Automatic Translation in Cross-Lingual Access to Legislative Databases

Catherine Bounsaythip, Aarno Lehtola, Jarno Tenni
VTT Information Technology P. Box 1201, FIN-02044 VTT, Finland
Phone: +358 9 456 5957. Fax: +358 9 456 6027.
Email: {catherine.bounsaythip, aarno.lehtola, jarno.tenni}@vtt.fi
http://www.vtt.fi/tte

Abstract:
This paper considers the use of controlled languages for query translation in a legislative document retrieval system. Problem statement and analysis of the approach are described. The use of controlled languages is motivated by the fact that precision is very important in our case. In many information retrieval systems, the use of unrestricted language resources such as general purpose machine translation or bilingual lexica, provides better recall at the expense of precision. Ambiguities and polysemy make the search engine retrieve irrelevant documents as semantic knowledge is missing. Controlled languages help to better specify the word sense according to the domain of interest. Thus ambiguities are avoided and polysemy is specified according to the domain. We will implement our idea in the area of VAT regulation in Europe.

Introduction

Nowadays, information on European legislation and intergovernmental agreements are scattered in distributed repositories in heterogeneous formats and in many languages. This information is technically accessible through information networks, but it is extremely difficult even for professionals to use it because of differences in document structures and languages. This is a common problem in cross-lingual information retrieval (IR) systems where queries are made in one language to a document collection in several different languages and the goal is to retrieve only those documents relevant to the query. Before retrievals can be performed, deep linguistic analysis and translation of the query appears necessary.

Natural language processing in IR systems is special in the sense that a pattern of term occurrences in a document generally suffices to determine the subject matter; as word order is largely irrelevant. Because of ambiguities and polysemy, query translation is not a trivial task. One way to ensure the performance of the system is to control the query construction. This approach is discussed in this paper where we present our machine translation software called Webtran.

Webtran is a machine translation system for controlled languages (CL) to be embedded in WWW-based information service systems (Lehtola et al. 1998a). It is designed to support fully automatic translation in online WWW services, such as online mail order catalogue or information retrieval from cross-lingual databases. The framework in which Webtran is involved, consists of an interface through which the user can make queries in one language to search for legislative texts from different EU databases of EU regulatory information. In **Figure 1**, we show the basic architecture of the system. As existing repositories are located in different countries and stored in different formats, it is necessary to convert the request into the formats of the targeted databases.

The user can make queries in his own language and his request is translated by Webtran Translator into the language of the target documents before being directed to the multilingual databases. Retrieved documents will be displayed in their original language. In the domain of

legislation, usually users prefer to have the texts in their original language so that the interpretation is more reliable. Moreover, it is usually out of question to translate the whole text in an automatic way, as some legal terms have different meanings according to country and according to the area of laws. In any case, translations of legal texts need to be authorised to avoid misinterpretations. Therefore, only the query terms will be handled by the Webtran translator. Possibly, Webtran can provide approximate translation of some meta-information related to the document (e.g., headers, titles, summary or keywords).

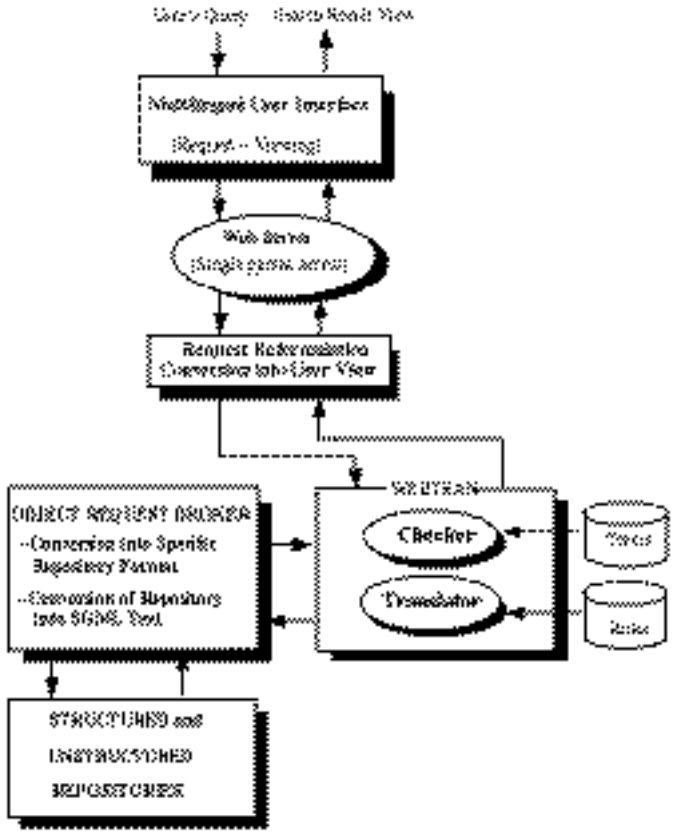

Figure 1 *: A cross-lingual IR architecture for accessing and viewing EU legislative databases.*

Cross-Lingual Information Retrieval

Language technology is important in cross-lingual document retrieval systems. In TITAN system (Kikui et al. 1996), the language processor contains language identifier (English / Japanese) and bilingual dictionaries. The user can make requests in Japanese or in English and the URLs found are displayed with their headers translated into the query language. In EMIR (Fluhr et al. 1996), SYSTRAN is used in the language processing part of the retrieving system.

Translation of queries and keywords does not need just multilingual machine-readable dictionaries as many ambiguous terms and polysemy may appear. Many approaches have been used, such as interlingua (Landauer 1990), alignment of large parallel text corpora in different languages (Davis and Dunning 1995), concept-based (Chen 1993) and controlled vocabulary (Soergel 1997).

In (Landauer 1990), an approach for fully automatic cross-language document retrieval was presented. Their system is based on a language-independent representation where no humanly constructed dictionary, thesaurus, or term bank are needed. The construction of the interlingua

is based on a statistical method using paragraph alignment of a sample collection of parallel texts. This is done once for a subject area. Each word in the sample is then assigned a vector value determined by the total pattern of usage of all the words in all the sample paragraphs. In the second step, a new document or query in any of the original languages is assigned a vector value that is an average of the values of the words it contains. Tests on a French-English corpus showed that the method works well, because the two languages are quite close to each other. This wouldn't work for example between Finnish and Swedish.

The approach to query translation in multilingual IR systems in (Davis and Dunning 1995) used evolutionary programming to optimise the construction of a query from bilingual dictionaries. The assignment of term weights is done by means of a population of potential weighting schemes to generate translated queries. Sentence-level alignments from a large parallel text collections were used to evaluate the correctness of a query translation. The approach is based on the consideration that translated queries are primarily derived by a mapping from a word set in the query language to a word set in the language of the derived query. They reported good results for the case where the original query is closely related to the document collection. Results are unclear for queries that are not closely related to the documents. Moreover evolutionary optimisation for discovering optimal queries using a parallel training corpus takes too much time for ``on-line" IR systems.

In (Chen 1993), the system was based on concept exploration. Concepts are extracted from the keywords used in the set of user-selected documents and Genetic Algorithm (GA) was used to perform concept optimisation. The optimisation is based on the relevance of each document to other documents in the user-selected set. A document which included more concepts shared by other documents had a higher score. The optimised chromosome contained relevant keywords which best described the initial set of documents. Then, the optimised concepts are put into a Hopfield Network to activate other relevant concepts, e.g., when the user selected a new document. The new keyword was then used to identify more relevant documents and the GA/HP process continued.

For performance and simplicity, many systems avoid sophisticated linguistic analysis of the documents by imposing a specialised ``controlled language" (Oard 1997). In (Soergel 1997), a multilingual thesaurus is built to relate the selected terms from each language to a common set of language_independent concept identifiers, and document selection is based on concept identifier matching. The user is assisted for specifying from a semantic field the term that best describes his intended meaning.

Webtran IR Approach

For low-cost services of the access to the legislative databases through WWW search engines, it is necessary that fully automatic translation achieves a reasonable performance. To do so, the approach adopted by Webtran is based on controlled vocabulary. This would help to relate terms from each language to a common set of language dependent concept identifiers. By the word concept we mean in this paper interrelated items in a conceptual model, that have been defined by humans for the target domain. At the language level a concept can be expressed by a term and its synonyms which can be single words or longer surface expressions. Term is the most obvious or most widely agreed expression of the synonyms. Then there can be semantically close expressions that are not accurate but approximately reflect the meaning of the term.

For example, the official term for the concept of "*avoiding payroll tax*" in Finnish is "ennakonpidätysvelvollisuudesta vapauttaminen", and one way of expressing it can be "ennakonpidätyksen välttäminen". Expressions of a term in different languages can also be

viewed as synonyms. **Table 1** shows an example of Finnish and Swedish surface expressions of a concept.

Finnish: "ennakonpidätysvelvollisuudesta vapauttaminen"
Swedish: "befrielse från skyldighet att verkställa förskottsinnehållning"
English (approx.): *acquitting from the responsibility of paying payroll tax.*

Table 1 : *Example of surface expressions of a concept as used in legislation.*

The use of concepts for query translation can enhance the retrieval performance. For example, an inexperienced user would likely make a request in Finnish about *avoiding payroll tax* to a Swedish database with: "ennakonpidätyksen välttäminen". The actual official term used in legal texts is: "ennakonpidätysvelvollisuudesta vapauttaminen". With a plain Boolean operator, the search may fail if the translation of "välttäminen" does not match in the target database. If semantically related words of the domain are not considered, precision of the retrieval is decreased. Kekäläinen and Järvelin (1998) have shown that expansion of queries into concepts and synonyms yielded better retrieval scores.

To achieve that, the system requires experts in legislation to define the conceptual models and relationships to surface expressions in the covered languages. This will be used for helping the construction of queries in a controlled way. The creation conceptual models can be done by analysing the existing repositories to create dictionaries of common elements or by aligning parallel texts in different EU languages.

Besides, for the end-users, it is not easy to find the proper term for making a request in the legislation domain and especially from foreign text databases. For instance, if the document is in Swedish, the system should help the user in finding correct search terminology by providing an automatic translation of search sentences from the user's native language to Swedish. If the documentation were not available in Swedish, the system should assist in translating the search terms provided by the users to proper search terms in the local language of the document database.

A help system will be developed to assist the user in defining the proper search term. In such a situation, a multilingual thesaurus can be used. One word chosen by the user can trigger inference of new words by the conceptual model. The help interface should be easy enough so that the user is not required to be trained in order to effectively select proper search terms and to exploit thesaurus relationships. These observations indicate that the user interface must be designed to adapt to the needs of each category of users (see, e.g., Lehtola et al. 1998b).

A user interface for developing controlled languages can be found in (Lehtola et al. 1998a). An IR user interface will be based on a WWW browser as the example shown in **Figure 2**. In this sketchy example, a click on button "HAE" would send commands to the translation component of Webtran. It is transparent to the user. The click of "APU" would trigger the opening of the help interface. This interface would share some functionality of the user interface built for controlled language designer described in (Lehtola et al. 1998a).

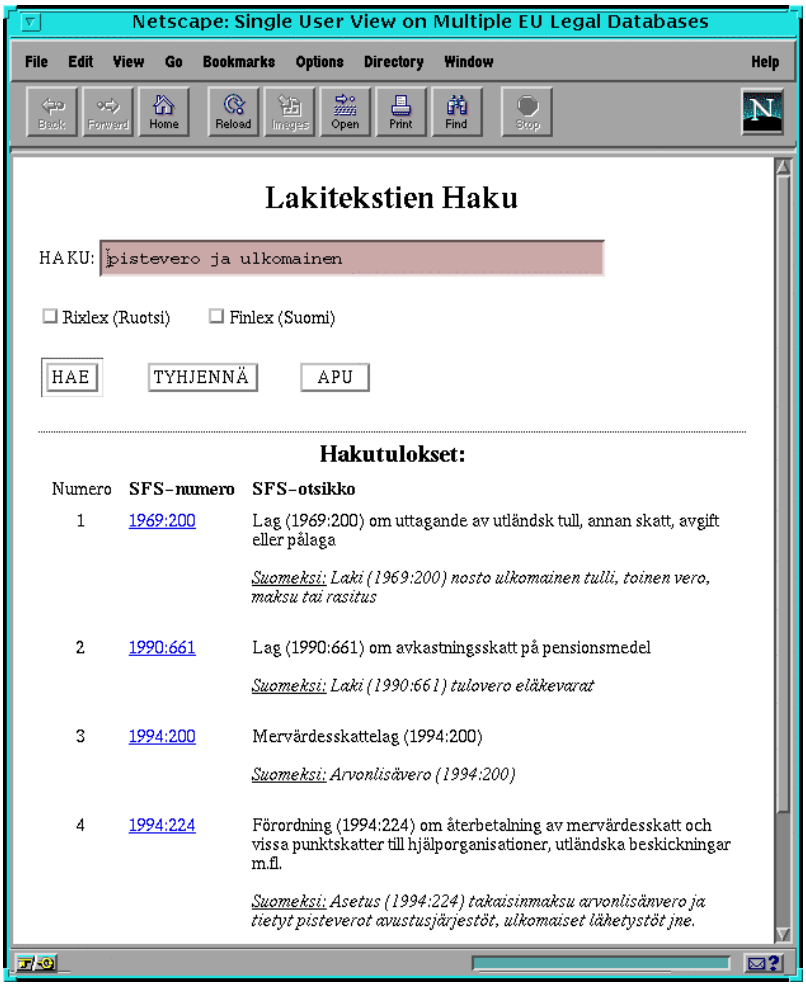

Figure 2 *: Sketch of a WWW-based cross-lingual information access interface. In this illustrative example, the query is made in Finnish to Swedish legislation databases.*

Conclusion

In this paper, we have described our ongoing work about using controlled languages for cross-lingual access to legislative databases. As a term may have different meanings in different areas of laws, controlled language should be designed for a specific domain of law. Prototype system will focus on VAT regulation texts from different European countries. We are now on the phase of gathering corpora in this domain in order to build controlled languages in Finnish and Swedish.

Acknowledgements

The authors would like to thank the Technical Development Centre of Finland (TEKES), Tieto Corporation Ltd., and Ellos Ltd., all from Finland, for supporting our work in many ways. Also, many thanks to Prof. Seppo Linnainmaa, Prof. Timo Honkela and, Kuldar Taveter for their useful comments on this paper.

References

Chen, H. (1994). GANNET: Information Retrieval Using Genetic Algorithms and Neural Nets. Working Paper, Center for Management of Information, College of Business and Public

Administration, University of Arizona, CMI-WPS.
http://ai.bpa.arizona.edu/papers/gannet93/gannet93.html

Davis, M. W., Dunning ,T. E. (1995), Query Translation Using Evolutionary Programming for Multi-lingual Information Retrieval, *Proceedings of the Fourth Annual Conference on Evolutionary Programming*, San Diego, CA, March.

Fluhr, C. Schmit, D. Ortet, P. Elkateb, F. Gurtner, K. (1996). Distributed Multilingual Information Retrieval, MULSAIC'96, Multilingual in Software Engineering: AI Contribution.
http://www.iit.nrcps.ariadne-t.gr/ costass/mulsaic.html

Kekäläinen, J., Järvelin, K. (1998). The Impact of Query Structure and Query Expansion on Retrieval Performance. In: Croft, W. B. & Moffat, A. & van Rijsbergen, C.J. & Wilkinson, R. & Zobel, J. (Eds.), *Proc. the 21th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (ACM SIGIR '98), Melbourne, Australia, August 23-28, 1998. New York, NY: ACM Press.

Kikui, G., Hayashi, Y. and Suzaki, S. (1996). Cross-Lingual Information Retrieval on the WWW, MULSAIC'96, Multilingual in Software Engineering: AI.
Paper: http://www.iit.nrcps.ariadne-t.gr/ costass/mulsaic.html.
TITAN URL: http://sting.navi.ntt.co.jp/titan/titan-e.html

Lehtola, A., Tenni, J., Bounsaythip, C. (1998a). Definition of a Controlled Language Based on Augmented Lexical Entries. *Proceedings of the Controlled Language Applications Workshop 98*, Carnegie Mellon, Pittsburg, USA, 21-22 May, 1998, pp. 16-29.

Lehtola, A., Tenni, J., Bounsaythip, C. (1998b). Controlled Language Technology in Multilingual User Interfaces. To appear in *Proceedings of the 4th ERCIM Workshop User Interface for All (UI4All)*, Special Theme: "Towards an Accessible Web", Stockholm, Sweden, 19-21 October, 1998.

Landauer, T. K. and Littman, M.L. (1990). Fully Automatic Cross-Language Document Retrieval Using Latent Semantic Indexing. *Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research*, pages 31-38.
UW Centre for the New OED and Text Research, Waterloo Ontario, October.
http://www.cs.duke.edu/~mlittman/docs/refer.html

Oard, D. (1997). Alternative Approaches for Cross-Language Text Retrieval. *AAAI Symposium on Crosslanguage and speech retrieval*, March 24-26.
http://www.ee.umd.edu/medlab/mlir/

Soergel, D. (1997). Multilingual Thesauri in Cross-Language Text and Speech Retrieval", *AAAI Symposium on Crosslanguage and speech retrieval*, March 24-26.
http://www.ee.umd.edu/medlab/mlir/