A Dataset of German Legal Documents for Named Entity Recognition

Abstract

We describe a dataset developed for Named Entity Recognition in German federal court decisions. It consists of approx. 67,000 sentences with over 2 million tokens. The resource contains 54,000 manually annotated entities, mapped to 19 fine-grained 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 legal documents were, furthermore, automatically annotated with more than 35,000 TimeML-based time expressions. The dataset, which is available under a CC-BY 4.0 license in the CoNNL-2002 format, was developed for training an NER service for German legal documents in the EU project Lynx.

1.1em

:::

1.1.1em

::: :::

1.1.1.1em

same

Elena Leitner, Georg Rehm, Julián Moreno-Schneider

DFKI GmbH, Alt-Moabit 91c, 10559 Berlin, Germany

{firstname.lastname}@dfki.de

We describe a dataset developed for Named Entity Recognition in German federal court decisions. It consists of approx. 67,000 sentences with over 2 million tokens. The resource contains 54,000 manually annotated entities, mapped to 19 fine-grained 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 legal documents were, furthermore, automatically annotated with more than 35,000 TimeML-based time expressions. The dataset, which is available under a CC-BY 4.0 license in the CoNNL-2002 format, was developed for training an NER service for German legal documents in the EU project Lynx.

Named Entity Recognition, NER, Legal Documents, Legal Domain, Corpus Creation, Corpus Annotation

Introduction and Motivation

Just like any other field, the legal domain is facing multiple challenges in the era of digitisation. Document collections are growing at an enormous pace and their complete and deep analysis can only be tackled with the help of assisting technologies. This is where content curation technologies based on text analytics come in rehm2016j. Such domain-specific semantic technologies enable the fast and efficient automated processing of heterogeneous document collections, extracting important information units and metadata such as, among others, named entities, numeric expressions, concepts and topics, time expressions, and text structure. One of the fundamental processing tasks is the identification and

categorisation of named entities (Named Entity Recognition, NER). Typically, NER is focused upon the identification of semantic categories such as person, location and organization but, especially in domain-specific applications, other typologies have been developed that correspond to task-, language- or domain-specific needs. With regard to the legal domain, the lack of freely available datasets has been a stumbling block for text analytics research. German newspaper datasets from CoNNL 2003 BIBREF0 or GermEval 2014 BIBREF1 are simply not suitable in terms of domain, text type or semantic categories covered.

The work described in this paper was carried out under the umbrella of the project Lynx: Building the Legal Knowledge Graph for Smart Compliance Services in Multilingual Europe, a three-year EU-funded project that started in December 2017 BIBREF2. Its objective is the creation of a legal knowledge graph that contains different types of legal and regulatory data BIBREF3, BIBREF4, BIBREF5. Lynx aims to help European companies, especially SMEs, that want to become active in new European countries and markets. The project offers compliance-related services that are currently tested and validated in three use cases (UC): (i) UC1 aims to analyse contracts, enriching them with domain-specific semantic information (document structure, entities, temporal expressions, claims, summaries, etc.); (ii) UC2 focuses on compliance services related to geothermal energy operations, where Lynx supports the understanding of regulatory regimes, including norms and standards; (iii) UC3 is a compliance solution in the domain of labour law, where legal provisions, case law, 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 – relevant for this paper – German BIBREF6.

Documents in the legal domain contain multiple references to named entities, especially domain-specific named entities, i. e., jurisdictions, legal institutions, etc. Legal documents are unique and differ greatly from newspaper texts. On the one hand, the occurrence of general-domain named entities is relatively rare. On the other hand, in concrete applications, crucial domain-specific entities need to be identified in a

reliable way, such as designations of legal norms and references to other legal documents (laws, ordinances, regulations, decisions, etc.). However, most NER solutions operate in the general or news domain, which makes them inapplicable to the analysis of legal documents BIBREF7, BIBREF8. Accordingly, there is a great need for an NER-annotated dataset consisting of legal documents, including the corresponding development of a typology of semantic concepts and uniform annotation guidelines. In this paper, we describe the development of a dataset of legal documents, which includes (i) named entities and (ii) temporal expressions.

The remainder of this article is structured as follows. First, Section SECREF3 gives a brief overview of related work. Section SECREF4 describes, in detail, the rationale behind the annotation of the dataset including the different semantic classes annotated. Section SECREF5 describes several characteristics of the dataset, followed by a short evaluation (Section SECREF6) and conclusions as well as future work (Section SECREF7).

Related Work

Until now, NER has not received a lot of attention in the legal domain, developed approaches are fragmented and inconsistent with regard to their respective methods, datasets and typologies used. Among the related work, there is no agreement regarding the selection of relevant semantic categories from the legal domain. In addition, corpora or datasets of legal documents with annotated named entities do not appear to exist, which is, obviously, a stumbling block for the development of data-driven NER classifiers.

dozier2010named describe five classes for which taggers are developed based on dictionary lookup, pattern-based rules, and statistical models. These are jurisdiction (a geographic area with legal authority), court, title (of a document), doctype (category of a document), and judge. The taggers were tested with

documents such as US case law, depositions, pleadings etc. cardellino2017low develop an ontology of legal concepts, making use of NERC (6 classes), LKIF (69 classes) and YAGO (358 classes). On the NERC level, entities were divided in abstraction, act, document, organization, person, and non-entity. With regard to LKIF, company, corporation, contract, statute etc. are used. Unfortunately, the authors do not provide any details regarding the questions how the entities were categorised or if there is any correlations between the different levels. They work with Wikipedia articles and decisions of the European Court of Human Rights. glaser2017named use GermaNER BIBREF9 and DBpedia Spotlight BIBREF10, BIBREF11 for the recognition of person, location and organization entities. References are identified based on the rules described by landthaler2016unveiling. The authors created an evaluation dataset of 20 court decisions.

Annotation of the Dataset

In the following, we describe the rationale behind the annotation of the dataset including the definition of the various semantic classes and the annotation guidelines.

Annotation of the Dataset ::: Named Entities vs. Legal Entities ::: Named Entity

An entity is an object or set of objects in the real world and can be referenced in a text with a proper name, noun or pronoun BIBREF12. The examples (UNKREF6–UNKREF8) show corresponding sentences that contain the named mention `John', the nominal mention `the boy' and the pronominal mention `he'. This distinction between names on the one hand and pronominal or nominal mentions on the other can also be applied to the broad semantic set of named entities from the legal domain, see (UNKREF9–UNKREF11). Thus, (UNKREF6, UNKREF9) contain actual named entities.

John is 8 years old.

The boy is 8 years old.

He is 8 years old.

The BGB regulates the legal relations between private persons.

The law regulates the legal relations [...].

It regulates the legal relations [...].

Annotation of the Dataset ::: Named Entities vs. Legal Entities ::: Legal Entity

Basically, legal entities are either designations or references. A designation (or name) is the title of a legal document. In law texts, the title is strictly standardised and consists of a long title, short title and an abbreviation BIBREF13. The title of the Act on the Federal Constitutional Court is: `Gesetz über das Bundesverfassungsgericht (Bundesverfassungsgerichtsgesetz – BVerfGG)', where `Gesetz über das Bundesverfassungsgericht' is the long title, `Bundesverfassungsgerichtsgesetz' is the short title, and `BVerfGG' is the abbreviation. A reference to a legal norm is also fixed with rules for short and full references BIBREF13. Designations or references of binding individual acts such as regulations or contracts, however, are not uniformly defined.

Annotation of the Dataset ::: Named Entities vs. Legal Entities ::: Personal Data

A fundamental characteristic of the published decisions, that are the basis of our dataset, is that all personal information have been anonymised for privacy reasons. This affects the classes person, location and organization. Depending on the respective federal court, different rules were used for this

anonymisation process. Named entities were replaced by letters or abbreviated (UNKREF14), sometimes ellipsis were used (UNKREF15, UNKREF16). Some anonymised locations are mentioned with terms such as “street”, “place”, “avenue”, etc. that are part of this named entity (UNKREF16).

Fernsehmoderator G. PER

`television presenter G.'

Firma X... UN

`company X...'

in der A-Straße STR in ... ST

`in the A-Street in ...'

Annotation of the Dataset ::: Semantic Classes

We defined 19 fine-grained semantic classes. The (proto)typical classes are person, location and organization. In addition, we defined more specific semantic classes for the legal domain. These are the coarse-grained classes legal norm, case-by-case regulation, court decision and legal literature. The classes legal norm and case-by-case regulation include designations and references, while court decision and legal literature include only references.

In the process of developing the typology and annotation guidelines, the fine-grained classes continent KONT (which belongs to location), university UNI, institute IS and museum MUS (which belonged to

organization) were eliminated due their low frequency in the corpus (less than 50 occurrences). This is why university, institute and museum were subsumed under the fine-grained class organization. Continent was integrated into landscape.

The specification of the 19 fine-grained classes was motivated by the need for distinguishing entities in the legal domain. A first distinction was made between standards and binding acts. Standards, which belong to legal norm, are legal rules adopted by a legislative body in a legislative process. We can distinguish further between law, ordinance (German national standards) and European legal norm. Binding acts (circulars, administrative acts, contracts, administrative regulations, directives, etc.) belong to the category of case-by-case-regulation. It includes regulation (arrangements or instructions on subjects) and contract (agreements between subjects). In addition, court decision and legal literature, which are important in the decision making process, were put into their own categories.

Within person, we distinguish between judge and lawyer, key roles mentioned frequently in the decisions. Locations are categorised in terms of their size in country, city and street. Organizations are divided based on their role in the process, into public or social organization, state institution, (private) economic company, mostly as a legal entity, and court as an organ of jurisprudence.

Annotation of the Dataset ::: Semantic Classes ::: Person

The coarse-grained class person PER contains the fine-grained classes judge RR, lawyer AN and person PER (such as accused, plaintiff, defendant, witness, appraiser, expert, etc.), who are involved in a court process and mentioned in a decision. In example (UNKREF19), the same surname occurs twice in a sentence, one as judge and one as person.

Zwar ist Paul Kirchhof RR mit dem Vizepräsidenten Kirchhof PER als dessen Bruder in der Seitenlinie im

zweiten Grade verwandt...

`Although Paul Kirchhof is related to the Vice President Kirchhof as his brother in the second-degree sidelines...'

Annotation of the Dataset ::: Semantic Classes ::: Location

The coarse-grained class location LOC contains names of topographic objects, divided into country LD, city ST, street STR and landscape LDS. Country (UNKREF21) includes countries, states or city-states and city (UNKREF22) includes to cities, villages or communities. Street (UNKREF23) refers to avenues, squares, municipalities, attractions etc., i. e., named entities within a city or a village. Landscape (UNKREF24) includes continents, lakes, rivers and other geographical objects.

... hat bislang nur das Land Mecklenburg-Vor-ForestGreen!50l pommern LD Gebrauch gemacht.

`So far, only the state of Mecklenburg-Vorpommern has made use of it.'

Dem Haftbefehl liegt eine Entscheidung des Berufungsgerichts in Bukarest ST vom 18. Februar 2016 zugrunde ...

`The arrest warrant is based on a decision of the Appeal Court in Bucharest of 18 February 2016 ...'

Zwar legt der Bezug auf die Grenzwertüberschreitung 2015 insbesondere in der Cornelius-GreenYellowl straße STR ...

`Admittedly, the reference to the exceedance of the 2015 threshold applies in particular to

Corneliusstraße ...'

... aus der Region um den Fluss Main LDS stammen bzw. dort angeboten werden ...

`... come from the region around the river Main or are offered there...'

Annotation of the Dataset ::: Semantic Classes ::: Organization

The coarse-grained class organization ORG is divided into public/social, state and economic institutions. Social and public institutions such as parties, associations, centres, communities, unions, educational institutions or research institutions are grouped into the fine-grained class organization ORG (UNKREF26). Institution INN (UNKREF27) contain public administrations, including federal and state ministries and the constitutional bodies of the Federal Republic of Germany at the federal and state level, i. e., the Federal Government, the Federal Council, the Bundestag, the state parliaments and governments. Company UN (UNKREF28) includes commercial legal entities.

Der FC Bayern München ORG schloss den Beschwerdeführer ... aus dem Verein aus ...

`Bayern Munich closed the complainant ... from the club ....'

Die Landesregierung Rheinland-Pfalz INN hat von einer Stellungnahme abgesehen.

`The state government of Rhineland-Palatinate refrained from commenting.'

... eingeführte Smartphone-Modellreihe des US-amerikanischen Unternehmens Apple UN ...

`... introduced smartphone model series of the US company Apple ...'

Court designations play a central role in decisions, which is why they are collected in their own class court GRT. These are designations of federal, supreme, provincial and local courts. The designations of the courts at the country level are composed of the names of the ordinary jurisdiction and their location (UNKREF29). Furthermore, brands are often discussed in decisions of the Federal Patent Court. They are subsumed under brand MRK, which can be contextual and semantically ambiguous, such as `Becker' from (UNKREF30), which has evolved from a personal name.

Diesen Anspruch hat das LSG Mecklenburg-RubineRed!50l Vorpommern GRT mit Urteil vom 22.2.2017 verneint ...

`This claim was rejected by the LSG Mecklenburg-Vorpommern by judgment of 22.2.2017 ...'

Vorliegend stehen sich die Widerspruchsmarke Becker Mining MRK und die angegriffene Marke Becker MRK gegenüber.

`In the present case, the opposing brand Becker Mining and the challenged brand Becker face each other.'

Annotation of the Dataset ::: Semantic Classes ::: Legal Norms

Norms are divided according to their legal status into the fine-grained classes of law GS, ordinance VO and European legal norm EUN. Law is composed of the standards adopted and designated by the legislature (Bundestag, Bundesrat, Landtag). Ordinance includes standards adopted by a federal or provincial government or by a ministry. European legal norm includes norms of European primary or

secondary legislation, European organizations and other conventions and agreements.

Example (UNKREF32) includes a reference to the `Part-Time and Limited Term Employment Act' and the designation 'Basic Law'. The complex reference consists of the reference to the particular section of the law, its name and abbreviation (in brackets), date of issue, the reference in parenthesis and the details of the most recent change. Cases such as this one are a full reference. Example (UNKREF33), on the other hand, shows a short reference consisting of information on the corresponding section of the law and the abbreviated name of the statutory order.

... § 14 Absatz 2 Satz 2 des Gesetzes über Teil- RedOrange!70I zeitarbeit und lRedOrange!70Ibefristete Arbeitsverträge RedOrange!70I(Tz-RedOrange!70I BfG) vom 21. Dezember 2000 (Bundesgesetz-RedOrange!70I blatt Seite 1966), zuletzt geändert durch GesetzRedOrange!70I vomRedOrange!70IRedOrange!70I20.RedOrange!70IDezemberRedOrange!70I2011 (Bundesgesetzblatt IRedOrange!70I Seite 2854) RedOrange!70I GS, ist nach Maßgabe der Gründe mit dem Grundgesetz GS vereinbar.

`... section 14 paragraph 2 sentence 2 of the Law on Part-Time and Limited Term Employment Act (TzBfG) of 21 December 2000 (Federal Law Gazette I, page 1966), as last amended by the Law of 20 December 2011 (Federal Law Gazette I, page 2854), shall be published in accordance with the reasons compatible with the Basic Law.'

... Neuregelung in § 35 Abs. 6 StVO VO...

`... new regulation in sec. 35 para. 6 StVO...'

Annotation of the Dataset ::: Semantic Classes ::: Case-by-case Regulation

The class case-by-case regulation REG contains individual binding acts. These include regulation VS and contract VT. Regulation is an internal order or instruction from a superordinate authority to a subordinate, regulating their activities. In addition to administrative regulations, these include guidelines, circulars and decrees. In contrast to legal norm, these rules have no direct effect on the citizen. The class contract includes public contracts, international treaties and collective agreements. Some designations and references from these classes are similar to legal norm (UNKREF35, UNKREF36).

... insbesondere durch die Richtlinien zur Be-Peach!70l wertung des Grundvermögens –BewRGr– vom Peach!70l 19. September 1966 (BStBl I, S. 890) VS.

`... in particular by the Guidelines for the Valuation of Real Estate – BewRGr – of 19 September 1966 (BStBl I, p. 890).'

... fand der Manteltarifvertrag für die Beschäf-Goldenrod!70l tigten der Mitglieder der TGAOK VT (BAT/Goldenrod!70l AOK-Neu VT) vom 7. August 2003 Anwendung.

`... the Collective Agreement for the Employees of Members of TGAOK (BAT/AOK-New) was applied of 7 August 2003 ...'

Annotation of the Dataset ::: Semantic Classes ::: Court Decision

The class court decision RS includes references to decisions. It does not have any subclasses, the coarsed and fine-grained versions are identical. In court decision, the name of the official decision-making collection, the volume and the numbered article are cited. Often mentioned are also the court, if necessary the decision type, date and file number. Example (UNKREF39) cites decisions of the Federal Constitutional Court (BVerfG) and the Federal Social Court (BSG). Decisions of the BVerfG are

referenced with regard to pages, while decisions of the BSG are sorted according to paragraphs, numbers and marginal numbers.

Annotation of the Dataset ::: Semantic Classes ::: Legal Literature

Legal literature LIT also contains references, but they refer to legal commentaries, legislative material, legal textbooks and monographs. The commentary in example (UNKREF39) includes the details of an author's and/or publisher's name, the name of a legal norm, a paragraph and a paragraph number. Multiple authors are separated by a slash. Textbooks and monographs are cited as usual (author's name, title, edition, year of publication, page number). References of legislative materials consist of a title and reference marked with numbers.

... vgl zB BVerfGE 62, 1, 45 RS; BVerfGEDandelion!70I 119, 96, 179 RS; BSG SozR 4–2500 § 62 NrDandelion!70I 8 RdNr 20 f RS; Hauck/Wiegand, KrV 2016,Tan!60!Bittersweet!70!whiteI 1, 4 LIT ...

`... cf. i.e. BVerfGE 62, 1, 45; BVerfGE 119, 96, 179; BSG SozR 4–2500 § 62 Nr 8 RdNr 20 f; Hauck/Wiegand, KrV 2016, 1, 4 ...'

Description of the Dataset

The dataset, which also includes annotation guidelines, is freely available under a CC-BY 4.0 license. The named entity annotations adhere to the CoNLL-2002 format BIBREF14, while time expressions were annotated using TimeML BIBREF15.

Description of the Dataset ::: Original Source Documents

Legal documents are a rather heterogeneous class, which also manifests in their linguistic properties, including the use of named entities and references. Their type and frequency varies significantly, depending on the text type. Texts belonging to specific text type, which are to be selected for inclusion in a corpus must contain enough different named entities and references and they need to be freely available. When comparing legal documents such as laws, court decisions or administrative regulations, decisions are the best option. In laws and administrative regulations, the frequencies of person, location and organization are not high enough for NER experiments. Court decisions, on the other hand, include person, location, organization, references to law, other decision and regulation.

Court decisions from 2017 and 2018 were selected for the dataset, published online by the Federal Ministry of Justice and Consumer Protection. The documents originate from seven federal courts: Federal Labour Court (BAG), Federal Fiscal Court (BFH), Federal Court of Justice (BGH), Federal Patent Court (BPatG), Federal Social Court (BSG), Federal Constitutional Court (BVerfG) and Federal Administrative Court (BVerwG).

From the table of contents, 107 documents from each court were selected (see Table). The data was collected from the XML documents, i. e., it was extracted from the XML elements Mitwirkung, Titelzeile, Leitsatz, Tenor, Tatbestand, Entscheidungsgründe, Gründen, abweichende Meinung, and sonstiger Titel. The metadata at the beginning of the documents (name of court, date of decision, file number, European Case Law Identifier, document type, laws) and those that belonged to previous legal proceedings was deleted. Paragraph numbers were removed. The extracted data was split into sentences, tokenised using SoMaJo BIBREF16 and manually annotated in WebAnno BIBREF17.

The annotated documents are available in CoNNL-2002. The information originally represented by and through the XML markup was lost in the conversion process. We decided to use CoNNL-2002 because our primary focus was on the NER task and experiments. CoNNL is one of the best practice formats for

NER datasets. All relevant tools support CoNNL, including WebAnno for manual annotation. Nevertheless, it is possible, of course, to re-insert the annotated information back into the XML documents.

Description of the Dataset ::: Annotation of Named Entities

The dataset consists of 66,723 sentences with 2,157,048 tokens (incl. punctuation), see Table . The sizes of the seven court-specific datasets varies between 5,858 and 12,791 sentences, and 177,835 to 404,041 tokens. The distribution of annotations on a per-token basis corresponds to approx. 19–23 %. The Federal Patent Court (BPatG) dataset contains the lowest number of annotated entities (10.41 %).

The dataset includes two different versions of annotations, one with a set of 19 fine-grained semantic classes and another one with a set of 7 coarse-grained classes (Table). There are 53,632 annotated entities in total, the majority of which (74.34 %) are legal entities, the others are person, location and organization (25.66 %). Overall, the most frequent entities are law GS (34.53 %) and court decision RS (23.46 %). The other legal classes (ordinance VO, European legal norm EUN, regulation VS, contract VT, and legal literature LIT) are much less frequent (1–6 % each). Even less frequent (less than 1 %) are lawyer AN, street STR, landscape LDS, and brand MRK.

The classes person, lawyer and company are heavily affected by the anonymisation process (80 %, 95 % and 70 % respectively). More than half of city and street, about 55 %, have also been modified. Landscape and organization are affected as well, with 40 % and 15 % of the occurrences edited accordingly. However, anonymisation is typically not applied to judge, country, institution and court (1–5 %).

The dataset was originally annotated by the first author. To evaluate and potentially improve the quality of

the annotations, part of the dataset was annotated by a second linguist (using the annotation guidelines specifically prepared for its construction). We selected a small part that could be annotated in approx. two weeks. For the sentence extraction we paid special attention to the anonymised mentions of person, location or organization entities, because these are usually explained at their first mention. The resulting sample consisted of 2005 sentences with a broad variety of different entities (3 % of all sentences from each federal court). The agreement between the two annotators was measured using Kappa on a token basis. All class labels were taken into account in accordance with the IOB2 scheme BIBREF18. The inter-annotator agreement is 0.89, i. e., there is mostly very good agreement between the two annotators. Differences were in the identification of court decision and legal literature. Some unusual references of court decision (consisting only of decision type, court, date, file number) were not annotated such as `Urteil des Landgerichts Darmstadt vom 16. April 2014 – 7 S 8/13 –'. Apart from missing legal literature annotations, author names and law designations were annotated according to their categories (i. e., `Schoch, in: Schoch/Schneider/Bier, VwGO § 123 Rn. 35', `Bekanntmachung des BMG gemäß §§ 295 und 301 SGB V zur Anwendung des OPS vom 21.10.2010').

The second annotator had difficulties annotating the class law, not all instances were identified (`§ 272 Abs. 1a und 1b HGB', `§ 3c Abs. 2 Satz 1 EStG'), others only partially (`§ 716 in Verbindung mit' in `§ 716 in Verbindung mit §§ 321 , 711 ZPO'). Some titles of contract were not recognised and annotated (`BAT', `TV-L', `TVÜ-Länder' etc.).

This evaluation has revealed deficiencies in the annotation guidelines, especially regarding court decision and legal literature as well as non-entities. It would also be helpful for the identification and classification to list well-known sources of law, court decision, legal literature etc.

Description of the Dataset ::: Annotation of Time Expressions

All court decisions were annotated automatically for time expressions using a customised version of HeidelTime BIBREF19, which was adapted to the legal domain BIBREF20. This version of Heideltime achieves an F$_1$ value of 89.1 for partial identification and normalization. It recognizes four TIMEX3-types of time expressions BIBREF21: DATE, DURATION, SET, TIME. DATE describe a calendar date (`23. July 1994', `November 2019', `winter 2001' etc). It also includes expressions such as `present', `former' or `future'. DURATION describes time periods such as `two hours' or `six years'. SET describes a set of times/periods (`every day', `twice a week'). TIME describes a time expression (`13:12', `tomorrow afternoon'). Expressions with a granularity less than 24 hours are of type TIME, all others are of type DATE. The distribution of TIMEX3 types in the legal dataset is shown in Table with a total number of 35,119 time expressions, approx. 94 of which are of type DATE.

...vgl. BGH, Beschluss vom blue$<$TIMEX3 redtid=Purple"t14" redtype=Purple"DATE" redvalue=Purple”1999-02-03”blue$>$3. Februar 1999blue$<$/TIMEX3$>$ – 5 StR 705/98, juris Rn. 2 ...

Evaluation

The dataset was thoroughly evaluated, see leitner2019fine for more details. As state of the art models, Conditional Random Fields (CRFs) and bidirectional Long-Short Term Memory Networks (BiLSTMs) were tested with the two variants of annotation. For CRFs, these are: CRF-F (with features), CRF-FG (with features and gazetteers), CRF-FGL (with features, gazetteers and lookup). For BiLSTM, we used models with pre-trained word embeddings BIBREF22: BiLSTM-CRF BIBREF23, BiLSTM-CRF+ with character embeddings from BiLSTM BIBREF24, and BiLSTM-CNN-CRF with character embeddings from CNN BIBREF25. To evaluate the performance we used stratified 10-fold cross-validation. As expected, BiLSTMs perform best (see Table). The F$_1$ score for the fine-grained classification reaches 95.46 and 95.95 for the coarse-grained one. CRFs reach up to 93.23 F$_1$ for the fine-grained classes and 93.22 F$_1$ for the coarse-grained ones. Both models perform best for judge, court and law.

Conclusions and Future Work

We describe a dataset that consists of German legal documents. For the annotation, we specified a typology of characteristic semantic categories that are relevant for court decisions (i. e., court, institution, law, court decision, and legal literature) with corresponding annotation guidelines. A functional service based on the work described in this paper will be made available through the European Language Grid BIBREF26.

In terms of future work, we will look into approaches for extending and further optimizing the dataset. We will also perform additional experiments with more recent state of the art approaches (i. e., with language models); preliminary experiments using BERT failed to yield an improvement. We also plan to replicate the dataset in one or more other languages, such as English, Spanish, or Dutch, to cover at least one more of the relevant languages in the Lynx project. We also plan to produce an XML version of the dataset that also includes the original XML annotations.

Acknowledgements

This work has been partially funded by the project Lynx, which has received funding from the EU's Horizon 2020 research and innovation programme under grant agreement no. 780602, see http://www.lynx-project.eu.