SPIDER Retrieval System at TREC7

Martin Braschler[**], Bojidar Mateev[*], Elke Mittendorf[*],
Peter Schäuble[*],Martin Wechsler[**]
[*] Swiss Federal Institute of Technology (ETH), CH-8092 Zürich
[**] Eurospider Information Technology, Schaffhauserstr. 18,
CH-8006 Zürich

Abstract

This year the Zurich team participated in two tracks: the automatic-adhoc track and the crosslingual track. For the adhoc task we focused on improving retrieval for short queries. We pursued two aims. First, we investigated weighting functions for short queries—explicitely without any kind of automatic query expansion. Second we developed rules that automatically decide for which queries automatic expansion works fine and for which it does not.

For the cross-language track, we approached the problem of retrieving documents from a multilingual document pool containing documents in all TREC CLIR languages. Our method uses individual runs for different language combinations, followed by merging their results into one final ranked list. We obtained good results without sophisticated machine translation or costly linguistic resources.

1 Introduction

For this year's adhoc runs we pursued ideas that were introduced by users with the hope to unite the weighting schemes based on probabilistic models with weighting schemes which take into account the user's subjective expectations. That is, we re-investigated the influence of feature-frequency weighting and the influence of proximity in particular for short queries.

Our runs for this year's cross-language track introduce a couple of new advancements compared to last year's submission: thesaurus filtering techniques and merging of results from multiple CLIR runs. The filtering techniques help us to automatically identify bad entries in the thesauri by comparing entries across different thesauri. The merging of results from individual crosslingual runs works by using document alignments computed for the TREC collections. The runs do not use any costly linguistic resources, and, in contrast to last year, we did not use any machine translation system.

2 General system description and reference method

Our indexing vocabulary Φ consists of single words $\varphi_i \in \Phi$ reduced by the Porter algorithm. A valid single word has a minimal length of three characters for the adhoc runs and two characters for the CLIR runs. Stopwords are removed.

The Retrieval Status Values (RSV) are obtained by

$$\mathrm{RSV}(q, d_j) = \sum_{\varphi_i \in \Phi(q,d_j)} a_{ij} \cdot b_i, \tag{1}$$

where $\Phi(q, d_j)$ denotes indexing features occurring in both the query q and the document d_j, a_{ij} denotes the weight of the document feature and b_i denotes the weight of the query feature.

As a **baseline** we used Lnu.ltn weighting as reported in [7] but applied minor changes. For adhoc experiments and final runs we used residual inverse document frequency (measure for the deviation from Poisson distribution) as suggested in [3]. The document length consisted of the number of (unique) stemmed and unstemmed features (this normalization has implementational reasons and unfortunately deteriorates the effectiveness).

Whe will refer in the next sections to **pseudo feedback** as choosing of features from the top ranked documents according to the Rocchio re-weighting function.

3 Adhoc Retrieval

Our focus in this year's adhoc track was on short queries. In last year's TREC we could see an impressive improvement on short queries based on the so-called "pseudo" feedback on top ranked documents [8]. Unfortunately users are often not satisfied with a "pseudo" feedback strategy. They are unhappy finding documents ranked on the top which do not contain the majority of the query features. This problem is even more severe when users are looking for a document containing proper names (person names, organization names etc.). Our approach is:

1. to investigate how good the retrieval functions can be **without pseudo feedback**, and to try to find a retrieval function which is closer to the user's usual expectations.

2. to **re-investigate the feature-frequency weighting** by looking for an alternative of the usual logarithmic feature-frequency weights, such as in the Lnu.ltn weighting scheme. The weighting should reward a document containing all query features.

We included **proximity** information and the coordination level match of features into the weighting scheme. In a second approach we designed rules for finding in which for which queries the proximity-based method works better than the baseline Lnu.ltn plus pseudo feedback. We tried to design the rules in such a way that the proximity method is chosen for queries where the user normally expects that all query features occur and for the rest of the queries the Lnu.ltn method with "pseudo" feedback is chosen.

3.1 Experiments and submitted adhoc runs

The baseline Lnu.ltn is described in Section 2. For the adhoc "pseudo" feedback run we chose the top 15 features (query expansion) from the seven top ranked documents. The parameters for the Rocchio formula were $\alpha = 7$ and $\beta = 3$. The slope for the Lnu.ltn was $s = 0.1$.

3.2 Method (ETHAR0)

The first method that we submitted uses the following strategy to rank the documents:

1. The documents, which have the maximal coordination-level match that is achieved by any document, are ranked before the documents with lower coordination-level match.

2. Documents with maximal coordination-level are ranked using a proximity weighting.

   (a) Determine the smallest window in the document that has the maximal coordination level. Documents with the narrowest window are ranked higher than documents with a larger window.

   (b) Documents ranked equally by (a) are further ranked according to how far is this window from the beginning of the document.

3. Documents with lower coordination-level are ranked according to the Lnu.ltn weighting scheme (disregarding the particular coordination level).

3.3 Method (ETHAC0)

We noticed that there are several queries for which pseudo feedback works significantly better than, e.g., the proximity and Lnu.ltn based method ETHAR0 and vice versa. Our intention is to find simple patterns which indicate when a short query weighted using a method based on proximity information outperforms a "pseudo" feedback method using a Lnu.ltn scheme. Our idea for finding such patterns is based on part-of-speech (POS) information. This information can be automatically obtained using a POS tagger (in our case the Brill tagger [2]). We compare the result list from method described in Section 3.2 with a result list obtained from baseline Lnu.ltn plus "pseudo" feedback. We focus on finding patterns for queries with noun phrases. The idea is that in such cases proximity might be a very useful information. Users are satisfied when—for a query consisting of noun phrases—they find documents which contain the same or only slightly different phrases.

The method decides for each query—based on a set of rules—which ranking method to use, either the ETHAR0 method or the Lnu.ltn plus "pseudo" feedback. The decision rules have been optimized on TREC6 queries (301–351). The rules are:

Choose method ETHAR0 if the query consists of

1. two rare nouns (the information whether or not a noun is rare is taken from the WordNet lexical database),

2. three nouns,

3. a nonstopword adjective followed by one or two nouns.

Choose method Lnu.ltn plus "pseudo" feedback otherwise.

According to these rules for 20 TREC6 short adhoc topics and for 25 TREC7 short adhoc topics the (coordination-level and proximity-based) method ETHAR0 is chosen.

On the one hand there is a large variety in the structure of noun phrases, on the other hand we have only a limited set of training queries (TREC6). It was obvious that under such conditions the rules for the choice method ETHAC0 might not be very robust. Moreover, the POS tagger does not perform as well on title-like short queries as it does on complete sentences. Despite of the possible lack of robustness we gave it a try.

3.4 Method ETHAB0

In this method the retrieval status values are determined in a first pass by Robertson–Sparck Jones (RSJ) [5] weighting, i.e., the ranking of documents is based primarily on inverse document frequency weights. After determining classes of documents that have the same RSJ retrieval status value. The documents within one RSJ-class are fine-ranked according to the Lnu.ltn weighting. Note that the RSJ-classes of documents are large because the queries are short.

3.5 Experiments

We report on experiments with the methods described above (Lnu.ltn baseline, Lnu.ltn plus pseudo-feedback, ETHAR0, ETHAC0, ETHAB0). A result list was produced for each of the methods for both TREC6 (301–350) and TREC7 (351–400).

Table 1 shows the average precision over all queries for the methods on both topic sets. In addition, Figure 1 and Figure 2 provide boxplots that visualize the distribution of the average precision per query.

Runs/Queries	TREC6 (301-350)	TREC7 (351-400)
Lnu.ltn	0.2202	0.1631
Lnu.ltn with pseudo feedback	0.2366	0.1853
ETHAR0	0.2251	0.1597
ETHAC0	0.2473	0.1645
ETHAB0	0.2306	0.1601

Table 1: Average precision for all queries

3.6 Discussion of Results

Unfortunately the results on TREC-6 and TREC-7 data are not consistent. The two methods that emphasise feature occurrence before feature frequencies, ETHAR0 and ETHAB0, yield a higher average precision on TREC-6 than standard Lnu.ltn, where high feature frequencies can overweigh pure occurrence of a feature. On TREC-7 the results are vice versa, standard Lnu.ltn outperforms our two new weighting schemes.

The tagger-based query classification method (used for ETHAC0), which is able to distinguish those TREC-6 queries for which pseudo feedback works better than the proximity-based method ETHAR0 and vice versa, fails for TREC-7 queries. The submitted method ETHAC0 is worse than the pseudo-feedback method, on which it is based.

In summary, in this year's TREC we have taken a chance and tried rather nonorthodox methods to emphasize occurrence over frequency, to emphasize proximity of features and to decide whether or not to use pseudo feedback. We know that the design of our methods has to be more robust.

4 Cross-Language Information Retrieval

Our participation this year consists of three runs, all using the German topics, and retrieving documents from the full pool of documents in all four languages (German, French, Italian

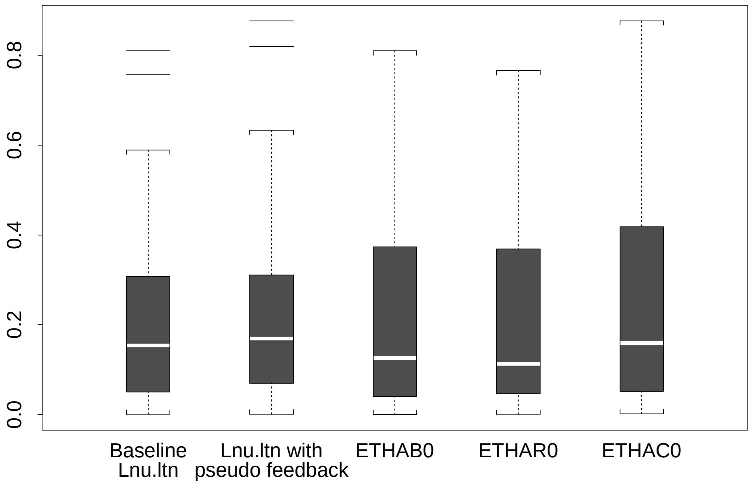

Figure 1: Experiments on TREC6 (topics 301–350), showing the distribution of average precision per query (y-axis).

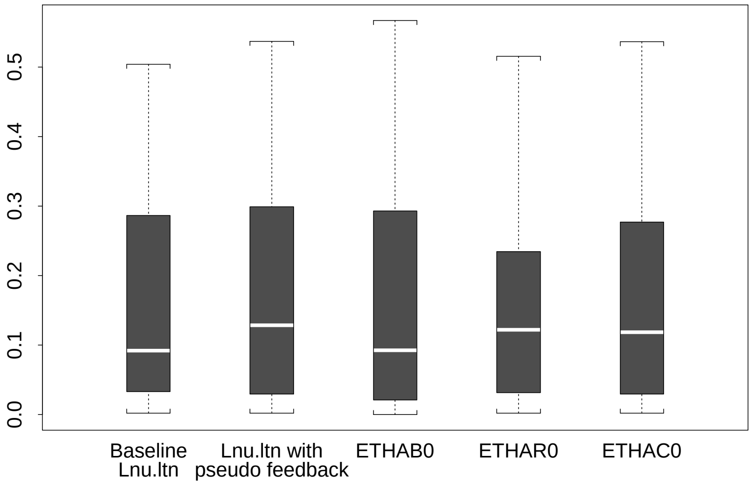

Figure 2: Experiments on TREC7 (topics 351–400), showing the distribution of average precision per query (y-axis).

and English). The runs were produced by doing individual runs for language pairs (e.g. German to French), and merging the results to form the final ranked list. Our focus this year was therefore both on improving such individual runs and on solving the problem of merging their results.

Merging individual ranked lists gives us the flexibility to use all the resources we had available for specific language pairs, instead of being restricted to a least common denominator across the languages. Consequently, not all the individual runs were produced in the same way.

4.1 Experiments and submitted runs

The weighting scheme for all runs was set to Lnu.ltn, as described in Section 2. The slope was set to $s = 0.24$. The parameters for the Rocchio formula were $\alpha = 1$ and $\beta = 0.81$, with the 20 best terms taken from the top 10 documents for pseudo-feedback.

4.2 Monolingual run

- German→German
  The German monolingual run was produced by doing an initial retrieval run using all configurations as described above. We then used pseudo-feedback [8] to expand the query by terms coming from the top ranked documents. For stemming, the German stemmer that is included with NIST ZPRISE system was adapted.

4.3 Crosslingual runs

For our crosslingual runs, we used similarity thesauri (for French, Italian) and a wordlist (a simplistic bilingual dictionary, for English). A similarity thesaurus is a data structure that is automatically derived from appropriate training data. While originally developed for monolingual query expansion, it has been successfully applied to the problem of crosslingual retrieval by using multilingual documents as training data. These documents are obtained through a document alignment process [1] that finds pairs of similar documents in different languages and creates a single multilingual document by joining them. We applied this alignment process to the TREC collections for all three language combinations described below.

The thesaurus itself is built by using co-occurrence statistics from the training data to determine the similarity of every pair of terms in the collection. The most similar pairs are stored to disk. The similarity can be calculated by exchanging the role of documents and terms in conventional weighting schemes. This way, the similarity of terms is determined through the sets of documents that indexes them. A similarity thesaurus can cover very large vocabularies; it usually however also contains various low quality entries. More details can be found in [6].

We also used a German→English wordlist we assembled from various free sources on the Internet as a simplistic form of a bilingual dictionary. While the resulting list is rather large (141,240 entries, or 85,931 unique "head" entries), it contains many questionable or even wrong entries, since we were not concerned with any clean-up of the data. We believe that through coupling the wordlist with our corpus-based alignment techniques the resulting retrieval process gets robust with respect to such "noise".

- German→French

  The German→French crosslingual run uses a similarity thesaurus built on the SDA data. Compared to the experiments described in [4], we had more data available from SDA than last year, also covering time periods not present in the SDA texts used for this year's track. Our complete document pool consisted of all German and French SDA news texts from 1988 to March of 1998 (1988-90 of this pool comes from the official track data). We exploited this by using thesaurus merging techniques to filter the entries in individual thesauri built over different time periods. While this allows us to generalize the resulting thesaurus, this also means that the fact that for the CLIR track the thesaurus can be built on the collection itself is not exploited as much as for last year's experiments. We think, however, that using a generalized thesaurus gives a more realistic scenario.

  The similarity thesauri are employed in much the same way as outlined in [1]. A pseudo-translation is produced using the similarity thesaurus, and is then combined with terms coming from a selection process on the top aligned documents, using the same document alignments as computed for the creation of the thesaurus. The initial monolingual run used to determine the aligned documents was produced as outlined above.

- German→Italian

  The German→Italian run is very similar to the German→French run. Again, we had additional SDA data available (the complete pool consisting of all Italian SDA news texts from end of 1989 to March of 1998), which allowed us to build several thesauri and employ our thesaurus filtering techniques. The pseudo-translation coming from the thesaurus is also combined with terms from the top aligned documents.

- German→English

  Because we did not think that a German→English similarity thesaurus with satisfactory quality can be derived from the differently focused international AP news and the national SDA news, we used the Internet wordlist for German→English query translation. This also nicely demonstrates the ability to use different resources for different language pairs in our approach. Query translation itself was done using a simplistic word-by-word dictionary lookup procedure, but was again complemented with terms coming from the top aligned documents. This combination helps offset many of the sense ambiguity problems associated with simple dictionary lookup.

4.4 Merging

The ranked lists of the four individual runs were then merged using the technique described in [1]. By again making use of the document alignments from the individual collections of the track data, we produced tables giving the relations between scores of the individual runs. It is then possible to map these scores to a common range using linear regression. After rescaling the scores, merging is done by simply sorting a joined version of both ranked lists. Repeating this for every language, we ultimately produced the final ranked lists that were submitted.

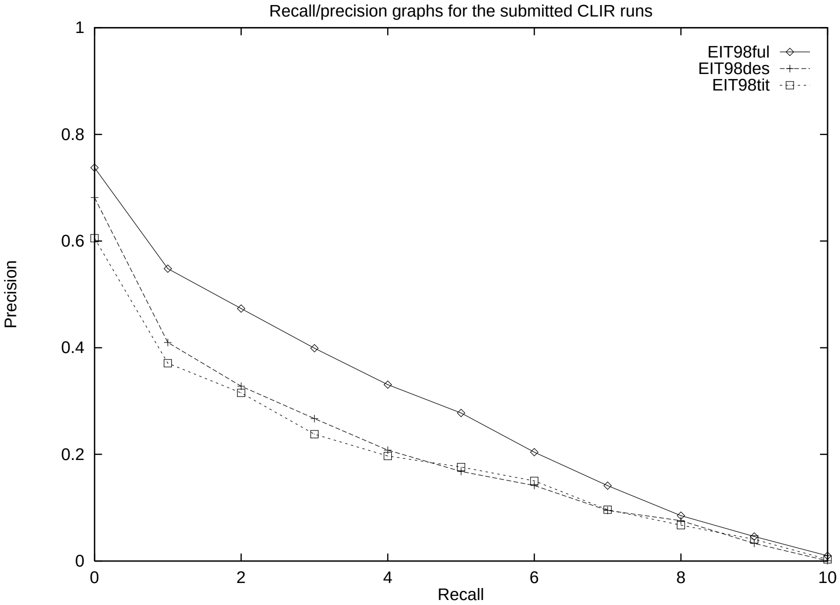

Figure 3: Comparison of the submitted cross-language runs.

4.5 Results

We submitted three runs for TREC-7, all using the German topics and the full document pool. The difference in the runs was in the topic fields used: EIT98ful used all topic fields ("full topics"), EIT98des used the title and description fields, and EIT98tit used the title field only.

We feel the results (Table 2) are very encouraging given the fact that we used no costly linguistic resources to produce the runs. The similarity thesauri were derived completely automatically from training data, whereas the wordlist was taken from free Internet sources. As mentioned, however, the presented approach is flexible towards incorporating further resources should they be available.

Clearly, the run using the full topics (EIT98ful) is outperforming the other two submissions (see Figure 3. We believe that long queries are beneficial to corpus-based techniques like the similarity thesaurus that have a broad vocabulary coverage, but also contain bad entries that may have an adverse effect if little input is available for the translation process. The fact that the full queries perform so much better than the other two query sets also shows that our corpus-based techniques suffer less from a word ambiguity problem than purely dictionary-based approaches. The latter approaches will produce very long output if the query length increases; this due to every word potentially having more than one translation. Such very long queries are not likely to perform well. The similarity thesaurus however allows to translate queries by using terms similar to the overall query concept instead of individual words, thus even allowing to perform query reduction.

Work is clearly needed for the case of shorter queries, which, as mentioned in the section about adhoc retrieval, is the usual case for a lot of applications. The future direction of work in this area will likely be the incorporation of more linguistic resources.

Merging seems to have done well. Problems we have encountered include when there are only a few score pairs available for the linear regression because only few documents of the result lists have been aligned. This can lead to an instability of the process. The method also seems to tend to prefer one run over the other for top ranked documents, giving a somewhat unbalanced mix at the top of the merged result list. We intend to look further into this effect.

run	above median	on median	below median	avg. prec
EIT98ful	17	0	11	0.2767
EIT98des	8	2	18	0.1962
EIT98tit	7	2	19	0.1841

Table 2: Results of the cross-language runs.

References

[1] 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.

[2] E. Brill. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging. *Computational Linguistics*, 1995.

[3] K. W. Church. One term or two? In *ACM SIGIR Conference on R&D in Information Retrieval*, pages 310–318, 1995.

[4] B. Mateev, E. Munteanu, P. Sheridan, M. Wechsler, and P. Schäuble. ETH TREC-6: Routing, Chinese, Cross-Language and Spoken Document Retrieval. In *Proceedings of the sixth Text REtrieval Conference (TREC-6)*, 1997.

[5] S. E. Robertson. The Probability Ranking Principle in IR. *Journal of Documentation*, 33(4):294–304, 1977.

[6] P. Sheridan, M. Braschler, and P. Schäuble. Cross-Language Information Retrieval in a Multilingual Legal Domain. In *The First European Conference on Research and Advanced Technology for Digital Libraries*, pages 253–268, 1997.

[7] A. Singhal, C. Buckley, and M. Mitra. Pivoted Document Length Normalization . In *ACM SIGIR Conference on R&D in Information Retrieval*, pages 21–29, 1996.

[8] E. M. Voorhees and D. Harman. Overview of the Sixth Text REtrieval Conference (TREC-6). In *Proceedings of the sixth Text REtrieval Conference (TREC-6)*, 1997.