AUTOMATED EXTRACTION OF SEMANTIC INFORMATION FROM GERMAN LEGAL DOCUMENTS

Bernhard Waltl / Jörg Landthaler / Elena Scepankova / Florian Matthes / Thomas Geiger / Christoph Stocker / Christian Schneider

Research Associates, Technical University of Munich, Department of Informatics, Software Engineering for Business Information Systems
Boltzmannstraße 3, 85748 Garching bei München, DE
b.waltl@tum.de / joerg.landthaler@tum.de / elena.scepankova@tum.de

Professor, Technical University of Munich, Chair for Software Engineering for Business Information Systems
Boltzmannstraße 3, 85748 Garching bei München, DE
matthes@tum.de; https://wwwmatthes.in.tum.de/

DATEV eG, Department for Portals and Collaboration (EM45)
Fürther Straße 111, 90329 Nürnberg, DE
thomas.geiger@datev.de / christoph.stocker@datev.de

DATEV eG, Department for Content – Taxes and Law (BC222)
Obere Kanalstraße 2–4, 90329 Nürnberg, DE
christian.schneider@datev.de

Keywords: ***Legal Data Science, Text Mining, Semantic Analysis, Structured Information***

Abstract: *Based on a collaborative data science environment, and a large document corpus (> 130'000 documents from German tax law) we demonstrate the extraction of semantic information. This paper shows the potential of rule-based text analysis to automatically extract semantic information, such as the year of dispute in cases. Additionally, it demonstrates the extraction of legal definitions in laws and the usage of terms in a defining context. Based on an iterative and interdisciplinary process, legal experts, software engineers, and data scientists evaluate and continuously refine the model used for the computer-supported extraction.*

1. Introduction

The application of data science methods in the domain of legal informatics is well established. Thereby, the provision of decision support for legal experts has always been in the focus of legal data science. This support covers a wide range from information retrieval to artificial intelligence. Due to the increasing possibilities of text mining capabilities, support for data, time, and knowledge intensive work is becoming more and more attractive. Thereby, not only the work of legal experts can be supported, but also of persons involved in the pre-processing and preparation of documents for subsequent processes, such as publishers and editors. In any case, the interdisciplinary field of legal informatics requires the collaboration of computer scientists and experts from the legal domain.

This paper describes the results of applying state-of-the-art natural language processing technologies on a given corpus of legally relevant documents from the domain of German tax law. An overview of relevant related literature is given in Section 2. The collaborative data science environment, the used document corpus (>130.000 documents), and the interdisciplinary and explorative data science method are introduced in Section 3. Section 4 presents the basic idea of rule based text annotation and describes the necessity of the interdisciplinary collaboration (see Section 4.1.). Results of the extraction of relevant metadata from a given document, such as the year of dispute (Streitjahr), are shown in Section 4.2. Continuing with the extraction of semantic information, Section 4.3. presents the possibilities and limitations of the extraction of legal definitions or valid interpretations of legal terms within laws and cases.

2. Related work

The extraction of data from unstructured information, i.e. structuring text through computer-supported analysis, is very attractive for the legal domain. Nowadays, the extraction of data, such as the analysis regarding the concrete year of dispute within a case document, or the analysis of legal definitions, is a relevant but mostly manual task. This manual editorial process ensures, on the one hand, the required high quality of labelled documents, but is, on the other hand, time-consuming and expensive. The basic idea is to support humans involved in the labelling process by trained and adapted algorithms determining relevant data, such as the year of dispute, automatically.

This basic idea of supporting the editorial process to extract particular data automatically is not new. However, considering the creation of genuine digital data, the increasing capabilities of algorithms, and the availability of computational power (processing time and memory) the applicability of those methods is making them highly attractive for practitioners. Still, the training of those methods is a manual and intensive task that requires a theoretical foundation. A plethora of approaches exist with different focusses and research questions.

In 2010 Maat and Winkels [Maat/Winkels 2010] proposed a taxonomy of classes for sentences occurring in laws. Based on Hart's famous book «The concept of law» [Hart 1961] they differentiated not only paragraphs, i.e. articles, of laws, but also sentences based on their function within the law. Thereby, the taxonomy differentiates between core rules, rules regarding definitions, etc. Using regular expressions (regex), they formalized patterns to identify those. Ultimately, the classifier reached a precision of 91%. Prior, in 2006, Maat et al. [Maat et al. 2006] had already extracted explicit references within norms with a remarkable parser accuracy of 95%; although, the linguistic variety was smaller.

Wyner et al. tried to extract arguments from legal cases [Wyner et al. 2010]. Thereby, they developed a context-free grammar that allowed the expressions of the rules to identify those expressions. Additionally, they differentiated between four classes of sentences in the context of arguments: premises, conclusions, non-argumentative information, and final decisions. The results were diverse, ranging from a precision of 59% and a recall of 70% for the identification of premises to a precision of 89% and a recall of 80% for non-argumentative information.

In 2015, Grabmeier et al. [Grabmair et al. 2015], created a powerful framework based on the Apache UIMA to annotate and classify legal texts based on linguistic and semantic features. Thereby, they used a rule based text annotation technology (Apache Ruta) allowing a more expressive specification of linguistic patterns than plain regular expressions. For their work, they used the system to extract semantic information out of decisions regarding vaccine injuries. They derived 49 rules to annotate on a sentence, but also a sub-sentence level.

The dissertation of Walter [Walter 2009] focused on the extraction of legal definitions from the federal constitutional court (BVerfG) only. He modelled the linguistic variety in detail, also considering the difficult problem of negations within legal definitions and linguistic patterns indicating a definition. Although the results and the technology used have not been convincing, his linguistic work sets a base line for further research in this direction.

3. Research Method

3.1. Data Science Environment

The analysis of legal documents has been done in an existing legal data science environment, that allows the collaboration on legal documents. The environment is a web application implemented with a Java back-end. From a software engineering point of view, the application consists of an implementation of a state-of-the-art pipes & filters architecture (see Figure 3.1).

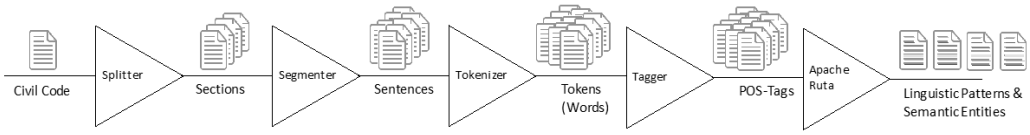
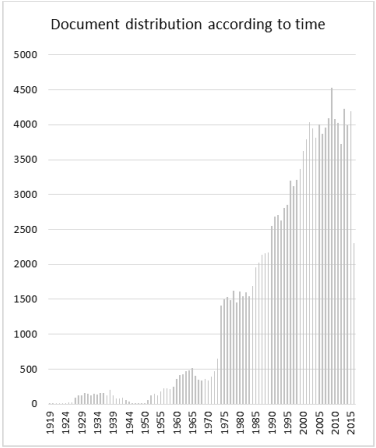


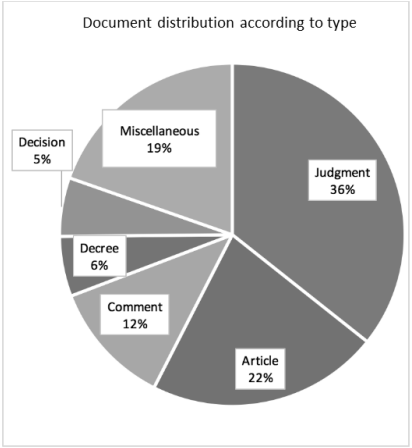
Figure 3.1: Pipes filters architecture for semantic analysis of legal documents consisting of unstructured data, i.e. text.

Besides the standard components for natural language processing for segmentation, tokenization, POS-tagging, etc., the environment allows the definition of complex linguistic patterns based on semantic annotations, which are applied to indexed legal documents. In [Waltl/Matthes/Grass 2016] we already showed that it is possible to extract legal definitions from laws, e.g., §90 German Civil Code. The definition of those legal definitions requires the specification of patterns more complex than regular expressions. Although basic vocabulary and keywords can be extracted from text documents quickly and accurately with regular expressions, they do not allow to consider linguistic information of words, such as part-of-speech information. More details on the data science environment, the different components and the base line architecture to perform computational intensive data analysis processes on large text corpora can be found in [Waltl/Matthes/Grass 2016].

3.2. Legal Document Corpus

The document corpus provided by the DATEV eG[1] consists of more than 130.000 different documents related to German tax law. The documents cover a time span of almost 100 years. The oldest documents indexed are from 1919, whereas the latest document in the corpus was published in July 2016. The corpus is fully digitized and available in XML but also in JSON format, in which each document is represented by a single file. Furthermore, the corpus consists of more than 40 different types, such as judgments (dt. Urteile), articles (dt. Aufsätze), laws (dt. Gesetze), etc. The corpus is a selection of the documents stored in the DATEV legal information database LEXinform.





[1] https://www.datev.de/, accessed 29.12.2016.

4. Structuring Unstructured Information in the German Tax Law

4.1. Rule-based text annotation as an interdisciplinary task

The computer-supported semantic analysis of legal documents has been an interdisciplinary challenge for a long time. This interdisciplinary challenge is hardly effected by the circumstance that algorithms for data and text mining have become more powerful and are easier to integrate in information systems. It is essential for legal data scientists to collaborate with legal practitioners. How a collaborative process for the semantic analysis involving legal practitioners and data scientists, that has proven to be successful in our project setting, could look like is shown in Figure 4.1.

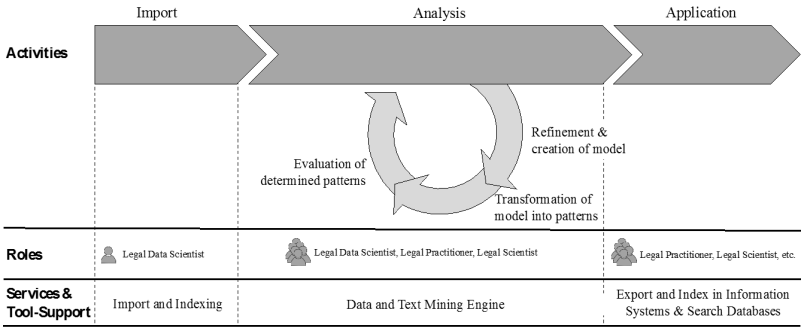


Figure 4.1: Reference process for computer-supported semantic analysis of legal texts.

The process consists of three main phases: import, analysis, and application. Within the import phase the document corpus is digitized, if necessary, and the data format harmonized, i.e. transformed, such that the documents share a common data format which can be accessed in the subsequent analysis phase. Thereby, especially the data format of the unstructured content of a document, has to be specified. This affects how the content can be analyzed and structured. Ultimately, this also influences the way how results can be integrated into information systems during the application phase. Therefore, the import phase is crucial for the quality of the subsequent phases. The problem with the quality of the content can become problematic when parsing OCR processed documents, which was not the case with this research project.

The analysis phase of the process mainly consists of three iterative steps, namely the creation and refinement of a linguistic model, which is the base line for the extraction of meta-data and legal definitions. The subsequent step transforms the model into a set of linguistic pattern definitions. Thereby, the integration of keywords and the use of a highly specialized segmenter, considering law specific abbreviations, tagging information, ontologies, and structural information (i.e. zoning) are possible. Finally, the patterns are applied to documents and the results can be evaluated. To perform the evaluation, the results can either be visualized by highlighting the corresponding parts within the documents, or the annotations of a document can be downloaded to be compared with a reference corpus, i.e. gold standard.

The third phase addresses the export and integration of the extracted information into other applications where it can be consumed by users. Thereby, various scenarios of visualizations in interactive user interfaces, considerations during search and navigation, or export to third-party applications, such as compliance databases or legal database information systems, via a tailored and extendable REST API, are possible.

We applied this method for two different approaches and used cases for the semantic analysis of legal documents: extraction of meta-data from cases from German tax law and determination of legal definitions, or rather, interpretative contexts of terms in the legal domain. The first use case is discussed in Section 4.2. and the focus is set on the extraction of the year of dispute from a case of German tax law. The second use case,

described in Section 4.3., sketches our approach to determine legal definitions in laws and the usage of relevant legal terms in a defining context.

The two use cases illustrate the potential of legal data science from a text mining perspective. On the one hand, the extraction of meta data is not only beneficial for end-users searching the database but also for editorial staff within companies preparing or publishing legal data. On the other hand, the extraction of more complex data, such as legal definitions, is highly desirable for end-users of legal information databases. This refers to the existing idea of a semantic analysis of texts. Thereby, semantics can be considered as additional information that represents the intention, meaning, or result of a hermeneutical interpretation process. The text mining approach described in this paper is an attempt to create this additional information element, i.e. annotation, by capturing its linguistic environment.

4.2. Extracting (meta-)data from legal documents

Cases by German courts are published in plain pdf or, more recently, in xml format. Those documents usually consist of different information objects. The most important part of the document for our purposes is the text of the judgment. This text is structured trough different sections, namely guiding principles (dt. «Orientierungssatz/Leitsatz»), elements of an offense (dt. «Tatbestand»), and reasoning (dt. «Gründe»). The sections are structured with paragraphs, which do not have titles or any other additional information. In addition to the text, cases contain information about the court, that decided this particular case, and the date when the judgment was published.

However, there is information contained in the text, that is not explicitly stated and whose extraction is a time-consuming and knowledge-intensive task. Currently, the structuring of unstructured information is done manually. On the one hand, this ensures high quality and accuracy, on the other hand, this task is labor intensive and therefore expensive. For a different domain, namely the classification of Flickr pictures, Wang et al. [Wang et al. 2012] already showed that tagging quality can be improved significantly by combining an interactive human and computer labelling system.

4.2.1. Approach

The implementation was done by training algorithms and models such that they can determine the year of dispute (dt. «Streitjahr») given a case document. The year of dispute is highly relevant within the domain of tax law, since it determines the relevant legislation.

To determine the year of dispute, it is required to constrain the search space by taking structural information into account, i.e. zoning. Since the year of dispute is generally stated within the section of the elements of an offense (dt. «Tatbestand») the algorithm, i.e. the input document for the analysis pipeline, only analyzes this particular section of a case. This restriction of the search space decreases the false positive rate and therefore contributes to the accuracy of the overall approach.

After the integration of zoning information, the following steps are performed to detect the year of dispute:

1. Constrain search space to section «Tatbestand»
2. Determine dates within the text
    a. Differentiate between specific dates, e.g., «21.10.2010» or «21. September 2010», dates referring to a whole year, e.g., «2010», and timespans «2000 bis 2009».
3. Determine indicating sentences, such as [Antragssätze] (Der Kläger beantragt…)
    a. If those sentences contain whole years or timespans, mark those as year of disputes.
4. Determine, based on particular linguistic patterns expressed in Apache Ruta, contexts that allow conclusions about the year of dispute.

a. **Pre-Indicators:** Linguistic features, i.e. tokens, words, patterns, indicating that the following date is likely to be the year of dispute.
Examples: «auf die im **Streitjahr** 2006 zugeflossenen Erstattungszinsen», «den **Einkommensteuerbescheid** 2006 vom 11.12.2007», «die **Kindergeldfestsetzung für den Zeitraum** von Oktober 2003 bis Dezember 2004 und von Januar 2006 bis Juni 2006», etc.
b. **Post-Indicators:** Linguistic features, i.e. tokens, words, patterns, indicating that the date mentioned before is likely to be the year of dispute.
Examples: «Erwerbsunfähigkeitsrente im Jahre 2005 **(Streitjahr)**»
c. **Clamp-Indicators:** Linguistic features, i.e. tokens, words, patterns, indicating that the date mentioned between two features is likely to be the year of dispute.
Examples: «**ab Januar** 2008 **Kindergeld** zu bewilligen», «**Bescheidfür** 1997 und 1998 **über Einkommensteuer**»

4.2.2. Evaluation

During the evaluation phase 100 different case documents were randomly selected and verified. The results are shown in Table 1 below. Note, that a case can have multiple years of dispute.

		Prediction Outcome	
		YOD	**No YOD**
Actual Outcome	**YOD**	186	21
	No YOD	11	-

F1	$\frac{2 * 186}{2 * 186 + 21 + 11} \approx \mathbf{92}\ \%$
Precision	$\frac{186}{186 + 11} \approx \mathbf{94}\ \%$
Recall	$\frac{186}{186 + 21} \approx \mathbf{90}\ \%$

Table 1: Quality assessment of extraction of year of dispute (YOD) in 100 randomly selected cases.

With a precision of 94% and a recall of 90%, the performance of the detection mechanism works fairly good. This result could be improved even further by providing a more comprehensive set of rules and vocabulary. Even though in many cases the year of dispute can be found through this approach, there are particular cases that do not allow to fully extract the desired information on the linguistic level with high confidence. In complex cases where many facts are described over many years, the algorithm can easily make mistakes and provide misleading (or insufficient) results. Therefore, this approach should be used to support processes of, e.g., editorial staff, rather than to make autonomous decisions.

4.2.3. Critical reflection

The main challenge during the extraction of the year of dispute from case documents is the linguistic variety in which those information are lexically and syntactically encoded; considering especially the huge time span and different types of German courts that are publishing the documents. Finally, the combination of structural (zoning) information and keywords indicating the occurrence has allowed us to create Ruta expressions (rules) to extract the desired information.

4.3. Determining legal definitions and defining contexts

The determination of legal definitions has been in the focus of legal data science for several years. Different approaches already showed, that it is possible – at least up to a certain degree – to automatically extract legal definitions from laws [Waltl/Matthes/Waltl/Grass 2016] and from judgments from the German constitutional court [Walter 2009]. However, in many cases it is not necessary to extract a legal definition from the law, which is often vague and undetermined, but to find phrases and sentences indicating a defining context within a case. The extraction of those defining contexts enables the understanding of valid interpretations of a

term. This issue can be illustrated considering § 9 EStG «Werbungskosten». Within those disputes additional interpretations or context-depending definitions are introduced by German courts. Consequently, the extraction of definitions and interpretations from those cases helps users to understand valid meanings of this term. This meaning can evolve over time and throughout courts, of course [Waltl 2016]. Nevertheless, algorithms can – up to a certain degree – be trained, such that they can automatically make suggestions for those defining contexts [Maat/Winkels 2010].

4.3.1. Approach

To determine legal definitions and defining contexts, a taxonomy was created differentiating between legal definitions in a narrow sense (source: statutory texts), contexts that extend definitions (source: cases), and interpretation of legal terms (source: commentaries and articles). This taxonomy also reflects the priority of a definition considering its origin. From a linguistic point of view one could also differentiate between constructive vs. non-constructive, inclusive vs. exclusive, enumerating vs. describing, and comprehensive vs. open legal definitions. Within this research project we focused on the extraction of legal definitions from laws and from cases.

Based on this two-dimensional differentiation, several examples, i.e. sentences, from laws and judgments can illustrate the overall approach:

1. ***Law: legal definitions (narrow sense)***
   Examples: «Sachen **im Sinne des Gesetzes** sind nur körperliche Gegenstände.» (§ 90 BGB), «Die Anfechtung muss in den Fällen der §§ 119, 120 ohne schuldhaftes Zögern (**unverzüglich**) erfolgen, [...].» (§ 121 BGB), «Zum Inland **im Sinne dieses Gesetzes** gehört auch der der Bundesrepublik Deutschland zustehende Anteil am Festlandsockel, soweit [...].» (§ 1 Abs. 1 Satz 2 EStG), «**Der** Gesamtbetrag der Einkünfte, [...], **ist dasEinkommen.**» (§ 2 Abs. 4 EStG)
2. ***Judgment: term declaration and defining contexts (with different sub-levels)***
   Examples: «Nach § 8 Abs. 1 EStG sind alle Güter, die in Geld oder Geldeswert bestehen und [...], Einnahmen.», «Dagegen liegt dann kein Arbeitslohn vor, wenn [...], nicht [...] gewährt wird.», «Der Aggregatszustand der Gegenstände ist unbeachtlich, so dass auch der elektrische Strom als Ware i.S.d. Vorschrift gilt.», «Hersteller sei demnach jeder, in dessen Organisationsbereich das Produkt entstanden ist.»

The determined patterns have been expressed in Apache Ruta and applied to the legal document collection. Again, structural information has been considered to restrict the analysis to relevant parts within judgments, namely to the chapter «Reasoning» (dt. «Gründe»). This restriction reduces the noise by avoiding potential false positives.

4.3.2. Critical reflection

The main challenge during the classification of phrases or sentences is the demarcation between sentences with and without defining content. This problem is a little easier to solve for laws since the linguistic variety is smaller and the field has been better studied throughout literature, but it is still far from trivial, and in many cases, not definitely determinable.

5. Conclusion and Outlook

In this research paper, we have shown the potential of legal data science for text analysis of cases from German tax law. Thereby, we used an existing data science environment to explore two concrete use cases: i) the computer-supported extraction of the year of dispute for German cases and ii) the extraction of legal definitions and defining contexts of legal terms in judgments. During this project, we jointly worked together in an

interdisciplinary team of researchers from a university and practitioners from the industry. We have briefly sketched our interdisciplinary approach within the paper.

Both use cases were implemented using a rule-based approach to determine the desired information. We have chosen a combination of structural and linguistic indicators to analyze the documents. Thereby, we have written custom annotators to classify relevant sentences within the cases. In addition, we have provided a comprehensive set of rules, which are logically connected. For the extraction of concrete info as in i), we have achieved a precision of 94% and a recall of 90%. The extraction of legal definitions as in ii), is more difficult and we not only lack the necessary technologies but it also requires more work from the field of legal theory to be performed. A legal, theoretical foundation, targeted at and suitable for a subsequent analysis through algorithms, is still missing. It can be considered as one of the major challenges for legal theory, to improve existing theories such that they can be used during legal data science tasks.

6. References

Matthias Grabmair/Kevin D. Ashley/Ran Chen/Preethi Sureshkumar/Chen Wang/Eric Nyberg/Vern R. Walker, Introducing LUIMA: An Experiment in Legal Conceptual Retrieval of Vaccine Injury Decisions Using a UIMA Type System and Tools, ICAIL Proceedings, 2015.

Herbert Lionel Adolphus Hart, The concept of law, Oxford University Press, 1961.

Emile de Maat/RadboudWinkels, Automated Classification of Norms in Sources of Law, Semantic processing of legal texts, 2010.

Emile de Maat/RadboudWinkels/Tom van Engers, Automated detection of reference structures in law, JURIX, 2006.

Stephan Walter, Definition extraction from court decisions using computational linguistic technology, Formal Linguistics and Law, vol. 212, 2009.

Bernhard Waltl, Computer-gestützte Analyse des Bedeutungswandels rechtlicher Begriffe, Wien: OCG, 2016.

Bernhard Waltl/Florian Matthes/Tobias Waltl/Thomas Grass, LEXIA: A data science environment for Semantic analysis of german legal texts, in: Erich Schweighofer/Franz Kummer/Walter Hötzendorfer/Georg Borges (Hrsg.), Netzwerke / Networks –Tagungsband des 19. Internationalen Rechtsinformatik Symposions IRIS 2016, OCG, Wien/Bern 2016.

Meng Wang/Bingbing Ni/Xian-Sheng Hua/Tat-Seng Chua, Assistive Tagging: A Survey of Multimedia Tagging with Human-Computer Joint Exploration, ACM Comput. Surv., vol. 44, no. 4, 2012.

Adam Wyner/Raquel Mochales-Palau/Marie-Francine Moens/David Milward, Approaches to text mining arguments from legal cases, Semantic processing of legal texts, 2010.