A CLIR Interface to a Web Search Engine

Philipp Daumke   Stefan Schulz   Kornél Markó
Freiburg University Hospital, Department of Medical Informatics (http://www.imbi.uni-freiburg.de/medinf)

Abstract
Medical document retrieval presents a unique combination of challenges for the design and implementation of retrieval engines. We introduce a method to meet these challenges by implementing a multilingual retrieval interface for biomedical content in the World Wide Web. To this end we developed an automated method for interlingual query construction by which a standard Web search engine is enabled to process non-English queries from the biomedical domain in order to retrieve English documents.

Morpho-Semantic Indexing
To cope with various morphological processes in different languages we developed a term normalization methodology, called Morpho-Semantic Indexing (henceforth, MSI). MSI uses a special type of dictionary with entries consisting of subwords, i.e. semantically minimal units. Subwords are grouped into equivalence classes (represented by Morpheme identifiers (MIDs)) which capture intralingual as well as interlingual synonymy. A morphosyntactic parser extracts subwords from texts and assigns MIDs in a three step MSI procedure. The result is a morpho-semantically normalized expression in a language independent representation. Its usefulness for cross-language indexing and retrieval has already been proven.

Retrieval interface
Using domain and language specific corpora (cf. Figure 1A) we created a target list (B) of data including surface words, word bigrams and trigrams (data) as well as their frequencies freq(data) within these corpora. All target queries are subsequently translated to a set of MIDs (C). This data is encoded in a table FreqTab (D), each record being a quadruple (data, freq(data), MSI(data), language(data)). A user can choose his favoured target language on a web interface. These user queries (E) are sent to our query construction tool. Again, these queries are initially transformed to a set of corresponding MIDs (F). Using several heuristics we further process these MIDs to a list of possible MID combinations called partitions which consist of subqueries (G). All subqueries are now compared to MSI(data) in FreqTab at which all matching records are returned (H, I). Out of these records we generate possible target queries by a set of combination rules, together with a score which can be used as a measure for its lexical relevance (J). Finally, these target queries are sent separately to a standard Web search engine and merged by a simple frequency heuristics.

Conclusion, Related Work
First test runs on our retrieval interface show promising performance. Future work will focus on problems in the query formulation process such as matching problems due to a lack of coherence of our subword lexicon and language specific problems[1].

References