Semantic access to multilingual legal information

Paolo Curtoni, Luca Dini, Vittorio Di Tomaso[1]
Laurens Mommers[2], *Wim Peters*[3], *Paulo Quaresma*[4]
Erich Schweighofer[5], *Daniela Tiscornia*[6]

[1]*Celi S.R.L.*
curtoni/dini/ditomaso@celi.it
[2]*Center for Law in Information Society, Leiden University*
l.mommers@law.leidenuniv.nl
[3]*NLP Group, Department of Computer Science,University of Sheffield*
w.peters@dcs.shef.ac.uk
[4]*Departamento de Informática, Universidade de Évora*
pq@di.uevora.pt
[5]*Arbeitsgruppe Rechtsinformatik, Universität Wien*
erich.schweighofer@univie.ac.at
[6]*Istituto di Teoria e Tecniche per l'Informazione Giuridica del CNR*
tiscornia@ittig.cnr.it

1. Introduction

Today, search engines for legal information retrieval do not include legal knowledge into their search strategies. These strategies include keyword and metadata search, but do not address the semantics of the keywords, which would allow, for instance, conceptual query expansion. In other words, there is no semantic relationship between information needs of the user and the information content of documents apart from text pattern matching. Often, query formulation by either legal practitioners or laymen users is only an imperfect description of an information need [Matthijssen 1999].

The EU funded eContent project LOIS (Lexical Ontologies for Legal Information Sharing) (EDC 22161) aims to remedy this semantic lacuna by means of the development of a multi-language legal thesaurus, whose structure is based on existing de facto standards for semantic thesaurus construction. The main task of Lois is the development and connection of 6 legal

WordNets based on the EuroWordNet (EWN) framework [Vossen et al. 1997]. From the start, the project integrated a number of methodologies, in order to cope with the acquisition and combination of multilingual domain specific terminology and existing general language repositories. Our architecture ensures the coverage of the semantic peculiarities of the legal domain, and facilitates the capture of essential semantic differences between the legal systems involved.

The paper is structured in the following way: the first sections describe the Lois data base, starting from the methodological choices (section 1), the building process (section 2) and the current state (section 3); section 4 describes the technical aspects; section 5 outlines some methodological questions under discussion and section 6 discusses potential applications of LOIS compared to the traditional tools and in section 7, one of these potential applications is presented in the form of a Question/Answering System.

2. Choice of database structure

As its methodological starting point, Lois adopts the structure of two widely known and used thesauri. WordNet [Fellbaum 1998] is a lexical database which has been under constant development at Princeton University. EuroWordNet (EWN) [Vossen et al. 1997] is a multilingual lexical database with wordnets for eight European languages, which are structured along the same lines as the Princeton WordNet. Both thesauri are organized around the notion of a *synset*. A synset is a set of one or more uninflected word forms (lemmas) with the same part-of-speech that can be interchanged in a certain context. For example, {*case, cause, causa, law suit*} form a noun synset because they can be used to refer to the same concept. A synset is often further described by a gloss, explaining the meaning of the concept. Synsets can be related to each other by semantic relations, of which the most important are hypernymy/hyponymy (between specific and more general concepts), meronymy (between parts and wholes), and antonymy (between semantically opposite concepts). Cross-lingual equivalence relations are made explicit in the so-called Inter-Lingual-Index (ILI). Each synset in the monolingual wordnets has at least one equivalence relation

with a record in this ILI. Language-specific synsets from different languages that are linked to the same ILI-record by means of a synonym relation are considered conceptually equivalent.

The ILI is the superset of all concepts from all wordnets, and the concepts from indigenous wordnets are linked into one or more ILI records by means of equivalence relations. These relations indicate complete equivalence, near equivalence, or equivalence as a hyponym or hypernym. The network of equivalence relations determines the interconnectivity of the indigenous wordnets.

In principle, the ILI is an unordered list of concepts, i.e., it does not have any internal structuring. The reason behind this is that we assume that each language imposes its own language-specific structural constraints on the concepts. Therefore, any ordering of ILI concepts needs to be retrieved from knowledge bases that link into the ILI. ILI concepts enter into relations with each other by means of:

- the equivalence relations between indigenous concepts and ILI concepts;
- traversal through the relations within the indigenous wordnets.

In addition to the two existing WordNets, a number of additional sources were used to establish the set of concepts and the links between legal terms in different languages. First, Eurovoc, the EU thesaurus, helped in establishing relations as it is a multilingual thesaurus that covers part of the (European) legal domain. Second, the phrase database Eurodicautom, which supports European translators in keeping track of translations of words and phrases in different languages, also helped establishing potential candidates for the translations of certain terms. Third, the Italian legal wordnet (JWN), discussed in the next section, played a vital role in the project, as it determined the starting set of legal concepts to be translated in the other languages. Fourth, the existing corpus of EU legislation provided a valuable source of concepts.

The LOIS WordNet, including its interlingual index, has significant advantages in comparison with existing thesauri on EU law and politics. All existing thesauri are focussed on documentation and lack sufficient granularity for semantic access to EU law. The most prominent EU thesaurus is *Eurovoc*. Eurovoc is a multilingual thesaurus – a controlled vocabulary –

covering the policy fields of the EU. It provides a means of indexing the documents in the documentation systems. The latest version - Eurovoc 4.2 - exists in 16 official languages of the European Union (Spanish, Czech, Danish, German, Greek, English, French, Latvian, Italian, Hungarian, Dutch, Polish, Portuguese, Slovene, Finnish and Swedish). Eurovoc has a hierarchical structure with inter-lingual relations. As the focus is on socio-economic issues, depth in law is quite low and the structure is not appropriate to EU law. The classification codes (or headings) of the *Register of the Community law in force* represent much higher quality for legal purposes. Inter-lingual relations exist and also a relatively fine-grained hierarchical structure is present. Depth is still not sufficient. The quite powerful *descriptors of the European Court of Justice* are more a list of legal sentences (“Rechtssätze”) than a proper thesaurus.

As Eurovoc is the most salient ‘competitor’ for LOIS, we explain some additional differences between the two. First, compared to Eurovoc, LOIS not only has hierarchial and synonymy relations, but also includes (near) equivalence and part-of relations, and other WordNet semantic relations, in order to contain more semantic knowledge on the meaning of a concept. LOIS is specifically aimed at the legal domain, whereas Eurovoc has a broader scope (European policy issues). Moreover, because of the lack of semantic precision with which Eurovoc was drafted (apparent from inaccurate hierarchical and synonymy relations), it is only suitable for retrieving *related* terms. The LOIS knowledge base has relatively precise synonymy and hierarchical relations, so that it is more suitable for retrieval purposes.

Beside the dynamic application of Lois in the searching process as a means of conceptual query expansion, the semantic connotation of Lois allows a deep and refined semantic tagging in the editing phase, capable to express sense distinction, polysemy disambiguation and context dependence and to check ontological consistency [Schweighofer et al. 2005].

3. Building the wordnets

In LOIS’s initial phase, we needed to pinpoint a nucleus of pilot concepts offering a reference structure for the building of

the wordnets in the other languages [Dini et al. 2005]. To allow greater sharing, only the general level of doctrine definitions could prove effective. This entailed the inclusion of common sense concepts that are used in doctrine, but are not confined solely to the domain of legal terminology. Thus we “translated” only a nucleus of common sense knowledge from JWN, the Italian legal wordnet [Bertagna et al. 2004], in order to bootstrap the localization into other languages. The Italian concepts were manually selected from the frequency list of Italian Legislation corpus. Their selection was based on the assessment of experts. Descriptions (glosses) were extracted from legal handbooks.

The identified core allows the integration and homogenization of local lexicons.

In the second step of LOIS development the emphasis was on the detection of legal conceptual terminology, i.e. terminology that is specific to the legal domain, as opposed to the common sense concepts above. In order to identify this legislative knowledge, a parallel corpus was created from the European Directives in the EU languages. Semi-automatic alignment techniques enabled the selection of a multilingual set of legal terms. Legal terms were only selected if they had an explicit definition in the text. This criterion sets these terms apart from the lexical terms described above. Once alignment had been established, conceptual equivalence was assumed and each set of corresponding terms in different languages were automatically linked to one unique identifier. The relation *implemented_as* defines the link between a European legal concept and its implementation in national legislation. As to legal concepts from European legislation, the unique *Identifier* acts as the Interlingual Index item. Automatic extraction of legal concepts from national legislation (limited to the consumer law domain) is still in progress.

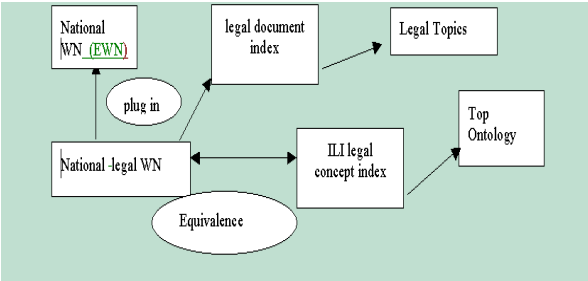
4. Content of LOIS

In correspondence with the two building approaches described in the previous section, the main module of each Lois national wordnet is composed of:

- An indigenous *lexical data base*, which conceptualizes general language entities pertaining to legal theory and legal dogmatic, a set of patterns (models) in line with

which law is formed and operates, and which is structured according to the EWN methodology;
- a *legislative data base,* populated by concepts defined in European and national legislation and structured according to purely legal (supra)national models.

The entries of the two types of legal knowledge link into the interlingual database component: the inter lingual index (ILI).



Moreover, synsets in the National Legal WN are (or shall be) linked by *plug-in* relations (such as eq-plug-in, hyper-plug-in, see [Magnini and Speranza 2002] to the general language modules, developed within the *EuroWordNet Project*. Overall, LOIS will consist of a number of modules that directly or indirectly link into EWN modules through each individual language component (see figure above for a simplified view on the database structure).

5. Technical aspects

From the technical point, in the task of identifying a structure to represent the lexical database, we had to address two issues:

- the structure of the lexical database had to be compatible with the workflow defined in the project, where different legal experts are working concurrently on a single structure, adding entries and modifying relations among entries in real time;
- the structure had to be compatible with CLIR, a multilingual search engine for the legal domain with advanced capability of query expansion and query

translation, and also with other applications, which may use the lexical database as a resource.

While the first aspect naturally led to the implementation of the Lois lexical database as a relational database, which is centrally maintained and accessed through a web interface, the second aspect led to developing a more portable structure, which could be easily deployed in different search environments. Such a structure supports the inferences used for query expansion and query translation made available from the lexical network as defined above, including the mechanism of *plug-in* and *implemented_as* relations, which further complicate the navigation in the lexical graph. Since the lexical database forms a rather complex graph, we had to face efficiency issues in order to improve the response time of the inference engine (the number of visited nodes significantly grows with respect to the query length and the required precision of expansion).

In the current system, as realised in CLIR, the LOIS database is exported from the relational structure as an inference engine which is DBMS independent. To improve response time (but increasing space occupation) each synset is exported as a complete subgraph, containing all the connected synsets, either through internal relation or through eq_links. At export time it is possible to decide the kind of relations available for expansion and/or translation and the maximum depth of each subgraph. The inference engine is realized in java and it has been tested on lexical graphs of more than one million synsets.

6. Open Methodological Questions

The database currently holds 5,500 synsets, which originate from European Community definitions, national legislation and lexical data bases. The expansion of the lexicon requires the integration of the bottom-up strategy described above with a top-down validation, in order to expand the coverage and consolidate the structure of the overall model.

The ILI forms the platform for the integration of external knowledge resources [Doerr 2003]. These resources will function as meta-ordering principles of the ILI concepts. Inclusion of an increasing number of these ordering principles, such as, for instance, a general top-level and a domain specific Core Legal

Ontology (CLO), will allow greater complexity and refinement in knowledge representation and ontology comparison.

One of the open questions that external ordering principles are expected to support is the management of the semantic/equivalence relations via the ILI in the integration of legislative and lexical/common sense knowledge. With respect to legal concepts from national legislation, the ILI can be automatically generated, i.e., for each legal concept a corresponding ILI equivalent is created. If a legal concept from a European directive is implemented in indigenous legislation, and the local legal concepts are deemed (legally) equivalent to their European counterparts, then an equivalence relation between the two local concepts may also be established. In all other cases, the creation of semantic links between local synsets does not necessarily imply the creation of equivalence relations with the ILI, except in cases where concepts from more than one indigenous wordnet coincide, in which case these will all be related to one ILI record. Within this architecture, the semantic structures peculiar to each wordnet will be preserved, and will overlap through the ILI. External ontologies such as the DOLCE2.1-Lite-Plus + CLO [Gangemi et al. 2005] will structure the ILI concepts, classifying concepts according to explicit and consistent subsumption relations.

Polysemy detection is a further aspect that an ontology may solve, as pointed out by [Gangemi et al. 2002] and [Vossen et al. 1997]. Polysemy (one term has more than one meaning) is expressed in LOIS by the association of one synset to each sense of a polysemic word.

To assign for each sense of a word in the source language the right equivalent in the target language (or to create a new synset when a sense in source or target language is missing), ontological distinctions can be necessary to make meaning commitments explicit [Sowa 2004]. For instance, one of the typical ambiguities of legislation is the distinction between the regulatory and physical existence of legal phenomena. The Italian term *contratto* is, in terms of CLO concepts, a *legal description*, an *information content* and a *physical object* (the material support of the information content. A legal institution, for instance the *Prime Minister*, is a figure, created by norms, but it is also a social role: in complex figures, like organizations or institutions, an endurant (a physical person) plays a *delegate*, or *representative* role of the figure.

In addition to the association of word senses of polysemic words with ontologically well-formed concepts, a key step in the process of the methodological refinement of ontological categorization will be the consistent distinction of degrees of equivalence between contexts in which the word occurs. The importance of contexts is stressed in the field of computational terminology [Kerremans and Temmerman 2003] in which there is consensus on the necessary anchoring of term extraction, term definition and inter-term relation identification on the contexts of use. The traditional 'standardisation oriented' and 'concept centred' approach, where (ideally) only one term is assigned to a concept, has proved to fail in cross-lingual conceptualizations.

In law, legislative definitions are contexts which have a prescriptive force. This fact influences the determination of the number of senses of terms, and the equivalence setting between legal concepts and lexical concepts.

The most common situation is the 'apparent polysemy' generated by the integration of the legal and lexical databases, because meanings of the legal concepts are usually more specific than the meanings of corresponding lexical items, and legal senses can display degrees of ontological overlap or even taxonomic ordering. The lexical sense can be considered to be a prototypical description of the concept properties, over which the legislator's definition impose constraint, and therefore it is classified as a hypernym of all legal senses. For instance, from EU Legislation texts, obtained from Celex [3], four senses of 'worker' are defined:

1. any worker as defined in Article 3 (a) of Directive 89/391/EEC who habitually uses display screen equipment as a significant part of his normal work.
2. any person employed by an employer, including trainees and apprentices but excluding domestic servants;
3. any person carrying out an occupation on board a vessel, including trainees and apprentices, but excluding port pilots and shore personnel carrying out work on board a vessel at the quayside;
4. any person who, in the Member State concerned, is protected as an employee under national employment law and in accordance with national practice;

The corresponding lexical entry is defined in the lexical part as follows:

5. a person who works at a specific occupation.

7. Potential applications of LOIS

The use of the semantic information contained in the LOIS database will be of crucial importance for a wide variety of applications. Below we list a number of them.

7.1. Information retrieval applications for laymen and legal professionals

Applications in this field use semantic information in order to make a better selection of search results (increased precision and recall). The optimal use of semantic information depends on the projected user group and the type of use this group makes of it. The following uses are possible for laymen:

- query expansion: using synonyms and narrower terms of the search term in order to increase recall (by *adding* these terms to the query);
- query specification: using synonyms and narrower terms of the search term in order to increase precision (by *replacing* search terms with these terms);
- query explanation: using the glosses in the WordNet to find out what a specific search term, synonym or narrower term means;
- result explanation: using the glosses in the WordNet to find out what a specific term in a document means;
- term explanation: by browsing in the WordNet, the structure and characteristics of a domain can be scrutinized.

7.2. Applications for legal professionals

Professional uses are more diverse. As a concept-oriented semantic net such as WordNet still requires background knowledge in order to grasp its meaning, its use in a professional context is more appropriate. The types of uses for legal professionals are the following, in addition to the ones listed for laymen:

- comparison of concepts and conceptual nets: for the comparison of legal systems, it is vital to know the contexts in which similar concepts are used. For instance, the embedding of the English concept of ‘property’ is different from the Dutch concept of ‘eigendom’. By comparing the conceptual networks, it becomes possible to improve the methodological foudings of comparative law research;
- finding implementations of EU legislation: EU directives find their way into national legal systems. The comparison of implementations of EU directives in different countries can support the unity of legal systems in Europe and the effectivity of the European legal order;
- finding relevant legal documents in other legal systems. In legal practice, arguments can nowadays not only be derived from national case law, but also from authoritative foreign case law. In order to find such case law, it is necessary either to have substantial command of foreign languages (and the accompanying legal terminology) or to have a ‘bridge’ between different legal systems. A multi-lingual WordNet can form such a bridge.

7.3. Question/Answering Applications

In addition to the above-mentioned uses, the semantic information contained in LOIS can e an important source for a variety of applications in the field of natural language processing (NLP), such as information extraction and question-answering. One of the key-issues in this type of applications is the need for an ontology of concepts, allowing the semantic-pragmatic interpretation of users queries in the context of the domain knowledge [Quaresma 2005].

NLP techniques such as syntactic analysis of legal text and conceptual definitions, and inferencing through the LOIS conceptual hierarchies provide powerful mechanisms to obtain information from the legal domain and provide it in customised forms. The LOIS WordNet is currently being used as an important source of knowledge in the development of a question-answering system for juridical documents.

The system is composed by two major modules: (1) preliminary analysis of documents (information extraction); and (2) query processing (information retrieval). Both modules are composed by several sub-modules: (1) syntactical analysis: sentences are processed with a specialised parser; (2) semantic analysis: sentences are rewritten into DRS (discourse representation structures), a list of discourse referents and a set of conditions; and (3) semantic and pragmatic interpretation In the context of this paper, we will only describe briefly the semantic/pragmatic interpretation. In this phase the DRSs are processed, taking into account the LOIS ontology (for a complete description see Quaresma 2005 in the JURIX main conference).

This process receives as input a discourse representation structure, DRS, and it interprets it using rules obtained from the knowledge ontology. In order to obtain a good interpretation, the strategy is to search for the best explanation that supports the sentence logical form. Suppose the sentence ‘The user X sent a text message to the company Y’ is transformed into the following structure by the semantic analysis (in order to keep the example simple, some simplifications were made):

```
drs(entities:[A, B, C]
conditions:[user(A), name(A, 'X'),
text_message(B),
sent(A, B),
company(C),
name(C, 'Y'),
rel(to,B,C)])
```

Using information from the LOIS ontology it is possible to interpret the DRS and to obtain the following structure:

```
drs(entities:[A, B, C, D]
conditions:[user(A), person(A),
name(A, 'X'),
text_message(B), electronic_mail(B),
sent(A, B, C),
company(C),
name(C, 'Y'),
electronic_service(D),
used(A, D)])
```

The inference of this new structure was possible due to the information obtained from the LOIS ontology: (1) a user is a

person using an electronic service (consumer law CELEX EU directive nº 2002/58/CE); (2) an electronic mail is any text, voice, sound or image message (consumer law CELEX EU directive nº 2002/58/CE); (3) Action “to send” has a preferred interpretation with 3 arguments: “A sends B to C”. As can be seen from this example, the LOIS ontology allowed the interpretation of the sentence in the context of the legal domain. Using the new interpretation it is possible to answer user queries, like the following ones: (1) who has sent electronic mails to company X? or (2) who has used electronic services? Note that these queries could not easily be answered from the initial DRS structure.

As future work, this question-answering system should be able to use the cross-language aspects of LOIS, and to be able to answer queries made in one language with information conveyed in another language, i.e., to be a full cross-laguage question-answering system.

8. Conclusions

Effective access to EU legal information requires advanced linguistic interpretation of search queries and appropriate links to a powerful lexical ontology. Based on the existing work of our groups on LOIS, we presented an inventory of possible applications of the LOIS WordNet. In the future, prototypes of both the different uses of LOIS in the context of legal information retrieval and of LOIS’ use in question-answering systems will be developed.

9. References

[Bertagna et al. 2004] *Bertagna F., Sagri M.-T., Tiscornia D.*, Jur-WordNetp Global Wordnet Conference (GWC 2004), Brno 2004

[Dini et al. 2005] *Dini, L., Liebwald D., Mommers L., Peters W., Schweighofer E. and Voermans W.*, 'Cross-lingual legal information retrieval using a WordNet architecture'. In Proceedings of ICAIL '05, p. 163-167.ACM, Bologna, 2005

[Doer et al. 2003] *Doerr, M., Hunter, J., Lagoze, C.,* Towards a Core Ontology for Information Integration , in Journal of Digital Information, Volume 4 Issue 1, 2003

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

[Gangemi et al. 2002] *Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., Schneider, L.*, Sweetening Ontologies with DOLCE. In: proceedings of EKAW 2002

[Gangemi et al. 2005] *Gangemi, A., Sagri, M.-T., Tiscornia, D.,* A Constructive Framework for Legal Ontologies . In: Law and the Semantic Web (Benjamins, Casanovas, Breuker and Gangemi eds.) Springer Verlag, 2005

[Kerremans and Temmerman 2004] *Kerremans K. and Temmerman R.*, Towards Multilingual, Termontological Support in Ontology Engineering. In; Proceeding of Termino 2004 , workshop on Terminology, (2004)

[Magnini and Speranza 2002] *Magnini, B. and Speranza, M.*, Merging Global and Specialized Linguistic Ontologies, in Proceedings of the Workshop Ontolex-2002 Ontologies and Lexical Knowledge Bases, LREC-2002, pp. 43-48

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

[Schweighofer et al. 2005] *Schweighofer, E., Liebwald, D.*, Advanced Lexical Ontologies and Hybrid Knowledge Based Systems: First Steps to a Dynamic Legal Electronic Commentary. In: Lehmann, Jos, Biasiotti, Maria Angela, Francesconi, Enrico, Sagri, Maria Teresa: LOAIT – Legal Ontologies and Artificial Intelligence Techniques. Nijmegen: Wolf Legal Publishers 2005, 71-81

[Quaresma and Rodrigues2005] *Quaresma, P., Rodrigues, I.*, A Question-Answering System for Legal Information Retrieval. In Proceedings of JURIX'05, Brussels, Belgium, December, 2005

[Sowa 2004], *Sowa J.*, Building, Sharing, and Merging Ontologies http://users.bestweb.net/~sowa/ontology/ontoshar.htm; Bertagna F. Sagri T. Tiscornia D. *Jur-WordNet* in Proceedings of the *Global Wordnet Conference (GWC 2004*), Brno

[Vossen et al. 1997] *Vossen, P., Peters, W. and Díez-Orzas, P.*, 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