Thomson Legal and Regulatory at CLEF 2001: monolingual and bilingual experiments

Hugo Molina-Salgado, Isabelle Moulinier, Mark Knutson, Elizabeth Lund, Kirat Sekhon
TLR
610 Opperman Drive
Eagan, MN 55123
USA
Isabelle.Moulinier@westgroup.com

Thomson Legal and Regulatory participated in the monolingual track for all five languages and in the bilingual track with Spanish-English runs. Our monolingual runs for Dutch, Spanish and Italian use settings and rules derived from our runs in French and German last year. Our bilingual runs compared merging strategies for query translation resources.

Introduction

Thomson Legal and Regulatory (TLR) participated in CLEF-2001 with two goals: reuse of rules and settings inside a family of languages for monolingual retrieval, and start our effort on bilingual retrieval.

In our monolingual runs, we considered Dutch and German as being one family of languages, while French, Spanish and Italian formed another. We used the parameters we derived from our runs at CLEF-2000 for German and French for each language in their respective family. In addition, we investigated the use of phrases for French and Spanish document retrieval.

Our first attempt at the bilingual track was from Spanish queries to English documents. In that task, we experimented with combining various resources for query translation. Our submitted runs used similarity thesauri and a machine-readable dictionary to translate a Spanish query into a single English query. We also compared our official runs with the merging of individual runs, one per translation resource.

In this paper, we briefly present our search engine and the settings common to all experiments. Then, we discuss our bilingual effort. Finally, we describe our participation in the monolingual track.

General System Description

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 [BCC93], 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 [Tur94].

WIN has also been modified to support non-English document retrieval. This included localization of tokenization rules (for instance, handling elision for French and Italian) and stemming. Stemming of non-English terms is performed using a third-party toolkit, the LinguistX platform® commercialized by Inxight[1]. A variant of the Porter stemmer is used for English.

The WIN engine supports various strategies for computing term beliefs and document scores. We used a standard tf-idf for computing term beliefs in all our runs. Among all the variants, we retained document scoring and portion scoring. Document scoring assigns a score to the document as a whole. This was used in our bilingual runs. Portion scoring finds the best dynamic portion in a document and combines the score of the best portion to the score of the whole document. Portion scoring can be considered as an approximation of paragraph scoring (we used this setting last year for French), when documents have no paragraph. We used portion scoring in our monolingual runs.

A WIN query consists of concepts extracted from natural language text. Normal WIN query processing eliminates stopwords and noise phrases (or introductory phrases), recognizes phrases or other important concepts for special handling, and detects misspellings. Many of the concepts ordinarily recognized by WIN are specific to English documents, in particular within the legal domain. Query processing in WIN

[1] Information can be found at http://www.inxight.com/products_sp/linguistx/index.html.

usually relies on various resources: a stopword list, a list of noise phrases (“Find cases about…”, “A relevant document describes…”) , a dictionary of (legal) phrases, and a list of common misspelled terms.

We used stopword and noise phrases lists for all languages, while for French and monolingual Spanish, we also used a phrase dictionary. We used our French and German stopword lists from last year, the Dutch list given on the CLEF homepage, and compiled Spanish and Italian stopword lists from various sources on the Web. For all languages, we extracted introductory phrases from the query sets of previous CLEF and TREC conferences. As we had no Italian speaker in our team, our introductory list in Italian is very limited and simple.

Finally, we submitted two sets of runs: runs including only the title and description fields from the CLEF topic, and runs including the whole topic. The former runs are labeled with ‘td’ and doubled weighted the title fields. The latter are labeled with ‘tdn’ and used a weight of 4 for the title field, 2 for the description field, and 1 for the narrative.

Spanish-English bilingual retrieval experiments and results

In our bilingual runs, we concentrated on query translation and more specifically the combination of various translation resources. We used three main resources, a machine-readable dictionary (MRD) that we downloaded from http://www.freedict.com and two different similarity thesauri. Coverage of these resources is reported in Table 1.

We implemented a variant of the similarity thesaurus approach described in [PBS97] for multilingual retrieval. We used a parallel corpus, the UN parallel text corpus produced by the Linguistic Data Consortium. We generated two different thesauri: a unigram thesaurus and a bigram thesaurus. Our intent with the bigram thesaurus was to capture some phrase translation. We limited the number of bigrams by constraining bigrams to not contain stopwords, and by frequency thresholding. We used at most 15 translations from each thesaurus, and also used a threshold on the similarity to filter out translations that we thought would not be helpful. This threshold was determined on training data from CLEF 2000. We used all translations from the MRD. In all cases, multiple translations of the same Spanish term were grouped as the same concept given a translation source.

Table 1: Coverage of the translation resources used

	Dictionary	Unigram Thesaurus	Bigram Thesaurus
Spanish	19,466 terms	33,674 terms	42,081 bigrams

We investigated two main approaches to combine our translation resources: *a priori* merging, i.e. combining translations during query construction, and *a posteriori* merging, i.e. merging runs produced by queries translated from a single resource. For *a posteriori* merging, we used a score-based and a rank-based technique to generate the merged score. The score based technique relies on a feature of the WIN engine. WIN computes the best score a document can achieve for a given query. We used that maximum score to normalize individual runs. Normalized runs are merged in a straightforward manner. The rank based technique is also fairly simple. The score in the merged result list is a function of the ranks in the original lists. Here, we report experiments using the sum of the logarithms of the document rank in each run.

Our official runs relied on the *a priori* approach, combining translations during construction. Runs tlres2entdw and tlres2entdnw[2] combined only the unigram thesaurus to the dictionary, while runs tlres2entdb and tlres2entdb combined both thesauri with the dictionary. In

Table 2, we also report *a posteriori* merging using the following conversion: b refers to the bigram thesaurus, u to the unigram thesaurus and d to the dictionary. The different scoring methods are indicated by a s for score and r for rank. Thus, the run labeled b+u+d_r_tdn refers to combining both thesauri and the MRD using rank-based merging. In all cases, terms not found in any resource were left intact in the query.

[2] The only difference between runs tlrdetdw and tlrdetdnw and between runs tlres2entdb and tlres2entdnb is whether the narrative field is used or not.

Table 2 Results from our bilingual Spanish to English experiments

			Performance of individual queries				
Run	Avg. Prec.	R-Prec.	Best	Above	Median	Below	Worst
Official runs							
Tlres2entdw	0.3846	0.3914	5	26	1	13	2
Tlres2entdb	0.3909	0.3914	6	26	1	12	2
Tlres2entdnw	0.4264	0.4234	6	29	1	9	2
Tlres2entdnb	0.4338	0.4291	6	30	1	8	2
Unofficial runs Comparison to the median is indicative. If these runs had been included, the median will be different.							
u+d_s_td	0.3164	0.3198	4	27	0	15	1
b+u+d_s_td	0.2921	0.3008	4	25	0	17	1
u+d_s_tdn	0.3875	0.3918	4	30	0	12	1
b+u+d_s_tdn	0.3813	0.3833	4	30	0	12	1
u+d_r_td	0.3210	0.3107	3	29	0	14	1
b+u+d_r_td	0.2862	0.2841	3	25	0	18	1
u+d_r_tdn	0.3636	0.3656	4	29	0	13	1
b+u+d_r_tdn	0.3197	0.3138	3	25	0	18	1

Figure 1 summarizes the impact of combining resources, as it shows runs using individual resources as well as our official runs using all fields in the CLEF topics. Run b_tdn used only the bigram thesaurus, run u_tdn the unigram thesaurus, while run d_tdn used the MRD. We did not report runs using only the title and description fields from the CLEF topics, as they showed the same behavior.

Figure 1 Summary of our Spanish to English bilingual runs

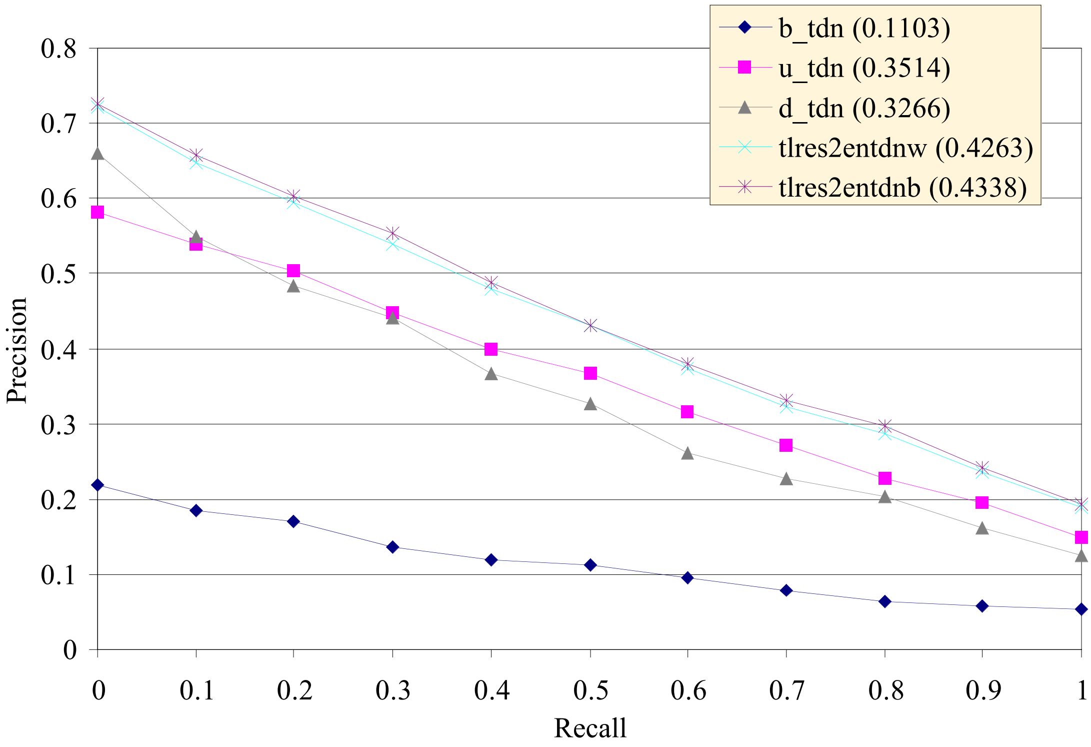

Table 2 shows that our official runs performed well in this year's evaluation. The results show a slight advantage in using the bigram thesaurus in combination with the other resources. However, the bigram thesaurus on its own shows very poor performance. There are two main reasons for that behavior. First, the coverage of the bigram thesaurus was poor: only a limited number of bigrams were found in the queries, some queries having no bigram. Second, English bigrams sometimes were not present in the retrieval collection, as thesauri were constructed on a non-related corpus. This resulted in some queries returning very few or no documents.

The poor performance of the bigram thesaurus also impacted our *a posteriori* merging. Indeed, runs using the bigram thesaurus show lower average precision than runs only using the unigram thesaurus and the MRD. The score-based technique performed better than the rank-based technique. The rank-based technique reported here is based on the product of the ranks. As a result, documents with very different ranks in the individual runs are penalized.

Finally, *a priori* merging performed better than both techniques for *a posteriori* merging. One reason is that the number of non-translated terms diminished when resources are combined *a priori*. Further analysis is needed to better understand the difference in behaviors.

Monolingual retrieval experiments and results

In our monolingual runs, we considered two aspects: families of languages, and the use of a phrase dictionary. We used the same rules for Dutch and German on the one hand, and French, Spanish and Italian on the other. In addition, we introduce a phrase dictionary in some of our Spanish and French runs.

German and Dutch were considered as compounding languages. Using the LinguistX morphological analyzer allowed us to detect compound words and break them at indexing and search time. We used a structured query and loose noun phrases to represent compounds.

For French, Spanish and Italian, we allowed the LinguistX morphological analyzer to generate several stems (we did not disambiguate using part-of-speech tags). Multiple stems were grouped as a single concept (using a OR/SYN or a SUM node for instance) in the structured query. For French and Spanish, we generated a dictionary of roughly 1000 noun phrases. We extracted noun phrases from the French and

Spanish document collections, we then derived some rules to filter out proper nouns like "*Bill Clinton"* and phrases we thought non-content bearing such as "*année dernière"* or *"premier trimestre"*. Finally, we manually filtered the 1500 most frequent noun phrases to remove noisy phrases not captured by our simple rules. Examples of phrases are "*unión europea*" and "*casque bleu*".

Table 3 summarizes our results. Runs marked with the sign † are unofficial runs; for these runs the comparison to the median is indicative. We have included a corrected value for all our official runs. During the analysis of our results, we realized that while our documents were ranked correctly according to the scores in the engine, some of the scores reported were incorrect due to a conversion error in Java. This influenced our performance, since the evaluation program trec_eval resorts documents based on their score.

Using a phrase dictionary was neither harmful, nor helpful. We observed that phrases from the dictionary were found in only one fifth of the queries. For those queries, there is no clear emerging behavior: some perform better using phrases, while others do not. The difference in precision per query between the two runs is usually very small.

Table 3: Summary of all monolingual runs

			Performance of individual queries				
Run	Avg. Prec.	R-Prec.	Best	Above	Median	Below	Worst
German and Dutch							
Tlrdetd (corrected)	0.4205 (0.4418)	0.4151 (0.4269)	1	33	1	14	0
Tlrdetdn (corrected)	0.4581 (0.4701)	0.4570 (0.4606)	4	38	3	5	0
Tlrnltd (corrected)	0.3775 (0.3797)	0.3731 (0.3756)	6	31	2	11	0
Tlrnltdn†	(0.3999)	(0.3886)	(3)	(38)	0	(7)	(1)
French, Spanish and Italian							
Tlrfrtd (corrected)	0.4339 (0.4557)	0.4386 (0.4531)	4	17	4	23	1
Tlrfrtdn (corrected)	0.4516 (0.4684)	0.4436 (0.4638)	7	15	10	16	1
Tlrfrtdnpc (corrected)	0.4503 (0.4698)	0.4388 (0.4596)	6	18	10	14	1
Tlrestd (corrected)	0.5195 (0.5302)	0.5132 (0.5175)	3	26	10	10	0
Tlrestdpc (corrected)	0.5180 (0.5299)	0.5095 (0.5169)	3	25	11	10	0
Tlrestdn† (corrected)	(0.5559)	(0.5351)	(6)	(38)	(1)	(4)	(0)
Tlrestdnpc (corrected)	0.5347 (0.5559)	0.5280 (0.5361)	5	33	2	8	1
Tlrittd (corrected)	0.4306 (0.4375)	0.4325 (0.4292)	3	13	4	25	2

Our results for compounding languages are in the better half of the participants for these runs, so are our Spanish results. We believe that reusing of settings in a family of languages is indeed helpful. We need to perform some further analysis to confirm that belief.

Our Italian run was hindered by the lack of a good noise phrase list, as some of our structured queries still contained terms like *information* or *document*.

While we used last year’s settings for French, we did not achieve the performance we were aiming for. So far, we have identified two reasons. First, our noise phrase list for French missed capturing some of the patterns used in this year’s topics. When we manually cleaned the topics, we observed an improvement in the average precision. Some topics, however, benefited from non-content bearing terms that were not very frequent in the collection (for instance *énumérant* in queries 59 and 71). Next, while we originally intended to consider a term with multiple stems as a single concept, we realized that our scoring was overweighing such a term. Changing the behaviour would also have helped our French runs.

Final remarks

One of the problems in our bilingual runs was the coverage of the translation resources. Many translated queries still included original Spanish terms. In order to solve that problem, we can either use a MRD with a wider coverage (20,000 entries is a rather limited dictionary), or try to get a better coverage from the similarity thesauri. Better coverage may be achieved by using a parallel/comparable corpus in the same domain as the retrieval collections, if not the retrieval collections themselves (see [SBS97]). We will be investigating alignments of documents in related collections in the future.

Our monolingual runs contain no query expansion or pseudo-relevance feedback. Once we have refined the list used in query processing, for instance adding a list for misspelled terms, we will focus on automatic query expansion to try and enhance our searches.

References

[CCB92] 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

[SBS97] P. Sheridan, M. Braschler, and P. Schäuble. Cross-lingual information retrieval in a multilingual legal domain. In *Proceedings of the First European Conference on Research and Advanced Technology for Digital Libraries*, 1997.

[TTYF95] P. Thompson, H. Turtle, B. Yang and J. Flood, "TREC-3 Ad Hoc Retrieval and Routing Experiments using the WIN System," in *Overview of the 3rd Text Retrieval Conference (TREC-3),* NIST Special Publication 500-225, Gaithersburg, MD, April 1995.

[Tur90] H. Turtle. *Inference Networks for Document Retrieval*. PhD Thesis, Computer Science Department, University of Massassuchets, Amherst, 1990.

[Tur94] H. Turtle. Natural language vs. Boolean query evaluation : a comparison of retrieval performance. In *Proceedings of the 17th Annual International Conference on Research and Development in Information Retrieval*, Dublin, 1994