Legal Knowledge and Information Systems
E. Francesconi et al. (Eds.)

doi:10.3233/FAIA220472

Extracting References from German Legal Texts Using Named Entity Recognition[1]

Silvio PEIKERT [a,2], Celia BIRLE [a,3], Jamal AL QUNDUS [b,4]
Le Duyen Sandra VU [a,5] Adrian PASCHKE [a,6]
[a] *Fraunhofer Institut für Offene Kommunikationssysteme, Berlin, Germany*
[b] *Middle East University, Amman, Jordan*

Abstract. Information extraction tasks are particularly challenging in specific contexts such as the legal domain. In this paper, Named Entity Recognition is used to make legal texts more accessible to domain experts and laymen. This paper focuses on extracting law references and citations of court decisions, which occur in various syntactic formats. To investigate this task a reference data set is constructed from a large collection of German court decisions and different NER-techniques are compared. Pattern matching, probabilistic sequence labeling (CRF), Deep Learning (BiLSTM) and transfer learning using a pretrained language model (BERT) are applied to extract references to laws and court decisions. The results show that the BERT based approach achieves F1 scores around 0.98 for both tasks and outperforms methods from prior work, which achieve F1 scores of 0.89 (CRF for law references) respectively 0.82 (CRF for court decisions) on the same data set.

Keywords. Named Entity Recognition, Knowledge extraction, Legal data

1. Introduction

Lawyers search for laws and past decisions in each new case, either to estimate a possible outcome or to use them as arguments or counter-arguments. “Search” implies the recognition of entities in the form of norms and decisions, either manually or automatically. Due to the large amount as well as the structure and complexity of the data available, exploring and manual recognition are complex and time consuming [1].

Named Entity Recognition (NER) systematically extracts semantic document features in order to support downstream tasks such as information retrieval. NER is not a simple task and is a research area of several AI disciplines, e.g. [2,3]. Legal texts and text documents in general contain a multitude of references to external entities, which provide important background information. These references can be used to define semantic and machine understandable representations of documents.

[1] partially funded by the German Federal Ministry of Education and Research (BMBF), grant no. 03COV03F
[2] Corresponding Author: silvio.peikert@fokus.fraunhofer.de, ORCID: 0000-0001-5716-1540
[3] ORCID:0000-0002-5652-831X
[4] ORCID:0000-0002-8848-1632
[5] ORCID:0000-0003-3307-675X
[6] ORCID:0000-0003-3156-9040

Beyond common entity types such as **person**, **location** or **organization** which can be extracted using many general purpose NER approaches, citations and references to **court decisions**, **laws**, **regulations** and **legal literature** are of particular importance in the legal domain. Although these references generally follow certain syntactic rules, in practice, a variety of deviating formats are encountered. Additionally, these references contain many special characters and abbreviations, which further complicates algorithmic NER. The following text fragment from a German court decision contains a reference to a **law** and two different *court decisions*:

> "(...) ist eine solche Entscheidung des erkennenden Gerichts gemäß **§ 238 Abs. 2 StPO** herbeizuführen (*BGH, Beschlüsse vom 14. Dezember 2010 - 1 StR 422/10, StV 2011, 458*; *vom 9. November 2017 - 1 StR 554/16*)."

The ability to automatically recognize and resolve such references enables more efficient information retrieval systems. These features improve accessibility by providing links and context information and more specific queries by exploiting semantic context. Furthermore, citation graphs can be used to identify documents of special importance using ranking algorithms.

The described applications require three steps: (1) the automatic recognition of references, (2) the separation of these references into fragments and identifiers and (3) the lookup of the references in a knowledge base. This paper focuses on the automatic recognition of references to **laws** and **court decisions**.

A data set for these two tasks is constructed from German court decisions and different NER approaches are compared. The approaches include pattern matching, probabilistic sequence labeling using conditional random fields (CRF) [4], deep learning using bidirectional LSTMs (BiLSTM) [5] and transfer learning based on a pretrained multitask language model (BERT) [6].

2. Related Work

NER efforts in the legal domain range from automatically identifying legal parties in court files, sometimes used to automate anonymization preceding publication in order to comply with data privacy standards [7,8], to using NER to build ontologies [9,10], to automatically annotate legal documents [11]. These applications rely on the recognition of common named entities such as **person**, **location**, **organization**.

Other types of applications, e.g. those trying to build citation graphs of legal documents in order to find legal precedent and other connected material [12,13,14,15,16] or those automating summaries of legal texts [17], consider additional entities, namely cross references to **legal norms and regulations**, **court decisions** and **legal literature**.

[18] applied different legal sentence classes to investigate the applicability of machine learning to different document types, while [19] considered legal contracts from German legal data and provided software for legal entity linking and extraction; they employed a two-stage process using NER and NED (Named Entity Disambiguation). Their best performance for NER reached an F1 score of 92%.

[20] also deals with the extraction of general legal named entities (not including citations) from German legal documents and achieved better performance by using BERT models.

[21] demonstrated on Canadian documents that the use of context might improve results further.

[12] focused on legal cross references. They used data from Luxembourg and extracted references using complex regular expressions. [22] defined a set of semantic classes and applied sequence labeling to evaluate a number of classifier models. They used approaches based on CRF and BiLSTM.

The approach described in section 3 builds on results from [22] and uses similar methods. This paper considers additional data sources and approaches based on pre-trained language models.

3. Methodology

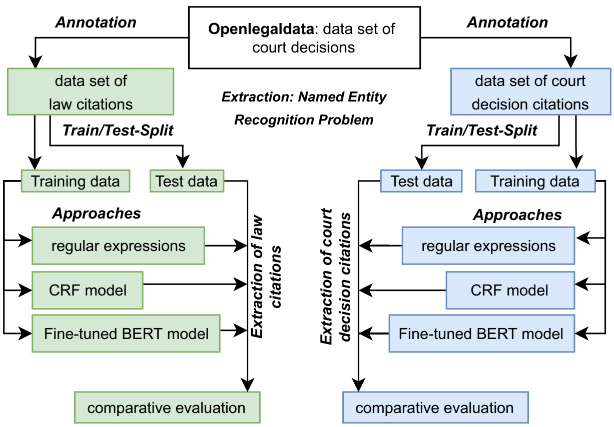

Figure 1. Overview of the Methodology

The detection and extraction of citations of law texts and court decisions can be modeled as a NER task. Two data sets, one for each NER task, were constructed. The extraction of these citations is approached using pattern matching, conditional random fields, bidirectional Long-Short Term Memory Networks and finally, transfer learning with BERT. Figure 1 summarizes the workflow used to compare these methods.

To construct a benchmark data set *Open Legal Data* [7] was used. For both reference types 100 documents from 2019 were randomly selected. and cleaned from their HTML encoding. The two data sets were then manually annotated using Docanno[8]. The resulting data sets include 3.018 law references and 1.297 citations of decisions.

The annotated data sets were converted into the CoNLL 2002 format with its BIO scheme. Each data set contains a single class - GS (*Gesetz*=law) and RS (*Rechtsprechung*=court decision) respectively. For the citations of court decisions an additional more detailed annotation was created. It includes subclasses for file reference number, date and print source, which are easily identifiable using regular expressions. Table 1 illustrates both annotations.

[7] https://static.openlegaldata.io/dumps/de/2020-12-10/cases.jsonl.gz
[8] https://github.com/doccano/doccano

Table 1. CoNLL 2002: Example of the BIO-schemes for references to court decisions

BIO	Detailed BIO	BIO	Detailed BIO
Beschluss B-RS	BGH, B-RS	StR I-RS	4 I-RS-AZ
vom I-RS	Urteil I-RS	124/16, I-RS	StR I-RS-AZ
23. I-RS	vom I-RS	JurionRS I-RS	421/00, I-RS-AZ
August I-RS	22. I-RS-DT	2016, I-RS	NJW I-RS-FS
2016 I-RS	Februar I-RS-DT	26140 I-RS	2001, I-RS-FS
- I-RS	2001 I-RS-DT	III. O	1874, I-RS-FS
2 I-RS	- I-RS	27 O	1876 I-RS-FS

Both types of references follow fairly specific rules, which make a pattern matching approach promising. In practice it is more challenging than anticipated to cover all the variations, e.g. the version of the text is quoted or not, variing levels of hierarchy of the referenced text and so on. Our attempt to define a rule set to extract law references using the language processing toolkit Spacy[9] resulted in 34 different patterns. Detecting citations of court decisions was also attempted using pattern matching. Parts of the references to court decisions (date of the decision, file reference number, print publication) follow clear syntactic patterns and can be reliably identified. However, not all citations include these features, since citations may be incomplete, referring to full citations mentioned earlier, or may be arranged differently. These are incompletely or not detected.

CRFs and BiLSTMs are machine learning approaches to perform NER. [22] has already shown that these approaches can solve the NER task studied in this paper. For the law reference extraction task, CRF and BiLSTM trained on the LER corpus [10] as used by [22] have been used. These results were compared to a CRF model trained on the data set described above. For the decision citation task, the pretrained BiLSTM model was compared to two CRF models trained on the data set described above. One was trained on the shorter BIO annotations and one on the more detailed BIO annotation.

Finally, a BERT-model was used for both tasks: The pretrained bert-base-german-cased model from huggingface[11] was fine-tuned on the data sets for both NER tasks as described in [6]. During fine-tuning the token representations generated by BERT are fed into an output layer for sequence tagging and the resulting network is trained. We used 20 epochs and a batch size of 16 for this process.

4. Results and Discussion

The results for both tasks are summarized in Table 2. A simple pattern matching approach achieves F1 of 83.24 % for the law reference detection task. Results using a pretrained CRF are in a similar range and a CRF trained on the training portion of the presented data set increases F1 to 89.44 %. This increase of performance can be attributed to better generalisation due to a better data set coverage. BiLSTM achieves worse results due to a high number of false positives. By far the best results are achieved using BERT (F1 score of 98.82 %).

[9] https://spacy.io/
[10] https://github.com/elenanereiss/Legal-Entity-Recognition
[11] https://huggingface.co/bert-base-german-cased

Table 2. Extraction Results of the compared approaches

Approach	Law References			References to Court Decisions		
	Precision	Recall	F1 Score	Precision	Recall	F1 Score
Pattern Matching	83.53 %	82.94 %	83.24 %	-	-	-
Pretrained CRF	84.88 %	78.78 %	81.72 %	-	-	-
CRF	89.94 %	88.90 %	89.44 %	-	-	-
CRF: BIO	-	-	-	86.29 %	77.54 %	81.68 %
CRF: Detailed BIO	-	-	-	81.36 %	72.73 %	76.80 %
Pretrained BiLSTM	24.7 %	76 %	37.34 %	73.97 %	73.09 %	74.24 %
Fine-tuned BERT	**98.53 %**	**99.11 %**	**98.82 %**	**96.92 %**	**98.26 %**	**97.58 %**

The extraction of decision citations using pattern matching did not prove promising and was not further pursued. The pretrained BiLSTM achieved acceptable results (F1 74.24 %). A CRF trained on the data set improved F1 to 81.68 %. The use of a more detailed annotation scheme did not yield improved performance. The best results were also obtained using a BERT. The model achieved a F1 score of 96.19 %.

The data set presented covers more variations, e.g. it includes references to European court decisions, but has a lower number of samples than the data set used for the pretrained models. Despite the low number of samples CRF achieved performance comparable or superior to a pattern matching approach. By far the best performance for both tasks was achieved by BERT, since transfer learning approaches are especially suited for problems with limited but diverse data.

5. Conclusion

Legal references are specialised entities relevant to the specific domain of legal texts. Standard NER approaches, such as the work of Glaser [19], focus on more common entities like Location, Organisation and so on. The results for these common entities are not comparable to the results of domain specific entities such as references to legal documents. Using specialized NER approaches has the potential to achieve more reliable solutions.

NER helps to semantically enrich legal documents. A reliable automation of this task enables more sophisticated information systems. The obtained results suggest that a reliable solution may be accomplished by extending existing data sets and considering transfer learning methods which have proven successful for similar tasks.

References

[1] Al Qundus J, Paschke A, Gupta S, Alzouby AM, Yousef M. Exploring the impact of short-text complexity and structure on its quality in social media. Journal of Enterprise Information Management. 2020.

[2] Hoppe T, Al Qundus J, Peikert S. Ontology-based Entity Recognition and Annotation. In: Proceedings of the Conference on Digital Curation Technologies (Qurator 2020); 2020. Available from: http://ceur-ws.org/Vol-2535/paper_4.pdf.

[3] Al Qundus J. Generating trust in collaborative annotation environments. In: Proceedings of the 12th international symposium on open collaboration companion; 2016. p. 1-4.

[4] Lafferty J, McCallum A, Pereira FC. Conditional random fields: Probabilistic models for segmenting and labeling sequence data; 2001. .

[5] Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural networks. 2005;18(5-6):602-10.

[6] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Bidirectional Encoder Representations from Transformers; 2016. .

[7] Oksanen A, Tamper M, Tuominen J, Mäkelä E, Hietanen A, Hyvönen E. Part III, Article No. 2. In: Perugunelli G, Faro S, editors. Semantic Finlex: Transforming, Publishing, and Using Finnish Legislation and Case Law As Linked Open Data on the Web. No. 317 in Frontiers in Artificial Intelligence and Applications. Netherlands: IOS PRESS; 2019. p. 212-28.

[8] Glaser I, Schamberger T, Matthes F. Anonymization of German Legal Court Rulings. In: Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law. New York, NY, USA: Association for Computing Machinery; 2021. p. 205–209. Available from: https://doi.org/10.1145/3462757.3466087.

[9] C Cardellino LAA M Teruel, Villata S. A Low-cost, High-coverage Legal Named Entity Recognizer, Classifier and Linker. In: Proc. of the 16th International Conference on Artificial Intelligence and Law (ICAIL-2017).; 2017. p. 9-18. Available from: https://hal.archives-ouvertes.fr/hal-01541446v1.

[10] Song F, De La Clergerie E. Clustering-based Automatic Construction of Legal Entity Knowledge Base from Contracts. In: Proceedings - 2020 IEEE International Conference on Big Data, Big Data 2020; 2020. p. 2149-52.

[11] Tamper M, Oksanen A, Tuominen J, Hietanen A, Hyvönen E. Automatic Annotation Service APPI: Named Entity Linking in Legal Domain. In: Harth A, Presutti V, Troncy R, Acosta M, Polleres A, Fernández JD, et al., editors. The Semantic Web: ESWC 2020 Satellite Events. Cham: Springer International Publishing; 2020. p. 208-13.

[12] Adedjouma M, Sabetzadeh M, Briand LC. Automated detection and resolution of legal cross references: Approach and a study of Luxembourg's legislation. In: 2014 IEEE 22nd International Requirements Engineering Conference (RE); 2014. p. 63-72.

[13] Shulayeva O, Siddharthan A, Wyner A. Recognizing cited facts and principles in legal judgements. Artificial Intelligence and Law. 2017;25(1):107-26. Cited By :31.

[14] Resck Domingues LE, Ponciano JR, Nonato LG, Poco J. LegalVis: Exploring and Inferring Precedent Citations in Legal Documents. IEEE Transactions on Visualization and Computer Graphics. 2022.

[15] Correia FA, Almeida AAA, Nunes JL, Santos KG, Hartmann IA, Silva FA, et al. Fine-grained legal entity annotation: A case study on the Brazilian Supreme Court. Information Processing and Management. 2022;59(1).

[16] Sadeghian A, Sundaram L, Wang DZ, Hamilton WF, Branting K, Pfeifer C. Automatic Semantic Edge Labeling over Legal Citation Graphs. Artif Intell Law. 2018 jun;26(2):127–144. Available from: https://doi.org/10.1007/s10506-018-9217-1.

[17] Galgani F, Compton P, Hoffmann A. Citation Based Summarisation of Legal Texts. In: Anthony P, Ishizuka M, Lukose D, editors. PRICAI 2012: Trends in Artificial Intelligence. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012. p. 40-52.

[18] Glaser I, Scepankova E, Matthes F. Classifying semantic types of legal sentences: Portability of machine learning models. In: Legal Knowledge and Information Systems. IOS Press; 2018. p. 61-70.

[19] Glaser I, Waltl B, Matthes F. Named entity recognition, extraction, and linking in German legal contracts. In: IRIS: Internationales Rechtsinformatik Symposium; 2018. p. 325-34.

[20] Zöllner J, Sperfeld K, Wick C, Labahn R. Optimizing small BERTs trained for German NER. arXiv preprint arXiv:210411559. 2021.

[21] Donnelly J, Roegiest A. The Utility of Context When Extracting Entities From Legal Documents. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management. CIKM '20. New York, NY, USA: Association for Computing Machinery; 2020. p. 2397–2404. Available from: https://doi.org/10.1145/3340531.3412746.

[22] Leitner E, Rehm G, Moreno-Schneider J. Fine-grained Named Entity Recognition in Legal Documents. In: Acosta M, Cudré-Mauroux P, Maleshkova M, Pellegrini T, Sack H, Sure-Vetter Y, editors. Proceedings of the 15th International Conference on International Conference on Semantic Systems. No. 11702 in Lecture Notes in Computer Science. Karlsruhe, Germany: Springer; 2019. p. 272-87. 10/11 September 2019.