Issues When Building CLIR Applications

Outline of Invited Talk at CLEF 2004 Workshop

G. Thurmair, linguatec

The talk will discuss some design issues which influence the quality of crosslingual retrieval.

1. On the **Analysis** Side, index creation is a key point. The index should be based on base forms, not on text forms or stems; therefore state-of-the-art stemmers are not the best tools to use. Also, there should be one index per language, to avoid clashes of terms occurring in several languages; this implies to run a language-detection component before indexing.
2. On the **Query** side, the quality of search crucially depends on the resource used for translation. While several techniques have been proposed, preference should be given to a multilingual conceptual network, to be used for a term-to-term translation. The closer this network matches the domain to be searched the better the search quality will be.
3. Once the search **result** is available, the questions of ranking of multilingual documents, as well as retranslation into the query language need to be addressed. Technology components to support these tasks must be provided.
4. The CLIR application requires a new kind of linguistic **resource** which merges requirements of terminology, formal NLP type dictionaries, conceptual hierarchies, and multilinguality like in Machine Translation. Such a multifunctional resource must be designed and maintained, and the paper ends with a list of requirements for such a resource.

As a result, a good CLIR system shows significant differences as compared to a conventional IR implementation.