Thomson Legal and Regulatory Experiments at CLEF-2005

Isabelle Moulinier and Ken Williams
Thomson Legal and Regulatory
610 Opperman Drive
Eagan, MN 55123, USA
{Isabelle.Moulinier,Ken.Williams}@thomson.com

Abstract

For the 2005 Cross-Language Evaluation Forum, Thomson Legal and Regulatory participated in the Hungarian, French, and Portuguese monolingual search tasks as well as French-to-Portuguese bilingual retrieval. Our Hungarian participation focused on comparing the effectiveness of different approaches toward morphological stemming. Our French and Portuguese monolingual efforts focused on different approaches to Pseudo-Relevance Feedback (PRF), in particular the evaluation of a scheme for selectively applying PRF only in the cases most likely to produce positive results. Our French-to-Portuguese bilingual effort applies our previous work in query translation to a new pair of languages. All experiments were performed using our proprietary search engine. We remain encouraged by the overall success of our efforts, with our main submissions for each of the four tasks performing above the overall CLEF median. However, none of the specific enhancement techniques we attempted in this year's forum showed significant improvements over our initial results.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]: Indexing methods; Linguistic processing; H.3.3 [**Information Search and Retrieval**]: Query formulation; relevance feedback; H.3.4 [**Systems and Software**]: Performance evaluation; J.5 [**Arts and Humanities**]: Language translation

General Terms

Experimentation, Performance, Languages, Algorithms

Keywords

Natural Language Processing, Bilingual Information Retrieval, Hungarian Language, French Language, Portuguese Language, Machine Translation

1 Introduction

Thomson Legal and Regulatory participated in the Hungarian, French, and Portuguese monolingual search tasks as well as French-to-Portuguese bilingual retrieval.

Our Hungarian participation further evaluates the configuration developed in prior participations for compounding languages such as German or Finnish. We rely on morphological stemming to normalize derivations and factor compound terms. As morphological stemming may generate multiple stems for a given term, we compare the effectiveness of selecting a single stem with selecting all stems.

In our CLEF 2004 participation, we applied pseudo-relevance feedback blindly to all queries, even though this approach can be detrimental to some query results. In this participation, we take a first step toward selectively applying pseudo-relevance feedback. We apply our simple approach to our French and Portuguese runs.

Finally, our bilingual runs extend our previous work to two more languages. Our approach relies on query translation, where queries are translated term by term using translation resources built from parallel corpora.

We describe our experimental framework in Section 2, and present our monolingual and bilingual runs in Sections 3 and 4, respectively.

2 Experimental framework

The cornerstone of our experimental framework is our proprietary search engine which supports Boolean and Natural language search. Natural language search is based on an inference network retrieval model similar to INQUERY [1] and has been shown effective when compared to Boolean search on legal content [6]. For our CLEF experiments, we extended the search experience by incorporating the pseudo-relevance feedback functionality described in Section 2.3.

2.1 Indexing

Our indexing unit for European languages is a word. We identify words in sequences of characters using localized tokenization rules (for example, apostrophes are handled differently for French or Italian).

Each word is normalized for morphological variations. This includes identifying compounds if needed. We use the Inxight morphological stemmer [3] to perform such normalization which, in addition, can be configured to handle missing case and diacritic information.

Morphological stemming can produce multiple stems for a given term. We have introduced the option of selecting a single stem or keeping all stems. If candidate stems include compound terms, we select the stem with the fewest compound parts. If candidate stems are simple terms, we select the first one.

We do not remove stopwords from indices, as indexing supports both full-text search and natural language search. Stopwords are handled during search.

2.2 Search

Once documents are indexed, they can be searched. Given a query, we apply two steps: query formulation and document scoring.

Query formulation identifies “concepts” from natural language text by removing stopwords and other noise phrases, and imposes a Bayesian belief structure on these concepts. In many cases, each term in the natural language text represents a concept, and a flat structure gives the same weight to all concepts. However, phrases, compounds or misspellings can introduce more complex concepts, using operators such as “natural phrase,” “compound,” or “synonym.” The structure is then used to score documents.

Scoring takes evidence from each document as a whole, as well as from the best dynamic portion of each document. The best portion is computed dynamically based on proximal concept occurrences. Each concept contributes a belief to the document (and portion) score. We use a standard *tf-idf* scheme for computing term beliefs in all our runs. The belief of a single concept is given by:

$$bel_{term}(Q) = 0.4 + 0.6 \cdot tf_{norm} \cdot idf_{norm}$$

where

$$tf_{norm} = \frac{\log(tf + 0.5)}{\log(tf_{max} + 1.0)} \quad \text{and} \quad idf_{norm} = \frac{log(C + 0.5) - log(df)}{log(C + 1.0)}$$

tf is the number of occurrences of the term within the document, tf_{max} is the maximum number of occurrences of any term within the document, df is the number of documents containing the term and C the total number of documents in the collection. The various constants in the formulae have been determined by prior testing on manually-labeled data. tf_{max} is a weak indicator of document length.

2.3 Pseudo-relevance feedback

We have incorporated a pseudo-relevance feedback module into our search system. We follow the approach outlined by Haines and Croft [2].

We select terms for query expansion using a Rocchio-like formula and add the selected terms to the query. The added terms are weighted either using a fixed weight or a frequency-based weight.

$$sw = \alpha \cdot qf \cdot idf_{norm} + \frac{\beta}{|R|} \sum_{d \in R} (tf_{norm} \cdot idf_{norm}) - \frac{\gamma}{|\overline{R}|} \sum_{d \in \overline{R}} (tf_{norm} \cdot idf_{norm}) \quad (1)$$

where qf is the query weight, R is the set of documents considered relevant, $\overline{R}$ the set of documents considered not relevant, and $|X|$ denotes the cardinality of set X. The α, β and γ weights are set experimentally. The sets of documents R and $\overline{R}$ are extracted from the document list returned by the original search: R correspond to the top n documents and $\overline{R}$ to the bottom m, where n and m are determined through experiments on training data.

3 Monolingual experiments

Our monolingual participation focuses on normalization for Hungarian and pseudo-relevance feedback for French and Portuguese.

3.1 Hungarian experiments

As mentioned in Section 2.1, we use a morphological stemmer to identify compounds and normalize terms. The stemmer can be configured to allow for missing case and diacritic information. In addition, we can select to use one stem, or all stems.

At search time, compounds are treated as "natural phrases," i.e. as words within a proximity of 3. In addition, multiple stems are grouped under a single operator so that terms with multiple stems do not contribute more weight than terms with one single stem. Finally, we used the Hungarian stopword list developed by the Université de Neuchâtel [7].

We submitted two runs, each with its own indexing scheme:

- Run tlrTDhuE keeps all stems and allows for missing case and diacritic information.
- Run tlrTDhuSC keeps a single stem per term and does not correct missing information.

Run	MAP (Above/Equal/Below Median)	R-Prec	Reciprocal Rank
tlrTDhuE	0.2952 (27/0/23)	0.3210	0.5872
tlrTDhuSC	0.2964 (30/0/20)	0.2999	0.6070

Table 1: Performance for our official Hungarian runs comparing the effectiveness of the two stemming choices.

As illustrated by Table 1, there is no overall significant difference between the two runs, still we observe marked differences on a per-query basis: tlrTDhuSC outperforms tlrTDhuE on 25 queries and underperforms on 20 queries (differences range from a few percent to over 50%). This, we believe, is due to two factors: concepts in queries differ depending on the stemming approach; so do terms in the indices.

3.2 Pseudo-relevance feedback experiments

Pseudo-relevance feedback (PRF) is known to be useful on average but can be detrimental to the performance of individual queries. This year, we took a first step towards predicting whether or not PRF would aid individual queries.

We followed the following methodology: we selected our parameters for PRF using training data from previous CLEF participations for both French and Portuguese. We then manually derived a simple prediction rule that identifies those queries where PRF was very detrimental. Our decision rule is composed of two components: the score of the top ranked document and the maximum score any document can achieve for a given query, computed by setting the tf_{norm} factor in belief scores to 1. Our prediction rule is of the form:

```
if maxscore >= Min_MS_Value
   and (maxscore <  MS_Threshold or bestscore >= Min_TD_Value)
 Apply PRF
else
 Don't apply PRF
```

Using training data, we searched for the best parameters in this three-dimensional space (`Min_MS_Value`, `MS_Threshold`, and `Min_TD_Value`).

Our French and Portuguese results, reported in Table 2, show that PRF applied to all queries improved performance (although the difference is not always statistically significant) but that PRF applied to selected queries did not provide additional improvement.

It is interesting to note that PRF, selective or not, degrades the Reciprocal Rank measure, i.e. the average rank of the first relevant document. This indicates that our PRF setting decreases precision in the top-ranked documents, although it increases recall overall. A comparative summary is provided in Table 3.

Run	MAP (Above/Equal/Below Median)	R-Prec	Reciprocal Rank	Recall
tlrTDfr3	0.3735 (23/2/25)	0.3879	0.7014	0.8912
tlrTDfrRF2	0.4039† (35/0/15)	0.4012	0.6806	0.9141
tlrTDfrRFS1	0.4† (33/1/16)	0.3990	0.6806	0.9119
tlrTfr3	0.2925	0.3027	0.6163	0.7789
tlrTfrRF2	0.3073	0.3313	0.5729	0.8232
tlrTfrRFS1	0.3046	0.3278	0.5729	0.8215
tlrTDpt3	0.3501 (30/0/20)	0.3734	0.7542	0.8729
tlrTDptRF2	0.3742 (31/3/16)	0.3904	0.6704	0.9016
tlrTDptRFS1	0.3584 (31/3/16)	0.3805	0.6718	0.8939
tlrTpt3	0.2712	0.3141	0.6816	0.7358
tlrTptRF2	0.2844†	0.3215	0.6682	0.7544
tlrTptRFS1	0.2830	0.3208	0.6515	0.7544

Table 2: Official runs for French and Portuguese. Runs ending in 3 correspond to the base run without PRF. Runs ending in 2 are the PRF runs using the following configuration: add 5 terms from the top 10 documents; terms are selected with $\alpha = \beta = 1$ and $\gamma = 0$; expansion uses a fixed weight of 0.5 for each added term. Runs ending in 1 are PRF runs using the prediction rule. † indicates a statistically significant difference using the Wilcoxon signed-rank test and a p-value of 0.05.

Although it performed reasonably well on our initial training data, our PRF selection rule often applied PRF when it was detrimental, and failed to apply it when it would have helped. Table 4 gives more details on the prediction effectiveness or lack thereof. The number of queries for which PRF degraded performance is not unexpected as we did not intend to cover all cases with our heuristic. What is surprising is the low number of cases where our prediction rule prevented PRF from helping performance. We believe that the parameters we selected over-fitted the training data. Retrospectively, this is not all that surprising as we use raw values rather than proportions or normalized values.

Compared Runs	# queries degraded	No change	# queries improved
tlrTDfr3 vs. tlrTDfrRF2	11	0	39
tlrTDfr3 vs. tlrTDfrRFS1	11	5	34
tlrTfr3 vs. tlrTfrRF2	23	2	25
tlrTfr3 vs. tlrTfrRFS1	23	3	24
tlrTDpt3 vs. tlrTDptRF2	21	0	29
tlrTDpt3 vs. tlrTDptRFS1	18	9	23
tlrTpt3 vs. tlrTptRF2	17	2	31
tlrTpt3 vs. tlrTptRFS1	17	3	30

Table 3: Comparison between base runs and PRF runs using the MAP measure.

Compared runs	Correct	Misses	Errors
tlrTDfr3 vs. tlrTDfrRFS1	0	5	11
tlrTfr3 vs. tlrTfrRFS1	0	1	23
tlrTDpt3 vs. tlrTDptRFS1	3	6	18
tlrTpt3 vs. tlrTptRFS1	0	1	17

Table 4: Effectiveness of our prediction rule. Correct corresponds to cases when the prediction rule correctly avoided applying PRF. Misses corresponds to cases when PRF would have helped but was not applied. Errors corresponds to cases when the rule applied PRF and the performance degraded.

4 French to Portuguese bilingual experiments

Our 2005 bilingual experiments follow the approach we established during our CLEF 2004 participation. We performed bilingual search by translating query terms. Translation resources were trained from parallel corpora using the GIZA++ statistical machine translation package [5].

We created a bilingual lexicon by training the IBM Model 3 on the Europarl parallel corpus [4] as we found Model 3 to provide better translations than Model 1. We selected at most three translations per term, excluding translations with probabilities smaller than 0.1. During query formulation, translations were grouped under a *SUM[1] operator so that concepts are given the same importance regardless of the number of translations. In addition, translations were weighted by their probabilities.

Table 5 summarizes our bilingual runs. We submitted runs with and without pseudo-relevance feedback. The PRF runs show a behavior similar to our monolingual runs as reciprocal rank degrades but recall improves. Five times out of 6, the prediction rule predicted correctly that PRF should not be applied. However the number of cases when PRF was applied and performance dropped was also high (around 20).

The bilingual runs achieved between 60 and 65% of the average precision of monolingual runs. This performance is comparable to our results with German to French search, but not as promising as our training runs, which reached 80%.

[1] A *SUM node averages the beliefs of its children.

Run	MAP (Above/Equal/Below Median)	R-Prec	Reciprocal Rank	Recall
tlrTDfr2pt3	0.2209 (26/0/24)	0.2525	0.7147	0.7063
tlrTDfr2ptRF2	0.2318 (28/0/22)	0.2614	0.5442	0.7401
tlrTDfr2ptRFS1	0.2358 (29/0/21)	0.2689	0.5566	0.7415
tlrTfr2pt3	0.1741	0.2080	0.4807	0.6332
tlrTfr2ptRF2	0.1799	0.2056	0.3993	0.6563
tlrTfr2ptRFS1	0.1778	0.2045	0.4456	0.6582

Table 5: Official runs for French to Portuguese search. Runs ending in 3 correspond to the base run without PRF. Runs ending in 2 are the PRF runs use the following configuration: add 5 terms from the top 10 documents; terms are selected with $\alpha = \beta = 1$ and $\gamma = 0$; expansion uses a fixed weight of 0.5 for each added term. Runs ending in 1 use the prediction rule prior to applying PRF.

5 Conclusion

We remain encouraged by the overall success of our efforts, with our main submissions for each of the four tasks performing above the overall CLEF median. However, none of the specific enhancement techniques we attempted in this year's forum showed significant improvements over our initial results.

For monolingual retrieval in Hungarian, a highly morphological language, we explored two techniques for morphological stemming in order to identify compound terms and normalize them, but were unable to find significant differences between the results.

For monolingual retrieval in French and Portuguese, where we have previously shown pseudo-relevance feedback (PRF) to increase overall performance, we attempted to find a heuristic to identify specific queries for which PRF would be helpful. So far we have been unable to achieve this to a significant degree. In the future, we intend to explore additional techniques such as the use of machine learning including feature engineering as in [8] and methods for using normalized values rather than raw values to prevent over-fitting.

For bilingual retrieval from French to Portuguese, we achieve good performance relative to other submissions, but perhaps like other forum participants, we remain disappointed in the bilingual performance relative to the same queries performed in a monolingual setting. We need to better understand the differences between our official and training runs. In addition, we plan on investigating the usefulness of translation disambiguation.

References

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

[2] D. Haines and W.B. Croft. Relevance feedback and inference networks. In *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2–11, 1993.

[3] http://www.inxight.com/products/oem/linguistx.

[4] P. Koehn. Europarl: A multilingual corpus for evaluation of machine translation. Draft, 2002.

[5] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

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

[7] http://www.unine.ch/info/clef/.

[8] E. Yom-Tov, S. Fine, D. Carmel, A. Darlow, and E. Amitay. Juru at trec 2004: Experiments with prediction of query difficulty. In E. M. Voorhees and L. P. Buckland, editors, *The Thirteenth Text Retrieval Conference (TREC 2004)*. NIST Special Publication: SP 500-261, 2004.