NLU and corresponding common main methods. Our initial experiments with a NLU pipeline for relation extraction and an outlook on forthcomming further experiments are described in Section 3. Section 4 contains our conclusions.

2. NLP and NLU – Overview and Methods

Both NLP and NLU focus on making sense of unstructured text data, but there is a difference between the two. NLP is primarily concerned with how computers are programmed to process language and to enable `natural´ communication between computers and humans. NLP processes are more of a statistical or pattern matching process to derive information from unstructured text data. NLU, on the other hand,

prioritizes the ability to understand human language and, thus, refers to how unstructured data is reorganized for machines to be able to `understand´ and analyse it [Ru06]. Initial NLU approaches analyse rules and grammatical characteristics to understand language. More recent approaches make use of Machine Learning techniques [Le22]. NLP and NLU often supply complementary solutions to a problem since they have different theoretical backgrounds, one statistical and one rule-based or Machine Learning-based. But some researchers suggest to view NLU as a subarea of NLP. Fig. 1 adopted from MacCartney's presentation on `Understanding Natural Language Understanding´ at the Inaugural Meeting of the ACM Special Interest Group on AI of the Bay Area Chapter in 2014 contains a corresponding visualization of this view on the two disciplines and identifies for each discipline common problems and application areas [Ma14].

So-called `NLU-pipelines´ refer to a series of steps in which complementary NLP/NLU processing tasks are performed on a given input text or a text corpus in order to provide particular application results. Typical results are a summary, the overall topic, the category, the sentiment of the input text or information being extracted such as named entities and relations among entities.

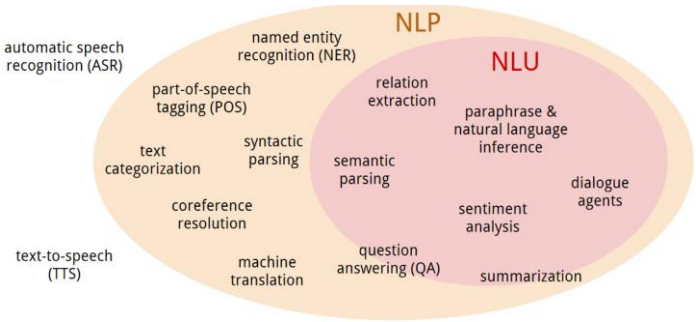

Fig. 1: Terminology NLP vs. NLU including typical applications (copied from [Ma14])

Typically, the first steps of an NLU pipeline perform pre-processing tasks to prepare the input text and, henceforth, enable that the application-specific goal can be achieved by the further processing steps. A common step that follows the text pre-processing is Part-Of-Speech Tagging (POS) which attempts to associate words and symbols in a text with word categories. A brief overview of common pre-processing techniques and POS techniques are contained in the next two paragraphs. This is followed by overviews of embedding-based methods and symbol-based methods. These two types of methods are primarily applied in the legal domain [Zh20]which subsumes the particular application domain of this research.

Common Pre-processing methods. The common pre-processing methods of most NLP tasks are tokenization, stemming, and lemmatization [AHN19], [Ma14]. The purpose of tokenization is to break chunks of language input into sets of tokens that correspond to paragraphs, sentences, and words. After the tokenization step, the text is converted to lower-case followed by an elimination of numbers, punctuations, and stop words such as

`and´, `the´, `a´, `an´ and similar words [Na18]. That is, basically everything which is redundant and does not convey any meaningful insight for the data gets eliminated [AHN19]. Stemming is the process of reducing infected or derived words to their word stem, base or root form. It basically affixes to suffixes and prefixes or to the roots of words known as a lemma. A stemmer removes the endings of many words, e.g., `consolidate´, `consolidated´ and `consolidating´ would be converted to `consolid´ [AHN19]. Lemmatization is the process of reducing inflected forms of a word while still ensuring that the reduced form belongs to the language. This reduced form or root word is called a lemma. For example, `organizes´, `organized´ and `organizing´ are all forms of the lemma `organize´. The inflection of a word allows to express different grammatical categories like tense (`organized´ vs. `organize´), number (`trains´ vs. `train´). Lemmatization is necessary because it helps to reduce the inflected forms of a word and enables to analyse them as a single item. It can also help to normalize the text. As a result, the content becomes more understandable and obtains a clear meaning [AHN19], [Gh20].

Part-Of-Speech Tagging (POS). POS is a common NLU technique which explores the role of a particular word in a sentence. The technique uses eight so-called `parts of speech´: noun, pronoun, adjective, verb, adverb, preposition, conjunction, and interjection. Single words get mark-ups in the text with pre-defined tags such as `N´ for noun, and `V´ for verb. The POS tags, for example, enable keyword extraction based on filtering nouns that typically carry the most significant information [Gh20].

Embedding-based methods aka representation learning. In the legal domain these methods focus on the representation of legal facts and knowledge in an embedding space. In order to solve application-specific tasks they may use deep learning methods. However, these methods require a large amount of prepared training data. Word Embedding methods can fill the gap between text and vectors and convert text into a format that can be processed by AI techniques [Zh20]. Word embeddings are words which are represented by a vector or an array of real numbers. Through the embedding process sentences are transcribed into an array of words and each word of the sentence is transformed into an embedding. Semantic similarity measurement methods can explore the connections between the words and compute similarities like the human memory. Words represented by a vector offer many advantages. They facilitate designing and training deep neural networks since the input consists of machine processible vectors instead of words. Several mathematical techniques for processing numbers are available and can be applied to perform classification, feature extraction, etc.

Symbol-based methods aka structured prediction methods. These methods use knowledge from the respective application domain to solve the NLU task. For example, in the legal domain, especially parts of the symbolic legal knowledge, such as events and relationships, can provide interpretability for lawyers [Zh20]. Two main approaches that use symbol-based methods are information extraction and relation extraction. Information extraction addresses the general problem of detecting entities referred in texts, the relations between them and the events they participate in. Informally, the goal is to detect elements such as `who´ did `what´ to `whom´, `when´ and `where´. It is the general primary goal to

convert a large amount of text into a formal representation of specific fine-grained facts. The structured data obtained from the input text refer to events, entities, facts or relationships between entities presented in the text. This structured information allows computers to perform logic inference or computation on the data, which is challenging if only raw text representation is used [TNS16]. Named Entity Recognition (NER) is a method often used to analyse the text for specific information like names, places, etc. [Gh20]. Relation Extraction focuses on discovering the semantic relations among entities in a text. Various Relation Extraction techniques extract relationship instances that belong to a set of relationship types. These techniques can be grouped into rule-based approaches, supervised approaches, and semi-supervised approaches [Ba16]. Rule-based approaches extract pre-defined relationship types from manually-crafted rules. Supervised approaches use manually labelled documents where the labels describe the type of relationship between each recognized pair of entities. A manually labelled collection of documents is used to train classifiers which, henceforth, are capable to identify all trained relationship types in any dataset. Semi-supervised approaches make use of known relations to recognize new relationships. From the textual contexts of the established relationships, the method derives new approaches and patterns, which in turn are used to derive new relationships.

3. Experiments with an NLU Pipeline for Relation Extraction

The goal of this research is to evaluate the information extraction capabilities of NLU techniques in the domain of corporate environmental compliance management. Through respective experiments we seek to answer two questions: 1. to which extent can useful information be extracted from text documents that are typically reviewed by corporate compliance managers and 2. what specific NLU pipeline is capable to perform this task. Of the planned series of experiments with different NLU pipelines, in a first experiment we used the common NLU pipeline displayed in Fig. 2 which extracts relations from text documents. The raw text of the document is in the initial steps of the pipeline parsed into sentences through a sentence segmenter and further split into words using a tokenizer. Then, each sentence is tagged with POS tags. In the next step, entities are extracted from the text. Finally, rule-based relation extraction is applied to identify relations between different entities in the text and to provide detected relations as tuples [BKL09]. The tuples consist of three elements referred to as triples that are visualized in the form of Knowledge Graphs [Ke22]. Consider for example the triple `(Berlin, capital, Germany)´ that might be extracted from a short text about Germany. Typically, the first element of a triple corresponds to the subject, the second corresponds to the relation, and the third corresponds to the object [Ma14].

The pipeline was implemented based on the programming language Python resulting a first version program referred in the following as `Compliance Management Information Extractor´ abbreviated CMI Extractor. Various general-purpose NLP/NLU packages and other common Python packages were used for specific steps of the pipeline.

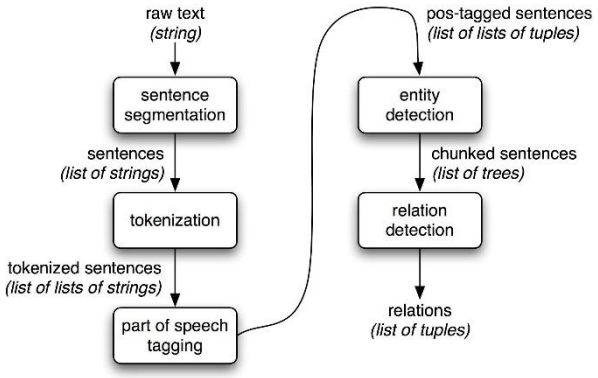

Fig. 2: Relation Extraction pipeline of the CMI Extractor (copied from [Le19])

In order to test and demonstrate the capabilities of the CMI Extractor a test run with the following six relatively short and easy to understand sentences about Elon Musk were performed: *Musk married Riley. Musk is the CEO of Tesla. Musk has a net worth of US$245 billion. Musk is the wealthiest person in the world. Musk founded The Boring Company. Musk graduated in 1995 with a Bachelor of Arts degree in economics.*

The CMI Extractor was able to correctly extract the relations contained in each of the six sentences. The resulting knowledge graph is displayed in Fig. 3. However, it is crucial to bear in mind that the test document consists of simple sentences which represent an ideal input that does not raise complicated processing challenges for the pipeline.

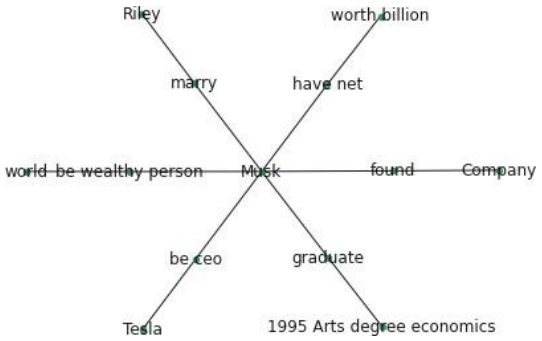

Fig. 3: Knowledge Graph of test run with ideal text input

Obviously, it is much harder for the CMI Extractor to obtain meaningful relations from legal text documents of today's environmental legislation. In order to obtain corresponding insights, the experiments described in Tab. 1 were performed with text segments of two specific valid directives of the European environmental legislation.

Run	Text	Results / Comments
1	ideal text comprising six simple short	Six meaningful relations

	declarative sentences with facts about Elon Musk; 42 words	extracted; see Fig. 3
2	List item (27) of EU Directive 2020/2184; 264 words; text consists of two complete sentences and a numbered list of 11 list items; list items are comma-separated descriptions containing expert terminology	Four complete relations extracted which require further investigations in order to obtain useful benefits for compliance specialists; see Fig. 4
3	*`Article 2 – Scope´* of EU Directive 2011/65; 261 words; about 80% of the text consists of a numbered list; list items are separated by semicolons; list items consist of several lines of comma separated descriptions containing expert terminology	One complete strange and meaningless relation extracted
4	*`Article 7 - Obligations of manufacturers´* of EU Directive 2011/65; 519 words; same characteristics as described for run 3	Two strange and meaningless relations extracted
5	*`Article 8 - Obligations of authorized representatives´* of EU Directive 2011/65; 162 words; same characteristics as described for run 3	Three strange and meaningless relations extracted
6	*`Article 9 - Obligations of importers´* of EU Directive 2011/65; 432 words; same characteristics as described for run 3	Four strange and meaningless relations extracted
7	*`Article 10 - Obligations of distributors´* of EU Directive 2011/65; 308 words; same characteristics as described for run 3	Five strange and meaningless relations extracted

Tab. 1: Experiments with the CMI Extractor

In the second test run the CMI Extractor was tested with a text fragment of the European Parliament's directive 2020/2184 which targets `the quality of water intended for human consumption´ [Eu20]. Through a random choice from page 8 the list item with number 27 was chosen that consists of 8 sentences and 264 words. In this experiment the CMI Extractor extracted four complete and one incomplete relation (i.e. triples) that are displayed in the knowledge graph of Fig. 4. As opposed to the above experiment the meanings of the extracted relations are less obvious to understand. Using the evaluation framework of a recent Japanese study [TNS16], the extracted relations are to be judged as both *incoherent relation*s and *uninformative relations*. In fact, further investigations are required in order to obtain benefits for compliance specialists from this extraction result.

Fig. 4: Knowledge Graph of test run with a text segment of an EU directive

For the test runs 3 to 7, we used randomly chosen articles of the EU Directive 2011/65 that focusses on `restriction of the use of certain hazardous substances in electrical and electronic equipment´ [Eu11]. In every of these five input texts the CMI Extractor was only able to find some strange and even `more incoherent´ and `more uninformative´ relations as in the second run. It can be assumed that compliance management specialists will not be able to draw any helpful information from these relations. One of the possible reasons for this result is that the input text segments (i.e. articles of the directive) are significantly different from the above ideal text input about Elon Musk and also most common text documents in general. Each of the five test text segments consists to a large degree (~80%) of a numbered list with semicolons separating the different list items. The list items themselves consist of several lines of comma separated enumerations of terms. Many of these terms belong to the expert terminology of the domain of environmental legislation. Even with a revised NLU pipeline that was able to treat semicolons similar to end of sentence points, the extraction performance of the CMI Extractor did not improve.

The above described experimental results confirm the expectation that an NLU pipeline of general-purpose text processing components will only have very limited capabilities to extract useful information from domain-specific text documents. Hence, in the ongoing phase of this research we are exploring approaches to develop a next CMI Extractor that is capable to deal with both 1. the specific language style of environmental legislation documents and 2. the specific terminology of environmental legislation. Consequently, a systematic study of the different types of environmental legislation documents is being prepared and existing dictionaries for the work of corporate environmental compliance management (e.g., [DHS19]) are being evaluated and possibly extended.

One of the promising options to achieve the targeted improvement of the CMI Extractor is to make use of the open source Python package LexNLP which is focused on natural language processing and machine learning for legal and regulatory text [BKD21]. LexNLP supports, among others, the building of unsupervised and supervised models such as word embedding models and tagging models. The package also [BKD21] ´[…] includes pre-trained models based on thousands of unit tests drawn from real documents available from the SEC EDGAR database as well as various judicial and regulatory proceedings´. We also intend to experiment with the popular language processing tool GTP-3 [ZL21] from OpenAI that is considered to be a foundation model. Foundation models can even be

trained on multiple forms of data at the same time. Experiments of our future research will also include to train GTP-3 by use of hand crafted/curated domain knowledge together with industry partners in order to develop a CMI Extractor version with advanced information extraction capabilities. The future refinement of the CMI Extractor will also address the capability to deal with multiple languages.

4. Conclusions

In recent years the business world has been paying increasing attention to the new possibilities that the latest advancements of text processing technologies provide for the digitalization of corporate processes. However, relatively little work of researchers and AI-based software start-ups is devoted to the new possibilities that these advancements offer to the work field of corporate environmental compliance management. Our research attempts to fill this gap by developing and evaluating NLU/NLP-based assistance tools which ultimately extract important information from environmental legislation documents. The extracted information, for example, can be displayed in Knowledge Graphs, thus enabling compliance managers to make fast assessments about the relevance of the document. Furthermore, the extracted information can also be used to generate recommendations for the users based on AI techniques such as machine learning methods. For example, when a legislation document is of relevance for a company and compliance measures are required then recommendations for these measures can be generated from the extracted information and further domain specific knowledge. With respect to this goal, our research is still in an infancy state and will therefore in the future address further experiments with different NLU pipelines. It is expected that from the corresponding experimental results useful recommendations can be obtained for law making regarding meta data and syntactic rules for the legal documents to enable proper NLU support.

Bibliography

[AHN19] Anandarajan, M.; Hill, C.; Nolan, T.: Practical Text Analytics. Maximizing the value of text data. Springer-Verlag, Cham, 2019.

[Ba16] Batista, D. S.: Large-scale semantic relationship extraction for information discovery. Doctoral Dissertation, Liboa, 2016.

[BKD21] Bommarito II, M. J.; Katz, D. M.; Detterman, E. M.: LexNLP: Natural language processing and information extraction for legal and regulatory texts. In (Vogl, R. Ed.): Research handbook on big data law. Edward Elgar Publishing, Northampton, Massachusetts, USA, pp. 216–227, 2021.

[BKL09] Bird, S.; Klein, E.; Loper, E.: Natural language processing with Python. O'Reilly, Sebastopol, California, 2009.

[DA19] DALE, R.: Law and Word Order: NLP in Legal Tech. Natural Language Engineering 1/25, pp. 211–217, 2019.

[DHS19] Deng, Q.; Hine, M., Shaobo, J.; Sujit, S.: Inside the Black Box of Dictionary Building for Text Analytics: A Design Science Approach. J. of Int. Technology and Information Management 3/27, pp. 119–159, 2019.

[Eu11] European Parliament: DIRECTIVE 2011/65/EU on the restriction of the use of certain hazardous substances in electrical and electronic equipment, 2011.

[Eu20] European Parliament: DIRECTIVE 2020/2184 on the quality of water intended for human consumption, 2020.

[Gh20] Ghavami, P.: Big data analytics methods. Analytics techniques in data mining, deep learning and natural language processing. De Gruyter, Boston, Berlin, 2020.

[Ha19] Haney, B. S.: Applied Natural Language Processing for Law Practice. SSRN Electronic Journal, pp. 1–44, 2019.

[Ke22] Kejriwal, M.: Knowledge Graphs: A Practical Review of the Research Landscape. Information 4/13, p. 161, 2022.

[Le19] Learntek: Named Entity Recognition with NLTK. https://www.learntek.org/blog/named-entity-recognition-with-nltk/, accessed April 14th, 2022.

[Le22] Lexalytics: Machine Learning for Natural Language Processing and Text Analytics. https://www.lexalytics.com/resources/wp-content/uploads/sites/3/2019/02/Lexalytics_Machine_Learning_Natural_Language_Processing_Whitepaper.pdf, accessed 22 Apr 2022.

[Ma14] MacCartney: Understanding Natural Language Understanding. ACM SIGAI Bay Area Chapter Inaugural Meeting, 2014.

[Na18] Nay, J.: Natural Language Processing and Machine Learning for Law and Policy Texts. SSRN Electronic Journal, pp. 1–35, 2018.

[Ru06] Rune Sætre: GeneTUC: Natural Language Understanding in Medical Text. Doctoral thesis for the degree of doktor ingeniør, Trondheim, 2006.

[Th15] Thimm, H.: IT-Supported Assurance of Environmental Law Compliance in Small and Medium Sized Enterprises. Int. Journal of Computer and Information Technology 2/4, 297-305, 2015.

[TNS16] Tran, C.-X.; Nguyen, M.-L.; Satoh, K.: A study of open information extraction from legal texts: Proc. First Int. Workshop on Scientific Document Analysis (SCIDOCA), p. 7, 2016.

[Zh20] Zhong, H. et al.: How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence. In (Jurafsky, D. et al. Eds.): Proc. 58th Meeting of the Assoc. f. Comp. Linguistics. Assoc. f. Comp. Linguistics, Stroudsburg, PA, pp. 5218–5230, 2020.

[ZL21] Zhang, M.; Li, J.: A commentary of GPT-3 in MIT Technology Review 2021.