[*First International Conference on Language Resources & Evaluation*, Granada, Spain, 28-30 May 1998]

Problems and Techniques of Cross Language Information Retrieval

Gregory Grefenstette
Xerox Research Centre Europe
6 chemin de Maupertuis
38240 Meylan, FRANCE
[Gregory.Grefenstette@xrce.xerox.com]

Abstract

Cross Language Information Retrieval concerns the problem of finding foreign language documents after using your own native language to write your information request. It supposes that documents are indexed in their original language, and that you do not have the possibility to translate all the documents into your native language but must work with these indexes. We present here a short overview of the particular problems caused by this situation and some techniques that have been proposed to attack it

Introduction

Cross Language Information Retrieval (CLIR) covers the research and techniques for finding documents that are written and indexed in a foreign language using a native language query to express the information request. It involves elements of Machine Translation (MT): finding possible translations of the words and terms appearing in the original request; and elements of Information Retrieval: how words should be normalized in order to match stored indexes, how terms should be weighted in a query. It also involves novel elements such as automatically identifying the language of a document given its text (so that same language documents can be indexed together), and deciding how many of the translations found for each query term should be retained. Cross Language Information Retrieval (Grefenstette, 1998) is both easier and harder than Machine Translation. It is easier because Machine Translation systems must both (1) chose only one translation alternative for each input term, and (2) produce a syntactically correct output for each input sentence; information retrieval systems, on the other hand, function using a *bag of words* technique[1]: the words in a query do not have be in any syntactic order, and the more words in the query, the better the result. A CLIR system producing a foreign language query from a native language query then does not have to produce syntactically correct output in the target language, and can retain more than one translation alternative for each original query term.

At the same time, CLIR is harder than MT since MT systems (Hauenschild & Heizmann, 1997) have been most successful when they are designed for restricted domains. Information retrieval (Frakes & Baeza-Yates, 1992), on the other hand, has always concentrated on domain independent techniques, employing techniques that are meant to be applicable to any text type or subject. Cross Language Information Retrieval methods, then, are not limited to certain domains and need a treatment of vocabulary that is much more robust than machine translation.

Language Identification

The WWW, with all its chaotic and rapid development and all its different language documents that have suddenly become freely available to anyone with a computer and modem, has been the stimulus for CLIR. The Web can be seen a large multilingual distributed database. Given a large database with many different languages in it, from a text indexing point of view, one should be able to identify the language of a document before it is indexed, since the index terms usually undergo some type of morphological normalisation before they indexed. When one searches for *dogs,* one wants to find documents with *dog,* too. Though this is rarely the case with current Web browsers, both techniques for identifying languages (Grefenstette, 1995) and tools for normalizing identified text (Grefenstette & Segond, 1997) exist[2]. CLIR techniques usually suppose that the document collection has been partitioned and indexed in separate languages.

The Three Problems of CLIR

After the problems of language identification are resolved, there are three new problems that a CLIR must.solve in order to use a query written in one language to find documents written in another. First, it must know how a term given in one language might be written in the target language. What are the possible translations of each original query term? Secondly, it must decide how many of the possible

[1] Classical information retrieval systems consider both documents and queries as simple bags of words, and try to match up bags which have the most similar items.

[2] For example, Altavista identifies the language of documents, but performs no normalisation of its index terms.

translation alternatives should be kept. Can we eliminate some of the translation possibilities? And third, it must deal with the problem caused by retaining more than one translation probability because information retrieval systems, in their bag of words approach, will give more importance to a term that contributes many query alternatives than to a term than contributes only one translation.

Finding translations

There are two solutions to finding translation terms: One involves using a bilingual dictionary which lists the possible translation terms (Hull & Grefenstette, 1996). Another involves using parallel corpora of text (Sheridan & Ballerini, 1996; Landauer *et al* 1990) Accessing dictionaries[3] gives rise to a number of subsidiary problems: spelling variants, derivational variants, coverage of vocabulary, treatment of proper names. In the case of using parallel corpora, the dictionary problem can be avoided in the following way. The original language query can be posed on original language documents, and relevant documents retrieved. The parallel, target language documents corresponding to the results of this first retrieval can be collated into one huge bag of target language words that can serve as new target language query without using an explicit translation dictionary.

Pruning Translation Alternatives

Once target language terms are found using a method such as the above, it is sometimes useful to eliminate some translations which would introduce noise into the query. For example, among the translations returned the French *voiture* in a common bilingual dictionary, we find archaic translations such *carriage.* It is possible to leave such words in place if one knows ahead of time that the corpus itself will act as a filter (Fluhr et al., 1998) since the odd translations will never appear in the target corpus. More elaborate techniques (Davis, 1998) involve using again a parallel corpus to do a first query using the original language query on original language documents, and then using the parallel target language documents to filter out query alternatives, only keeping those alternatives that appear in the highly ranked parallel target language documents. The retained translation terms then serve to create a new target language query on a new, unseen and monolingual, target language database.

Weighting Translation Alternatives

When more than one translation alternative is retained for an original term, the IR system that will treat the target language query must account for this reduplication. Hull (1998) proposes a weighted Boolean retrieval technique that answers this problem.

Conclusion

We have briefly sketched some of the important and unique problems that CLIR poses, being at the cross roads of both MT and IR. This is a new area of research made pertinent by the appearance of the inherently multilingual World-Wide Web.

References

Davis, M. (1998) "On the Effective use of Large Parallel Corpora in CLIR" in Grefenstette (ed) *Cross Language Information Retrieval*, pp. 11-21

Frakes, W. & Baeza-Yates, R. (1992) *Information Retrieval: Data Structures and Algorithms.* New Jersey: Prentice Hall.

Fluhr, C., Schmit, D., Ortet, P., Elkateb, F. Gurtner, K. & Radwan, K. (1998) "Distributed Cross Lingual Information Retrieval" in Grefenstette (ed) *Cross Language Information Retrieval*, pp.41-50

Grefenstette, G. (1995) "Comparing Two Language Identification Schemes" in *Proceedings of the 3rd International Conference on the Statistical Analysis of Textual Data* Dec 11-13, Rome: JADT'95

Grefenstette, G.& Segond, F.(1997). "Multilingual Natural Language Processing" in *International Journal of Corpus Linguistics*, 2(1). John Benjamin Publishers

Grefenstette, G. Hull, D. A., Gaussier, E. &. Schulze, B. M. (1997) "Xerox TREC-6 Site Report: Cross Language Text Retrieval" *TREC-6 Conference Working Notes,* INSTN, Gaithersburg, MD.

Hull D. "A weighted Boolean Model for CLIR" in. Grefenstette, ed. *Cross Language Information Retrieval Boston,* :Kluwer Academic p. 119-135.

Hauenschild, C. & Heizmann, S., editors (1997). *Machine Translation and Translation Theory* Berlin: Mouton de Gruyter.

Hull, D. & Grefenstette, G.(1996). "Querying across languages: A dictionary-based approach to multilingual information retrieval" In *SIGIR Proceedings*, Zurich, Switzerland, ETH.

Landauer, T.K. & Littman, L.M. "Fully Automatic Cross-Language Document Retrieval Using Latent Semantic Indexing" in 6th *Conference of the University of Waterloo Centre for the New Oxford Dictionary and Text Research*, Waterloo. pp 31-38.

Sheridan, P. & Ballerini, J.P. (1996) "Experiments in Multilingual Information Retrieval using the SPIDER system" in *SIGIR'96,* Zurich, pp. 58-65.

[3] Multilingual dictionary resources themselves are difficult to find, though some organisations, such as the ELRA, are making them more easily available to the research and industrial community.