Mono- and Crosslingual Retrieval Experiments at the University of Hildesheim

René Hackl, Thomas Mandl, Christa Womser-Hacker

University of Hildesheim, Information Science, Marienburger Platz 22
D-31141 Hildesheim, Germany
mandl@uni-hildesheim.de

Abstract.
In this year's participation we continued to evaluate open source information retrieval software. We used mainly the system Lucene and experimented with some of the most effective optimization strategies applied in CLEF. The effectiveness of open source and other free tools can be enhanced by these optimization strategies. For most languages, blind relevance feedback leads to considerable improvement. Indexing strategies with n-grams have not led to improvements within Lucene.

1 Introduction

In the CLEF 2004 campaign, we tested an adaptive fusion system based on the MIMOR model with several mono- and multi-lingual tasks. As a basic retrieval system we employed the open source system Lucene.
Our main goal is to measure the quality of open source product in comparison to the best systems at CLEF. We exploit some of the most promising optimization techniques applied at CLEF in order to observe the potential for improvement of standard IR systems like Lucene. This work contributes to the practical application of the results from CLEF.
Lucene has proved to be very efficient in CLEF as well as in other projects (e.g. cf. Hackl, Mandl & Schwantner 2004) and is becoming increasingly popular. We expect Lucene to be employed in many more contexts. Therefore, we intend to continue testing its effectiveness within the CLEF campaign.
Our basic fusion approach MIMOR is described in more detail in Womser-Hacker 1997 and Mandl & Womser-Hacker 2004a. MIMOR has already been applied to CLEF experiments (Hackl, Kölle et al. 2004).

2 Cross Language Retrieval Experiments

The tools we employed this year include Lucene 1.4-final[1] and Java™-based snowball[2] language analyzers. Last year we had also evaluated the MySQL's full text indexing and search module, but due to their poor performance they were excluded this year. For this year's participation we focussed on different indexing methods such as different stemmers and n-gram-techniques.
We took part in the monolingual tracks for Russian and Finnish, the bilingual track English to Russian and the multilingual track.
Firstly, we ran some preliminary monolingual experiments on the collections from 2003 without query expansion (Table 1). Note that we did not index the LA Times 1994, as well as the 1994 volumes of the French collections as they were not needed for this year.
Secondly, we had planned to try out every combination of indexing methods for a language, to find out whether there is some fusion potential. This can be seen in the table for Finnish ("Finnish all"), where all result lists from the different indices were merged into a single one.
Russian character handling in Java led to problems which caused a very low performance.
For evaluation of the test runs we used our beta stage Java clone of the official trec_eval program.

[1] Lucene: http://jakarta.apache.org/lucene/docs/index.html
[2] Snowball: http://jakarta.apache.org/lucene/docs/lucene-sandbox/snowball/

Table 1. Test runs with data from 2003, English and French collections from 1995 only

Language	Indexing	Recall	Average Precision
English	4-gram	516 / 1006	0.1256
English	5-gram	516 / 1006	0.1083
English	6-gram	507 / 1006	0.1034
English	snowball stemmer	497 / 1006	0.1608
English	lucene stemmer	499 / 1006	0.1690
Finnish	4-gram	391 / 483	0.2237
Finnish	5-gram	403 / 483	0.2261
Finnish	6-gram	391 / 483	0.2036
Finnish	snowball stemmer	450 / 483	0.4853
Finnish	lucene stemmer	N/A	N/A
Finnish	Fusion of all	452 / 483	0.3218
French	4-gram	548 / 946	0.1242
French	5-gram	549 / 946	0.1077
French	6-gram	560 / 946	0.1050
French	snowball stemmer	563 / 946	0.1498
French	lucene stemmer	525 / 946	0.1504
Russian	4-gram	98 / 151	0.0652
Russian	5-gram	98 / 151	0.0620
Russian	6-gram	96 / 151	0.0642
Russian	snowball stemmer	71 / 151	0.0810
Russian	lucene stemmer	88 / 151	0.1336

For the submitted runs we used the title and descriptor topic fields, which were also mandatory. We applied pseudo-relevance feedback for all tasks. For runs involving Russian, we also created one run without BRF. To translate the queries we used the internet service freetranslation.com[3] which provided some surprisingly good translations from English to Russian. The results can be seen in Table 2.

Table 2. Results for runs in CLEF 2004

Runs	Optimization	Recall	Average Precision
UHImlt1	BRF 5 10 RSV	1031 / 1826	0.1974
UHImlt2	BRF 5 10 KL	973 / 1826	0.1849
UHIenru1	BRF 5 10 RSV	77 / 123	0.1353
UHIenru2	BRF 5 10 KL	73 / 123	0.1274
UHIenru3	no BRF	53 / 123	0.0484
UHIru1	BRF 5 10 RSV	88 / 123	0.1553
UHIru2	BRF 5 10 KL	82 / 123	0.1420
UHIru3	no BRF	56 / 123	0.0459
UHIfi1	BRF 5 10 KL	349 / 413	0.4699
UHIfi2	BRF 5 10 RSV	367 / 413	0.5042

For Finnish, the performance is quite high. The snowball stemmer works very well.
For Russian, our results are very bad due to some encoding problems. BRF still worked well for Russian under these circumstances. Also, the test runs had indicated that the Lucene stemmer seems very capable of dealing with Russian and it held up to that expectation.
The multilingual runs suffered severely from the obstacles that led to the bad results for Russian. We do also have only limited insight into the usefulness of intertran.com as a translation tool for Finnish.

[3] http://www.freetranslation.com

3 Conclusion

This year's Russian tracks posed some challenges we could not easily overcome. Despite working with Java only and unicode-based character sets, the Russian stopwords could not be eliminated. We did not have the resources to work on a more sophisticated approach.

4 Outlook

Our system is far from well adapted the task. It has about 30 weighting parameters. This year, we could only experiment with a few ones.
Furthermore, in future years we intend to exploit the observed relation between the number of named entities in topics and retrieval performance (cf. Mandl & Womser-Hacker 2004b).

Acknowledgements

We would like to thank the Jakarta and Apache projects' teams for sharing Lucene with a wide community as well as the providers of Snowball. Furthermore, we acknowledge the work of several students from the University of Hildesheim who implemented MIMOR as part of their course work.

References

Hackl, René (2004): Multilinguales Information Retrieval im Rahmen von CLEF 2003. Master Thesis, University of Hildesheim, Information Science.

Hackl, René; Kölle, Ralph; Mandl, Thomas; Ploedt, Alexandra; Scheufen, Jan-Hendrik; Womser-Hacker, Christa (2004): Multilingual Retrieval Experiments with MIMOR at the University of Hildesheim. To appear in: Peters, Carol; Braschler, Martin; Gonzalo, Julio; Kluck, Michael (eds.): Evaluation of Cross-Language Information Retrieval Systems. Proceedings of the CLEF 2003 Workshop. Berlin et al.: Springer [Lecture Notes in Computer Science]

Hackl, René; Mandl, Thomas; Schwantner, Michael (2004): Evaluierung und Einsatz des open source Volltext-Retrievalsystems Lucene am FIZ Karlsruhe. In: Ockenfeld, Marlies (ed.): Information Professional 2011: Strategien – Allianzen – Netzwerke. Proceedings 26. DGI Online-Tagung. Frankfurt a.M. 15.-17. Juni. pp. 147-153.

Mandl, Thomas; Womser-Hacker, Christa (2004a): A Framework for long-term Learning of Topical User Preferences in Information Retrieval. In: New Library World. vol. 105 (5/6). pp. 184-195.

Mandl, Thomas; Womser-Hacker, Christa (2004b): Analysis of Topic Features in Cross-Language Information Retrieval Evaluation. In: 4th International Conference on Language Resources and Evaluation (LREC) Lisbon, Portugal, May 24-30. Workshop Lessons Learned from Evaluation: Towards Transparency and Integration in Cross-Lingual Information Retrieval (LECLIQ). pp. 17-19.

Womser-Hacker, C. (1997): Das MIMOR-Modell. Mehrfachindexierung zur dynamischen Methoden-Objekt-Relationierung im Information Retrieval. Habilitationsschrift. Universität Regensburg, Informations-wissenschaft.