# Experiments with the Eurospider Retrieval System for CLEF 2000

Martin Braschler, Peter Schäuble

Eurospider Information Technology AG
Schaffhauserstr. 18, 8006 Zürich, Switzerland
{braschler|schauble}@eurospider.com

This paper describes the experiment setup that we used for our CLEF participation and gives a preliminary analysis of the results that were obtained. We participated in the multilingual and monolingual tasks with three runs each. For our experiments, we investigated query translation using different approaches, as well as document translation. A main focus was the use of so-called similarity thesauri for query translation. Our approach produced promising results, and shows potential for future adaptations.

## Introduction

This paper describes our experiments conducted for CLEF 2000. We will start by outlining our system setup, including details of the collection and indexing. The paper continues with a description of the particular characteristics of the individual experiments, followed by a preliminary analysis of our results. The paper closes with a discussion of our findings.

Eurospider participated in the multilingual and monolingual retrieval tasks. For multilingual retrieval, we investigated both document and query translation, as well as a combination of the two approaches. For translation, we used similarity thesauri, a bilingual wordlist and a machine translation system. Various combinations of these resources were tested and are discussed in the following.

## Multilingual Retrieval

The goal of the multilingual task in CLEF is to pick a topic language, and use the queries to retrieve documents regardless of their language. I.e., a mixed result list has to be returned, potentially containing documents in all languages. The CLEF test collection consisted of newspapers for German (Frankfurter Rundschau, Der Spiegel), French (Le Monde), Italian (La Stampa) and English (LA Times).

We submitted three runs for this task, labeled EITCLEFM1, EITCLEFM2, and EITCLEFM3. They represent increasingly complex experiments. All runs use the German topics and all topic fields. We spent our main effort to produce these multilingual experiments. In contrast, the monolingual runs were base runs for the multilingual work, and were sent in mainly to have a comparison base.

We investigated both query translation (abbreviated »QT« in the following) and document translation (»DT«). Technologies used for query translation were similarity thesauri (»ST«), a bilingual wordlist and a commercially available machine translation (»MT«) system. For document translation, we used the same machine translation system.

These key technologies will be described in the following:

Similarity Thesaurus: The similarity thesaurus is an automatically calculated data structure, which is built on suitable training data. It links terms to lists of their statistically most similar counterparts (Qiu and Frei, 1993). If multilingual training data is used, the resulting thesaurus is also multilingual. Terms in the source language are then linked to the most similar terms in the target language (Sheridan et al., 1997). Such a thesaurus can be used to produce a »pseudo-translation« of the query by substituting the source language terms with those terms from the thesaurus that are most similar to the query as a whole.

We used training data provided by the Schweizerische Depeschenagentur (SDA, the Swiss national news wire) to build German/French and German/Italian similarity thesauri. Some of this data is well known as part of the TREC6-8 CLIR test collection. All in all, we used a total of 11 years of news reports. While SDA produces German, French and Italian news reports, it is important to note that these stories are not actual translations. They are written by different editorial staff in different places, to suit the interests of the different audiences. Therefore, the SDA training collection is a comparable corpus (as compared to a parallel corpus, which contains actual translations of all items). The ability of the similarity thesaurus calculation process to deal with comparable corpora is a major advantage, since these are usually much easier to obtain than the rare parallel corpora.

Unfortunately, we were not able to obtain suitable German/English training data to also build a German/English thesaurus. Instead, we opted to use a bilingual German/English wordlist. As will be shown below, this likely was a big disadvantage.

Bilingual wordlist: As just mentioned, we used a German/English bilingual wordlist for German/English crosslingual retrieval. We assembled this list from various free sources on the Internet. This means that the wordlist is simplistic in nature (just translation pairs, no additional information such as grammatical properties or word senses) and noisy (i.e. there is a substantial amount of incorrect entries).

Machine translation system: For a limited number of language pairs, commercial end-user machine translation products are available nowadays. Since these systems are very cheap and run on standard PC hardware, we decided to try and link such a product with both our translation component and our retrieval software. We therefore used MT to translate the document collection, enabling us to use the translated documents in our retrieval system, and also to translate the queries, combining those with the translation output from the similarity thesaurus.

We used the standard RotondoSpider retrieval system developed at Eurospider for indexing and retrieval. Additional components were used for query translation and blind feedback.

Indexing of German documents and queries used the Spider German stemmer, which is based on a dictionary coupled with a rule set for decompounding of German nouns.

Indexing of French documents and queries used the Spider French rule-based stemmer. French accents were retained, since we decided that the quality of the data from Le Monde ensured consistent use of accenting.

Indexing of Italian documents and queries used the Spider Italian rule-based stemmer. There was a simple preprocessing that replaced the combination »vowel + quote« with an accented vowel, since the La Stampa texts use this alternative way of representation for accented characters. This simple rule produces some errors if a word was actually quoted, but the error rate was considered too small to justify the development of a more sophisticated rule.

Indexing of English documents used an adapted version of the Porter rule-based stemmer.

The Spider system was configured to use a straight Lnu.ltn weighting scheme for retrieval, as described in (Singhal et al., 1996).

The ranked lists for the three multilingual runs were obtained as follows:

EITCLEFM1: We built one large unified index containing all the German documents plus all the English, French and Italian documents in their German translations as obtained by MT. It is then possible to perform straight monolingual German retrieval on this combined collection. An added benefit is the avoidance of the merging problem: since only one search has to be performed on one index, merging of multiple ranked lists is not necessary.

EITCLEFM2: Our second submission has an entirely different focus. Instead of document translation, we used only query translation for this experiment. We obtained individual runs for every language pair (German/German, German/French, German/Italian, and German/English). For every language pair, we used two different translation strategies (or in the case of German/German, two different retrieval strategies). For retrieval of the French and Italian documents, we translated the German queries both using an appropriate similarity thesaurus and using the MT system. For search on the English collection, we again used the MT system, but additionally used the German/English bilingual wordlist. The two German monolingual runs were a simple, straightforward retrieval run, and a run that was enhanced through blind relevance feedback (for a discussion of blind feedback and some possible enhancements to it, see e.g. Mitra et al., 1998). The choice of relevance feedback was to »imitate« the expansion effect of the similarity thesaurus for the other languages. We expanded the query by the twenty statistically best terms from the top 10 initially retrieved documents.

The two runs per each language are merged by adding together the ranks of a document in both individual runs to form a new score. In order to boost documents with high ranks, we used the logarithms of the ranks of the documents in both experiments.

$$\text{new\_score} = \mathit{log}(\text{ rank\_in\_run\_1 }) + \mathit{log}(\text{ rank\_in\_run\_2 })$$

The step resulted in four runs, one per language combination. These were then merged by taking a document each in turn from every run, thus producing the final ranked list.

EITCLEFM3: The last multilingual experiment combines elements from both the QT and DT-based runs. To produce the final ranked list, these two runs are merged by setting the score to the sum of the logarithms of the ranks, as described above.

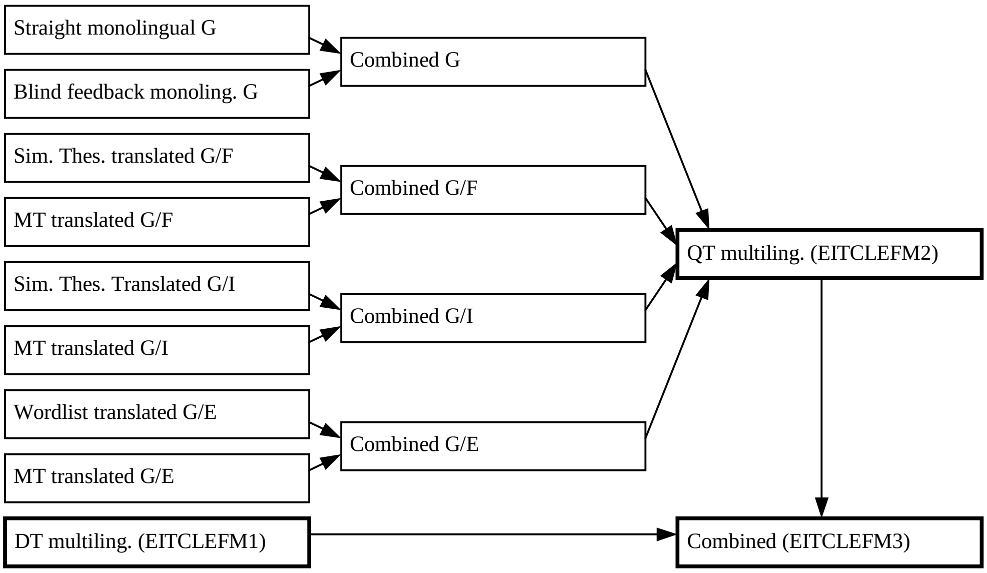

**Figure 1: Procedure to obtain the multilingual experiments.**

## Monolingual Retrieval

We also submitted three runs for the monolingual task named EITCLEFGG, EITCLEFFF and EITCLEFII (German, French and Italian monolingual, respectively). These runs all use the full topics (all fields). As mentioned earlier, they were produced mainly to serve as baselines for comparison. The main work was invested into the multilingual experiments.

EITCLEFGG: This was our German monolingual submission. It is the straight retrieval run that was used to produce the EITCLEFM2 run (see above).

EITCLEFFF and EITCLEFII: These two runs were also obtained through straight monolingual retrieval using the French and Italian queries, respectively.

## Results

Looking at the results, the document translation-based run outperforms the query translation-based run. However, looking at the individual parts that make up the QT-based run, we notice that the translation using the bilingual wordlist performs very badly. It seems likely that the actual difference would be a lot smaller if a good English similarity thesaurus was available.

| *Runs against Multilingual Collection* | *Average Precision* |
|---|---:|
| EITCLEFM1 | 0.2816 |
| EITCLEFM2 | 0.2500 |
| EITCLEFM3 | 0.3107 |

**Table 1: Average precision numbers for the multilingual experiments.**

The combined run produces the best results, and does so on a very consistent basis. As shown in table 2, the vast majority of queries improves, often substantially, in terms of average precision when compared to either the DT-only or QT-only run. The picture is less conclusive for the comparison between DT-only and QT-only. We think that this shows that whereas both approaches have strengths, they nicely mix in the combined run to boost performance.

| *Comparison*<br>*Avg. Prec. per Query* | *better;*<br>*diff. > 10%* | *better;*<br>*diff. < 10%* | *worse;*<br>*diff. < 10%* | *worse;*<br>*diff. > 10%* |
|---|---:|---:|---:|---:|
| EITCLEFM3 (combined) vs. EITCLEFM1 (DT-only) | 16 | 16 | 6 | 2 |
| EITCLEFM3 (combined) vs. EITCLEFM2 (QT-only) | 19 | 12 | 4 | 5 |
| EITCLEFM1 (DT-only) vs. EITCLEFM2 (QT-only) | 14 | 10 | 5 | 11 |

**Table 2: Comparison of average precision numbers for individual queries.**

We also looked at the individual language pairs and at the impact of the different query translation strategies.

| *Runs against German Collection* | *Average Precision* |
|---|---:|
| Straight | 0.4030 |
| Blind Feedback | 0.3994 |

**Table 3: Average precision numbers for the German monolingual runs.**

It seems like the blind feedback loop did not help boost performance. In any case, the difference is so slight that it can be considered meaningless. A per-query analysis shows that most queries are affected very little by the feedback, and that the number of queries with a substantial increase or decrease in average precision is exactly the same. This reinforces the conclusion that the feedback was not helpful in this case.

| *Runs against French Collection* | *Average Precision* |
|---|---:|
| Monolingual | 0.3884 |
| MT G/F | 0.3321 |
| Similarity Thesaurus G/F | 0.2262 |
| Combined G/F | 0.3494 |

**Table 4: Average precision numbers for runs against the French collection.**

The French MT-based run outperforms the similarity thesaurus-based run quite substantially. However, a sizable part of the difference can be attributed to five queries that failed completely using the thesaurus (we consider a query a complete failure if the result has an average precision < 0.01). For the rest of the queries, the similarity thesaurus performed well, even outperforming the MT-based run by more than 10% for eight queries in terms of average precision. The combined run gives a modest improvement over the MT run. 20 queries benefit from the combination, whereas the performance of the remaining 14 queries falls.

| ***Runs against Italian Collection*** | ***Average Precision*** |
|---|---:|
| Monolingual | 0.4319 |
| MT G/I | 0.3306 |
| Similarity Thesaurus G/I | 0.2568 |
| Combined G/I | 0.3636 |

**Table 5: Average precision numbers for runs against the Italian collection.**

In Italian, the similarity thesaurus is closer to the performance of the MT-based run. Again, a big part of the difference is due to 7 queries failing completely when using the thesaurus. The combination is quite an improvement over the MT-only run, gaining 10% in average precision.

| ***Runs against English Collection*** | ***Average Precision*** |
|---|---:|
| Monolingual | 0.3879 |
| MT G/E | 0.3753 |
| Wordlist G/E | 0.1414 |
| Combined G/E | 0.2809 |

**Table 6: Average precision numbers for runs against the English collection.**

In English, the good performance of the MT-based run is striking. This probably is due to the main effort in MT research still going into language combinations involving English. The poor performance of the run using the bilingual wordlist is also noteworthy. While this might be partly due to shaky quality of the input sources, we think that it underscores how important word sense disambiguation is, something which MT and the similarity thesaurus try to address, but which is lacking from the wordlist. It seems obvious that bilingual wordlists/dictionaries are not competitive without a serious investment of effort in that direction.

We are pleased to see that our runs compare very favorably when compared to other entries in CLEF. Table 7 shows an analysis of per-query performance compared to the median performance of all participants. Especially the multilingual runs performed strongly. The monolingual runs are more mixed, which was to be expected, since we did not tune them specifically for performance. While German seems to perform nicely, probably due to the compound analysis in the Spider stemming, the results for French and Italian indicate room for improvement.

| | ***Best*** | ***Above*** | ***Median*** | ***Below*** | ***Worst*** | ***# queries*** |
|---|---:|---:|---:|---:|---:|---:|
| EITCLEFM1 | 1 | 29 | 0 | 10 | 0 | 40 |
| EITCLEFM2 | 1 | 22 | 2 | 15 | 0 | 40 |
| EITCLEFM3 | 7 | 23 | 1 | 9 | 0 | 40 |
| EITCLEFGG | 6 | 17 | 6 | 8 | 0 | 37 |
| EITCLEFFF | 0 | 7 | 5 | 22 | 0 | 34 |
| EITCLEFII | 3 | 7 | 7 | 17 | 0 | 34 |

**Table 7: Officially submitted runs compared to median of all submitted runs (on individual query basis).**

## Conclusions

Overall, we think the performance of the similarity thesaurus is very remarkable. While it did not produce results equal to the MT-based runs, it is important to note that we were in a »worst-case scenario«: the thesauri were built on a comparable corpus (no real translations, as opposed to a parallel corpora), and there was absolutely no overlap in training data and the test collection. This means that similar requirements for other translation scenarios can be quite easily matched. I.e., it would be easy to build similarity thesauri with comparable performance for a multitude of additional language pairs, even exotic ones, simply by gathering suitable training data, such as a sufficient amount of texts from a national newspaper each. Also, the performance of the similarity thesaurus will get a sizeable boost when the problems can be addressed that led to a complete failure in translation of a number of queries. We should be able to do this by increasing the size of the thesaurus, which again is only a matter of processing more training data. Note also that the thesaurus is very suited for situations in which the query length is much shorter, such as Web searches. As shown during the Eurosearch project (for a short description of Eurosearch, see Braschler et al., 1998), the expansion effect of the thesaurus is very beneficial for the short queries. Machine translation system traditionally have problems with short, key-word style queries.

Document translation gave us some good results, and was feasible for a collection of the size of the CLEF test collection. This means that DT should not be discounted for reasonably static collections with limited size. Note, however, that some of the advantage we found for DT versus query translation may be due to the inadequate performance from the wordlist we used for English. Also, QT clearly remains the only possibility for huge or highly dynamic collections.

## Acknowledgements

Thanks go to the providers of the CLEF data and all the assessors doing the judgement work. Also, we appreciate the data that was provided to us by SDA for building the similarity thesauri. Additional thanks go to Patrick Frey for tuning the rules of the German stemmer and Min-Yen Kan for a lot of scripts originally written for last year's TREC that proved a big help in doing these experiments.

## References

(Braschler et al., 1998) M. Braschler, C. Peters, E. Picchi, P. Schäuble. Cross-Language Web Querying: The EuroSearch Approach. In Proceedings of the 2nd European Conference on Research and Advanced Technology for Digital Libraries, pages 701 - 702, 1998.

(Mitra et al., 1998) M. Mitra, A. Singhal, and C. Buckley. Improving Automatic Query Expansion. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 206 - 214, 1998.

(Qiu and Frei, 1993) Y. Qiu and H. Frei. 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.

(Sheridan et al., 1997) P. Sheridan, M. Braschler, and P. Schäuble. 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.

(Singhal et al., 1996) A. Singhal, C. Buckley, and M. Mitra. Pivoted Document Length Normalization. In Proceedings of the 19th ACM SIGIR Conference on Research and Development in Information Retrieval, pages 21 - 29, 1996.