CITATION ANALYSIS OF THE CZECH CASE-LAW: FIRST STEP TOWARDS THE GOLD STANDARD

Jakub Harašta / Matěj Myška / Michal Malaník / Jakub Míšek

Research fellow, Masaryk University, Faculty of Law, Institute of Law and Technology
Veveří 70, 61180 Brno, CZ
jakub.harasta@law.muni.cz; http://cyber.law.muni.cz

Senior Assistant Professor, Masaryk University, Faculty of Law, Institute of Law and Technology
Veveří 70, 61180 Brno, CZ
matej.myska@law.muni.cz; http://cyber.law.muni.cz

Ph.D. Candidate, Masaryk University, Faculty of law, Department of Legal Theory
Veveří 70, 61180 Brno, CZ
michmal@mail.muni.cz

Ph.D. Candidate, Masaryk University, Faculty of law, Institute of Law and Technology
Veveří 70, 61180 Brno, CZ
jkb.misek@mail.muni.cz; http://cyber.law.muni.cz

Keywords: ***Natural Language Processing, Gold Standard, Citation Analysis, Methodology***

Abstract: *In this paper, the authors aim to present the starting point of the development of the gold standard corpus of the Czech case law. The Czech Republic largely lacks the standardisation and prior work allowing to do this automatically. Therefore, the manually annotated gold standard corpus has to be created. Our final goal is to use the gold standard corpus for automated annotation of the case law of the Constitutional Court, Supreme Court and Supreme Administrative Court. We explain our motivation, basic background, previous language-specific research and small-scale experiments used in developing the initial methodology.*

1. Introduction

In this paper, the authors aim to present an initial phase of research aimed to building of resources necessary for citation and sentiment analysis of the case law in the Czech Republic. The authors aim to develop the gold standard corpus of case law references. This is required because the missing standard of citations disallows to process citations automatically. With missing standardized structure of citations and mostly missing previous work aimed at the Czech environment, the knowledge bottleneck mentioned by Wyner[1] requires manual annotation to develop the gold standard corpus containing references, citations and sentiments. Such corpus can then be utilized for re-evaluation of the research and assessment of techniques used to further processing. Our ultimate goal is to annotate the rest of the case law automatically.

2. Background – Ignorantia legis neminem excusat

The civil and common legal systems are converging in their approach towards the use of case law. This claim is materialized even within the new Czech Civil Code[2]. Our research aims to leverage this claim from several

[1] See Wyner/Peters/Katz, A Case Study on Legal Case Annotation, in *Ashley* (Ed.), Legal Knowledge and Information Systems (Frontiers in Artificial Intelligence and Applications 2013), 165–174.

[2] See Art. 13, act No. 89/2012 Sb., the Civil Code: «*Anyone seeking legal protection may reasonably expect that his legal case will be decided similarly to another legal case that has already been decided and that coincides in essential aspects with his legal case;*

aspects aiming towards the real availability and understandability of case law.

Ignorantia legis neminem excusat is understood as the notion of formal availability of the law. The actual knowledge of its addressees is in the sense of this principle irrelevant. However, with growing availability of judicial decisions and missing tools to evaluate what the law is, the principle becomes more and more formal and actually detached from reality.

Actual realization of this principle remains largely only possible by the use of proprietary information systems and is therefore unavailable to a large audience and consequently the law becomes «*entirely unpredictable for its addressees*»[3]. Moreover, not even proprietary information systems allow us to evaluate the relevance of existing case law and only rarely allow us to identify the most important cases without extensive prior knowledge[4].

We believe that the citation analysis of the case law is necessary, because it directly influences the understandability of law. Law is textual in its nature[5] and is directly created by the network of sources[6]. Citing cases allows us to argue by the collective legal knowledge and to create an environment of legal certainty. But without a tool to map these citations, we cannot achieve the fundamental ideal of democratic legal state in its very essence, because the formal availability of the law remains strictly formal.

Citing cases allows us to rely in our argumentation on the authority of existing court decisions[7]. This kind of referencing has certain conditions for both practitioners and judges, as it requires such citation, which allows other participants to reasonably search for cited decision. However, this is not sufficient to ensure automatic extraction of such citation. This can be efficiently achieved by the use of well-known standardized citations with firm syntax. However, with regard to the availability of law, most efficient is the use of vendor-neutral citations[8]. The Czech Republic currently does not have a single standardized way of citing case law that would assist in automation of citation analysis. Eventually, achieving the creation of a network of inter-related decisions of individual courts or the whole judiciary, or assessing[9] the importance of case law based on number of inward citations, directly contributes to the understandability of law.

---

where the legal case has been decided differently, anyone seeking legal protection has the right to a persuasive explanation of the reasons for such a variance. »

3 See Decision of Czech Constitutional Court No. Pl. ÚS 77/06, para. 39: «*The requirement of foreseeability of the law as a part of the rule of law principle ceases to be fulfilled when the amending legal act is a part of another legal act in the formal sense, whose content is in no relation with the amended legal act. Orientation of the legal norms addressees in the legal system without the use of instruments of information technology becomes totally impossible.*» (The translation above has been prepared by the authors.)

4 See Winkels/Ruyter, Survival of the Fittest: Network Analysis of Dutch Supreme Court Cases, in Palmirani et al (Ed.), AI Approaches to the Complexity of Legal Systems. Models and Ethical Challenges for Legal Systems, Legal Language and Legal Ontologies, Argumentation and Software Agents. DOI: 10.1007/978-3-642-35731-2-7 (Lecture Notes in Computer Science 2011), 106–115.

5 See Widdison, New Perspectives in Legal Information Retrieval, International Journal of Law and Information Technology 2002, 41–70.

6 See Post/Eisen, How Long is the Coastline of the Law? Thoughts on the Fractal Nature of Legal Systems, The Journal of Legal Studies 2000, 545–584 (545).

7 See Decision of Supreme Administrative Court No. 6 Ads 94/2007-73: «*When then the regional court on page 13 of its judgment ventured to cite case law, he did so in an entirely unacceptable way. The sense of quoting case law is based on the use of argumentative conclusions already highlighted by the recognized authority of the judicial nature and in the possibility resulting from this – shortening its own justification. But to fully meet its purpose and to be in compliance with the requirement reviewability of the decision such a link must be to a unequivocally specified decision in a way that will allow the parties its reasonable traceability* [internal citation omitted]. *In the Czech legal environment it is customary in the decision citation of Czech judicial authorities to indicate at least the court that issued the decision, the date of its release, reference number or case number, and source according to which the decision was cited, or whether or not it was published* […].» (The translation above has been prepared by the authors.)

8 See Widdison, International Journal of Law and Information Technology 2002, 41–70.

9 See Fowler et al, Network Analysis and the Law: Measuring the Legal Importance of Precedents at the U.S. Supreme Court, Political Analysis 2007, 234–346.

Due to the growing availability of ICTs and the growing demand for openness of administrative bodies, we now have a previously unseen number of court decisions directly available through various platforms. Therefore, we need a way to distinguish between the individual decisions and the cases with possible normative or argumentative effect outside of the proceedings. As a matter of fact, the Constitutional Court, Supreme Court and Supreme Administrative Court are publishing their case law in its entirety. Discussing prior restraints on attorney speech with regard to published and unpublished cases[10] is thus far off limits.

Our goal is to create the means to assess the number of citations of individual decisions and their relevance and to evaluate sentiment of these citations. This would, together with annotating the use of literature, lead to improvement in legal argumentation. Methodology used to create those means have to be publicly available and the same applies to the initial gold standard corpus. This will allow us to create a sufficiently reasoned structure that may be beneficial to courts, legal education and legal professionals.

3. Materials and Further Work

During this phase of research, we obtained the set of cases of the Supreme Administrative Court of the Czech Republic (unaccounted for in Kříž et al.[11], which is so far the most valuable related work in the Czech Republic). As of 6th November 2015, this set contains 38 164 decisions.

The methodology of annotation is currently being drafted, together with the annotation manual containing rules and examples of annotation. Because we want to achieve the possibility of automated annotation, it is of utmost importance that our annotations are correct. We believe that in some cases we will be able to ensure actual validity of the data – annotated categories will be correct, which is the case of explicit references and their attributes. However, in other cases we have to ensure reliability – only when annotated categories are annotated by various annotators consistently, annotations are reliable. Most probably, this will be the case of sentiment analysis. Of course, this whole notion arises from the assumption that high reliability eventually implies validity[12].

To enhance the reliability of our gold standard corpus, every version of annotation manual undergoes a small-scale experiment. In these small-scale experiments, one random decision from our set is chosen. To account for possible specifics of the Supreme Administrative Court in referencing etc., a second decision is randomly chosen from the Constitutional Court or Supreme Court. Both decisions are then annotated by 2 post-docs in law and 5 law Ph.D. candidates to achieve as unambiguous guidelines as possible. So far we did not utilize undergraduates, because in general they lack the prior knowledge for fast reference recognition due to lack of awareness of legal journals and reviews available in the Czech Republic or ability to recognize specific identifiers as implying certain attributes of reference. Inter annotators agreement for first experiments allows us to assess whether we should work with the notion of validity or reliability and therefore whether we actually need to ensure that single document is annotated by multiple annotators. As pointed out in previous research, $\kappa = 0.85$ suggests the annotation task to be easy[13], and eventually allows us to set differently tiered annotations to ease the cognitive load on annotators[14]. These small-scale experiments on drafts of manual allow us to decide, whether we work with the notion of validity (explicit references and their attributes) or reliability (sentiment analysis) in annotations (incomplete references to previous explicit references, implicit references to previous explicit references etc.).

[10] Comp. Tusk, No-Citation Rules as a Prior Restraint on Attorney Speech, Columbia Law Review 2003, 1202–1235.

[11] See Kříž et al, Statistical Recognition of References in Czech Court Decisions, in *Gelbukh/Espinoza/Galicia-Haro* (Ed.), Human-Inspired Computing and Its Applications. DOI: 10.1007/978-3-319-13647-9-6 (Lecture Notes in Computer Science 2014), 51–61.

[12] See Artstein/Poesio, Inter-coder Agreement for Computational Linguistics, Computational Linguistics 2008, 555–596.

[13] See Kříž et al, Statistical Recognition of References in Czech Court Decisions, 51–61 (55).

[14] See Annotation exercise in Wyner/Peters/Katz, A Case Study on Legal Case Annotation, 2013.

4. Conclusion

As mentioned above, the research is currently in its methodological phase – developing the annotation manual and the methodology that will eventually be used for creation of the gold standard corpus. As the Czech case law is published in various ways and under different designation, our aim is to lower the threshold for its cognition. The gold standard corpus will be used for automated annotation, which can be achieved through machine learning or tools to standardize references[15] – this currently remains unresolved.

The task remains labour intensive as it requires a significant amount of resources. The available previous work is either legal specific, but focused on different legal cultures or languages, or language specific, but lacking sufficient legal expertise[16]. Our research is therefore not to be understood as an exercise in machine learning, but as a necessary step towards efficient information retrieval[17] and towards the understandability of the cases constituting the law.

5. References

Artstein/Poesio, Inter-coder Agreement for Computational Linguistics, Computational Linguistics 2008, 555–596.

Fowler et al, Network Analysis and the Law: Measuring the Legal Importance of Precedents at the U.S. Supreme Court, Political Analysis 2007, 234–346.

Geist, Using Citation Analysis Techniques for Computer-Assisted Legal Research in Continental Jurisdictions, 01.05.2009. Available at SSRN: http://ssrn.com/abstract=1397674.

Kříž et al, Statistical Recognition of References in Czech Court Decisions, in *Gelbukh/Espinoza/Galicia-Haro* (Ed.), Human-Inspired Computing and Its Applications. DOI: 10.1007/978-3-319-13647-9-6 (Lecture Notes in Computer Science 2014), 51–61.

Opijnen, Canonicalizing Complex Case Law Citations, in *Winkels* (Ed.), Legal Knowledge and Information Systems (Frontiers in Artificial Intelligence and Application 2010), 97–106.

Post/Eisen, How Long is the Coastline of the Law? Thoughts on the Fractal Nature of Legal Systems, The Journal of Legal Studies 2000, 545–584.

Tusk, No-Citation Rules as a Prior Restraint on Attorney Speech, Columbia Law Review 2003, 1202–1235.

Widdison, New Perspectives in Legal Information Retrieval, International Journal of Law and Information Technology 2002, 41–70.

Winkels/Ruyter, Survival of the Fittest: Network Analysis of Dutch Supreme Court Cases, in *Palmirani et al* (Ed.), AI Approaches to the Complexity of Legal Systems. Models and Ethical Challenges for Legal Systems, Legal Language and Legal Ontologies, Argumentation and Software Agents. DOI: 10.1007/978-3-642-35731-2-7 (Lecture Notes in Computer Science 2011), 106–115.

Wyner/Peters/Katz, A Case Study on Legal Case Annotation, in *Ashley* (Ed.), Legal Knowledge and Information Systems (Frontiers in Artificial Intelligence and Applications 2013), 165–174.

[15] Similar to Opijnen, Canonicalizing Complex Case Law Citations, in *Winkels* (Ed.), Legal Knowledge and Information Systems (Frontiers in Artificial Intelligence and Application 2010), 97–106.

[16] See Kříž et al, Statistical Recognition of References in Czech Court Decisions, 51–61.

[17] See Geist, Using Citation Analysis Techniques for Computer-Assisted Legal Research in Continental Jurisdictions, 1 May 2009.