Thomson Legal and Regulatory experiments for CLEF 2002

Isabelle Moulinier and Hugo Molina-Salgado
Thomson Legal and Regulatory
Research and Development Group
610 Opperman Drive, Eagan, MN 55123, USA
{Isabelle.Moulinier,Hugo.Salgado}@westgroup.com

Abstract

Thomson Legal and Regulatory participated in the monolingual, the bilingual and the multilingual tracks. Our monolingual runs added Swedish to the languages we had submitted in previous participations. Our bilingual and multilingual efforts used English as the query language. We experimented with dictionaries and similarity thesauri for the bilingual task, while we used machine translations in our multi-lingual runs. Our various merging strategies had limited success compared to a simple round robin.

1 Introduction

For CLEF-2002, Thomson Legal and Regulatory (TLR) participated in monolingual, bilingual, and multilingual retrieval. Our monolingual experiments benefited from previous efforts. We added Swedish to the languages we submitted last year (Dutch, French, German, Italian and Spanish). In addition, we tried to improve our Italian runs by refining language resources. Our bilingual runs were from English to either French or Spanish. We translated query concepts using a combination of similarity thesauri and machine-readable dictionaries. Translated queries were structured to take into account multiple translations as well as translations of word pairs rather than words. In our multilingual experiments, we used a machine translation system rather than our bilingual approach. We mostly focused on merging strategies, using CORI, normalization or round-robin.

We give some background to our experiments in Section 2. Sections 3, 4 and 5 respectively present our monolingual, bilingual, and multilingual experiments.

2 Background

2.1 Previous research

Our participation at CLEF-2002 benefits from our earlier work, as well as from the work of others. Our bilingual effort relies on similarity thesauri for translating query terms from English to French or Spanish. In addition to translating words [6], we also translate word pairs which loosely capture noun and verb phrases. This differs from our approach last year when we generated word bigrams rather than pairs [3]. In addition, we follow Pirkola's approach for handling multiple translations. By taking advantage of query structures available in INQUERY, Pirkola [4] has shown that grouping translations for a given term is a better technique than allowing all translations to contribute equally. This has been developed further by Sperer and Oard [7].

One of the main issues in multilingual retrieval remains collection merging. In our experiments, we use simple merging techniques like round robin, normalized scores, as well as a variant of the CORI algorithm [1]. This is similar to Savoy's work at CLEF-2001 [5] and others.

2.2 The WIN system

The WIN system is a full-text natural language search engine, and corresponds to TLR/West Group's implementation of the inference network retrieval model. While based on the same retrieval model as the INQUERY system

[2], WIN has evolved separately and focused on the retrieval of legal material in large collections in a commercial environment that supports both Boolean and natural language searches [8].

In addition, WIN has shifted from supporting mostly English content to supporting a large number of Western-European languages as well. This was performed by localizing tokenization rules (mostly for French and Italian) and adopting morphological stemming. Stemming of non-English terms is performed using a third-party toolkit, the LinguistX platform commercialized by Inxight. A variant of the Porter stemmer is used for English.

2.2.1 Document Scoring

WIN supports various strategies for computing term beliefs and scoring documents. We used a standard tf-idf for computing term beliefs in all our runs. The document is scored by combining term beliefs using a different rule for each query operator [2]. The final document score is an average of the document score as a whole and the score of the best portion. The best portion is dynamically computed based on query term occurrences.

2.2.2 Query formulation

Query formulation identifies concepts in natural language text, and imposes a structure on these queries. In many cases, each term represents a concept, and a flat structure gives the same weight to all concepts. The processing of English queries eliminates stopwords and other noise phrases (such as "Find cases about", or "Relevant documents will include"), identifies (legal) phrases based on a phrase dictionary and detects common misspellings.

In the experiments reported below, we use our standard English stopword and noise phrase lists, but do not identify phrases or misspellings. We have expanded the English noise phrase list with noise phrases extracted from queries used in previous years. Our German, French, Spanish, and Dutch runs use the same stopword lists as last year, but noise phrase patterns have been updated to cover query sets from CLEF-2001. Our Italian stopword and noise phrase list was validated by a native speaker, while our Swedish resources were extracted from the web and from available query sets.

Concept identification depends on text segmentation. In our experiments, we follow two main definitions for a concept: a concept is an indexing unit (typically a word) or a concept is a construct of indexing units. Constructs are expressed in terms of operators (average, proximity, synonym, etc.) and indexing units. For instance, we use a construct when a term has multiple translations, or when we identify word pairs.

3 Monolingual experiments

Our approach for monolingual runs is similar to last year's. We have revised the Italian stopword and noise phrase lists with the help of a native speaker. Our stemming procedure, although still based on the LinguistX toolkit, has been altered slightly to limit the occurrence of multiple stems.

German, Dutch, and Swedish are all compounding languages. However, the LinguistX platform does not support compound breaking for Swedish. We thus index and search using compound parts only German and Dutch content. Swedish is treated as a non-compounding language.

For all languages, we allow the stemmer to generate several stems for each term, as we do not rely on part-of-speech tagging for disambiguation. Multiple stems were grouped under a single concept in the structure query.

Results from our official runs are reported in Table 1. All runs used the title and description fields from the topics. Our results are comparable to those of previous years. Introducing revised stopword and noise phrase lists for Italian allows us to achieve good performance.

While most languages achieve an average precision in the same range (between 0.4 and 0.5), the figures for Swedish are much lower. We suspect that not breaking compounds may be the main cause, since previous work with German and Dutch has shown that retrieval performance was enhanced by compound breaking.

4 Bilingual experiments

Our bilingual runs were from English queries to Spanish and French collections. As in our previous work [3], we used a combination of similarity thesauri and machine-readable dictionaries. The machine-readable dictionaries were downloaded from the Internet (freedict).

We implemented a variant of the similarity thesaurus approach described in [6] for multilingual retrieval. We constructed two similarity thesauri: a word thesaurus and a word pair thesaurus. Both similarity thesauri were

Run ID	Lang.	Avg. Prec.	R-Prec.	Above Median	Median	Below Median
tlrde	German	0.4221	0.4294	21	6	23
tlres	Spanish	0.4993	0.4816	31	3	16
tlrfr	French	0.4232	0.4134	17	8	25
tlrit	Italian	0.4159	0.4072	24	7	18
tlrnl	Dutch	0.4141	0.4211	27	3	20
tlrsv	Swedish	0.2439	0.2700	17	11	21

Table 1. Summary of all monolingual experiments using the title and description fields. Comparison to the median is expressed in the number of queries above, equal, and below.

trained on a collection merging the UN parallel text corpus produced by the Linguistic Data Consortium, and an European Union (E.U.) parallel corpus that we have at TLR.

Using a part-of-speech tagger, we restricted the set of words to nouns, verbs, adjectives and adverbs. Word pairs were generated using sliding windows centered only on nouns, and components in pairs were ordered alphabetically. Terms, words, or pairs, were considered as translations when their similarity was above a predefined threshold. This threshold was chosen as the best configuration on CLEF-2001 data.

While we identified noise phrase patterns in our official runs, stopwords were expected to have a different part-of-speech (like auxiliary, prepositions, etc). We later added a stopword list in conjunction with noise phrase patterns.

Table 2 reports our official runs. The translation resources for our official runs were a combination of the word similarity thesaurus and the dictionary.

Run ID	Lang.	Avg. Prec.	R-Prec.	Above Median	Median	Below Median
tlren2es	English/Spanish	0.2873	0.2857	12	4	34
tlren2fr	English/French	0.3198	0.3440	13	3	34

Table 2. Summary of our official bilingual experiments using the title and description fields. The median was computed from all submitted runs.

Table 3 summarizes our unofficial runs. These runs used an explicit stopword list, instead of relying on part-of-speech tags. We also translated word pairs after we completed training the word pairs similarity thesauri. The last runs use automatic translation and are part of our multilingual run.

Run Description	Lang.	Avg. Prec.	R-Prec.
Stopwords	English/Spanish	0.3123	0.3047
Stopwords + Pairs	English/Spanish	0.3118	0.3038
Machine Translation	English/Spanish	0.3391	0.3414
Stopwords	English/French	0.3263	0.3474
Stopwords + Pairs	English/French	0.3257	0.3605
Machine Translation	English/French	0.3513	0.3543

Table 3. Summary of our unofficial bilingual runs using the title and description fields. Stopwords is the same as the official runs but we use an explicit stopword list. Pairs correspond to the combination of the translation from all three sources. The Machine Translation runs use Babelfish.

A comparison of Tables 2 and 3 shows that using an explicit list of stopwords helps enhance the average precision. We have identified inaccuracies in part-of-speech tagging as one of the main reasons. Inaccuracies are often caused by inadequate context, or by the lack of a specific tag in one of the languages, e.g. auxiliary versus verb.

Our approach using similarity thesauri has some shortcomings in comparison with the machine translation approach. In particular, it is very dependent on the parallel corpus used for training. In our experiments, using E.U. material lead to some E.U.-oriented translations. For instance, European is translated into the French terms européen and communauté, and the Spanish terms europeo, comunidad and constitutivo. One way of addressing that issue may be to filter out corpus-specific terminology.

Unlike our results with bigrams at CLEF-2001 [3], translating word pairs provides little advantage over translating individual words. One plausible hypothesis is that the window used to generate word pairs (we used a window of 9 centered on a noun) and the query structure are not compatible (we used a phrase node, i.e. a proximity of 3).

5 Multilingual experiments

During our multilingual experiments, we translated queries only. We used the indices generated for the monolingual runs for German, French, English, Italian and Spanish. Queries were translated from English to the other languages using Babelfish.

Our main focus was merging, although we have not been very successful so far. We tried a variety of merging approaches:

- round robin, i.e. a rank-based approach that alternates documents from each collection. In our setting, documents with identical score were given the same rank.
- raw score, which may or may not be comparable across collections
- CORI, where the collection score is estimated by the maximum score a (translated) query can achieve on that collection, not the original collection score in Callan et al [1].

$$cori\_score = score\_within\_collection_i * (1 + nb\_lang * \frac{collection\_score_i - avg\_collection\_score}{avg\_collection\_score})$$

- normalized score, where the $maximum\_score\_within\_collection_i$ is the score of the document at rank 1.

$$norm\_score = \frac{score\_within\_collection_i - 0.4}{maximum\_score\_within\_collection_i - 0.4}$$

  0.4 represents the minimum score any document can achieve in the belief network retrieval model.

- collection-weighted normalized score, where $collection\_score_i$ is the same as in the CORI approach above, and $maximum\_collection\_score$ is the maximum of these scores.

$$weighted\_norm\_score = norm\_score * \frac{collection\_score_i}{maximum\_collection\_score}$$

Our official run tlren2multi used round robin. Table 4 reports results from the different merging approaches. As reported too often, we found it hard to outperform the round robin approach. Our collection-weighted normalized score is the only merging approach to perform better but the difference is not significant. Our results with the CORI merging strategy are comparable to those obtained by Savoy [5]. It is possible that the CORI algorithm is impacted by our choice of $collection\_score_i$. More analysis is required to assess the difference between the original CORI and our version.

Run ID	Avg. Prec.	R-Prec.	Above Median	Median	Below Median
tlren2multi (round robin)	0.2049	0.2803	17	4	29
raw score	0.1883	0.2521			
$cori\_score$	0.1023	0.1489			
$norm\_score$	0.1827	0.2496			
$weighted\_norm\_score$	0.2160	0.2794			

Table 4. Summary of our multilingual experiments using the title and description fields. English was the query language. The median was computed from all submitted runs.

There are two issues with multilingual retrieval, the quality of the individual runs and the effectiveness of the merging strategy. The quality of the individual runs can easily be assessed by comparing their performance to the performance of monolingual runs. As can be seen in Table 5, using translated queries leads to an average degradation of 25% in performance (performance is measured in terms of average precision).

How to quantify the effectiveness of merging strategies remains an open issue. We can observe the following properties in an attempt to measure the effectiveness of merging. In Table 6, we observe that merging better

Collection language	Monolingual	Translated from English
German	0.4221	0.2849 (-32.5%)
French	0.4232	0.3513 (-17.0%)
Spanish	0.4993	0.3391 (-32.1%)
Italian	0.4159	0.3212 (-22.8%)

Table 5. The impact of translation in multilingual retrieval. The percentages reflect differences in average precision when we compare retrieval using an English query with retrieval using a query in the collection language.

individual runs (the monolingual column vs. the translated column) leads to better performance. We can also compare the average of the individual run performances with the performance of the multilingual runs, and find that the average of individual runs is higher that any multilingual run. These observations tend to indicate that merging also deteriorates the effectiveness of multilingual runs, but do not tell us how much so.

Merging strategy	Monolingual	Translated
round robin	0.2948	0.2049 (-30.5%)
raw score	0.3230	0.1883 (-41.7%)
$cori\_score$	0.1354	0.1023 (-24.5%)
$norm\_score$	0.2663	0.1827 (-31.4%)
$weighted\_norm\_score$	0.3042	0.2160 (-29.0%)
Average of individual runs	0.4007	0.3077 (-23.2%)

Table 6. Average precision of merging strategies. The monolingual column uses results from our monolingual runs (English, German, French, Spanish and Italian). The translated column refers to English queries translated to the collection language. The row Average of individual runs does not rely on merging.

The poor performance of our English monolingual run (around 25% average precision) [1] had a noticeable impact on multilingual runs. We found that round robin, $cori\_score$ and $weighted\_norm\_score$ were not affected as much as raw score and $norm\_score$ by the English run. We expected round robin to be more sensitive to English documents, since one fifth of the documents are English. In effect, our modified version of round robin limited that effect for 40 queries, and aggravated it for 10 others. As could be expected, raw score was mislead by the higher score of English documents for a large number of queries. $norm\_score$ suffers a similar problem: it is mislead when document scores are close to the highest document score in the retrieved list.

6 Conclusion

Our participation at CLEF-2002 has mixed results. On the one hand, we consider that our monolingual runs successful, even though we intend to evaluate how much improvement can be achieved by relevance feedback. On the other hand, our bilingual and multilingual runs did not lead to the expected results. For instance, we did not find any evidence that translating word pairs was helpful in our bilingual runs. We also encountered an over-fitting problem when training similarity thesauri on the E.U. corpus. Finally, we are still in the process of investigating alternative merging algorithms, since our current approach has shown limited success.

References

[1] J. P. Callan, Z. Lu, and W. B. Croft. Searching distributed collections with inference networks. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 21–28, Seattle, WA, 1995.

[2] W. B. Croft, J. Callan, and J. Broglio. The inquery retrieval system. In *Proceedings of the 3rd International Conference on Database and Expert Systems Applications*, Spain, 1992.

[1] We suspect that this run encountered problems. We have not yet identified what the issues are.

[3] H. Molina-Salgado, I. Moulinier, M. Knutson, E. Lund, and K. Sekhon. Thomson legal and regulatory at clef 2001: monolingual and bilingual experiments. In *Workshop Notes for the CLEF 2001 Workshop*, Darmstadt, Germany, 2001.

[4] A. Pirkola. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In *Proceedings of the 21th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 55–63, Melbourne, Australia, 1998.

[5] J. Savoy. Report on clef-2001 experiments. In *Workshop Notes for the CLEF 2001 Workshop*, Darmstadt, Germany, 2001.

[6] P. Sheridan, M. Braschler, and P. Schuble. Cross-lingual 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, Pisa, Italy, 1997.

[7] R. Sperer and D. W. Oard. Structured translation for cross-language information retrieval. In *Proceedings of the 23th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 120–127, Athens, Greece, 2000.

[8] H. Turtle. Natural language vs. boolean query evaluation: a comparison of retrieval performance. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 212–220, Dublin, Ireland, 1994.