Mpro-IR in Clef 2001

Bärbel Ripplinger

IAI
Martin-Luther-Str. 14
66111 Saarbrücken, Germany
babs@iai.uni-sb.de *

Abstract

The objective of this year's CLEF participation was to evaluate an improved German component, focusing on the impact decomposition information has on performance.

1 Introduction

The MPRO-IR system is a CLIR system based on query translation and focuses rather on a better recall than on a balanced recall and precision figure. To improve the recall, the system tries to take advantage of a sophisticated linguistic processing component whose results are used in the monolingual retrieval modules. Based on the output of a morpho-syntactic analysis which provides the full range of morphological information, not only inflection which would correspond to the power of a Porter-like stemmer but also derviation and decomposition of compound nouns are exploited. This information is used for indexing, query expansion, search and document ranking.

The objective of this year's CLEF participation was to evaluate an improved German component, therefore only one official monoligual German run had been submitted. The investigations focused on how an improved linguistic processing affects the performance compared to last year's result. A new morpho-syntactic analysis for German has been applied which uses a far better tagging and lemmatisation component based on a morpheme lexicon with 85.500 entries (morphemes, stems as well as word forms) compared to 42.000 entries used for last years experiments.

2 Experiment Settings

As found out last year, the number of documents retrieved by MPRO-IR had been low compared to other systems. Because this was mainly due to the restriction to one sentence as search window, we did not apply this limitation in this year's experiment. Another reason was that we were obliged to submit a run using title and description as query input (finding all meaning bearing words of the description in one sentence would make no sense).

Even the underlying architecture of MPRO-IR requires that each word has to occur in a document to be relevant, no query preprocessing was done, i.e. fixed phrases such as 'find documents about', 'find reports on', etc. were not deleted. However, because these phrases hardly occur in a document, we took account by weaken the requirement above, i.e. not every queried term has to occur in a document to be relevant. In consequence, the calculation of the rank has been changed from last year by calculating the weight not only on basis of the linguistic information used to retrieve a particular document but considering additionally the number of queried terms found within this document. The query was morpho-syntactically analysed, using the information extracted for meaning bearing words (those having as part-of-speech noun, verb, or adjective) such as lexical base form, derivational root, and decomposition to search for in German document.

*Working now for Eurospider Information Technology AG, Email address: ripplinger@eurospider.ch.

3 Results

The overall result of our run shows a lower retrieval performance of MPRO-IR compared to the other systems. In spite of a higher number of documents retrieved, the result is even worser than last year. One reason is certainly that corrupted lexcions were used for document and query analysis (unfortunately there was no time to redo the corpus analysis). Insofar, the results have no significance to our aim expecting that an improved linguistic analysis positively affects the performance. Furthermore, there is reason to suppose that the worse results are due to the decision not to preprocess the queries, and instead changing the search algorithm plus the ranking, and thus undermine MPRO-IR's philosophy.

The investigation of the results per query in more detail shows more or less the same findings as last year: Most hits could be retrieved by using precise lexcial base forms, and derivational information. Compositional information was also valuable to detect syntactic variants of German compounds. However, because the lexicon has now 50% more entries, the number of wrong compound analyses has increased which is mainly due to the current state of the morpheme lexicon. Not all entries are examined in respect to allowed and forbidden compounding, information which has to be explicitly encoded.

Acknowledgements

I'm indebted to Peter Schäuble for allowing me to use EUROSPIDER resources to carry out this experiment.

References

[1] Maas, D. *Multilinguale Textverarbeitung mit MPRO.* In G. Lobin et al.(eds): **Europäische Kommunikationskybernetik heute und morgen,** KoPäd, München, 1999. http://www.iai.uni-sb.de/global/memos.html

[2] Ripplinger, B. MPRO-IR– *A Cross-language Information Retrieval Component Enhanced by Linguistic Knowledge.* In *Proceedings of* RIAO 2000, Paris, 2000.