Computational Linguistic Support for Legal Ontology Construction

Stephan Walter
Dept. of Computational Linguistics and Phonetics
Universität des Saarlandes
66123 Saarbrücken, Germany
stwa@coli.uni-sb.de

Manfred Pinkal
Dept. of Computational Linguistics and Phonetics
Universität des Saarlandes
66123 Saarbrücken, Germany
pinkal@coli.uni-sb.de

ABSTRACT

The potential of knowledge-based technological support for work in the legal domain has become widely recognized in recent time. This paper argues for an approach that is meant to complement existing work on top-level, statute-based ontologies for the law by extracting conceptual knowledge from verdict texts. Starting from a corpus-based survey on the definition and modification of concepts in German court verdicts we are currently working out a detailed description of the linguistic structures and semantic mechanisms involved. Based on this we want to develop tools that automatically recognize and process portions of verdicts that contain definitions. The resulting system will support the corpus-based extension of (hand-made) top-level legal ontologies and the update of legal terminological knowledge bases with regard to new jurisdiction. The research described in the paper is conducted at the Saarbrücken computational linguistics (CL) group within the project CORTE (computational linguistic methods for legal terminology).[1]

1. Definitions in Verdicts

Most ontologies presently used in legal information management formalize high-level expert knowledge and conceptual knowledge contained within definitions occuring in statute texts. The scientific discussion has so far focused on general characteristics of statute-based ontologies, and on the specification of a core ontology for the law (e.g. [5], [6], [3]).

But the definitions contained in statutes never fully specify how the relevant concepts are to be applied when deciding cases. First, definitions in laws texts are formulated in natural language, and natural language expressions are very often inherently vague and ambiguous. So even if a natural language definition may effectively decide the applicability of a concept for a large set of clear cases, there will almost inevitably also be a lot of borderline cases, where the applicability of the terms that define the concept is itself unclear. Second, reality is complex and constantly changing. At the time a definition is given, it may simply be impossible to predict all the complexities and developments that may occur in the domain of the regulation. Definitions within verdicts serve e.g. to adapt the degree of precision of a concept to the needs of the respective context, or to bridge the gap between what is available and what is required by rendering given concepts definitely applicable (or definitely inapplicable) to unanticipated cases. Because it provides this possibility, the open-textured character (cf. [4]) of many legal concepts is often seen as an essential feature of any functioning legal system.

Texts that document court decisions are therefore in large parts devoted to pinning down whether certain concepts apply or not. In this process the fundamental knowledge defined in statutes is brought to bear on reality by working out a body of auxiliary knowledge that remains binding beyond the case at hand. However comparably little is included in present ontologies about how concepts are used and modified within verdicts, and what auxiliary concepts are additionally created.

2. The Need for NLP support

A first survey that we conducted based on a random selection of about 150 definitions in German court verdicts has shown a broad variety of formulations that German judges employ when introducing and modifying concepts, as well as a high degree of syntactic complexity. No simple account, for instance in terms of keyword spotting or pattern matching, will suffice to extract the relevant information from all such contexts. Deeper linguistic analyses of the constituent sentences are needed for this task. Moreover much of the encountered linguistic variation is actually functional. Quite often specific formulations of definitions are not just picked due to stylistic considerations. Rather their choice is closely tied to certain purposes, for instance expressing certain kinds of definitions (e.g. by modifying existing usage or by giving necessary and sufficient conditions), or degrees of commitment to a definition (e.g. only for typical cases). Such linguistic cues need to be found and taken into account in the automatic extraction of concept-related information. But even their empirical description is still a largely open task (which will itself require linguistic analysis of a large set of corpus examples).

Natural Language Processing (NLP) techniques will be of great use in approaching these topics. For any serious large-scale application that attempts to keep track of concept introduction and modification in verdicts, such automatic support will in fact be

ICAIL '05, June 6-11, 2005, Bologna, Italy.

[1] funded by the DFG (German Science Foundation), PI 154/10-1

indispensable if only due to the enormous and steadily increasing amount of text that has to be considered. Yet little research has been conducted so far on the use of NLP in this area, and the tools needed are not readily available. In principle large scale grammars exist as well as sophisticated parsing technology. But efficiency aside, the massive ambiguity detected by such systems poses a major problem for their practical application. Within such systems, almost any sentence is given a variety of theoretically possible readings. At some point the intended reading has to be selected, and there is no simple way of doing this automatically.

To avoid processing difficulties some systems trade precision for robustness. Such systems do not build on full linguistic analysis. Instead they often use statistical methods and normally do not pursue many alternative readings. Results may therefore be comparably unreliable and essential information for further processing steps may be missing. Ambiguities for instance may simply be overlooked. Given the need for reliably accessible rich structural information together with the high linguistic complexity generally encountered in the domain of legal text, an approach is needed that deals with the ambiguity problem without losing too much.

3. The CORTE project

This section reports our current work within the project CORTE (computational linguistic methods for legal terminology). For our work, we have access to a corpus of more than 8 million German legal documents provided by *juris* GmbH, Saarbrücken.

In order to analyze these documents grammatically, we build on a semantically-oriented parsing system that has been developed in the COLLATE project (*Computational Linguistics and Language Technology for Real Life Applications*, funded by the German Ministry for Education and Research) at the Saarbrücken CL group (cf. [2], initially it was applied to newspaper texts). The system balances depth of linguistic analysis with robustness of the analysis process and is therefore able to provide relatively detailed linguistic information for large amounts of text. To deal with the problem of ambiguity it makes use of syntactic underspecification. Under certain conditions it commits only to the established common parts of alternative syntactic analyses. That way later processing steps can access at least partial information without having to decide for one syntactic reading. Most important for our purposes however is the fact that the system is semantically-oriented. It not only analyzes the grammatical structure of the input, but also provides an abstract representation of its meaning (a so-called partially resolved dependency structure or PREDS). Active and passive sentences for instance receive identical representations, so that their common semantic content becomes accessible for further processing.

We are presently adapting the PREDS-parsing system to the domain of legal documents. At the same time we are working on the data-driven exploration of the phenomenon of definitions (in the widest sense) in German legal text. Starting off from a collection of definitions compiled relying on legal expert knowledge, we have devised an annotation scheme for marking up the functional parts of these definitions. This scheme will be extended to encode information regarding external relations such as rhetorical and argumentative function of definitions and citation structure, and it will be applied in the collection of further data. At the same time we are working out detailed linguistic analyses of definition instances and we relate them to the analyses in functional terms. Our aim in this is to develop a taxonomy of definition types according to semantic functions and syntactic realization.

The syntacto-semantic information made accessible by the PREDS system will facilitate the automatic recognition and extraction of definitions by providing an additional level of structure besides the syntactic surface. Extracted definitions can then be used to validate our taxonomy. More importantly, the information contained in the PREDS constructed will be used to organize the collected extraction results within a semi-structured knowledge base. In particular it will serve to automatically segment and classify extracted definitions according to the taxonomy we are developing, based on linguistic cues. The resulting knowledge base will contain the extracted text passages along with rich additional information that allows the user to navigate through the collected definitions according to his needs, e.g. sorted by concept defined, grouped by type of definition, or following citations.

A very promising (yet also more ambitious) topic we are going to address is that of using the information provided by the PREDS-based definition extraction system to actually update and enlarge existing formalized ontologies. Languages based on description logics (DL) (cf. [1]) have emerged as the standard framework for the specification of such formalized ontologies. The central question to be pursued is therefore how to model the semantic effect of definitions within this formalism. Moreover, due to the organization of DL knowledge bases around atomic concepts that are incrementally characterized semantically by adding constraints, the framework is especially interesting for the modeling of "open-texture", i.e. underdefined or vague concepts and their incremental specification. Building on a linguistically well-founded understanding of definitions together with automatic definition extraction methods, it will be possible to approach this topic empirically.

4. REFERENCES

[1] Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., and Patel-Schneider, P.F. (Eds.) *The Description Logics Handbook: Theory, Implementations, and Applications.* Cambridge University Press, 2003.

[2] Braun, C. Parsing German text for syntacto-semantic structures. In *Prospects and Advances in the Syntax/Semantics Interface*, Lorraine-Saarland Workshop Series, Nancy, France, 2003, 99-102.

[3] Breuker, J. and Hoekstra, R. Core concepts of law: taking common-sense seriously. In *Proceedings of Formal Ontologies in Information Systems (FOIS-2004).* IOS-Press, Amsterdam, The Netherlands, 2004.

[4] Hart, H.L.A. *The concept of Law*. Oxford University Press, London, 1961.

[5] Valente, A. *Legal Knowledge Engineering: A Modelling Approach*. IOS Press, Amsterdam, The Netherlands, 1995.

[6] van Kralingen, R. *Frame-based Conceptual Models of Statute Law.* Computer / Law Series. Kluwer Law International, The Hague, The Netherlands, 1995.