Legal Knowledge and Information Systems
A. Wyner and G. Casini (Eds.)

doi:10.3233/978-1-61499-838-9-177

Toward Linking Heterogenous References in Czech Court Decisions to Content

Jakub HARAŠTA [a,1], Jaromír ŠAVELKA [b]
[a] *Institute of Law and Technology, Faculty of Law, Masaryk University, Czechia*
[b] *Graduate Student, Intelligent Systems Program, University of Pittsburgh, USA*

Abstract. In this paper we present initial results from our effort to automatically detect references in decisions of the courts in the Czech Republic and link these references to their content. We focus on references to case-law and legal literature. To deal with wide variety in how references are expressed we use a novel distributed approach to reference recognition. Instead of attempting to recognize the references as a whole we focus on their lower level constituents. We assembled a corpus of 350 decisions and annotated it with more than 50,000 annotations corresponding to different reference constituents. Here we present our first attempt to detect these constituents automatically.

Keywords. case law analysis, reference recognition, conditional random fields, information extraction

1. Introduction and Challenge

Information extraction from unstructured (textual) data such as court decisions is challenging. References to other documents provide a rich set of information that could be useful in many practical applications, especially in legal information retrieval (IR). In our work we assess the possibility of recognizing the references automatically. We focus on references to case-law and scholarly literature. We intentionally leave aside references to statutory law and regulations because these appear to be quite uniform and significantly less challenging. In addition to detecting references we explore the possibility of detecting the piece of content they relate to in a decision (often quotation or a paraphrase from the referred document).

In the Czech Republic the courts do not observe a single citation standard such as the Bluebook in the United States. Instead there are multiple court-specific standards. On top of that the standards are not strictly enforced and they are subject to changes. Consider the following example of three different references that all refer to the same decision:[2]

> Decision of Constitutional Court of the Czech Republic Pl. ÚS 4/94 published in Collection of Decisions and Decrees of the Constitutional Court of the Czech Republic, year 1994, no. 46.

> Docket no. Pl. ÚS 4/94, Collection of decisions, volume 2, decision no. 46, published as no. 214/1994 Sb.

[1] Corresponding Author: jakub.harasta@law.muni.cz.

[2] All of the references are referring to an important decision of the Constitutional court that established test of proportionality within the Czech law. All references are translated to English.

Decision of October 12, 1994, docket no. Pl. ÚS 4/94, N 46/2 SbNU 57, 214/1994 Sb.

As can be seen the identifiers include docket numbers, vendor-specific identifiers, court reports or ID under which a case is listed in a specific legal information database.

2. Related Work

Significant amount of work was done in the area of reference recognition for purpose of bringing references under a set of common standards (for use in Italian legislation see [6]) or to account for multiple variants of the same reference and vendor-specific identifiers (see [5]). Both [6] and [5] are based on use of regular expressions. Language specific (Czech) work in [2] focused on detecting and classifying references to other court decisions and acts.

Our work is also focused on the content carried by a reference. Content of a reference is usually ignored; with exception of [8], [9], and [7]. [8] allows to determine which sentences near a reference are the best ones to represent the Reason for Citing. [9] uses metadata obtained by [8] that allow to explore so called Reason for Citing to create semantic-based network. [7] uses manual annotation with subsequent automated reference recognition and detection of topics of paragraphs using GATE framework [1].

3. Task

3.1. Specification

Because of the lack of a single citation standard we decided to understand references as consisting of smaller units. The smaller units are more uniform and therefore better suited for automatic detection. Some references may contain many of these units whereas other references may only have some of them. The units may appear in almost any order within a single reference.

For references to case-law the following constituents were identified:

- **c:id** - a unique court decision identifier,
- **c:court** - the court that issued the referred decision,
- **c:date** - the date on which the decision was issued,
- **c:type** - the type of the decision (e.g., decision, decree, opinion).

References to scholarly literature consist of these elements:

- **l:title** - the title of the referred work,
- **l:author** - the author or multiple authors of the referred work,
- **l:other** - other information of interest, such as place or year of publication.

Both types of references may also contain the following elements:

- **POI** - a pointer to a specific place in the decision or literary work (e.g., a page),
- **content** - the content associated with the reference (e.g., quotation, paraphrase).

References can also be expressed implicitly. In this way the courts usually refer decisions or scholarly literature that have been referred earlier in the decision. Since this occurs quite often we have created a special **implicit** constituent.

It is possible to encounter two tests for assessing impartiality of judge even in the case-law of *European Court of Human Rights*: subjective test stemming from personal convictions of judge deciding given case, objective test following whether sufficient assurances exist to exclude any legitimate doubt (comp. decision in **case of Saraiva de Carvalho v. Portugal**, decided on April 22, 1994, **application no. 15651/89**, eventually **Gautrin and others v. France**, decided on May 20, 1998, **application no. 21257/93**). In this context, it is worth mentioning that the Constitutional court noted that procedure of exclusion of judge from deciding upon case is one of the procedural guarantees of impartiality of court. When assessing whether the objective aspect of doubt on impartiality is present, even appearance may play a role in this regard [eg. decision of the *European Court of Human Rights* **Piersack v. Belgium**, decided on October 1, 1982, **application no. 8692/79**, §30; compare also decision **Wettstein v. Switzerland** decided on December 21, 2000, **application no. 33958/96**, §42 – 44, and decision of the *Constitutional Court* **docket no. III. ÚS 441/04** decided on January 12, 2005, published as **N 6/36 SbNU 53**].

Figure 1. Sample of annotated decision. Types of annotations are as follow: *c:court*, content, **c:id**, c:type, c:date, POI

	c:id	c:type	c:date	c:court	l:author	l:title	l:other	POI	implicit	content
annotations count	12043	5964	5449	4305	3426	2406	2609	3760	1202	10129
average per doc	34.41	17.04	15.57	12.30	9.79	6.87	7.45	10.74	3.43	28.94
gold count	6237	2992	2687	2236	1863	1176	1251	1854	483	4903
average gold per doc	17.82	8.55	7.68	6.39	5.32	3.36	3.57	5.30	1.38	14.01
strict agreement (inter)	70.62	80.72	86.27	73.61	73.88	50.25	44.92	73.56	27.62	26.51
overlap agreement (inter)	81.08	84.10	88.31	77.03	83.22	69.12	70.37	80.45	38.60	59.54
agreement (gold)	80.36	87.58	91.22	82.74	81.88	57.61	59.06	79.25	41.64	32.96

Table 1. Summary statistics of the data set.

3.2. Data Set

The data set consists of 350 decisions of the top-tier courts in the Czech Republic (160 Supreme Court, 115 Supreme Administrative Court, 75 Constitutional Court).[3] The shortest decision has 4,746 characters whereas the longest decision has 537,470 characters (average 36,148.68).

Decisions were annotated by thirteen annotators who were paid for their work. The annotators were trained to follow the annotation manual by means of dummy runs (i.e., annotation of documents that are not included in the data set). To ensure high quality of the resulting gold data set the three most knowledgeable annotators were appointed curators of the data set. Each document was then further processed by one of the curators. A curator could not be assigned a document that he himself annotated. The goal of the curators was to evaluate correctness of each annotation and to fill-in missing annotations. The result of their work is the gold data set.

The annotators generated 51,293 annotations (i.e., approximately 146.6 annotations per document). The detailed counts are shown in the first two rows of Table 1. The numbers correspond to all the annotations where each document was processed by two annotators. The second (gold count) and the third (avg gold per doc) rows of Table 1 provide details of the gold data set created by the curators. These entries do not contain duplicate annotations as opposed to the first two rows.

[3]The decisions were downloaded from publicly available online databases with exception of 8 cases. These were unavailable from public database of respective court and were retrieved from commercial information systems.

We report three types of inter-annotator agreement in the bottom three rows of Table 1. The strict agreement is the percentage of the annotations where the annotators agree exactly (i.e., the start and end character offsets are the same). The overlap agreement relaxes the exact matching condition—it is sufficient if the two annotations overlap by at least one character. The agreement (gold) reports the percentage of the annotations that were evaluated as correct by the curators.

3.3. Detecting Reference Constituents Automatically

We attempted to recognize the constituents of references automatically. This corresponds to detecting text spans representing the types described in Section 3.1 and summarized in Table 1. As a prediction model we use conditional random fields (CRF).[4] A CRF is a random field model that is globally conditioned on an observation sequence O. The states of the model correspond to event labels E. We use a first-order CRF in our experiments (observation O_i is associated with E_i). [3,4] We train a CRF model for each of the 10 labels. Although this is certainly suboptimal, we use the same training strategy and features for all the models. We reserve fine-tuning of models for future work.

In tokenization we consider an individual token to be any consecutive sequence of either letters, numbers or whitespace. Each character that does not belong to any of these constitutes a single token. Each of the tokens is then a data point in a sequence a CRF model operates on. Each token is represented by a small set of relatively simple features. Specifically, the set includes:

- *position* – position of a token within a document.
- *lower* – a token in lower case.
- *stem* and *aggressive stem* – two types of token stems.[5]
- *sig* – a feature representing a signature of a token.
- *length* – token's length.
- *islower* – true if all the token characters are in lower case.
- *isupper* – true if all the token characters are in upper case.
- *istitle* – true if only the first of the token characters is in upper case.
- *isdigit* – true if all the token characters are digits.
- *isspace* – true if all the token characters are whitespace.

For each token we also include *lower*, *stem* and *aggressive stem*, *sig*, *islower*, *isupper*, *istitle*, *isdigit*, and *isspace* features from the five preceding and five following tokens. If one of these tokens falls beyond the document boundaries we signal this by including *BOS* (beginning of sequence) and *EOS* (end of sequence) features.

4. Results and Discussion

4.1. Results

To evaluate the performance we use a 10-fold cross-validation. Table 2 summarizes the results of the experiments. The first two rows report the number (and average per docu-

[4] We use the CRFSuite which is available at `www.chokkan.org/software/crfsuite/`

[5] A stemmer for Czech implemented in Python by Luís Gomes was used for stemming. The stemmer is available at `http://research.variancia.com/czech_stemmer/`

	c:id	c:type	c:date	c:court	l:author	l:title	l:other	POI	implicit
predicted count	5891	2967	3001	1936	1454	786	909	1779	158
average per doc	16.83	8.48	8.57	5.53	4.15	2.25	2.60	5.08	0.45
strict agreement (gold)	65.22	75.95	75.00	56.81	64.88	43.22	49.07	70.02	6.86
overlap agreement (gold)	70.86	78.40	75.14	57.86	74.77	58.31	61.20	75.03	21.53

Table 2. Results of automatic detection of reference constituents

ment) of annotations of each type that were automatically generated for the whole data set. The third row reports the agreement with the gold standard where the equality of annotations is measured strictly (i.e., the start and end offsets both need to match exactly). The fourth row reports the agreement where the annotations to be considered equal just need to overlap by at least one character.

As one would expect the performance of the models correlates to the performance of human annotators. In case of the elements constituting references to literary works the performance of our models matches the humans. This is almost the case of the POI element as well.

4.2. Result analysis

The counts of detected elements closely correlate with the counts of annotations created by humans (compare first row of Table 2 with the third row of Table 1). The only exception is the implicit element which has been automatically recognized in only 158 cases whereas the humans found 483 instances of this element. This clearly suggests that our models struggled to recognize the implicit type of reference.

Overall the performance of the trained models was decent. It appears that in case of the l:author, l:title, l:other and POI the models almost matched human performance (compare the fourth row of Table 2 with the seventh row of Table 1). The models trained to recognize the c:id, c:type, c:date, and c:court constituents perform somewhat worse than human annotators.

It may be quite surprising to see the relatively low performance for elements such as c:date or c:court. Indeed the task of detecting dates or court mentions should not be that challenging. However, our models need to deal with a situation where we detect only certain dates and court mentions—only those that are part of references. Therefore the models may get confused by seeing mentions that appear to be of the relevant types but they are not. This problem could be mitigated in later stages where the constituents would be linked together to form references.

4.3. Grouping Constituents into References and Linking References to Content

Eventually we would like to use the automatically recognized constituents as building blocks for references. Grouping the constituents presents an interesting research problem in its own right. We already have the annotations that group the elements into individual references. A reference is essentially a set of a number of constituents. The detailed statistics on the references are reported in Table 3.

Finally we would like to connect each reference to a content element. The evaluation of the human annotator's effort is summarized in Table 4. The top two rows are the same types of measures as the ones used for references. The only difference is that the content is used as an additional constituent in all the four reference types.

	c:ref (expl)	c:ref (impl)	l:ref (expl)	l:ref (impl)
reference count	7570	1040	2497	293
average per doc	21.63	2.97	7.13	0.84
gold count	3753	429	1228	122
average gold per doc	10.72	1.23	3.51	0.35
strict agreement (inter)	45.96	20.66	33.23	22.58
overlap agreement (inter)	88.60	35.00	85.18	50.51

Table 3. Statistics of references in the data set

	c:ref (expl)	c:ref (impl)	l:ref (expl)	l:ref (impl)
strict agreement (inter)	42.18	12.42	30.65	19.49
overlap agreement (inter)	89.05	35.67	86.30	51.19

Table 4. Agreement on references between human annotators

5. Conclusions and Future Work

We presented early results from our ongoing effort to automatically detect references in Czech case-law. With regard to future work, our annotation task also involved marking polarity of references, which was not discussed in this paper. As such, it needs to undergo similar evaluation as other types of annotation. Another partial task for automation is the creation of whole references from lower level constituents. Moreover, successful statistical recognition is only a single step in our research. Ultimate goal is to allow for creation of citation network of the Czech top-tier court decisions and leverage this network to investigate the concept of 'importance' of court decisions and scholarly works.

6. Acknowledgements

The authors would like to thank annotators and editors – František Kasl, Adéla Kotková, Pavel Loutocký, Jakub Míšek, Daniela Procházková, Helena Pullmannová, Petr Semenišin, Nikola Šimková, Tamara Šejnová, Michal Vosinek, Lucie Zavadilová, and Jan Zibner. JH gratefully acknowledges the support from the Czech Science Foundation under grant no. GA17-20645S.

References

[1] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, Valentin Tablan, GATE: an Architecture for Development of Robust HLT Applications. Proceedings of the 40th Annual ACL meeting, pp. 168-175.
[2] Vincent Kríž, Barbora Hladká, Jan Dědek and Martin Nečaský. Statistical Recognition of References in Czech Court Decisions. Proceedings of MICAI 2014, Part I, pp. 51–61.
[3] John Lafferty, Andrew McCallum, Fernando Pereira, and others. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning*, ICML, Vol. 1. 282–289.
[4] Naoaki Okazaki. *CRFsuite: a fast implementation of Conditional Random Fields*. (2007).
[5] Marc van Opijnen. Canonicalizing Complex Case Law Citations. JURIX 2010, pp. 97–106.
[6] Monica Palmirani, Raffaella Brighi and Matteo Massini. Automated Extraction of Normative References in Legal Texts. Proceedings of ICAIL 2003, pp. 105–106.
[7] Yannis Panagis and Urška Šadl. The Force of EU Case Law: A multidimensional Study of Case citations. Proceedings of JURIX 2015, pp. 71–80.
[8] Patent US6856988. Automated system and method for generating reasons that a court case is cited.
[9] Paul Zhang and Lavanya Koppaka. Semantics-Based Legal Citation Network. Proceedings of ICAIL 2007, pp. 123–130.