Legal Knowledge and Information Systems
E. Schweighofer (Ed.)

doi:10.3233/FAIA210327

Signal Phrase Extraction: A Gateway to Information Retrieval Improvement in Law Texts

Michael VAN DER VEEN [a] and Natalia SIDOROVA [a]
[a] *Eindhoven University of Technology, The Netherlands*

Abstract. NLP-based techniques can support in improving understanding of legal text documents. In this work we present a semi-automatic framework to extract signal phrases from legislative texts for an arbitrary European language. Through a case study using Dutch legislation, we demonstrate that it is feasible to extract these phrases reliably with a small number of supporting domain experts. Finally, we argue how in future works our framework could be utilized with existing methods to be applied to different languages.

Keywords. information retrieval, legislative texts, signal phrase extraction

1. Introduction

Legislative texts are complex and information-dense by their nature. Automatic analysis of these texts is an active research field with applications in different areas ranging from document annotation and legal text generation [1] to rule extraction [2] and verdict prediction [3]. One of the complicating factors in a wider employment of NLP-techniques in the legal domain arises from the fact that all countries have their legislation in their national language(s). Many NLP-based techniques, like rule extraction or event mining [4], can potentially be implemented as multi-language tools. They are however based on the use of signal words or linguistic patterns [5], which are language specific. There are many efforts to provide support both for specific languages [6,7] and across multiple languages [8]. Still multiple gaps need to be filled to achieve this goal.

In this paper we focus on the problem of automated generation of categorized lists of signal words and linguistic patterns used in the legal domain and indicating causal or temporal relationships. These signal words and phrases are necessary for e.g. legal text annotation and rule extraction. Signal words and phrases used in the legal area often differ from the ones in regular language usage. For example, “mits” (provided that) is rarely used in modern spoken and written Dutch, but it is very common in legislative texts. Our goal is to develop a general framework for extracting signal words and phrases from legislative texts in a given language using language-independent techniques. We demonstrate the use of our approach on the example of the Dutch language.

In Section 2, we introduce our semi-automatic framework. In Section 3 we apply and evaluate our framework on Dutch legislation. We draw conclusions and discuss future work in Section 4.

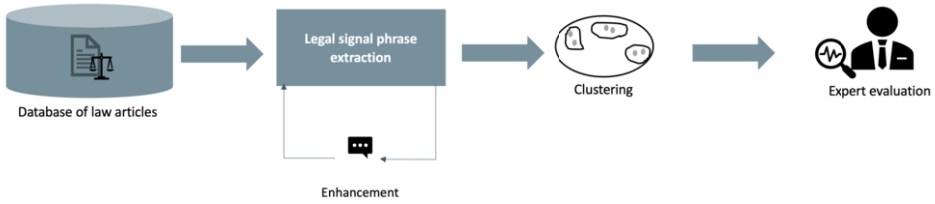

Figure 1. Framework to extract relevant signal words and phrases

2. Methodology and Framework

Our framework aims at the semi-automatic identification and categorisation of words and phrases indicating conditions and causal or temporal relationship between activities in legislation documents. We call these words and phrases *signal phrases*. Figure 1 illustrates the steps of the framework.

Step 1 First, we extract potential signal phrases. The extraction is based on thresholds for the total *frequency* of *n-grams* with $n \leq 3$ in all the laws and for the *coverage*, which is here the percentage of laws in which the signal phrase occurs. Sufficiently high coverage ensures that signal phrases are general for legal texts and span across multiple legal domains and laws. This reduces the number of false positives in the form of expressions frequent within certain legal areas but not used in other areas and therefore not carrying any causal or temporal meaning. Stop words are discarded from the search results. The choice of thresholds and the parameter 3 for n-grams was based on the input obtained from domain experts, who were enquired to deliver a list of typical signal phrases they expect to see in legal documents. Their lists were bundled and analysed on their total frequency and coverage and on their properties such as POS-tag. Our search strategy consisted of first selecting thresholds that would guarantee that all the key phrases provided by experts would be included in the search results. We do that to minimize the number of false negatives. Then, with each step we added more constraints on POS tags (based on our expert curated list' properties), e.g., including VERBS with coverage > 0.9. This process is repeated, until not too many phrases (< 400) are selected, while maintaining an adequate recall score.

Step 2 Phrases extracted in Step 1 are embedded and then clustered to their respective category. In our framework, we use the Universal Sentence Encoder introduced in [9], as it is able to handle multiple languages and n-grams with $n > 1$. We define a number of categories for the Dutch language based on [10], e.g., conditional, temporal and opposing. In further analysis, conditional phrases could be translated to different forms of implications, e.g., $A \longrightarrow B$, $\neg A \longrightarrow B$. We predefined a centroid per category to make sure that clustering leads to interpretable results. The embedded phrases are clustered around these pre-defined centroids using the cosine distance.

Step 3 Finally, interview sessions with domain experts were conducted. The main purpose of these sessions is to remove false-positive phrases. We also check the consistency of answers. Each interviewee receives a set of phrases consisting of two parts: one is the same for all of them (in order to check the consistency of answers) and the other is distinct. Additionally, each phrase is to be evaluated by two interviewees. Phrases are to be presented to interviewees in the context of their usage, in order to facilitate the work of experts.

3. Evaluation and Results

For this case study 1413 Dutch laws were utilized, scraped from the Dutch government website[1]. In the first step of our framework we initally set *frequency*: 1000 and *coverage*: 0.25 to include the expert curated phrase list of 36 items. After our first search we found 1453 phrases. Finally, after selecting phrases with POS-tags ADP, ADV, VERB(coverage$>$ 0.9) and SCONJ(coverage$>$ 0.35), 322 phrases were selected for our next step.

After the extraction step, we embedded the phrases and clustered them. To evaluate whether our embeddings worked correctly, we made a subset of phrases for which synonyms that originate from an online database[2] exist. Using the purity measure described in [11], we checked whether synonym phrases were assigned to the same cluster. This resulted in a perfect score of 1.0, which indicates an adequate embedding quality.

We conducted interviews with 5 experts. Each of them received 72 or 73 phrases from all clusters found in [10]. In the subset creation, we ensured that the consistency amongst the interviewees could be measured by including the same 10 randomly selected phrases to the set of each interviewee. Due to time constraints, we were not able to ensure that each phrase was evaluated twice. To check the consistency of evaluation of the 10 overlapping phrases we used the lower bound on the error relative to the (unknown) ground truth [12]. When the error rate is lower than 0.10, we can assume that the results consistently propagate to the non-overlapping phases [12]. The results of our interviews show that for classifying true positive (TP) phrases, we have an error rate of 0.08. The error rate for categorization was 0.24. This means that the selection of TP phrases can be considered as reliable. However, the evaluation of clusters is less sound. In future work, an experiment setting where at least 2 experts evaluate each phrase is required. Whenever they are in conflict, more analysis on context and semantics could prove useful One of the phrases where the experts were in conflict was ”op basis van” (based on). Some experts denoted this phrase as an explaining phrase, while others state that it is a referencing phrase. Both explanations are possible, depending on the context in which this phrase occurs and this shows that context information should be included in the analysis.

The experts selected $204/322(0.634)$ phrases as TPs. Several TPs were close to our predefined thresholds, which indicates that potentially there could be several false negatives. False positives were mostly phrases that are commonly used, but not specific enough for this research. Examples of such phrases are ”door” (by) and ”bedoeld” (meant). In the example for ”door”, we found that this phrase indicates a resource, which is not considered in the current setting of our research, as we focus on causal and temporal relations. It could be considered in future work since it maybe be important to extract such information. In the example ”bedoeld”, we found that this n-gram is too short to be recognized as relevant, ”als bedoeld” was considered a TP by experts.

From the selected TPs $104/204(0.510)$ were assigned automatically to the correct cluster. The clusters indicating examples and conditions were misclassified most often. This is probably due to the context-dependent nature of typical phrases in these clusters. It could also be caused by the fact that the embeddings used were trained on a regular corpus rather than a corpus specific for the Dutch legal domain. Such phrases sometimes have different meanings in regular language than in legislation.

[1] https://wetten.overheid.nl
[2] https://synoniemen.net

4. Conclusion

In this work we proposed a framework to semi-automatically mine signal phrases from legislative texts. This method combines automated processes with domain knowledge provided by experts. Furthermore, our case study demonstrated that a relatively small number of domain experts is required to filter out false positives consistently. Classification of clusters into categories requires more domain experts and further analysis. The quality of the language model used to generate embeddings is critical for successful automatic clustering of signal phrases. Identifying nuances in logical and temporal structures inside each cluster of signal words requires a collaboration of experts in law and in logics.

In future works we plan to integrate our technique with several others, namely [13]. We also plan to enhance our framework by using EU legislative texts, which are published in all 24 official languages of the EU. We expect that this will allow us to reduce the number of false positives and false negatives, as well as facilitate categorisation and interpretation of signal phrases.

References

[1] Dale R. Law and Word Order: NLP in Legal Tech. Natural Language Engineering. 2019;25(1):211-7.
[2] Dragoni M, Villata S, Rizzi W, Governatori G. Combining NLP approaches for Rule Extraction from Legal Documents. In: 1st Workshop on MIning and REasoning with Legal texts (MIREL 2016); 2016. .
[3] Medvedeva M, Vols M, Wieling M. Using machine learning to predict decisions of the European Court of Human Rights. Artificial Intelligence and Law. 2020;28(2):237-66.
[4] Filtz E, Navas-Loro M, Santos C, Polleres A, Kirrane S. Events Matter: Extraction of Events from Court Decisions. In: Legal Knowledge and Information Systems. IOS Press; 2020. p. 33-42.
[5] van der Aa H, Di Ciccio C, Leopold H, Reijers HA. Extracting Declarative Process Models from Natural Language. In: International Conference on Advanced Information Systems Engineering. Springer; 2019. p. 365-82.
[6] Koeva S, Obreshkov N, Yalamov M. Natural Language Processing Pipeline to Annotate Bulgarian Legislative Documents. In: Proceedings of The 12th Language Resources and Evaluation Conference; 2020. p. 6988-94.
[7] Waltl B, Landthaler J, Scepankova E, Matthes F, Geiger T, Stocker C, et al. Automated Extraction of Semantic Information from German Legal Documents. In: IRIS: Internationales Rechtsinformatik Symposium; 2017. .
[8] Doncel VR, Ponsoda EM. LYNX: Towards a Legal Knowledge Graph for Multilingual Europe. Law in Context A Socio-legal Journal. 2020 Dec;37(1):175-8.
[9] Yang Y, Cer D, Ahmad A, Guo M, Law J, Constant N, et al. Multilingual Universal Sentence Encoder for Semantic Retrieval. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations; 2020. p. 87-94.
[10] Bovenhoff M, Zeijl W. Basisboek taal. Pearson Education; 2009.
[11] Sanderson M. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008. ISBN-13 978-0-521-86571-5, xxi+ 482 pages. Natural Language Engineering. 2010;16(1):100-3.
[12] Smyth P. Bounds on the mean classification error rate of multiple experts. Pattern Recognition Letters. 1996;17(12):1253-7.
[13] Ferraro G, Lam HP, Tosatto SC, Olivieri F, Islam MB, van Beest N, et al. Automatic Extraction of Legal Norms: Evaluation of Natural Language Processing Tools. In: JSAI International Symposium on Artificial Intelligence. Springer; 2019. p. 64-81.