Extracting Parallel Sub-Sentential Fragments from Non-Parallel Corpora

Dragos Stefan Munteanu
University of Southern California
Information Sciences Institute
4676 Admiralty Way, Suite 1001
Marina del Rey, CA, 90292
dragos@isi.edu

Daniel Marcu
University of Southern California
Information Sciences Institute
4676 Admiralty Way, Suite 1001
Marina del Rey, CA, 90292
marcu@isi.edu

Abstract

We present a novel method for extracting parallel sub-sentential fragments from comparable, non-parallel bilingual corpora. By analyzing potentially similar sentence pairs using a signal processing-inspired approach, we detect which segments of the source sentence are translated into segments in the target sentence, and which are not. This method enables us to extract useful machine translation training data even from very non-parallel corpora, which contain no parallel sentence pairs. We evaluate the quality of the extracted data by showing that it improves the performance of a state-of-the-art statistical machine translation system.

1 Introduction

Recently, there has been a surge of interest in the automatic creation of parallel corpora. Several researchers (Zhao and Vogel, 2002; Vogel, 2003; Resnik and Smith, 2003; Fung and Cheung, 2004a; Wu and Fung, 2005; Munteanu and Marcu, 2005) have shown how fairly good-quality parallel sentence pairs can be automatically extracted from comparable corpora, and used to improve the performance of machine translation (MT) systems. This work addresses a major bottleneck in the development of Statistical MT (SMT) systems: the lack of sufficiently large parallel corpora for most language pairs. Since comparable corpora exist in large quantities and for many languages – tens of thousands of words of news describing the same events are produced daily – the ability to exploit them for parallel data acquisition is highly beneficial for the SMT field.

Comparable corpora exhibit various degrees of parallelism. Fung and Cheung (2004a) describe corpora ranging from noisy parallel, to comparable, and finally to very non-parallel. Corpora from the last category contain "... disparate, very non-parallel bilingual documents that could either be on the same topic (on-topic) or not". This is the kind of corpora that we are interested to exploit in the context of this paper.

Existing methods for exploiting comparable corpora look for parallel data at the sentence level. However, we believe that very non-parallel corpora have none or few good sentence pairs; most of their parallel data exists at the sub-sentential level. As an example, consider Figure 1, which presents two news articles from the English and Romanian editions of the BBC. The articles report on the same event (the one-year anniversary of Ukraine's Orange Revolution), have been published within 25 minutes of each other, and express overlapping content.

Although they are "on-topic", these two documents are non-parallel. In particular, they contain no parallel sentence pairs; methods designed to extract full parallel sentences will not find any useful data in them. Still, as the lines and boxes from the figure show, some parallel fragments of data do exist; but they are present at the sub-sentential level.

In this paper, we present a method for extracting such parallel fragments from comparable corpora. Figure 2 illustrates our goals. It shows two sentences belonging to the articles in Figure 1, and highlights and connects their parallel fragments.

Although the sentences share some common meaning, each of them has content which is not translated on the other side. The English phrase *reports the BBC's Helen Fawkes in Kiev*, as well

Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 81–88,
Sydney, July 2006.

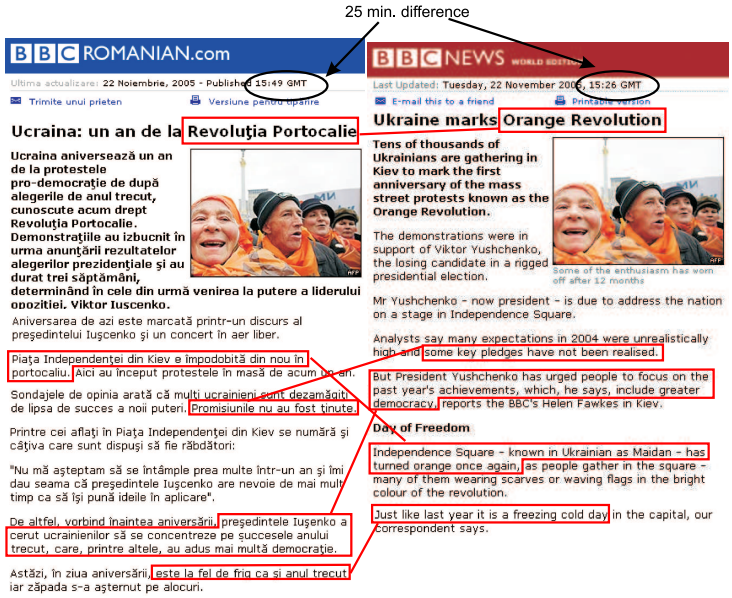


Figure 1: A pair of comparable, non-parallel documents

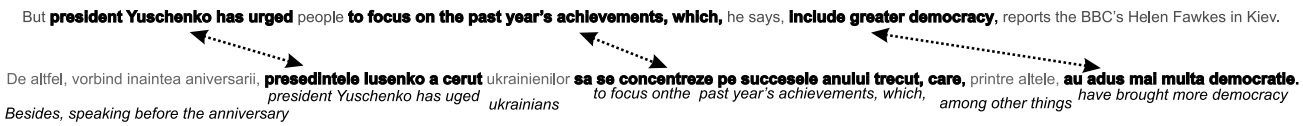


Figure 2: A pair of comparable sentences.

as the Romanian one *De altfel, vorbind inaintea aniversarii* have no translation correspondent, either in the other sentence or anywhere in the whole document. Since the sentence pair contains so much untranslated text, it is unlikely that any parallel sentence detection method would consider it useful. And, even if the sentences would be used for MT training, considering the amount of noise they contain, they might do more harm than good for the system's performance. The best way to make use of this sentence pair is to extract and use for training just the translated (highlighted) fragments. This is the aim of our work.

Identifying parallel subsentential fragments is a difficult task. It requires the ability to recognize translational equivalence in very noisy environments, namely sentence pairs that express different (although overlapping) content. However, a good solution to this problem would have a strong impact on parallel data acquisition efforts. Enabling the exploitation of corpora that do not share parallel sentences would greatly increase the amount of comparable data that can be used for SMT.

2 Finding Parallel Sub-Sentential Fragments in Comparable Corpora

2.1 Introduction

The high-level architecture of our parallel fragment extraction system is presented in Figure 3.

The first step of the pipeline identifies document pairs that are similar (and therefore more likely to contain parallel data), using the Lemur information retrieval toolkit[1] (Ogilvie and Callan, 2001); each document in the source language is translated word-for-word and turned into a query, which is run against the collection of target language documents. The top 20 results are retrieved and paired with the query document. We then take all sentence pairs from these document pairs and run them through the second step in the pipeline, the candidate selection filter. This step discards pairs which have very few words that are translations of each other. To all remaining sentence pairs we apply the fragment detection method (described in Section 2.3), which produces the output of the system.

We use two probabilistic lexicons, learned au-

[1] `http://www-2.cs.cmu.edu/$\sim$lemur`

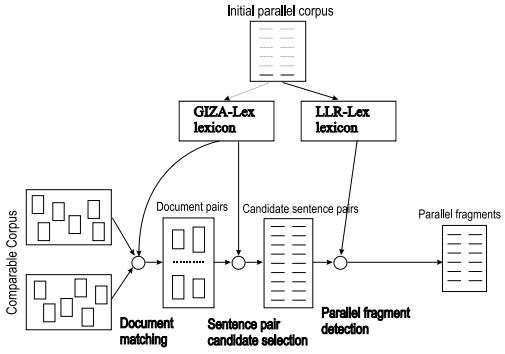


Figure 3: A Parallel Fragment Extraction System

tomatically from the same initial parallel corpus. The first one, *GIZA-Lex*, is obtained by running the GIZA++[2] implementation of the IBM word alignment models (Brown et al., 1993) on the initial parallel corpus. One of the characteristics of this lexicon is that each source word is associated with many possible translations. Although most of its high-probability entries are good translations, there are a lot of entries (of non-negligible probability) where the two words are at most related. As an example, in our *GIZA-Lex* lexicon, each source word has an average of 12 possible translations. This characteristic is useful for the first two stages of the extraction pipeline, which are not intended to be very precise. Their purpose is to accept most of the existing parallel data, and not too much of the non-parallel data; using such a lexicon helps achieve this purpose.

For the last stage, however, precision is paramount. We found empirically that when using *GIZA-Lex*, the incorrect correspondences that it contains seriously impact the quality of our results; we therefore need a cleaner lexicon. In addition, since we want to distinguish between source words that have a translation on the target side and words that do not, we also need a measure of the probability that two words are *not* translations of each other. All these are part of our second lexicon, *LLR-Lex*, which we present in detail in Section 2.2. Subsequently, in Section 2.3, we present our algorithm for detecting parallel sub-sentential fragments.

2.2 Using Log-Likelihood-Ratios to Estimate Word Translation Probabilities

Our method for computing the probabilistic translation lexicon *LLR-Lex* is based on the the Log-Likelihood-Ratio (LLR) statistic (Dunning, 1993), which has also been used by Moore (2004a; 2004b) and Melamed (2000) as a measure of word association. Generally speaking, this statistic gives a measure of the likelihood that two samples are not independent (i.e. generated by the same probability distribution). We use it to estimate the independence of pairs of words which cooccur in our parallel corpus.

If source word f and target word e are independent (i.e. they are not translations of each other), we would expect that $p(e|f) = p(e|\neg f) = p(e)$, i.e. the distribution of e given that f is present is the same as the distribution of e when f is not present. The LLR statistic gives a measure of the likelihood of this hypothesis. The LLR score of a word pair is low when these two distributions are very similar (i.e. the words are independent), and high otherwise (i.e. the words are strongly associated). However, high LLR scores can indicate either a positive association (i.e. $p(e|f) > p(e|\neg f)$) or a negative one; and we can distinguish between them by checking whether $p(e, f) > p(e)p(f)$.

Thus, we can split the set of cooccurring word pairs into positively and negatively associated pairs, and obtain a measure for each of the two association types. The first type of association will provide us with our (cleaner) lexicon, while the second will allow us to estimate probabilities of words *not* being translations of each other.

Before describing our new method more formally, we address the notion of word cooccurrence. In the work of Moore (2004a) and Melamed (2000), two words cooccur if they are present in a pair of aligned sentences in the parallel training corpus. However, most of the words from aligned sentences are actually unrelated; therefore, this is a rather weak notion of cooccurrence. We follow Resnik et. al (2001) and adopt a stronger definition, based not on sentence alignment but on word alignment: two words cooccur if they are linked together in the word-aligned parallel training corpus. We thus make use of the significant amount of knowledge brought in by the word alignment procedure.

We compute $LLR(e, f)$, the LLR score for words e and f, using the formula presented by Moore (2004b), which we do not repeat here due to lack of space. We then use these values to compute two conditional probability distributions: $P^+(e|f)$, the probability that source word f trans-

[2] `http://www.fjoch.com/GIZA++.html`

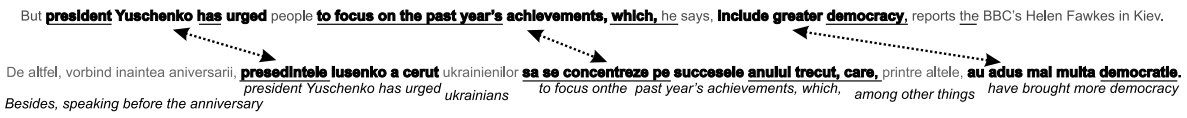


Figure 4: Translated fragments, according to the lexicon.

lates into target word e, and $P^-(e|f)$, the probability that f does not translate into e. We obtain the distributions by normalizing the LLR scores for each source word.

The whole procedure follows:

- Word-align the parallel corpus. Following Och and Ney (2003), we run GIZA++ in both directions, and then symmetrize the alignments using the refined heuristic.
- Compute all LLR scores. There will be an LLR score for each pair of words which are linked at least once in the word-aligned corpus
- Classify all $LLR(e, f)$ as either $LLR^+(e, f)$ (positive association) if $p(e, f) > p(e)p(f)$, or $LLR^-(e, f)$ (negative association) otherwise.
- For each f, compute the normalizing factors $\sum_e LLR^+(e, f)$ and $\sum_e LLR^-(e, f)$.
- Divide all $LLR^+(e, f)$ terms by the corresponding normalizing factors to obtain $P^+(e|f)$.
- Divide all $LLR^-(e, f)$ terms by the corresponding normalizing factors to obtain $P^-(e|f)$.

In order to compute the $P(f|e)$ distributions, we reverse the source and target languages and repeat the procedure.

As we mentioned above, in *GIZA-Lex* the average number of possible translations for a source word is 12. In *LLR-Lex* that average is 5, which is a significant decrease.

2.3 Detecting Parallel Sub-Sentential Fragments

Intuitively speaking, our method tries to distinguish between source fragments that have a translation on the target side, and fragments that do not. In Figure 4 we show the sentence pair from Figure 2, in which we have underlined those words of each sentence that have a translation in the other sentence, according to our lexicon *LLR-Lex*. The phrases "*to focus on the past year's achievements, which,*" and "*sa se concentreze pe succesele anului trecut, care,*" are mostly underlined (the lexicon is unaware of the fact that "achievements" and "succesele" are in fact translations of each other, because "succesele" is a morphologically inflected form which does not cooccur with "achievements" in our initial parallel corpus). The rest of the sentences are mostly not underlined, although we do have occasional connections, some correct and some wrong. The best we can do in this case is to infer that these two phrases are parallel, and discard the rest. Doing this gains us some new knowledge: the lexicon entry *(achievements, succesele)*.

We need to quantify more precisely the notions of "mostly translated" and "mostly not translated". Our approach is to consider the target sentence as a numeric *signal*, where translated words correspond to positive values (coming from the P^+ distribution described in the previous Section), and the others to negative ones (coming from the P^- distribution). We want to retain the parts of the sentence where the signal is mostly positive. This can be achieved by applying a smoothing filter to the signal, and selecting those fragments of the sentence for which the corresponding filtered values are positive.

The details of the procedure are presented below, and also illustrated in Figure 5. Let the Romanian sentence be the source sentence F, and the English one be the target, E. We compute a word alignment $F \rightarrow E$ by greedily linking each English word with its best translation candidate from the Romanian sentence. For each of the linked target words, the corresponding signal value is the probability of the link (there can be at most one link for each target word). Thus, if target word e is linked to source word f, the signal value corresponding to e is $P^+(e|f)$ (the distribution described in Section 2.2), i.e. the probability that e is the translation of f.

For the remaining target words, the signal value should reflect the probability that they are not

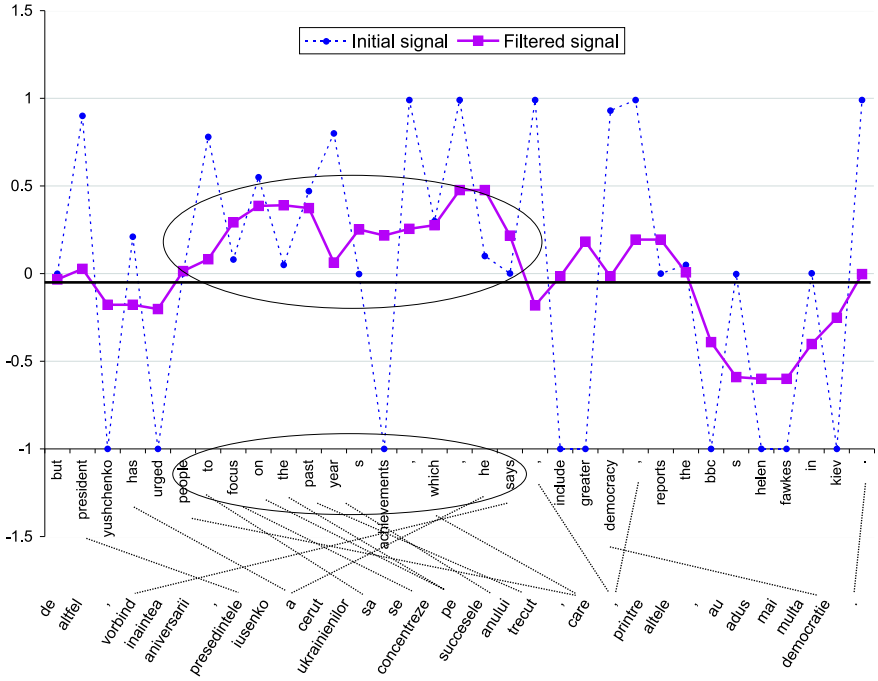


Figure 5: Our approach for detecting parallel fragments. The lower part of the figure shows the source and target sentence together with their alignment. Above are displayed the initial signal and the filtered signal. The circles indicate which fragments of the target sentence are selected by the procedure.

translated; for this, we employ the P^- distribution. Thus, for each non-linked target word e, we look for the source word least likely to be its non-translation: $f_0 = argmin_{f \in F} P^-(e|f)$. If f_0 exists, we set the signal value for e to $-P^-(e|f_0)$; otherwise, we set it to -1. This is the *initial signal*. We obtain the *filtered signal* by applying an averaging filter, which sets the value at each point to be the average of several values surrounding it. In our experiments, we use the surrounding 5 values, which produced good results on a development set. We then simply retain the "positive fragments" of E, i.e. those fragments for which the corresponding filtered signal values are positive.

However, this approach will often produce short "positive fragments" which are not, in fact, translated in the source sentence. An example of this is the fragment "*, reports*" from Figure 5, which although corresponds to positive values of the filtered signal, has no translation in Romanian. In an attempt to avoid such errors, we disregard fragments with less than 3 words.

We repeat the procedure in the other direction ($E \rightarrow F$) to obtain the fragments for f, and consider the resulting two text chunks as parallel. For the sentence pair from Figure 5, our system will output the pair:

people to focus on the past year's achievements, which, he says

sa se concentreze pe succesele anului trecut, care, printre

3 Experiments

In our experiments, we compare our fragment extraction method (which we call *FragmentExtract*) with the sentence extraction approach of Munteanu and Marcu (2005) (*SentenceExtract*). All extracted datasets are evaluated by using them as additional MT training data and measuring their impact on the performance of the MT system.

3.1 Corpora

We perform experiments in the context of Romanian to English machine translation. We use two initial parallel corpora. One is the training data for the Romanian-English word alignment task from the Workshop on Building and Using Parallel Corpora[3] which has approximately 1M English words. The other contains additional data

[3] http://www.statmt.org/wpt05/

Source	Romanian		English	
	# articles	# tokens	# articles	# tokens
BBC	6k	2.5M	200k	118M
EZZ	183k	91M	14k	8.5M

Table 1: Sizes of our comparable corpora

from the Romanian translations of the European Union's acquis communautaire which we mined from the Web, and has about 10M English words.

We downloaded comparable data from three online news sites: the BBC, and the Romanian newspapers "Evenimentul Zilei" and "Ziua". The *BBC* corpus is precisely the kind of corpus that our method is designed to exploit. It is truly non-parallel; as our example from Figure 1 shows, even closely related documents have few or no parallel sentence pairs. Therefore, we expect that our extraction method should perform best on this corpus.

The other two sources are fairly similar, both in genre and in degree of parallelism, so we group them together and refer to them as the *EZZ* corpus. This corpus exhibits a higher degree of parallelism than the BBC one; in particular, it contains many article pairs which are literal translations of each other. Therefore, although our sub-sentence extraction method should produce useful data from this corpus, we expect the sentence extraction method to be more successful. Using this second corpus should help highlight the strengths and weaknesses of our approach.

Table 1 summarizes the relevant information concerning these corpora.

3.2 Extraction Experiments

On each of our comparable corpora, and using each of our initial parallel corpora, we apply both the fragment extraction and the sentence extraction method of Munteanu and Marcu (2005). In order to evaluate the importance of the *LLR-Lex* lexicon, we also performed fragment extraction experiments that do not use this lexicon, but only *GIZA-Lex*. Thus, for each initial parallel corpus and each comparable corpus, we extract three datasets: *FragmentExtract*, *SentenceExtract*, and *Fragment-noLLR*. The sizes of the extracted datasets, measured in million English tokens, are presented in Table 2.

Initial corpus	Source	FragmentExtract	SentenceExtract	Fragment-noLLR
1M	BBC	0.4M	0.3M	0.8M
1M	EZZ	6M	4M	8.1M
10M	BBC	1.3M	0.9M	2M
10M	EZZ	10M	7.9M	14.3M

Table 2: Sizes of the extracted datasets.

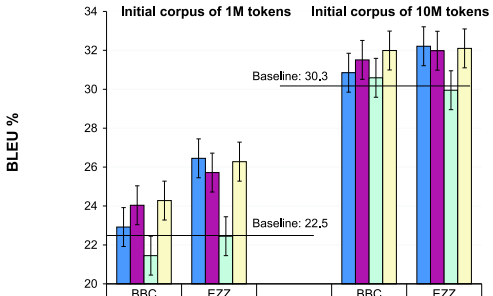
3.3 SMT Performance Results

We evaluate our extracted corpora by measuring their impact on the performance of an SMT system. We use the initial parallel corpora to train *Baseline* systems; and then train comparative systems using the initial corpora plus: the *FragmentExtract* corpora; the *SentenceExtract* corpora; and the *FragmentExtract-noLLR* corpora. In order to verify whether the fragment and sentence detection method complement each other, we also train a *Fragment+Sentence* system, on the initial corpus plus *FragmentExtract* and *SentenceExtract*.

All MT systems are trained using a variant of the alignment template model of Och and Ney (2004). All systems use the same 2 language models: one trained on 800 million English tokens, and one trained on the English side of all our parallel and comparable corpora. This ensures that differences in performance are caused only by differences in the parallel training data.

Our test data consists of news articles from the Time Bank corpus, which were translated into Romanian, and has 1000 sentences. Translation performance is measured using the automatic BLEU (Papineni et al., 2002) metric, on one reference translation. We report BLEU% numbers, i.e. we multiply the original scores by 100. The 95% confidence intervals of our scores, computed by bootstrap resampling (Koehn, 2004), indicate that a score increase of more than 1 BLEU% is statistically significant.

The scores are presented in Figure 6. On the *BBC* corpus, the fragment extraction method produces statistically significant improvements over the baseline, while the sentence extraction method does not. Training on both datasets together brings further improvements. This indicates that this corpus has few parallel sentences, and that by going to the sub-sentence level we make better use of it. On the *EZZ* corpus, although our method brings improvements in the BLEU score, the sen-



	BBC	EZZ		BBC	EZZ
Plus SentenceExtract	22.92	26.45		30.85	32.21
Plus FragmentExtract	24.04	25.72		31.51	31.98
Plus Fragment-noLLR	21.45	22.45		30.59	29.95
Plus Fragment+Sentence	24.28	26.28		31.99	32.1

Figure 6: SMT performance results

tence extraction method does better. Joining both extracted datasets does not improve performance; since most of the parallel data in this corpus exists at sentence level, the extracted fragments cannot bring much additional knowledge.

The *Fragment-noLLR* datasets bring no translation performance improvements; moreover, when the initial corpus is small (1M words) and the comparable corpus is noisy (BBC), the data has a negative impact on the BLEU score. This indicates that *LLR-Lex* is a higher-quality lexicon than *GIZA-Lex*, and an important component of our method.

4 Previous Work

Much of the work involving comparable corpora has focused on extracting word translations (Fung and Yee, 1998; Rapp, 1999; Diab and Finch, 2000; Koehn and Knight, 2000; Gaussier et al., 2004; Shao and Ng, 2004; Shinyama and Sekine, 2004). Another related research effort is that of Resnik and Smith (2003), whose system is designed to discover parallel document pairs on the Web.

Our work lies between these two directions; we attempt to discover parallelism at the level of fragments, which are longer than one word but shorter than a document. Thus, the previous research most relevant to this paper is that aimed at mining comparable corpora for parallel sentences.

The earliest efforts in this direction are those of Zhao and Vogel (2002) and Utiyama and Isahara (2003). Both methods extend algorithms designed to perform sentence alignment of parallel texts: they use dynamic programming to do sentence alignment of documents hypothesized to be similar. These approaches are only applicable to corpora which are at most "noisy-parallel", i.e. contain documents which are fairly similar, both in content and in sentence ordering.

Munteanu and Marcu (2005) analyze sentence pairs in isolation from their context, and classify them as parallel or non-parallel. They match each source document with several target ones, and classify all possible sentence pairs from each document pair. This enables them to find sentences from fairly dissimilar documents, and to handle any amount of reordering, which makes the method applicable to truly comparable corpora.

The research reported by Fung and Cheung (2004a; 2004b), Cheung and Fung (2004) and Wu and Fung (2005) is aimed explicitly at "very non-parallel corpora". They also pair each source document with several target ones and examine all possible sentence pairs; but the list of document pairs is not fixed. After one round of sentence extraction, the list is enriched with additional documents, and the system iterates. Thus, they include in the search document pairs which are dissimilar.

One limitation of all these methods is that they are designed to find only full sentences. Our methodology is the first effort aimed at detecting sub-sentential correspondences. This is a difficult task, requiring the ability to recognize translationally equivalent fragments even in non-parallel sentence pairs.

The work of Deng et. al (2006) also deals with sub-sentential fragments. However, they obtain parallel fragments from parallel sentence pairs (by chunking them and aligning the chunks appropriately), while we obtain them from comparable or non-parallel sentence pairs.

Since our approach can extract parallel data from texts which contain few or no parallel sentences, it greatly expands the range of corpora which can be usefully exploited.

5 Conclusion

We have presented a simple and effective method for extracting sub-sentential fragments from comparable corpora. We also presented a method for computing a probabilistic lexicon based on the LLR statistic, which produces a higher quality lexicon. We showed that using this lexicon helps improve the precision of our extraction method.

Our approach can be improved in several aspects. The signal filtering function is very simple; more advanced filters might work better, and eliminate the need of applying additional

heuristics (such as our requirement that the extracted fragments have at least 3 words). The fact that the source and target signal are filtered separately is also a weakness; a joint analysis should produce better results. Despite the better lexicon, the greatest source of errors is still related to false word correspondences, generally involving punctuation and very common, closed-class words. Giving special attention to such cases should help get rid of these errors, and improve the precision of the method.

Acknowledgements

This work was partially supported under the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022.

References

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Percy Cheung and Pascale Fung. 2004. Sentence alignment in parallel, comparable, and quasi-comparable corpora. In *LREC2004 Workshop*.

Yonggang Deng, Shankar Kumar, and William Byrne. 2006. Segmentation and alignment of parallel text for statistical machine translation. *Journal of Natural Language Engineering*. to appear.

Mona Diab and Steve Finch. 2000. A statistical word-level translation model for comparable corpora. In *RIAO 2000*.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Pascale Fung and Percy Cheung. 2004a. Mining very non-parallel corpora: Parallel sentence and lexicon extraction vie bootstrapping and EM. In *EMNLP 2004*, pages 57–63.

Pascale Fung and Percy Cheung. 2004b. Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. In *COLING 2004*, pages 1051–1057.

Pascale Fung and Lo Yuen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *ACL 1998*, pages 414–420.

Eric Gaussier, Jean-Michel Renders, Irina Matveeva, Cyril Goutte, and Herve Dejean. 2004. A geometric view on bilingual lexicon extraction from comparable corpora. In *ACL 2004*, pages 527–534.

Philipp Koehn and Kevin Knight. 2000. Estimating word translation probabilities from unrelated monolingual corpora using the EM algorithm. In *National Conference on Artificial Intelligence*, pages 711–715.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *EMNLP 2004*, pages 388–395.

I. Dan Melamed. 2000. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–249.

Robert C. Moore. 2004a. Improving IBM word-alignment model 1. In *ACL 2004*, pages 519–526.

Robert C. Moore. 2004b. On log-likelihood-ratios and the significance of rare events. In *EMNLP 2004*, pages 333–340.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4).

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

Franz Joseph Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–450.

P. Ogilvie and J. Callan. 2001. Experiments using the Lemur toolkit. In *TREC 2001*, pages 103–108.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL 2002*, pages 311–318.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. In *ACL 1999*, pages 519–526.

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

Philip Resnik, Douglas Oard, and Gina Lewow. 2001. Improved cross-language retrieval using backoff translation. In *HLT 2001*.

Li Shao and Hwee Tou Ng. 2004. Mining new word translations from comparable corpora. In *COLING 2004*, pages 618–624.

Yusuke Shinyama and Satoshi Sekine. 2004. Named entity discovery using comparable news articles. In *COLING 2004*, pages 848–853.

Masao Utiyama and Hitoshi Isahara. 2003. Reliable measures for aligning Japanese-English news articles and sentences. In *ACL 2003*, pages 72–79.

Stephan Vogel. 2003. Using noisy bilingual data for statistical machine translation. In *EACL 2003*, pages 175–178.

Dekai Wu and Pascale Fung. 2005. Inversion transduction grammar constraints for mining parallel sentences from quasi-comparable corpora. In *IJCNLP 2005*, pages 257–268.

Bing Zhao and Stephan Vogel. 2002. Adaptive parallel sentences mining from web bilingual news collection. In *2002 IEEE Int. Conf. on Data Mining*, pages 745–748.