CLIR Research at the University of Tampere: issue Editorial

Ari Pirkola
Department of Information Studies
University of Tampere
Finland

Cross-language information retrieval (CLIR) refers to an information retrieval task where the language of queries is other than that of the retrieved documents. A user may present a query in his/her native language and in response the system retrieves documents in another language. Query translation can be done using thesauri, corpora, dictionaries or machine translation systems. An overview of different approaches to CLIR is in Oard and Diekema (1998).

This special issue of *Information Research* presents recent CLIR research done at the *University of Tampere* (UTA). We have adopted the dictionary-based approach. In *dictionary-based cross-language retrieval*, queries are translated by means of electronic dictionaries by replacing source language query keys with their target language equivalents. The main problems associated with dictionary-based CLIR are (1) untranslatable query keys due to the limitations of general dictionaries, (2) the processing of inflected words, (3) phrase identification and translation, and (4) lexical ambiguity in source and target languages. The category of untranslatable keys involves compound words, special terms, and cross-lingual spelling variants, i.e., equivalent words in different languages which differ slightly in spelling, particularly proper names and technical terms.

Earlier CLIR research done at UTA is summarized in Pirkola *et al.*. (2001). We have devised a *query structuring* technique. Many studies have shown that it is an effective disambiguation method, in particular in the case of long queries. We have tested different translation methods using *dictionary combinations* of general and domain specific dictionaries. Earlier CLIR research at UTA also involves the development of *morphological typology* of languages for IR. A set of computable variables characterizing the morphological features of natural languages were developed to be used in CLIR system development and evaluation.

The articles presented in this special issue deal with n-gram based matching of cross-lingual spelling variants and monolingual morphological variants, compound word processing in CLIR, and target language query disambiguation based on word frequency statistics of a document collection.

N-gram matching for untranslatable keys

General dictionaries only include the most commonly used proper names and technical terms. Most of them are untranslatable. A common method to handle untranslatable words in dictionary-based CLIR is to pass them as such to the final target language query. In many languages proper names and technical terms are spelling variants of each other. This allows the use of n-gram matching to find the target language correspondents of the source language query keys. In n-gram matching, query keys and index terms are split into the substrings of length n. The n-gram sets of keys and index terms are compared, and the best matching words are then used as query keys.

In their paper, Pirkola, Keskustalo, Leppänen, Känsälä, and Järvelin examine the effectiveness of digrams combined both of adjacent and *non-adjacent* characters of words in cross- and monolingual word form matching. They present a novel n-gram matching technique (called the *targeted s-gram matching* technique), in which n-grams are classified into categories on the basis of the number of skipped characters. The n-grams belonging to the same category are compared with one another, but not to n-grams belonging to different categories. The effectiveness of the technique is examined empirically. The results show that for Finnish-English, German-English, and Swedish-English cross-lingual spelling variants the technique outperforms the conventional n-gram matching technique using adjacent characters as n-grams.

Compound words in CLIR

Natural languages are productive systems that unlimitedly generate new compounds words. Some languages, such as Finnish, German, and Swedish are characterized by high percentage of compounds. Typically translation dictionaries of such languages include the lexicalised compounds. Most compounds are untranslatable, however. It is thus obvious that effective dictionary look-up and the searching of compound words in CLIR cannot solely be based on full compounds but also on their component parts. For good retrieval performance, compound splitting with a morphological analyzer and separate translation of component words are necessary. To promote phrase recognization, the component translations are recombined with a proximity operator in the target language query.

In her paper, Hedlund considers linguistic features of Finnish, German, and Swedish compound words, as well as CLIR compound processing, i.e., the decomposition of full compounds into their component parts, the normalization and translation of the component parts, and the recombination of component translations with proximity operators. In the paper, the effects of compound processing on CLIR performance are tested.

The use of average term frequency in CLIR

In their paper, Pirkola, Leppänen, and Järvelin examine the utilization of average term frequency in CLIR. Dictionaries typically give many mistranslated words for source language keys. Many of the mistranslated words are general words whose average term frequency is low and document frequency high. The opposite holds for more specific terms which often are important keys in queries. These facts as a starting point, the researchers developed a query key goodness scheme (called the *RATF formula*, where RATF stands for *relative average term frequency*). The use of the RATF formula in CLIR is tested empirically. Query keys are weighted based on their RATF values. The assumption is that in this way good keys are given more weight than bad keys. Also other key weighting methods associated with RATF which particularly are suited for CLIR are tested. The results indicate that RATF-based CLIR queries often perform better than undisambigauated CLIR queries, but the results also indicate the limitations of the use of RATF in CLIR.

Acknowledgements

I would like to thank Editor-in-Chief, Prof. Tom Wilson, for his help in preparing this special issue.

References

- Oard, D. and Diekema, A. (1998) Cross-Language Information Retrieval. *Annual Review of Information Science and Technology (ARIST)*, **33**, 223-256.
- Pirkola, A., Hedlund, T., Keskustalo, H., and Järvelin, K. (2001) Dictionary-based cross-language information retrieval: problems, methods, and research findings. *Information Retrieval*, **4**(3/4), 209-230.

How to cite this paper:

Pirkola, A. (2002) "CLIR Research at the University of Tampere: issue Editorial"] *Information Research*, **7**(1) [Available at: http://InformationR.net/ir/7-2/CLIR.html]

Last updated: 20 January 2002