Experiments with the Eurospider Retrieval System for CLEF 2001

Martin Braschler, Bärbel Ripplinger, Peter Schäuble

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

Abstract. Eurospider participated in both the multilingual and monolingual retrieval tasks for CLEF 2001. Our multilingual experiments, the main focus of this year's work, combine multiple approaches to cross-language retrieval: machine translation, similarity thesauri, and machine-readable dictionaries. We experimented with both query translation and document translation. The monolingual experiments focused on the use of two fundamentally different stemming components: one commercially and one linguistically motivated stemmer.

1 Introduction

This paper describes our experiments conducted for CLEF 2001. Much of the work for this year builds directly on ideas we already applied to last year's experiments [1]. First, we present our system setup, and outline some details of the collection and indexing. This is followed by a description of the particular characteristics of the individual experiments, including a comparison to last year, and a preliminary analysis of our results. The paper closes with a discussion of our findings.

Eurospider participated in the multilingual and German and French monolingual retrieval tasks. Our experiments in multilingual retrieval try to combine as many approaches to translation as possible in order to obtain a robust system that delivers good results in the widest range of situations possible: we used similarity thesauri, machine-readable dictionaries and a machine translation system. We tried both document and query translation. The focus of the monolingual experiments was an investigation into various aspects of stemming.

2 System Setup

For our runs, we used the standard Eurospider retrieval system, a core part of all Eurospider commercial products, enhanced by some experimental multilingual information access (MLIA) components.

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

Indexing of French documents and queries for the multilingual task used the French Spider stemmer. French accents were retained.

Some more in-depth experiments regarding stemming for German and French were carried out for the monolingual task.

Indexing of Italian documents and queries used the Spider Italian rule-based stemmer. For the La Stampa documents, there was a simple preprocessing that replaced the combination "vowel + quote" with an accented vowel, to normalize the alternative way of representation for accented characters in this subcollection. This simple rule produces some errors if a word was intentionally quoted, but the error rate was considered too small to justify the development of a more sophisticated replacement process. This heuristic was not necessary for the AGZ/SDA Italian texts.

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

Indexing of the Spanish documents used a new experimental stemmer specifically developed for this task.

The Spider system was configured to use a straight Lnu.ltn weighting scheme for retrieval, as described in [5].

The CLEF multilingual test collection consists of newspaper and newswire articles for German (Frankfurter Rundschau, Der Spiegel, SDA), French (Le Monde, ATS), Italian (La Stampa, AGZ), English (LA Times) and, new in 2001, Spanish (EFE). There are additional documents in Dutch and German, which are used for special subtasks that we did not participate in.

3 Multilingual Retrieval

We spent our main effort on our experiments for the multilingual task. The goal of this task in CLEF is to pick a topic language, and use the queries to retrieve documents independent of their language. I.e., a mixed result list has to be returned, potentially containing documents in all languages (English, French, German, Italian and Spanish).

We submitted four runs for this task, labeled EIT01M1N, EIT01M2N, and EIT01M3D/EIT01M3N. They represent increasingly complex experiments. All runs use the German topics; the "N" runs use all topic fields, whereas the "D" run uses title+description only.

We investigated both query translation (also abbreviated "QT" in the following) and document translation ("DT"). Technologies used for query translation were similarity thesauri ("ST"), machine-readable dictionaries ("MRD") and a commercially available machine translation ("MT") system. For document translation, only the MT system was used.

Following is a description of these key technologies.

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 [2]. 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 [4]. 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.

Because some of the data that was newly added to the CLEF collections overlaps with the training data of the thesauri we used for our 2000 CLEF experiments, we had to rebuild all thesauri to make sure that the training data is completely disjoint from the CLEF collection.

For German/French and German/Italian, we used training data provided by the Schweizerische Depeschenagentur (SDA), which is from a different time period than the SDA data in CLEF. For German/Spanish, we aligned German SDA with Spanish texts from Reuters and Agence France Presse (AFP). Since this thesaurus is only used to search the Spanish subcollection (EFE), the use of all SDA data was acceptable. For German/English, we used German SDA data aligned to English Associated Press (AP) data.

There was a considerable difference in the amount of training data available to build the thesauri. While the training data for German/French and German/Italian was substantial (roughly 10 and 9 years of newswire articles, respectively), we started from scratch for German/English (we used no German/English ST in 2000) and German/Spanish. This means that the resulting thesauri for these latter language combinations were not as well refined as for the earlier language pairs. We expected this to have a significant impact on retrieval quality.

In all cases, training used comparable corpora, not parallel corpora that contain real translations.

Machine-Readable Dictionaries: This year, we added general-purpose MRDs to our experiments. These dictionaries were used for pre-translation of queries, without a proper integration into the weighting mechanism of the system. Therefore, we used a heuristic to decide on the number of potential translations generated for ambiguous terms. This lack of integration limited the control over terms left untranslated due to gaps in the dictionary.

Table 1. Size of the machine-readable dictionaries used for the multilingual experiments

Language Pair	# of Entries
German - English	486,851
German - French	70,161
German - Italian	7,953
German - Spanish	36,636

Machine translation system: For a limited number of language pairs, commercial end-user machine translation products are available. Since some of these systems are inexpensive and run on standard PC hardware, we decided to loosely combine such a product with both our translation component and our retrieval software. We used MT to translate both the document collection and the queries.

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

EIT01M1N: This run is based on one large, unified index containing all German documents plus the MT-translations of all English, French, Italian and Spanish documents. Because we had no direct German/Spanish machine translation available, we used a two-step German/English/Spanish translation in this case. We then performed straight German monolingual retrieval on this index. An added benefit is the avoidance of the merging problem that typically arises when results are calculated one language at a time. Since only one search has to be performed on one index, a single ranked list is obtained.

EIT01M2N: This is an experiment based on query translation. We obtained individual bilingual runs for each language pair (German/French, German/Italian, German/Spanish, and German/English). For each pair, we used three different translation strategies: similarity thesaurus, machine translation, and machine-readable dictionaries. The ranked lists obtained were then merged to produce the bilingual results. In a last step, these bilingual results, plus a monolingual German run, were merged into the final multilingual result.

EIT01M3D/EIT01M3N: The two runs are related: EIT01M3D used only the title and description fields, whereas EIT01M3N used all topic fields. The two experiments combine all elements described for the EIT01M1N (DT-based) and EIT01M2N (QT-based) runs. EIT01M3N is the result of merging these two runs, whereas EIT01M3D is the result of merging the two corresponding title+description runs (which were not submitted as official experiments).

3 Monolingual Retrieval

Our interest in the monolingual track was to investigate the effects of stemming for the German and the French language. We had the opportunity this year to use the MPRO morpho-syntactic analysis for some research experiments. This analysis component contains elaborate linguistic information, among them base forms and compound analysis for German [3].

By standard, the Eurospider system uses stemming procedures that have been adapted over the years specifically to the needs communicated by customers of Eurospider's commercial retrieval systems. We were interested to see how such a "commercial" approach compares to a more linguistically motivated alternative.

We submitted three runs for the German monolingual task: EIT01GGSSN, an all-topic-fields run using the original Spider stemmer; EIT01GGLUN, a run using the MPRO morpho-syntactic analysis, and EIT01GGLUD, a variant of the second run, using topic+description fields only.

For French monolingual, we also submitted three runs, EIT01FFFN, a run using a new experimental variant of the French Spider stemmer, and EIT01FFLUN, a run using the MPRO analysis. Our third run, EIT01FFFD, was mistakenly lost during the submission process, and therefore had to be disqualified from the official evaluation. While analyzing our results, we found a bug in EIT01FFFN. After fixing this bug (missing accents in queries), performance improved significantly.

The concentration on stemming means that we did not use some "enhancements", such as blind feedback, that probably would have increased overall performance, but that we felt make it harder to investigate the impact of stemming. The runs therefore are simplistic: removal of stopwords, stemming, and then straight retrieval.

4 Results

Multilingual: One of the major obstacles when performing our experiments for this year was a lack of suitable training data to build the German/English and German/Spanish similarity thesauri. We built the thesauri for these two languages even though we expected their quality to be inadequate. The reason was two-fold: one, to investigate the effects of using (too) little training data, and two, to build a system that treats all languages equally.

Analysis of the QT-based run (EIT01M2N) shows that both the German/English and German/Spanish components performed poorly, and therefore hurt overall performance. Unfortunately, the German/French and German/Italian thesauri also did not perform as well as last year. We assume this to be due to the exclusion of the SDA data from 1994 (which was used in the CLEF document collection). However, the German/French and German/Italian thesauri performed much better than their English and Spanish counterparts due to their much larger training sets.

The dictionary-based components we introduced into the QT-based run suffered from a similar problem, with the Italian dictionary being very small. Again, we expected this to hurt overall performance, but the Italian dictionary was used to allow consistent handling of the languages.

Comparing the three translation methods used for each language pair (MT, ST, MRD), machine translation generally performed best. There are, however, big performance differences between more "popular" language combinations (German/English) and less "popular" ones (German/Italian).

The similarity thesaurus did significantly worse. Like last year, we observe that a sizable part of the difference is caused by a subset of queries that completely fail to retrieve anything relevant. The remaining queries performs well, but the average performance suffers from the outliers.

We observed last year that the combination of machine translation with similarity thesauri substantially outperformed the use of a single strategy. This year, the combination generally gives only performance comparable to machine translation alone, probably due to the less appropriate quality of the thesauri. However, we observed an increase in recall, and also better performance in the high precision range.

The dictionary-based translations overall performed similarly to the similarity thesaurus-based translations. When combined with machine translation, the dictionary gave no advantage, instead negatively affecting retrieval performance.

Table 2. Average precision numbers for the multilingual experiments

Runs against Multilingual Collection	Average Precision
EIT01M1N (DT; TDN)	0.3099
EIT01M2N (QT; TDN)	0.2773
EIT01M3D (Combination; TD)	0.3128
EIT01M3N (Combination; TDN)	0.3416

The combined run produces the best results, and does so on a consistent basis. As shown in table 3, the majority of queries improves, often substantially, in terms of average precision when compared to the DT-only or QT-only run. The picture is less conclusive for the comparison between DT-only and QT-only. This seems to indicate that the mixture works well and boosts performance.

Table 3. Comparison of average precision numbers for individual queries

Comparison Avg. Prec. per Query	better; diff.>10%	better; diff.<10%	worse; diff.<10%	worse; diff.>10%
EIT01M3N (comb.) vs. EIT01M1N (DT)	20	21	9	0
EIT01M3N (comb.) vs. EIT01M2N (QT)	20	16	9	5
EIT01M1N (DT) vs. EIT01M2N (QT)	17	8	9	16

For French, we observed good performance of the MT translations. The similarity thesaurus performed appropriately, but not as well as last year. This may be due to the mismatch between the time period covered in the training data and the CLEF test set. The performance of dictionary-based translations was adequate, thanks to the large dictionary for French.

Combining MT with ST benefits mainly long queries, because the number of queries failing completely was higher for the short queries. In both cases, long and short, combination helped in high precision situations. Further combination with MRD brought no additional improvement.

For English, MT performed well, as expected. The similarity thesaurus performed poorly, because the English/German thesaurus had the least appropriate training data available (time shift and too little volume). This means that combination with ST and dictionary negatively affected the English component.

The Italian ST outperformed the thesauri for other languages. While not as good as the Italian MT translations, a full 13 queries performed at least 10% better based on the ST translations than when using the MT system. Combining MT with ST outperformed MT alone, especially in high precision situations. The Italian dictionary was the smallest, and consequently of no additional benefit.

Our work in Spanish was started specifically for CLEF. The performance of the German/Spanish thesaurus was adequate, given the little time available. Having training data from the same time period as the CLEF test data proved to be an advantage, even though the pool of training data was not as big as necessary to achieve the same quality as the French or Italian thesauri. The Spanish dictionary did not contribute positively to the overall performance of the German/Spanish component run.

We are pleased to see that our runs compare favorably when compared to other entries in CLEF. Table 4 shows an analysis of per-query performance compared to the median performance of all participants. All runs are above a "theoretical median": the average of the median average precision values. Especially the combination runs performed strongly and were among the best entries for CLEF 2001.

Table 4. Officially submitted runs (multilingual task) compared to median of all submitted runs (on individual query basis)

Run	Best	Above	Median	Below	Worst	Avg. Prec vs. Theor. Median
EIT01M1N	1	23	0	25	1	+0.0351
EIT01M2N	1	21	6	22	0	+0.0024
EIT01M3D	1	28	1	20	0	+0.0379
EIT01M3N	1	36	1	12	0	+0.0667

Monolingual: For German, the elaborate morpho-syntactic analysis of MPRO seems to bring a slight improvement over the more conventional Spider stemmer. However, the number of queries affected positively and negatively by over 10% in average precision is equal. We intend to conduct an in-depth analysis on the difference of the two approaches to stemming in the future. All German runs performed well when compared to the median performance.

For French, the broken run EIT01FFFN performs poorly. When fixed, performance improves substantially, and outperforms the MPRO analysis component.

Table 5. Average precision numbers for the monolingual experiments

Runs against Multilingual Collection	Average Precision
EIT01GGSN (German; TDN)	0.4285
EIT01GGLUN (German; TDN)	0.4408
EIT01GGLUD (German; TD)	0.4132
EIT01FFFN (French; TDN) (official/broken)	0.3848
EIT01FFFN (French; TDN) (unofficial/fixed)	0.4712
EIT01FFLUN (French; TDN)	0.4471

Taking into account the simplicity of the monolingual experiments, we consider the performance to be satisfactory. The German performed well, beating the median of all CLEF submissions consistently, while the fixed French run also topped median performance.

Table 6. Comparison of average precision numbers for individual queries

Comparison Avg. Prec. per Query	better; diff.>10%	better; diff.<10%	worse; diff.<10%	worse; diff.>10%
EIT01GGLUN vs. EIT01GGSN	8	19	14	8
EIT0FFLUN vs. EIT01FFFN (official/broken)	19	10	14	4
EIT01FFLUN vs. EIT01FFFN (unofficial/fixed)	4	14	20	8

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

Run	Best	Above	Median	Below	Worst	Avg. Prec vs. Theor. Median
EIT01GGSN	3	27	4	15	0	+0.0625
EIT01GGLUN	1	30	6	12	0	+0.0748
EIT01GGLUD	1	28	4	16	0	+0.0472
EIT01FFFN (brk)	1	13	5	28	2	-0.0787
EIT01FFFN (cor)	(2)	(23)	(4)	(20)	(0)	+0.0076
EIT01FFLUN	4	17	7	21	0	-0.0164

5 Summary

This year, we tried a combination of three different translation strategies: machine translation, similarity thesauri and machine-readable dictionaries. The results when using the thesauri were not as remarkable as last year, because of the lack of appropriate training data for some language combinations. Adding the similarity thesaurus to machine translation showed potential in the same areas we identified already last year, i.e. substantial benefit in recall and the high precision range for a number of queries.

The general-purpose dictionaries we introduced this year did not improve the performance of our experiments.

The monolingual experiments concentrated on an investigation into stemming behavior. We tested both the standard Spider stemmer, which is commercially motivated, and stemming based on the MPRO morpho-syntactic component. Based on our CLEF 2001 results, we plan to conduct an in-depth analysis.

6 Acknowledgements

We thank IAI Saarbrücken for the opportunity to use MPRO. Anne Göhring worked on the French and Spanish Spider stemmers, and Ron Caneel updated the similarity thesaurus components. Some of the tools for converting result lists were originally written by Min-Yen Kan for the Eurospider TREC-8 experiments.

References

1. Braschler, M., Schäuble, P.: Experiments with the Eurospider Retrieval System for CLEF 2000. In Proceedings of CLEF 2000. Lecture Notes in Computer Science, Springer Verlag, 2001.
2. 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.
3. Ripplinger, B.: The Use of NLP Techniques in CLIR. In Proceedings of CLEF 2000. Lecture Notes in Computer Science, Springer Verlag, 2001.
4. 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.
5. 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.