Dictionary-Based Cross-Language Information Retrieval: Principles, System Design and Evaluation

Turid Hedlund
Department of Information Studies
Faculty of Information Sciences
University of Tampere
hedlund@shh.fi

The research problems of the thesis relate to the Scandinavian language Swedish. When the research work on this thesis started, there was very limited knowledge on information retrieval or cross-language information retrieval research in Swedish. The linguistic features of this and other compound rich languages indicate that research focusing on languages of other types than English is of great importance. One problem was also the lack of automated dictionary-based systems for query translation of Scandinavian languages and other compound rich languages. Firstly, cross-language information retrieval problems for non-English languages, particularly Swedish are discussed. In the article the need to extend research on information retrieval techniques to undertreated languages is demonstrated. Secondly, one of the main problems identified for Swedish, the frequent presence of compounds is discussed in detail and solutions are proposed. Retrieval efficiency may be improved by splitting not directly translatable compounds into constituents using morphological analysis programs and by normalising the constituents into base form before translation using machine-readable dictionaries. This solution is tested for 80 cross-language information retrieval queries.

Thirdly, this thesis deals with bilingual natural language information retrieval techniques where English is the target or document language and Swedish, Finnish and German are source or query languages. The system design of the UTACLIR, an extendable bilingual dictionary-based query translation system, is presented. The approach is to apply linguistic tools in an automated dictionary-based system able to handle several languages.

Fourthly, the performance of the system is evaluated in international evaluation campaigns and shown effective. The automated CLIR process is also tested for the performance of its components. The tests with structuring of the queries indicate that structuring is a good way to reduce the effect of ambiguity caused by several dictionary translation equivalents for a source language word. This is true for all the source languages, but is particularly notable for Finnish and German where the translation dictionaries used in the study were comprehensive. Compound handling for the compound rich source languages Swedish, German and Finnish is found beneficial to the system performance. An n-gram based algorithm was implemented in the process in order to solve the problem of untranslatable words, such as proper names. The process was particularly successful for the Finnish language where proper names usually appear in inflected forms and where matching to the target language document index therefore is difficult.

Electronic version of thesis: *http://acta.uta.fi/pdf/951-44-5790-0.pdf*