CONTROL: CLEF-2003 with Open, Transparent Resources Off-Line

Monica Rogati and Yiming Yang
Computer Science Department, Carnegie Mellon University
Pittsburgh, Pennsylvania
{mrogati, yiming}@cs.cmu.edu

Abstract: Corpus-based approaches to CLIR have been studied for many years. However, using commercial MT systems for CLEF has been considered easier and better performing. Our goal is to be one of the CLEF participants who show that the hypothetical performance drop is not large enough to justify the loss of control and transparency, especially for research systems. We participated in two bilingual runs and the small multilingual run using software and data that are free to obtain, transparent and modifiable.

1 Introduction

Over the past years, a necessary condition for a good cross- or multi-lingual performance in CLEF appeared to be the use of commercial MT systems, be it purchased or freely available online (Systran etc.)[1,3,11]. While using black boxes to cross the language barrier allowed researchers to concentrate on important issues such as stemming, query pre- and post-processing, combining black boxes outputs, and multilingual merging, [1,3,11] we believe that query translation does play an essential role in CLIR, and that understanding, control and transparency are crucial in a research system. Online MT systems can be upgraded, lose their free status, or change parameters at will, making past experiments irreproducible. If such a dependence is permitted, research in IR in general can be similarly reduced to pre- and post- processing of Google I/O. Our goal is to attempt to move away from basing the core of our CLIR research system on a module that cannot be fully understood and modified, to which future access might not be guaranteed, and in which external changes are allowed and sometimes not even detected. The main challenge, however, is to do so while sacrificing as little performance as possible.

Our initial attempt to reach this goal (CLEF 2001) was disappointing in this respect, mainly because we disallowed using translation resources entirely and relied on the temporal correspondence between CLEF documents to produce a “parallel” corpus. In CLEF 2003 we relaxed the independence requirement to using transparent data and code, freely available or available for a modest one-time fee, which we can store locally, easily modify, recompile and process, and which cannot change in uncontrollable or undetectable ways. We participated in two bilingual tasks (DE->IT, IT->ES), and the small multilingual task, which involved four languages.

Our general approach was to rely on parallel corpora and GIZA++ [8] for query translation, and on Lemur [9] for retrieval. All these resources (as well as the stemmers we used where applicable) fulfill the criteria outlined above. Moreover, with the exception of LDC data, which we did not use in the official runs but did use in preliminary experiments, all these resources are free of charge and publicly available.

In section 2 we discuss the parallel data and preprocessing (stemming, stopping etc.). In section 3 we discuss our approach to bilingual retrieval in general as well as approaches for situations where a parallel corpus between the two languages does not exist.

2 Data Description and Preprocessing

We have used the European Parliament proceedings 1996-2001 [6]. It includes versions in 11 European languages: Romance (French, Italian, Spanish, Portuguese), Germanic (English, Dutch, German, Danish, Swedish), Greek and Finnish. Sentence aligned parallel corpora (English-X) have been prepared by the author of [6]. We have also prepared German-Italian and Italian-Spanish versions for the two bilingual CLEF tasks we participated in, by detecting almost identical English sentences and aligning the corresponding non-English sentences. Table 1 shows the size of the relevant parallel corpora post-processing, after stopping and stemming. Some sentence pairs have been eliminated after becoming empty post-processing. Note that our quick intersection of X-EN to Y-EN parallel corpora by taking only sentences where the English versions were close resulted in losing about ¾ of the corpus. A much better approach would have been to follow [6]’s procedure, most likely resulting in a corpus of comparable size with the English versions.

Table 1: Size of the European Parliament parallel corpora (in sentences)

DE-IT	IT-ES	DE-EN	FR-EN	IT-EN	ES-EN
128505	150910	659773	674770	687890	738772

We have also experimented with several other corpora, including Hansard set A for French (available from the Linguistic Data Consortium). Although the sentence-aligned version was much larger (2.7M sentences), preliminary experiments on CLEF '01 and '02 datasets showed a consistent performance drop (usually around 10%). As a result, Hansard has not been used for CLEF 2003.

We preprocessed the parallel corpora and CLEF documents by eliminating punctuation, stopwords, and document sections disallowed in the task description. We have used the Porter stemmer for English and the rule-based stemmers and stopword lists kindly provided by J. Savoy [10]. After stemming, we have used 5-grams as a substitute for German word decompounding.

3 Bilingual Retrieval

Our main focus in bilingual retrieval has been query translation without the use of commercial MT systems, including Systran. In this section we will discuss our bilingual retrieval system using a parallel corpus, as well as the challenge of handling language pairs for which parallel corpora do not exist.
Conceptually, our approach consists of several steps:

1. Parallel corpora and test documents preprocessing
2. Dictionary generation from parallel corpora
3. Pseudo-Relevance Feedback in the source language
4. Query translation
5. Pseudo-Relevance Feedback in the target language
6. Retrieval

3.1 Dictionary Generation and Query Translation

We have used GIZA++ [8] as an implementation of IBM Model 1 [2]. GIZA++ takes a parallel corpus and generates a translation probability matrix. The number of training iterations was 10. Although GIZA++ implements the more sophisticated translation models discussed in [2], we have not used them for efficiency reasons, and because word order is not a factor during retrieval.
Query translation was done on a word-by-word basis. A significant difference from MT or online dictionary based approaches is that instead of using a rank-based cutoff (i.e. the first or first two variants for each word) we are using all translations weighted by their translation probability:

$$q_t = q_s \bullet M_{st}$$

where q_t is the query in the target language, q_s is the query in the source language, and M_{st} is the translation matrix. M was pruned to 50 translations per word for efficiency reasons.

This is similar to IBM and BBN CLIR approaches [4,5] except the translation is not integrated in the retrieval model; only the query is translated. This approach has the welcome side effect of a very focused query expansion.

3.2 Pseudo-Relevance Feedback and Retrieval

We have used the Lemur toolkit [9] to implement weighted query expansion, and we modified the retrieval interface to accept weighted queries as input. After query expansion is done in Lemur , the resulting query vector (q_s , words + weights) is extracted for future translation. After translation, q_t is loaded into Lemur for a new round of query expansion in the target language, followed by retrieval.

PRF and retrieval parameters we tuned include the number of documents to be considered relevant, the number of new query words added, the relative weight of added queries (usually 0.5) and term weighting method. There is one such parameter set for each pre- and post- translation query expansion, and for each language pair. However, experiments on CLEF 2001 and 2002 indicated that post-translation query expansion hurts performance by diluting the query in some languages, so the second set of parameters were set to 0 for the bilingual runs.

3.3 Handling language pairs with no available parallel corpora

The bilingual task this year was more challenging, in that we were aware of no Italian-Spanish or German-Italian parallel corpora. However, since most parallel corpora have English as one of the languages we had the option of using English as a pivot language in two ways:

1. to create a new parallel corpus if there is significant overlap (as described in Section 2). This is the least likely situation, but it does happen in the case where there is an underlying text translated in multiple languages, as it happened with the European Parliament corpus.
2. to translate first *to* English, then *from* English. This is where keeping and using translation probabilities is very useful. In traditional MT approaches, where the query is translated as a sentence twice, the (binary) mistakes accumulate, making the original meaning difficult to preserve. We believe the original meaning is easier to preserve when the entire query vector is translated, taking into account the translation probabilities:

$$q_t = q_s \bullet M_{s2EN} \bullet M_{EN2t}$$

where q_t is the query in the target language, q_s is the query in the source language, and M_{X2Y} is the translation matrix for language X to language Y.

3.4 Official Runs (German-Italian and Italian-Spanish)

All our official runs use the Title and Description fields. Relevant parameters are pre-translation feedback documents/terms, whether a new parallel corpus was created or if English was used as a pivot language during translation.

Table 2 : Official Bilingual Runs

Run Name	**Task**	**Feedback docs/terms**	**Parallel/Pivot**	**Avg. Precision**
cmuG2Icombfb	G2I	10/150	Pivot	0.3439
cmuG2Icomb	G2I	0/0	Pivot	0.3124
cmuG2Iparafb	G2I	10/150	Parallel	**0.4117**
cmuG2Ipara	G2I	0/0	Parallel	0.3669
cmuI2Scombfb	I2S	15/80	Pivot	**0.4269**
cmuI2Scomb	I2S	0/0	Pivot	0.4114
cmuI2Sparafb	I2S	15/80	Parallel	0.2921
cmuI2Spara	I2S	0/0	Parallel	0.4154

It is hard to draw conclusions from the official runs without more extensive experimentation on CLEF 2003 data (to be completed in the final version of the working notes). In particular, we are seeking an explanation for the extremely low relative performance of cmuI2Sparafb. If the run is not buggy, feedback performance is very unstable from one translation method to the other, and from language to language. This would not be a complete surprise, since feedback performance varied dramatically for French and Spanish on our system for CLEF 2001 and 2002 , and is the main reason why our runs are duplicated with their “low feedback” alternative in both bilingual and multilingual tasks.

4 Multilingual Retrieval

By using English as the query language we have leveraged the parallel corpora that had English as one of the languages. We have experimented with several parallel corpora, but chose the European Parliament proceedings as the corpus for our CLEF submission. We performed bilingual retrieval as described in Section 3, and we used Lemur for English monolingual retrieval. We then merged the results using the two methods described in Section 4.1. The number of feedback documents and words were tuned for each language.

4.1 Merging strategies

We examined two simple merging strategies: normalizing the individual scores and two step RSV [7].
The first strategy consists of normalizing the first N document scores to fit in the [0,1] interval, then using the normalized scores to produce the final ranked document list. This strategy is easy, requires no training but it has been proved inferior to regression-based models or two-step RSV.

Two-step RSV is a reindexing-based method: top ranked documents from each collection are translated to the topic language, then reindexed. Note that this is fundamentally different from translating the test collection, which we would like to avoid. Only top documents are translated, instead of a large test collection. However, the disadvantage of this method is that translation and reindexing need to be done online. Document caching can somewhat alleviate this problem when there are many queries.

Translation is done on a word-by-word basis, using the translation matrix built from the parallel corpus. We use only the first two translations for efficiency; however, we allocate S slots to each untranslated word and distribute the translated words proportionally to their normalized translation probabilities. Due to lack of running time, official runs had S=3.

4.2 Official Runs (Small Multilingual)

All our official runs use the Title and Description fields.

Table 3 : Official Multilingual Runs

Run Name	Feedback docs/terms pre- and post translation	Norm/2step merging	Avg. Pr.
cmuM4fb	EN: 5/30, FR:10/20-5/20, ES:5/20-10/20, DE:15/20-10/30	Norm	0.2921
cmuM4fbre	EN: 5/30, FR:10/20-5/20, ES:5/20-10/20, DE:15/20-10/30	2step	**0.3710**
cmuM4lowfb	EN: 5/30, FR:0/0-5/20, ES:0/0-10/20, DE:5/20-10/30	Norm	0.3398
cmuM4lowfbre	EN: 5/30, FR:0/0-5/20, ES:0/0-10/20, DE:5/20-10/30	2step	**0.3773**

Note that in this case the feedback made little difference among the best runs. The merging strategy had a significant impact, with the two-step RSV being better as expected.

5 Conclusion and Future Work

Our main goal in participating in this year’s CLEF was to prove that freedom from opaque, uncontrollable commercial systems does not have to mean poor CLIR performance for European languages. Many conceptual or implementation-related improvements can be made. They include better solutions for using a pivot language, especially when the domains do not match; better morphological processing, pseudo-relevant regression for merging etc.

6 References

[1] Braschler, M., Gohring, A. and Shauble, P. Eurospider at CLEF 2002. 2002. In C. Peters(Ed.), *Results of the CLEF2002 cross-language evaluation forum*, (to appear).

[2] Brown, P.F, Pietra, D., Pietra, D, Mercer, R.L. 1993. 2002.The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics, 19: 263-312

[3] Chen, A. Cross-language Retrieval Experiments at CLEF-2002. 2002. In C. Peters(Ed.), *Results of the CLEF2002 cross-language evaluation forum*, (to appear).

[4] Franz, M. and McCarley, J.S. 2002. Arabic Information Retrieval at IBM.TREC 2002 proceedings

[5] Fraser, A., Xu, J., Weischedel, R. 2002. TREC 2002 Cross-lingual Retrieval at BBN. TREC 2002 proceedings

[6] Koehn, P. Europarl: A Multilingual Corpus for Evaluation of Machine Translation. Draft, Unpublished.

[7] Martinez-Santiago, Martin M. and Urena, A. 2002. SINAI on CLEF 2002: Experiments with merging strategies. In C. Peters(Ed.), *Results of the CLEF2002 cross-language evaluation forum*, (to appear).

[8] Och, F. J. and Hermann N. 2000. Improved Statistical Alignment Models. In *Proc. of the 38th Annual Meeting of the Association for Computational Linguistics*, pp. 440-447, Hongkong, China

[9] Ogilvie, P and Callan, J. 2001. Experiments using the Lemur toolkit. In *Proceedings of the Tenth Text Retrieval Conference (TREC-10)*.

[10] Savoy, J. 1999. A stemming procedure and stopword list for general French corpora. Journal of the American Society for Information Science, 50(10), 944-952.

[11] Savoy, J. 2002. Report on CLEF-2002 Experiements: Combining multiple sources of evidence. In C. Peters(Ed.), *Results of the CLEF2002 cross-language evaluation forum*, (to appear).