Customizing Parallel Corpora at the Document Level

Monica ROGATI and Yiming YANG
Computer Science Department, Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
mrogati@cs.cmu.edu, yiming@cs.cmu.edu

Abstract

Recent research in cross-lingual information retrieval (CLIR) established the need for properly matching the parallel corpus used for query translation to the target corpus. We propose a document-level approach to solving this problem: building a custom-made parallel corpus by automatically assembling it from documents taken from other parallel corpora. Although the general idea can be applied to any application that uses parallel corpora, we present results for CLIR in the medical domain. In order to extract the best-matched documents from several parallel corpora, we propose ranking individual documents by using a length-normalized Okapi-based similarity score between them and the target corpus. This ranking allows us to discard 50-90% of the training data, while avoiding the performance drop caused by a good but mismatched resource, and even improving CLIR effectiveness by 4-7% when compared to using all available training data.

1 Introduction

Our recent research in cross-lingual information retrieval (CLIR) established the need for properly matching the parallel corpus used for query translation to the target corpus (Rogati and Yang, 2004). In particular, we showed that using a general purpose machine translation (MT) system such as SYSTRAN, or a general purpose parallel corpus - both of which perform very well for news stories (Peters, 2003) - dramatically fails in the medical domain. To explore solutions to this problem, we used cosine similarity between training and target corpora as respective weights when building a translation model. This approach treats a parallel corpus as a homogeneous entity, an entity that is self-consistent in its domain and document quality. In this paper, we propose that instead of weighting entire resources, we can select individual documents from these corpora in order to build a parallel corpus that is tailor-made to fit a specific target collection. To avoid confusion, it is helpful to remember that in IR settings the true *test data* are the queries, not the target documents. The documents are available off-line and can be (and usually are) used for training and system development. In other words, by matching the training corpora and the target documents we are not using test data for training.

(Rogati and Yang, 2004) also discusses indirectly related work, such as query translation disambiguation and building domain-specific language models for speech recognition. We are not aware of any additional related work.

In addition to proposing individual documents as the unit for building custom-made parallel corpora, in this paper we start exploring the criteria used for individual document selection by examining the effect of ranking documents using the length-normalized Okapi-based similarity score between them and the target corpus.

2 Evaluation Data

2.1 Medical Domain Corpus: Springer

The Springer corpus consists of 9640 documents (titles plus abstracts of medical journal articles) each in English and in German, with 25 queries in both languages, and relevance judgments made by native German speakers who are medical experts and are fluent in English. We split this parallel corpus into two subsets, and used the first subset (4,688 documents) for training, and the remaining subset (4,952 documents) as the test set in all our experiments. This configuration allows us to experiment with CLIR in both directions (EN-DE and DE-EN). We applied an alignment algorithm to the training documents, and obtained a sentence-aligned parallel corpus with about 30K sentences in each language.

2.2 Training Corpora

In addition to Springer, we have used four other English-German parallel corpora for training:

- NEWS is a collection of 59K sentence aligned news stories, downloaded from the web (1996-2000), and available at http://www.isi.edu/~koehn/publications/de-news/
- WAC is a small parallel corpus obtained by mining the web (Nie et al., 2000), in no particular domain
- EUROPARL is a parallel corpus provided by (Koehn). Its documents are sentence aligned European Parliament proceedings. This is a large collection that has been successfully used for CLEF, when the target corpora were collections of news stories (Rogati and Yang, 2003).
- MEDTITLE is an English-German parallel corpus consisting of 549K paired titles of medical journal articles. These titles were gathered from the PubMed online database (http://www.ncbi.nlm.nih.gov/PubMed/).

Table 1 presents a summary of the five training corpora characteristics.

Name	Size (sent)	Domain
NEWS	59K	news
WAC	60K	mixed
EUROPARL	665K	politics
SPRINGER	30K	medical
MEDTITLE	550K	medical

Table 1. Characteristics of Parallel Training Corpora

3 Selecting Documents from Parallel Corpora

While selecting and weighing entire training corpora is a problem already explored by (Rogati and Yang, 2004), in this paper we focus on a lower granularity level: individual documents in the parallel corpora. We seek to construct a custom parallel corpus, by choosing individual documents which best match the testing collection. We compute the similarity between the test collection (in German or English) and each individual document in the parallel corpora for that respective language. We have a choice of similarity metrics, but since this computation is simply retrieval with a long query, we start with the Okapi model (Robertson, 1993), as implemented by the Lemur system (Olgivie and Callan, 2001). Although the Okapi model takes into account average document length, we compare it with its length-normalized version, measuring per-word similarity. The two measures are identified in the results section by "Okapi" and "Normalized".

Once the similarity is computed for each document in the parallel corpora, only the top N most similar documents are kept for training. They are an approximation of the domain(s) of the test collection. Selecting N has not been an issue for this corpus (values between 10-75% were safe). However, more generally, this parameter can be tuned to a different test corpus as any other parameter. Alternatively, the document score can also be incorporated into the translation model, eliminating the need for thresholding.

4 CLIR Method

We used a corpus-based approach, similar to that in (Rogati and Yang, 2003). Let L1 be the source language and L2 be the target language. The cross-lingual retrieval consists of the following steps:

1. Expanding a query in L1 using blind feedback
2. Translating the query by taking the dot product between the query vector (with weights from step 1) and a translation matrix obtained by calculating translation probabilities or term-term similarity using the parallel corpus.
3. Expanding the query in L2 using blind feedback
4. Retrieving documents in L2

Here, blind feedback is the process of retrieving documents and adding the terms of the top-ranking documents to the query for expansion. We used simplified Rocchio positive feedback as implemented by Lemur (Olgivie and Callan, 2001). For the results in this paper, we have used Pointwise Mutual Information (PMI) instead of IBM Model 1 (Brown et al., 1993), since (Rogati and Yang, 2004) found it to be as effective on Springer, but faster to compute.

5 Results and Discussion

5.1 Empirical Settings

For the retrieval part of our system, we adapted Lemur (Ogilvie and Callan, 2001) to allow the use of weighted queries. Several parameters were tuned, none of them on the test set. In our corpus-

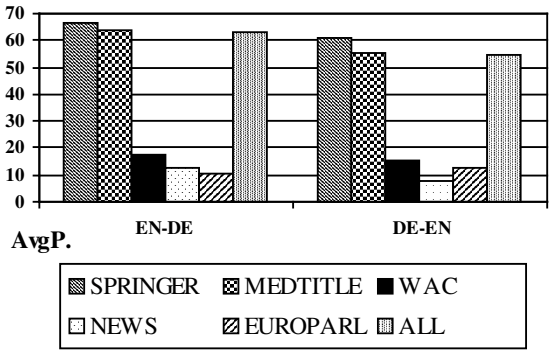
based approach, the main parameters are those used in query expansion based on pseudo-relevance, i.e., the maximum number of documents and the maximum number of words to be used, and the relative weight of the expanded portion with respect to the initial query. Since the Springer training set is fairly small, setting aside a subset of the data for parameter tuning was not desirable. We instead chose parameter values that were stable on the CLEF collection (Peters, 2003): 5 and 20 as the maximum numbers of documents and words, respectively. The relative weight of the expanded portion with respect to the initial query was set to 0.5. The results were evaluated using mean average precision (AvgP), a standard performance measure for IR evaluations.

In the following sections, DE-EN refers to retrieval where the query is in German and the documents in English, while EN-DE refers to retrieval in the opposite direction.

5.2 Using the Parallel Corpora Separately

Can we simply choose a parallel corpus that performed very well on news stories, hoping it is robust across domains? Natural approaches also include choosing the largest corpus available, or using all corpora together. Figure 1 shows the effect of these strategies.

Figure 1. CLIR results on the Springer test set by



using PMI with different training corpora.

We notice that choosing the largest collection (EUROPARL), using all resources available without weights (ALL), and even choosing a large collection in the medical domain (MEDTITLE) are all sub-optimal strategies.

Given these results, we believe that resource selection and weighting is necessary. Thoroughly exploring weighting strategies is beyond the scope of this paper and it would involve collection size, genre, and translation quality in addition to a measure of domain match. Here, we start by selecting individual documents that match the domain of the test collection. We examine the effect this choice has on domain-specific CLIR.

5.3 Using Okapi weights to build a custom parallel corpus

Figures 2 and 3 compare the two document selection strategies discussed in Section 3 to using all available documents, and to the ideal (but not truly optimal) situation where there exists a "best" resource to choose *and this collection is known*. By "best", we mean one that can produce optimal results on the test corpus, with respect to the given metric In reality, the true "best" resource is unknown: as seen above, many intuitive choices for the best collection are not optimal.

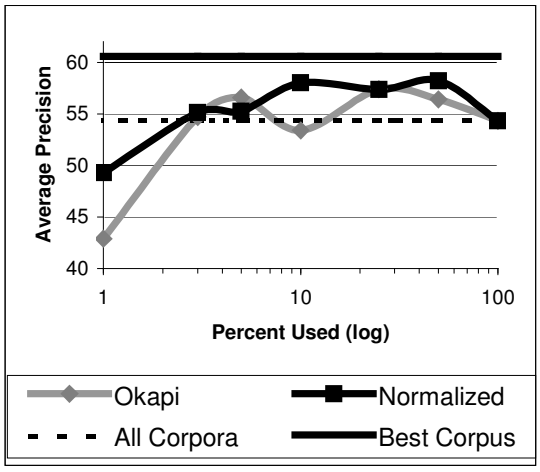


Figure 2. CLIR DE-EN performance vs. Percent of Parallel Documents Used. "Best Corpus" is given by an oracle and is usually unknown.

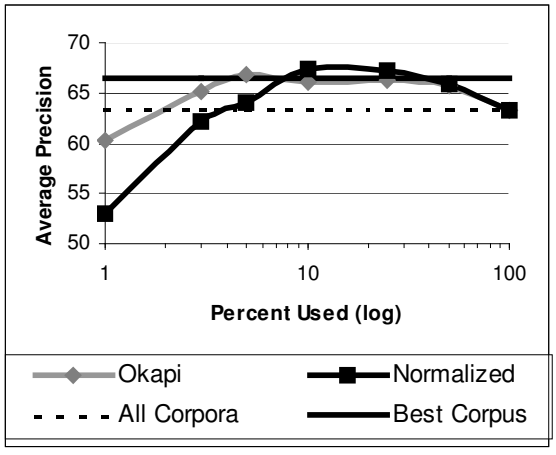


Figure 3. CLIR EN-DE performance vs. Percent of Parallel Documents Used. "Best Corpus" is given by an oracle and is usually unknown

Notice that the normalized version performs better and is more stable. Per-word similarity is, in this case, important when the documents are used to train translation scores: shorter parallel documents are better when building the translation matrix. Our strategy accounts for a 4-7% improvement over using all resources with no weights, for both retrieval directions. It is also very close to the "oracle" condition, which chooses the best collection in advance. More importantly, by using this strategy we are avoiding the sharp performance drop when using a mismatched, although very good, resource (such as EUROPARL).

6 Future Work

We are currently exploring weighting strategies involving collection size, genre, and estimating translation quality in addition to a measure of domain match. Another question we are examining is the granularity level used when selecting resources, such as selection at the document or cluster level.

Similarity and overlap between resources themselves is also worth considering while exploring tradeoffs between redundancy and noise. We are also interested in how these approaches would apply to other domains.

7 Conclusions

We have examined the issue of selecting appropriate training resources for cross-lingual information retrieval. We have proposed and evaluated a simple method for creating a customized parallel corpus from other available parallel corpora by matching the domain of the test documents with that of individual parallel documents. We noticed that choosing the largest collection, using all resources available without weights, and even choosing a large collection in the medical domain are all sub-optimal strategies. The techniques we have presented here are not restricted to CLIR and can be applied to other areas where parallel corpora are necessary, such as statistical machine translation. The trained translation matrix can also be reused and can be converted to any of the formats required by such applications.

8 Acknowledgements

We would like to thank Ralf Brown for collecting the MEDTITLE and SPRINGER data.

This research is sponsored in part by the National Science Foundation (NSF) under grant IIS-9982226, and in part by the DOD under award 114008-N66001992891808. Any opinions and conclusions in this paper are the authors' and do not necessarily reflect those of the sponsors.

References

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

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

Nie, J. Y., Simard, M. and Foster, G.. 2000. Using parallel web pages for multi-lingual IR. In C. Peters(Ed.), *Proceedings of the CLEF 2000 forum*

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

Peters, C. 2003. Results of the CLEF 2003 Cross-Language System Evaluation Campaign. *Working Notes for the CLEF 2003 Workshop*, 21-22 August, Trondheim, Norway

Robertson, S.E. and all. 1993. Okapi at TREC. In *The First TREC Retrieval Conference,* Gaithersburg, MD. pp. 21-30

Rogati, M and Yang, Y. 2003. Multilingual Information Retrieval using Open, Transparent Resources in CLEF 2003 . In C. Peters (Ed.), *Results of the CLEF2003 cross-language evaluation forum*

Rogati, M and Yang, Y. 2004. Resource Selection for Domain Specific Cross-Lingual IR. In Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'04).