Domain-Specific Cross Language Retrieval: Comparing and Merging Structured and Unstructured Indices

Jens Kürsten & Maximilian Eibl
Chemnitz, University of Technology
Faculty of Computer Science, Chair Media Informatics
Straße der Nationen 62
09111 Chemnitz, Germany
[jens.kuersten | eibl] at informatik.tu-chemnitz.de

Abstract

This year, we participated in all *Monolingual*, *Bilingual* and *Multilingual tasks* of the *Domain-Specific track*. We used a redesigned version of our retrieval system prototype from 2006, which is based on the Lucene API [1]. A plugin to access the online translation services Google Translate [2] and PROMT [3] was implemented for the cross-language experiments. Furthermore, we tried to figure out the differences between plain and structured indices and also applied a data fusion approach for both index schemes. In comparison to the median of all participants of the *Monolingual tasks* we achieved average performance for our german and english and strong performance for our russian runs. The results of the cross-language tasks were robust compared to our own monolingual experiments and better than the average of the results submitted by all participants.

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.3.7 Digital Libraries

Keywords

Evaluation, Cross-Language Information Retrieval, Multi-Indexing, Combination and Fusion, Experimentation

1 Introduction and outline

This year, we submitted experiments for all *Mono- and Cross-lingual tasks* of the *Domain-Specific track*. Therefore, we redesigned our last years retrieval system prototype (see [4] for a short description). The main reason for that was the need for better extensibility of the system. Furthermore, we decided to use a multi-index approach for our experiments, because of the promising results of our last years participation in the *Monolingual task* and further tests on the training data. But we did not use the sophisticated optimization approaches based on local clustering (see [4] for more details) this time.

The main goal of our experiments was to achieve stable performance in all monolingual tasks as well as robust results for the cross-lingual tasks. Therefore, an online translation plug-in was implemented, which allows to apply some well-known online translation services within our framework like Babel Fish [5], Google Translate [2], PROMT [3] and Reverso [6].

The outline of the paper is as follows. Sections 2, 3 and 4 describe our setup for the monolingual, bilingual and multilingual configurations. In Sect. 5 we compare the results of our submitted runs. In the final section, we conclude our observations and discuss the results as well as our future work.

2 Monolingual configurations

In our experiments, we tried to improve the basic retrieval performance by combining different index schemes in a multi-index. Therefore, we created two different indices for each of the three languages of the task. We used the structure of the corresponding corpus for one index. For the creation of the other index we simply threw away the complete structure of the documents from the collection. We submitted one run using the structured index, another one using the plain index and finally a multi-index run combining both of them. That was the main setup for all Monolingual tasks. The data fusion of the multi-index configurations was realized with the z-score operator, which had been introduced in [7].

The general configuration of our system was as follows. We used a classic language processing chain for processing the topics, i.e. a stopword filter with the stopword lists provided by [8] and a stemming algorithm depending on the language (see the following subsections) as well as a standard tokenizer. A standard pseudo-relevance feedback approach has been used to improve retrieval performance. We also used our frequency-based topic pre-processor from last year.

2.1 English

For this task, we merged a number of indices in each run. The main difference between the configurations is the structure of the data in the indices, as mentioned above. Additionally, two different english stemming approaches were combined with data fusion in each of the submitted runs. The first was the Porter stemmer from the Snowball Project [9] and the second the Krovetz stemmer, which is described in [10].

2.2 German

For the *Monolingual German task* we only used the German2 stemmer from the Snowball Project [9]. We did not use any decompounding algorithm or decompounding stemmer as we had done last year, since we run short on time for the adaptation of the code to work within our new retrieval framework. Furthermore, we were not able to use the thesaurus for query expansion for the same reason. We did some additional experiments after the submission deadline. The results are also shown in the section 4.

2.3 Russian

In our experiments for the *Monolingual Russian task* we used an analyzer and a stemmer, which are part of an outdated version of the Lucene API [11]. Again, we combined the two different index schemes that were mentioned before.

3 Cross-lingual configurations

We developed a plug-in that is capable to access an online translation service to receive translations for the cross-lingual experiments. Namely, Google Translate [2] and PROMT [3] had been used, because they performed best in some preliminary runs. Additionaly, we used the bilingual thesauri that were provided for the tasks. The configuration of the cross-lingual runs is based on the combination of the monolingual runs on the corresponding target collection.

3.1 Bilingual configurations

The configurations we submitted for the *Bilingual task* are summarized in table 1. The merged monolingual run of the corresponding target collection is the basis of each of the configurations.

Table 1: Configurations of submitted bilingual runs

identifier	*language pair*	*translation*
CUT_DS_BILI_RU2EN_MERGED	RU-EN	Google Translate
CUT_DS_BILI_DE2EN_MERGED	DE-EN	Google Translate
CUT_DS_BILI_DE2EN_MERGED_THES	DE-EN	Google Translate + thesaurus
CUT_DS_BILI_RU2DE_MERGED	RU-DE	PROMT
CUT_DS_BILI_RU2DE_MERGED_THES	RU-DE	PROMT + thesaurus
CUT_DS_BILI_EN2DE_MERGED	EN-DE	Google Translate
CUT_DS_BILI_EN2DE_MERGED_THES	EN-DE	Google Translate + thesaurus
CUT_DS_BILI_EN2RU_MERGED	EN-RU	Google Translate
CUT_DS_BILI_DE2RU_MERGED	DE-RU	PROMT
CUT_DS_BILI_DE2RU_MERGED_THES	DE-RU	PROMT + thesaurus

3.2 Multilingual configurations

The configurations we submitted for the *Multilingual task* are summarized in table 2. We used an extension of our translation plug-in to translate into the languages of all target collections.

Table 2: Configurations of submitted multilingual runs

identifier	*source language*	*translation*
CUT_DS_MULTI_EN2X_MERGED	EN	Google Translate
CUT_DS_MULTI_DE2X_MERGED	DE	Google Translate + PROMT
CUT_DS_MULTI_RU2X_MERGED	RU	Google Translate + PROMT

4 Results

This section summarizes the results of our experiments according to the corresponding task of the *Domain-Specific track*.

4.1 Monolingual

The results for the *Monolingual English task* are shown in table 3. As we expected, the merged run performed best. An interesting observation is that the experiment based on the structured index performs very bad. Whether this is due to an unbalanced weighting scheme or to an inappropriate structure of the collection is currently under investigation. Interestingly the merged run performes only slightly better than the plain one.

Table 3: Results of the monolingual english experiments

identifier	*index structure*	*MAP*	*GMAP*
CUT_DS_MONO_EN_UNSTRUCT	plain	0.2952	0.2208
CUT_DS_MONO_EN_STRUCT	structured	0.1850	0.1124
CUT_DS_MONO_EN_MERGED	merged	0.2985	0.2218

Table 4 summarizes the results of our experiments for the Monolingual German task. Generally speaking, the results for this task turned out to be very poor in contrast to our last year's experiments. In order to further investigate the unsatisfactory results we completed an additional run subsequent to the official evaluation. This run is shown in the last row of table 4. The main alteration of this run was the integration of a thesaurus-based query expansion. Though MAP and GMAP could be slightly raised, this run as well did not perform in accordance to our expectations.

Concerning the performance of the structured index, the same conclusions as for the *Monolingual English task* can be drawn.

Table 4: Results of the monolingual german experiments

identifier	*index structure*	*MAP*	*GMAP*
CUT_DS_MONO_DE_UNSTRUCT	plain	0.2887	0.2192
CUT_DS_MONO_DE_STRUCT	structured	0.2631	0.1687
CUT_DS_MONO_DE_MERGED	merged	0.2991	0.2189
CUT_DS_MONO_DE_MERGED_THES[1]	merged	0.3495	0.2854

We also submitted experiments for the *Monolingual Russian task*. The results are shown in table 5. Again, the merged run performs best and the experiment with the structured index worst. The general performance of the runs is worse compared to the other monolingual tasks. But the evaluation results of the past years share this observation, which can be seen in [12] and [13].

Table 5: Results of the monolingual russian experiments

identifier	*index structure*	*MAP*	*GMAP*
CUT_DS_MONO_RU_UNSTRUCT	plain	0.1283	0.0108
CUT_DS_MONO_RU_STRUCT	structured	0.0898	0.0096
CUT_DS_MONO_RU_MERGED	merged	0.1312	0.0119

4.2 Bilingual

In this section, we present the results of our bilingual experiments and compare their performance to the corresponding monolingual runs. The performance of the experiments on the english target collection are compared in table 6, the results of the runs on the german target collection are shown in table 7. Table 8 lists the results corresponding to the russian target collection.

Table 6: Results of the bilingual experiments (english target collection)

identifier	*language pair*	*MAP*	*GMAP*
CUT_DS_MONO_EN_MERGED	EN-EN	0.2985	0.2218
CUT_DS_BILI_RU2EN_MERGED	RU-EN	0.2646 (-12.36%)	0.1502
CUT_DS_BILI_DE2EN_MERGED	DE-EN	0.1988 (-33.40%)	0.1453
CUT_DS_BILI_DE2EN_MERGED_THES	DE-EN	0.2027 (-32.10%)	0.1504

The results show that the russian source topics performed best for the english target collection. The robustness of our cross-lingual retrieval is approved by the small decrease of 12.36% in performance (in comparison to our best monolingual run). One can also see that using the provided bilingual thesaurus enhances the performance.

Table 7: Results of the bilingual experiments (german target collection)

identifier	*language pair*	*MAP*	*GMAP*
CUT_DS_MONO_DE_MERGED	DE-DE	0.2991	0.2189
CUT_DS_BILI_RU2DE_MERGED	RU-DE	0.1883 (-37.04%)	0.0327
CUT_DS_BILI_RU2DE_MERGED_THES	RU-DE	0.2047 (-31.56%)	0.0388
CUT_DS_BILI_EN2DE_MERGED	EN-DE	0.2012 (-32.73%)	0.0984
CUT_DS_BILI_EN2DE_MERGED_THES	EN-DE	0.2721 (-09.03%)	0.1601

[1] not officially submitted experiment

For the german target collection the english source topics achieved the best results in our experiments. Again, the use of the provided bilingual thesauri improves performance in all cases and the small gap (9.03%) between the best monolingual and bilingual experiments shows the robustness of the cross-language retrieval.

Table 8: Results of the bilingual experiments (russian target collection)

identifier	*language pair*	*MAP*	*GMAP*
CUT_DS_MONO_RU_MERGED	RU-RU	0.1312	0.0119
CUT_DS_BILI_EN2RU_MERGED	EN-RU	0.1142 (-12.96%)	0.0177
CUT_DS_BILI_DE2RU_MERGED	DE-RU	0.0938 (-28.51%)	0.0091
CUT_DS_BILI_DE2RU_MERGED_THES	DE-RU	0.0935 (-28.74%)	0.0092

The bilingual experiments on the russian target collection perform not that good, but compared to our monolingual runs the results are acceptable. Again, the gap of 12.96% to the best monolingual run is very small. In contrast to the runs on the other target collections, the utilization of the thesaurus for translation does not improve retrieval performance here.

4.3 Multilingual

The results of our multilingual experiments are summarized in table 9. It can be seen that the multilingual retrieval is still a hard task for scientific data collections. We show the performance of an additional experiment, which was not officially submitted. The performance of this experiment is best, because we changed the data fusion approach from z-score to simple sum-score merging.

Table 9: Results of the multilingual experiments

identifier	*source language*	*MAP*	*GMAP*
CUT_DS_MULTI_EN2X_MERGED	EN	0.0833	0.0399
CUT_DS_MULTI_DE2X_MERGED	DE	0.0842	0.0494
CUT_DS_MULTI_RU2X_MERGED	RU	0.0508	0.0080
CUT_DS_MULTI_EN2X_MERGED_ADD[2]	EN	0.1058	0.0503

5 Conclusions and future work

In our experiments we achieved fairly robust cross-lingual retrieval results. Nevertheless, our monolingual retrieval experiments did not meet our expectations and performed significantly worse than last year. This was mainly due to major changes in the system architecture: Some of the language processing algorithms were not ready this year. Additional experiments next to the official runs included thesaurus-based query expansion. Here, a slight increase in performance for the Monolingual tasks could be achieved.
In the future we will improve the system and implement some language processing algorithms that we already used last year. Furthermore, we have to investigate our weighting scheme to use the collection structure and implement some kind of adaptive weighting, for example. Besides, we will do some work on our translation scheme to achieve better performance in the cross-lingual tasks. Especially for the *Multilingual task*, we will focus our research on more efficient data fusion approaches.

References

[1] The Apache Software Foundation (1998-2007). Lucene. Retrieved August 16, 2007, from Lucene Web site: `http://lucene.apache.org`

[2] not officially submitted experiment

[2] Google (2007). Google Translate BETA. Retrieved August 16, 2007, from Google Web site:
`http://www.google.com/translate_t`

[3] PROMT, Ltd. (2003-2007). PROMT online-translator. Retrieved August 16, 2007, from PROMT Web site:
`http://www.online-translator.com/text.asp`

[4] Kürsten, J. & Eibl, M. (2006). Monolingual Retrieval Experiments with a Domain-Specific Document Corpus at the Chemnitz Technical University. In Working Notes for the CLEF 2006 Workshop, 20-22 September, Alicante, Spain. Retrieved August 16, 2007, from CLEF Web site:
`http://www.clef-campaign.org/2006/working_notes/workingnotes2006/kuerstenCLEF2006.pdf`

[5] Overture Services, Inc. (2006). Babel Fish online-translator. Retrieved August 16, 2007, from Altavista Web site:
`http://babelfish.altavista.com/babelfish/tr`

[6] Softissimo (2007). Reverso online-translator. Retrieved August 16, 2007, from Reverso Web site:
`http://www.reverso.net/text_translation.asp`

[7] Savoy, J. (2004). Data Fusion for Effective European Monolingual Information Retrieval. In Working Notes for the CLEF 2004 Workshop, 15-17 September, Bath, UK. Retrieved August 16, 2007, from CLEF Web site:
`http://www.clef-campaign.org/2004/working_notes/WorkingNotes2004/22.pdf`

[8] University of Neuchâtel, IIUN - computer science department (2007). CLEF and Multilingual information retrieval. Retrieved August 16, 2007, from IIUN Web site:
`http://members.unine.ch/jacques.savoy/clef/index.html`

[9] Porter, M. (2001-2007). The Snowball Project. Retrieved August 16, 2007, from Snowball Web site:
`http://snowball.tartarus.org`

[10] Krovetz, B. (1993). Viewing morphology as an inference process. In Proceedings of the 16th Annual International ACM SIGIR Conference, pages 191-202.

[11] The Apache Software Foundation (1998-2007). Lucene. Retrieved August 16, 2007, from Lucene Web site:
`http://svn.apache.org/viewvc/lucene/java/branches/lucene_1_4_2_dev/`

[12] Di Nunzio, G. M. & Ferro, N. (2005). Appendix A - Results of the Core Tracks and Domain-Specific Tracks. In Working Notes for the CLEF 2005 Workshop, 21-23 September, Vienna, Austria. Retrieved August 16, 2007, from CLEF Web site:
`http://www.clef-campaign.org/2005/working_notes/workingnotes2005/appendix_a.pdf`

[13] Di Nunzio, G. M. & Ferro, N. (2006). Appendix C - Results of the Domain Specific Track. In Working Notes for the CLEF 2006 Workshop, 20-22 September, Alicante, Spain. Retrieved August 16, 2007, from CLEF Web site:
`http://www.clef-campaign.org/2006/working_notes/workingnotes2006/Appendix_Domain%20Specific.pdf`