Efficient Extraction of Pseudo-Parallel Sentences from Raw Monolingual Data Using Word Embeddings

Benjamin Marie **Atsushi Fujita**
National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Souraku-gun, Kyoto, 619-0289, Japan
{bmarie, atsushi.fujita}@nict.go.jp

Abstract

We propose a new method for extracting *pseudo-parallel* sentences from a pair of large monolingual corpora, without relying on any document-level information. Our method first exploits word embeddings in order to efficiently evaluate trillions of candidate sentence pairs and then a classifier to find the most reliable ones. We report significant improvements in domain adaptation for statistical machine translation when using a translation model trained on the sentence pairs extracted from in-domain monolingual corpora.

1 Introduction

Parallel corpus is an indispensable resource for statistical and neural machine translation. Generally, using more sentence pairs to train a translation system makes it able to produce better translations. However, for most language pairs and domains, parallel corpora remain scarce due mainly to the cost of their creation (Germann, 2001).

In the last two decades, numerous methods have been proposed to extract parallel sentences from comparable corpora. In addition to comparable corpora in large quantity, to the best of our knowledge, all previous methods heavily rely on document-level information and/or lexical translation models, such as those for statistical machine translation (SMT) systems (Zhao and Vogel, 2002; Fung and Cheung, 2004; Munteanu and Marcu, 2005; Tillmann and Xu, 2009) and manually-created bilingual lexicon (Utiyama and Isahara, 2003). The most successful approaches use cross-lingual information retrieval techniques (Abdul Rauf and Schwenk, 2011; Ștefănescu et al., 2012) to extract sentence pairs from comparable documents. Using such document pairs has the strong advantage that it drastically reduces the search space; we need to consider only sentence pairs in each document pair instead of scoring all sentence pairs in the two monolingual corpora. However, in many cases, we do not have access to document-level information. Only Tillmann and Xu (2009) have explored this scenario using efficient caching strategies to extract useful sentence pairs from nearly one trillion candidates in comparable data. Yet, their approach is tightly related to the exploitation of accurate lexical translation models and does not allow us to introduce other features. The reliance on lexical translation models implies that we must have already access to parallel data sufficiently large for obtaining accurate estimates. Nevertheless, the most useful sentence pairs for SMT are actually the ones that contain infrequent or even unseen tokens in these parallel data. Relying only on lexical translation models thus seems rather inadequate to extract sentence pairs containing numerous infrequent or unseen tokens, and may actually be more prone to extract sentence pairs that contain words and phrases for which we already have accurate translation probability estimates.

This paper proposes a new method that exploits word embeddings to efficiently extract *pseudo-parallel* sentences[1] from raw monolingual data without using any document-level information. We report significant improvements of translation quality in a domain adaptation scenario for SMT.

2 Sentence pair extraction

During the sentence pair extraction, we do not assume an access to document-level information. Our method thus has to be efficient in evaluating

[1] As in previous work, we regard the sentence pairs extracted by our method as "pseudo-parallel" because they are not necessarily parallel. As shown by Goutte et al. (2012), even very noisy parallel corpora may be useful for SMT.

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 392–398
Vancouver, Canada, July 30 - August 4, 2017.
https://doi.org/10.18653/v1/P17-2062

trillions of sentence pairs hypothesized from two monolingual corpora, each containing millions of sentences. To achieve this computationally challenging task, we need a fast way to compute some similarity between the source and target sentences, without relying on large lexical translation models that may not be available or accurate enough in some low-resourced conditions.

2.1 Step 1: Filtering with sentence embeddings

Assuming the availability of large-scale monolingual data, our method exploits word embeddings (Mikolov et al., 2013b) that are fast to estimate. First, word embeddings for each language are learned from the given monolingual data. This enables us to evaluate arbitrary sentence pair given all the words it contains, which is not fully guaranteed by a lexical translation model as some tokens may be out-of-vocabulary (OOV). We then proceed to the projection of all the source word embeddings to the target embedding space, following Mikolov et al. (2013a),[2] in order to represent both source and target words in the same space.

To compute the similarity between arbitrary sentence pairs, we represent each sentence by averaging the embeddings of its constituent words.[3]

As a result of this first step, our method keeps for each source sentence the n closest target sentences (n being small, for instance with a value of 100) according to the similarity score.

2.2 Step 2: Refining with a classifier

Given a far smaller search space, this second step evaluates and re-ranks the remaining sentence pairs, incorporating more complex features to train a classifier. We use a total of five features.

For each sentence pair, we use the score computed in the first step and a more accurate similarity score based on alignments between word embeddings, following the work in Kajiwara and Komachi (2016). They found out that the average of the cosine similarity between all the best word pairs, for each source word, taken from the sentence pair, shown in Eq. (1), was a good indicator of similarity between two sentences.

$$S(\mathbf{x},\mathbf{y}) = \frac{1}{|\mathbf{x}|}\sum_{i=1}^{|\mathbf{x}|}\max_{j}\phi(x_i^{emb}, y_j^{emb}) \quad (1)$$

where $\mathbf{x}$ and $\mathbf{y}$ are respectively the source and target sentences, $|\mathbf{x}|$ the length of $\mathbf{x}$, and ϕ the cosine similarity between the embeddings in the target language space of the i-th word in $\mathbf{x}$, i.e., x_i^{emb}, and the j-th word in $\mathbf{y}$, i.e., y_j^{emb}. The computation of this score can be highly costly, depending on the sentence length and the number of dimensions of the word embeddings. Thus, we compute this score only for the source to target direction, unlike Kajiwara and Komachi (2016).

In many situations, we may also have an access to a lexical translation model trained on some parallel data. We therefore incorporate the scores proposed by Tillmann and Xu (2009), but considering one probability for each translation direction, instead of summing them up, so that our classifier can optimize their weight separately.

$$P(\mathbf{x}|\mathbf{y}) = \sum_{i=1}^{|\mathbf{x}|}\frac{1}{|\mathbf{x}|}\log\left(\frac{1}{|\mathbf{y}|}\sum_{j=1}^{|\mathbf{y}|}p(x_i^{tok}|y_j^{tok})\right) \quad (2)$$

$$P(\mathbf{y}|\mathbf{x}) = \sum_{j=1}^{|\mathbf{y}|}\frac{1}{|\mathbf{y}|}\log\left(\frac{1}{|\mathbf{x}|}\sum_{i=1}^{|\mathbf{x}|}p(y_j^{tok}|x_i^{tok})\right) \quad (3)$$

where x_i^{tok} is the i-th token in $\mathbf{x}$, y_j^{tok} the j-th token in $\mathbf{y}$ and p the probability given by an already estimated lexical translation model.

Our last feature is the length ratio of the source and target sentences (Munteanu and Marcu, 2005).

To assign a real-valued score to each sentence pair in order to filter and rank them, we train a Maximum Entropy (ME) classifier, following Munteanu and Marcu (2005). ME classifier suits particularly well our situation, since we deal with a small number of dense features and have hundred millions of sentence pairs to classify quickly.

Positive examples for training the classifier can be obtained straightforwardly: we use true sentence pairs sampled from parallel data, different from the one used to train the lexical translation model. As for negative examples, Munteanu and Marcu (2005) randomly paired sentences from

[2] Despite the availability of more accurate methods (Coulmance et al., 2015; Duong et al., 2016) we choose this method considering its low computational cost and its reasonable need of external resources to estimate the translation matrix, i.e., only a small bilingual dictionary.

[3] As shown by Adi et al. (2016), this can be effective to encode sentence-level information such as content and length, while being computationally more efficient than other methods, such as inducing paragraph vectors (Le and Mikolov, 2014) and using LSTM auto-encoders (Li et al., 2015). Our decision also relies on the promising accuracy of linear projection of word (not sentence) embeddings across different languages (Mikolov et al., 2013a).

their parallel data using two constraints: a length ratio not greater than two, and a coverage constraint that considers a negative example only if more than half of the words of the source sentence has a translation in the given target sentence according to some bilingual lexicon. However, from a large parallel corpus, one can easily retrieve another target sentence, almost identical, containing most of the words that the true target sentence also contains. In this case, the negative example will be almost as semantically close as the positive one, weakening the discriminative power of the features based on word embeddings. To circumvent this problem, we generate negative examples, as many as positive examples, without using this coverage constraint.

Having assigned a score for each sentence pair, we make a pseudo-parallel corpus selecting the target sentence with the best score for each source sentence and retaining only the sentence pairs with a score above some threshold, th. This pseudo-parallel corpus can then be used to train a new phrase table.

3 Experiments

We evaluated our method in a scenario of domain adaptation for phrase-based SMT (PBSMT). In this scenario, we assumed a lot of general-domain parallel data to train a general-domain phrase table and a lot of in-domain monolingual data as our source of in-domain pseudo-parallel sentences.

3.1 Data and SMT system

We experimented with the French–English language pair, both translation directions, on the medical domain. We used `Moses` (Koehn et al., 2007) to train, tune, and test our PBSMT systems. The general-domain phrase table was trained on Europarl V7[4] (1.99M sentences). The in-domain monolingual data were prepared by applying the `NLTK`[5] sentence segmenter to the concatenation of all the monolingual corpora provided for the WMT'14 medical translation task.[6] As the source of extracting in-domain sentence pairs, we randomly sampled 1M sentences (33M tokens) from the French data and 5M sentences (164M tokens) from the English data. Given pseudo-parallel sentences extracted by our method from these data (see Section 3.2), we trained an in-domain phrase table. `Moses` exploits the two phrase tables, i.e., general-domain and in-domain ones, with its multiple decoding path ability. The PBSMT systems used one language model trained on the entire target in-domain monolingual data concatenated to the target side of Europarl and News Crawl data provided by WMT'15.[7] The development and test data used to tune and evaluate the PBSMT systems were excerpts of the EMEA parallel corpus (Carpuat et al., 2012).

3.2 Parameters for sentence pair extraction

We used `word2vec`[8] to learn word embeddings with the parameters `-cbow 1 -window 10 -negative 15 -sample 1e-4 -iter 15 -min-count 1`, specifying 800 and 300 dimensions for the source and target languages, respectively,[9] on the same data used to train the language models. The translation matrix used to project the source word embeddings to the target embedding space was trained on a bilingual lexicon containing the 5k[10] most frequent French tokens,[11] from Europarl, and their most probable single token in English given by the Europarl phrase table. The first step of our method evaluated five trillion (1M×5M) sentence pairs and retained the 100 closest target sentences for each source sentence.

The second step then dealt with only 100M (1M×100) sentence pairs. The lexical translation probabilities used to compute our features were given by the Europarl lexical translation models. We used `Scikit-learn`[12] to train the ME classifier, with default parameters, on 5k positive and 5k negative examples[13] randomly generated from the MultiUn corpus.[14] According to the classifier's score, only the 1-best target sentence for each source sentence was retained. We discarded sentence pairs having a score lower than a thresh-

[4] http://statmt.org/europarl/
[5] http://www.nltk.org/
[6] http://statmt.org/wmt14/medical-task/
[7] http://statmt.org/wmt15/
[8] http://word2vec.googlecode.com/
[9] Mikolov et al. (2013a) observed that a more accurate projection is obtained when using a greater number of dimensions on the source side than that for the target side.
[10] Vulić and Korhonen (2016) demonstrated that 5k word pairs is enough to train a useful translation matrix.
[11] We extracted sentence pairs regarding French and English as source and target languages, respectively, but used the resulted parallel corpus for both translation directions.
[12] http://scikit-learn.org/
[13] We chose this number empirically through observing the classification accuracy on a set of held-out sentence pairs.
[14] http://opus.lingfil.uu.se/MultiUN.php

System		Fr→En BLEU	Fr→En #OOV	En→Fr BLEU	En→Fr #OOV	#extracted pairs	speed (#pairs/sec)
Only general-domain phrase table		25.9	3,134	23.1	3,099	-	-
Baseline (Tillmann and Xu (2009))		27.2	2,729	24.7	2,661	121k	1.22M
Proposed method	($th = 0.955$)	**28.0**	2,607	**25.4**	2,533	121k	14.46M
	($th = 0.700$)	**28.6**	1,985	**26.4**	1,955	361k	
Proposed method w/ cov. constraint	($th = 0.600$)	26.1	3,064	23.2	3,077	11k	19.21M

Table 1: BLEU scores (Papineni et al., 2002) averaged over 3 tuning runs, obtained when added an in-domain phrase table to the system, created either by the baseline method or by our work with or without the coverage constraint activated (denoted "w/ cov. constraint"). Bold scores indicate statistical significance ($p < 0.01$) of the score over the baseline system, measured by approximate randomization using `MultEval` (Clark et al., 2011). We also present the number of OOV tokens in the test set and the number of sentence pairs actually used to train the in-domain phrase table. The speed of the method to evaluate sentence pairs from monolingual data was measured with 100 CPU threads (Xeon E5-2600) on 1 trillion sentence pairs randomly sampled.

old value. We examined $\{0.5, 0.6, 0.7, 0.8, 0.9\}$ as the threshold value through tuning PBSMT systems, and determined 0.7 to be optimal.

We regarded the method proposed by Tillmann and Xu (2009) as a baseline, because it does not rely on document-level information, as ours. Unlike our method, in addition to the constraint based on length ratio, this method also used the coverage constraint. As discussed in Section 2.2, this constraint speeds up the extraction, but sacrifices source sentences with numerous OOV due to its heavy reliance on a bilingual lexicon learned from parallel data. To measure the effect of the coverage constraint, we also activated it in some of our experiments using our method. Then, as for our method, we discarded sentence pairs having a score lower than a threshold value and found the threshold value of -10 to be the best among $\{$-15, -12, -10, -7$\}$.

3.3 Results

Table 1 presents the results. Both the baseline and our methods outperformed the system using only the general-domain phrase table in both translation directions. This may be explained by the presence of highly parallel sentences in the in-domain monolingual data, from Wikipedia articles for instance, that can be retrieved by both methods.

Our method significantly outperformed the baseline, with 1.4 and 1.7 BLEU points gains respectively for Fr→En and En→Fr. Our method, with the optimal threshold of 0.7, extracted 361k sentence pairs from the in-domain monolingual data, while the baseline method extracted only 121k sentence pairs due presumably to the use of the coverage constraint that might remove source sentences with a high OOV ratio. Less OOV tokens remained with the system using our method, highlighting the positive effect of exploiting word embeddings in addition to lexical translation models. Activating the coverage constraint on our method was harmful and was significantly worse than the baseline. This constraint excludes candidate sentence pairs by relying only on general-domain lexical translation models, while our classifier is trained to use word embeddings that are more robust but unhelpful to discriminate the remaining candidates. Therefore, the optimal threshold value allowed the extraction of only 11k sentence pairs. In contrast, without this constraint, even with a high threshold value of 0.955 that retrieved as many sentence pairs as the baseline method, the extracted sentence pairs resulted in a significantly higher BLEU score than the baseline method, with a slightly better lexical coverage. Last but not least, our method is 11.9 times faster than the baseline method.

4 Feature contribution

To evaluate the impact of the features used during classification, we performed a feature ablation experiment. The results for the EMEA translation

Feature set	*th*	Fr→En	En→Fr
all	0.7	28.6	26.4
-avg. emb.	0.7	28.8	26.1
-max. al. emb.	0.7	29.0	26.1
-max. al. emb.	0.8	28.4	25.6
-lex. prob.	0.8	28.3	26.0
-length	0.6	28.9	26.4

Table 2: Results (BLEU) obtained without using some of the features during the classification (see Section 2.2). The features removed, independently, are the following: averaged word embeddings (avg. emb.), maximum alignment between embeddings (max. al. emb.), lexical translation probabilities (lex. prob.) and the length ratio of the source and target sentences (length). The "*th*" column indicates the threshold value for the classifier's score above which we retain the sentence pairs. This value was selected among the values {0.5,0.6,0.7,0.8,0.9} with respect to the BLEU score on the development data, through the tuning of the PBSMT system, for each configuration.

task are reported in Table 2.

For both translation directions, the features that have the most important were the ones based on lexical translaiton probabilities and alignments between embeddings. For instance, in En→Fr translation, removing them led to a significant drop of 0.4 and 0.8 BLEU points, respectively.

For the Fr→En translation direction, surprisingly, we observed improvements on the test set for all configurations, except when removing either of the above two types of features. However, we did not observe such improvements for the En→Fr translation direction; removing any feature(s) consistently led to a lower or equal BLEU score. Feature ablation did not improve the performance on the development set for both translation directions, neither.

5 Classifier accuracy

To better understand the performance of our method, we also evaluated the accuracy of the classifier used in step 2 (see Section 2.2). Note that this evaluation does not intend to show how well the classifier retrieves useful pseudo-parallel sentences. We cannot directly evaluate it, as we do not have an evaluation data set that contains *gold pseudo-parallel sentences* at hand.

A set of in-domain *truly parallel sentences* was used for our evaluation. We selected the 50k first source sentences from the held-out in-domain EMEA parallel corpus,[15] and used each one of them to make two sentence pairs in order to obtain a positive and a negative example. For the positive example, the source sentence is associated to its correct translation from the EMEA corpus, while for the negative example, we associated the source sentence with a target sentence randomly extracted from the EMEA corpus. The classifier has then to decide if the sentence pair is correct or incorrect.

The classifier is the same one that was presented in Section 3.2 and trained on the MultiUn parallel data. On our EMEA evaluation data set, this classifier achieves an accuracy of 85.98%. This high accuracy highlights the potential of our method in retrieving highly, or truly, parallel sentences if such kinds of sentence pairs exist in the monolingual data exploited by our approach.

6 Conclusion and future work

We presented a method for extracting pseudo-parallel sentences from a pair of large monolingual corpora, without relying on any document-level information. Our domain adaptation experiments showed that our method outperformed the state-of-the-art method by more efficiently extracting more useful sentence pairs from in-domain monolingual data. In addition to the improved BLEU scores, our method provides a better handling of OOV, ignored by other methods that strongly rely on already trained lexical translation models.

Our method can further be speeded up by some approximation, such as local sensitive hashing, or by using a smaller number of dimensions for word embeddings. We leave the study of their impact to our future work. We believe that our work is also useful for other downstream tasks that need comparable or pseudo-parallel sentences, such as parallel phrase extraction (Hewavitharana and Vogel, 2016) and adaptation of neural machine translation systems (Luong and Manning, 2015; Freitag and Al-Onaizan, 2016).

Acknowledgments

We would like to thank all the reviewers for their valuable comments and suggestions.

[15] We used the EMEA training data provided by the same workshop on domain adaptation (Carpuat et al., 2012) that released the development and test data used in our experiments.

References

Sadaf Abdul Rauf and Holger Schwenk. 2011. Parallel sentence generation from comparable corpora for improved smt. *Machine Translation* 25(4):341–375. https://doi.org/10.1007/s10590-011-9114-9.

Yosshi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2016. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *Proceedings of ICLR 2017*. https://arxiv.org/pdf/1608.04207v3.pdf.

Marine Carpuat, Hal Daumé III, Alexander Fraser, Chris Quirk, Fabienne Braune, Ann Clifton, Ann Irvine, Jagadeesh Jagarlamudi, John Morgan, Majid Razmara, Aleš Tamchyna, Katharine Henry, and Rachel Rudinger. 2012. Domain adaptation in machine translation: Final report. In *2012 Johns Hopkins Summer Workshop Final Report*. http://hal3.name/damt/.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of ACL-HLT*. Portland, Oregon. http://aclweb.org/anthology/P11-2031.

Jocelyn Coulmance, Jean-Marc Marty, Guillaume Wenzek, and Amine Benhalloum. 2015. Trans-gram, fast cross-lingual word-embeddings. In *Proceedings of EMNLP*. Lisbon, Portugal. https://doi.org/10.18653/v1/D15-1131.

Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2016. Learning crosslingual word embeddings without bilingual corpora. In *Proceedings of EMNLP*. Austin, Texas. http://aclweb.org/anthology/D16-1136.

Markus Freitag and Yaser Al-Onaizan. 2016. Fast domain adaptation for neural machine translation. *CoRR* abs/1612.06897. https://arxiv.org/abs/1612.06897.

Pascale Fung and Percy Cheung. 2004. Mining very-non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and em. In *Proceedings of EMNLP*. Barcelona, Spain. http://aclweb.org/anthology/W04-3208.

Ulrich Germann. 2001. Building a statistical machine translation system from scratch: How much bang for the buck can we expect? In *Proceedings of the ACL Workshop on Data-Driven Methods in Machine Translation*. Toulouse, France. http://aclweb.org/anthology/W01-1409.

Cyril Goutte, Marine Carpuat, and George Foster. 2012. The impact of sentence alignment errors on phrase-based machine translation performance. In *Proceedings of AMTA*. San Diego, USA. http://www.mt-archive.info/AMTA-2012-Goutte.pdf.

Sanjika Hewavitharana and Stephan Vogel. 2016. Extracting parallel phrases from comparable data for machine translation. *Natural Language Engineering* 22(4):549–573. https://doi.org/10.1017/S1351324916000139.

Tomoyuki Kajiwara and Mamoru Komachi. 2016. Building a monolingual parallel corpus for text simplification using sentence similarity based on alignment between word embeddings. In *Proceedings of COLING*. Osaka, Japan. http://aclweb.org/anthology/C16-1109.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL*. Prague, Czech Republic. http://aclweb.org/anthology/P07-2045.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. *CoRR* abs/1405.4053. https://arxiv.org/abs/1405.4053.

Jiwei Li, Minh-Thang Luong, and Dan Jurafsky. 2015. A hierarchical neural autoencoder for paragraphs and documents. *CoRR* abs/1506.01057. https://arxiv.org/abs/1506.01057.

Minh-Thang Luong and Christopher D. Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of IWSLT*. Da Nang, Vietnam. http://www.mt-archive.info/15/IWSLT-2015-luong.pdf.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation. *CoRR* abs/1309.4168. http://arxiv.org/abs/1309.4168.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*. https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics* 31(4):477–504. http://aclweb.org/anthology/J05-4003.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*. Philadelphia, USA. http://aclweb.org/anthology/P02-1040.

Dan Ștefănescu, Radu Ion, and Sabine Hunsicker. 2012. Hybrid parallel sentence mining from comparable corpora. In *Proceedings of EAMT*. Trento, Italy. http://www.mt-archive.info/EAMT-2012-Stefanescu.pdf.

Christoph Tillmann and Jian-ming Xu. 2009. A simple sentence-level extraction algorithm for comparable data. In *Proceedings of HLT-NAACL*. Boulder, Colorado. http://aclweb.org/anthology/N09-2024.

Masao Utiyama and Hitoshi Isahara. 2003. Reliable measures for aligning japanese-english news articles and sentences. In *Proceedings of ACL*. Sapporo, Japan. http://aclweb.org/anthology/P03-1010.

Ivan Vulić and Anna Korhonen. 2016. On the role of seed lexicons in learning bilingual word embeddings. In *Proceedings of ACL*. Berlin, Germany. http://aclweb.org/anthology/P16-1024.

Bing Zhao and Stephan Vogel. 2002. Adaptive parallel sentences mining from web bilingual news collection. In *Proceedings of IEEE ICDM*. Maebashi, Japan. http://dl.acm.org/citation.cfm?id=844380.844785.