67th IFLA Council and General Conference
August 16-25, 2001

Code Number: **099-183(WS)-E**
Division Number: **VI**
Professional Group: Information Technology **Workshop**
Joint Meeting with: -
Meeting Number: 183
Simultaneous Interpretation: -

Multilingual access for information systems

Carol Peters
Istituto di Elaborazione della Informazione, CNR
Pisa, Italy
E-mail: carol@iei.pi.cnr.it

Páraic Sheridan
MNIS-TextWise Labs, Syracuse
New York, NY, USA
E-mail: paraic@textwise.com

Abstract:

With the rapid growth of the global information society, the concept of library has evolved to embrace all kinds of information collections, on all kinds of storage media, and using many different access methods. The users of today's information networks and digital libraries, no longer restricted by geographic or spatial boundaries, want to be able to find, retrieve and understand relevant information wherever and in whatever language it may have been stored. For this reason, much attention has been given over the past few years to the study and development of tools and technologies for multilingual information access (MLIA). The tutorial will provide participants with an overview of the main issues of interest in this sector. Topics covered will include: character encoding, specific requirements of particular languages and scripts, localization and presentation issues, techniques for cross-language retrieval, the importance of resources.

1. Introduction

The global information society has radically changed the way in which knowledge is acquired, disseminated and exchanged and is rapidly bringing about a revolution in the library world. Users of internationally distributed networked collections need to be able to find, retrieve and understand relevant information in whatever language and form it may have been stored. Many users have some foreign language knowledge, but their proficiency may not be good enough to formulate queries that appropriately express their information needs. Such users will benefit enormously if they can enter their queries in their native language, because they are able to examine and extract information from relevant documents even if they are not translated. Monolingual users, on the other hand, can use translation aids to help them understand their search results in a second language.

For this reason, much attention has been given over the past years to the study and development of methodologies and tools for multilingual information access (MLIA) and cross-language information retrieval (CLIR). This is a complex multidisciplinary area in which natural language processing and information retrieval techniques converge. The aim of the tutorial will be to contribute to an awareness of the issues involved and the different components needed to build effective multilingual interfaces and cross-language search tools for digital library systems. This paper provides a brief outline of the main topics that will be covered. For a more detailed discussion, the reader is referred to [1].

2. Multilingual Text Processing

In information retrieval, a representation of the text to be searched is usually obtained by extracting 'indexing features' from a document collection or from the text of a user's query. In a simple approach, this extraction process consists of four basic steps: conversion of characters, extraction of words (tokenization), removal of 'stopwords', and normalization of remaining content words. While these processing steps have been studied extensively in the context of retrieval of English texts, new challenges are presented when access to information in multiple languages is involved.

Language Recognition: Since processing texts to extract indexing features often involves steps which use language-specific knowledge, it is important to first establish the language of the text, if that is not already known. To date many different approaches have been used to address the problem of language identification in general texts. Approaches which have been tried range from approaches relying on the presence of individual characters [2] in texts, or the presence of particular character N-grams [3] or even on the presence of given words [4]. A common approach to language identification specifically for multilingual access uses language-specific stopwords to identify the language of texts [5].

Character Encoding: While most of the Western European languages are all covered under the standard ISO-8859-1 (Latin-1) encoding scheme, multilingual access to non-Roman languages involves addressing the issue of document encoding. The encoding of a language, or more specifically the encoding of the character set used to represent the alphabet of a given language's script, specifies the mapping between the written script and its binary representation. A character encoding is therefore specific to a given alphabet and in many cases there exist multiple mappings for a given alphabet (for example the Cyrillic alphabet used in Russian). Depending on the number of characters needed in the representation of a language, an encoding scheme can be based on a single byte (e.g. German) or must require a double-byte encoding (e.g. Chinese).

In an attempt to provide a single encoding scheme for mapping all of the world's languages, the UNICODE consortium (www.unicode.org) has designed the UNICODE standard. This provides a character encoding system designed to support the interchange, processing and display of the written texts

of the diverse languages of the modern world. In processing texts for multilingual access, it is common for UNICODE compliant systems to use standard libraries to convert the native encoding of texts (e.g. Shift-JIS for Japanese) into a UNICODE format (e.g. UTF-8) as a single standard representation.

Language Specific Tokenization: Once the language of a text has been determined and the character encoding has been standardized, the next step is to identify the specific words being used. While in many languages this is straightforward because of the use of spaces to delimit words, many languages compound or concatenate words together to form new compound words. In the most difficult cases, no spaces are used between words in text (e.g. Japanese, Chinese) so that the tokenization process must determine all word boundaries. In this case, a dictionary or lexicon of valid words in the language is typically used to determine legal words. A process is then used whereby a sentence of text is scanned in order to find the set of words from the dictionary which provides full coverage of the characters found in the text. In the tokenization stage, punctuation is also removed from words and hyphens between word segments are processed.

In order to reduce the number of indexing features to be included in the representation of a text, words which have little content value are often discarded; so-called 'stopwords', e.g. *the* and *at* in English. Since between 30% and 50% of the words of a text may be included in a stopword list, the removal of such words can have a significant impact on the retrieval index. Stopwords in any given language are usually easily determined on the basis of either parts-of-speech (e.g. determiners, prepositions) or of high frequency within a sample text.

Word Normalization: The final step in processing text for retrieval indexing involves the normalization of content words remaining after stopword removal. The most common form of normalization involves reducing words to a *stem* form by removing suffixes or inflections. In the simplest case, a stemming algorithm simply removes standard suffixes (e.g. '-s', '-es', '-ation' in English) in an iterative process until the shortest form remains. The best known algorithm of this type was developed by Porter [6] for English and similar algorithms have been developed for other languages [5]. An alternative is to perform a more linguistically motivated *morphological* analysis of the text in order to identify the root forms of the words. Word normalization results in greater efficiency in both the text processing and text retrieval processes. This is particularly true when dealing with European languages, which have a much richer inflectional morphology than English.

In the normalization phase, it is also common, especially in the context of multilingual access, to attempt to identify multi-word phrases as individual index features so that phrases can be translated as a unit rather than as individual words. In many cases, a word-by-word translation of a phrase does not render a true translation (e.g. translate *'fast food'* into French or German). Phrases can be identified in a text by matching against a dictionary or lexicon of known phrases, or by using statistical procedures which recognise frequently co-occurring words as potential phrases.

Summing-up: The steps used in multilingual text processing will depend on the languages involved and on the overall approach being used within a given system for multilingual information access. The index features which result from the text processing phase must be compatible with the resource that is being used to match queries in one language to documents in possibly many other languages. It is therefore important to have an understanding of the different approaches to cross-language information retrieval and the kinds of resources used in each approach.

3. Approaches to Cross-Language Text Retrieval

Basically, in cross-language text retrieval the task is to develop methods which successfully match queries against documents over languages and rank the retrieved documents in order of relevance. In monolingual

retrieval, the traditional way to do this is through some kind of word matching and weighting; with cross-language text retrieval we have the additional problem of matching (and weighting) words across languages. This implies employing some kind of resource in order to translate from the language of the query to that of the documents or vice versa, and addressing the problem of sense disambiguation, already present in monolingual retrieval but greatly increased when mapping over languages. Three main approaches have been experimented: machine translation; knowledge-based techniques (i.e. thesauri or dictionaries); corpus-based techniques. Each of these methods has given promising results but also has disadvantages associated with it.

Machine Translation: Full machine translation (MT) is not viewed as a realistic answer to the problem of matching documents and queries over languages. The goal of an MT system is to produce a readable and reliable target language version of a source text, whereas cross-language retrieval aims at finding sufficient similarities between a source language query and a target language document in order to be able to claim that the document is more or less relevant to the information needs expressed by the query. The translating of entire collections of documents into another language (that of the query) is thus not only very expensive, but also involves a number of tasks that are redundant from the purely retrieval viewpoint, e.g. encoding of linguistic, semantic and pragmatic information.

Efforts using MT systems have thus concentrated on attempting to translate the queries rather than the documents. However, queries are usually sets of words with little or no syntactic structure. Therefore, input cannot be parsed by an MT system and traditional methods of word-sense disambiguation cannot be applied as there is no semantically coherent text. Accurate translation is thus not possible but also not necessary. There is no need for a linearly coherent and unique output, in fact multiple translations of query terms can provide a form of query expansion that can improve performance. It has been shown that simpler and less resource costly techniques can work at least as effectively and that, for query translation, dictionary-based techniques can outperform commercial MT systems [7].

Multilingual Thesauri: Early experiments showed that multilingual thesauri can give acceptable results for cross-language retrieval and there are now a number of thesaurus-based systems available commercially. A multilingual thesaurus for indexing and searching with a controlled vocabulary can be seen as a set of monolingual thesauri that all map to a common system of concepts. With a controlled vocabulary, there is a defined set of concepts used in indexing and searching. In this way, the problem of ambiguity is eliminated. Users can use a term in their own language to find the corresponding concept identifier in order to retrieve documents in another language. In the simplest system, this can be achieved through manual look-up in a thesaurus that includes, for each concept, corresponding terms from several languages and has an index for each language. In more sophisticated systems, the mapping from term to descriptor would be done internally [8].

With the controlled vocabulary approach, appropriate terms from the vocabulary must be assigned to each document in the collection. Traditionally this was done manually by experts in the field. This is expensive. Methods are now being developed for the (semi)automatic assignation of these indicators. The fact remains that thesauri are expensive to build, costly to maintain and difficult to update. Furthermore, it has been found to be quite difficult to train users to effectively exploit the thesaurus relationships.

In any case, the current trend is away from controlled vocabulary searching in favour of free text searching even though, from many viewpoints, cross-language free-text searching is a more complex task. It requires that each term in the query be mapped to a set of search terms in the language of the texts, possibly attaching weights to each search term expressing the degree to which occurrence of a search term in a text would contribute to the relevance of the text to the query term. The greater difficulty of free-text cross-language retrieval stems from the fact that one is working with actual usage while in controlled-

vocabulary retrieval usage can, to some extent, be dictated. On the other hand, the query potential is greater than with a controlled vocabulary.

Using Dictionaries: Many free-text cross-language systems use bilingual machine-readable dictionaries (MRDs) as their transfer resource. Such resources are becoming increasingly available both commercially and on-line. As they have generally been prepared for human use, they require some kind of pre-processing before they can be used in an automatic system. This essentially implies analysing the mark-up information to identify the different lexical information: headwords, parts-of-speech, sense division, translation equivalents, etc.

It has been demonstrated that straightforward dictionary-based query translation, where each term or phrase in the query is replaced by a list of all its possible translations, represents an acceptable first pass at cross-language information retrieval although such -- relatively simple -- methods clearly show performance below that of monolingual retrieval. Automatic MRD query translation has been found to lead to a drop in effectiveness of 40-60% of monolingual retrieval [9][10]. There are three main reasons for this: (i) general purpose dictionaries do not normally contain specialised vocabulary; (ii) failure to translate multiword terms; (iii) the problem of ambiguity.

Perhaps the greatest problem using MRDs is coping with ambiguity. In word-by-word dictionary translation, each word is replaced by all possible translation equivalents. When the query term is polysemous and thus in itself ambiguous, this can result in a large set of target search terms, many of which are spurious and will contribute to the retrieval of irrelevant documents. In sentence or document translation, the context provides information that can be used for disambiguation; the shortness of the average query means that there is a lack of context for this scope. Current research work gives considerable attention to this question.

It has been shown that both syntactic and statistical methods can significantly reduce the effects of ambiguity and bring the effectiveness of cross-language retrieval near the level of monolingual retrieval. Well-formed queries can be tagged by part-of-speech taggers to eliminate grammatical homonyms and thus reduce the number of incorrect target terms generated by the dictionary. In particular, query expansion techniques have been shown to help considerably in reducing ambiguity. Basically, such techniques add new terms, selected according to a given criteria, in order to make the query more precise. See, for example, [11], [12].

Corpus-based Techniques: Corpus-based approaches analyse large collections of texts on a statistical basis and automatically extract the information needed to construct application-specific translation techniques. The collections analysed may consist of parallel (translation equivalent) or comparable (domain-specific) sets of documents. The main approaches that have been experimented using corpora are vector space and probabilistic techniques.

The first tests with parallel corpora were on statistical methods for the extraction of multilingual term equivalence data which could be used as input for the lexical component of MT systems. The problem with using parallel texts as training corpora is that test corpora are usually domain-specific and costly to acquire -- it is difficult to find already existing translations of the right kind of documents and translated versions are expensive to create. For this reason, there has been a lot of interest in the potential of comparable corpora.

A comparable document collection is one in which documents are aligned on the basis of the similarity between the topics they address rather than because they are translation equivalent. The requirement is that they are similar in genre, register, and period. The basic idea underlying the use of such corpora is that the words used to describe a particular topic will be related semantically across languages.

The best known cross-language strategy using comparable corpora is the multilingual similarity thesaurus approach. [13] reports results using a reference corpus created by aligning news stories from the Swiss news agency (SDA) in German and Italian by topic label and date and then merging them to build the "similarity thesaurus". German queries were then tested over a large collection of Italian documents. The results of this approach are promising, especially when used on a domain-specific collection [14].

A strong disadvantage of corpus-based techniques is that they tend to be very application dependent. New reference corpora are needed for new domains.

Summing-up: With the current-state-of-the-art, all the above approaches if implemented in a well-designed, tested and tuned system can be expected to achieve approximately 80% of monolingual effectiveness in the general domain. However, as can be seen from this brief overview, any single method for cross-language retrieval presents limitations. Whatever the method chosen, the resources used to provide the means for mapping between query and collection are a major factor towards successful retrieval. Already existing resources -- such as electronic bilingual dictionaries -- are normally inadequate for the purpose; the building of specific resources such as thesauri and training corpus is expensive and such resources are generally not fully reusable; a new multilingual application will require the construction of new resources or considerable work on the adaptation of previously built ones.

It should also be noted than most systems currently in use concentrate on pairs rather than multiples of languages. This is hardly surprising. The situation is far more complex when we attempt to achieve effective retrieval over a number of languages than over a single pair; it is necessary to study some kind of interlingual mechanism -- at a more or less conceptual level -- in order to permit multiple cross-language transfer. In a conceptual interlingua, terms and phrases from multiple languages which refer to the same concept are mapped into a language-independent scheme. In this way it is possible to match to equivalent terms in all languages and to achieve CLIR in any of the language combinations, not just pair-wise. However, the building of a such a resource is not an easy task and much work remains to be done before we can talk about truly multilingual retrieval systems.

4. Cross-Language System Evaluation Campaigns

System evaluation activities play an important role in stimulating system development. This is particularly true for cross-language retrieval systems which are still very much in the experimental stage. There are currently several international activities in this area.

- TREC - Text Retrieval Conference Series (http://trec.nist.gov/) includes a cross-language track this year for English and French to Arabic.
- CLEF – Cross language Evaluation Forum (http://www.clef-campaign.org/). CLEF sponsored by the European Commission as part of the DELOS Network of Excellence for Digital Libraries is an evaluation activity for European languages and represents the continuation of the CLIR track begun at TREC in 1997. CLEF 2001 has four tasks for multilingual, bilingual, domain-specific and monolingual (non-English) text retrieval evaluation. This year's multilingual document collection has comparable newspaper corpora for six languages (Dutch, English, French, German, Italian, Spanish) and topics in 10 European languages and 3 Asian ones.
- NTCIR: NACSIS Test Collection for Information Retrieval (http://www.rd.nacsis.ac.jp/) hosted by the National Institute for Informatics, Tokyo. NTCIR includes cross-language tasks for Chinese – English and Japanese – English.

These activities provide important forums for system developers to meet, exchange ideas and experiences and compare results. For the reports on the most recent research in the MLIA area, the reader is advised to refer to the latest proceedings of these initiatives [15],[16],[17].

8. Conclusions

We have given a very rapid overview of some of the issues that must be considered when building a system that provides access and retrieval functionality for document collections in multiple languages. Much progress has been made in this sector in recent years and substantial momentum has been built. Current efforts are focused in such areas as combining multiple sources of translational evidence to improve cross-language matching of queries and documents, multilingual access to so-called 'low-density' languages – those for which linguistic resources are not readily available in electronic form, multilingual access to multimedia content – particularly spoken documents, and presentation of results from multilingual searches – including summarization of content across multiple documents in different languages.

However, it is noticeable that although MLIA and CLIR research has made significant advances over the last few years, most real-world applications that handle documents in multiple languages still provide only very simple access tools, usually not going beyond a controlled vocabulary search on selected fields, and seldom for more than two languages. A recent survey at a meeting of European Digital Library projects sponsored by the European Commission confirmed that although most projects actually handled documents in several languages, very few of them had already implemented any tools to enable search over more than one language collection at a time. It is evident that a considerable effort is now needed to transfer the results achieved by the research world to the application community. We hope that our tutorial will be one step in this direction.

References

1. Peters, C., Sheridan, P.: Multilingual Information Access. In M. Agosti, F. Crestani, G. Pasi (eds.). Lectures on Information Retrieval, Lecture Notes in Computer Science 1980, pp51-80, Springer Verlag, 2001.
2. Ziegler, D.: The Automatic Identification of Languages Using Linguistic Recognition Signals. PhD Thesis, State University of New York, Buffalo, 1991
3. Damashek, M.: Guaging Similarity with N-grams: Language-independent Categorization of Text. Science, Vol. 267 (No. 10) 1995
4. Souter, C., Churcher, G., Hayes, J., Johnson, S.: Natural Language Identification using Corpus-based Models. Hermes Journal of Linguistics, Vol. 13, pp. 183-203, Faculty of Modern Languages, Aarhus School of Business, Denmark, 1994
5. Wechsler, M., Sheridan, P., Schäuble, P.: Multi-Language Text Indexing for Internet Retrieval. In Proceedings of the 5th RIAO Conference, Computer-Assisted Information Searching on the Internet, Montreal, Canada, June 1997.
6. Porter, M.F.: An Algorithm for Suffix Stripping. Program, Volume 14 (No. 3), pp 130-137, 1980.
7. Ballestreros, L., Croft, W.B.: Resolving Ambiguity for Cross-language Retrieval. In Proceedings of the 20th International ACM SIGIR Conference on Research and Development in Information Retrieval, Philadelphia, PA, pp 84–91, 1997.
8. Soergel, D.: Multilingual Thesauri in Cross-Language Text and Speech Retrieval. In Working Notes of AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, Stanford, CA, pp 164–170, 1997.
9. Hull, D.A., Grefenstette, G.: Querying Across Languages. A Dictionary-based Approach to Multilingual Information Retrieval. In Proc. of 19th ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland, pp 49–57, 1996.
10. Ballesteros, L., Croft, W.B.: Dictionary-based methods for cross-lingual information retrieval. In Proceedings of the 7th International DEXA Conference on Database and Expert Systems Applications, pp 791–801, 1996

11. Ballesteros, L., Croft, W.B.: Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval. In Working Notes of AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, CA, pp 1–8, 1997
12. Adriani, M., van Rijsbergen, C.J.: Term Similarity-Based Query Expansion for Cross-Language Information Retrieval. In Lecture Notes in Computer Science, 1696, 1999.
13. Sheridan, P., Ballerini, J.P.: Experiments in Multilingual Information Retrieval using the SPIDER System, In Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland, pp 58-65, 1996.
14. Sheridan, P., Braschler, M., Schäuble, P.: Cross-Language Information Retrieval in a Multilingual Legal Domain. In ECDL'97 Proceedings, Pisa, Italy, pp 253–268, 1997
15. Voorhees, E.M., Harman, D.K. (eds.). The Eighth Text Retrieval Conference (TREC-8), US National Institute of Standards and Technology, 2000
16. Peters C. (ed.). Cross-Language Information Retrieval and Evaluation: Proc. of the CLEF 2000 Workshop. Lecture Notes in Computer Science, 2069, Springer Verlag, 2001
17. Kando, N., Aihara, K., Eguchi, K, Kato, H. Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization, ISBN 4-924600-89-X, 2001.