

Mono- and Bilingual Retrieval Experiments with a Social Science Document Corpus

René Hackl, Thomas Mandl

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

Abstract

This paper reports on our participation in CLEF 2005's domain-specific retrieval track. The experiments were based on previous experiences with the GIRT document corpus and were run in parallel to the multi-lingual experiments for CLEF 2005. We optimized the parameters of the system with one corpus from 2004 and applied these settings to the domain specific task. In that manner, the robustness of our approach over different document collection was assessed.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software

General Terms

Measurement, Performance, Experimentation

Keywords

Domain specific, Social Science, Bilingual retrieval, Thesaurus

1 Introduction

In previous CLEF campaigns, we tested an adaptive fusion system based on the MIMOR model (Mandl & Womser-Hacker 2004) within the domain specific GIRT track (Hackl et al. 2003). For CLEF 2005, the parameter optimization was based on a French document collection. The parameter settings were applied to the four language document collection of the multilingual task of CLEF 2005 (Hackl et al. 2005).

In addition, we applied almost the same settings to the domain specific track in order to test the robustness of our system over different collections.

Robustness has become an issue in information retrieval research recently. It has been noted often, that the variance between queries is worse than the variance between systems. There are often very difficult queries which few systems solve well and which lead to very bad results for most systems (Harman & Buckley 2004). Thorough failure analysis can lead to substantial improvement. For example, the absence of named entities are a factor which can make queries more difficult overall (Mandl & Womser-Hacker 2004). As a consequence, a new evaluation track for robust retrieval has been established at the Text Retrieval Conference (TREC). This track does not only measure the average precision over all queries but also emphasizes the performance of the systems for difficult queries. To perform well in this track is more important for the systems to retrieve at least a few documents for difficult queries than to improve the performance in average (Voorhees 2005). In order to allow a system evaluation based on robustness more queries than for a normal ad-hoc track are necessary. The concept of robustness is extended in TREC 2005. Systems need to perform well over different tracks and tasks (Voorhees 2005).

For multilingual retrieval, robustness would also be an interesting evaluation concept because the performance between queries differs greatly (Mandl & Womser-Hacker 2004). Robustness in multilingual retrieval could be interpreted in three ways:

- Stable performance over all topics instead of high average performance (like at TREC)
- Stable performance over different tasks (like at TREC)
- Stable performance over different languages (focus of CLEF)

For the participation in the domain specific track in 2005, we tested the stability of our ad-hoc system for the domain specific track.

2 Domain Specific Mono- and Cross-lingual Retrieval Experiments

Our system was optimized with the French collection of CLEF 2004. The optimization procedure is described in detail in Hackl et al. 2005. The GIRT runs were produced with only slightly different settings.

Previous experiences with the GIRT corpus showed that blind relevance feedback does not lead to good results (Kluck 2004). Our test runs confirmed that fact and blind relevance feedback was not applied for the submitted runs. Instead, term expansion was based the thesaurus available for the GIRT data. This thesaurus was developed by the Social Science Information Centre (Kluck 2004). For the query terms, the fields Broader, Narrower and Related term were extracted from the thesaurus and added to the query for the second run. The topic title weights were set to ten, topic description weights to three and the thesaurus terms were weighted with one. This weighting scheme was adopted from the ad-hoc task.

For the second mono-lingual run UHIGIRT2, we added terms from the multilingual European terminology database Eurodicautom¹ which was also used for the ad-hoc experiments. However, Eurodicautom contributed terms for very few queries. Most often, it returned "out of vocabulary".

As bilingual GIRT run, we submitted one English-to-German run. The query and the thesaurus terms were translated by ImTranslator². In addition, the document field "english-translation" was indexed.

Table 1. Results from the CLEF 2005 Workshop. EDA = Euradicautom

RunID	Languages	Run Type	Fields used	Retrieved	Relevant docs.	Avg. Prec.
UHIGIRT1	Monolingual German	Lucene stemmer	TD	1400	2682	0.220
UHIGIRT2	Monolingual German	Lucene stemmer, IZ thesaurus, EDA	TD	1335	2682	0.193
UHIGIRT3	English-German	Lucene stemmer, IZ thesaurus, EDA ImTranslator	TD	1159	2682	0.178

Although, our system has been tested with Russian data at earlier CLEF campaigns and at the ad-hoc task this year, the Russian social science RSSC collection could not be used because it was provided later than the rest of the data.

3 Conclusion and Outlook

For next year, we intend to implement for multi-lingual runs for the domain specific task. The thesaurus use led to a drop in performance. For the future, we intend to develop a more sophisticated strategy to apply thesaurus terms.

References

Hackl, René; Kölle, Ralph; Mandl, Thomas; Womser-Hacker, Christa (2003): Domain Specific Retrieval Experiments at the University of Hildesheim with the MIMOR System. In: Peters, Carol; Braschler, Martin; Gonzalo, Julio; Kluck, Michael (Eds.): Advances in Cross-Language Information Retrieval: Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002, Rome, Italy, September 2002. Berlin et al.: Springer [LNCS 2785] pp. 343-348.

Hackl, René; Mandl, Thomas; Womser-Hacker, Christa (2005): Ad-hoc Mono- and Multilingual Retrieval Experiments at the University of Hildesheim. In this volume

¹ <http://europa.eu.int/eurodicautom/Controller>

² <http://freetranslation.paralink.com/>

- Harman, Donna; Buckley, Chris (2004): The NRRC reliable information access (RIA) workshop. In: Proceedings of the 27th annual international conference on Research and development in information retrieval (SIGIR). pp. 528-529.
- Kluck, Michael (2004): The GIRT Data in the Evaluation of CLIR Systems - from 1997 until 2003. In: Comparative Evaluation of Multilingual Information Access Systems: 4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003, Trondheim, Norway, August 21-22, 2003, Revised Selected Papers. pp. 376-390
- Mandl, Thomas; Womser-Hacker, Christa (2004): 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 (2005): The Effect of Named Entities on Effectiveness in Cross-Language Information Retrieval Evaluation. In: Proc ACM SAC Symposium on Applied Computing (SAC). Information Access and Retrieval (IAR) Track. Santa Fe, New Mexico, USA. March 13.-17. pp. 1059-1064.
- Voorhees, Ellen (2005): The TREC robust retrieval track. In: ACM SIGIR Forum 39 (1) pp. 11-20.