CMU PRF using a Comparable Corpus: CLEF Working Notes

Monica Rogati (mrogati@cs.cmu.edu) and Yiming Yang (yiming@cs.cmu.edu)
Computer Science Department, Carnegie Mellon University

Abstract: We applied a PRF (Pseudo-Relevance Feedback) system, for both the monolingual task and the German(->English task. We focused on the effects of extracting a comparable corpus from the given newspaper data; our corpus doubled the average precision when used together with the provided parallel corpus. The PRF performance was lower for the queries with few relevant documents. We also examined the effects of the PRF first-step retrieval in the source language half of the parallel corpus vs. the entire document collection .

1. Introduction

For its first year at CLEF, the CMU group applied a PRF (Pseudo-Relevance Feedback) system, for both the monolingual task and the German->English task. We focused on the effects of extracting a comparable corpus from the given newspaper data; our corpus doubled the average precision when used together with the provided parallel corpus. The PRF performance was lower for the queries with few relevant documents. We also examined the effects of the PRF first-step retrieval in the source language half of the parallel corpus (official runs), when compared to the entire document collection (unofficial runs). This provides a relative upper bound (modulo the document collection) for the bilingual PRF method, since the entire collection is not available as bilingual text.

Section 2 briefly presents the PRF system; section 3 discusses the comparable corpus, and section 4 details the experimental setup and results.

2. The CMU Pseudo-Relevance Feedback system

The Pseudo-Relevance Feedback procedure is well known approach to query expansion in Information Retrieval. Its uses for both monolingual and translingual IR have been previously explored [3]. For the monolingual case, the algorithm assumes the top K retrieved documents are relevant and expands the original query using words selected from these documents. To cross the language barrier using PRF, a parallel bilingual collection is used for retrieval in the query (source) language, followed by query expansion/substitution using the corresponding documents in the target language.

A good parallel collection that closely matches the statistical profile of the target collection is essential for the success of this method. Given such a collection, the parameters that need to be tuned are the number of top relevant document used (K) , the number of words in the new query (E) and the weighting scheme for the retrieval engine (in our case, SMART). Section 4 contains more details about the experimental setup and the parameter values.

3. The comparable corpus

Intrigued by the IBM success in TREC 7 & 8 in the CLIR track [1], we adapted their approach to the extraction of a comparable corpus.

A web-based parallel collection was provided by the CLEF organizers; however, we believed that a parallel corpus that closely matched the document collection would be beneficial to the PRF performance. In the absence of such corpus, a comparable corpus that is derived from the given German and English newspapers could still be a useful resource. To obtain such a corpus, we used the statistical machine translation-based methodologies from IBM [1], adapted to our own resources and goals. As our results section shows, the comparable corpus doubled the 11-pt. average precision on the 2001 CLEF queries.

The fundamental assumption that underlines the generation of the comparable corpus is that the data itself is “comparable”; more specifically, that the articles in the newspapers contain the same events and ideas. This proved to be somehow difficult with the CLEF data, where the articles come from newspapers with very different characteristics. A similar mismatch has been previously observed when using the SDA data; we believe the LA

Times/German newspapers mismatch to be more pronounced. The results are even more encouraging when this mismatch is taken into account.

The algorithm for generating the comparable corpus

```
1) Divide the German and English newspaper articles into overlapping N days long windows.
2) Initialize a dictionary (similarity thesaurus) D
3) While the results are improving,
     a. For each window,
          i. Break the articles into fixed-size (P) paragraphs
          ii. Do a word-by-word translation of the paragraphs, using dictionary D and fertility F[1]
          iii. Use each paragraph in one language as a query, and retrieve the top matching paragraph
               among the ones in the other language
          iv. Repeat, switching languages
          v. If two paragraphs retrieved each other with a score above a certain threshold S, consider
             them “mates” and add them to the comparable corpus C
     b. Extract a dictionary D’ from C using CHI-SQUARED (see below) as a similarity measure
        between words
     c. D = D’
```

The CHI-SQUARED statistic is measure we found useful in several contexts, which captures the crosslingual lexical associations based on co-occurrence in the training corpus.

$$\textit{CHI-SQUARED}(t,s) = N(AD-BC)^2/[(A+C)(A+B)(D+B)(D+C)]$$

where
t = term in the target language
s = term in the source language
N = number of document pairs
A = number of documents where t occurs in the target language document and s occurs in the corresponding source language document
B = number of documents where t occurs in the target language document and s DOES NOT occur in the corresponding source language document
C = number of documents where t DOES NOT occur in the target language document and s occurs in the corresponding source language document
D = N-A-B-C

4. Experimental Setup and Results

We used the Porter stemmer and the SMART stopword list for the English collection. For the German collection, we used Morphix [2] as a German stemmer / compound analysis tool. We also used a short corpus-derived German stopword list. Morphix significantly improved the performance in the early stages of system development , and was used in all subsequent experiments.

The PRF parameters were tuned using the CLEF-200 data. Good empirical values were 15-25 for K (number of top documents considered relevant), 200-300 for E (number of query words after expansion), and ltc and ntc term weighting. The same values proved to be best for the CLEF-2001 queries, with the exception of the term weighting scheme (ntc performed significantly worse than ltc on the new queries).

We used one week windows overlapping by half when generating the comparable corpus, because some of the newspapers were published weekly and a more fine-grained distinction was not needed. The best results were

[1] The fertility is the number of words used to translate one word in the other language. This is different for every language pair.

obtained when the starting dictionary was not initialized (i.e. the first retrieval step was based on names and cognates). The resulting corpus had cca. 20000 paragraph pairs. A paragraph size of 250 bytes (plus the bytes necessary to keep the entire last word) worked best.

The quality of the comparable corpus was fairly low (most paragraphs were far from being translations of each other). This is understandable given that the only thing linking the German and English articles was the time period; the culture, continent and the newspapers' political goals and interpretations were different.

4.1 Official runs and results

Run Name	Task (DE)	Avg Precision	Weighting scheme	K	E	Query
CMUmll15e200td	ML	0.2467	ltc	5	200	Title+desc
CMUmll5e300td	ML	0.2397	ltc	5	300	Title+desc
CMUmll15e300tdn	ML	0.3057	ltc	15	300	Title+desc+narr
CMUbnn25e2td15	BL	0.1007	ntc	25	200	Title+desc
CMUbll25e3tdn25	BL	0.2013	ltc	25	300	Title+desc+narr
CMUbnn25e2tdn15	BL	0.1041	ntc	25	200	Title+desc+narr

There were 6 runs: 3 monolingual (DE->DE) and 3 bilingual (DE->EN). The results varied widely, and the term weighting was critical for the bilingual runs.

The monolingual runs were obtained by using the German (source language) half of the parallel corpus for the first retrieval step, in order to be consistent with the bilingual runs and provide a collection-dependent upper bound for them. Subsequent experiments revealed a significant difference between this procedure and the one using the entire collection (see figure below).

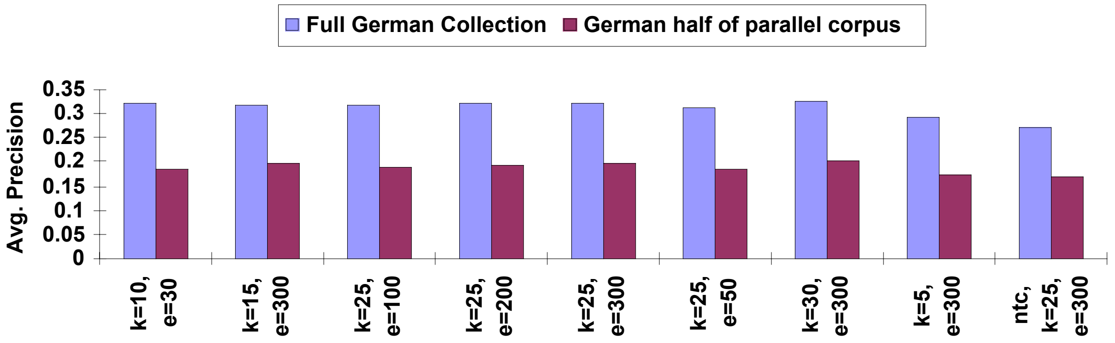

Figure 1: Monolingual Performance Using the Entire Collection

Another important factor that affected our performance was the number of relevant documents. When a query only has one or two relevant documents, the basic assumption of the PRF idea is violated. Specifically, PRF assumes the first K documents to be relevant, which is false for at least 75% of the CLEF-2001 queries (if K>=20), even with a perfect search engine.

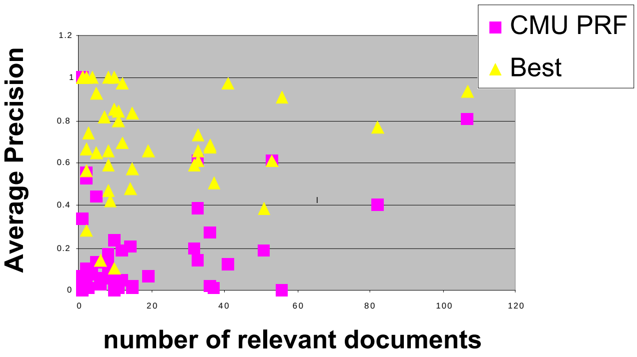

Figure 2:Correlations between avg. precision and the number of relevant documents

Our method's sensitivity to the low number of relevant documents is illustrated in the figure above, where PRF is compared to the best method for each query. The correlation between the fictional "best" method and the number of relevant documents was practically non-existent (-0.07), while CMU-PRF was comparatively more affected by the low number of relevant documents (the correlation is 0.26). We do not know how the individual runs from which the best result was selected were affected by the low number of relevant documents.

4.2 Unofficial results and experiments

After the release of the relevance judgments, we conducted several experiments to examine the effect different parameters had on the bilingual PRF system. The most impressive was the gain obtained from using the comparable corpus in addition to the given parallel corpus (the performance was doubled).

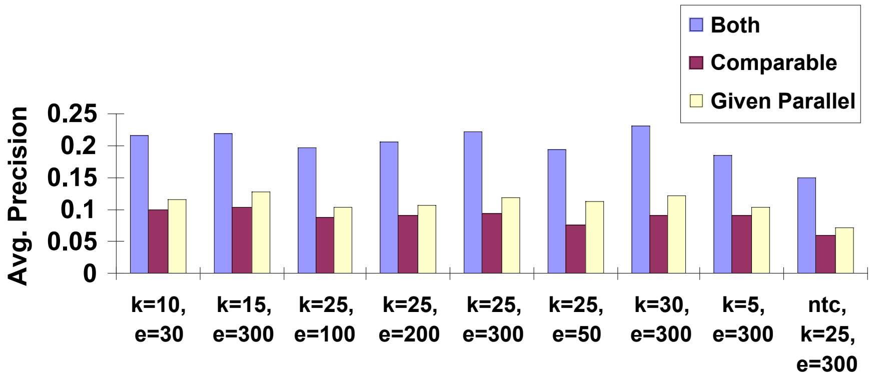

Figure 3: The Comparable Corpus Doubles the Performance

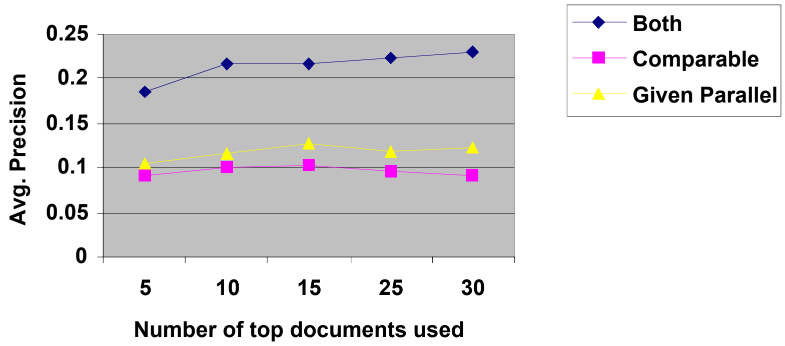

Figure 4: The effect of the Number of Documents Used for Query Expansion

The number of documents used for query expansion did not significantly affect the average precision, although the effects on individual queries remain to be examined.

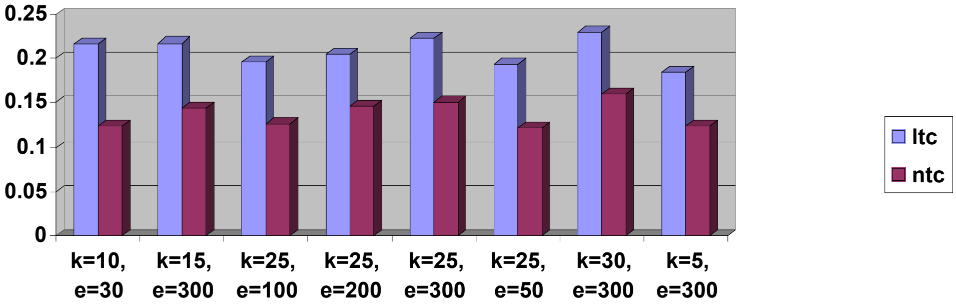

Figure 5: The effect of two term weighting schemes

Unlike the CLEF-2000 queries, this year's queries were sensitive to the term weighting scheme. The difference in performance between the two schemes shown above is significant.

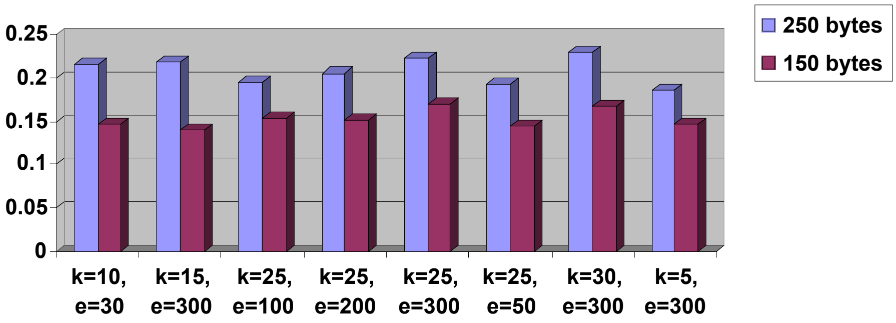

Figure 6: The effect of two comparable corpus paragraph sizes

The fixed paragraph size of the comparable corpus was another important factor. Paragraphs that were too short produced unreliable co-occurrence statistics.

5. Conclusion and Future Work

The most important finding in our experiments was the effect of the comparable corpus had on the average performance of the PRF method. Since the quality and quantity of parallel or comparable text is crucial to this method, we plan to gather more such data from the web. Preliminary results showed improvements over the official and unofficial CLEF runs and will be discussed elsewhere.

We are also planning to expand our CLEF toolkit to other methods previously implemented at CMU, such as GVSM, LSI, EBT and LLSF [3].

6. References

[1] M. Franz et al. Ad hoc and Multilingual Information Retrieval at IBM. In *The Seventh Text REtrieval Conference (TREC-8)*

[2] G. Neumann. Morphix Software Package. http://www.dfki.de/~neumann/morphix/morphix.html

[3] Y. Yang et al. Translingual Information Retrieval: Learning from Bilingual Corpora. In *AI Journal Special Issue: Best of IJCAI 1997*

[4] J. Xu and W.B. Croft. Query Expansion Using Local and Global Document Analysis. In *Proceedings of the Nineteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*