Constructing a Chinese–Japanese Parallel Corpus from Wikipedia

Chenhui Chu, Toshiaki Nakazawa, Sadao Kurohashi

Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
E-mail: {chu, nakazawa}@nlp.ist.i.kyoto-u.ac.jp, kuro@i.kyoto-u.ac.jp

Abstract

Parallel corpora are crucial for statistical machine translation (SMT). However, they are quite scarce for most language pairs, such as Chinese–Japanese. As comparable corpora are far more available, many studies have been conducted to automatically construct parallel corpora from comparable corpora. This paper presents a robust parallel sentence extraction system for constructing a Chinese–Japanese parallel corpus from Wikipedia. The system is inspired by previous studies that mainly consist of a parallel sentence candidate filter and a binary classifier for parallel sentence identification. We improve the system by using the common Chinese characters for filtering and two novel feature sets for classification. Experiments show that our system performs significantly better than the previous studies for both accuracy in parallel sentence extraction and SMT performance. Using the system, we construct a Chinese–Japanese parallel corpus with more than 126k highly accurate parallel sentences from Wikipedia. The constructed parallel corpus is freely available at `http://orchid.kuee.kyoto-u.ac.jp/~chu/resource/wiki_zh_ja.tgz`.

Keywords: Chinese–Japanese, Parallel Corpus, Wikipedia

1. Introduction

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

Many studies have been conducted on constructing parallel corpora from comparable corpora, such as bilingual news articles (Zhao and Vogel, 2002; Utiyama and Isahara, 2003; Munteanu and Marcu, 2005; Tillmann, 2009; Abdul-Rauf and Schwenk, 2011) and patent data (Utiyama and Isahara, 2007; Lu et al., 2010). Recently, some researchers try to construct parallel corpora from Wikipedia (Smith et al., 2010; Ştefănescu and Ion, 2013).

While most studies are interested in language pairs between English and other languages, we focus on Chinese–Japanese, where parallel corpora are very scarce. This paper describes our efforts to improve a parallel sentence extraction system for constructing a Chinese–Japanese parallel corpus from Wikipedia. The system is inspired by (Munteanu and Marcu, 2005) and (Chu et al., 2013b), which both mainly consist of a parallel sentence candidate filter and a binary classifier for parallel sentence identification. The main contributions of this paper are in two aspects:

- Using common Chinese characters[1] (Chu et al., 2013a) for the filter to solve the domain dependent problem caused by the lack of an open domain dictionary.

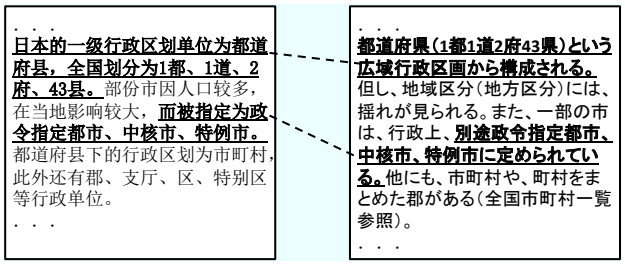

Figure 1: Example of aligned Chinese (left) and Japanese (right) article pairs via interlanguage links from Wikipedia, both desribe the topic of "Japan" (Parallel sentences are linked with dashed lines).

- Improving the classifier by introducing two novel feature sets.

Experiments show that our system performs significantly better than the previous studies for both accuracy in sentence extraction and SMT performance. Using the system, we construct a large scale Chinese–Japanese parallel corpus with more than 126k highly accurate parallel sentences from Wikipedia.

2. Chinese–Japanese Wikipedia

Wikipedia[2] is a free, collaborative and multilingual encyclopedia. Chinese and Japanese Wikipedia are in the top 20 language editions of Wikipedia, with more than 740k and 887k articles respectively (24th December 2013).

A special characteristic of Wikipedia is that article alignment is established via interlanguage links. As parallel sentences trend to appear in similar article pairs, Wikipedia can be a valuable resource for constructing parallel corpora. Figure 2 shows an example of aligned article pairs via interlanguage links from Chinese and Japanese Wikipedia, where there are parallel sentences. Our task is to identify the parallel sentences from the aligned article pairs.

[1] Common Chinese characters can be seen as a kind of cognates.

[2] http://en.wikipedia.org/wiki/Wikipedia

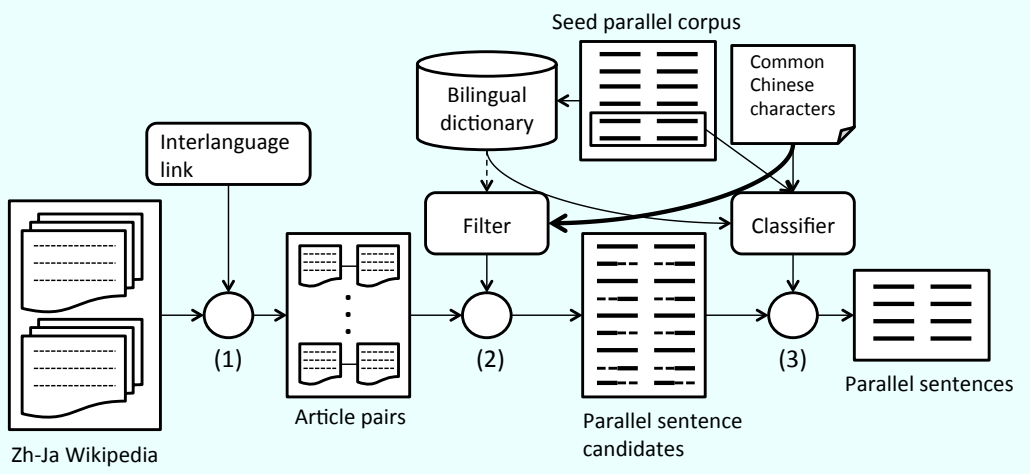

Figure 2: Parallel sentence extraction system.

3. Parallel Sentence Extraction System

The overview of our parallel sentence extraction system is presented in Figure 2. We first align articles on the same topic in Chinese and Japanese Wikipedia via the interlanguage links ((1) in Figure 2). Then we generate all possible sentence pairs by the Cartesian product from the aligned articles, and discard the pairs that do not fulfill the conditions of a filter to reduce the candidates keeping more reliable sentences ((2) in Figure 2). Finally, we use a classifier trained on a small number of parallel sentences from a seed parallel corpus to identify the parallel sentences from the candidates ((3) in Figure 2).

Our system differs from previous studies on the strategy of the filter and the features used for the classifier, which will be described in Section 3.1. and Section 3.2. in detail.

3.1. Parallel Sentence Candidate Filtering

A parallel sentence candidate filter is necessary, because it can remove most of the noise introduced by the simple Cartesian product sentence generator, and reduce computational cost for parallel sentence identification. Previous studies use a filter with sentence length ratio and dictionary–based word overlap conditions (Munteanu and Marcu, 2005; Chu et al., 2013b). Although the sentence length ratio condition is domain independent, the word overlap condition is domain dependent[3]. Wikipedia is an open domain database, so using a domain dependent condition for filtering may decrease the performance of our system. In the scenario that an open domain dictionary is unavailable, do we have any alternatives that are robust against the domain diversity and effective to filter the noise?

3.1.1. Common Chinese Characters

Different from other language pairs, Chinese and Japanese share Chinese characters. In Chinese the Chinese characters are called Hanzi, while in Japanese they are called Kanji. Hanzi can be divided into two groups, Simplified Chinese (used in mainland China and Singapore) and Traditional Chinese (used in Taiwan, Hong Kong and Macau). The number of strokes needed to write characters has been largely reduced in Simplified Chinese, and the shapes may be different from those in Traditional Chinese. Because Kanji characters originated from ancient China, many common Chinese characters exist between Hanzi and Kanji. Chu et al., (2012b) created a Chinese character mapping table between Traditional Chinese, Simplified Chinese and Japanese[4]. Table 1 gives some examples of common Chinese characters in that mapping table with their Unicode.

Meaning	snow	love	begin
TC	雪 (U+96EA)	愛 (U+611B)	發 (U+767C)
SC	雪 (U+96EA)	爱(U+7231)	发(U+53D1)
Kanji	雪 (U+96EA)	愛 (U+611B)	発 (U+767A)

Table 1: Examples of common Chinese characters (TC denotes Traditional Chinese and SC denotes Simplified Chinese).

Since Chinese characters contain significant semantic information, and common Chinese characters share the same meaning, they can be valuable linguistic clues for many Chinese–Japanese SMT tasks, such as phrase alignment (Chu et al., 2011) and bilingually motivated word segmentation (Chu et al., 2012a). Also, they can be powerful linguistic clues to identify parallel sentences. Chu et al., (2013b) proposed using common Chinese character features for a parallel sentence classifier, which significantly improve the accuracy of classification.

3.1.2. Common Chinese Characters for Filtering

Since common Chinese characters are domain independent and effective to filter the noise introduced by the simple Cartesian product sentence generator, here we propose to use them for the filter. We compare 4 different filtering strategies: dictionary–based word overlap (Word), common

[3]Because the dictionary is automatically generated using a word alignment tool from a seed parallel corpus, which is domain specific.

[4]http://nlp.ist.i.kyoto-u.ac.jp/member/chu/pubdb/LREC2012/kanji_mapping_table.txt

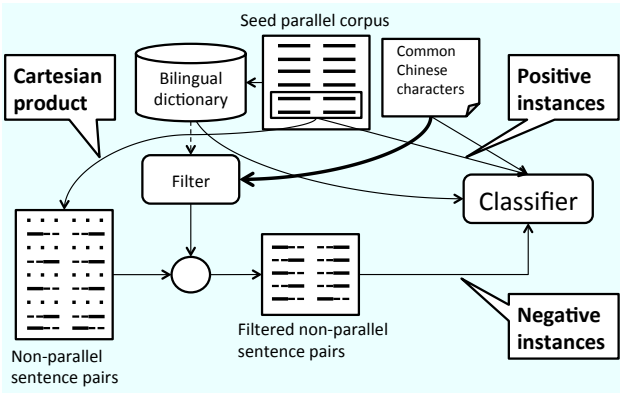

Figure 3: Parallel sentence classifier.

Chinese character overlap (CC) and their logical combination. We define:

- Word filter: Using dictionary–based word overlap.
- CC filter: Using common Chinese character overlap[5].
- Word and CC filter: Using logical conjunction of word and common Chinese character overlaps.
- Word or CC filter: Using logical disjunction of word and common Chinese character overlaps.

We report the performance of using different filtering strategies in Section 4.3..

3.2. Parallel Sentence Identification by Binary Classification

Since the quality of the extracted sentences is determined by the accuracy of the classifier, the classifier becomes the core component of the extraction system. In this section, we first describe the training and testing process, then introduce the features we use for the classifier.

3.2.1. Training and Testing

We use a Support Vector Machine (SVM) classifier. Training and testing instances for the classifier are created following the method of (Munteanu and Marcu, 2005). We use a small number of parallel sentences from a seed parallel corpus as positive instances. Negative instances are generated by the Cartesian product of the positive instances excluding the original positive instances, and they are filtered by the same filtering method used in Section 3.1.. Moreover, we randomly discard some negative instances for training when necessary[6], to guarantee that the ratio of negative to positive instances is less than five for the performance of the classifier. Figure 3 illustrates this process.

3.2.2. Features

Basic Features. The basic features are the ones proposed in (Munteanu and Marcu, 2005):

- Sentence length, length difference and length ratio.
- Word overlap: Percentage of words on each side that have a translation on the other side (according to the dictionary).
- Word alignment features extracted from the word alignment results of the sentences used as instances for the classifier.

Chinese Character Features. The Chinese character features are the ones proposed in (Chu et al., 2013b):

- Number of Chinese characters on each side.
- Percentage of Chinese characters out of all characters on each side.
- Ratio of Chinese character numbers on both sides.
- Number of common Chinese character n–grams (from 1–gram to 4–gram).
- Percentage of common Chinese character n–grams out of all Chinese character n–grams on each side.

Inspired by the above features, we propose two novel feature sets.

Non–CC Word Features. Chinese–Japanese parallel sentences often contain alignable words that do not consist Chinese characters, such as foreign words, numbers and punctuation etc., which we call Non–Chinese character (Non–CC) words. Note that for Non-CC words, we do not consider Japanese kana. Non–CC words can be helpful clues to identify parallel sentences. We use the following features:

- Number of Non–CC words on each side.
- Percentage of Non–CC words out of all words on each side.
- Ratio of Non–CC word numbers on both sides.
- Number of same Non–CC words.
- Percentage of same Non–CC words out of all Non–CC words on each side.

Content Word Features. The word overlap feature proposed in (Munteanu and Marcu, 2005) has the problem that function words and content words are handled in the same way. Function words often have a translation on the other side, thus erroneous parallel sentence pairs with a few content word translations are often produced by the classifier. Therefore, we add the following content word features:

- Percentage of content words out of all words on each side.
- Percentage of content words on each side that have a translation on the other side (according to the dictionary).

We determine a word as content or function word using a predefined part–of–speech (POS) tag sets of function words for Chinese and Japanese respectively.

[5]We used 1–gram common Chinese character overlap with threshold of 0.1 for Chinese and 0.3 for Japanese.

[6]Note that we keep all negative instances for testing.

4. Experiments

We evaluated classification accuracy, and conducted extraction and translation experiments to verify the effectiveness of our proposed parallel sentence extraction system. In all our experiments, we preprocessed the data by segmenting Chinese and Japanese sentences using a segmenter proposed by Chu et al. (2012a) and JUMAN (Kurohashi et al., 1994) respectively.

4.1. Data

The seed parallel corpus we used is the Chinese–Japanese part of ASPEC[7] (Asian Scientific Paper Excerpt Corpus), which is provided by JST[8] and NICT[9]. 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).
Also, we downloaded Chinese[10] (20120921) and Japanese[11] (20120916) Wikipedia database dumps. We used an open–source Python script[12] to extract and clean the text from the dumps. Since the Chinese dump is mixed of Traditional and Simplified Chinese, we converted all Traditional Chinese to Simplified Chinese using a conversion table published by Wikipedia[13]. We aligned the articles on the same topic in Chinese and Japanese via the interlanguage links, obtaining 162k article pairs (2.1M Chinese and 3.5M Japanese sentences respectively).

4.2. Classification Accuracy Evaluation

We evaluated classification accuracy using two distinct sets of 5k parallel sentences from the seed parallel corpus for training and testing respectively. We report the results using word overlap for filtering, for easier comparison to previous studies.

4.2.1. Settings

- Word alignment tool: GIZA++[14].
- Dictionary: Top 5 translations with translation probability larger than 0.1 created from the seed parallel corpus.
- Classifier: LIBSVM[15] with 5–fold cross–validation and radial basis function (RBF) kernel.
- Sentence length ratio threshold: 2.
- Word overlap threshold: 0.25.
- Classifier probability threshold: 0.9.

[7] http://orchid.kuee.kyoto-u.ac.jp/ASPEC
[8] http://www.jst.go.jp
[9] http://www.nict.go.jp
[10] http://dumps.wikimedia.org/zhwiki
[11] http://dumps.wikimedia.org/jawiki
[12] http://code.google.com/p/recommend-2011/source/browse/Ass4/WikiExtractor.py
[13] http://svn.wikimedia.org/svnroot/mediawiki/branches/REL1_12/phase3/includes/ZhConversion.php
[14] http://code.google.com/p/giza-pp
[15] http://www.csie.ntu.edu.tw/~cjlin/libsvm

Features	Precision	Recall	F–measure
Munteanu+ 2005	96.65	83.56	89.63
Chu+ 2013	97.05	93.52	95.25
+Non–CC	97.38	93.64	95.47
+Content	**98.34**	**95.94**	**97.12**

Table 2: Classification results.

4.2.2. Results

We compared the following features:

- Munteanu+ 2005: Only using the features proposed in (Munteanu and Marcu, 2005).
- Chu+ 2013: Further using the Chinese character features proposed in (Chu et al., 2013b).
- +Non–CC: Further using the Non–CC word features.
- +Content: Further using the content word features.

Results are shown in Table 2. We can see that our proposed Non–CC word and content word overlap features further improve the accuracy.

4.3. Extraction and Translation Experiments

We extracted parallel sentences from Wikipedia and evaluated Chinese–to–Japanese MT performance using the extracted sentences as training data.

4.3.1. Settings

- Tuning and testing: Two distinct sets of 198 parallel sentences. These sentences were randomly selected from the sentence pairs extracted from Wikipedia by our system with different methods, and the erroneous parallel sentences were manually discarded[16]. Note that the sentences in the tuning and testing sets are not included in the training data.
- Decoder: Moses (Koehn et al., 2007) with default options, except for the distortion limit (6→20).
- Language model: 5–gram LM trained on the Japanese Wikipedia (12.5M sentences) using SRILM toolkit[17].

The other settings are the same as the ones used in the classification experiments.

4.3.2. Results

Parallel sentence extraction and translation results using different methods are shown in Table 3. "Munteanu+ 2005", "Chu+ 2013" "+Non–CC" and "+Content" denote the different features described in Section 4.2.2.. "Word", "CC", "Word and CC" and "Word or CC" denote the 4 different filtering strategies described in Section 3.1.2..
We can see that our proposed Non–CC word and content word features improve MT performance significantly. "CC filter" shows better performance than "Word filter", which

[16] To get the 396 sentences for tuning and testing, 404 sentences were manually discarded.
[17] http://www.speech.sri.com/projects/srilm

Features	Filter	# extracted sentences	BLEU–4(dev)	BLEU–4(test)
Munteanu+ 2005	Word	122,569	36.90	35.18
Chu+ 2013	Word	146,797	38.31	$36.27^{\dagger}$
+Non–CC	Word	161,046	38.00	$36.79^{\dagger}$
+Content	Word	164,993	38.98	$37.39^{\dagger\ddagger}$
+Non–CC	CC	115,985	39.18	$37.70^{\dagger\ddagger}$
+Content	CC	126,811	**39.40**	$\mathbf{37.82}^{\dagger\ddagger}$
+Non–CC	Word and CC	78,962	37.47	35.36
+Content	Word and CC	80,598	36.95	36.14
+Non–CC	Word or CC	178,085	38.76	$36.53^{\dagger}$
+Content	Word or CC	184,103	38.70	$36.41^{\dagger}$

Table 3: Parallel sentence extraction and translation results ("†" and "‡" denote that the result is significantly better than "Munteanu+ 2005" and "Chu+ 2013" respectively at $p < 0.05$).

indicates that for open domain data such as Wikipedia, using common Chinese characters for filtering is more effective than a domain specific dictionary. "Word and CC filter" decreases the performance, because the number of extracted sentences decreases significantly. "Word or CC filter" also shows bad performance, and we suspect the reason is the increase of erroneous parallel sentence pairs.

For the best method of "+Content with CC filter", we manually estimated 100 sentence pairs that were randomly selected from the extracted sentences. We found that 64% of them are actual translation equivalents, while the other erroneous parallel sentences only contain little noise.

5. Related Work

As parallel sentences trend to appear in similar article pairs, many studies first conduct article alignment from comparable corpora, then identify the parallel sentences from the aligned article pairs (Utiyama and Isahara, 2003; Fung and Cheung, 2004; Munteanu and Marcu, 2005). This study extracts parallel sentences from Wikipedia. Wikipedia is a special type of comparable corpora, because article alignment is established via interlanguage links. Approaches without article alignment also have been proposed (Tillmann, 2009; Abdul-Rauf and Schwenk, 2011; Ştefănescu et al., 2012; Chu et al., 2013b). These studies directly retrieve candidate sentence pairs, and select the parallel sentences using some filtering methods.

The way of parallel sentence identification can be specified with two different approaches: binary classification (Munteanu and Marcu, 2005; Tillmann, 2009; Smith et al., 2010; Ştefănescu et al., 2012; Chu et al., 2013b) and translation similarity measures (Utiyama and Isahara, 2003; Fung and Cheung, 2004; Abdul-Rauf and Schwenk, 2011). We adopt the binary classification approach with novel features sets.

Few studies have been conducted for extracting parallel sentences from Wikipedia (Smith et al., 2010; Ştefănescu and Ion, 2013). Previous studies are interested in language pairs between English and other languages such as German and Spanish. We focus on Chinese–Japanese, where parallel corpora are very scarce.

6. Conclusion

In this paper, we improved a parallel sentence extraction system by using the common Chinese characters for filtering and two novel feature sets for classification. The Experimental results on Wikipedia showed that our proposed methods are more effective than the previous studies.

The parallel sentences extracted by the method of "+Content with CC filter" which showed the best performance in Section 4.3., and the tuning and testing sets used in the translation experiments are available at `http://orchid.kuee.kyoto-u.ac.jp/~chu/resource/wiki_zh_ja.tgz`.

7. Acknowledgements

The first author is supported by Hattori International Scholarship Foundation[18]. We also thank the anonymous reviewers for their valuable comments.

8. 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, and Sadao Kurohashi. 2011. Japanese-chinese phrase alignment using common chinese characters information. In *Proceedings of MT Summit XIII*, pages 475–482, Xiamen, China, September.

Chenhui Chu, Toshiaki Nakazawa, Daisuke Kawahara, and Sadao Kurohashi. 2012a. Exploiting shared Chinese 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 2012)*, pages 35–42, Trento, Italy, May.

Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. 2012b. Chinese characters mapping table of Japanese,

[18] http://www.hattori-zaidan.or.jp

Traditional Chinese and Simplified Chinese. In *Proceedings of the Eighth Conference on International Language Resources and Evaluation (LREC 2012)*, pages 2149–2152, Istanbul, Turkey, May.

Chenhui Chu, Toshiaki Nakazawa, Daisuke Kawahara, and Sadao Kurohashi. 2013a. Chinese-japanese machine translation exploiting chinese characters. *ACM Transactions on Asian Language Information Processing (TALIP)*, 12(4):16:1–16:25, October.

Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. 2013b. Chinese–japanese parallel sentence extraction from quasi–comparable corpora. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 34–42, Sofia, Bulgaria, August. Association for Computational Linguistics.

Dan Ştefănescu and Radu Ion. 2013. Parallel-wiki: A collection of parallel sentences extracted from wikipedia. In *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2013)*, pages 117–128, Samos, Greece, March.

Dan Ştefănescu, Radu Ion, and Sabine Hunsicker. 2012. Hybrid parallel sentence mining from comparable corpora. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT 2012)*, pages 137–144, 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.

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.

Bin Lu, Tao Jiang, Kapo Chow, and Benjamin K. Tsou. 2010. Building a large english-chinese parallel corpus from comparable patents and its experimental application to smt. In *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC 2010*, pages 42–49, Valletta, Malta, May.

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

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.

Masao Utiyama and Hitoshi Isahara. 2007. A japanese-english patent parallel corpus. In *Proceedings of MT summit XI*, pages 475–482.

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.