# Chemnitz at CLEF 2009 Ad-Hoc TEL Task: Combining Different Retrieval Models and Addressing the Multilinguality

Jens Kürsten
Chemnitz University of Technology
Faculty of Computer Science, Dept. Computer Science and Media
09107 Chemnitz, Germany
jens.kuersten@cs.tu-chemnitz.de

**Abstract**

In this paper we report our efforts for the participation in the CLEF 2009 Ad-Hoc TEL task. In our second participation we were able to test and evaluate a new feature of the *Xtrieval framework*, which was the accessibility of the three core retrieval engines Lucene, Lemur and Terrier. This year we submitted 24 experiments in total, 12 each for the monolingual and bilingual subtasks. We compared our baseline experiments to combined runs, where we used two different retrieval models, namely the vector space model (VSM) used in Lucene and the Bose-Einstein model for randomness (BB2) available in the Terrier framework. We found that an almost constant improvement in terms of mean average precision over all provided collections is achievable. Furthermore we tried to benefit from the multilingual contents of the collections by running combined multilingual experiments for both subtasks. The evaluation showed that the used approach achieves small improvements in the monolingual setting of the task. Unfortunately, we were not able to confirm this finding in the bilingual setting, where the multilingual experiments were outperformed by the standard bilingual runs, especially on the English target collection.

## Categories and Subject Descriptors

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

## Keywords

Evaluation, Experimentation, Cross-Language Information Retrieval

## 1 Introduction and outline

The *Xtrieval* framework [3],[2] was used to prepare and run this year's retrieval experiments in the *Ad-Hoc track TEL* setting. The core retrieval functionality was provided by *Lucene*[1] and the *Terrier framework* [4]. For the *TEL task* three different multilingual corpora with content mainly in German, English and French were provided by *The European Library*. Each collection consists of approximately one million library records. These library records only contain sparse information and have descriptions in multiple languages.

[1] http://lucene.apache.org

We conducted monolingual experiments on each of the collections and also submitted experiments for the bilingual task. For the translation of the topics the Google AJAX language API[2] was accessed through a JSON[3] programming interface.

The remainder of the paper is organized as follows. Section 2 describes the general setup of our system. The individual configurations and the results of our submitted experiments are presented in section 3. In sections 4 and 5 we summarize the results and conclude our observations.

## 2 Experimental setup

This year we were able to choose from various retrieval models and combine the results in the retrieval stage by applying our implementation of the *Z-Score* operator [5]. We also used a standard top-k pseudo-relevance feedback algorithm in the retrieval stage, where the values for the top most frequent terms that were obtained from the top documents differed according to the language and used retrieval model. We used the vector space model (VSM) shipped with Lucene and the Bose-Einstein model for randomness (BB2) available in the Terrier framework. We submitted two monolingual baseline runs for all provided collections. Additionally we submitted one monolingual merged experiment and another one in which we tried to benefit from the multilingual character of the collections. The merged monolingual experiments for each collection formed the baseline for two bilingual experiments, where the topics were translated from two different source languages to the corresponding target collection. For two additional bilingual experiments on each target collection we also tried to access the multilingual content of the collections.

We submitted 9 experiments in which we tried to benefit from the multilingual character of the collections. Therefore we created multiple indexes for each target collection using appropriate stemming and stopword removal for the four most frequent languages. During the retrieval we queried these four indexes and combined the results into one final result list. We needed to translate the topics for all those experiments to the according language of the index, which makes those experiments somewhat multilingual. In table 1 we denote the experiments that had multilingual character and present the boost values for the combination in the multilingual result set for each of the experiments in column 'IDs'. These values were chosen according to the occurrence frequency of the language in the corresponding target collection. All runs in the column 'IDs' correspond to an experiment in column 'refer ID' and are directly comparable to this experiment, because we used identical system configurations except for the translation component and the multilingual indexes.

Table 1: Settings for Multilingual Experiments

| *IDs* | *refer ID* | *Collection* | *Topic Translation* | *Boosting* |
|---|---|---|---|---|
| cut4++, cut7++, cut8++ | cut1 | TELONB | X2DE | 0.568536 |
| | | TELONB | X2EN | 0.098114 |
| | | TELONB | X2FR | 0.023721 |
| | | TELONB | X2IT | 0.014716 |
| cut12++, cut15++, cut16++ | cut9 | TELBL | X2EN | 0.595822 |
| | | TELBL | X2FR | 0.088845 |
| | | TELBL | X2DE | 0.052568 |
| | | TELBL | X2ES | 0.028257 |
| cut20++, cut23++, cut24++ | cut17 | TELBNF | X2FR | 0.567183 |
| | | TELBNF | X2EN | 0.118061 |
| | | TELBNF | X2DE | 0.038291 |
| | | TELBNF | X2IT | 0.020950 |

[2]http://code.google.com/apis/ajaxlanguage/documentation
[3]http://json.org

# 3 Configurations and Results

The detailed setup of our experiments and their evaluation results are presented in the following subsections.

## 3.1 Monolingual Experiments

We submitted 12 monolingual experiments in total, whereof 4 were submitted for each target collection in German, English and French. For all experiments a language-specific stopword list was applied[4]. We used the stemmers from Snowball[5] for English and French and applied a special n-gram stemmer[6] for German.

In table 2 the retrieval performance of our experiments is reported in terms of mean average precision (MAP) and the absolute rank of the experiment in the evaluation. We compare the two baseline runs to one combined experiment per target collection. Furthermore we compare the performance of the first baseline run per collection (cut1, cut9, cut17) to the corresponding multilingual experiment (cut4++, cut12++, cut20++).

Table 2: Experimental Results for the Monolingual Task

| *ID* | *IR Core / Model* | *QE* | *Lang* | *MAP* | *Rank* |
|---|---|---|---|---|---|
| cut1 | Terrier / BB2 | 10 / 70 | MONO-DE | 0.2602 | 10/35 |
| cut2 | Lucene / VSM | 10 / 70 | MONO-DE | 0.2641 | 9/35 |
| cut3 | merged cut1 & cut2 | 10 / 70 | MONO-DE | **0.2789** | 2/35 |
| cut4++ | Terrier / BB2 | 10 / 70 | MONO-DE | 0.2713 | 5/35 |
| cut9 | Terrier / BB2 | 3 / 11 | MONO-EN | 0.3864 | 9/46 |
| cut10 | Lucene / VSM | 10 / 7 | MONO-EN | 0.3672 | 15/46 |
| cut11 | merged cut9 & cut10 | both | MONO-EN | **0.4071** | 2/46 |
| cut12++ | Terrier / BB2 | 3 / 11 | MONO-EN | 0.3914 | 7/46 |
| cut17 | Terrier / BB2 | 0 / 0 | MONO-FR | 0.2470 | 9/35 |
| cut18 | Lucene / VSM | 0 / 0 | MONO-FR | 0.2399 | 12/35 |
| cut19 | merged cut17 & cut18 | 0 / 0 | MONO-FR | **0.2583** | 2/35 |
| cut20++ | Terrier / BB2 | 0 / 0 | MONO-FR | 0.2540 | 3/35 |

The evaluation of our experiments allows to draw some interesting conclusions. First the overall performance in terms of MAP on the German and French collection were quite similar, while the experiments on the English collection achieved much better results. Interestingly this seemed not to be a flaw in our configuration since we achieved identical position in the ranking over all submitted experiments. Another important observation was that our combined experiments (where different retrieval models were used) always performed better than the baseline run on each of the target collections. However the overall gain was not very large. Furthermore one can conclude that our multilingual approach also worked consistently well by slightly improving MAP (compare cut1 to cut4++, cut9 to cut12++ and cut17 to cut20++).

## 3.2 Cross-lingual Experiments

We submitted 12 experiments for the bilingual subtask, whereof 4 were submitted for each target collection. Two experiments per target collection correspond to the combined monolingual run on that collection. Though two different source topic languages were translated in those experiments. The remaining two runs per target collection had again multilingual character. We translated the topic from the source language to the four most common languages in the target collections, queried the four indexes and combined the results in a multilingual result set. Again the general

[4] http://members.unine.ch/jacques.savoy/clef/index.html
[5] http://snowball.tartarus.org
[6] http://www-user.tu-chemnitz.de/w̃ags/cv/clr.pdf

configuration was equal to the corresponding monolingual reference run for comparability. In table 3 we report the evaluation results for each of the bilingual experiments in terms of MAP and the rank over all submitted experiments. Additionally we report our best monolingual experiment for each target collection as baseline for comparison.

Table 3: Experimental Results for the Bilingual Task

| *ID* | *Model* | *QE* | *Lang* | *MAP* | *Rank* |
|---|---|---|---|---|---|
| cut3 | merged cut1 & cut2 | 10 / 70 | MONO-DE | 0.2789 | 2/35 |
| cut5 | Lucene/VSM & Terrier/BB2 | 10 / 70 | BILI-EN2DE | **0.2583** | 1/26 |
| cut6 | Lucene/VSM & Terrier/BB2 | 10 / 70 | BILI-FR2DE | 0.2552 | 3/26 |
| cut7++ | Terrier/BB2 | 10 / 70 | BILI-EN2DE | 0.2580 | 2/26 |
| cut8++ | Terrier/BB2 | 10 / 70 | BILI-FR2DE | 0.2444 | 4/26 |
| cut11 | merged cut9 & cut10 | 10 / 7 & 3 / 11 | MONO-EN | 0.4071 | 2/46 |
| cut13 | Lucene/VSM & Terrier/BB2 | 10 / 7 & 3 / 11 | BILI-DE2EN | **0.4046** | 1/43 |
| cut14 | Lucene/VSM & Terrier/BB2 | 10 / 7 & 3 / 11 | BILI-FR2EN | 0.4029 | 2/43 |
| cut15++ | Terrier/BB2 | 3 / 11 | BILI-DE2EN | 0.3427 | 9/43 |
| cut16++ | Terrier/BB2 | 3 / 11 | BILI-FR2EN | 0.3332 | 11/43 |
| cut19 | merged cut17 & cut18 | 0 / 0 | MONO-FR | 0.2583 | 2/35 |
| cut21 | Lucene/VSM + Terrier/BB2 | 0 / 0 | BILI-EN2FR | 0.2320 | 4/26 |
| cut22 | Lucene/VSM + Terrier/BB2 | 0 / 0 | BILI-DE2FR | 0.2119 | 9/26 |
| cut23++ | Terrier/BB2 | 0 / 0 | BILI-EN2FR | 0.2255 | 5/26 |
| cut24++ | Terrier/BB2 | 0 / 0 | BILI-DE2FR | **0.2557** | 1/26 |

The evaluation results of our bilingual experiments were very strong. The retrieval performance of our best monolingual runs compared to our best bilingual runs decreased only about 0.6% on the English collection, about 1% on the French collection and about 7,5% on the German collection. We still contribute those results to the quality of the Google translation service. Another finding was that the experiments in which we tried to benefit from the multilinguality of the collections also performed quite good in the bilingual setting. In fact one of those experiments performed best on the French collection and on the German collection it performed almost as good as the best experiment. Only on the English collection we could not benefit from the multilinguality, where those two experiments were clearly outperformed by the standard bilingual runs.

# 4 Result Analysis - Summary

The following list provides a summary of the analysis of our retrieval experiments for the *Ad-Hoc TEL task* at CLEF 2009:

- *Combining retrieval models:* Our experiments showed that combining different retrieval models results in a small but consistent gain in terms of MAP over all target collections.
- *Monolingual task:* The submitted monolingual experiments achieved strong performance on all target collections. Interestingly the MAP on the French and German collection is almost the same, while the performance is much better on the English collection
- *Bilingual task:* Probably due to the used translation service our bilingual experiments performed very good and achieved top results on each target collection. The gap to our best corresponding monolingual runs ranged between 0.6% and 7.5%.
- *Addressing the multilinguality of the collections:* We experimented with multilingual configurations and compared them to a baseline experiment. We found that our approach to combine multiple indexed collections works quite good except for the bilingual configurations on the English target collection.

## 5 Conclusion and Future Work

In our second participation in the *CLEF Ad-Hoc TEL task* we were able to chose from a wide selection of retrieval models. The *Xtrieval framework* supports three different retrieval cores now, namely *Lucene*, *Lemur* and *Terrier*. By combining results from *Lucene* and *Terrier* we achieved constant gains in terms of mean average precision on all collections over our baseline runs. Again we found that the translation service provided by Google seems to be extremely superior to any other approach or system. We used this service for translating our bilingual and multilingual experiments and got very strong retrieval performance for all those runs. In the future we will further investigate the numerous retrieval models and try to help to develop an open-source retrieval framework for information retrieval evaluation as it was proposed by Ferro and Harman [1].

## Acknowledgments

We would like to thank Jaques Savoy and his co-workers for providing numerous resources for language processing. Also, we would like to thank Giorgio M. di Nunzio and Nicola Ferro for developing and operating the DIRECT system[7].

This work was partially accomplished in conjunction with the project *sachsMedia*, which is funded by the *Entrepreneurial Regions*[8] program of the German Federal Ministry of Education and Research.

## References

[1] Nicola Ferro and Donna Harman. Dealing with multilingual information access: Grid experiments at trebleclef. *Post-proceedings of the Fourth Italian Research Conference on Digital Library Systems (IRCDL 2008)*, pages 29–32, 2008.

[2] Jens Kürsten, Thomas Wilhelm, and Maximilian Eibl. Extensible retrieval and evaluation framework: Xtrieval. *LWA 2008: Lernen - Wissen - Adaption, Würzburg, October 2008, Workshop Proceedings*, October 2008.

[3] Jens Kürsten, Thomas Wilhelm, and Maximilian Eibl. The xtrieval framework at clef 2007: Domain-specific track. In C. Peters, V. Jijkoun, Th. Mandl, H. Müller, D.W. Oard, A. Peñas, V. Petras, and D. Santos, editors, *LNCS - Advances in Multilingual and Multimodal Information Retrieval*, volume 5152, pages 174–181, Berlin, 2008. Springer Verlag.

[4] Iadh Ounis, Christina Lioma, Craig Macdonald, and Vassilis Plachouras. Research directions in terrier: a search engine for advanced retrieval on the web. *Novatica/UPGRADE Special Issue on Next Generation Web Search*, pages 49–56, 2007.

[5] Jaques Savoy. Data fusion for effective european monolingual information retrieval. *Working Notes for the CLEF 2004 Workshop, Bath, UK*, September 2004.

[7] http://direct.dei.unipd.it

[8] The Innovation Initiative for the New German Federal States