Multilingual access modalities to legal resources based on semantic disambiguation

G. Peruginelli, E. Francesconi

ITTIG-CNR – Institute of Legal Information Theory and Techniques
Italian National Reserach Council, Italy

Abstract. An effective access to multilingual legal materials is strictly linked to the peculiarities of legal language as a technical language closely related to the diverse legal systems. This paper proposes an approach for a coherent cross-language information retrieval system based on semantic document indexing able to contextualize queries for terms disambiguation and translation.

1 Multilingualism in the law domain: an overview

Today there is a strong need for worldwide sharing of legal information as internationalization and increasing globalization of market economy and social patterns of life have created a situation where the need for legal information from foreign countries and from different legal systems is greater than ever before. Such requirement is not new, but it is getting increasingly complex to meet under the pressure of the rapid cross-border transactions occurring between people of different legal cultures and languages. It is no doubt that the exchange of information is largely dependent on language, to be intended not only as a system of symbols, but also as a mean of communication [1], a tool for mediating between different cultures. As regards the language of the law, such languages properties have a major impact on the exchange of legal information. Cross-Language Information Retrieval (CLIR) refers to a functionality implying the ability of a system to process a query for information in any language, search a multi-language collection and return the most relevant documents. As such, CLIR offers a practical approach towards worldwide sharing of knowledge for its potential to make information accessible across language barriers. The difficult task to effectively access multilingual legal material through information retrieval systems is definitively to match and weight legal terms across languages [2]. This generally implies translating from the language of the query to that of the material to be found or viceversa, and addressing the problem of word disambiguation which is greatly increased when mapping over legal languages. In fact crossing the language barrier between search requests and documents implies facing the problems of the system-bound nature of legal terminology and devising methods to map concepts between different legal systems. It is a matter of fact that in the last decades research and developments on CLIR have progressed rapidly, important cooperative initiatives have been undertaken at international level1 and

issues of multilingual querying, presentation and retrieval have been extensively tackled mainly in the area of general domain information. A rather limited number of studies and applications have been produced in domain-specific areas and cross-language retrieval of legal information has received limited attention. In a multilingual access environment information is searched, retrieved and presented effectively without constraints due to the different languages and scripts used in the material to be searched and in the metadata, that is descriptive and semantic information allowing the retrieval of indexed documents to be found. One main question arising in the context of CLIR of law material is how should the language barrier between the search requests and documents be crossed. This involves decisions about what to translate: search requests into the language of the documents or documents into the language of the request, or even both. Besides this fundamental question, one crucial issue regards the best approach to adopt in carrying out translation and how far translating terms can be successful when dealing with legal information. From a practical point of view the approach for large collections is usually based on the most economical method, consisting in simply translating the query at retrieval time into the document (or metadata) languages, although it would be possible to translate all of the documents into the query language. This presupposes that the query can be translated in a reasonably accurate fashion and that monolingual retrieval systems are available for all of the document languages. Although many experiments have been carried out in general domain information using query translation techniques, in the real world they pose a number of problems related to the need of contextualization and interpretation, which are increased in the law domain [3].

2 Key components of cross-language legal information retrieval

Retrieving information over languages implies facilities such as multiple language recognition, translation, manipulation of information of queries and documents, cross-language search and retrieval, display and merging of results [4]. Basically, these components reflect different sides of the problem of multilingual access, covering technical and linguistic aspects. In such a context the system-bound nature of legal terminology, the complexity of legal languages, legal translations issues and comparative law aspects are major issues having important implications for effective retrieval of law across languages. Legal translation is an essential function for cross-language retrieval systems. One major question concerns the translation strategy to be adopted in order to ensure that users access legal information independently of the language used in a query. The relation between law and language can significantly broaden the scope of legal translation theory. In fact, while it can be assessed that everyday language already implies a formalized way of communication, legal language introduces a supplementary system of formalisation [5]. Although legal translation demands precision and certainty, it is bound to use abstractions, whose meanings derive from particular changing cultural and social contexts. These contexts generate a certain degree of ambi-

guity, which increases when the legal cultures and systems are vastly different from each other. On a practical level the problems raised by legal translation are strictly connected with those related to the variety and diversity of legal systems and as such to comparative law. Retrieval systems to legal information across different legal systems represent a practical approach to the confrontation and exchange of legal cultures. The whole process of interaction between legal languages can be identified as finding equivalents across legal systems. If no acceptable equivalents can be found in the target-language, subsidiary solutions must be sought, such as no translation and use of source terms, paraphrasing, creating a neologism with explanatory notes. Pure linguistic problems are likely to be encountered due to legal false friends.[1] Despite the difficulties in establishing the equivalence of legal concepts belonging to different legal systems, a compromise has been adopted in trying to favour the integration of diverse legal cultures, while respecting each national legal system. What is needed is the identification of a common ground, namely common legal concepts and facts which, although not perfectly coinciding with those belonging to other systems, are conceptually close. It is up to legal users, once the material has been examined, to perceive the differences and peculiarities which make these resources unique. It is to be underlined that this does not necessarily lead to noise or unsuccessful searches, but allows for a first-phase search in context, useful to give evidence of the existence or non-existence of a specific concept in other legal systems.

3 Towards solutions for multilingual retrieval of law

Knowledge-based systems can greatly contribute to cross-language retrieval through the structure and function of thesauri and ontologies. In fact these tools have the potential to manage the complexities of terminology in language and provide conceptual relationships, ideally through an embedded classification/ontology [6]. In the domain of law efforts are starting to be made in this direction. These are represented for example by the Lexical Ontologies for Legal Information Sharing (LOIS) project [7], Jurwordnet[2] [8], DALOS project [9] and by a number of linguistic tools like the Legal Taxonomy Syllabus (LTS)[3], Eu-

[1] For examples the terms "administrative tribunals" cannot be translated in French as "tribunaux administratifs". The English word for the French tribunal is Court and the administrative tribunals are administrative commissions which are comparable, mutatis mutandis, to the French "autorits administratives indpendantes".

[2] The law lexicon is characterized by both taxonomic vertical and associative horizontal relations and it has been developed by the Institute of Theories and Techniques for Legal Information (ITTIG-CNR)

[3] LTS consists of both a database and a software development within the European project "Uniform terminology for European Private Law" and is coordinated by the Dipartimento di scienze giuridiche of the University of Turin. Available at: `http://www.eulawtaxonomy.org/index_en.html`

rovoc Thesaurus[4] and Jurivoc[5], the legal thesaurus of the Swiss Federal Court. But in practice, aligning law vocabularies of two or more languages is a hard process. Ideally a multilingual legal thesaurus should include all concepts needed in searching by any user in any of the source languages, but difficulties arise in making the systems of legal concepts the same for all languages as a different language often suggests a different way of classifying law material and a system needs to be hospitable to all of these. In such a context what cross language retrieval of legal information systems should manage is mapping each query term from the source language to its possible multiple equivalents in the target language. However each of these equivalents may have other meanings in the target language or may not have a precise equivalent, requiring to be mapped to broader or narrower terms, but this can lead to distorting the meaning of the original query. Multiple meanings can be disambiguated through users interaction, but the success of this approach depends on the quality of the hierarchy of concepts, the provision of well-structured cross-references, and on the interface of the system. The adoption of a common metadata format, where to accommodate semantic classification of legal documents by using categories of law, can ensure a successful legal mapping across languages and systems. Categories of law of a specific legal system, in fact, represent the way how retrieval can be satisfactorily achieved. As often there is no conceptual nor content similarity between the categories of law (i.e. trade law, constitutional law, criminal law) of the different legal systems, mapping between such law categories is necessary to reach proper contextualisation of the query in the diverse legal systems. An example illustrates the need for such mapping. The concepts related to property rights, such as the development of property law, land law, property questions on insolvency, intellectual property, etc. according to UK law belong to the field of property law, whereas in the Italian legal system these legal facts are regulated by private law, agricultural law and industrial law. Below an illustration is given of a possible approach to a coherent multilingual legal information access based on categories of law and on full text and metadata indexing.

4 A possible approach for accessing multilingual legal resources

Let us consider an information system offered to the users where a full text and metadata indexes in a multi-language environment are available. In this context two are the different advanced search modalities that can be envisaged:

1. *metadata-based document querying* (MBDQ);
2. *keyword-based document querying* (KBDQ), combined with category (*category-based document querying* (CBDQ)).

[4] It is the multilingual and polythematic thesaurus of the European Union. `http://europa.eu/eurovoc/`

[5] `http://www.bger.ch/it/index/juridiction/jurisdiction-inherit-template/jurisdiction-jurivoc-home.htm`

Case 1. Advanced search: the user submits a query filling in the fields related to the adopted metadata schema (for example DC metadata set, taken here as reference).

Case 2. Simple search: the user submits a query, filling an unqualified text box using keywords. Moreover, in order to make the query more focused, the user may choose a legal category of the legal system associated with a language domain.

Dealing with querying and retrieving multi-language documents, basically involves the problem of query translation. As discussed in Section 2, especially in legal domain, a word in the query language can be ambiguous, having therefore different translations in a target language, each corresponding to a legal category in the target legal system (i.e. the Italian word "dolo" has two different translations into English: "fraud" and "malice", respectively belonging to private law and criminal law). The right sense of an ambiguous word in query language can be obtained only by word contextualization, giving the right sense to the context in terms of a legal category. A legal category in the legal system of the query language can be mapped to the correspondent legal category in the target legal system, therefore the right translation of the ambiguous word can be obtained. If more than one category in the target legal system corresponds to the original legal category, more than one translations of the ambiguous word are selected. Therefore, in both the modalities of querying (MBDQ and KBDQ+CBDQ), the identification of a legal category is essential in order to identify the right translation of an ambiguous word. The procedures used to obtain these results in MBDQ and in KBDQ+CBDQ modalities are described respectively in Section 4.1 and 4.2 (Fig. 1 can be used as reference).

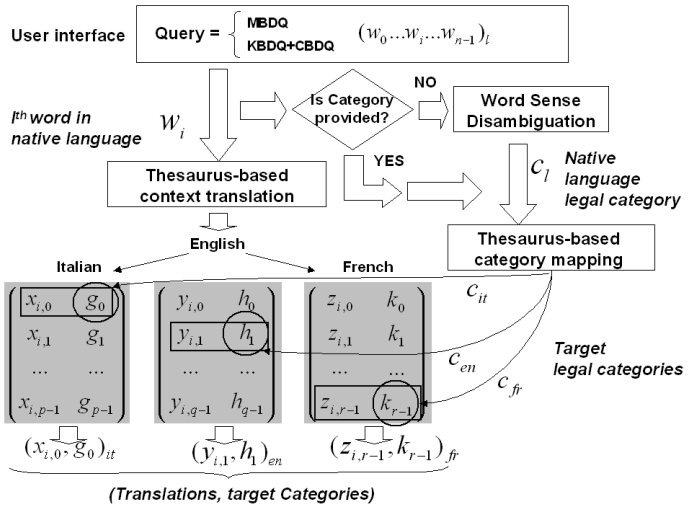

Fig. 1. Query translations of in MBDQ and KBDQ+CBDQ modalities

4.1 Query based on metadata (MBDQ)

MBDQ represents an "advanced search" modality of querying a qualified document index. The user first of all is required to choose a legal system, thus implicitly identifying a language for queries, and a legal category, identifying the right translations of possible ambiguous words. Then each metadata field is filled with a set of words $(w_0, w_1, ..., w_{n-1})_l$, representing a context expressed in the query native language l, that has to be translated by a *thesaurus-based context translation* procedure. Not every field has to be translated. In fact, bibliographic metadata (as for example the Dublin Core metadata set) can be divided into query language-dependent and query language-independent metadata. For example *dc:title* field is query language-independent since, for example, the title of a document has to be queried in its native language, independently from the query-language. Therefore only the contents of query language-dependent metadata have to be translated. While in a multi-language environment the semantic classification (*dc:subject*) is usually query language-independent (or neutral [10]), within a multi-language legal domain this is not true (Section 2). For this reason a semantic category has to be translated, by mapping it from a legal system to different target ones. Also the content of the widely used access point *dc:description* field (the document abstract) is query language-dependent: the information contained is often expressed using a semi-technical language; therefore a *dc:description* field can be considered as important to translate as the *dc:subject* . The contents of *dc:subject* and *dc:description* fields, submitted in a native language are translated in a "pivot" language (English) [11]. Then, from the "pivot" language, the query is translated again to the other languages used by the retrieval system. The use of a "pivot" language in a N-language environment allows the reduction of the number of bilingual thesauri from a factor N^2 to a factor N, and also allows the solution of the problem of the non-availability of some bilingual thesauri. As discussed in Section 2 the main problem with translation is that a single word (w_i) or expression in the native language can have different translations in a target language, depending on the context. For example, let us assume, without loosing generality, that w_i be an ambiguous single word of the context $(w_0, w_1, ..., w_{n-1})_l$ in the *dc:description* field in query native language l. According to Fig. 1, different English translations $\{y_{i,0}, y_{i,1}, ..., y_{i,q-1}\}$ can be associated to w_i, each one corresponding to as many legal categories $\{h_0, h_1, ..., h_{q-1}\}$. For example, being the language l= Italian and w_i= "dolo", possible translations in English are $y_{i,0}$="fraud" related to law category h_0="private law" and $y_{i,1}$="malice" related to law category h_1="criminal law". The right translation can be obtained only by knowing the sense, namely the category h_j, of the context in the query native language, where w_i is contained. Such a context, or legal category, is required and is provided by the user using a *dc:subject* field. When a category c_l (Fig. 1) is selected, the problem arises of different classification schemes in different languages, corresponding to different legal systems (Section 2). The problem can be solved by using a *thesaurus-based category mapping*. In fact, when the category c_l is submitted as a query parameter, the category c_l is mapped in the corresponding, or

the closest, categories in the "pivot" language, and from it to the other languages considered by the retrieval system ($c_l \Rightarrow c_{en} \Rightarrow \{c_{it}, c_{fr}\}$), using a classification schema. In accordance with Fig. 1 and without loosing generality, let us assume that only one legal category $c_{en} = h_1$ in the English legal system corresponds to the legal category c_l ($c_l \Rightarrow c_{en} = h_1$). Consequently, the English translation $y_{i,1}$ (Fig. 1) can be selected (in our example, the English word $y_{i,1}$="malice", related to law category h_1="criminal law" is selected as the right translation of the Italian word w_i="dolo"). If more than one category of the target legal system can be associated to c_l, all the corresponding translations of the current w_i are selected. When all the words of the current context are translated in *dc:description*, we obtain the translation of the submitted context $(w_0, w_1, ..., w_{n-1})_l$ from language l to retrieval system target languages. The category c_l is also mapped to the corresponding categories in the target languages. Now queries in different languages are ready to be dispatched to the related domain language indexes (Fig. 2).

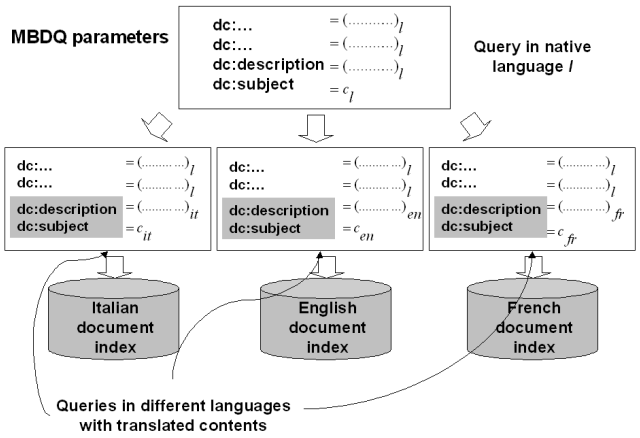

Fig. 2. MBDQ: results of query translation in different languages (in grey metadata whose content is translated)

4.2 Query based on keywords and legal categories (KBDQ+CBDQ)

A query based on keywords and legal categories represents the "simple search" modality of querying our multilingual retrieval system. In this modality the user is provided only with an unqualified text box to be filled with a context $(w_0, w_1, ..., w_{n-1})_l$ of words in a native language l. Words identifying the context will be translated into the target languages of the retrieval system (*thesaurus-based context translation*). Moreover, the user may provide a legal category of the query legal system. If the user selects a legal category c_l, among the values of *dc:subject* in the query legal system, a procedure of *thesaurus-based category*

mapping is executed, as described in Section 4.1, obtaining the correspondences of c_l in target legal systems (Fig. 1). If the user fills only the unqualified text box without choosing any value in *dc:subject*, since category is essential for translation, the right sense to the query context can be provided by a procedure of automatic word sense disambiguation, which assigns a legal category to a context as described in Section 5. The legal category thus identified in native query language, is then mapped to the related legal categories in target legal systems (*thesaurus-based category mapping*). At the end of the process, the right translations of ambiguous words can be obtained, as discussed in Section 4.1 (Fig. 1), and as many different queries as target languages considered can be dispatched to the different language indexes (Fig. 3).

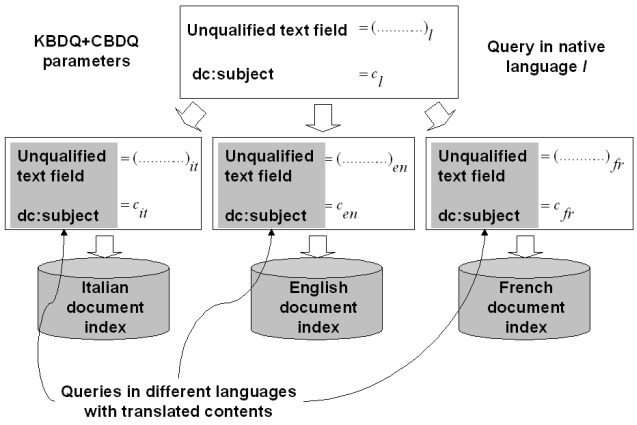

Fig. 3. KBDQ+CBDQ: results of query translation in different languages (in grey metadata whose content is translated).

5 Automatic word sense disambiguation

The problem of assigning the right meaning to a word in context is a problem of assigning the right sense to the context itself out of the various meanings that can be assigned to the ambiguous word. According to the literature lots of methods have been used to solve the problem of automatic disambiguation:

- Thesaurus-based disambiguation [12];
- Disambiguation based on sense definitions [13];
- Disambiguation based on translation in a second-language dictionary [14];
- Bayesian disambiguation [15].

In our retrieval system word disambiguation is a problem of context categorization with respect to the legal categories considered within a legal system.

Moreover [15] context categorization is the same problem of document categorization, once we view contexts as documents and word sense as categories. For these reasons in our system we can use different document categorization methods described in literature [16], trained with labelled documents of different legal categories of a particular legal system and language. At the end of the training phase each category profile (for example a vector of weighted terms relevant to it [17]) can also be considered as a context profile to be used for disambiguation function. While experimental results on legal document categorization have been carried on (see [17] [18]) a similar experiment on a set of previously categorized query is to be carried out. It is important to execute automatic word disambiguation prior to translation, because, as discussed, correct word translation depends on contextualization activity of words in their native language.

6 Conclusions

The approach analyzed in this contribution fully reflects the problems illustrated so far, as legal information retrieval is strongly conditioned to the legal orders' specificity, that is to the concepts on which they are based. It is a matter not so much of handling the diversity of languages in which these concepts are expressed, rather considering and managing the peculiarities of the law environment, that is the historical and cultural heritage of a given legal system, whose comparison with other legal orders is often hard, if not impossible. Therefore the real problem is how to establish a correspondence among concepts of diverse legal systems expressed in different languages. A comparative analysis of legal concepts and, parallel to this, the study of translation theory and practice to be intended as search of functional equivalents, are fundamental activities to reach a satisfactory mediation among different legal identities, thus ensuring intercultural communication and at the same time increasing the value of diversity, to be intended as a strength and a challenging factor of integration. Europe is a typical example of this phenomenon: it is praised for its strategies in language policy as a modern relevant experiment of institutional and political innovation which is in the position to open new forms of coexistence and cooperation. In this context multilingual legal information retrieval systems do represent the necessary tools to encourage multilingualism in the law domain and have the chance to make it effective. In particular, in this article an approach is proposed to offer the users a single point of access into multilanguage document collections where categories of law are the key factors to point to relevant material irrespective of the language used in a query. This is done through techniques able to translate legal queries to different target languages, disambiguating ambiguous words if needed. Basically, the approach gives the benefit of accessing multi-language legal documents respecting the identity and the peculiarities of different legal systems.

References

1. L. Wittgenstein, *Philosophical investigations.* Oxford : Blackwell, 1997.

2. R. Sacco, "Droit et langue," in *Rapports italiens au XV Congrès international de droit comparé*, 1998. Milano.
3. E. Francesconi and G. Peruginelli, "Opening the legal literature portal to multilingual access," in *Proceedings of the Dublin Core Conference*, pp. 37–44, 2004.
4. C. Peters and E. Picchi, "Across languages, across cultures: issues in multilinguality and digital libraries," *D-Lib Magazine*, 1997. Retrieved May 11, 2009, from (http://www.dlib.org/dlib/may97/peters/05peters.html).
5. H. Prakken and G. Sartor, "On the relation between legal language and legal argument: assumptions, applicability and dynamic priorities," in *Proceedings of the Fifth International Conference on Artificial Intelligence and Law*, pp. 1–10, New York: ACM, 1995.
6. D. Soergel, "Multilingual thesauri in cross-language text and speech retrieval," in *Working notes of AAAI Symposium on Cross-Language Text and Speech Retrieval*, 24-26 March 1997.
7. W. Peters, M. Sagri, and D. Tiscornia, "The structuring of legal knowledge in lois," *Artificial Intelligence and Law*, vol. 15, pp. 117–135, 2007.
8. A. Gangemi, M. Sagri, and D. Tiscornia, "A constructive framework for legal ontologies," in *Law and the Semantic Web* (Benjamins, Casanovas, Breuker, and Gangemi, eds.), Springer Verlag, 2005.
9. T. Agnoloni, L. Bacci, E. Francesconi, W. Peters, S. Montemagni, and G. Venturi, "A two-level knowledge approach to support multilingual legislative drafting," in *Law, Ontologies and the Semantic Web* (J. Breuker, P. Casanovas, M. Klein, and E. Francesconi, eds.), vol. 188 of *Frontiers in Artificial Intelligence and Applications*, pp. 177–198, IOS Press, 2009.
10. W. Lee, S. Sugimoto, M. Nagamori, T. Sakaguchi, and K. Tabata, "A subject gateway in multiple languages: a prototype development and lessons learned," in *DC*, pp. 59–66, 2003.
11. F. Sebastiani, "Interactive query expansion with automatically generated category-specific thesauri," in *Text Databases and Document Management: Theory and Practice, Idea Group Publishing*, pp. 103–117, Amita G. Chin (ed.), 2001. Hershey, US.
12. D. Yarowsky, "Word sense disambiguation using statistical models of rogets categories trained on large corpora," in *International Conference on Computational Linguistics*, pp. 454–460, 1992.
13. M. Lensk, "Automatic sense disambiguation," in *Proceedings of the SIGDOC Conference*, pp. 24–26, 1986.
14. I. Dagan and A. Itai, "Word sense disambiguation using a second language monolingual corpus," *Computational Linguistic*, no. 20, pp. 563–569, 1994.
15. W. A. Gale, W. K. Church, and D. Yarowsky, "A method for disambiguating word sense in a large corpus," *Computer and Humanities*, vol. 5, no. 26, pp. 415–439, 1993.
16. F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, 2002.
17. E. Francesconi and G. Peruginelli, "Access to italian legal literature: Integration between structured repositories and web documents," in *Proceedings of the Dublin Core Conference*, pp. 99–107, 2003.
18. E. Francesconi and G. Peruginelli, "Retrieval of italian legal literature: a case of semantic search using legal vocabulary," in *Proceedings of the Dublin Core Conference*, pp. 97–106, 2005.