A Fast Method for Parallel Document Identification

Jessica Enright and **Grzegorz Kondrak**
Department of Computing Science
University of Alberta
Edmonton, AB, T6G 2E8, Canada
{enright,kondrak}@cs.ualberta.ca

Abstract

We present a fast method to identify homogeneous parallel documents. The method is based on collecting counts of identical low-frequency words between possibly parallel documents. The candidate with the most shared low-frequency words is selected as the parallel document. The method achieved 99.96% accuracy when tested on the EUROPARL corpus of parliamentary proceedings, failing only in anomalous cases of truncated or otherwise distorted documents. While other work has shown similar performance on this type of dataset, our approach presented here is faster and does not require training. Apart from proposing an efficient method for parallel document identification in a restricted domain, this paper furnishes evidence that parliamentary proceedings may be inappropriate for testing parallel document identification systems in general.

1 Introduction

Parallel documents are documents that are mutual translations. There are a number of reasons one might want to either identify parallel documents, or confirm that a pair of documents are in fact parallel. Most prominently, one could use pairs of automatically detected parallel documents to build parallel corpora. Parallel corpora have many uses in natural language processing, and their dearth has been identified as a major bottleneck (Diab, 2004). They have been employed in word sense disambiguation (Diab and Resnik, 2002), automatic construction of bilingual dictionaries (McEwan et al., 2002), and inducing statistical machine translation models (Koehn et al., 2003). In addition to building parallel corpora, one can envision other uses for parallel document identification, such as cross-language information retrieval (Chen and Nie, 2000).

Much work on identifying pairs of parallel documents focuses on the use of external features of the documents, rather than content. Chen and Nie (2000) describe PTMiner, a cross-language information retrieval system. They consider a number of factors in determining if a pair of documents are parallel, including document size, date, URL, and language flag. For example, if a document is available in both French and English, it is common for the French document's URL to contain $.fr$ and the English to contain $.en$ In addition to these measures, they consider website structure.

McEwan et al. (2002) find parallel documents which they then use to automatically build a bilingual dictionary. In their system, they first generate a set of candidate pairs based on manual selection, or advanced search engine use. They then filter the pairs to remove non-parallel pairs. First, they confirm that one of each pair is in each of the desired languages using tuned lists of stop-words, then they compare the documents based on length in tokens, and HTML markup. Resnik and Smith (2003) use a similar idea of candidates and filters in their STRAND system. STRAND filters the documents based on aligning them by length in tokens and location of HTML markup in the documents.

Apart form the work done on external metrics, Patry and Langlais (2005) investigated a number of content-based metrics. They consider several document features, including the numbers, proper names and punctuation contained within, as well as document length, and alignment scores between candidate pairs. The features are then used to train an Ada-Boost classifier, which makes decisions based on edit-distance and cosine scores. They experimented with several combinations of features, one of which achieved 100% correctness when tested on 487 out of 488 parallel documents that constitute the English-Spanish portion of the EUROPARL corpus. They conclude that a bag-of-words approach is inferior to one that considers feature order.

In this work, we demonstrate that a much simpler approach can achieve equally good results. Our method does not depend on hand-coded linguistic knowledge and requires no training data, which may be unavailable for some language pairs. In addition, thanks to its simplicity, our method is very fast.

2 Parallel document identification

One can consider the parallel document identification problem to be as follows:

> Given one document d_A in language A, and a set of documents D_B in language B, identify exactly one document $d_B \in D_B$ that is the parallel, or translation, of d_A.

We initially designed a cognate-based approach to the problem, which employed a combination of orthographic word similarity measures to identify cognates such as French *nombres* and English *numbers* between documents. In order to make the method computationally feasible, potential cognates were filtered based on word order, location in the document, frequency, and length. However, we found that a faster and simpler procedure, which is described below, performed extremely well, eliminating the need for a more sophisticated approach.

We propose to identify parallel documents by counting the number of unique words that appear in both documents. The documents are treated as bags of words, that is, their word order is not considered. From each document, we extract a set of words that are at least 4 characters long and have frequency 1. Given a document in language A, we select the document in language B that shares the largest number of these words. An implementation based on hash tables ensures speed.

Since identical words of frequency 1 are almost certainly cognates, this method can be seen as an extremely conservative approach to cognate detection. In practice, most of unique identical words are proper nouns.

3 Experimental setup

We performed experiments on two different parliamentary corpora. The English-French Canadian Hansards from the 36th sitting of the Canadian Parliament (Germann, 2001) was selected as the development dataset. In testing on the Canadian Hansards, English was used as the Language A, and French as the Language B. Our approach correctly identified all parallel documents.

In order to allow for a direct comparison with the work of Patry and Langlais (2005), we adopted the EUROPARL corpus of parliamentary proceedings (Koehn, 2002) as our test dataset. However, rather than focusing on a single language pair, we performed tests on all 110 language pairs involving the following 11 languages: German, English, Greek, Finnish, Swedish, Dutch, French, Danish, Italian, Spanish and Portuguese. Diacritics were stripped from the documents of all languages. Since Greek utilizes a different script from the rest of the documents. we used a straightforward context-free mapping to convert every Greek character to its nearest roman equivalent.

Some of the 488 documents available in EUROPARL were missing in Finnish, Swedish, Greek and Danish. In particular, Greek had 392 documents, Danish had 487 documents, and Swedish and Finnish had 433 each. In such cases, the parallels of those missing documents were excluded from the language A for that test.

The EUROPARL documents range in size from 114 tokens (13 lines) to 138,557 tokens (11,101 lines). The mean number of tokens is 59,387 (2,826 lines). Each orientation of each language pair was tested. For example, for the language pair English-Dutch, tests were run twice - once with English as language A and Dutch as language B, and once the other way around. The results for a given language pair are not necessarily symmetric. Henceforth when referring to a language pair, we list the language A as the first one.

For each document and each language pair, an individual test was run. An individual test consisted of finding, for a given document in language A, its parallel in the language B set. Since we did not take advantage of the pigeon-hole constraint, the individual tests were independent from each other.

No changes were made to the approach once testing on the EUROPARL corpus began, in order to avoid adapting it to work on any particular data set.

4 Results

In total, only 20 of the 49872 tests did not produce the correct result (0.04% error rate). There was one incorrect selection in the English-Spanish language pair, one in the English-German pair, as well as in each of 18 language pairs involving Danish or English as a Language A. All of the incorrect results can be traced to mistranslation, or to missing/truncated documents. In particular, one of the documents is severely truncated in Danish and English, one of the German documents missing a portion of its text, and the Spanish version of one of the documents contains a number of phrases and sentences of English, apparently belonging to the English version of the text.

Effectively, when this method fails it is because the input does not match the problem definition. Recall that the problem was defined as selecting a document d_B from a set of documents D_B in language B that is the correct parallel to d_A, a document in language A. Failure cases occurred because there was no correct parallel to the d_A in D_B. In fact, each of the "incorrect" results is a manifestation of an editorial error in the EUROPARL corpus. One could see this approach being used as an aid to identifying fragmentary documents and mistranslations in parallel corpora.

Encouraged by the excellent accuracy of our method, we decided to try an even simpler approach, which is based on words of frequency 1 in the entire set of documents in a given language, rather than in a single document. For every document from a language A, we select as its parallel the document from language B that shares the most of those words with it. However, the results obtained with this method were clearly inferior, with the error rates ranging from 2.9% for Dutch to 27.3% for Finnish.

5 Discussion

The implications of this work are two-fold. First, it shows a simple, fast, and effective method for identifying parallel documents. Second, it calls into question the usefulness of parliamentary proceedings for the evaluation of parallel document identification schemes.

The method described in this paper is sufficiently simple as to be used as a baseline for comparison with other methods. No information is shared between trials, no word similarity measures are used, and word order is ignored. The method does not incorporate any language-specific linguistic knowledge, and it has shown itself to be robust across languages without any alterations. The only constraint is that the languages must share an alphabet, or can be converted into a common alphabet. Furthermore, it requires no training phase, which would likely have to be repeated for every pair of languages.

Our method achieves 99.9% accuracy on the English-Spanish language pair, which roughly matches the best result reported by Patry and Langlais (2005) (who apparently removed one document pair from the collection). However, their method requires a training phase on aligned parallel documents, making it time consuming and inconvenient to adapt their approach to a new language pair, even in cases where such document-aligned corpora are available. In addition, their top accuracy value corresponds to only one of several combination of features — the results with classifiers based on other combinations of features were lower.

We implemented our method using hash tables, which store the words occurring in a document together with their frequencies. This makes the entire search for a parallel document roughly linear in the total number of words in all the documents. Average total wall-clock time spent for one test with one language A document and 488 language B documents was 59.4 seconds. on a AMD Athlon(tm) 64 Processor 3500+. Profiling showed that on average 99.7% of the wall-clock time was spent on I/O operations, with the remainder taken by hash table lookups and string equality checks. Clearly, little speed improvement is possible. In contrast to the speed of our approach, the approach used by Patry and Langlais (2005) requires not only the time to train a classifier, but also the time to compute edit distance between many document pairs.

In addition to yielding a simple, accurate and fast method for parallel document identification, our results suggest that relatively "clean" collections of parliamentary proceedings of the EUROPARL type may be inappropriate for testing parallel document identification schemes in general. If a very simple approach can achieve near perfect accuracy in such a domain, perhaps the task is too easy. Future general parallel document identification systems should be tested on more challenging datasets.

6 Future Work

While the approach presented here has been very successful thus far, there are a number of extensions that could be made to make it more applicable in general. More work could allow it to deal with cases of missing parallel documents, datasets with fewer proper names, and even yield knowledge of the difficulty of the problem in general.

First, the problem definition could be expanded to include cases where there is no valid parallel for a given language A document in the language B document set. This could take the form of establishing a score or significance threshold. For example, if there were no document in the language B set that shared more than the minimum number of unique words with the document d_A in language A, then the approach might return no parallel for that document.

Second, it might be revealing to run further tests with this approach on other types of text than parliamentary proceedings. What types of text would require a more sophisticated approach? The answer to that question might have implications for the range of text types that ought to be used to comprehensively test parallel document identification systems.

The exact matching of words is a critical feature of our approach, which enables it to perform quick comparisons of documents by representing them as sets of low-frequency words stored in hash tables. However, it is also a limitation because many cross-language cognates are not orthographically identical. A system relying on non-binary word similarity measures rather than on total identity of words would be more complex and slower, but also more robust across different domains of text.

7 Conclusion

We have presented a viable, simple method for identification of homogeneous parallel documents. This method uses less resources and time than other content-based methods, a valuable asset when many languages lack linguistic resources. In addition to showing the effectiveness of our approach, the results of the experiments suggest that parliamentary proceedings may be inappropriate for parallel document identification scheme testing.

Acknowledgments

We would like to thank Colin Cherry and other members of the NLP research group at University of Alberta for their helpful comments and suggestions. This research was supported by the Natural Sciences and Engineering Research Council of Canada.

References

Jiang Chen and Jian-Yun Nie. 2000. Parallel web text mining for cross-language IR. In *In Proc. of RIAO*, pages 62–77.

Mona Diab and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proc. of ACL*, pages 255–262.

Mona Diab. 2004. Relieving the data acquisition bottleneck for word sense disambiguation. In *Proc. of ACL*, pages 303–310.

Ulrich Germann. 2001. Aligned Hansards of the 36th Parliament of Canada, Release 2001-1a. Available at http://www.isi.edu/natural-language/download/hansard/.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proc. of HLT-NAACL*, pages 48–54.

Philipp Koehn. 2002. Europarl: A multilingual corpus for evaluation of machine translation. Available at http://people.csail.mit.edu/koehn/.

Craig J. A. McEwan, Iadh Ounis, and Ian Ruthven. 2002. Building bilingual dictionaries from parallel web documents. In *Proc. of ECIR*, pages 303–323.

Alexandre Patry and Philippe Langlais. 2005. Automatic identification of parallel documents with light or without linguistic resources. In *Proc. of Canadian AI*, pages 354–365.

Philip Resnik and Noah A. Smith. 2003. The web as a parallel corpus. *Comput. Linguist.*, 29(3):349–380.