Learning and Verification of Legal Ontologies by Means of Conceptual Analysis

Erich Schweighofer

Centre for Computers and Law
DEICL/AVR, Faculty of Law, University of Vienna
Schottenbastei 10-16/2/5, 1010 Wien, Austria
Erich.Schweighofer@univie.ac.at
rechtsinformatik.univie.ac.at

Abstract. A combination of intellectual input, NLP tools and appropriate ontological representation may overcome the existing bottleneck of legal knowledge acquisition of legal ontologies. Such semi-automatic tools rely on easily available input, extensive iterative semiautomatic checking and refining of this knowledge. Preliminary results using the tools of SOM/GHSOM, KONTERM and GATE show the feasibility of this method. However, it remains to be seen if a sufficient number of legal writers will adapt to this new workbench.

Keywords: Legal ontologies, text analysis, learning, conceptual indexing

1 Introduction

In law, indexing was and is still a very important tool for coping with the vast body of legal materials. Since the advent of information retrieval, legal full text search has been added to the methods of legal research. However, an index of concepts or legal sources is still considered as the best access to the sequential structure of handbooks, textbooks or collections of materials. Such indices may be also used for the production of summaries of cases (head notes) identifying the important parts of court decisions. Huge reference systems on legal materials also exist either based on citations (e.g. the Austrian index [1] or thesauri (e.g. the Swiss thesaurus [2]).

In previous papers, we have argued for the creation of a dynamic electronic legal commentary [3]. Differently to a handbook (or commentary) as the most advanced traditional form of explicit knowledge representation, the dynamic electronic legal commentary is based on legal ontologies as major knowledge base and integrates semiautomatic means of semantic indexing. A complete knowledge representation of a legal domain requires many resources, either that of legal experts or those of ICT. Given the dynamic change in law, means of semiautomatic creation and verification of ontologies are thus highly needed for cheaper and faster efforts of compilation and analysis.

Such an approach consist in “working together” between legal experts and ontological tools. Only experts can easily produce the extensive input and check the vast semiautomatic output. However, up to now, legal writers still prefer intellectual analysis without semiautomatic means.

Legal ontologies should be the core of such a knowledge base. However, these ontologies are either too broad and shallow (e.g. LOIS and DALOS) or too small and deep (e.g. LRI Core) in order to meet the standards of semantic indexing. Thus, we propose the development and refinement of such an ontology by means of conceptual analysis.

The remainder of this paper is organized as follows: section 2 describes related work, section 3 gives an overview on the method. In section 4, the status of implementation and problems are discussed. Section 5 contains conclusions and future work.

2 Related work

The main components of legal knowledge are the legal retrieval system (or legal information system) as a huge text corpus with a (mostly) textual representation of the legal order and meta knowledge about the text corpus. Computationally speaking, meaningful semantic indexing is linked to a legal text corpus. Such indexing exists in legal brains, legal books but also legal knowledge bases. Legal structuring as such is done by lawyers, in their minds, and is presented and made explicit in their argumentations and writings. As a product of this process, a legal commentary is considered as the highest level of this endeavour.

The semantic web can be considered as an extension to the current web in providing a common framework that allows data to be shared and reused [4]. Semantic search may also improve disappointing results of present legal information retrieval [5].

Thesauri (or legal dictionaries) are getting more importance now as a traditional tool for representation of knowledge about legal language use. A thesaurus for indexing contains a list of every important term in a given domain of knowledge and a set of related terms for each of these terms [6]. A lexical ontology builds up from this basis with works on glossaries and dictionaries, extends the relations and makes this knowledge computer-usable in order to allow intelligent applications. More advanced representations may formalize complex legal rules and conceptual structures.

Ontologies [7] constitute an explicit formal specification of a common conceptualization with term hierarchies, relations and attributes that makes it possible to reuse this knowledge for automated applications.

Legal knowledge representation remains the most important and challenging task of legal ontologies [8]. The frame-based ontology FBO of [9] and [10] as well as the functional ontology FOLaw [11] can still be considered as important work on formalisation. More advanced work exists in the development of a core legal ontology called LRI-Core [12] or the impressive standard for the development of a legal ontology called LKIF Core Ontology (Legal Knowledge Interchange Format) [13].

Quite many projects were focused on conceptual information retrieval (see e.g. Iuriservice [14], LOIS (Lexical Ontologies for legal Information Serving) project [15]. The Legal Taxonomy Syllabus [16], DALOS [17] or the Comprehensive Legal Ontology (CLO) [3].

Such powerful ontologies can only be built if resources of robust NLP and machine–learning are exploited. We share the view of [18] that such technologies are "the key to any attempt to successfully face what we termed the acquisition paradox". However, we argue that the quite huge experience of semi-automatic text analysis and conceptual indexing in law (see e.g. the projects KONTERM/LabelSOM/GHSOM [19, 20], SALOMON [21], FLEXICON [22], SMILE [23], or Support Vector Machines [24]) should be taken into account and reused.

The automated linking of documents constitutes the most advanced work in semantic indexing (e.g. AustLII [25], CiteSeer [26]). It has to be noted that the task is easier due to more formalized language and a controlled vocabulary.

3 Idea and Method

This method adds the idea of an ontological workbench for the lawyer to the already existing tools. A combination of expert knowledge, easy access to intellectual input and the use of semiautomatic refinement empowers this method for contributing to solve the "scaling up"-problem. It should be noted that legal writing consists to a large degree in structuring, refining and representing the content of legal text corpora in an abridged and more abstract way. So far, legal writers are still not much in favor of semi-automatic analysis as a tool for improving efficiency to this very time-consuming process.

Our method combines intellectual input, corpus-based methods of verification and refinement as well as text categorization, conceptual analysis and text extraction. Using our expertise of as a lawyer with extensive practice and an academic in legal informatics, we are developing a workbench for other lawyers for NLP techniques and text analysis.

Due to this corpus-based approach on legal analysis, all (tentative) results have to be checked against a legal text corpus. In our case, the millions of documents of the Austrian legal retrieval system RIS (Rechtsinformationssystem des Bundes) [27] and related private databases RDB and LexisNexis are used for improvement, refinement and verification of the ontological representation.

As a start, we give a sketchy picture for a sufficient granularity of an ontological representation of a jurisdiction: about 10 000 thesaurus entries, 5 000 citations, up to 200 document types, a classification structure (e.g. RIS classification or EUR-Lex classification codes), 100 text extraction and summarization rules, and, as representation of the dynamic legal electronic commentary, an indefinite number of concepts, rules and procedures. It is an enormous body of knowledge and it should be clear that a stepwise approach has to be taken, e.g. a start with descriptor or citation lists that will later be transformed into ontological representations. The final version, the dynamic legal electronic commentary, will take some time to finish.

The target – the ontological representation – should consist of the following meta data that has to be maintained in a database with different types of knowledge units (or tables):

Thesaurus entries: header, definition (with sources), examples (with sources), relations (synonym, homonym, polysem, hyponym, hyperonym, antonym etc.), classification, other information.

Citations: header, identification (abbreviation or number), synonyms, classification, author, other information.

Document types: header, identification (abbreviation), use, format, other information.

Classification: header, code, definition, relations, other information.

Extraction and summarization rules: header, rule, definition, relations, other information.

Concepts: header, definition (with sources), related thesaurus entries and citations, relations (synonym, homonym, polysem, hyponym, hyperonym, antonym etc.), classification, legal conceptual structure (ontological model), other information.

Rules: header, quasi-logical expression, source, type, classification, legal conceptual structure (ontological model), other information.

Procedures: header, decision tree, source, type, classification, legal conceptual structure (ontological model), other information.

For the start, we have collected available information on legal meta knowledge from traditional sources. Concept lists were taken from the table of contents and indices of text books and commentaries. A quite complete citation list was provided by the Federal High Court of Administration and improved using the reference book. The list of document types reflects the present status of documents in the legal information system RIS. Text extraction rules were intellectually created by studying the linguistic styles and patterns of Austrian laws, judgments and literature. We took also advantage of the experience with the LOIS project. With this method, it was quite easy to achieve a sufficient but still rough representation of conceptual structure of the Austrian legal order. For easier re-use, this information was incorporated in a relational database. Data may still quite incomplete at the beginning but must be sufficient for semiautomatic analysis. An XML representation is also available for later incorporation in higher representations, e.g. the knowledge base of the dynamic electronic legal commentary.

As tools of semi-automatic analysis, we have implemented the modified GHSOM method of classification, the KONTERM method of conceptual analysis, and the GATE methods of ANNIE and JAPE [28].

The modified GHSOM method is based on the self-organising map, a general unsupervised tool for ordering high-dimensional data in such a way that alike input items are mapped close to each other. In order to use the self-organising map to explore text documents, we represent the various texts as the histogram of its words with a *TFxIDF* vector representation. The methods LabelSOM can properly describe the common similarities of the cluster. An extension to the SOM architecture, the GHSOM [20] can automatically represent the inherent hierarchical structure of the documents. An extension for legal purposes allows the manual refinement of vector weights of the documents with data enrichment tools. The produced output consists in structured maps of clusters with cluster descriptions. These descriptions were used for refinement of the thesaurus, in particular for completeness and for synonyms.

The KONTERM method [3] produces structured lists of term occurrences with a description of the various meanings. These representations were incorporated in the description of homonyms and polysems of thesaurus entries.

The GATE JAPE tool (Regular Expressions Over Annotations) is implemented for a similar purpose. It is much more powerful in bigger text environments but does not allow so sophisticated representations of meanings as the KONTERM method.

The GATE ANNIE (A Nearly New Information Extraction System) tool supports a more detailed analysis: segmentation of documents (tokenizer), words, gazetteer, sentence splitter and semantic tagger.

These methods have one big advantage and two important disadvantages. The huge text corpus of materials can be explored and analyzed with much higher accuracy, speed and efficiency. All ontological concepts can be checked for meanings, definitions and relations in the legal information system. However, the semiautomatic output is quite voluminous and analysis takes some time. Further, it represents only an intermediate step in the process of analysis. It must be mentioned that these tools are by far not sufficiently adapted for a legal environment. Legal experts may refuse the use for the simple reason of an inconvenient interface. However, in the hands of a supportive and experienced expert, such tools prove to be very helpful and may substitute other research.

4 Implementation Details and Problems

The test environment consists of the Austrian legal order, its textual representation in the retrieval system RIS *(Rechtsinformationssystem des Bundes*, Austrian legal information system) and the "rough ontology" of thesaurus entries, citations and extraction rules.

This very "rough ontology" was checked and refined by selective document corpora analyzed with GHSOM, KONTERM and GATE tools. For easier checking of results, subfields like telecommunications law or state aid law were selected. The output was then used for extension and enlargement of the knowledge representation.

The work is still ongoing but some preliminary remarks can be made. The output improves very much the ontological representation. Further analytical work is much supported by ontological representation, faster browsing, reading and text extraction. However, the workload of checking the output remains significant.

Thus, such efforts of knowledge representation may only be justified if they support the production of other knowledge products like handbooks and commentaries. Only a symbiosis of these efforts may produce the required "scaling-up" of legal ontologies. It has to be noted that without the day-to-day input of legal authors the quality of knowledge representation may not be sufficient. Later, automated applications of legal reasoning can be also envisaged.

As a preliminary result, the problem of these methods is not its usability but its acceptance by legal authors. Refinement of methods and improved interface will play a decisive role and will be part of future research.

5 Conclusions and Future Work

Next steps for a dynamic electronic legal commentary require semantic indexing of legal information systems and extraction of ontological information of these huge data warehouses. A combination of intellectual input, NLP tools and appropriate ontological representation may overcome the existing bottleneck of legal knowledge acquisition of legal ontologies. Preliminary results based on the test environment of Austrian law using the tools of SOM/GHSOM, KONTERM and GATE show the feasibility of this method. However, refinement and adaptation still require important personal resources in practice. Success of this method will depend on the willingness of legal writers to modify working habits and include this approach in their methods of legal structural analysis.

References

1. Index 2006, Rechtsprechung und Schrifttum, Jahresübersicht 2006. Band 59, Begründet von Franz Hohenecker. Manz, Wien (2007)
2. Jurivoc. Dreisprachiger Thesaurus des Schweizerischen Bundesgerichts. `http://www.bger.ch/de/index/juridiction/jurisdiction-inherit-template/jurisdiction-jurivoc-home.htm` (2009).
3. Schweighofer, E.: Computing Law: From Legal Information Systems to Dynamic Legal Electronic Commentaries, In: Magnusson Sjöberg, C., Wahlgren, P. (eds.), Festskrift till Peter Seipel pp. 569-588. Norsteds Juridik AB, Stockholm (2006)
4. Berners-Lee, T. et al.: The Semantic Web. Scientific American Vol. 284, No. 5, 34-53 (2001)
5. Blair, D. C., Maron, M. E.: An Evaluation of Retrieval Effectiveness for a Full-text Document-retrieval System. Comm ACM, Vol. 28, 289-299 (1985)
6. ISO: Documentation. Guidelines for the establishment and development of monolingual thesauri, ISO 2788 (1986)
7. Gruber, T.R.: A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition vol. 5/2, 199-220 (1993)
8. Bench-Capon, T.J.M., Visser, P.R.S.: Ontologies in Legal Information Systems: The Need for Explicit Specifications of Domain Conceptualisations. In: Proceedings of the 6th ICAIL, pp. 132-141. ACM Press, New York, NY (1997)
9. Kralingen, R.W. van: Frame-based Conceptual Models of Staute Law. Ph.D. Thesis, University of Leiden, The Hague (1995)
10. Visser, P.R.S.: Knowledge Specification for Multiple Legal Tasks: A Case Study of the Interaction Problem in the Legal Domain. Computer Law Series Vol. 17, Kluwer Law International, The Hague (1995)
11. Valente, A.: Legal knowledge engineering: A modelling approach. IOS Press, Amsterdam (1995)
12. Breuker, J. and Hoekstra, R.: DIRECT: Ontology-based Discovery of Responsibility and Causality in Legal Case Descriptions. In: Proceedings of the 17th JURIX. IOS Press, Amsterdam et al. (2004)
13. Hoekstra, R., Breuker, J., De Bello, M., Boer, A.: The LKIF Core Ontology of Basic Legal Concepts. In: Casanovas, P., Biasiotti, M. A., Francesconi, E., Sagri, M. T. (eds.) Proceedings of LOAIT 07, II. Workshop on Legal Ontologies and Artificial Intelligence Techniques, pp. 43-64. `http://www.ittig.cnr.it/loait/LOAIT07-Proceedings.pdf` (2007)
14. Casellas, N., Casanovas, P., Vallbé, J.-J., Poblet, M., Blázquez, M., Contreras, J., López-Cobo, J.-M., Benjamins, R.: Semantic Enhancement for Legal Information Retrieval: IURISERVICE performance. In: Eleventh International Conference on Artificial Intelligence and Law, pp. 49-57. ACM Press, New York (2007).
15. Dini, L., Liebwald, D., Mommers, L., Peters, W., Schweighofer, E., Voermans, W.: LOIS Cross-lingual Legal Information Retrieval Using a WordNet Architecture. In: Proc Tenth Int Conf on Artificial Intelligence & Law, pp. 163-167. ACM Press, New York (2005).

16. Ajani, G., Lesmo, L., Boella, G., Mazzei, A., Rossi, P.: Terminological and Ontological Analysis of European Directives: multilinguism in Law. In: Eleventh International Conference on Artificial Intelligence and Law, pp. 43-48. ACM Press, New York (2007).
17. Francesconi, E., Spinosa, P., Tiscorina, D.: A linguistic-ontological support for multilingual legislative drafting: the DALOS Project. In: Casanovas, P., Biasiotti, M. A., Francesconi, E., Sagri, M. T. (eds.) Proceedings of LOAIT 07, II. Workshop on Legal Ontologies and Artificial Intelligence Techniques, pp. 103-112. http://www.ittig.cnr.it/loait/LOAIT07-Proceedings.pdf (2007)
18. Lenci, A., Montemagni, S., Pirrelli, V., Ventur, G.: NLP-based ontology learning from legal texts. A case Study, In: Casanovas, P., Biasiotti, M. A., Francesconi, E., Sagri, M. T. (eds.) Proceedings of LOAIT 07, II. Workshop on Legal Ontologies and Artificial Intelligence Techniques, pp. 103-112. http://www.ittig.cnr.it/loait/LOAIT07-Proceedings.pdf (2007)
19. Schweighofer, E.: Legal Knowledge Representation, Automatic Text Analysis in Public International and European Law. Kluwer Law International, The Hague (1999)
20. Schweighofer, E. et al.: Improvement of Vector Representation of Legal Documents with Legal Ontologies. In: Proceedings of the 5th BIS, Poznan University of Economics Press, Poznan (2002)
21. Moens, M.-F. et al.: Abstracting of Legal Cases: The SALOMON Experience. In: Proceedings of the 6th ICAIL pp. 114-122. ACM Press, New York (1997)
22. Smith, J.C. et al.: Artificial Intelligence and Legal Discourse: The Flexlaw Legal Text Management System. Artificial Intelligence and Law Vol. 3/1-2, 55-95 (1995).
23. Brüninghaus, S. and Ashley, K.D.: Improving the Representation of Legal Case Texts with Information extraction Methods. In: Proceedings of the 8th ICAIL pp. 42-51. ACM Press, New York (2001)
24. Gonçalves, T. and Quaresma, P.: Is linguistic information relevant for the classification of legal texts? In: Proceedings of the 10th ICAIL, pp. 168-176. ACM Press, New York (1995)
25. AustLII website. http://www.austlii.edu.au
26. CiteSeer website. http://citeseer.ist.psu.edu/cs
27. RIS website: http://www.ris.bka.gv.at
28. GATE (General Architecture for Text Engineering) Engineering) website. http://gate.ac.uk/