CLEF 2008 Ad-Hoc Track: On-line Processing Experiments with Xtrieval

Jens Kürsten, Thomas Wilhelm and Maximilian Eibl
Chemnitz University of Technology
Faculty of Computer Science, Dept. Computer Science and Media
09107 Chemnitz, Germany
[jens.kuersten | thomas.wilhelm | maximilian.eibl] at cs.tu-chemnitz.de

Abstract

This article describes our first participation at the *Ad-Hoc track*. We used the *Xtrieval* framework [2], [3] for the preparation and execution of the experiments. We regard our experiments as on-line or live experiments since the preparation of all results including indexing and retrieval took us less than 4 hours in total. This year, we submitted 18 experiments in total, whereof only 4 were pure monolingual runs. In all our experiments we applied a standard top-k pseudo-relevance feedback algorithm. The translation of the topics for the multilingual experiments was realized with a plug-in to access the Google AJAX language API[2]. The performance of our monolingual experiments was slightly below the average for the German and French collection and in the top 5 for the English collection. Our bilingual experiments performed very well (at least in the top 3) for all target collections.

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 [2],[3] was used to prepare and execute this years retrieval experiments for the *Ad-Hoc track*. The core retrieval functionality is provided by Apache Lucene[1]. For the *Ad-Hoc track* 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.

We conducted monolingual experiments on each of the collections and also submitted experiments for the bilingual subtasks. For the translation of the topics the Google AJAX language API[2] was accessed through a JSON[3] programming interface. We also tried to identify all languages of each record in the collections by using the language detector, which is also available through the Google AJAX language API. But unfortunately we used the wrong document identifiers in the indexing stage and we did not realize that until we tried to submit our experiments 6 hours before the deadline. Due to that mistake we could not use this feature for the evaluation, because we had to rebuild all three indexes within a few hours.

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 sum up our observations.

2 Experimental setup

We think that our experiments for this year's *Ad-Hoc track* could be called *on-line* or *live* retrieval experiments. As already mentioned in the introduction, we used the wrong document identifiers for indexing, which resulted in completely useless experiments. We had 6 hours to fix this problem and to re-run all or at least some feasible experiments. Therefore we had to rectify and verify the indexing process. Additionally, we had to implement a simple retrieval algorithm, because our more sophisticated approach using language detection stored all language-specific information on indexing time and thus was not available for our final experiments.

Nevertheless, we used different stemming approaches for German and English and combined the results in the retrieval stage by applying our implementation of the *Z-Score* operator [4]. We also used a standard top-k pseudo-relevance feedback algorithm in the retrieval stage. Our baseline retrieval experiment was compared to three additional experiments for each monolingual subtask and one additional experiment for each bilingual subtask.

3 Configurations and Results

The detailed setup of our experiments 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 different stemmers for each language: Porter[5] and Krovetz [1] for English, Snowball[5] and a n-gram variant decompounding stemmer[6] for German and again the Snowball[5] implementation of a stemmer for French. We applied top-k (k = 10) pseudo-relevance feedback in all our experiments.

Besides a baseline experiment, which simply returns everything regardless in which language the description library record is stored, we also tried to implement a more sophisticated retrieval algorithm. In that retrieval algorithm we translate the query into the top 10 (in terms of occurrence) languages and merge these multilingual terms into a single query. We used three different weights for this query. In the first setup we weighted all topic languages equally. For the second and third configuration we used the distribution (x) of the language in the corresponding collection. In the second we weighted the topic languages with x and in the last configuration we simply used *1-x*. For the experiments with x as language weight, we want to boost documents in languages with high occurrence frequency since they will probably have more relevant documents for a specific topic. In contrast to that in the experiments with *1-x* as language weight, we assume that documents in all language might contain relevant documents and therefore push up documents in languages with low occurrence frequency in the whole collection.

In table 1, the retrieval performance of our experiments is presented in terms of mean average precision (map) and the absolute rank of the experiment in the evaluation. We compare the baseline run with experiments using different language weights (lw).

The results show that our simple (and pure monolingual) configuration always outperformed the experiments with translation and language weights. The overall performance of our experiments is also not very promising,

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

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

[3] http://json.org

[4] http://members.unine.ch/jacques.savoy/clef/index.html

[5] http://snowball.tartarus.org

[6] http://www-user.tu-chemnitz.de/˜wags/cv/clr.pdf

except for one monolingual English experiment. The results also show, that the experiments with *lw=x*, which means the weight is equivalent to the occurrence of the language in the collection, significantly outperformed the other weighting schemes for all collections.

Table 1: Experimental Results for the monolingual subtask

id	*lang*	*lw*	*map*	*rank*
cut_merged_simple	DE	-	0.2109	17/30
cut_multi10_wx_plusplus	DE	x	0.1795	19/30
cut_multi10_w1_plusplus	DE	1	0.1113	26/30
cut_multi10_w1minusx_plusplus	DE	1-x	0.1073	28/30
cut_merged_simple	EN	-	0.3562	4/37
cut_multi10_wx_plusplus	EN	x	0.2484	30/37
cut_multi10_w1minusx_plusplus	EN	1-x	0.1620	34/37
cut_multi10_w1_plusplus	EN	1	0.1453	35/37
cut_merged_simple	FR	-	0.1981	22/29
cut_multi10_wx_plusplus	FR	x	0.1629	26/29
cut_multi10_w1minusx_plusplus	FR	1-x	0.0929	28/29
cut_multi10_w1_plusplus	FR	1	0.0915	29/29

3.2 Cross-lingual Experiments

We submitted 6 experiments for the bilingual subtask, whereof 2 were submitted for each target collection. Again, we ran a baseline experiment and translated each topic with Google's translation service. Also, one experiment was submitted for each target collection using the language weights with *lw=1* (see section *Monolingual Experiments* for a detailed description). In table 2 we compare each of the bilingual experiments with respect to the performance of the corresponding monolingual experiment.

Table 2: Experimental Results for the bilingual subtask

id	*lang*	*lw*	*map*	*rank*
cut_merged_simple	DE	-	0.2109	17/30
cut_merged_simple_en2de	EN→DE	-	0.1852 (-12.19%)	2/17
cut_multi10_w1_plusplus	DE	1	0.1113	26/30
cut_merged_simple_multi10_w1_en2de	EN→DE	1	0.1126 (+01.17%)	8/17
cut_merged_simple	EN	-	0.3562	4/37
cut_merged_simple_de2en	DE→EN	-	0.3416 (-04.10%)	1/24
cut_multi10_w1_plusplus	EN	1	0.1453	35/37
cut_merged_simple_multi10_w1_de2en	DE→EN	1	0.1475 (+01.51%)	14/24
cut_merged_simple	FR	-	0.1981	22/29
cut_merged_simple_en2fr	EN→FR	-	0.1754 (-11.46%)	3/16
cut_multi10_w1_plusplus	FR	1	0.0915	29/29
cut_merged_simple_multi10_w1_en2fr	EN→FR	1	0.1270 (+38.80%)	8/16

The evaluation results of our bilingual experiments show strong performance for our baseline configurations. For these experiments the decrease in retrieval performance varies between 4 and 12 percent in comparison to the corresponding monolingual experiment. This is probably due to quality of the translation. Another interesting observation can be made by analyzing our experiments on the language weights. The bilingual experiments perform just as well as the monolingual experiments, which is actually what we did expect. Only the experiment on the French collection achieved a remarkably better performance just by translating from English (instead of French) to all nine other languages.

4 Result Analysis - Summary

The following list provides a summary of the analysis of our retrieval experiments for the *Ad-Hoc track* at CLEF 2008:

- *On-line Processing for Retrieval:* Running (= indexing and retrieving) all listed experiments in less than 4 hours was one of most interesting experiences for us in this years evaluation. This fact impressively shows the performance and adaptability of the *Xtrieval* framework.

- *Monolingual:* The performance of our monolingual experiments was slightly below the average for the German and French collection and very good for the English collection. The multilingual experiments (++) performed quite bad, mainly because we used 10 languages for querying the multilingual collections.

- *Bilingual:* Probably due to the used translation service our bilingual experiments performed very well and achieved top results on each target collection. The performance of some multilingual experiments could be improved just by using another query language. But most of these experiments produced almost the same results as they did when the language of the query and the language of the target collection were the same.

5 Conclusion and Future Work

This year, we participated in the *Ad-Hoc track* for the first time and we had to tackle a real bad problem on the day of the submission deadline. Therefore, we regard our experiments as on-line or live experiments. An important observation in all our experiments for this years CLEF campaign was that the translation service provided by Google seems to be extremely superior to any other approach or system. This should motivate the cross-language community to investigate and improve their current approaches. In the future we will try to use only 3 or 4 main languages for multilingual experiments on the collections and we assume that we can outperform our best experimental result from this work. Furthermore we will rebuild our indexes with help of language detection as we had planned and completed for the participation in this year.

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] Robert Krovetz. Viewing morphology as an inference process. In *SIGIR '93: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 191–202, New York, NY, USA, 1993. ACM.

[2] 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, Berlin, 2008. Springer Verlag.

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

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

[3] 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, to appear.

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