Eurospider at CLEF 2002

Martin Braschler, Anne Göhring, Peter Schäuble
Eurospider Information Technology AG
Schaffhauserstrasse 18, CH-8006 Zürich, Switzerland
{martin.braschler|anne.goehring|peter.schauble}@eurospider.com

Abstract. For the CLEF 2002 campaign, Eurospider participated in the multilingual and German monolingual tasks. Our main focus was on trying new merging strategies for our multilingual experiments. In this paper, we describe the setup of our system, the characteristics of our experiments, and give a first analysis of some of the results.

1 Introduction

In the following, we describe our experiments carried out for CLEF 2002. Much of the work for this year's campaign builds again on the foundation laid in the previous two CLEF campaigns [1] [2]. Eurospider participated both in the main multilingual track and in the German monolingual track. The main focus of the work was on the multilingual experiments, continuing our successful practice of combining multiple approaches to cross-language retrieval into one system. Using a combination approach makes our system more robust against deficiencies of any single CLIR approach. The main effort for this year was spent on the problem of merging multiple result lists, such as obtained either from searches on different subcollections (representing the different languages of the multilingual collection) or from searches using different retrieval methods. We feel that merging remains an unsolved problem after the first two CLEF campaigns. We introduced a new merging method based on term selection after retrieval, and we also implemented a slightly updated version of our simple interleaving merging strategy.
The remainder of this paper is structured as follows: we first discuss the system setup that we used for all experiments. This is followed by a closer look at the multilingual experiments, including details of this year's new merging strategies. The next section describes our submissions for the German monolingual track. The paper closes with conclusions and an outlook.

2 System setup

All experiments used a system consisting of the core software that is included in the "relevancy" system developed by Eurospider Information Technology AG. This core system is the same that is used in all commercial products of Eurospider. Prototypical enhancements included in the CLEF system are mainly in the components for multilingual information access (MLIA).

As an additional experiment, we collaborated this year with Université de Neuchâtel in order to produce a special multilingual run. The results for this run are based both on output from the Eurospider system and on output produced by the Neuchâtel group. More details will be given later in this paper.

Indexing: The following table gives an overview of indexing methods used for the respective languages:

Table 1. Overview of indexing setup for all languages.

Language	Stopwording	Stemming	Convert diacritical chars
English	yes	English "Porter-like" stemmer	-
French	yes	French Spider stemmer	no
German	yes	German Spider stemmer, including decompounding	yes
Italian	yes	Italian Spider stemmer	no
Spanish	yes	Spanish Spider stemmer	yes

Ranking: the system was configured to use straight Lnu.ltn weighting, as described in [3]. An exception was made for some of the monolingual runs, which either used BM25 weighting [4] or a mix of both Lnu.ltn and BM25. All runs used a blind feedback loop, expanding the query by the top 10 terms from the 10 best ranked documents.

We decided to make indexing and weighting of all languages as symmetric as possible. We did not use different weighting schemes for different languages, or different policies with regard to query expansion. By choosing this approach, we wanted to both mirror a realistic, scalable approach as well as avoid overtraining to characteristics of the CLEF multilingual collection.

Submitted Runs:

The following table summarizes the main characteristics of our official experiments:

Table 2. Main characteristics of official experiments.

Run tag	Track	Topic lang.	Topic fields	Run type	Translation	Merging strategy	Expansion	Weighting
EIT2MNU1	Multilingual	DE	TDN	Automatic	Documents (MT)	N/A	Blind Feedback 10/10	Lnu.ltn
EIT2MNF3	Multilingual	DE	TDN	Automatic	Queries (ST+MT), Documents (MT)	Term Sel	Blind Feedback 10/10	Lnu.ltn
EIT2MDF3	Multilingual	DE	TD	Automatic	Queries (ST+MT), Documents (MT)	Term Sel	Blind Feedback 10/10	Lnu.ltn
EIT2MDC3	Multilingual	DE	TD	Automatic	Queries (ST+MT), Documents (MT)	Interleave coll	Blind Feedback 10/10	Lnu.ltn
EAN2MDF4	Multilingual	EN	TD	Automatic	Queries (MT+Uni Neuchâtel)	Logrank	Blind Feedback 10/10+Uni Neuchâtel	Lnu.ltn + Uni Neuchâtel
EIT2GDB1	German Monolingual	DE	TD	Automatic	N/A	N/A	Blind Feedback 10/10	BM25
EIT2GDL1	German Monolingual	DE	TD	Automatic	N/A	N/A	Blind Feedback 10/10	Lnu.ltn
EITGDM1	German Monolingual	DE	TD	Automatic	N/A	Logrank	Blind Feedback 10/10	Lnu.ltn + BM25
EITGNM1	German Monolingual	DE	TDN	Automatic	N/A	Logrank	Blind Feedback 10/10	Lnu.ltn + BM25

3 Multilingual retrieval

As for last year, we again spent our main effort on the multilingual track. The goal of this track in CLEF is to select a topic language, and use the queries in that language to retrieve documents in five different languages from a multilingual collection. A single result list has to be returned, potentially containing documents in all languages.

A system working on such a task needs to bridge the gap between the language of the search requests and the languages used in the document collection. For translation, we have successfully used combination approaches, integrating more than one translation method, in the past two campaigns. In 2001, we attempted our most ambitious combination yet [2], combining three forms of query translation (similarity thesaurus [5] [6], machine translation and machine-readable dictionaries) with document translation (through machine translation). This year, we shifted the focus of our experiments somewhat, and used a slightly simpler combination approach,

discarding the machine-readable dictionaries from the system. The experiments used either a combination of document translation DT and two forms of query translation QT (machine translation MT and similarity thesaurus ST) or document translation only.

More than last year, we concentrated on the problem of producing the final, multilingual result list that is to be returned by the system in answer to a search request. Our combination approach, as indeed most approaches used in the CLEF 2000 and 2001 campaigns, produces various intermediate search results, either due to using different translation methods, or due to handling only a subset of the five languages at a time. These intermediate search results need then be "merged" to produce the multilingual result list necessary for submission to CLEF.

It seems to be generally agreed among the community of active participants in CLEF that merging is an important problem in designing truly multilingual retrieval systems, but as pointed out in the discussions at the CLEF 2001 workshop, not much progress has been made on this topic in the 2001 campaign.

3.1 Merging

In our previous participations to CLEF, we have stuck to two very simple merging strategies: rank-based merging, and interleaving.

For merging, we essentially distinguish two scenarios:

1. Both retrieval results were calculated on the same search space. The two result lists will essentially be a reordering of each other (with some extra items appearing at the bottom of the lists).
2. The sets of documents in the result lists are disjoint. This is the case if the runs were produced through retrieval on disjoint search spaces, e.g. one search on the English part of the multilingual CLEF collection, and another search on the French part.

Some merging strategies apply to both scenarios, whereas some strategies can only be used for scenario one. Rank-based merging can only be applied if the search spaces are shared among the lists to be merged. Interleaving is a more general strategy and can be used for both scenarios.

The main difficulty in merging is the lack of comparability of scores across different result lists. Result lists obtained from different collections, or through different weightings, are not directly comparable. The retrieval status value RSV that the weighting scheme attaches to every document is only used for sorting the list, and is only valid in the context of the query, weighting and collection used.

Both merging strategies described below address the problem by not using the RSV scores at all in determining the new rank of a document in the merged result list.

3.2 Rank-based Merging

For rank-based merging, calculation of a new RSV value for the merged list is based on the ranks of the documents in the original result lists. To calculate the new RSV of a document, its ranks in all the result lists are added. Clearly, the strategy only applies if the search space of all runs is shared, and therefore a substantial "overlap" in the documents retrieved for the individual runs exists. Since we feel that there is more importance in a rank difference between highly ranked documents than in a similar difference among lower ranked documents, we introduced a logarithmic dampening of the rank value, thus boosting the influence of highly ranked documents.

3.3 Interleaving

As an alternative that applies in both merging scenarios (same search space and disjoint search space), we have in the past used interleaving: the merged result list is produced by taking one document in turn from all individual lists. If the collections are not disjoint, duplicates will have to be eliminated after merging, for example by keeping the most highly ranked instance of a document.

3.4 New merging strategies

We introduced two new merging strategies for this year's experiments. The first is a slight update of the interleaving strategy. The second is more elaborate, and presents an attempt to guess how well a query "hit" a specific subcollection.

3.5 Collection size-based interleaving

One main deficiency of interleaving as described above is that all result lists are handled equally, taking the same number of documents from each. It is extremely difficult to determine the number of relevant documents to be expected in the individual subcollections, but we have observed that the ratio of relevant items is quite stable across the different languages in CLEF. Consequently, we have used a simple update to the straight interleaving method: since the subcollections of the CLEF multilingual collection vary considerably in size, we take the portion of documents taken from any one result list to be proportional to the size of the corresponding subcollection.

3.6 Feedback merging

The second new strategy aims to predict the amount of relevant information contributed by each subcollection for a specific search request. It does this by carrying out an initial retrieval step, and then analyzing the top ranked documents from the result set, building an "ideal" query to retrieve that set of documents. This query is then compared to the original query, determining the overlap as an indication for the degree to which the concepts of the original query are represented in the retrieval result. The better such representation, the higher is the estimate of relevant documents. The result lists are then finally merged in proportion to these estimates. The biggest advantage of this method is its query dependence: whereas all the other methods described above use fixed ratios for merging the different result sets, this method determines an "optimal" ratio per query.

3.7 Results

The results for our officially submitted multilingual experiments are detailed in the following, and a first analysis is given.

Table 3: Key performance figures for the multilingual experiments

Run tag	Average Precision	Precision @ 10 docs	Relevant retrieved
EIT2MNU1	0.3539	0.6560	5188
EIT2MNF3	0.3554	0.6520	5620
EIT2MDC3	0.3400	0.5860	5368
EIT2MDF3	0.3409	0.6040	5411
EAN2MDF4	0.3480	0.6260	5698

Total number of relevant documents: 8068
"Virtual best performance": 0.4834
"Virtual median performance": 0.2193

Table 4: Comparison of performance against median for multilingual experiments

Run Tag	Best	Above Median	Median	Below Median	Worst
EIT2MNU1	4	34	2	9	1
EIT2MNF3	5	34	3	8	0
EIT2MDC3	1	36	3	10	0
EIT2MDF3	2	37	1	10	0
EAN2MDF4	1	45	0	4	0

As can be seen from Table 3, there is very little difference between runs EIT2MDC3 and EIT2MDF3, which differ only in the merging strategy employed. Additionally to having essentially equivalent average precision values, the two runs also differ only very slightly when compared on a query-by-query basis. Clearly, the merging strategy based on feedback, which we newly introduced for this year, had little impact, even though it allows different merging ratios for different queries. When looking at the ratios which were actually used for merging, we see that the new method indeed chooses fairly different ratios for individual queries. Why this does not result into more difference is not immediately clear to us and will require further analysis.

The two runs based on full queries (including the narrative field), EIT2MNU1 and EIT2MNF3, also show little variation in performance as measured by average precision. Closer analysis shows some striking differences however: EIT2MNF3, which is based on a combination of query translation and document translation, outperforms EIT2MNU1 by more than 20% for 14 queries, whereas the reverse is only true for 4 of the queries. Also, EIT2MNF3 retrieves more than 8% more relevant documents than run EIT2MNU1. We believe that this shows that the combination approach boosts reliability of the system by retrieving extra items, but that we have not found the ideal combination strategy this year that would have maximized this potential.
The potential to retrieve more relevant items by using combination approaches is also demonstrated by our final multilingual entry, EAN2MDF4, which was produced by merging output from the Eurospider system with results kindly provided to us by Université de Neuchâtel (Prof. J. Savoy). This run retrieves the most relevant items among all our multilingual entries.
Compared to median performance, all five multilingual runs perform very well. All runs have around 80% of the queries performing on or above the median, with the Eurospider/Neuchâtel combined run outperforming the median in more than 90% of the cases. The "virtual median performance", obtained by an artificial run that mirrors the median performance among all submissions for every query, is outperformed by more than 50% for all experiments. The best run obtains slightly below 75% of the "virtual best performance", which is obtained by an artificial run combining the best entries for every query.

4 Monolingual retrieval

For monolingual retrieval, we restricted ourselves to the German document collection. We fine-tuned our submissions compared to last year, and experimented with two different weighting schemes.
In contrast to last year, we used blind feedback, which was found to be beneficial in CLEF 2001 by several groups. Our German runs used the Spider German stemmer coupled with splitting of German compound nouns (decompounding). We chose the most aggressive splitting method available in the system, in order to split a maximum number of compound nouns.

Table 5: Key performance figures for the monolingual experiments

Run tag	Average precision	Precision @ 10	Relevant retrieved
EIT2GDB1	0.4482	0.5160	1692
EIT2GDL1	0.4561	0.5420	1704
EIT2GDM1	0.4577	0.5320	1708
EIT2GNM1	0.5148	0.5940	1843

Total number of relevant documents: 1938
Virtual best performance: 0.6587
Virtual median performance: 0.4244

Table 6: Comparison of performance against median for monolingual experiments

Run Tag	Best	Above Median	Median	Below Median	Worst
EIT2GDB1	1	28	2	19	0
EIT2GDL1	3	27	5	15	0
EIT2GDM1	1	30	5	14	0
EIT2GNM1	6	32	4	8	0

The results show very little difference between EIT2GDB1 and EIT2GDL1, which used the BM25 and Lnu.ltn weighting scheme, respectively. Query-by-query analysis confirms little impact from choosing between the two alternatives. Not surprisingly, merging the two runs (into EIT2GDM1) leads again to very similar performance. On an absolute basis, the runs all perform well, with all runs having around 60%-75% of queries with performance above the median. All runs also outperform the "virtual median performance". For German monolingual, this seems to be a harder benchmark, as "virtual median performance" is around 65% of "virtual best performance", whereas for the multilingual track it was only roughly 45% of the "optimum".

5 Conclusions and Outlook

Our main focus this year were experiments on the problem of merging multiple result lists coming from the different language-specific subcollections in the multilingual task. We introduced a new method based on feedback merging, which showed little impact in practice. We will now evaluate the results more closely to determine why the new strategy hardly affected retrieval behavior.

As in last year, we again used a combination translation approach for the multilingual experiments. The results confirm last year's good performance. We can again conclude that combination approaches are robust; 80-90% of all queries performed on or above median performance in our systems. This good performance was achieved even though we avoided special configuration for individual languages, in order to more accurately reflect situations were only few details about the collections to be searched are known in advance.
For the monolingual experiments, we compared the impact of choosing between two of the most popular high-performance weighting schemes: BM25 and Lnu.ltn. We could not detect a meaningful difference, either in overall performance or when comparing individual queries. Consequently, combining the two weightings gives no clear advantage over using either one individually.

Acknowledgements

We thank Jacques Savoy from Université de Neuchâtel for providing us with his runs that we used for merging in the experiment labeled "EAN2MDF4".

References

[1] Braschler, M., Schäuble, P.: Experiments with the Eurospider Retrieval System for CLEF 2000, CLEF 2000, pages 140-148 (Carol Peters (Ed.), Cross-Language Information Retrieval and Evaluation, Workshop of the Cross-Language Evaluation Forum, CLEF 2000, Lisbon, Portugal, September 2000, Revised Papers. Lecture Notes in Computer Science, Vol. 2069, Springer, 2001, ISBN 3-540-42446-6)
[2] Braschler, M., Ripplinger, B., Schäuble, P.: Experiments with the Eurospider Retrieval System for CLEF 2001, CLEF 2001, pages 102-110 (Carol Peters, Martin Braschler, Julio Gonzalo, Michael Kluck (Eds.), Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Darmstadt, Germany, September 3-4, 2001. Revised Papers. Lecture Notes in Computer Science, Vol. 2406, Springer, 2002, ISBN 3-540-44042-9)
[3] Singhal, A., Buckley, C., Mitra, M.: Pivoted Document Length Normalization. In Proceedings of the 19th ACM SIGIR Conference on Research and Development in Information Retrieval, pages 21-29, 1996.
[4] Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., Gatford, M.: Okapi at TREC-3, Proceedings of the Third Text REtrieval Conference (TREC.3), NIST Special Publication 500-226, pages 109-126, 1994.
[5] Qiu, Y., Frei, H.: Concept Based Query Expansion. In Proceedings of the 16th ACM SIGIR Conference on Research and Development in Information Retrieval, Pittsburgh, PA, pages 160 - 169, 1993.
[6] Sheridan, P., Braschler, M., Schäuble, P.: Cross-language information retrieval in a multilingual legal domain. In Proceedings of the First European Conference on Research and Advanced Technology for Digital Libraries, pages 253 - 268, 1997.