Differentiation and Empirical Analysis of Reference Types in Legal Documents

Bernhard WALTL [a,1], Jörg LANDTHALER [a], and Florian MATTHES [a]
[a] *Software Engineering for Business Information Systems, Technische Universität München, Germany*

Abstract.
This paper proposes an extensible model distinguishing between reference types within legal documents. It differentiates between four types of references, namely fully-explicit, semi-explicit, implicit, and tacit references.

We conducted a case study on German laws to evaluate both: the model and the proposed differentiation of reference types. We adapted text mining algorithms to determine and classify the different references according to their type. The evaluation shows that the consideration of additional reference types heavily impacts the resulting network structure by inducing a plethora of new edges and relationships. This work extends the approaches made in network analysis and argues for the necessity of detailed differentiation between references throughout legal documents.

Keywords. References, reference types, citations, citation types, natural language processing, regular expression, data analysis, text mining, legal data science

Introduction

Throughout legal systems various complementary dimensions inducing networks exist. Network structures can emerge on and throughout different levels such as nations, companies, organizations, institutions, people (roles), ..., and legal documents. The latter is in particular interesting for this research. Although many different attempts have already been made to describe, model, analyze, visualize, or evaluate networks arising from legal texts, surprisingly less effort has been spent on the differentiation of reference types between legal documents. This paper's contribution narrows this gap by providing a constructive and extensible differentiation of four different reference types. In Section 2 we present the results and the evaluation of the reference analysis in German legislative texts, showing that many different relationships beside the well-studied citations exist and that those can be automatically determined.

1. A Conceptual Framework for Reference Types in Legal Texts

We seize on the differentiation of reference types according to Albrecht Berger [6]. We show how and which technology can assist within the detection of the proposed reference

[1]Corresponding Author: Bernhard Waltl, Software Engineering for Business Information Systems, Boltzmannstr. 3, 85748 Garching bei München, Germany; E-mail: b.waltl@tum.de.

types and briefly discuss a generic tool-support to examine and explore legal data with respect to automatically determined references.

Many attempts have already been made to analyze, extract and visualize the network structure throughout legal texts. Rather less effort has been spent on the differentiation between reference types. Beside the well-known citation that can be determined using basic technology, e.g., regular expressions (see also [1]), there exist three more reference types that heavily impact the interpretation of legal texts. Table 1 presents the different reference types, a short description, a selection of illustrative examples, and additional literature providing detailed information and further readings.

Reference Type	**Description**	**Example(s)**	**Literature**
Full-explicit reference (FR)	The referenced norm, respectively document, is explicitly stated and provides the full information about paragraph and document.	§81 Abs. 1 Satz 3; §§32 und 34; §126 Abs. 1 Satz 2 Nr. 3 der Grundbuchordnung;	[2,3,4,1,5]
Semi-explicit reference (SR)	The reference norm, respectively document, is named but provides no explicit information about the referenced article or document.	"[...] finden die Vorschriften über die Hypothek entsprechende Anwendung [...]" (see §1192 BGB)	[6,7,2]
Implicit reference (IR)	The referencing norm uses a term, that is legally defined in another – not-explicitly mentioned – norm.	"Wer eine fremde Sache beschädigt oder zerstört [...]" (see §228 BGB); The term "Sache" is defined in §90 BGB.	[7,8,9,6]
Tacit reference (TR)	The connection between the norms emerges due to systemic interpretation and cannot not be determined by exclusively analyzing the norm text.	"lex posterior derogat lex inferior"; Connections between general part (book 1) and specific part (book 2) of the BGB.	[8,6,2]

Table 1. Structured consolidation of different reference types present in legal documents.

This conceptual framework serves as a base line for the used technology stack and implementation. Thereby, we will discuss if and how the different reference types can be determined using algorithms.

2. Empirical Analysis of German Laws

This Section summarizes the analysis and evaluation on a subset of German federal laws. Thereby, we implemented a prototype to perform the analysis and selected ten (out of more than 6,000) German laws containing the most tokens (i.e. words).

2.1. Empirical Analysis of Reference Types: Dataset, Analysis, and Evaluation

Based on German laws we have analyzed the usage and occurrence of the various reference types. Thereby, we have selected ten federal laws containing the most tokens out of more than 6,000 available federal laws. Table 2 summarizes the key findings.

Table 2 shows that the German Civil Code (BGB) contains 2,072 FR, of which 1.918 are inbound and 154 are outbound. In addition, there exist 411 SR, (≙ 19.84% compared to FR) and 2,570 IR (≙ 124.03% compared to FR). This analysis shows that the mere

Law	#T ↓	#§	FR (in, out)	SR	SR (rel)	IR	IR (rel)
BGB	185,751	2,381	2,072 (1,918; 154)	411	19.84%	2,570	124.03%
SGB 5	147,621	456	4,678 (4,220; 458)	52	1.11%	426	9.11%
KAGB	113,166	356	3,157 (2,781; 376)	64	2.03%	3,701	117.23%
KredWG	91,145	208	2,657 (2,234; 423)	37	1.39%	1,393	52.43%
HGB	90,877	643	1,733 (1,514; 219)	102	5.89%	496	28.62%
ZPO	90,421	1,003	927 (794; 133)	83	8.95%	94	10.14%
SGB 6	84,683	413	1,165 (901; 264)	78	6.70%	344	29.53%
AMG	77,002	216	2,281 (2,112; 169)	34	1.49%	420	18.41%
StPO	74,887	644	1,757 (1,426; 331)	38	2.16%	48	2.73%
StGB	62,986	518	1,313 (1,234; 79)	4	0.30%	48	3.66%

Table 2. Analysis of the reference types on German laws. The table shows the law, number of tokens (#T), number of articles (#§), full-explicit references (FR, inbound and outbound), semi-explicit references (SR), semi-explicit references relative to FR (SR rel), implicit references (IR), and implicit references relative to FR (IR rel).

consideration of FR neglects a huge part of the emerging links between norms of a law. 2,981 (= 411 + 2,570) references are additionally induced by linguistic and semantic relationships. Similar conclusions can be drawn for the Capital Investment Law (KAGB). Thereby, the law heavily uses concepts and terms, that are legally defined within the law. The evaluation showed that those terms are mainly specific abbreviations, such as AIF, OGAW, or terms like “Ausgabepreis”, “Rücknahmepreis”. The usage of abbreviations and highly specified terminology makes the evaluation difficult since the demarcation between legal definition and domain specific term becomes ambiguous.

We manually derived the regular expressions and respective Apache Ruta scripts on the product liability act and the general part of the German Civil Code. Thereby, we have created the expressions and rules to determine full-explicit, semi-explicit and implicit references (i.e., legal definitions). We evaluated the precision and recall on a subset (n = 100 ≙ 19%) of the German Penalty Law (StGB) articles with respect to full-explicit references (precision: 98%; recall 97%), semi-explicit references (precision: 80%; recall 80%), implicit references (precision: 93%; recall 93%). We additionally evaluated a subset (n = 50 ≙ 23%) of the articles of the banking act (KWG) articles, with respect to full-explicit references (precision: 89%; recall 88%), semi-explicit references (precision: 82%; recall 60%), and implicit references (precision: 96%; recall 92%). The results are satisfying but could be further improved, e.g., the recall of semi-explicit references. This can be achieved by investing more time and effort in defining additional and more accurate Apache Ruta pattern definitions.

Table 2 shows that German laws differ heavily by the amount of FR, SR, and IR. However, considering those heavily impact the resulting network structure, since various additional relationships, i.e. links, are induced.

3. Conclusion and Outlook

Beside the well-studied citations several additional reference types exist throughout legal documents. We argued, that beside full-explicit references, it is necessary to consider at least three additional reference types to comprehensively analyze the network emerging within legal documents. Consequently, in order to fully capture links between

legal documents at least those four reference types have to be considered. Using existing data and text mining methods we proposed a technology stack that is suitable to determine those references based on linguistics, e.g., regular expressions, or more elaborate semantic features, e.g., Jape grammar, Apache Ruta.

We prototypically implemented regular expressions and Apache Ruta scripts to determine and evaluate the detection of references according to their type. Using publicly available data from German legislation we analyzed laws regarding the occurrence of full-explicit, semi-explicit and implicit references. The results show that beside the full-explicit reference numerous semi-explicit and implicit references exist in legal documents. In the German Civil Code the number of references induced by terminology, i.e. implicit references, is even higher than the full-explicit references (124%). Although, the evaluation has shown that accuracy drops for laws that are domain specific, such as the Capital Investment Law (KAGB), the results are promising and additional effort in training the patterns would be necessary to ensure sufficient accuracy.

The differentiation of reference types helps in understanding the network structures arising within legal documents and can be used in subsequent applications, such as recommender systems.

Acknowledgment

This research was sponsored in part by the German Federal Ministry of Education and Research (BMBF) (project "Software Campus (TU München)", grant no. 01IS12057). The authors thank everyone involved in Lexalyze for valuable discussions and remarks.

References

[1] R. Winkels, A. Boer, B. Vredebregt, and A. van Someren, "Towards a Legal Recommender System," in *Frontiers in Artificial Intelligence*, 2014, vol. Volume 271: Legal Knowledge and Information Systems, pp. 169–178. [Online]. Available: http://ebooks.iospress.nl/volumearticle/38453

[2] Bundesministerium der Justiz und Verbraucherschutz, *Handbuch der Rechtsförmlichkeit*, Berlin, 2008.

[3] R. Boulet, P. Mazzega, and D. Bourcier, "A Network Approach to the French System of Legal codes - Part I: Analysis of a Dense Network," *CoRR*, vol. abs/1201.1262, 2012.

[4] Michael J. Bommarito II, Daniel Katz, and Jon Zelner, "Law as a seamless web?: comparison of various network representations of the United States Supreme Court corpus (1791-2005)," in *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, Barcelona, Spain, 2009.

[5] J. Landthaler, B. Waltl, and F. Matthes, "Unveiling References in Legal Texts - Implicit versus Explicit Network Structures," *Jusletter IT*, 2016.

[6] A. Berger, *Die Erschliessung von Verweisungen bei der Gesetzesdokumentation*, ser. Informationssysteme. München-Pullach: Verlag Dokumentation, 1971, vol. Bd. 3.

[7] A. G. Debus, *Verweisungen in deutschen Rechtsnormen*. Duncker & Humblot, 2008.

[8] K. Larenz and C.-W. Canaris, *Methodenlehre der Rechtswissenschaft*. Berlin [u.a.]: Springer, 1995.

[9] Paul Zhang and Lavanya Koppaka, "Semantics-based legal citation network," in *Proceedings of the 11th international conference on Artificial intelligence and law*, Stanford, California, 2007, pp. 123–130.

[10] B. Waltl, F. Matthes, T. Waltl, and T. Grass, "LEXIA: A data science environment for Semantic analysis of german legal texts," *Jusletter IT*, 2016.