Accurate Parallel Fragment Extraction from Quasi–Comparable Corpora using Alignment Model and Translation Lexicon

Chenhui Chu, Toshiaki Nakazawa, Sadao Kurohashi
Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo-ku
Kyoto, 606-8501, Japan
{chu,nakazawa}@nlp.ist.i.kyoto-u.ac.jp kuro@i.kyoto-u.ac.jp

Abstract

Although parallel sentences rarely exist in quasi–comparable corpora, there could be parallel fragments that are also helpful for statistical machine translation (SMT). Previous studies cannot accurately extract parallel fragments from quasi–comparable corpora. To solve this problem, we propose an accurate parallel fragment extraction system that uses an alignment model to locate the parallel fragment candidates, and uses an accurate lexicon filter to identify the truly parallel ones. Experimental results indicate that our system can accurately extract parallel fragments, and our proposed method significantly outperforms a state–of–the–art approach. Furthermore, we investigate the factors that may affect the performance of our system in detail.

1 Introduction

In statistical machine translation (SMT) (Brown et al., 1993; Koehn et al., 2007), since translation knowledge is acquired from parallel data, the quality and quantity of parallel data are crucial. However, except for a few language pairs, such as English–French, English–Arabic, English–Chinese and several European language pairs, parallel data remains a scarce resource. As non–parallel corpora are far more available, extracting parallel data from non–parallel corpora is an attractive research field.

Most previous studies focus on extracting parallel sentences from comparable corpora (Zhao and Vogel, 2002; Utiyama and Isahara, 2003; Munteanu and Marcu, 2005; Tillmann, 2009; Smith et al., 2010; Abdul-Rauf and Schwenk, 2011). Quasi–comparable corpora that contain far more disparate very–non–parallel bilingual documents that could either be on the same topic (in–topic) or not (out–topic) (Fung and Cheung, 2004), are available in far larger quantities than comparable corpora. In quasi–comparable corpora, there are few or no parallel sentences. However, there could be parallel fragments in comparable sentences that are also helpful for SMT.

Previous studies for parallel fragment extraction from comparable sentences have the problem that they cannot extract parallel fragments accurately. Some studies extract parallel fragments relying on a probabilistic translation lexicon estimated on an external parallel corpus. They locate the source and target fragments independently, making the extracted fragments unreliable (Munteanu and Marcu, 2006). Some studies develop alignment models for comparable sentences to extract parallel fragments (Quirk et al., 2007). Because the comparable sentences are quite noisy, the extracted fragments are not accurate.

In this paper, we propose an accurate parallel fragment extraction system. We locate parallel fragment candidates using an alignment model, and use an accurate lexicon filter to identify the truly parallel ones. Experimental results on Chinese–Japanese corpora show that our proposed method significantly outperforms a state–of–the–art approach, which indicate the effectiveness of our parallel fragment extraction system. Moreover, we investigate the factors that may affect the performance of our system in detail.

2 Related Work

(Munteanu and Marcu, 2006) is the first attempt to extract parallel fragments from comparable sentences. They extract sub-sentential parallel fragments by using a Log–Likelihood-Ratio (LLR) lexicon estimated on an external parallel corpus and a smoothing filter. They show the effectiveness of fragment extraction for SMT. This study has the drawback that they do not locate the source

International Joint Conference on Natural Language Processing, pages 1144–1150,
Nagoya, Japan, 14-18 October 2013.

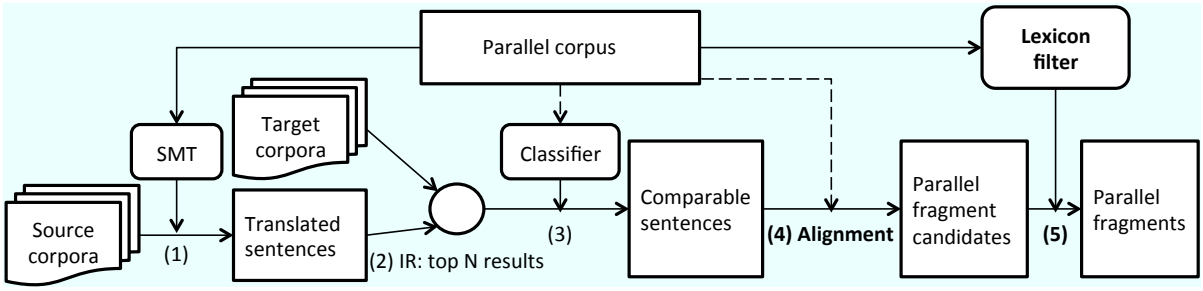

Figure 1: Parallel fragment extraction system.

and target fragments simultaneously, which cannot guarantee that the extracted fragments are translations of each other. We solve this problem by using an alignment model to locate the source and target fragments simultaneously.

Quirk et al. (2007) introduce two generative alignment models for extracting parallel fragments from comparable sentences. However, the extracted fragments slightly decrease MT performance when appending them to in–domain training data. We think the reason is that because the comparable sentences are quite noisy, the alignment models cannot accurately extract parallel fragments. To solve this problem we only use alignment models for parallel fragment candidate detection, and use an accurate lexicon filter to guarantee the accuracy of the extracted parallel fragments.

Besides the above studies, there are some other efforts. Hewavitharana and Vogel (2011) propose a method that calculates both the inside and outside probabilities for fragments in a comparable sentence pair, and show that the context of the sentence helps fragment extraction. However, the proposed method only can be efficient in a controlled manner that supposes the source fragment was known, and search for the target fragment. Another study uses a syntax–based alignment model to extract parallel fragments from noisy parallel data (Riesa and Marcu, 2012). Since their method is designed for noisy parallel data, we believe that the method cannot accurately extract parallel fragments from comparable sentences.

3 Proposed Method

3.1 System Overview

Figure 1 shows an overview of our parallel fragment extraction system. We first apply comparable sentence extraction using a combination method of (Abdul-Rauf and Schwenk, 2011) (1)(2) and (Munteanu and Marcu, 2005) (3), which were originally used for extracting parallel sentences from comparable corpora. We translate the source sentences to target language with a SMT system trained on a parallel corpus (1). Then we use the translated sentences as queries for IR. We retrieve the top 10 target documents for each source sentence using Indri[1], and use all sentences in the documents as comparable sentence candidates (2). Next, we identify the comparable sentences from the candidates using a classifier trained on a part of a parallel corpus[2] following (Munteanu and Marcu, 2005) (3).

As the noise in comparable sentences will decrease MT performance, we further apply parallel fragment extraction. We apply two steps to accurately extract parallel fragments. We first detect parallel fragment candidates using bidirectional IBM models (Brown et al., 1993) with symmetrization heuristics (Koehn et al., 2007) (4). The generative alignment models proposed by Quirk et al. (2007) may be more efficient for parallel fragment candidate detection, we leave this for future work. Then we filter the candidates with probabilistic translation lexicon to produce accurate results (5). We present the details of our proposed method in following sections.

3.2 A Brief Example

Figure 2 shows an example of comparable sentences extracted from Chinese–Japanese quasi–comparable corpora by our system. The alignment results are computed by IBM models. We notice that the truly parallel fragments "lead ion selective electrode" and "potentiometric titration method" are aligned, although there are some incorrectly aligned word pairs. We think this kind

[1] http://www.lemurproject.org/indri

[2] In our experiments, we used 5k parallel sentences.

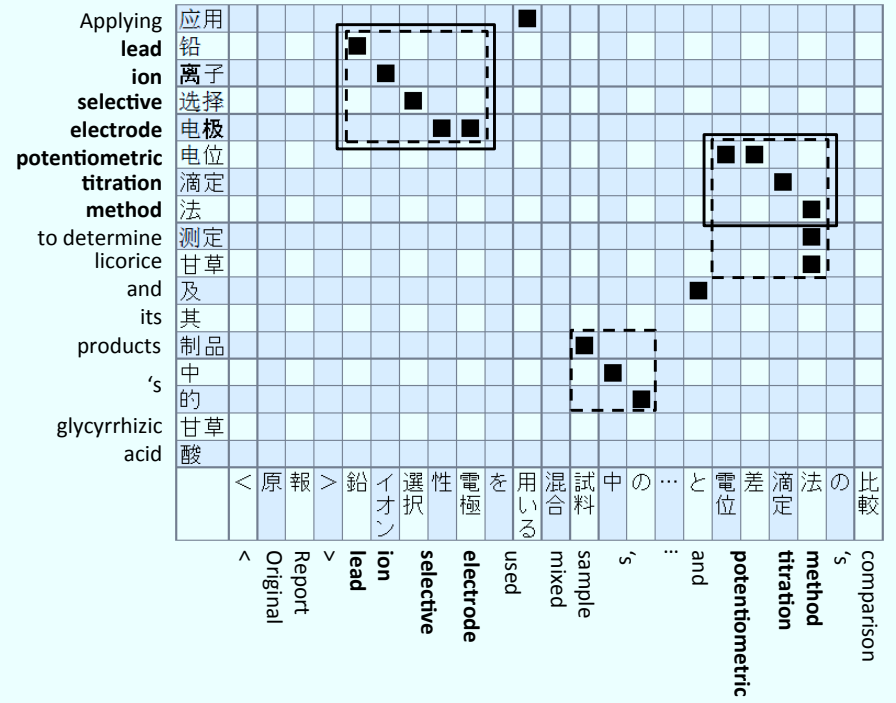

Figure 2: Example of comparable sentences with alignment results computed by IBM models (Parallel fragment candidates are in dashed rectangles, parallel fragments are in rectangles with solid line border).

of alignment information can be helpful for fragment extraction. What we need to do is develop a method to identify the true parallel fragments from the aligned fragments.

3.3 Parallel Fragment Candidate Detection

We treat the longest spans that have monotonic and non–null alignment as parallel fragment candidates. The reason we only consider monotonic ones is that based on our observation, ordering of IBM models on comparable sentences is unreliable. Quirk et al. (2007) also produce monotonic alignments in their generative model. Monotonic alignments are not sufficient for many language pairs. In the future, we plan to develop a method to deal with this problem. The non–null constraint can limit us from extracting incorrect fragments. Similar to previous studies, we are interested in fragment pairs with size greater than 3. Taking the comparable sentences in Figure 2 as an example, we will extract the fragments in dashed rectangles as parallel fragment candidates.

3.4 Lexicon–Based Filter

The parallel fragment candidates cannot be used directly, because many of them are still noisy as shown in Figure 2. Aiming to produce accurate results, we use a lexicon–based filter. We filter a candidate parallel fragment pair with a probabilistic translation lexicon. The lexicon–pair may be extracted from a parallel corpus, or from comparable corpora using some state–of–the–art approaches such as (Vulić et al., 2011). In this study, we use the lexicon extracted from a parallel corpus. Different lexicons may have different effects for filtering. Here, we compare three types of lexicon. The first lexicon we use is the IBM Model 1 lexicon, which is obtained by running GIZA++[3] that implements sequential word–based statistical alignment model of IBM models.

The second lexicon we use is the LLR lexicon. Munteanu and Marcu (2006) show that the LLR lexicon performs better than the IBM Model 1 lexicon for parallel fragment extraction. One advantage of the LLR lexicon is that it can produce both positive and negative associations. Munteanu and Marcu (2006) develop a smoothing filter applying this advantage. We extract the LLR lexicon from a word–aligned parallel corpus using the same method as (Munteanu and Marcu, 2006).

The last lexicon we use is the SampLEX lexicon. Vulić and Moens (2012) propose an associative approach for lexicon extraction from par-

[3]http://code.google.com/p/giza-pp

allel corpora that relies on the paradigm of data reduction. They extract translation pairs from many smaller sub–corpora that are randomly sampled from the original corpus, based on some frequency–based criteria of similarity. They show that their method outperforms IBM Model 1 and other associative methods such as LLR in terms of precision and F-measure. We extract SampLEX lexicon from a parallel corpus using the same method as (Vulić and Moens, 2012).

Aiming to gain new knowledge that does not exist in the lexicon, we apply a smoothing filter similar to (Munteanu and Marcu, 2006). For each aligned word pair in the fragment candidates, we set scores to the words in two directions according to the extracted lexicon. If the aligned word pair exists in the lexicon, we set the corresponding translation probabilities as scores. For LLR lexicon, we use both positive and negative association values. If the aligned word pair does not exist in the lexicon, we set the scores in both directions to -1. There is the one exception that the aligned words are the same number, punctuation or abbreviation. In this case, we set the scores to 1 without considering the existence of the word pair in the lexicon. After this process, we get *initial scores* for the words in the fragment candidates in two directions.

We then apply an averaging filter to the *initial scores* to obtain *filtered scores* in both directions. The averaging filter sets the score of one word to the average score of several words around it. We think the words with initial positive scores are reliable, because they satisfy two strong constraints, namely alignment by IBM models and existence in the lexicon. Therefore, unlike (Munteanu and Marcu, 2006), we only apply the averaging filter to the words with negative scores. Moreover, we add another constraint that only filtering a word when both the left and right words around it have positive scores, which can further guarantee accuracy. For the number of words used for averaging, we used 5 (2 preceding words and 2 following words). The heuristics presented here produced good results on a development set.

Finally, we extract parallel fragments according to the *filtered scores*. We extract word aligned fragment pairs with continuous positive scores in both directions. Fragments with less than 3 words may be produced in this process, and we discard them like previous studies.

4 Experiments

In our experiments, we compared our proposed fragment extraction method with (Munteanu and Marcu, 2006). We manually evaluated the accuracy of the extracted fragments. Moreover, we used the extracted fragments as additional MT training data, and evaluated the effectiveness of the fragments for MT. We conducted experiments on Chinese–Japanese data. In all our experiments, we preprocessed the data by segmenting Chinese and Japanese sentences using a segmenter proposed by Chu et al. (2012) and JUMAN (Kurohashi et al., 1994) respectively.

4.1 Data

4.1.1 Parallel Corpus

The parallel corpus we used is a scientific paper abstract corpus provided by JST[4] and NICT[5]. This corpus was created by the Japanese project "Development and Research of Chinese–Japanese Natural Language Processing Technology", containing 680k sentences (18.2M Chinese and 21.8M Japanese tokens respectively). This corpus contains various domains such as chemistry, physics, biology and agriculture etc.

4.1.2 Quasi–Comparable Corpora

The quasi–comparable corpora we used are scientific paper abstracts collected from academic websites. The Chinese corpora were collected from CNKI[6], containing 420k sentences and 90k articles. The Japanese corpora were collected from CiNii[7] web portal, containing 5M sentences and 880k articles. Most articles in the Chinese corpora belong to the domain of chemistry, while the Japanese corpora contain various domains such as chemistry, physics and biology etc. Note that since the articles in these two websites were written by Chinese and Japanese researchers respectively, the collected corpora are very–non–parallel.

4.2 Extraction Experiments

We first applied sentence extraction on the quasi–comparable corpora using our system, and 30k comparable sentences of chemistry domain were extracted. We then applied fragment extraction on the extracted comparable sentences. We compared our proposed method with (Munteanu and

[4]http://www.jst.go.jp
[5]http://www.nict.go.jp
[6]http://www.cnki.net
[7]http://ci.nii.ac.jp

Method	# fragments	Average size (zh/ja)	Accuracy
Munteanu+, 2006	28.4k	20.36/21.39	(1%)
Only (IBM Model 1)	18.9k	4.03/4.14	80%
Only (LLR)	18.3k	4.00/4.14	**89%**
Only (SampLEX)	18.4k	3.96/4.05	87%
External (IBM Model 1)	28.7k	4.18/4.33	81%
External (LLR)	26.9k	4.17/4.33	85%
External (SampLEX)	28.0k	4.11/4.23	82%

Table 1: Fragment extraction results (Accuracy was manually evaluated on 100 fragments randomly selected from fragments extracted by different methods, based on the number of exact match).

Marcu, 2006). We applied word alignment using GIZA++. External parallel data might be helpful for alignment models to detect parallel fragment candidates from comparable sentences. Therefore, we compared two different settings to investigate the influence of external parallel data for alignment to our proposed method:

- Only: Only use the extracted comparable sentences.
- External: Use a small number of external parallel sentences together with the comparable sentences (In our experiment, we used chemistry domain data of the parallel corpus described in Section 4.1.1, containing 11k sentences).

We also compared IBM Model 1, LLR and SampLEX lexicon for filtering. All lexicons were extracted from the parallel corpus.

Table 1 shows the results for fragment extraction. We can see that the average size of fragments extracted by (Munteanu and Marcu, 2006) is unusually long, which is also reported in (Quirk et al., 2007). Our proposed method extracts shorter fragments. The number of extracted fragments and the average size are similar among the three lexicons when using the same alignment setting. Using the external parallel data for alignment extracts more fragments than only using the comparable sentences, and the average size is slightly larger. We think the reason is that the external parallel data is helpful to improve the recall of alignment for the parallel fragments in the comparable sentences, thus more parallel fragments will be detected.

To evaluate accuracy, we randomly selected 100 fragments extracted by the different methods. We manually evaluated the accuracy based on the number of exact match. Note that exact match criteria has a bias against (Munteanu and Marcu, 2006), because their method extacts subsentential fragments which are quite long. We found that only one of the fragments extracted by "Munteanu+, 2006" is exact match, while for the remainder only partial matches are contained in long fragments. Our proposed method have a accuracy over 80%, while the remainder are partial matches. For the effects of different lexicons, LLR and SampLEX shows better performance than IBM Model 1 lexicon. We think the reason is the same one reported in previous studies that LLR and SampLEX lexicon are more accurate than IBM Model 1 lexicon. Also, LLR lexicon performs slightly better than SampLEX lexicon in this experiment. The accuracy of only using the comparable sentences for alignment are better than using the external parallel data, except for IBM Model 1 lexicon. We think the reason is that the external parallel data may have a bad effect on the precision of alignment for the parallel fragments in the comparable sentences.

4.3 Translation Experiments

We further conducted Chinese–to–Japanese translation experiments by appending the extracted fragments to a baseline system. For comparison, we also conducted translation experiments by appending the extracted comparable sentences. For decoding, we used the state–of–the–art phrase–based SMT toolkit Moses (Koehn et al., 2007) with default options, except for the distortion limit (6→20). The baseline system used the parallel corpus (680k sentences). We used another 368 and 367 sentences from the chemistry domain for tuning and testing respectively. We trained a 5–gram language model on the Japanese side of the parallel corpus using the SRILM toolkit[8].

[8]http://www.speech.sri.com/projects/srilm

System	BLEU
Baseline	38.64
+Sentences	39.16
+Munteanu+, 2006	38.87
+Only (IBM Model 1)	38.86
+Only (LLR)	$39.27^{\dagger}$
+Only (SampLEX)	$39.28^{\dagger}$
+External (IBM Model 1)	$\mathbf{39.63}^{\ddagger *}$
+External (LLR)	39.22
+External (SampLEX)	$39.40^{\dagger}$

Table 2: Results for Chinese–to–Japanese translation experiments ("†" and "‡"denotes the result is better than "Baseline" significantly at $p < 0.05$ and $p < 0.01$ respectively, "$*$" denotes the result is better than "+Munteanu+, 2006" significantly at $p < 0.05$).

Translation results evaluated on BLEU–4, are shown in Table 2. We can see that appending the extracted comparable sentences have a positive effect on translation quality. Adding the fragments extracted by (Munteanu and Marcu, 2006) has a negative impact, compared to appending the sentences. Our proposed method outperforms both "+sentences" and "Munteanu+, 2006", which indicates the effectiveness of our proposed method for extracting useful parallel fragments for MT.

We compared the phrase tables produced by different methods to investigate the reason for different MT performance. We found that all the methods increased the size of phrase table, meaning that new phrases were acquired from the extracted data. However, the noise contained in the data extracted by "+sentences" and "Munteanu+, 2006" produced many noisy phrase pairs, which may decrease MT performance. Our proposed method extracted accurate parallel fragments, which led to correct new phrases. Among all the settings of our proposed method, "+External (IBM Model 1)" showed the best performance. The reason for this is that it extracted more correct parallel fragments than the other settings, thus more new phrase pairs were produced.

Surprisingly, the translation performance after appending the fragments extracted by our proposed method only using the comparable sentences for alignment shows comparable results when using LLR and SampLEX lexicon for filtering, compared to the ones using the external parallel data for alignment. We think the reason is that the extracted fragments not only can produce new phrases, but also can improve the quality of phrase pairs extracted from the original parallel corpus. Because the fragments extracted only using the comparable sentences are more accurate than the ones using the external parallel data, they are more helpful to extract good phrase pairs from the original parallel corpus. This result indicates that external parallel data is not indispensable for the alignment model of our proposed method.

5 Conclusion and Future Work

In this paper, we proposed an accurate parallel fragment extraction system using alignment model together with translation lexicon. Experiments conducted on Chinese–Japanese data showed that our proposed method significantly outperforms a state–of–the–art approach and improves MT performance.

Our system can be improved in several aspects. Firstly, we only use IBM models for parallel fragment candidate detection, alignment models such as the ones proposed by (Quirk et al., 2007) could be more effective. Secondly, currently our proposed method cannot deal with ordering, an alignment model that is effective for ordering even on comparable sentences should be developed. Thirdly, although the experimental results indicate that external parallel data is not indispensable for the alignment model, we still use a parallel corpus for comparable sentence selection and lexicon filtering. An alternative method is constructing a large bilingual dictionary from comparable corpora, and use it for comparable sentence selection and lexicon filtering. Finally, although our proposed method is designed to be language and domain independent, the effectiveness for other language pairs and domains needs to be verified.

References

Sadaf Abdul-Rauf and Holger Schwenk. 2011. Parallel sentence generation from comparable corpora for improved smt. *Machine Translation*, 25(4):341–375.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Association for Computational Linguistics*, 19(2):263–312.

Chenhui Chu, Toshiaki Nakazawa, Daisuke Kawahara, and Sadao Kurohashi. 2012. Exploiting shared Chi-

nese characters in Chinese word segmentation optimization for Chinese-Japanese machine translation. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT'12)*, Trento, Italy, May.

Pascale Fung and Percy Cheung. 2004. Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. In *Proceedings of Coling 2004*, pages 1051–1057, Geneva, Switzerland, Aug 23–Aug 27. COLING.

Sanjika Hewavitharana and Stephan Vogel. 2011. Extracting parallel phrases from comparable data. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 61–68, Portland, Oregon, June. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Sadao Kurohashi, Toshihisa Nakamura, Yuji Matsumoto, and Makoto Nagao. 1994. Improvements of Japanese morphological analyzer JUMAN. In *Proceedings of the International Workshop on Sharable Natural Language*, pages 22–28.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504, December.

Dragos Stefan Munteanu and Daniel Marcu. 2006. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 81–88, Sydney, Australia, July. Association for Computational Linguistics.

Chris Quirk, Raghavendra Udupa U, and Arul Menezes. 2007. Generative models of noisy translations with applications to parallel fragment extraction. In *In Proceedings of MT Summit XI, European Association for Machine Translation*.

Jason Riesa and Daniel Marcu. 2012. Automatic parallel fragment extraction from noisy data. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 538–542, Montréal, Canada, June. Association for Computational Linguistics.

Jason R. Smith, Chris Quirk, and Kristina Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 403–411, Los Angeles, California, June. Association for Computational Linguistics.

Christoph Tillmann. 2009. A beam-search extraction algorithm for comparable data. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 225–228, Suntec, Singapore, August. Association for Computational Linguistics.

Masao Utiyama and Hitoshi Isahara. 2003. Reliable measures for aligning japanese-english news articles and sentences. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 72–79, Sapporo, Japan, July. Association for Computational Linguistics.

Ivan Vulić and Marie-Francine Moens. 2012. Subcorpora sampling with an application to bilingual lexicon extraction. In *Proceedings of COLING 2012*, pages 2721–2738, Mumbai, India, December. The COLING 2012 Organizing Committee.

Ivan Vulić, Wim De Smet, and Marie-Francine Moens. 2011. Identifying word translations from comparable corpora using latent topic models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 479–484, Portland, Oregon, USA, June. Association for Computational Linguistics.

Bing Zhao and Stephan Vogel. 2002. Adaptive parallel sentences mining from web abilingual news collections. In *Proceedings of the 2002 IEEE International Conference on Data Mining*, pages 745–748, Maebashi City, Japan. IEEE Computer Society.