Fine-Grained Named Entity Recognition in Legal Documents

Elena Leitner, Georg Rehm(✉), and Julian Moreno-Schneider

DFKI GmbH, Alt-Moabit 91c, 10559 Berlin, Germany
{elena.leitner,georg.rehm,julian.moreno_schneider}@dfki.de

Abstract. This paper describes an approach at Named Entity Recognition (NER) in German language documents from the legal domain. For this purpose, a dataset consisting of German court decisions was developed. The source texts were manually annotated with 19 semantic classes: *person*, *judge*, *lawyer*, *country*, *city*, *street*, *landscape*, *organization*, *company*, *institution*, *court*, *brand*, *law*, *ordinance*, *European legal norm*, *regulation*, *contract*, *court decision*, and *legal literature*. The dataset consists of approx. 67,000 sentences and contains 54,000 annotated entities. The 19 fine-grained classes were automatically generalised to seven more coarse-grained classes (*person*, *location*, *organization*, *legal norm*, *case-by-case regulation*, *court decision*, and *legal literature*). Thus, the dataset includes two annotation variants, i.e., coarse- and fine-grained. For the task of NER, Conditional Random Fields (CRFs) and bidirectional Long-Short Term Memory Networks (BiLSTMs) were applied to the dataset as state of the art models. Three different models were developed for each of these two model families and tested with the coarse- and fine-grained annotations. The BiLSTM models achieve the best performance with an 95.46 F_1 score for the fine-grained classes and 95.95 for the coarse-grained ones. The CRF models reach a maximum of 93.23 for the fine-grained classes and 93.22 for the coarse-grained ones. The work presented in this paper was carried out under the umbrella of the European project LYNX that develops a semantic platform that enables the development of various document processing and analysis applications for the legal domain.

Keywords: Language technology · LT · Natural Language Processing · NLP · Named Entity Recognition · NER · Legal processing · Curation technologies · Legal technologies · BiLSTM · CRF

1 Introduction

Named Entity Recognition (NER) is the automatic identification of named entities (NEs) in texts, typically including their assignment to a set of semantic categories [19]. The established classes (for newspaper texts) are *person* `PER`, *location* `LOC`, *organization* `ORG` and *other* `OTH` [3,36,37]. Research on NER has a history

M. Acosta et al. (Eds.): SEMANTiCS 2019, LNCS 11702, pp. 272–287, 2019.
https://doi.org/10.1007/978-3-030-33220-4_20

of more than 20 years and produced approaches based on linear statistical models, e.g., Maximum Entropy Models [1,10], Hidden Markov Models [27], among others. Nowadays, the state of the art results are produced by methods such as CRFs [2,4,16,17] and BiLSTMs [9,20,22,26]. For English news documents, the best models have a performance of approx. 90 F_1 [9,20,22,26,29,38], while the best models for German are not quite as good with approx. 80 F_1 [2,4,16,22]. Based on their very good performance on news documents, we examine the use of CRFs and BiLSTMs in legal documents.

1.1 Application and Project Context

The objective of the project LYNX (Building the Legal Knowledge Graph for Smart Compliance Services in Multilingual Europe), a three year EU project that started in December 2017, is the creation of a legal knowledge graph that contains different types of legal and regulatory data.[1] LYNX aims to help European companies, especially SMEs, that already operate internationally, facing to offer and to promote their products and services in other countries. The project will eventually offer compliance-related services that are currently tested and validated in three use cases. The first pilot is a legal compliance solution, where documents related to data protection are innovatively managed, analysed, and visualised across different jurisdictions. In the second pilot, LYNX supports the understanding of regulatory regimes, including norms and standards, related to energy operations. The third pilot is a compliance solution in the domain of labour law, where legal provisions, case law, administrative resolutions, and expert literature are interlinked, analysed, and compared to define legal strategies for legal practice. The LYNX services are developed for several European languages including English, Spanish and German [32].

Documents in the legal domain contain multiple references to NEs, especially NEs specific to the legal domain, i.e., jurisdictions, legal institutions, etc. Most NER solutions operate in the general or news domain, which makes them not completely suitable for the analysis of legal documents, because they are unable to detect domain-specific entities. The goal is to make knowledge workers, who process and make use of these documents, more efficient and more effective in their day to day work, this also includes the analysis of domain-specific NEs, see [5,31] for related approaches in the area of content curation technologies.

1.2 Research Questions

This article is dedicated to the recognition of NERs and their respective categories in German legal documents. Legal language is unique and differs greatly from newspaper language. This also relates to the use of *person*, *location* and *organization* NEs in legal text, which are relatively rare. It does contain such specific entities as designations of legal norms and references to other legal documents (laws, ordinances, regulations, decisions, etc.) that play an essential role.

[1] http://www.lynx-project.eu.

Despite the development of NER for other languages and domains, the legal domain has not been exhaustively addressed yet. This research also had to face the following two challenges. (1) There is no uniform typology of semantic concepts related to NEs in documents from the legal domain; correspondingly, uniform annotation guidelines for NEs in the legal domain do not exist either. (2) There are no freely available datasets consisting of documents from the legal domain, in which NEs have been annotated.

Thus, the research goal is to examine NER with a specific focus on German legal documents. This includes the elaboration of the corresponding *concepts*, the construction of a *dataset*, developing, evaluating and comparing state of the art *models for NER*. We address the following research questions:

1. Which state of the art approaches are in use for NER? Which approaches have been developed for NER in legal documents? Do these approaches correspond to the state of the art?
2. Which NE categories are typical for legal documents? Which classes are to be identified and classified? Which legal documents can be used for a dataset?
3. What performance do current models have? How are different categories recognized? Which categories are recognized better than others?

2 Related Work

NER in the legal domain, despite its high relevance, is not a well researched area. Existing approaches are inconsistent with regard to the applied methods, techniques, classifications and datasets, which makes it impossible to compare their results adequately. Nevertheless, the developed approaches make an important contribution and form the basis for further research.

The first work in which NER in the legal domain was explicitly defined as a term was described by Dozier et al. [13]. The authors examined NER in US case law, depositions, pleadings and other legal documents, implemented using simple lookups in a list of NEs, contextual rules, and statistical models. Taggers were developed for *jurisdiction*, *court*, *title*, *document type* (e.g., brief, memorandum), and *judge*. The *jurisdiction* tagger performed best with an F_1 of 92. The scores of the other taggers were around 82–85.

Cardellino et al. developed a tool for recognizing, classifying, and linking legal NEs [8]. It uses the YAGO and LKIF ontologies and elaborated four different levels of granularity: NER, NERC, LKIF and YAGO. A Support Vector Machine, Stanford NER [17] and a neural network (NN) were trained and evaluated on Wikipedia and decisions of the European Court of Human Rights. The best result on the Wikipedia dataset was achieved by the NN with F_1 scores for the NERC and YAGO classes of 86 and 69, respectively. For the LKIF classes, Stanford NER was better with F_1 score of 77. The performance was significantly worse on decisions. The F_1 scores varied according to the model and the level of granularity. Stanford NER was able to achieve a maximum F_1 score of 56 with the NERC classes.

Glaser et al. tested three NER systems [18]. The first, GermaNER [4], recognized *person*, *location*, *organization* and *other*. Temporal and numerical expressions were recognized using rule-based approaches, and references using the approach described in Landthaler et al. [23]. The second system was DBpedia Spotlight [11,28], developed for the automatic annotation of DBpedia entities. The third system, Templated, was designed by Glaser et al. [18]. It focused on NER in contracts created using templates. For GermaNER and DBpedia Spotlight a manually annotated corpus was created, which consisted of 500 decisions of the 8th Civil Senate of the German Federal Court of Justice and had reference to tenancy law. GermaNER and DBpedia-Spotlight were evaluated on 20 decisions from the created dataset and Templated was evaluated on five different contracts. GermaNER and DBpedia Spotlight achieved an F_1 of 80 and 87, respectively. The result of Templated NER was 92 F_1.

To adapt categories for the legal domain, the set of NE classes was redefined in the approaches described above. Thus, Dozier et al. [13] focused on legal NEs (e.g., *judge*, *lawyer*, *court*). Cardellino et al. [8] extended NEs on NERC level to *document*, *abstraction*, and *act*. It is unclear what belongs to these classes and how they were separated from each other. Glaser et al. [18] added *reference* [23]. However, this was understood as a reference to legal norms, so that further references (to decisions, regulations, legal literature, etc.) were not covered.

The research of NER in legal documents is also complicated by the fact that there are no freely available datasets, neither for English nor for German. Datasets for newspaper texts, which were developed in CoNNL 2003 or GermEval 2014, again are not suitable in terms of the type of text and the annotated entities. In this context, the need for a manually annotated dataset consisting of legal texts is enormous, requiring the development of a classification of legal categories and uniform annotation guidelines. Such a dataset consisting of documents from the legal domain would make it possible to implement NER with state of the art architectures, i.e., CRF and BiLSTM, and to analyze their performance.

3 A Dataset of Documents from the Legal Domain

3.1 Semantic Categories

Legal documents differ from texts in other domains, and from each other in terms of text-internal, and text-external criteria [7,12,15,21], which has a huge impact on linguistic and thematic design, citation, structure, etc. This also applies to NEs used in legal documents. In law texts and administrative regulations, the occurrence of typical NEs such as *person*, *location* and *organization* is very low. Court decisions, on the other hand, include these NEs, and references to national or supranational laws, other decisions, and regulations. Two requirements for a typology of legal NEs emerge from these peculiarities. First, the categories used must reflect those entities that are typical for decisions. Second, a typology must concern the entities whose differentiation in decisions is highly relevant.

Domain-specific NEs in legal documents can be divided into two basic groups, namely designations and references. For legal norms (i.e., for laws and ordinances) designations are headings for their standard legal texts, which provide information on rank and content [6, Rn. 321 ff.]. Headings are uniform and usually consist of a long title, short title and abbreviation, e.g., the title of the Medicinal Products Act of 12 December 2005 ‘Gesetz über den Verkehr mit Arzneimitteln (Arzneimittelgesetz – AMG)’ (Federal Law Gazette I p. 3394). The short title ‘Arzneimittelgesetz’ and the abbreviation ‘AMG’ are in brackets. The citation of the legal norms is also fixed. There are different citation rules for full and short citations [6, Rn. 168 ff.]. The designation and citation of binding individual acts such as regulations or contracts is not uniformly defined.

For our dataset consisting of court decisions, a total of 19 fine-grained classes were developed, which are based on seven coarse-grained classes (see Table 1). As a starting point, the well-researched newspaper domain was used for the elaboration of the typology. The annotation guidelines are based on the ACE guidelines [25] and NoSta-D Named-Entity [3]. The core NEs are typical classes like `PER`, `LOC`, and `ORG`, which are split into fine-grained classes.[2] The coarse- and fine-grained classifications correlate such that, e.g., the coarse-grained class of *person* `PER` under number 1 in Table 1 contains the fine-grained classes of *judge* `RR`, *lawyer* `AN` and other *person* `PER` (plaintiffs, defendants, witnesses, appraisers, etc.) under numbers 1 to 3. The *location* `LOC` includes the fine-grained classes of *country* `LD` (countries, states and city-states), *city* `ST` (cities, villages and communities), *street* `STR` (streets, squares, avenues, municipalities and attractions) and *landscape* `LDS` (continents, mountains, lakes, rivers and other geographical units). The coarse-grained class *organization* `ORG` is divided into public/social, state and economic institutions. They form the fine-grained classes of *organization* `ORG`, *institution* `INN`, and *company* `UN`. Designations of the federal, supreme, provincial and local courts are summarized in the fine-grained class *court* `GRT`. Furthermore, *brand*[3] `MRK` is a separate category.

A fundamental peculiarity of the published decisions is that all personal information is anonymised on account of data privacy reasons. This applies primarily to *person*, *location* and *organization*. NEs are replaced by letters (1) or dots (2).

(1) . . . das Land B. **LD** . . .

(2) . . . unter der Firma C . . . AG **UN** . . .

In addition to the typical categories, other classes specific to legal documents, i.e., court decisions, are also included in the categories. These are the coarse-grained classes of *legal norm* `NRM`, *case-by-case regulation* `REG`, *court decision* `RS`

[2] The coarse- and fine-grained classes `PER` and `ORG` are different despite their identical abbreviations.

[3] From an onomastical point of view, *brand* belongs to object NEs which also contain the coarse-grained class of *organization*. Despite terminological and typological inaccuracy, *brand* was intentionally categorized as a fine-grained class of *organization* and not as independent coarse-grained class (see Table 1).

and *legal literature* `LIT`. The *legal norm* and *case-by-case regulation* include NEs (3) and references (4), but the *court decision* and *legal literature* only references (5). *Legal norm* `NRM` is subdivided according to legal force into the fine-grained classes *law* `GS`, *ordinance* `VO` and *European legal norm* `EUN`. *Case-by-case regulation* `REG`, on the other hand, contains binding individual acts that are below each legal standard. These include the fine-grained classes *regulation* `VS` (administrative regulations, directives, circulars and decrees) and *contract* `VT` (public service contracts, international treaties, collective agreements, etc.). The last two coarse-grained classes, *court decision* `RS` and *legal literature* `LIT`, do not have any fine-grained classes. `RS` reflects references to decisions, and `LIT` summarizes references to legal commentaries, legislative materials, legal textbooks and monographs.

(3) ... ist nach Maßgabe der Gründe mit dem Grundgesetz **GS** vereinbar.

(4) Mit der Neuregelung in § 35 Abs. 6 StVO **VO** ...

(5) ... Klein, in: Maunz/ Schmidt-Bleibtreu/ Klein/ Bethge, BVerfGG, § 19 Rn. 9 **LIT** ...

3.2 Dataset Statistics and Distribution of Semantic Categories

The dataset Legal Entity Recognition (LER) consists of 750 German court decisions published online in the portal 'Rechtsprechung im Internet'.[4] The source texts were extracted from the XML documents and split into sentences and words by SoMaJo [30]. The annotation was performed manually by one Computational Linguistics student using WebAnno [14]. In terms of future work we plan to add annotations from two to three linguists so that we can report inter-annotator agreement. The dataset[5] is freely available for download under the CC-BY 4.0 license[6], in CoNLL-2002 format. Each line consists of two columns separated by a space. The first column contains a token and the second a tag in IOB2 format. The sentence boundary is marked with an empty line.

The dataset consists of 66,723 sentences and 2,157,048 tokens. The percentage of annotations (per-token basis) is approx. 19%. Overall, the dataset includes 53,632 annotated NEs. The dataset has two variants for the classification of legal NEs (Table 1). The *person*, *location* and *organization* make up 25.66% of all annotated instances. 74.34% are specific categories like the *legal norm* `NRM`, *case-by-case regulation* `REG`, *court decision* `RS` and *legal literature* `LIT`. The largest classes are the *law* `GS` (34.53%) and *court decision* `RS` (23.46%). Other entities, i.e., *ordinance*, *European legal norm*, *regulation*, *contract*, and *legal literature*, are less common (between 1 and 6% of all annotations).

[4] http://www.rechtsprechung-im-internet.de.
[5] https://github.com/elenanereiss/Legal-Entity-Recognition.
[6] https://creativecommons.org/licenses/by/4.0/deed.en.

Table 1. Distribution of coarse- and fine-grained classes in the dataset

Coarse-grained classes			#	%	Fine-grained classes			#	%
1	**PER**	Person	3,377	6.30	1	**PER**	Person	1,747	3.26
					2	**RR**	Judge	1,519	2.83
					3	**AN**	Lawyer	111	0.21
2	**LOC**	Location	2,468	4.60	4	**LD**	Country	1,429	2.66
					5	**ST**	City	705	1.31
					6	**STR**	Street	136	0.25
					7	**LDS**	Landscape	198	0.37
3	**ORG**	Organization	7,915	14.76	8	**ORG**	Organization	1,166	2.17
					9	**UN**	Company	1,058	1.97
					10	**INN**	Institution	2,196	4.09
					11	**GRT**	Court	3,212	5.99
					12	**MRK**	Brand	283	0.53
4	**NRM**	Legal norm	20,816	38.81	13	**GS**	Law	18,520	34.53
					14	**VO**	Ordinance	797	1.49
					15	**EUN**	European legal norm	1,499	2.79
5	**REG**	Case-by-case regulation	3,470	6.47	16	**VS**	Regulation	607	1.13
					17	**VT**	Contract	2,863	5.34
6	**RS**	Court decision	12,580	23.46	18	**RS**	Court decision	12,580	23.46
7	**LIT**	Legal literature	3,006	5.60	19	**LIT**	Legal literature	3,006	5.60
		Total	53,632	100			**Total**	53,632	100

4 Evaluation and Results

We used two tools for sequence labeling for our experiments: sklearn-crfsuite[7] and UKPLab-BiLSTM [35]. In total, 12 models were tested, i.e., three CRF and BiLSTM models with coarse- and fine-grained classes. For CRFs, the following groups of features and sources were selected and manually developed:

1. F: features for the current word in a context window between −2 and +2, which are case and shape features, prefixes, and suffixes;
2. G: for the current word, gazetteers of persons from Benikova et al. [4]; gazetteers of countries, cities, streets, landscapes, and companies from GOVDATA[8], the Federal Agency for Cartography and Geodesy[9] and Datendieter.de[10]; gazetteers of laws, ordinances and administrative regulations from the Federal Ministry of Justice and Consumer Protection[11,12]. A detailed description of the gazetteers can be found in the Github project;
3. L: lookup table for the word similarity in a context window between -2 and +2 as in Benikova et al. [4], which contains the four most similar words to the current word.

[7] https://sklearn-crfsuite.readthedocs.io.

[8] https://www.govdata.de/apps/-/details/liste-der-staatennamen.

[9] https://www.bkg.bund.de/DE/Produkte-und-Services/Shop-und-Downloads/Digitale-Geodaten/Geographische-Namen/geographische-namen.html.

[10] https://www.datendieter.de.

[11] https://www.gesetze-im-internet.de.

[12] http://www.verwaltungsvorschriften-im-internet.de.

Three models were designed to chain these groups of features and gazetteers: (1) CRF-F with features; (2) CRF-FG with features and gazetteers; and (3) CRF-FGL with features, gazetteers, and the lookup table; the model names reflect the three groups. As a learning algorithm, the L-BFGS method is used with L1 and L2 regularization parameters, set to the coefficient 0.1. The maximum number of iterations for optimizing the algorithm is set to 100.

For BiLSTM we also use three models: (1) BiLSTM-CRF [20]; (2) BiLSTM-CRF+ with character embeddings from the BiLSTM [22]; (3) BiLSTM-CNN-CRF with character embeddings from CNN [26]. As hyperparameters we used the values that achieved the best NER performance according to Reimers and Gurevych [34]. The BiLSTM models have two BiLSTM layers, each with a size of 100 units and a dropout of 0.25. The maximum number of epochs is 100. At the same time, the tool uses pre-trained word embeddings for German [33].

The results were measured with the micro-precision, -recall and -F$_1$ measures. In order to reliably estimate their performance, we evaluated the models using stratified 10-fold cross-validation. The dataset is shuffled, sentence-wise, and divided into ten mutually exclusive partial sets of similar size. One iteration uses one set for validation and the rest for training. We iterate ten times, so that each part of the dataset is used nine times for training and once for validation. The distribution of NEs in the training and validation set remain the same over the iterations. The cross-validation prevented overfitting during training and the stratification prevented measurement errors in unbalanced data.

4.1 CRF Models

For the fine-grained classes, CRF-FGL achieved the best performance with an F$_1$ score of 93.23 (Table 2). The recognition of legal NEs in the different classes had varied levels of success depending on the model. *Lawyer*, *institution*, *court*, *contract* and *court decision* reached the highest F$_1$ with CRF-F. With the CRF-FG better results could be achieved for *judge*, *city*, *regulation* and *legal literature*. This means that the gazetteers have had a positive impact on the recognition of these NEs. The remaining classes performed better with CRF-FGL. The concatenation of gazetteers and the lookup table for the word similarity has improved the results, but not as much as expected.

For the coarse-grained classes, the CRF-FG and CRF-FGL together achieved the best result with an F$_1$ value of 93.22 (Table 3). However, *person* was recognized better with CRF-FG and *location* and *organization* better with CRF-FGL. CRF-FG achieved the best result in the *case-by-case regulation* and *court decision*. With CRF-FGL, the values in the *legal norm* and *legal literature* increased. Compared to the fine-grained classes, the better balanced precision and recall were observed and the F$_1$ increased by max. 0.1 per model.

4.2 BiLSTM Models

For the fine-grained classes, two models with character embeddings have achieved the best result with an F$_1$ score of 95.46 (Table 4), confirming the positive impact

Table 2. Precision, recall and F_1 values of CRF models for fine-grained classes

Fine-grained classes	CRF-F			CRF-FG			CRF-FGL		
	Prec	Rec	F_1	Prec	Rec	F_1	Prec	Rec	F_1
Person	89.41	83.53	86.32	90.50	83.54	86.83	90.44	84.22	**87.18**
Judge	98.22	97.62	97.92	98.68	97.75	**98.21**	98.55	97.75	98.14
Lawyer	93.14	76.84	**83.73**	89.81	73.51	80.39	92.17	75.04	81.99
Country	96.73	90.42	93.44	97.03	91.98	94.40	96.93	92.62	**94.70**
City	88.99	77.37	82.70	88.27	81.77	**84.77**	88.09	81.82	84.67
Street	88.69	59.58	70.51	87.51	57.95	68.90	90.50	59.85	**71.30**
Landscape	94.34	61.14	73.43	92.63	64.09	75.25	93.33	65.27	**76.08**
Organization	86.82	71.25	78.20	86.71	71.95	78.56	88.84	72.72	**79.89**
Company	92.77	86.04	89.21	93.00	86.18	89.39	93.54	86.85	**90.01**
Institution	92.74	89.49	**91.07**	92.88	89.20	90.98	92.51	89.47	90.96
Court	97.23	96.35	**96.78**	97.03	96.35	96.69	97.19	96.33	96.75
Brand	85.85	56.91	67.85	90.33	56.20	68.82	88.40	58.07	**69.61**
Law	96.86	96.34	96.60	97.00	96.44	96.72	97.02	96.56	**96.79**
Ordinance	91.91	82.23	86.79	91.35	82.85	86.87	91.41	83.49	**87.26**
European legal norm	89.37	86.07	87.67	88.91	85.49	87.14	89.41	86.21	**87.76**
Regulation	83.83	71.38	77.00	84.34	71.03	**77.02**	84.42	70.66	76.85
Contract	90.66	87.72	**89.15**	90.18	87.42	88.76	90.53	87.67	89.06
Court decision	93.35	93.39	**93.37**	93.22	93.34	93.28	93.21	93.29	93.25
Legal literature	92.98	91.28	92.12	92.94	91.42	**92.17**	92.79	91.28	92.02
Total	94.28	91.85	93.05	94.31	91.96	93.12	94.37	92.12	**93.23**

Table 3. Precision, recall and F_1 values of CRF models for coarse-grained classes

Coarse-grained classes	CRF-F			CRF-FG			CRF-FGL		
	Prec	Rec	F_1	Prec	Rec	F_1	Prec	Rec	F_1
Person	94.20	89.43	91.74	94.54	89.99	**92.20**	94.22	90.20	92.16
Location	94.60	84.55	89.26	93.89	85.48	89.45	94.33	86.45	**90.18**
Organization	92.82	89.00	90.87	93.02	89.08	90.99	93.23	89.10	**91.11**
Legal norm	96.19	95.16	95.67	96.29	95.26	95.77	96.28	95.44	**95.86**
Case-by-case regulation	89.29	84.72	86.94	89.28	84.77	**86.96**	88.76	84.15	86.39
Court decision	93.19	93.26	93.23	93.28	93.23	**93.25**	93.08	93.08	93.08
Legal literature	92.72	91.15	91.92	92.99	91.14	92.06	93.11	91.13	**92.11**
Total	94.17	92.07	93.11	94.26	92.20	**93.22**	94.22	92.25	**93.22**

of character level information. A significant improvement with an increase in F_1 by 5–16 (compared to the BiLSTM-CRF without character embeddings) was found in *organization*, *company*, *ordinance*, *regulation* and *contract*. *Judge* and *lawyer* were recognized better by about 1 with the BiLSTM-CRF. *Person*, *country*, *city*, *court*, *brand*, *law*, *ordinance*, *European legal norm*, *regulation* and

Table 4. Precision, recall and F_1 values of BiLSTM models for fine-grained classes

Coarse-grained classes	BiLSTM-CRF			BiLSTM-CRF+			BiLSTM-CNN-CRF		
	Prec	Rec	F_1	Prec	Rec	F_1	Prec	Rec	F_1
Person	89.30	91.08	90.09	90.78	92.24	**91.45**	90.21	92.57	91.35
Judge	98.64	99.48	**99.05**	98.37	99.21	98.78	98.18	99.01	98.59
Lawyer	94.85	84.62	**88.19**	86.18	90.59	87.07	88.02	87.96	87.11
Country	94.66	95.98	95.29	96.52	96.81	**96.66**	95.09	97.20	96.12
City	81.26	86.32	83.48	82.58	89.06	**85.60**	83.21	87.95	85.38
Street	81.70	75.94	78.10	81.82	75.78	77.91	86.24	78.21	**81.49**
Landscape	78.54	79.08	77.57	78.50	80.20	78.25	80.93	81.80	**80.90**
Organization	79.50	74.72	76.89	82.70	80.18	81.28	84.32	81.00	**82.51**
Company	85.81	81.34	83.44	90.05	88.11	89.04	91.72	89.18	**90.39**
Institution	88.88	90.91	89.85	89.99	92.40	91.17	90.24	92.23	**91.20**
Court	97.49	98.33	97.90	97.72	98.24	**97.98**	97.52	98.34	97.92
Brand	78.34	73.11	75.17	83.04	76.25	**79.17**	83.48	73.62	77.79
Law	96.59	97.01	96.80	98.34	98.51	**98.42**	98.44	98.38	98.41
Ordinance	82.63	72.61	77.08	92.29	92.96	**92.58**	91.00	91.09	90.98
European legal norm	90.62	89.79	90.18	92.16	92.63	**92.37**	91.58	92.29	91.92
Regulation	75.58	68.91	71.77	85.14	78.87	**81.63**	79.43	78.30	78.74
Contract	87.12	85.86	86.48	92.00	92.64	**92.31**	90.78	92.06	91.40
Court decision	96.34	96.47	96.41	96.70	96.73	96.71	97.04	97.06	**97.05**
Legal literature	93.87	93.68	93.77	94.34	93.94	94.14	94.25	94.22	**94.23**
Total	93.80	93.70	93.75	95.36	95.57	**95.46**	95.34	95.58	**95.46**

contract were identified better with the BiLSTM-CRF+, and *street*, *landscape*, *organization*, *company*, *institution*, *court decision* and *legal literature* with the BiLSTM-CNN-CRF. Dependencies of the results on character embeddings produced by BiLSTM and CNN were also found. *Brand*, *ordinance* and *regulation* benefited significantly from the use of the BiLSTM. However, recognition of *street* and *landscape* improved with the character embeddings from the CNN.

For the coarse-grained classes, F_1 increased by 0.3−0.9 per model, and precision and recall were also more balanced (Table 5). The best result was produced by the BiLSTM-CRF+ with 95.95. The model had the highest values of more than 90 F_1 in almost all classes. An exception was the BiLSTM-CNN-CRF in *organization*, which increased F_1 by 0.3.

Table 5. Precision, recall and F_1 values of BiLSTM models for coarse-grained classes

Coarse-grained classes	BiLSTM-CRF			BiLSTM-CRF+			BiLSTM-CNN-CRF		
	Prec	Rec	F_1	Prec	Rec	F_1	Prec	Rec	F_1
Person	94.34	95.16	94.74	94.82	96.03	**95.41**	94.09	96.21	95.12
Location	90.85	92.59	91.68	92.60	94.05	**93.31**	91.74	93.45	92.57
Organization	91.82	90.94	91.37	92.87	92.89	92.87	93.80	92.65	**93.21**
Legal norm	97.04	96.50	96.77	97.93	98.04	**97.98**	97.71	97.87	97.79
Case-by-case regulation	86.79	84.15	85.43	90.72	90.53	**90.61**	90.11	90.80	90.43
Court decision	96.54	96.58	96.56	96.93	97.05	**96.99**	96.73	96.83	96.78
Legal literature	93.78	93.91	93.84	94.23	94.62	**94.42**	94.24	93.80	94.02
Total	94.86	94.49	94.68	95.84	96.07	**95.95**	95.71	95.87	95.79

4.3 Discussion

The BiLSTMs achieved superior performance compared to the CRFs. They produced good results even with the fine-grained classes covered poorly in the dataset. The CRF models, on the other hand, delivered values that were about 1–10 lower per class. In addition, some classes are characterized by bigger differences in precision and recall, indicating certain weaknesses of the CRFs. In particular, the recognition of *street* and *brand* with the BiLSTM models improved by values of at least 10. The values for *lawyer*, *landscape* and *ordinance* also increased by a value of 5.

The results also show that the two model families exhibit a similar performance due to the dataset or structure of the data. The models produce their best results with 95 F_1 score in the fine-grained classes *judge*, *court* and *law*. On the one hand, this depends on a smaller number of types compared to tokens in *judge* and *court*. On the other hand, the precise identification of *law* can be explained by its good coverage in the dataset and uniform citation. Incorrect predictions about boundaries are made if references had a different form such as in '§ 7 des Gesetzes (gemeint ist das VersAnstG)' instead of common '§ 7 VersAnstG', 'das zwölfte Kapitel des neunten Sozialgesetzbuches' instead of 'das Kapitel 12 des SGB XII'. There were also incorrect classifications of terms as a NE containing the word 'law', such as 'federal law', 'law of experience', 'criminal law', etc. The recognition of *country*, *institution*, *court decision*, and *legal literature* was also very good with scores higher than 90 F_1. This is also due to a smaller number of types in *country*, *institution* and uniform references of *court decision* and *legal literature*.

However, the recognition of *street*, *landscape*, *organization* and *regulation* is the lowest throughout, amounting to 69–80 with the CRF and 72–83 with the BiLSTM models, caused by inconsistent citation styles. The recognition of *street* and *landscape* is poor because they are covered in the dataset with only about 200 instances, but heterogeneously represented. The worst result, i.e., a maximum F_1 value of 69.61 with the CRFs and of 79.17 with the BiLSTMs, was observed in *brand*. These NEs were also expressed in different contexts, such as the *brand* NE 'Einstein's Garage' and the scientist Albert Einstein. It can be concluded that the differences in the recognition of certain NEs is firstly due to the unbalanced class distribution and secondly to the specifics of the legal documents, in particular because of the coverage in the corpus, the heterogeneity with regard to the form of names or references as well as the context.

Overall, the CRFs and BiLSTMs perform very well, producing state of the art results, which are significantly better than comparable models for newspaper text. This fact can, first, be explained by the size of the dataset which is larger than other NE datasets for German. Second, the form of legal NEs, which also includes references, differs a lot from NEs in newspaper text. The distribution of designations or references in the dataset consisting of documents from the legal domain is greater compared to *person*, *location* or *organization*. Third, the strictly regulated linguistic and thematic design (repeated use of NEs per one decision, repeated use of formulaic, template-like sentences, etc.) and the uniform reference style have had a positive impact on performance. The applied evaluation method made it possible to reliably estimate performance for unbalanced data. Unfortunately, it is not possible to compare our results with other systems for NER in legal documents because they are not freely available.

5 Conclusion

We describe and evaluate a set of approaches for the recognition of semantic concepts in German court decisions. In line with the goals, the characteristic and relevant semantic categories such as *legal norm*, *case-by-case regulation*, *court decision* and *legal literature* were worked out and a dataset of legal documents was built, instances of a total of 19 semantic classes were annotated. For the experiment, CRF and BiLSTM models were selected that correspond to the state of art, and tested with the two sets of classes. The results of both model families demonstrate the superiority of the BiLSTMs models with character embeddings with an F_1 score of 95.46 for the fine-grained classes and 95.95 for the coarse-grained classes. We found that the structure of the data involved in the training process strongly impacts the performance. To improve NER, it is necessary to extend or optimize the unbalanced data. This helps to minimize the specific influencing factors of the legal documents on models. Our results show that there is no universal model that recognizes all classes in the best way. Accordingly, an even better universal system could be built as an ensemble of different models that perform well for particular classes.

Acknowledgments. The project LYNX has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 780602. More information is available online at http://www.lynx-project.eu. This work was carried out under the umbrella of an internship of the first author at DFKI, which resulted in a Bachelor thesis at the University of Potsdam [24].

References

1. Bender, O., Och, F.J., Ney, H.: Maximum entropy models for named entity recognition. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, vol. 4, pp. 148–151. Association for Computational Linguistics (2003)
2. Benikova, D., Biemann, C., Kisselew, M., Padó, S.: GermEval 2014 named entity recognition shared task: companion paper. In: Proceedings of the KONVENS GermEval Workshop, Hildesheim, Germany, pp. 104–112 (2014)
3. Benikova, D., Biemann, C., Reznicek, M.: NoSta-D named entity annotation for German: guidelines and dataset. In: Calzolari, N., et al. (eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, 26–31 May 2014. pp. 2524–2531. European Language Resources Association (ELRA) (2014)
4. Benikova, D., Yimam, S.M., Santhanam, P., Biemann, C.: GermaNER: free open German named entity recognition tool. In: Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology, GSCL 2015, University of Duisburg-Essen, Germany, 30 September–2 October 2015, pp. 31–38 (2015)
5. Bourgonje, P., Moreno-Schneider, J., Rehm, G.: Domain-specific entity spotting: curation technologies for digital humanities and text analytics. In: Reiter, N., Kremer, G. (eds.) CUTE Workshop 2017 - CRETA Unshared Task zu Entitätenreferenzen. Workshop bei DHd 2017, Berne, Switzerland, February 2017
6. Bundesministerium der Justiz: Bekanntmachung des Handbuchs der Rechtsförmlichkeit. Bundesanzeiger Jahrgang **60**(160a), 296 (2008)
7. Busse, D.: Textsorten des Bereichs Rechtswesen und Justiz. Text- und Gesprächslinguistik. Ein internationales Handbuch zeitgenössischer Forschung **1**, 658–675 (2000)
8. Cardellino, C., Teruel, M., Alemany, L.A., Villata, S.: A low-cost, high-coverage legal named entity recognizer, classifier and linker. In: Proceedings of the 16th Edition of the International Conference on Artificial Intelligence and Law, ICAIL 2017, pp. 9–18. , ACM, New York (2017)
9. Chiu, J.P.C., Nichols, E.: Named entity recognition with bidirectional LSTM-CNNs. TACL **4**, 357–370 (2016)
10. Clark, A.: Combining distributional and morphological information for part of speech induction. In: Proceedings of the Tenth Conference on European chapter of the Association for Computational Linguistics, vol. 1, pp. 59–66. Association for Computational Linguistics (2003)
11. Daiber, J., Jakob, M., Hokamp, C., Mendes, P.N.: Improving efficiency and accuracy in multilingual entity extraction. In: I-SEMANTICS 2013–9th International Conference on Semantic Systems, ISEM 2013, Graz, Austria, 4–6 September 2013, pp. 121–124 (2013)
12. Deutsch, A.: 5. Schriftlichkeit im Recht: Kommunikationsformen/Textsorten. Handbuch Sprache im Recht **12**, 91–117 (2017)
13. Dozier, C., Kondadadi, R., Light, M., Vachher, A., Veeramachaneni, S., Wudali, R.: Named entity recognition and resolution in legal text. In: Francesconi, E., Montemagni, S., Peters, W., Tiscornia, D. (eds.) Semantic Processing of Legal Texts. LNCS (LNAI), vol. 6036, pp. 27–43. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-12837-0_2

14. Eckart de Castilho, R., et al.: A web-based tool for the integrated annotation of semantic and syntactic structures. In: Hinrichs, E.W., Hinrichs, M., Trippel, T. (eds.) Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities, LT4DH@COLING, Osaka, Japan, December 2016, pp. 76–84. The COLING 2016 Organizing Committee (2016)
15. Engberg, J.: Prinzipien einer Typologisierung juristischer Texte. Fachsprache Int. J. Spec. Commun. **15**(1/2), 31–38 (1993)
16. Faruqui, M., Padó, S.: Training and evaluating a german named entity recognizer with semantic generalization. In: Pinkal, M., Rehbein, I., im Walde, S.S., Storrer, A. (eds.) Semantic Approaches in Natural Language Processing: Proceedings of the 10th Conference on Natural Language Processing, KONVENS 2010, Saarland University, Saarbrücken, Germany, 6–8 September 2010, pp. 129–133. universaar, Universitätsverlag des Saarlandes/Saarland University Press/Presses universitaires de la Sarre (2010)
17. Finkel, J.R., Grenager, T., Manning, C.D.: Incorporating non-local information into information extraction systems by Gibbs sampling. In: Knight, K., Ng, H.T., Oflazer, K. (eds.) Proceedings of the Conference on ACL 2005, 43rd Annual Meeting of the Association for Computational Linguistics, 25–30 June 2005, University of Michigan, USA, pp. 363–370. The Association for Computer Linguistics (2005)
18. Glaser, I., Waltl, B., Matthes, F.: Named entity recognition, extraction, and linking in German legal contracts. In: IRIS: Internationales Rechtsinformatik Symposium, pp. 325–334 (2018)
19. Grishman, R., Sundheim, B.: Message understanding conference-6: a brief history. In: Proceedings of the 16th International Conference on Computational Linguistics, COLING 1996, Center for Sprogteknologi, Copenhagen, Denmark, 5–9 August 1996, pp. 466–471 (1996)
20. Huang, Z., Xu, W., Yu, K.: Bidirectional LSTM-CRF models for sequence tagging. CoRR abs/1508.01991 (2015)
21. Kjær, A.: Normbedingte Wortverbindungen in der juristischen Fachsprache (Deutsch als Fremdsprache). Fremdsprachen Lehren und Lernen **21**, 46–64 (1992)
22. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. In: NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, 12–17 June 2016, pp. 260–270 (2016)
23. Landthaler, J., Waltl, B., Matthes, F.: Unveiling references in legal texts - implicit versus explicit network structures. In: IRIS: Internationales Rechtsinformatik Symposium, pp. 71–78 (2016)
24. Leitner, E.: Eigennamen- und Zitaterkennung in Rechtstexten. Bachelor's thesis, Universität Potsdam, Potsdam, February 2019
25. Linguistic Data Consortium: ACE (Automatic Content Extraction) English Annotation Guidelines for Entities (2008)
26. Ma, X., Hovy, E.H.: End-to-end sequence labeling via bi-directional lstm-CNNs-CRF. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, Volume 1: Long Papers, Berlin, Germany, 7–12 August 2016 (2016)
27. Mayfield, J., McNamee, P., Piatko, C.: Named entity recognition using hundreds of thousands of features. In: Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003, vol. 4, pp. 184–187. Association for Computational Linguistics (2003)

28. Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: DBpedia spotlight: shedding light on the web of documents. In: Ghidini, C., Ngomo, A.N., Lindstaedt, S.N., Pellegrini, T. (eds.) Proceedings the 7th International Conference on Semantic Systems, I-SEMANTICS 2011, Graz, Austria, 7–9 September 2011. . ACM International Conference Proceeding Series, pp. 1–8. ACM (2011)
29. Passos, A., Kumar, V., McCallum, A.: Lexicon infused phrase embeddings for named entity resolution. In: Morante, R., Yih, W. (eds.) Proceedings of the Eighteenth Conference on Computational Natural Language Learning, CoNLL 2014, Baltimore, Maryland, USA, 26–27 June 2014, pp. 78–86. ACL (2014)
30. Proisl, T., Uhrig, P.: SoMaJo: State-of-the-art tokenization for German web and social media texts. In: Cook, P., Evert, S., Schäfer, R., Stemle, E. (eds.) Proceedings of the 10th Web as Corpus Workshop, WAC@ACL 2016, Berlin, 12 August 2016, pp. 57–62. Association for Computational Linguistics (2016)
31. Rehm, G., et al.: Event detection and semantic storytelling: generating a travelogue from a large collection of personal letters. In: Caselli, T., et al. (eds.) Proceedings of the Events and Stories in the News Workshop, co-located with ACL 2017, Vancouver, Canada, August 2017, pp. 42–51. Association for Computational Linguistics (2017)
32. Rehm, G., et al.: Developing and orchestrating a portfolio of natural legal language processing and document curation services. In: Aletras, N., et al. (eds.) Proceedings of Workshop on Natural Legal Language Processing (NLLP 2019), co-located with NAACL 2019, Minneapolis, USA, 7 June 2019, pp. 55–66 (2019)
33. Reimers, N., Eckle-Kohler, J., Schnober, C., Kim, J., Gurevych, I.: GermEval-2014: nested named entity recognition with neural networks. In: Faaß, G., Ruppenhofer, J. (eds.) Workshop Proceedings of the 12th Edition of the KONVENS Conference, Oktober 2014, pp. 117–120. Universitätsverlag Hildesheim (2014)
34. Reimers, N., Gurevych, I.: Optimal hyperparameters for deep LSTM-networks for sequence labeling tasks. CoRR abs/1707.06799 (2017)
35. Reimers, N., Gurevych, I.: Reporting score distributions makes a difference: performance study of LSTM-networks for sequence tagging. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 338–348. Association for Computational Linguistics (2017)
36. Sang, E.F.T.K., Meulder, F.D.: Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. In: Daelemans, W., Osborne, M. (eds.) Proceedings of the Seventh Conference on Natural Language Learning, CoNLL 2003, Held in cooperation with HLT-NAACL 2003, Edmonton, Canada, 31 May–1 June 2003, pp. 142–147. ACL (2003)
37. Tjong Kim Sang, E.F.: Introduction to the CoNLL-2002 shared task: language-independent named entity recognition. In: Proceedings of the 6th Conference on Natural Language Learning, COLING 2002, vol. 20, pp. 1–4. Association for Computational Linguistics, Stroudsburg (2002)
38. Tkachenko, M., Simanovsky, A.: Named entity recognition: exploring features. In: Jancsary, J. (ed.) 11th Conference on Natural Language Processing, KONVENS 2012, Empirical Methods in Natural Language Processing, Vienna, Austria, 19–21 September 2012. Scientific series of the ÖGAI, vol. 5, pp. 118–127. ÖGAI, Wien (2012)