Cross-lingual legal information retrieval using a WordNet architecture

Luca Dini
CELI
dini@celi.it

Doris Liebwald
University of Vienna
doris.liebwald@univie.ac.at

Laurens Mommers
Leiden University
l.mommers@law.leidenuniv.nl

Wim Peters
University of Sheffield
w.peters@dcs.shef.ac.uk

Erich Schweighofer
University of Vienna
erich.schweighofer@univie.ac.at

Wim Voermans
Leiden University
w.j.m.voermans@law.leidenuniv.nl

ABSTRACT

The LOIS project encompasses the construction of a large, multi-lingual WordNet for cross-lingual information retrieval in the legal domain. In this article, we set out how a hybrid approach, featuring lexically and legally grounded conceptual representations, can fit the cross-lingual information retrieval needs of both legal professionals and laymen. With respect to the legally grounded part of this WordNet, we focus on the automatic extraction of legal definitions from European directives.

1. INTRODUCTION

In the EU-funded e-Content LOIS project (Lexical Ontologies for legal Information Sharing), a multi-lingual legal WordNet is built for the purpose of facilitating information retrieval. Legal Information Retrieval (IR) research has stressed the importance of legal knowledge systems being sufficiently able to interpret and handle the semantics of a database. Today, such semantic processing is missing in real-life legal information systems [9]. Legal databases are syntactically structured text archives with powerful search engines. One of the main deficiencies of these systems is the lack of efficient representations of semantic relationships between information needs and the information content of documents. This is especially a problem for cross-lingual information retrieval. In this case, lack of knowledge of a certain language may prevent users from formulating queries, and thus from finding relevant results.

Luuk Matthijssen ascertained four theoretical limitations of information retrieval [7]: (1) the fact that the index of a database only partially describes its information contents, (2) the imperfect description of an information need by the query formulation, (3) the rough heuristics and tight closed world assumption of the matching function, and (4) the presence of the conceptual gap: the discrepancy between users' views of the subject matter of the stored documents in the context of their professional setting and the reduced formal view on these subjects as presented by information retrieval systems. Legal practitioners have to translate their information need - which they have in mind in the form of legal concepts - into a query, which must be put in technical database terms.

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

Contrary to other academic disciplines (such as biology and genetics), taxonomies are rarely inherent in law. Legal vocabularies contain open-textured terms, they are inherently dynamic, and the norms in which legal terms are used, are syntactically ambiguous. This allows for contradictions to arise from judicial problem solving [5]. A legal 'language', consisting of a complex structure of concepts, forms an abstraction from the text corpus as represented in legal databases. Such legal structural knowledge does not only contain interpretations of the meaning of legal terms, but also shows the (supposed) logical and conceptual structure. Bridging the gap between legal text archives and legal structural knowledge is the principal task of studying the law, and the key challenge in legal information retrieval.

Building an 'interface' between syntactic legal databases and professional or lay users requires the extraction of structural knowledge, preferably by automatic means. At present, legal ontologies are considered to be the most promising way for formalizing such knowledge. Modelling knowledge by using ontologies or advanced thesauri enhances the ability to extract and exploit information from documents, by establishing explicit semantic links among related items. An ontology is an explicit formal specification of a common conceptualisation [4]. A formal definition of term hierarchies, relations and attributes (the explicit description of concepts in the legal domain) opens the way for implementations, such as information retrieval systems. Formalization is a difficult task: on the one hand, it must be sufficiently powerful with regard to knowledge representation, on the other hand it must offer functionalities for automation.

In this article, we will present a method for building a large semantically enriched, multi-lingual terminological database, following the WordNet framework. First, we discuss the theoretical obstacles and building blocks for cross-lingual concept representation, resulting in a model for that purpose (section 2). Subsequently, we describe the structure of the LOIS database, and the way in which legal definitions can be extracted to fill the corresponding part of the database (section 3). Finally, in section 4, we provide conclusions.

2. DEFINING LEGAL CONCEPTS

The composition of the LOIS WordNet presents challenges for both *defining* legal concepts and for *linking* concepts from those

legal systems. First, explicit definitions ought to be given to represent legal concepts in a certain language, and those definitions should be linked to each other in meaningful ways. Second, representations of comparable legal concepts from different languages should be linked to each other in order to create the possibility of cross-lingual information retrieval. So, for instance, the English legal term 'property' should be defined, and be linked to (hierarchically) related English legal terms (if any), and to comparable terms in other languages (again, if any). In a normal (non-legal) translation, the English term 'property' would probably be translated into the Dutch term 'eigendom'. However, seen from a legal viewpoint, the concepts these terms refer to, only resemble each other to a very limited degree.

Legal terms often have (partly) explicit definitions that are authoritative, because they are introduced by a legislator. At least in this respect, legal terms differ from lexical terms (with corresponding lexical definitions). Such lexical definitions are systematically listed in dictionaries. They are meant primarily to make people grasp the meaning of a term. Legal definitions, however, are authoritative fixed meaning. According to Eijlander and Voermans there are basically four techniques to define a term in a legislative text [1]: (1) not defining a term at all, thereby leaving a term's lexical meaning intact; (2) defining a term by its context, which implicitly specifies a term's meaning; (3) defining a term by a reference, providing a link to an explicit definition in a different place; and (4) defining a term by providing an explicit definition, which can have different forms (generalizations, specifications, recursive definitions and abbreviations).

With respect to building the LOIS WordNet, especially the definition techniques listed under type 4 (definitions) are useful: generalizations (definitions by general descriptions), specifications (definitions by listing the elements that constitute a concept) and recursive definitions (definitions by listing along the lines of a decreasing set of elements constituting the concept) can be used as glosses for the corresponding terms, abbreviations can be used as synonyms for terms. As to the first and second definition techniques, they merely provide an indication that, instead of an explicit legal definition, the implicit lexical definition should be used. The third definition technique provides an indication that the scope of a definition is extended to a different legislative document; this information can be used in establishing the correct use of WordNet synsets in specific contexts.

In a context of cross-lingual information retrieval, the links between terms in different languages have to be established on the basis of their meaning. The deviations that exist between lexical meanings and legal meanings pose additional difficulties for this linking activity. First, differences between lexical meanings and legal meanings have to be made explicit. Second, legal meanings are defined relative to a legal system. Definitions of terms contain other terms. Those terms can have lexical or legal meanings that are quite different from similar terms in different languages.

2.1 A body of fixed meanings

In European Community legislation, a unique situation is created regarding legal meaning. All language versions of regulations and directives are deemed to be authentic. Thereby, they are *de iure* equivalent to each other. Thus, for instance, the meaning of the Dutch version of a European directive is deemed identical to the meaning of the English or Greek version of that directive. Although there can be objections to this principle, relating to practical translation difficulties and theoretical discussions about the meaning of 'meaning', the effect of the principle is that there is a common basis for assessing meanings of legal terms in different EU languages. Directives establish explicit links between Community legislation and national legislation, as they provide measures that should be implemented in national legislation. For this purpose, any directive contains a series of norms. One of these is a definition article, containing a list of term definitions.

Member States can either choose to implement these definitions literally, or they can opt for a different definition, for multiple definitions, or no definition at all. Their implementations should, of course, remain within the preconditions set by the directive. Thus, in a number of cases, an explicit implementation relation can be established between terms in directives and terms in national legislation. This implementation relation, in itself, does not say anything about the way in which a concept is implemented. It only says *that* a concept has been implemented. The implementation relation can be complemented by a relation stating the nature of the link between the original concept and the implemented concept(s). For instance, if the definition of the national legal concept is identical to the Community legislative concept, an equivalence relation can be established. If the definitions are almost identical, a near equivalence relation is assumed. If the national concept has a more specific definition than the Community concept, the former is a narrower term of the latter. If the national concept has a more general definition than the Community concept, the former is a broader term of the latter.

2.2 Cross-lingual legal concept comparison

Underlying the present research is a model of linking concepts from different legal systems in various languages. This model is based on the following assumptions. First, the meaning of legal terms is for the greater part established in authoritative legal documents. These documents consist of legislation, case law or doctrine (insofar as these document types are considered to be authoritative within a certain jurisdiction). Such legal documents contain terms, some of which are explicitly defined, whereas the meaning of other ones is established on the basis of everyday or contextual use. For explicit definitions, assembling definition elements is relatively easy, especially in continental law, where many of such elements are codified. Sometimes, additional elements have to be assembled from other sources; e.g., different parts of legislation, and discussions in authoritative case law or doctrine. A term with an assigned meaning (either a legal definition, or an everyday or contextual definition) is a *concept*. Thus, legal documents contain terms, and terms refer to concepts, which on their turn are constructed from definitions or definition parts found in legal documents.

The consequence of using legal definitions is that one term may be defined in multiple ways: different legal definitions may occur for a term such as 'consumer', and for the same term, a lexical definition may be provided. The term 'consumer' may have a different meaning in agricultural legislation than it has in consumer protection law, and again a different one in a dictionary. In order to establish the meaning of a certain term within a specific context, preference rules need to be established. The

preference rules partly depend on the legal system at hand: the specific characteristics of, e.g., continental law and common law systems will induce shifts in their application.

Relations among legal concepts may provide insight into the structure of legal systems. As such, they can facilitate retrieval of relevant, related information. As explained before with respect to Community directives, two types of relations are distinguished: structural and content relations. Structural relations reflect actual systemic connections between legal concepts; content relations reflect similarities or differences among the meanings of legal concepts. A structural relation that can be used in the current model is the *implemented_as* relation, providing a reference relation between a definition in a Community directive and a definition in a national legislative document. Content relations are currently all 'borrowed' from standard WordNet relations (especially hypernymy and hyponymy).

3. LOIS ARCHITECTURE

WordNet is an initiative of the linguist George Miller and was developed and is being maintained at Princeton University [2],[8]. It encompasses an English-language electronic lexical database inspired by psycho-linguistic and computational theories of human lexical memory. A WordNet serves to support automatic text analysis and AI applications, and to provide an intuitively usable enhanced dictionary. The database of the current version 2.0 contains about 150,000 words organized in 115,000 synsets for 200,000 word-sense pairs.

WordNet represents semantic relationships between terms by arranging them in a hierarchical structure. Words (nouns, verbs, adjectives and adverbs) and their short definitions are grouped by part-of-speech in their uninflected form into synonym sets (synsets), each representing a specific lexical concept. For example, {*case, cause, causa, law suit*} form a noun synset because these nouns can be used to refer to the same concept. Synsets are often further described by glosses, in this case: 'a comprehensive term for any proceeding in a court of law whereby an individual seeks a legal remedy'. Synsets are linked to each other by different semantic relations. The most important of these are hypernymy/hyponymy (between specific and more general concepts), meronymy (between parts and wholes), and antonymy (between semantically opposite concepts). For example, the synset above has a number of hyponyms, such as *civil suit, criminal suit* and *paternity suit.*

3.1 The LOIS database

The main task of the LOIS project is the development and connection of a WordNet with concepts in six European languages, based on the EuroWordNet (EWN) framework [10]. Using this framework assures compatibility of the LOIS WordNets with EWN, allowing them to function as an extension of EWN for the legal domain. Ten partners from six European countries (seven universities/research centres and three enterprises) participate in this project. Within the approved project duration of 24 months, around 5000 synsets are being localized for each language involved. The LOIS project primarily aims at providing easy access to European legal databases for legal experts as well as for laymen. Further research will focus on improved techniques for information retrieval, on providing document standards (common XML standard for the representation of legal documents), on the representation of legal documents), on the commercial use of public sector information, on showcase applications for test and demonstration purposes, and on product placement for integration of the result into commercial applications. To reach this goal, WordNets of six different languages (Italian, Dutch, Portuguese, German, Czech, English) will be localized and - according to the archetype EWN - cross-linked through an unstructured interlingual index (ILI).

The existing Italian legal WordNet 'JurWordNet' (JWN) [3], which was developed as an extension of the Italian part of EWN, provides the basis for the LOIS lexical database (the first module of the LOIS database). Before the start of manual localization, an automatic intersection of the 1695 synsets of the Italian JWN with EuroDicAutom was made (see http://europa.eu.int/eurodicautom/). Subsequently, a mapping was created between the English result list of 579 literals and Princeton WordNet 1.6. The WordNet structures of the different WordNets have been established analogously to the Italian JWN. Up till the present moment, the manual revision, adding of definitions, and integration have been going on.

The legislative database (the second module of the LOIS database) is based on legal definitions extracted from EU sources and, for the sub-domain of consumer protection law, also from the national implementations and other relevant national provisions. For this purpose, a tool was developed to extract legal definitions from European directives. Definitions of different language versions are getting automatically connected and national implementation measures can be added manually. As a result of the distinction of a lexical database and a legislative database, two different types of concepts are represented within LOIS: *lexical concepts*, designated by terms and the lexical meanings assigned to them, and *legal concepts*, designated by terms and their definitions from legal documents.

Regarding language internal relations, primarily, the lexical relations synonymy/antonymy and the taxonomic relations hypernymy/hyponymy are used. Equivalence relations between synsets in each language are made explicit in the ILI, whereas each synset in monolingual WordNets has – either directly or indirectly by related synsets – at least one equivalence relation with an ILI-record. For demonstration purposes, the sub-domain of consumer protection law will be further structured with other WordNet relations.

As to the relationship between the lexical database and the legislative database, the priority for searching each of them can be adjusted to the specific needs of the person using the search engine. The first tests on recall and precision will take place over a text corpus covering consumer protection law. This choice offers a lot of advantages: it is a manageable field of law interesting for both legal professionals and laymen, relevant documents are easily available on both a European level and within national jurisdictions, and the localization of a limited number of concepts will be sufficient to reach realizable results for validation of the approach.

Figure 1 shows a schematic presentation of the modular LOIS architecture, with the Italian legal database (IT) as example. The main LOIS module is the National Legal WordNet. This is composed of lexical and legal concepts. The first type consists of lexical concept representations. The second type covers legal terminology. These occur in national legislation, and therefore,

they are part of the National Legal WN (NC2 in figure 1), and in EU legislation, in which case they are, because of their pan-European character, part of the National Legal WN on the one hand, and the ILI on the other (NC1 in figure 1). Each National Legal WordNet concept representation has a number of information fields associated with it. These provide information on, e.g., language, orthography, definition and associated field of law. Any of these National Legal WordNet concept representations present in language specific synsets (LSS in figure 1 below) of the corresponding EWN language components are linked to these synsets by means of plug-in relations [6].

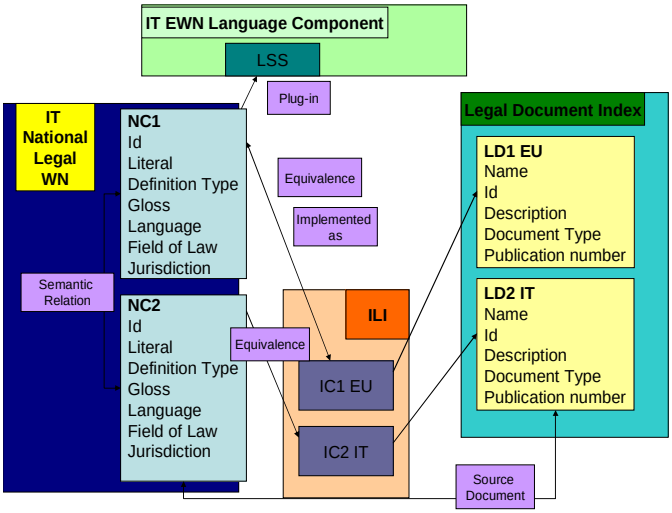

Figure 1. The LOIS database lay-out

All National Legal WordNet concept representations are linked into the interlingual index (ILI) by means of equivalence relations. Furthermore, an 'implemented as' relation has been introduced to indicate the link between EU concept representations and their nation-specific implementations. The Legal Document Index contains keys into national and European legislative texts in which the legal terms are explicitly used and defined. Currently, we anticipate to establish a consolidated legislative database, comprising current (thus, no historical) versions of statutes. Each legal document (LD in figure 1) has a number of information slots associated with it that further specify its nature. The main information is provided by the identity number that is taken from the EUR-Lex database for EU documents, and local categorizations for national documents.

Overall, the LOIS architecture will allow users to investigate a wide range of legal issues, such as the following:

- multiple senses of terms, due to different legislative sources (for instance, a legislative text amending a definition is considered as introducing a new sense of the concept);
- differences between definitions of concepts in EU and national legislation through the *implemented_as* relation;
- comparisons of national legal systems;
- lexical definitions of concepts, if no legal definition is given;
- *comparisons between* common language meaning and terminological legal meaning through available plug-in links.

3.2 Interlingual index and relations in LOIS

Cross-language retrieval presupposes a large number of highly reliable links between legal terms from different countries. In order to bootstrap an interlinked multilingual data set that will aid multilingual information retrieval and is maximally reliable, the LOIS consortium decided to look for bootstrapping data that have a maximum level of correlation between terms in different languages.

The most obvious bootstrapping candidate that allows us to capitalise on its inherent commonality is the set of EU directives, obtainable from http://europa.eu.int/eur-lex/lex/. Each of these directives has been translated into all EU languages, and the whole collection effectively forms a document-aligned multilingual corpus. We applied semi-automatic alignment techniques to link the legal terminology used in these directives across the languages covered by LOIS. The resulting paired terms can be regarded as conceptually equivalent, because they are each other's translational equivalents. In the structure of the database, the English term is chosen as the instantiation of the interlingual concept that underlies all language-specific lexicalizations.

The construction of the database relied on intra- and inter-document structural properties. From a document-internal perspective, the structure of the directives allows, up to a certain extent, the automatic extraction of terms and definitions. The use of different extraction techniques is justified on the basis of the fact that the directives we took into account were published in a quite wide range of time (from 1967 up till the present). Thus, they have been drafted using different legal conventions regarding the notation of definitions. For the task of automatic extraction, we used a mix of language-dependent and structure-dependent techniques. In particular the extraction algorithm was based on the following steps:

1. Identify a group of definitions (if present). This task is facilitated by the fact that in EU directives definitions are usually contained in article 2. As this is not always the case, we added a special heuristic to get rid of false matches.
2. Once a group of definitions has been identified with a reasonable precision, the group is divided into definition items, i.e., units of text containing both the term and its definition. The dividing process is based on paragraph division, which causes the loss of definitions extending on more than one paragraph. These definitions are usually only of limited interest from a legal point of view as, in most cases, they focus on enumerations of technical items such as chemical substances, species of micro-organisms, etc.
3. Each definition item is finally divided into a term-definition pair. This dividing process is based on several strategies such as: (a) the use of term marking punctuation such as quote, double quote, etc.; (b) the use of separation punctuation, such as column, comma, dash, etc.; (c) the use of linguistic formula, such as 'means', 'shall mean', etc.

To perform the structural analysis, we adopted a standard technique of automatically translating the directives from HTML to XHTML. After that, we used a set of XSL style sheets to perform all the computation needed. To perform the content analysis, given the quantity of the involved languages, we could not rely on any resource-based language understanding techniques. Therefore, for each language, a set of character-based

regular expressions has been encoded in order to recognize term-definition patterns.

As naive as it might appear, this two layer solution still allows the complete separation of data (structural and linguistic patterns) from code (java), so that the addition of a new language does not imply any code change. For each of the languages involved, we ran the extraction algorithm on about 1,000 directives per language, obtaining about 3,000 term-definition pairs to be submitted to manual validation. Inter-document properties that facilitate alignment are, for instance, the identical structures of the articles and sections of each directive in all languages. The alignment techniques produced juxtapositions of legal terms with their definitions. In case of ambiguity or uncertainty, human intervention was required to make a choice.

On this respect it is worth mentioning the fact that most of the problems for automatic extraction and alignment were due to either a scarce structural homogeneity across the monolingual corpora of directives or to a lack of alignment across directives translated into different languages (missing definitions or swapped positions in the list of definitions). However, in spite of these deficiencies, it must be recognized that thanks to the automatic extraction phase, the duration of the manual 'cleaning' phase can be reduced considerably. By combining automatic extraction techniques and manual cleaning it was possible to harvest a large number of authoritative legal definitions, thereby building a foundation for the interlingual index and part of the national legal WordNets.

Of the relations mentioned in subsection 2.2, the structural relation *implemented_as*, indicating the link between a concept in a directive and a concept in a national implementation measure, will be added manually during the work on the inventory of implementation measures in the domain of consumer law. Certain WordNet content relations can be based on European resources as well. First, all EU legislation (including directives) is organized by means of the Directory of Community Legislation, a high level classification of the documents. This classification is identical for all language versions involved, and consists of 20 main classes. We decided to concentrate in this first building phase on class 15: 'Environment, Consumers and Health Protection', in particular 'Consumers' (15.20). This Directory provides an opportunity to add authoritative hypernymy and hyponymy relations to the LOIS WordNet: synsets derived from individual directives can be categorized under the common denominators in the Directory.

4. CONCLUSIONS

Legal-theoretical obstacles prevent the mere one-to-one translation of legal terms between different languages. Therefore, in building a multi-lingual WordNet, we chose a hybrid approach, using both lexical and legal definitions. With this approach, cross-lingual information can be attained both on a more general, lexical level, and on a more specific, legal level. Lexical definitions, insofar as they cannot be extracted correctly from existing lexical databases, can be translated manually on the basis of the original lexically oriented JurWordNet and its English translation. Legal definitions can be based on the authoritative language versions of all European regulations and directives. These offer the possibility of of introducing an equivalence relation between legal concepts in different languages. An equivalence relation (for identical concepts from directives) and a near-equivalence relation (for related lexically defined concepts) establish links between concepts in different languages. If no equivalence or near-equivalence relation is present, analogous hierarchical structures can help in finding relations between terms in different languages, and thus, for instance, in comparative law research.

ACKNOWLEDGEMENTS

We would like to thank three anonymous referents for their comments on the full version of this paper. We would also like to thank the other partners in the LOIS project for their collaboration in the research underlying this paper. The LOIS project is funded through the eContent-programme of the European Commission. Please refer to www.loisproject.org for more information.

REFERENCES

[1] Eijlander, P. and W. Voermans (2000), *Wetgevingsleer*, Den Haag: Boom Juridische uitgevers.

[2] Fellbaum, C. (ed.) (1998), *WordNet: An Electronic Lexical Databas*e, Cambridge, Mass.: MIT Press.

[3] Gangemi, A, M.-T. Sagri, D. Tiscornia (2003), 'Jur-Wordnet, a Source of Metadata for Content Description in Legal Information', in: *Proceedings of the Workshop on 'Legal Ontologies & Web based legal information management'*, part of The International Conference of Artificial Intelligence and Law (ICAIL 2003), Edinburgh, June 24, 2003.

[4] Gruber, T.R. (1993), 'A Translation Approach to Portable Ontology Specifications', *Knowledge Acquisition* vol. 5/2 (1993), London et al.: Academic Press, pp. 199-220.

[5] Hart, H.L.A. (1961), *The Concept of Law*, Oxford: Clarendon Press.

[6] Magnini, B and Speranza, M. (2001), 'Integrating Generic and Specialized WordNets', in: *Proceedings of the Euroconference RANLP 2001*, Tzigov Chark, Bulgaria.

[7] Matthijssen, L. (1999), *Interfacing between Lawyers and Computers: An Architecture for Knowledge-based Interfaces to Legal Databases*, The Hague et al.: Kluwer Law International.

[8] Miller, G.A. et al. (1990), *Five Papers on WordNet*, CSL Report 43, Princeton University: Cognitive Science Laboratory (ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.ps).

[9] Schweighofer, E. (1999): *Legal Knowledge Representation, Automatic Text Analysis in Public International and European Law*, Law and Electronic Commerce, Volume 7, Kluwer Law International, The Hague.

[10] Vossen, P., Peters, W. and Díez-Orzas, P. (1997), *The Multilingual design of the EuroWordNet Database*, in: Mahesh, K. (ed.), *Ontologies and multilingual NLP, Proceedings of IJCAI-97 workshop*, Nagoya, Japan, August 23-29.