The Eurospider Retrieval System and the TREC-8 Cross-Language Track

Martin Braschler[1], Min-Yen Kan[2], Peter Schäuble[1], Judith L. Klavans[2]
[1] Eurospider Information Technology AG
Schaffhauserstrasse 18
CH-8006 Zürich
Switzerland
{braschle, schauble}@eurospider.ch

[2] Natural Language Processing Group
Columbia University
New York, NY 10027
USA
{min, klavans}@cs.columbia.edu

This year the Eurospider team, with help from Columbia, focused on trying different combinations of translation approaches. We investigated the use and integration of pseudo-relevance feedback, multilingual similarity thesauri and machine translation. We also looked at different ways of merging individual cross-language retrieval runs to produce multilingual result lists. We participated in both the CLIR main task and the GIRT sub task.

1. Introduction

The main aim of our participation in the cross-language track this year was to try different combinations of various individual *cross-language information retrieval* (CLIR) approaches. We reused the same corpus-based methods that we utilized last year with considerable success, while experimenting with using a number of off-the-shelf machine translation products.

We also revisited our merging approach, trying out an alternative strategy.

2. General system description

For all of our runs we used a Eurospider retrieval system, which evolved from a prototype originally created at the Swiss Federal Institute of Technology (ETH) in Zurich, with continuing development of the system now at Eurospider Information Technology AG. When indexing the different collections, we used different stemmers for the individual languages:

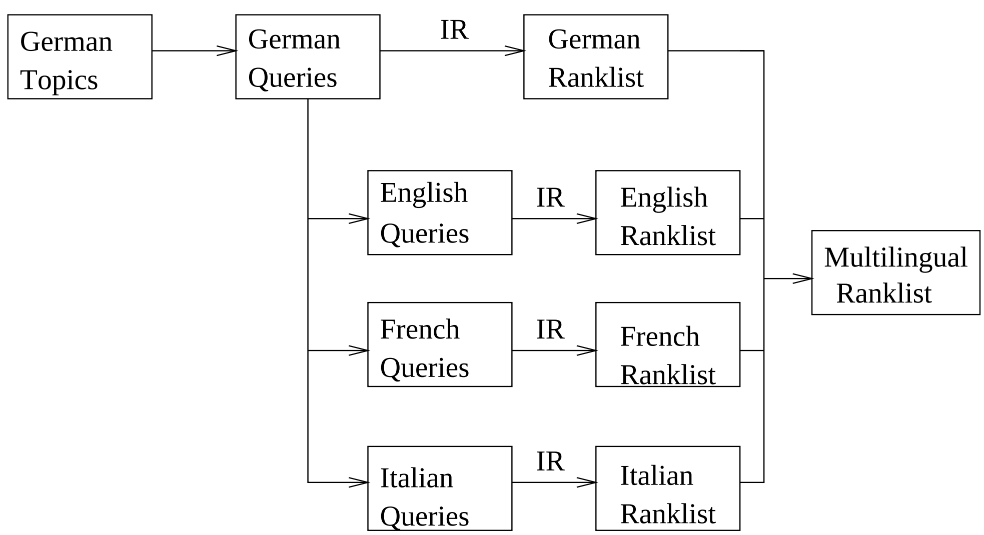
- German:
  - the stemmer distributed with the NIST PRISE retrieval system for our submissions to the main task, and;
  - the Eurospider stemmer featuring German word decomposition for the GIRT subtask submissions.
- French: the Eurospider French stemmer.

- Italian: the Eurospider Italian stemmer.
- English: the Porter English stemmer.

The retrieval status values are calculated using the *Lnu.ltn* weighting scheme as described in [6].

3. Main Task

The main task this year consisted of choosing a language in which topics are specified. Queries in that one language are then used to produce runs against all the documents in all languages (i.e. English, German, French and Italian documents). Our approach to this task is to initially produce runs using only a pair of languages, and then to merge these separate runs to produce the final, multilingual ranked list. For all submissions, we used German topics, and translated them into the other languages. This means that for all our main task submissions, we first had to obtain four runs (German monolingual, German → English, German → French and German → Italian).



The individual submissions differ in:

- the method used for query translation;
- the method used for merging of the runs, and;
- the fields of the topics used for creating the queries.

We submitted three runs for the main task, EIT99sta, EIT99mta and EIT99sal, which are explained in more detail in the following sections. All runs were automatic.

3.1. *EIT99sta* – Using only Automatic Compiled Resources

This run uses all three fields of the topics (title + description + narrative). It builds on methods we tested in TREC-7. The defining characteristic is that it does not use any costly, manually built linguistic resources. Instead, it uses only data structures automatically built from suitable training data.

This run uses two different methods to perform query translation: *pseudo relevance feedback* (PRF) and *similarity thesauri*. This year we used a German ↔ English similarity thesaurus, as opposed to using a manually built word list that as in TREC-7. Using the thesaurus for this run likely degraded performance, since the training data for German ↔ English was not well suited for thesaurus construction. However, we wanted to more clearly distinguish this run from run EIT99mta (see section 3.2), making this run completely free of manually constructed resources.

3.1.1. Pseudo relevance feedback

For pseudo relevance feedback, we use the fact that the TREC CLIR collections have similar content (i.e. news stories and articles), even though they are written in different languages. Therefore, we can calculate which items in these collections cover the most similar stories, using a process we call document alignment. Ultimately, we obtained three lists with pairs of the most statistically similar documents in the combined German-English collection, German-French collection and the German-Italian collection. We used these lists for the three cross-language runs. For more details on the document alignment process, see [2].

These lists are applied as follows: to retrieve French documents using a German query, we first run the German queries against the German documents, obtaining an initial result list. We then compare this list to our list of pairs of similar French and German documents. If any of the documents in the German result list have a similar French counterpart, they are replaced. If there is no matching pair, the document is discarded. Through the replacement step, we obtain a possibly shorter French result list. We then use the top documents from this list to do a pseudo relevance feedback loop, (i.e. we select the most significant terms from this set of documents using methods developed for relevance feedback — see also [3] and [5]). These terms form our French query, which we run against the French documents, to obtain a French result list.

Note that the multilingual collections used for document alignment do not necessarily have to be identical to the search collections, although they were in the case of our TREC experiments.

3.1.2. Similarity Thesaurus

A similarity thesaurus is a data structure that provides a list of terms in one language that are statistically similar to a head term in another language. Such a similarity thesaurus can be automatically built using suitable multilingual training data [4]. We built and used three such thesauri, German ↔ French, German ↔ Italian, and German ↔ English. The German ↔ French and German ↔ Italian thesauri were built on supersets of the TREC SDA data, enriched by additional years of SDA data, which was provided to us by SDA. The German ↔ English thesaurus was built using the TREC German SDA data and the TREC English AP collection.

In relevance feedback, the original query is usually expanded by terms coming from a term selection process. The new combined query is then reweighted. However, in the cross-language case, we cannot use the terms from the original query, since they are in the wrong language. We therefore replace the original query with terms selected from the similarity thesaurus. We can then apply term reweighting to combine the resulting terms from both methods, similar to the reweighting step in the classical relevance feedback case. For this run, we added the similarity thesaurus translations only in cases when we had few documents (less than three) coming from the PRF method.

3.1.3. Runs

The four individual runs used to produce the merged result list were obtained as follows:

German monolingual: German retrieval run, followed by pseudo relevance feedback, using the top 21 ranked documents for feedback[1]. This PRF loop is very similar to the multilingual case described above, only we can directly apply term selection without having to do document replacement. We used the NIST stemmer to index the documents and queries.

German → French: This run used PRF combined with a German/French similarity thesaurus built on a German/French SDA superset, as described above. We used the top 21 documents for the pseudo feedback loop.

German → Italian: This run used PRF combined with a German/Italian similarity thesaurus built on a German/Italian SDA superset, as described above. We used the top 21 documents for the pseudo feedback loop.

German → English: This run used PRF combined with a German/English similarity thesaurus built on the German SDA and English AP collections, as described above. We used the top 21 documents for the pseudo feedback loop.

3.1.4. Merging

For merging, we again used the document alignments from the pairs of the individual collections. We produced tables giving relations between scores of individual runs, making it possible to map these scores to a common range using linear regression. By repeatedly merging pairs of runs, we obtained the multilingual result lists that we submitted. This merging strategy is also detailed in [2].

3.2. EIT99mta – Adding Machine Translation Resources

As mentioned in the introduction, our aim was to test a combination of approaches to cross-language retrieval. We therefore added *machine translation* (MT) to the components used in the previous run.

This run used all topic fields, namely title + description + narrative.

3.2.1. Machine Translation

Machine translation is interesting for use in cross-language retrieval, since the majority of these systems utilize linguistic knowledge. However, we believe that MT cannot be the only solution to CLIR. This is because even though we invested considerable effort, we were not able to locate an off-the-shelf German ↔ Italian machine translation system. We think the fact that these two widely

---

[1] The somewhat strange number 21 is due to a minor bug. We intended to use 20 documents.

spoken European languages are not covered by commercially available software shows that very few language pairs seem to be economically viable, given the considerable effort required to build these systems[2]. In fact, nearly all systems we were able to locate on the consumer market translate either from or to English.

3.2.2. MT Systems Used

We used the following MT systems:

German → English:

- MZ Translator from Holtschke GmbH
- T1 Translator from Langenscheidt
- Systran web translation from Systran
- Power Translator 2000 from Pons

German → French:

- MZ Translator from Holtschke GmbH
- Systran web translation from Systran

German → Italian:

- Systran web translation, using English as a pivot language (German → English → Italian)

3.2.3. Runs

For all languages, we also used the translation coming from the similarity thesaurus (see section 3.1).

We created the EIT99mta submission by combining the output from all MT systems for a given target language with the similarity thesaurus output to create an intermediate query. This query was then used in the pseudo relevance feedback loop. Our internal tests showed that using all MT systems produced small performance gains over using just the best MT system (Systran or Power Translator, depending on query fields used and query language).

3.2.4. Merging

The four individual runs were merged using the same strategy as explained in section 3.1.4.

---

[2] This does not exclude the possibility that professional systems for corporate use for this language pair exist, but these are usually priced outside of the range of many potential customers. Holtschke GmbH is advertising a German ↔ Italian system, but it uses a very small dictionary compared to their other language combinations. A few months after our TREC experiments, we became aware of a new system by LHS for German ↔ Italian. We have not been able to obtain a copy in time for this paper.

3.3. EIT99sal – Experimental Run

Our last submission for the main task used only the title + description fields of the topics. This run used a combination of pseudo relevance feedback and similarity thesaurus. Apart from the different query length, there were three modifications with respect to this run: only 10 documents were used for PRF, the similarity thesaurus translation was employed for every query, and a different merging process was used.

The merging for the first two runs, EIT99sta and EIT99mta, both calculate a relation between the retrieval status values (RSVs) of the two runs to merge. For EIT99sal, we calculated a relation between the RSVs of one run, and the *rank* of a similar document in the other run. Since RSVs tend to fall logarithmically in our system, we used logarithmic regression to obtain the relation between RSVs and ranks of the two runs.

Initial tests showed that both merging methods resulted in similar performance.

3.4. Results

The following table shows the results we obtained for the three runs we submitted for the main task.

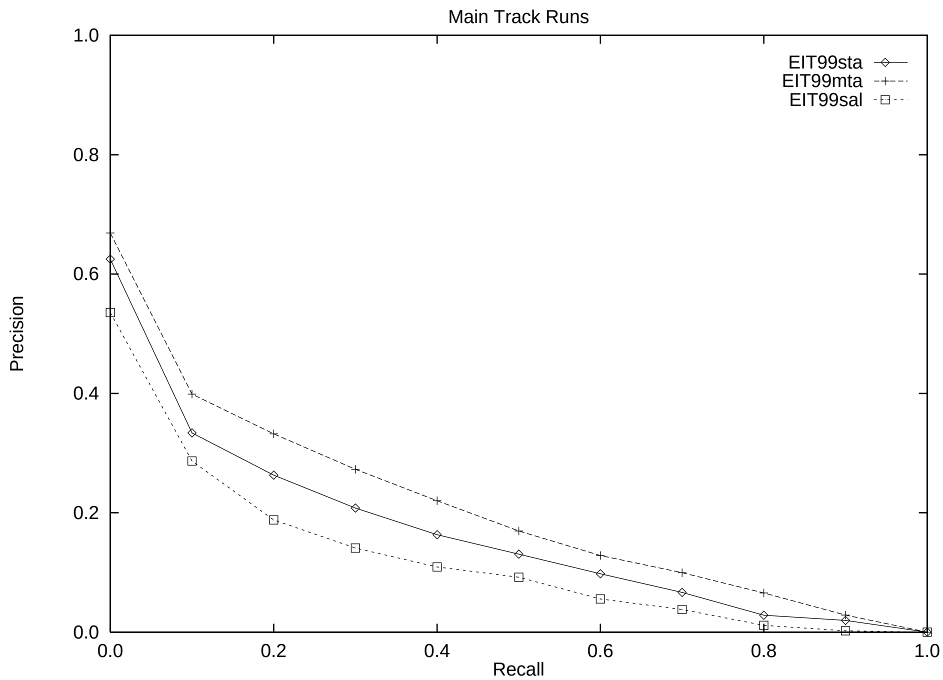
			Performance of individual queries				
Run	Avg. Prec.	R-Prec.	Best	Above	Median	Below	Worst
EIT99mta	0.1937	0.2415	2	10		16	
EIT99sta	0.1527	0.2006		9	1	14	4
EIT99sal	0.1108	0.1682	1	4	1	20	2

The results this year are somewhat mediocre, a surprise after our combination of similarity thesaurus and pseudo relevance feedback worked very well last year. We have also observed that with our new system, we don't reach the same level of performance as last year when we run the old TREC-7 queries. What exactly causes this problem is still unclear.

The rather big difference between average precision and R-precision shows that the precision seems to tail off quickly with higher recall. Some degradation may also be due to the fact that this year we used a German ↔ English similarity thesaurus, despite our belief that the training data (the German SDA in conjunction with the English AP) was not well suited for this purpose. We still need to examine these factors together with an analysis of the performance of the different merging strategies.

Not surprisingly our second run, EIT99mta, which combined most resources to produce a translation, performed best. However, for a sizeable number of queries, EIT99sta performed nearly as well, or even better (better results are obtained for 4 of 28 queries). Since this is a completely corpus-based run which did not use any manually built language resources, there are some queries which retrieve no or only very little relevant documents. These queries lower the run's performance considerably.

EIT99sal was an experimental run to examine a different method for merging. It used shorter queries, and performed poorer than the other two runs, since all the other methods we used benefited from the additional context of the longer queries.



4. GIRT Sub Task

For the GIRT subtask, the documents only exist in German, with the queries available in German, French and English. We did pure CLIR runs, ignoring both the English titles provided with the GIRT documents and the classification terms.

4.1. Runs

We submitted three runs:

- EIT99gfg, using a French ↔ German similarity thesaurus. The French queries are translated through obtaining the 20 most similar German terms from the thesaurus. No relevance feedback was used.
- EIT99geg, using an English ↔ German similarity thesaurus. Similar to the German ↔ English thesaurus, we doubted its quality due of the lack of suitable training data.
- EIT99gmt, a French ↔ German run, which used the Systran web translation to translate the queries.

All runs use all topic fields (title + description + narrative). All were automatic runs.

4.2.GIRT results

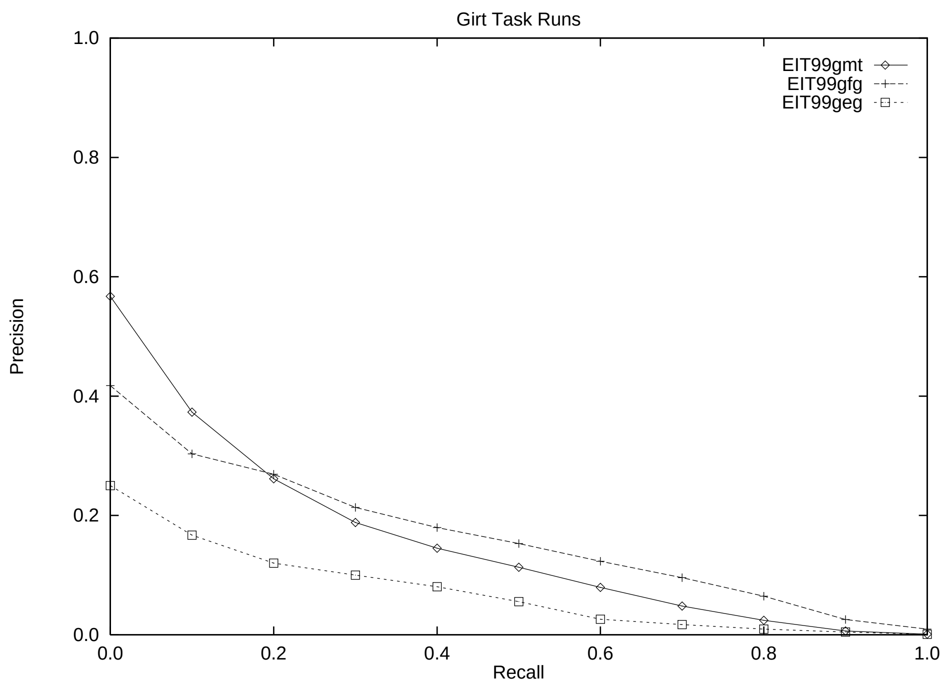
The following table shows the results we obtained for the three runs we submitted for the GIRT task.

			Performance of individual queries				
Run	Avg. Prec.	R-Prec.	Best	Above	Median	Below	Worst
EIT99gfg	0.1547	0.1844	5	4	1	14	4
EIT99gmt	0.1438	0.1965	5	3	1	15	4
EIT99geg	0.0624	0.1002	2	4		11	11

The numbers for above median and below median should be read with caution, since only few GIRT runs were submitted, making it hard to compare.

We observed however, that the results swing heavily to either side. Looking closer, we discovered that unfortunately we had a mismatch in stemming between the terms coming from our similarity thesaurus translation and the document collection. This is likely to have caused a fair number of good translated terms to be lost, which would explain the uneven performance on individual queries. We will try to analyze this problem further.

It is interesting to see that the French → German similarity thesaurus run (EIT99gfg) outperformed a high quality Systran MT run (EIT99gmt). Not surprisingly, the English → German run did much worse, again supporting our suspicion that the thesaurus is of inferior quality.



5. Thanks

Our thanks go to everyone that helped in preparing the document collections, queries and relevance assessments used in this year's CLIR track. We also thank SDA for providing us with the training data used to create the similarity thesauri.

References

[1] M. Braschler, B. Mateev, E. Mittendorf, P. Schäuble and M. Wechsler. SPIDER Retrieval System at TREC7. In *Information Technology: The Seventh Text REtrieval Conference (TREC-7),* pages 509-517, 1999.

[2] M. Braschler and P. Schäuble. Multilingual Information Retrieval Based on Document Alignment Techniques. In *Second European Conference on Research and Advanced Technology for Digital Libraries*, pages 183-197, 1998.

[3] D. K. Harman. Relevance Feedback and Other Query Modification Techniques. In *Frakes, W. B., Baeza-Yates, R.: Information Retrieval, Data Structures & Algorithms*, pages 241-261, 1992.

[4] Y. Qiu. Automatic Query Expansion Based on A Similarity Thesaurus. *PhD Thesis, Swiss Federal Institute of Technology (ETH)*, 1995.

[5] 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.

[6] 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.