CLEF 2001 Experiments Using KCSL's Retrieval System

Ilia Kaufman and Meena Ghanekar
KCSL Inc., 5160 Yonge Street, Suite 1012
Toronto, ON, Canada M2N 6L9

Abstract

We entered CLEF 2001 Evaluation Forum with our UniFind retrieval system that was developed at KCSL to satisfy corporate information retrieval needs. We participated in the following three tasks:

- Multilingual Information Retrieval (EN=>FR, DE, ES, IT, EN)
- Bilingual Information Retrieval (FR=>EN)
- Monolingual Information Retrieval (Spanish collection)

This is the first time we entered TREC/CLEF experiments and we used UniFind essentially without any modifications, except for the ranking of documents in the Multilingual tasks and for determining the relevancy cutoff points.

For the Multilingual tasks we first performed separate runs for each of the five languages, namely, English, French, Italian, German and Spanish. The results of these five runs were then merged and re-ranked based on the similarity values obtained in the individual runs. UniFind's quantitative process of similarity ranking and of extracting relevant documents is identical for all languages. Therefore, since the number of documents processed in each of the five individual runs was substantial (the smallest was French with more than 87,000 documents), at the merging and re-ranking stage, we used the original similarity values obtained separately for each language.

With respect to the relevancy cutoff points, our system selects them automatically and usually succeeds in eliminating irrelevant and marginally relevant documents. Since CLEF expects, by default, 1000 documents in each set, we decided to relax our cutoff strategy in order to return more documents from each run. Unfortunately, we didn't do it quite right, for most of our runs still returned well under 50 documents. This clearly contributed to lower Recall scores and we believe that in turn it resulted in lower overall scores for UniFind.

For query translation we used commercial MT (Machine Translation) software from Lernout and Hauspie.

Our system analyzes the query and all documents in the corpus to determine word usage, word morphology, sentence boundaries and a very detailed topological structure that accounts for the distribution of query words and sentences containing these words and their derivatives.

In addition, a sentence analysis is performed to determine both a position independent and a position dependent score for each sentence in a document. This step not only helps to improve the accuracy, relevancy, and quality of the results but also determines the most relevant part of a document that best relates to the query. Thus, our algorithms comprise Topological, Statistical and Linguistic analyses of queries and documents. The algorithms tend to relate to concepts as they are expressed in sentences, as well as, how they relate to a document as a whole. This process is conceptually identical for all sentence-based languages.

The test data as supplied by CLEF 2001, contained 749,877 documents, in five languages (we didn't process the Dutch corpus) and occupied approximately 2GB of disk space.

We submitted three sets of runs for each of the three tasks by automatically constructing the queries from the 50 topics in the selected language. Our three runs in each task were: title field only, title and description fields, and title with description and narrative fields.

We observed that in all three tasks the sets with queries consisting of title and description fields seem to have the best performance.

All of our runs were executed on Windows 2000 platform running on Pentium III CPU at 800 MHz with 1GB of RAM and 40 GB disk.