Parallel Sentence Extraction Based on Unsupervised Bilingual Lexicon Extraction from Comparable Corpora

Chenhui Chu[†,††], Toshiaki Nakazawa[††] and Sadao Kurohashi[†††]

Parallel corpora are crucial for statistical machine translation (SMT); however, they are quite scarce for most language pairs and domains. As comparable corpora are far more available, many studies have been conducted to extract parallel sentences from them for SMT. Parallel sentence extraction relies highly on bilingual lexicons that are also very scarce. We propose an unsupervised bilingual lexicon extraction based parallel sentence extraction system that first extracts bilingual lexicons from comparable corpora and then extracts parallel sentences using the lexicons. Our bilingual lexicon extraction method is based on a combination of topic model and context based methods in an iterative process. The proposed method does not rely on any prior knowledge, and the performance can be improved iteratively. The parallel sentence extraction method uses a binary classifier for parallel sentence identification. The extracted bilingual lexicons are used for the classifier to improve the performance of parallel sentence extraction. Experiments conducted with the Wikipedia data indicate that the proposed bilingual lexicon extraction method greatly outperforms existing methods, and the extracted bilingual lexicons significantly improve the performance of parallel sentence extraction for SMT.

Key Words: *Bilingual Lexicon Extraction, Parallel Sentence Extraction, Comparable Corpora*

1 Introduction

In statistical machine translation (SMT) (Brown, Della Pietra, Della Pietra, and Mercer 1993; Och and Ney 2003; Koehn 2010), translation knowledge is acquired from parallel corpora (sentence-aligned bilingual texts); therefore, the quality and quantity of parallel corpora are crucial. However, high quality parallel corpora of sufficient size are currently available only for a few language pairs such as languages paired with English and several European language pairs. Moreover, even for these language pairs, the available domains are limited. For the rest, comprising the majority of language pairs and domains, only a few or no parallel corpora are available. This scarceness of parallel corpora has become the major bottleneck for SMT.

Comparable corpora are a set of monolingual corpora that roughly describe the same topic in

[†] This work was done when the first author was a Japan Society for the Promotion of Science Research Fellow.
[††] Japan Science and Technology Agency
[†††] Kyoto University

different languages but are not exact translation equivalents. Exploiting comparable corpora for SMT is the key to addressing the scarceness of parallel corpora because comparable corpora are far more available than parallel corpora, and there is a large amount of parallel data contained in comparable texts. Figure 1 shows an example of Japanese-Chinese comparable texts from Wikipedia describing the French city Sète that contain parallel sentences and bilingual lexicons.

Many studies have been conducted to extract parallel sentences from comparable corpora for SMT. Parallel sentence extraction depends highly on bilingual lexicons because the word overlap between a sentence pair is a crucial criterion to identify truly parallel sentences from erroneous ones, and bilingual lexicons are required to calculate this. Previous studies have used either manually created lexicons (Utiyama and Isahara 2003; Fung and Cheung 2004; Adafre and de Rijke 2006; Lu, Jiang, Chow, and Tsou 2010) or lexicons generated from a seed parallel corpus (Zhao and Vogel 2002; Munteanu and Marcu 2005; Tillmann 2009; Smith, Quirk, and Toutanova 2010; Abdul-Rauf and Schwenk 2011; Stefanescu, Ion, and Hunsicker 2012; Stefanescu and Ion 2013; Ling, Xiang, Dyer, Black, and Trancoso 2013) to identify parallel sentences. However, manual construction of bilingual lexicons is very expensive and time-consuming, and high quality seed parallel corpora of sufficient size are only available for limited language pairs and domains. A more desirable method is extracting bilingual lexicons from comparable corpora automatically, and using them for parallel sentence extraction.

Here, we propose an unsupervised bilingual lexicon extraction based parallel sentence extraction system. We first extract bilingual lexicons from comparable corpora in an unsupervised manner. We then use them for parallel sentence extraction. The proposed system consists of two major components:

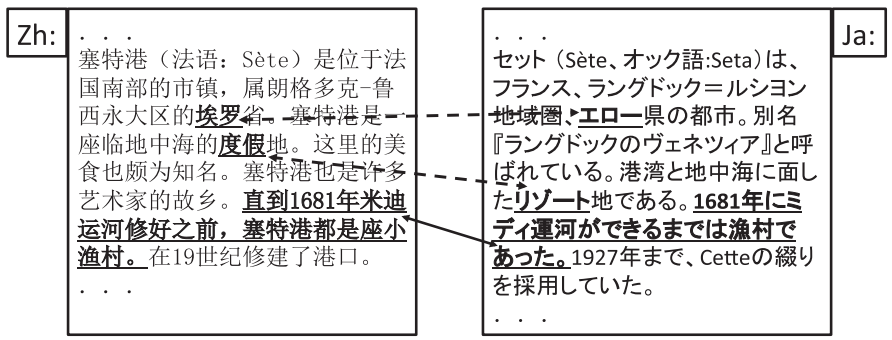

Fig. 1 Example of Japanese-Chinese comparable texts describing the French city Sète from Wikipedia (parallel sentences are linked with solid lines: bilingual lexicons are linked with dashed lines)

- Bilingual lexicon extraction: This is motivated by a method proposed in a previous study (Chu, Nakazawa, and Kurohashi 2014b) and is used to extract bilingual lexicons from comparable corpora. Chu et al. (2014b) only evaluated the accuracy of the lexicons without showing their application. In this study, we apply the extracted lexicons to parallel sentence extraction. Our bilingual lexicon extraction method is based on a combination of the topic model based method (TMBM) (Vulić, De Smet, and Moens 2011) and the context based method (CBM) (Rapp 1999) in an iterative process. TMBM and CBM are two main categories of methods proposed for bilingual lexicon extraction from comparable corpora. The proposed method maintains the advantages of TMBM, which does not require any prior knowledge, and can iteratively improve the accuracy of bilingual lexicon extraction through combination CBM.
- Parallel sentence extraction: This procedure is inspired by a previous study (Chu, Nakazawa, and Kurohashi 2014a), and is used to identify parallel sentences from comparable corpora. Chu et al. (2014a) used bilingual lexicons generated from a seed parallel corpus to identify parallel sentences. In this study, we extract bilingual lexicons from comparable corpora for parallel sentence extraction. Our parallel sentence extraction method uses a binary classifier for parallel sentence identification following (Munteanu and Marcu 2005). We use the extracted lexicons to calculate the word overlap features for the classifier to improve performance.

We conduct bilingual lexicon extraction experiments with Chinese-English, Japanese-English, and Japanese-Chinese Wikipedia data, and bilingual lexicon extraction based parallel sentence extraction experiments with Japanese-Chinese Wikipedia data. The experimental results show that the proposed bilingual lexicon extraction method considerable outperforms previously reported methods, and the extracted bilingual lexicons significantly improve the performance of parallel sentence extraction for SMT. The proposed system is language independent because it does not depend on language specific knowledge. This system can also be applied to comparable corpora other than Wikipedia in which article alignment has been established.

2 Related Work

Here, we review the literature of bilingual lexicon extraction and parallel sentence extraction separately. We then describe work related to bilingual lexicon extraction for parallel sentence extraction.

2.1 Bilingual Lexicon Extraction

2.1.1 Topic model based methods

The TMBM uses the distributional hypothesis on topics, stating that two words are potential translation candidates if they are frequent in the same cross-lingual topics and not observed in other cross-lingual topics (Vulić et al. 2011). TMBM trains a bilingual latent Dirichlet allocation (BiLDA) topic model on document-aligned comparable corpora and identifies word translations relying on word-topic distributions from the trained topic model. This method is attractive as it does not require any prior knowledge.

Vulić et al. (2011) first proposed this method. Later, Vulić and Moens (2012) extended this method to detect highly confident word translations using a symmetrization process and one-to-one constraints. They demonstrated a way to build a high quality seed dictionary using both BiLDA and cognates. Liu, Duh, and Matsumoto (2013) developed this method by converting document-aligned comparable corpora into a parallel topic-aligned corpus using BiLDA topic models and identifying word translations with the help of word alignment. Richardson, Nakazawa, and Kurohashi (2013) exploited this method in a transliteration task. Vulić and Moens (2013a) improved this method using BiLDA to learn the semantic word responses of words and identify word translations using the semantic word response vectors.

Our study differs from previous studies in that it uses a combination of TMBM and CBM. Vulić and Moens (2013b) also proposed a combination method that obtains an initial seed dictionary with a variant of TMBM. Their method increases the size of the seed dictionary iteratively using only CBM. Our method differs from the method proposed by (Vulić and Moens 2013b) in that it produces an initial seed dictionary for all source words in the vocabulary with TMBM and iteratively improves the quality using a combination of TMBM and CBM. We demonstrate that this combination outperforms both TMBM and CBM. In addition, Vulić and Moens (2013b) compared the effects of the size of the initial seed dictionary and showed that using all bilingual lexicons obtained by TMBM demonstrated the best or comparable performance relative to the best performing method, which is similar to our method as it iterates using a seed dictionary for all source words.

2.1.2 Context based methods

From the pioneering work of (Rapp 1995) and (Fung 1995), various studies have been conducted on CBM for extracting bilingual lexicons from comparable corpora. CBM is based on the distributional hypothesis on context, stating that words with similar meaning appear in similar contexts across languages. It usually consists of three steps: context vector modeling, vector

similarity calculation, and translation identification that considers a candidate with higher similarity score as a more confident translation. Gaussier, Renders, Matveeva, Goutte, and Dejean (2004) presented a geometric view of this process. Previous studies have used different definitions of context such as window-based context (Fung 1995; Rapp 1999; Koehn and Knight 2002; Haghighi, Liang, Berg-Kirkpatrick, and Klein 2008; Prochasson and Fung 2011; Tamura, Watanabe, and Sumita 2012), sentence-based context (Fung and Yee 1998), and syntax-based context (Garera, Callison-Burch, and Yarowsky 2009; Yu and Tsujii 2009; Qian, Wang, Zhou, and Zhu 2012). To quantify the strength of the association between a word and its context word, different association measures have been used, such as log likelihood ratio (Rapp 1999), term frequency - inverse document frequency (TF-IDF) (Fung and Yee 1998) and pointwise mutual information (Andrade, Nasukawa, and Tsujii 2010). Previous studies have also used different measures to compute the similarity between the vectors, such as cosine similarity (Fung and Yee 1998; Garera et al. 2009; Prochasson and Fung 2011; Tamura et al. 2012), Euclidean distance (Fung 1995; Yu and Tsujii 2009), the city-block metric (Rapp 1999), and Spearman rank order (Koehn and Knight 2002). Laroche and Langlais (2010) conducted a systematic study using different association and similarity measures for CBM.

Essentially, CBM requires a seed dictionary to project the source vector onto the vector space of the target language, which is one of the main concerns of the proposed method. In previous studies, a seed dictionary was usually created manually (Rapp 1999; Garera et al. 2009) and sometimes complemented with bilingual lexicons extracted from a parallel corpus (Fung and Yee 1998; Tamura et al. 2012), parallel sentences mined from comparable corpora (Morin and Prochasson 2011), or the Web (Prochasson and Fung 2011). In addition, some studies have attempted to create a seed dictionary using cognates (Koehn and Knight 2002; Haghighi et al. 2008); however, this cannot be applied to distant language pairs that do not share cognates, such as Chinese-English and Japanese-English. There are also some studies that have not required a seed dictionary (Rapp 1995; Fung 1995; Yu and Tsujii 2009). However, these studies show lower accuracy compared to conventional methods that do use a seed dictionary.

Our study differs from previous studies in that it uses a seed dictionary learned from comparable corpora in an unsupervised manner that is acquired automatically without prior knowledge.

2.2 Parallel Sentence Extraction

As parallel sentences tend to appear in similar article pairs, many studies first conduct article alignment from comparable corpora and then identify parallel sentences from aligned article pairs. Cross-lingual information retrieval technology is commonly used for article alignment

(Utiyama and Isahara 2003; Fung and Cheung 2004; Munteanu and Marcu 2005). Large-scale article alignment from the Web has also been studied (Resnik and Smith 2003; Uszkoreit, Ponte, Popat, and Dubiner 2010). Our study extracts parallel sentences from Wikipedia, which is a special type of comparable corpora because article alignment is established via interlanguage links. Approaches without article alignment have also been proposed (Tillmann 2009; Abdul-Rauf and Schwenk 2011; Stefanescu et al. 2012; Ling et al. 2013). These studies retrieve candidate sentence pairs directly and select parallel sentences using various filtering methods.

Parallel sentence identification methods can be classified into two different approaches, binary classification (Munteanu and Marcu 2005; Tillmann 2009; Smith et al. 2010; Stefanescu et al. 2012) and translation similarity measures (Utiyama and Isahara 2003; Fung and Cheung 2004; Abdul-Rauf and Schwenk 2011). Similar features such as word overlap and sentence length based features are used in both approaches. We believe that a machine learning approach can be more discriminative with respect to the features; thus, we adopt the binary classification approach.

Previous studies have extracted parallel sentences from various types of comparable corpora, such as bilingual news articles (Zhao and Vogel 2002; Utiyama and Isahara 2003; Munteanu and Marcu 2005; Tillmann 2009; Do, Besacier, and Castelli 2010; Abdul-Rauf and Schwenk 2011), patent data (Utiyama and Isahara 2007; Lu et al. 2010), social media (Ling et al. 2013), and the Web (Resnik and Smith 2003; Jiang, Yang, Zhou, Liu, and Zhu 2009; Hong, Li, Zhou, and Rim 2010). Recently, several studies have also been conducted to extract parallel sentences from Wikipedia (Adafre and de Rijke 2006; Smith et al. 2010; Stefanescu and Ion 2013).

2.3 Bilingual Lexicon Extraction for Parallel Sentence Extraction

We are aware of only one previous study that uses bilingual lexicon extraction for parallel sentence extraction (Smith et al. 2010). Smith et al. (2010) extracted bilingual lexicons from aligned Wikipedia articles on the basis of a supervised method. One drawback of their method is that manually created language specific training data, which is difficult to obtain, is required to achieve satisfactory results. Our study differs in that it uses an unsupervised bilingual lexicon extraction method that does not require manual efforts.

3 Bilingual Lexicon Extraction Based Parallel Sentence Extraction System

This study extracts bilingual lexicons and parallel sentences from Wikipedia. The overview of our bilingual lexicon extraction based parallel sentence extraction system is presented in Figure 2.

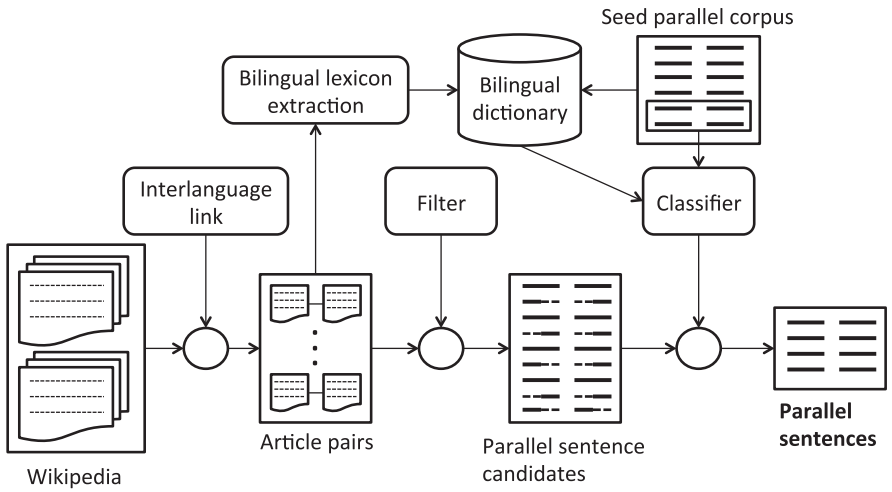

Fig. 2 Bilingual lexicon extraction based parallel sentence extraction system

We first align articles on the same topic in Wikipedia via the interlanguage links. Next, we extract bilingual lexicons from the aligned articles. From the same aligned articles, we generate all possible sentence pairs using the Cartesian product and discard pairs that do not pass a filter that reduces candidate pairs by keeping more reliable sentences. Sentence length ratio, dictionary-based word overlap (Munteanu and Marcu 2005), and cognate overlap conditions (Chu et al. 2014a) have been proposed for this filter. However, we simply use a sentence length ratio based filter. Finally, we use a classifier trained with a small number of parallel sentences from a seed parallel corpus to identify the parallel sentence from the candidates. We generate bilingual lexicons from the seed parallel corpus on the basis of the sequential word-based statistical alignment model of the IBM models (Brown et al. 1993). The generated lexicons and the bilingual lexicons extracted by the proposed method are combined to a bilingual dictionary used for the classifier to extract parallel sentences.

The details of the proposed bilingual lexicon extraction method and classifier are further described in Sections 3.1 and 3.2, respectively.

3.1 Proposed Bilingual Lexicon Extraction Method

An overview of the proposed bilingual lexicon extraction method is presented in Figure 3. We first apply TMBM to obtain bilingual lexicons from the aligned articles, which we call topical bilingual lexicons. The topical bilingual lexicons contain a list of translation candidates for a source word w_i^S, where a target word w_j^T in the list has a topical similarity score

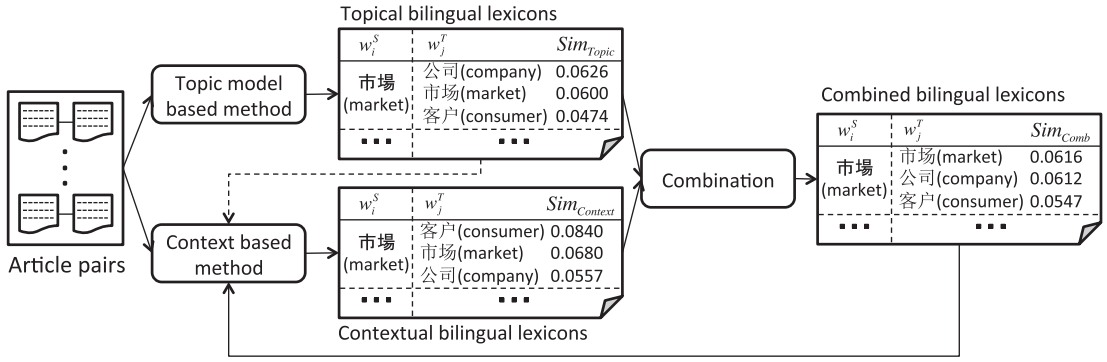

Fig. 3 Proposed bilingual lexicon extraction method

$Sim_{Topic}(w_i^S, w_j^T)$. Then, using the topical bilingual lexicons as an initial seed dictionary, we apply CBM to obtain bilingual lexicons, which we refer to as contextual bilingual lexicons. The contextual bilingual lexicons also contain a list of translation candidates for a source word, where each candidate has a contextual similarity score $Sim_{Context}(w_i^S, w_j^T)$. We then combine the topical bilingual lexicons with the contextual bilingual lexicons to obtain combined bilingual lexicons. The combination is achieved by calculating a combined similarity score $Sim_{Comb}(w_i^S, w_j^T)$ using the $Sim_{Topic}(w_i^S, w_j^T)$ and $Sim_{Context}(w_i^S, w_j^T)$ scores. After combination, the quality of the lexicons can be higher, i.e., the correct translation in the candidate list is assigned a high score and ranked higher. Therefore, we iteratively use the combined bilingual lexicons as the seed dictionary for CBM and perform combination to improve the contextual bilingual lexicons and further enhance the combined bilingual lexicons.

The proposed method not only retains the advantage of TMBM (i.e., it does not require any prior knowledge) but can also iteratively improve accuracy by a combination with CBM. The details of TMBM, CBM, and the combination method are further described in Sections 3.1.1, 3.1.2, and 3.1.3, respectively.

3.1.1 Topic model based method

In this section, we describe the TMBM used to calculate the topical similarity score Sim_{Topic} (w_i^S, w_j^T). We first train a BiLDA topic model (Mimno, Wallach, Naradowsky, Smith, and McCallum 2009), which is an extension of the standard LDA model (Blei, Ng, and Jordan 2003). Figure 4 shows the plate model for BiLDA, with D document pairs, K topics, and hyper-parameters α, β. Topics for each document are sampled from a single variable θ, which contains the topic distribution and is language-independent. Words of the two languages are

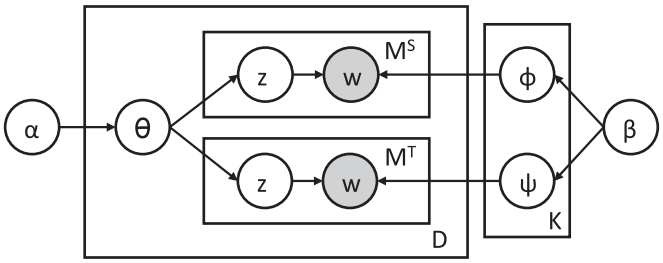

Fig. 4 BiLDA topic model

sampled from θ in conjunction with the word-topic distributions ϕ (for source language S) and ψ (for target language T).

Once the BiLDA topic model is trained and the associated word-topic distributions are obtained for both source and target corpora, we calculate the similarity of word-topic distributions to identify word translations. For similarity calculation, we use the *TI+Cue* measure (Vulić et al. 2011), which has demonstrated the best performance for identifying word translations. The *TI+Cue* measure is a linear combination of the *TI* and *Cue* measures, defined as follows,

$$Sim_{TI+Cue}(w_i^S, w_j^T) = \lambda Sim_{TI}(w_i^S, w_j^T) + (1-\lambda) Sim_{Cue}(w_i^S, w_j^T) \tag{1}$$

TI and *Cue* measures interpret and exploit the word-topic distributions in different ways; thus, combining them leads to better results.

The *TI* measure is the similarity calculated from source and target word vectors constructed over a shared space of cross-lingual topics. Each dimension of the vectors is a term frequency - inverse topic frequency score (*TF-ITF*). The *TF-ITF* score is computed in a word-topic space, which is similar to the *TF-IDF* score that is computed in a word-document space. *TF* measures the importance of a word w_i within a particular topic z_k, whereas the *ITF* of a word w_i measures the importance of w_i across all topics. Here, $n_k^{(w_i)}$ is the number of times word w_i is associated with topic z_k, W denotes the vocabulary, and K denotes the number of topics. Thus, we obtain the following.

$$TF_{i,k} = \frac{n_k^{(w_i)}}{\sum_{w_j \in W} n_k^{(w_j)}} \tag{2}$$

$$ITF_i = \log \frac{K}{1 + |\{k : n_k^{(w_i)} > 0\}|} \tag{3}$$

The *TF-ITF* score is the product of $TF_{i,k}$ and ITF_i. Then, the *TI* measure is obtained by calculating the cosine similarity of the K-dimensional source and target vectors. Let S^i be the

source vector for a source word w_i^S and T^j be the target vector for a target word w_j^T. Then, the cosine similarity is defined as follows.

$$Cos(w_i^S, w_j^T) = \frac{\sum_{k=1}^{K} S_k^i \times T_k^j}{\sqrt{\sum_{k=1}^{K}(S_k^i)^2} \times \sqrt{\sum_{k=1}^{K}(T_k^j)^2}} \tag{4}$$

The *Cue* measure is the probability $P(w_j^T|w_i^S)$, where w_j^T and w_i^S are linked via the shared topic space, defined as:

$$P(w_j^T|w_i^S) = \sum_{k=1}^{K} \psi_{k,j} \frac{\phi_{k,i}}{Norm_\phi}, \tag{5}$$

where

$$\phi_{k,i} = \frac{n_k^{(w_i)} + \beta}{\sum_{w_j \in W} n_k^{(w_j)} + W\beta} \tag{6}$$

and $\psi_{k,j}$ is similarly defined, and $Norm_\phi$ denotes the normalization factor given by $Norm_\phi = \sum_{k=1}^{K} \phi_{k,i}$ for a word w_i.

3.1.2 Context based method

Here, we describe the CBM used to calculate the contextual similarity score $Sim_{Context}(w_i^S, w_j^T)$. We use a window-based context and leave the comparison of different definitions of context as future work. Given a word, we count all its immediate context words with a window size of four (two preceding words and two following words). We build a context by collecting the counts in a bag of words fashion, i.e., we do not distinguish the positions at which the context words appear. The number of dimensions of the constructed vector is equal to the vocabulary size. We reweight each component in the vector by multiplying the *IDF* score (Garera et al. 2009), which is defined as follows.

$$IDF(t, D) = \log \frac{|D|}{1 + |\{d \in D : t \in d\}|} \tag{7}$$

Here, $|D|$ is the total number of documents in the corpus and $|\{d \in D : t \in d\}|$ denotes the number of documents wherein the term t appears. We model the source and target vectors using the method described above and project the source vector onto the vector space of the target language using a seed dictionary. The similarity of the vectors is computed using the cosine similarity (Equation 4).

Initially, we use the extracted topical bilingual lexicons (Section 3.1.1) as the seed dictionary. Note that the topical bilingual lexicons are noisy, especially for rare words (Vulić and Moens

2012). However, they provide comprehensible and useful contextual information in the target language for the source word (Vulić et al. 2011); thus, it is effective to use the lexicons as a seed dictionary for CBM.

Once contextual bilingual lexicons are extracted, we combine them with the topical bilingual lexicons. After combination, the quality of the lexicons can be improved. Therefore, we further use the combined lexicons as the seed dictionary for CBM, which can produce better contextual bilingual lexicons. Again, we combine the better contextual bilingual lexicons with the topical bilingual lexicons. By repeating these steps, both the contextual bilingual lexicons and combined bilingual lexicons can be improved iteratively.

Applying CBM and the combination once is defined as a single iteration. At iteration one, the topical bilingual lexicons are used as the seed dictionary for CBM. From the second iteration, the combined lexicons are used as the seed dictionary. In all iterations, we produce a seed dictionary for all source words in the vocabulary and use the top candidate to project the source context vector to the target language. We stop the iteration when a predefined number of iterations have been executed.

3.1.3 Combination

The TMBM measures the distributional similarity of two words on cross-lingual topics, whereas CBM measures the distributional similarity on contexts across languages. A combination of these methods can exploit both topical and contextual knowledge to measure distributional similarity, thereby making bilingual lexicon extraction more reliable and accurate. Here, we use a linear combination of the two methods to calculate a combined similarity score, which is defined as follows.

$$Sim_{Comb}(w_i^S, w_j^T) = \gamma Sim_{Topic}(w_i^S, w_j^T) + (1-\gamma) Sim_{Context}(w_i^S, w_j^T) \tag{8}$$

To reduce computational complexity, we only keep the top N translation candidates for a source word during all the steps in the proposed method. We first produce a top N candidate list for a source word using TMBM. We then apply CBM to calculate the similarity only for the candidates in the list. Finally, we conduct combination. Thus, the combination process is a type of re-ranking of candidates produced by TMBM. Note that both $Sim_{Topic}(w_i^S, w_j^T)$ and $Sim_{Context}(w_i^S, w_j^T)$ are normalized before combination, where the normalization is given by

$$Sim_{Norm}(w_i^S, w_j^T) = \frac{Sim(w_i^S, w_j^T)}{\sum_{n=1}^{N} Sim(w_i^S, w_n^T)}, \tag{9}$$

where N is the number of translation candidates for a source word.

3.2 Parallel Sentence Identification by Binary Classification

The quality of the extracted sentences is determined by the accuracy of the classifier; therefore, the classifier becomes the core component of the extraction system. Here, we first describe the training process and then introduce the features used for the classifier.

3.2.1 Training

We use a support vector machine classifier (Chang and Lin 2011). Training instances for the classifier are created following a previously reported method (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. These are filtered by the same filtering method used for parallel sentence candidate generation in the proposed system. Moreover, we randomly discard some negative instances for training when necessary to guarantee that the ratio of negative to positive instances is less than five for the performance of the classifier. Figure 5 illustrates this process.

3.2.2 Features

In this study, we reuse the features proposed by Munteanu and Marcu (2005) and Chu et al. (2014a). We divide the features to word overlap features that are related to bilingual lexicon extraction and other features.

Word Overlap Features. The word overlap feature proposed by Munteanu and Marcu (2005) has a problem, meaning that function and content words are handled in the same manner.

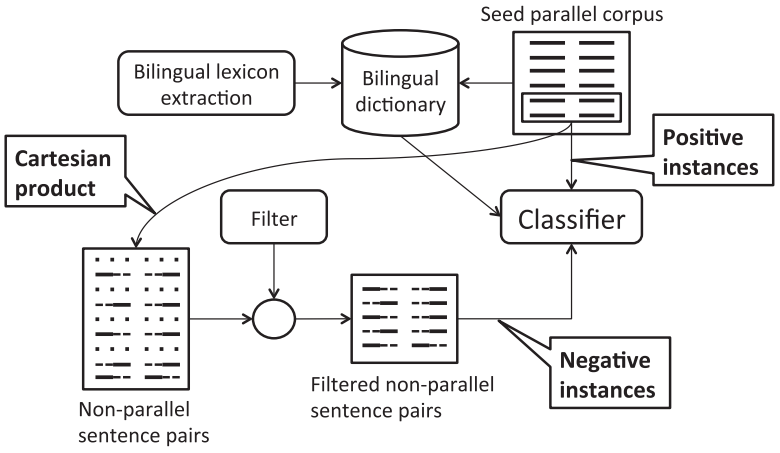

Fig. 5 Parallel sentence classifier

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 content word overlap following Chu et al. (2014a) and the following features.

- Percentage of words on each side that have a translation on the other side (according to the bilingual dictionary)
- Percentage of words that are content words on each side
- Percentage of content words on each side that have a translation on the other side (according to the bilingual dictionary)

We determine a word as a content or function word using predefined part-of-speech (POS) tag sets of function words.

Other Features. In addition to the word overlap features, the following features are used.

- Sentence length, length difference, and length ratio[1]
- Alignment features
  - Percentage and number of words that have no connection on each side
  - Top three largest fertilities[2]
  - Length of the longest contiguous connected span
  - Length of the longest unconnected substring

  The alignment features are extracted from the alignment results of the parallel and non-parallel sentences used as instances for the classifier. Note that alignment features may be unreliable when the quantity of non-parallel sentences is significantly larger than that of parallel sentences.
- Same word features. Parallel sentences often contain the same words, such as abbreviations and numbers. Such same words can be helpful clues to identify parallel sentences. We use the following features.
  - Percentage and number of words that are the same on each side

4 Experiments

We conducted bilingual lexicon extraction experiments and bilingual lexicon extraction based parallel sentence extraction experiments. We evaluated the proposed bilingual lexicon extraction method with the Chinese-English, Japanese-English, and Japanese-Chinese Wikipedia data.

[1] In our experiments, sentence length was calculated based on the number of words in a sentence.
[2] Fertility defines the number of words that a word is connected to in an alignment (Brown et al. 1993).

Bilingual lexicon extraction based parallel sentence extraction experiments were conducted with the Japanese-Chinese Wikipedia data.

4.1 Bilingual Lexicon Extraction Experiments

4.1.1 Data

We downloaded Chinese[3] (2012/09/21), Japanese[4] (2012/09/16), and English[5] (2012/10/01) Wikipedia database dumps. We used an open-source Python script[6] to extract and clean the text. Because the Chinese dump is a mixture of traditional and simplified Chinese, we converted all traditional Chinese to simplified Chinese using a conversion table published by Wikipedia.[7] We aligned the articles on the same topics in Chinese-English, Japanese-English, and Japanese-Chinese Wikipedia data via the interlanguage links. From the aligned articles, we selected $10k$ Chinese-English, Japanese-English, and Japanese-Chinese pairs as our training corpora. For Japanese-Chinese, we also conducted experiments using all aligned articles ($162k$ pairs). Using all aligned articles for Japanese-Chinese facilitates investigation of the effect of the size of the training data for the proposed method. In addition, we used the extracted bilingual lexicons for parallel sentence extraction performed on all aligned articles (4.2).

We preprocessed the Chinese and Japanese corpora using a tool proposed by Chu, Nakazawa, Kawahara, and Kurohashi (2012) and the JUMAN morphological analyzer (Kurohashi, Nakamura, Matsumoto, and Nagao 1994), respectively, for segmentation and POS tagging. The English corpora were POS tagged using the Lookahead POS Tagger (Tsuruoka, Miyao, and Kazama 2011). To reduce data sparsity and computational complexity, we retained only lemmatized noun forms. The Chinese-English data contained 112,682 Chinese and 179,058 English nouns. The Japanese-English data contained 47,911 Japanese and 188,480 English nouns. The Japanese-Chinese data contained 51,823 Japanese and 114,256 Chinese nouns for the $10k$ article pairs and 104,461 Japanese and 772,433 Chinese nouns for all article pairs. The Japanese vocabulary was smaller than the Chinese and English vocabularies because we retained only common, sahen (verbal) and proper nouns, and place, person, and organization names among all sub POS tags of nouns in JUMAN.

[3] http://dumps.wikimedia.org/zhwiki
[4] http://dumps.wikimedia.org/jawiki
[5] http://dumps.wikimedia.org/enwiki
[6] http://code.google.com/p/recommend-2011/source/browse/Ass4/WikiExtractor.py
[7] http://svn.wikimedia.org/svnroot/mediawiki/branches/REL1_12/phase3/includes/ZhConversion.php

4.1.2 Experimental settings

For BiLDA topic model training, we used the PolyLDA++ implementation proposed by Richardson et al. (2013).[8] We set the hyper-parameters α and β to $50/K$ and 0.01, respectively, following Vulić et al. (2011), where K denotes the number of topics. We trained the BiLDA topic model using Gibbs sampling with $1k$ iterations. For the combined *TI+Cue* method, we employed the Bilingual Lexicon Extractor using Topic Models toolkit created by Vulić et al. (2011).[9] Following their study, we set the linear interpolation parameter $\lambda = 0.1$. For the proposed method, we empirically set the linear interpolation parameter $\gamma = 0.8$[10] and performed 20 iterations.[11]

4.1.3 Evaluation criterion

We manually created Chinese-English, Japanese-English, and Japanese-Chinese test sets for the most frequent $1k$ source nouns[12] in the experimental data with the help of Google Translate.[13] For each source noun, if Google Translate provided correct translation, we used them. Otherwise, we performed manual translations. Note that some source nouns could have multiple translations, and we attempted to include all possible translations to the best of our knowledge. However, the test sets could be still incomplete, i.e., some translations of source words might be not registered. Following Vulić et al. (2011), we used the two metrics shown below to evaluate accuracy.

- Precision@1: Percentage of words where the top word from the list of translation candidates is the correct translation.
- Mean Reciprocal Rank (MRR) (Voorhees 1999): Here, w is a source word, $rank_w$ denotes the rank of its correct translation within the list of translation candidates, and V denotes the set of words used for evaluation. Then, MRR is defined as follows.

$$MRR = \frac{1}{|V|} \sum_{w \in V} \frac{1}{rank_w} \tag{10}$$

[8] https://bitbucket.org/trickytoforget/polylda

[9] http://people.cs.kuleuven.be/~ivan.vulic/software/BLETMv1.0wExamples.zip

[10] Because we did not have a held-out data set, we determined γ based on the Chinese-English test set. We compared the effects of different γ from 0.1 to 0.9 in intervals of 0.1; 0.8 showed the best performance. We applied the same parameter to the Japanese-English and Japanese-Chinese tasks. We recognize that determining all the parameters using held-out data is preferable; however, we leave that for future work.

[11] This iteration number was also determined empirically using the Chinese-English test set. Based on the experimental results (Figure 6), the accuracy of the proposed method greatly improves in the first few iterations, and then the performance becomes stable. We believe that accuracy would not improve with further iterations; therefore, we terminated our process at iteration 20.

[12] For Japanese-Chinese, the test sets were created for the most frequent $1k$ Japanese nouns that are limited to the sub POS tags listed in Section 4.1.1 in all article pairs.

[13] http://translate.google.com

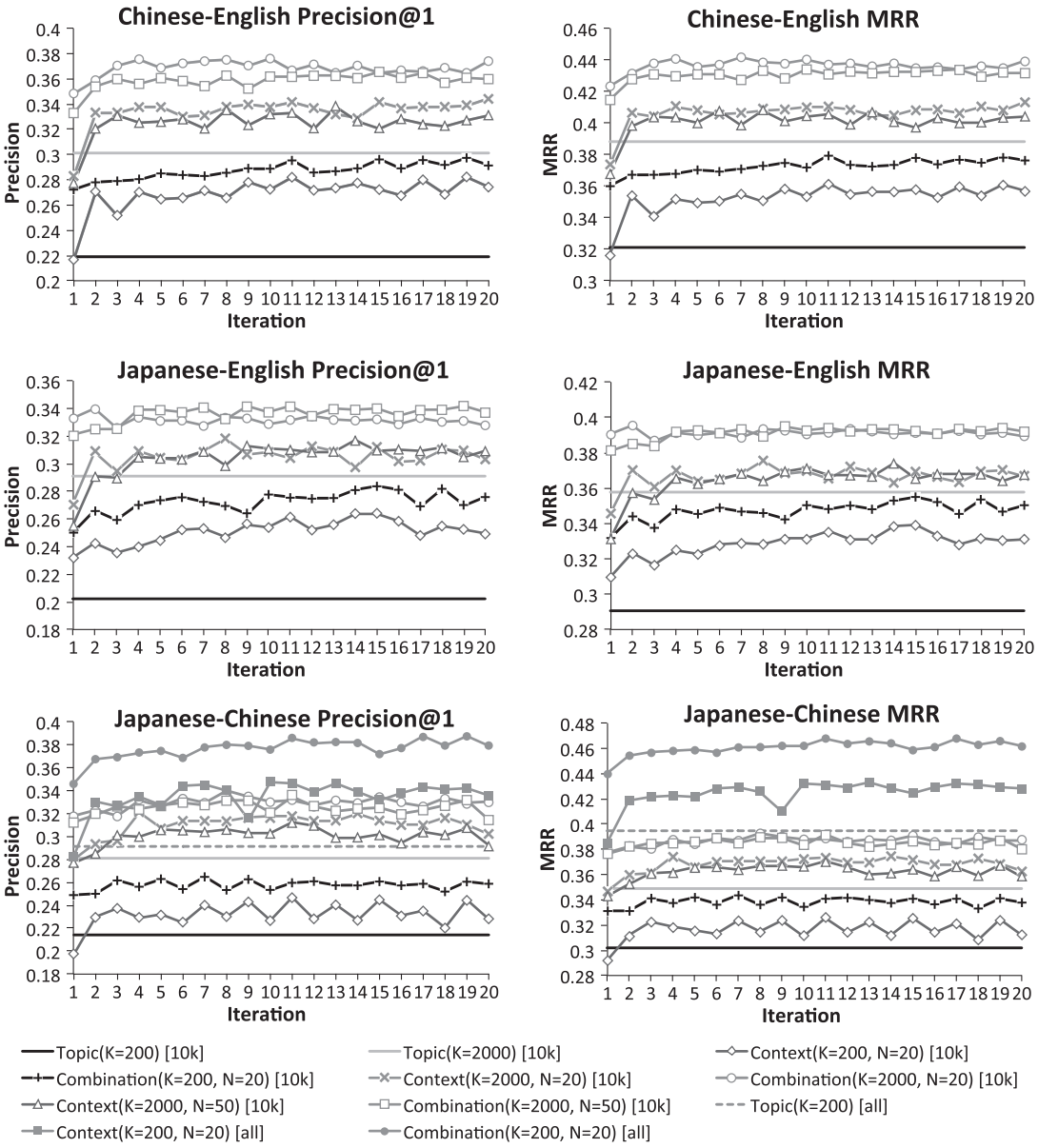

Fig. 6 Bilingual lexicon extraction results for Chinese-English, Japanese-English, and Japanese-Chinese on the test sets

We only used the top 20 candidates from the ranked list to calculate MRR. Note that for some source words, the correct translation might be not included in the list of top 20 candidates. In this case, we assume $rank_w$ to be infinity; thus, $\frac{1}{rank_w}$ is 0. We did not

discard these source words when calculating MRR, i.e., V is always $1k$. Moreover, if a source word has multiple translations in the test set and more than two are included in the candidate list, we used the most highly ranked translation to calculate MRR.

4.1.4 Results

The bilingual lexicon extraction results for the Chinese-English, Japanese-English, and Japanese-Chinese test sets are shown in Figure 6, where "Topic" denotes the lexicons extracted using only TMBM (Section 3.1.1), "Context" denotes the lexicons extracted using only CBM (Section 3.1.2), "Combination" denotes the lexicons obtained after applying the combination method (Section 3.1.3), "K" denotes the number of topics, "N" denotes the number of translation candidates for a word compared in the experiments, and "10k" and "all" denote using 10k and all article pairs as training data, respectively. For Chinese-English, Japanese-English, and the $10k$ Japanese-Chinese data, we used $K = 200$ and $K = 2000$[14] and $N = 20$ and $N = 50$,[15] For the Japanese-Chinese data that used all the articles, we used only $K = 200$[16] and $N = 20$.

Generally, it is evident that the proposed method can improve accuracy in both Precision@1 and MRR metrics compared with TMBM. CBM outperforms TMBM, which verifies the effectiveness of using the lexicons extracted by TMBM as a seed dictionary for CBM. The combination method performs better than both TMBM and CBM, which verifies the effectiveness of using both topical and contextual knowledge for bilingual lexicon extraction. Moreover, iteration can further improve the accuracy, especially in the first few iterations.

Regarding the different parameters used in our experiments, $2k$ topics is considerably better than 200 topics for both TMBM and the proposed method, which is similar to the results reported by Vulić et al. (2011). However, increasing the topic number can lead to higher computational complexity, which is not scalable for a large data set such as the Japanese-Chinese article data used in our experiments. Using 50 candidates decreases performance slightly than when using 20 candidates. Although using more candidates may increase the percentage of words where the correct translation is contained within the top N word list of translation candidates (Precision@N), it also increases the number of noisy pairs and thus decreases performance.

In our experiments, we compared two different sizes of Japanese-Chinese training data, i.e.,

[14] Vulić et al. (2011) studied the effect of the number of topics K on the performance of TMBM empirically. In our experiments, we compared $2k$ topics, which showed the best performance in (Vulić et al. 2011), to a small number of topics (200).

[15] We used 20 candidates to calculate MRR; thus, we did not examine using a number less than 20. On the other hand, because increasing it to 50 showed worse performance in our experiments, we believe that further increasing N to a number larger than 20 is not helpful.

[16] The reason for this is that $2k$ is not scalable for this large data set.

$10k$ and all article pairs. As can be seen in Figure 6, the TMBM performance obtained using all article pairs is much better than that obtained using $10k$ pairs, regardless of the number of topics used. This indicates that using more training data can improve the accuracy of TMBM. Relative to the proposed combination method, the improvements over TMBM and CBM are greater when using all article pairs than using $10k$ pairs, indicating that using more training data can also improve effectiveness.

Relative to the performance obtained with three language pairs, the performance of TMBM and the absolute values of improvement for the proposed method differ owing to the different characteristics of the data; however, the improvement curves are similar. This indicates that language independence of the proposed method.

We investigated the improved lexicons to examine the reasons for the performance improvement. We found that most improvements occurred for the case in which the Sim_{Topic} scores were similar, whereas the $Sim_{Context}$ scores are easy to distinguish. With the help of the $Sim_{Context}$ scores, the proposed method can find the correct translation. The left side of Table 1 shows an improved example of this type. Although TMBM can find topic related translations, it lacks the ability to distinguish candidates with highly similar word-topic distributions to the source word. This weakness can be solved with CBM. Moreover, a small number of improvements occur for the case in which both Sim_{Topic} and $Sim_{Context}$ scores are indistinguishable. The combination of the two methods successfully finds the correct translation, although this could be by chance. The right side of Table 1 shows such an improved example.

We also investigated the erroneous lexicons. We found that most errors occur when the correct translation is not included in the top N candidate list produced by TMBM. There are also some errors for words with correct translation that are included in the list; however, the proposed method fails to identify the translation. According to our investigation, most failures occur when either TMBM or CBM gives a significantly lower score to the correct translation than the scores given to the incorrect translations, whereas the other gives the highest or nearly

Table 1 Improved lexicon examples of "開発 (development)" (left) and "攻撃 (attack)" (right)

Candidate	Sim_{Topic}	$Sim_{Context}$	Sim_{Comb}
开发 (development)	0.0503	**0.2691**	**0.0941**
计划 (plan)	**0.0624**	0.1492	0.0798
研发 (R & D)	0.0519	0.1773	0.0770
测试 (test)	0.0561	0.1577	0.0764
里程碑 (milestone)	0.0494	0.0925	0.0580

Candidate	Sim_{Topic}	$Sim_{Context}$	Sim_{Comb}
攻击 (attack)	0.0557	0.0826	**0.0611**
部队 (troop)	0.0527	**0.0900**	0.0602
战斗 (fighting)	**0.0594**	0.0600	0.0595
士兵 (soldier)	0.0553	0.0659	0.0574
作战 (fighting)	0.0463	0.0713	0.0513

highest score to the correct translation. In this case, a simple linear combination of the two scores is not sufficiently discriminative, and incorporating both scores as features in a machine learning manner may be more effective.

4.2 Bilingual Lexicon Extraction Based Parallel Sentence Extraction Experiments

We conducted parallel sentence extraction and translation experiments to verify the effectiveness of the proposed system.

4.2.1 Data

Parallel sentence extraction experiments were conducted using all aligned articles in the Japanese-Chinese Wikipedia data (Section 4.1.1), containing $162k$ article pairs ($2.1M$ Chinese and $3.5M$ Japanese sentences).

We used the Japanese-Chinese section of the Asian Scientific Paper Excerpt Corpus as the seed parallel corpus.[17] This corpus is a scientific domain corpus provided by the Japan Science and Technology Agency[18] and the National Institute of Information and Communications Technology.[19] This corpus was created by the Japanese "Development and Research of Japanese-Chinese Natural Language Processing Technology" project and contains $680k$ sentences ($18.2M$ Chinese and $21.8M$ Japanese tokens).

4.2.2 Experimental settings

We used a sentence length ratio threshold of two as the filtering condition, i.e., the sentence pairs with sentence length ratio greater than two were discarded and not passed to the classifier. We used the LIBSVM toolkit (Chang and Lin 2011)[20] with five-fold cross-validation and a radial basis function kernel for the support vector machine classifier. The classification probability threshold was set to 0.9, i.e., we treated the sentence pairs with classification probability ≥ 0.9 as parallel sentences.[21] We used the GIZA++[22] word alignment tool, which implements the sequential word-based statistical alignment model of the IBM models (Brown et al. 1993) to

[17] http://lotus.kuee.kyoto-u.ac.jp/ASPEC
[18] http://www.jst.go.jp
[19] http://www.nict.go.jp
[20] http://www.csie.ntu.edu.tw/~cjlin/libsvm
[21] In our experiments, we compared the effects of different thresholds from 0.5 to 0.9 in intervals of 0.1; 0.9 showed the best performance. We suspect the reason for this is that lowering the threshold extracted additional sentences that contain noise, thereby affecting the SMT performance negatively.
[22] http://code.google.com/p/giza-pp

generate bilingual lexicons using the parallel sentences from the seed parallel corpus (hereafter referred to as seed parallel sentences) and calculate the alignment features. We compared four different settings for lexicon generation to investigate the effect of the number of seed parallel sentences on the proposed system:

- Baseline ($0k$): no parallel sentences were used in the seed parallel corpus, i.e., we did not use generated lexicons in the experiments.
- Baseline ($5k$): $5k$ parallel sentences from the seed parallel corpus were used.[23]
- Baseline ($10k$): $10k$ parallel sentences from the seed parallel corpus were used.
- Baseline ($680k$): all parallel sentences ($680k$) in the seed parallel corpus were used.

For the generated lexicons, we kept the top five translations with translation probability greater than 0.1 for each source word following Munteanu and Marcu (2005).[24] For the bilingual lexicon extraction based experiments, we compared the Japanese-Chinese bilingual lexicons extracted by TMBM (labeled "lexicon (TMBM)") and our best performing method (i.e., the combination method at iteration 17) shown in Figure 6 (labeled "lexicon (proposed)") to show the effect of bilingual lexicon extraction accuracy on parallel sentence extraction. To show the effect of the number of extracted bilingual lexicons on parallel sentence extraction, we empirically compared the following thresholds.[25]

- Freq100Top1: kept the lexicons for the source (Japanese) words whose frequencies were not less than 100 and the top candidate for each source word ($18,775$ lexicons).
- Freq100Top3: kept the lexicons for the source (Japanese) words whose frequencies were not less than 100 and the top three candidates for each source word ($56,325$ lexicons).
- Freq10Top1: kept the lexicons for the source (Japanese) words whose frequencies were not less than 10 and the top candidate for each source word ($52,357$ lexicons).
- Freq10Top3: kept the lexicons for the source (Japanese) words whose frequencies were not less than 10 and the top three candidates for each source word ($157,071$ lexicons).

We combined the lexicons generated from the seed parallel sentences with the extracted bilingual lexicons, further obtaining different dictionary settings (labeled "Baseline + lexicon"). The word overlap features were calculated according to the above mentioned different dictionary settings, thereby obtaining different classifiers that estimate the word overlap features using the different dictionaries while the other settings were the same. As using different parallel sentences for

[23] They were selected from the $10k$ sentences used as seed parallel sentences in the setting Baseline ($10k$).
[24] Note that the dictionary might contain noisy translation pairs and further cleaning might be helpful for our task (Aker, Paramita, Pinnis, and Gaizauskas 2014); however, we leave this as future work.
[25] Other combinations are also possible; however, we leave this as future work.

training the classifier might demonstrate different performance, we further compared the following settings.

- $5k$ Seed: used the same $5k$ parallel sentences from the seed parallel corpus as that used in the setting Baseline ($5k$). Note that the domain of these sentences differs from the Wikipedia data.
- $5k$ Extraction: used the $5k$ sentences with the highest classification probabilities selected from the sentences extracted using the classifier trained with $5k$ Seed. The domain of these sentences is the same as the Wikipedia data.[26]
- $2.5k$ Seed + $2.5k$ Extraction: used $2.5k$ sentences from $5k$ Seed and the $2.5k$ sentences with the highest classification probabilities from $5k$ Extraction.

In our experiments, we first compared the effect of bilingual lexicon extraction accuracy and number on parallel sentence extraction depending on Baseline ($0k$) and $5k$ Seed. We then compared the effect of the seed parallel sentence number on the basis of the best setting of the lexicons. Finally, we compared the effect of different parallel sentences for training the classifier.

We extracted parallel sentences from Wikipedia using the different classifiers and evaluated the Chinese-to-Japanese SMT performance using the extracted sentences as training data. For decoding, we used the state-of-the-art phrase-based SMT toolkit Moses (Koehn, Hoang, Birch, Callison-Burch, Federico, Bertoldi, Cowan, Shen, Moran, Zens, Dyer, Bojar, Constantin, and Herbst 2007) with the default options, except for the distortion limit ($6 \rightarrow 20$). We trained a 5-gram language model on the Japanese Wikipedia data ($10.7M$ sentences) using the SRILM toolkit[27] with interpolated Kneser-Ney discounting.[28] For tuning and testing, we used two distinct sets of 198 parallel sentences with 1 reference in (Chu et al. 2014a).[29] These sentences were randomly selected from the sentence pairs extracted from the same Japanese-Chinese Wikipedia data using different methods proposed by Chu et al. (2014a).[30] The erroneous parallel sentences were discarded manually because the tuning and testing sets for SMT require truly parallel sentences. Note that for training, we kept all the sentences extracted by different methods except for the sentences duplicated in the tuning and testing sets. Tuning was performed by minimum error rate training (Och 2003), which was re-run for every experiment.

[26] It would be straightforward if we had in-domain parallel sentences beforehand, however they are not always available. In the case of Japanese-Chinese Wikipedia domain, we did not have any parallel sentences available.
[27] http://www.speech.sri.com/projects/srilm
[28] Note that the Japanese sentences in the tuning and testing sets were not discarded from the data used for training the language model. Therefore, the n-grams with frequency 1 contained in the tuning and testing sets were also used for training the language model.
[29] http://lotus.kuee.kyoto-u.ac.jp/~chu/resource/wiki_zh_ja.tgz
[30] For more details of the different methods, we recommend the interested readers to refer to the original paper.

4.2.3 Results

Parallel sentence extraction and translation results obtained using different methods are shown in Tables 2, 3, and 4. Here, we report the Chinese-to-Japanese translation results on the test set using the BLEU-4 score (Papineni, Roukos, Ward, and Zhu 2002). In the tables, "# dictionary entries" denotes the number of dictionary entries for different methods and "# Sentences" denotes the number of sentences extracted by different methods after discarding the sentences duplicated in the tuning and testing sets, which were used as training data for SMT. For comparison, we conducted translation experiments using the seed parallel sentences used for lexicon generation as SMT training data (labeled "Seed ($5k$)," "Seed ($10k$)," and "Seed ($680k$)"). A significance test was performed using the bootstrap resampling method proposed by Koehn (2004).

Table 2 shows the effect of bilingual lexicon extraction number on bilingual lexicon extraction based parallel sentence extraction. As can be seen, the proposed method outperforms TMBM. The reason being that the lexicons extracted by the proposed method are more accurate than TMBM, which extracts more parallel sentences, leading to reduced out of vocabulary (OOV) word rates. Freq100 and Freq10 show comparable performance when we keep the same number of candidates. Although lowering the frequency can maintain more lexicons, it also introduces more noise because the extraction results are noisy for words with low frequencies, which leads to comparable results. Note that Top3 shows better performance than Top1. This could be because more correct lexicons are contained by keeping the top three candidates. However, increasing the number of candidates also introduces more noise; therefore, further increasing the number of candidates might decrease performance. Determining the best combination of

Table 2 Effect of bilingual lexicon extraction number on parallel sentence extraction and translation results ("†" and "‡" indicate that the result is significantly better than "Baseline" and "Baseline + lexicon (TMBM)," respectively, at $p < 0.05$)

Method	Threshold	# dictionary entries	# sentences	BLEU-4%	OOV%
Baseline ($0k$)	N/A	N/A	78,752	30.49	5.62
Baseline ($0k$) + lexicon (TMBM)	Freq100Top1	18,775	86,823	30.89	5.43
Baseline ($0k$) + lexicon (proposed)	Freq100Top1	18,775	88,499	32.99†‡	5.43
Baseline ($0k$) + lexicon (TMBM)	Freq100Top3	56,325	91,285	32.42†	5.29
Baseline ($0k$) + lexicon (proposed)	Freq100Top3	56,325	106,776	**33.63**†‡	4.91
Baseline ($0k$) + lexicon (TMBM)	Freq10Top1	52,357	85,921	31.30†	5.41
Baseline ($0k$) + lexicon (proposed)	Freq10Top1	52,357	88,928	31.49†	5.41
Baseline ($0k$) + lexicon (TMBM)	Freq10Top3	157,071	95,599	32.85†	5.15
Baseline ($0k$) + lexicon (proposed)	Freq10Top3	157,071	104,046	31.99†	4.98

Table 3 Effect of seed parallel sentence number on bilingual lexicon extraction based parallel sentence extraction and translation results (The threshold used for the extracted lexicons was "Freq100Top3", "†" and "‡" indicate that the result is significantly better than "Seed" and "Baseline," respectively, at $p < 0.05$)

Method	# dictionary entries	# sentences	BLEU-4%	OOV%
Baseline (0k)	0	78,752	30.49	5.62
Baseline (0k) + lexicon (proposed)	56,325	106,776	33.63‡	4.91
Seed (5k)			15.53	26.89
Baseline (5k)	23,446	22,849	28.10†	9.99
Baseline (5k) + lexicon (proposed)	78,561	47,191	32.22†‡	6.78
Seed (10k)			16.59	23.18
Baseline (10k)	32,607	30,115	28.67†	9.37
Baseline (10k) + lexicon (proposed)	87,523	50,440	31.56†‡	7.44
Seed (680k)			25.42	9.11
Baseline (680k)	204,254	74,852	34.44†	5.19
Baseline (680k) + lexicon (proposed)	258,124	95,644	**34.94**†	4.53

Table 4 Effect of parallel sentences used for training the classifier on bilingual lexicon extraction based parallel sentence extraction and translation results (experiments are based on Baseline (0k) + lexicon (proposed); "†" and "‡" indicate that the result is significantly better than "5k Extraction" and "2.5k Seed + 2.5k Extraction," respectively, at $p < 0.05$)

Training sentences	# sentences	BLEU-4%	OOV%
5k Seed	106,776	**33.63**†‡	4.91
5k Extraction	7,689	22.15	15.79
2.5k Seed + 2.5k Extraction	19,967	26.35	10.43
5k Extraction (same)	106,776	26.65	4.15
2.5k Seed + 2.5k Extraction (same)	106,776	31.47	4.96

frequency and number of candidates is planned for future work. Among all settings, Baseline + lexicon (proposed) with the threshold of Freq100Top3 demonstrates the best MT performance. Therefore, we adopted it for further bilingual lexicon extraction based experiments.

Table 3 shows the effect of seed parallel sentence number on bilingual lexicon extraction based parallel sentence extraction. Generally, the Seed systems do not perform well because they are trained on the parallel sentences from the seed parallel corpus that belong to the scientific domain. These differ from the tuning and testing sets, which are open domain data extracted from Wikipedia, leading to OOV word rates. The systems trained on the parallel sentences extracted from Wikipedia data perform significantly better than Seed because they consist of the same domain data as the tuning and testing sets, and the OOV word rates are significantly lower than

Seed. The Baseline + lexicon systems outperform the Baseline systems because combining the extracted bilingual lexicons to the Baseline dictionaries can help extract more parallel sentences, which leads to lower OOV word rates and thus higher SMT performance.

Focusing on the Baseline systems, we see that using a small number of parallel sentences (i.e., Baseline ($5k$) and Baseline ($10k$)) for lexicon generation is even worse than that without using it (i.e., Baseline ($0k$)), whereas using a larger number of sentences (i.e., Baseline ($680k$)) does help. The reason for the poor performance obtained using a small number of parallel sentences may be attributed to the word overlap feature gap between the sentences used for training the classifier and the Wikipedia data based on the generated lexicons. For Baseline ($5k$) and Baseline ($10k$), the sentences used for training the classifier have very high word overlap on the basis of generated lexicons because the lexicons are generated from the same sentences. However, the Wikipedia data have very low word overlap owing to the small size and domain difference of the generated lexicons. This leads to only a small number of sentences being extracted compared with the other settings. Baseline ($0k$) does not demonstrate this gap problem, thereby leading to the highest number of sentences extracted and better performance than Baseline ($5k$) and Baseline ($10k$). However, the quality of the extracted sentences is lower than the other settings because it does not use the word overlap features. The lexicon size of Baseline ($680k$) is larger, which can address the gap problem and guarantee the quality of the extracted sentences. Therefore, it shows the best performance.

The Baseline + lexicon (proposed) systems show better performance than Baseline systems, indicating the effectiveness of the proposed method. However, Baseline ($680k$) + lexicon (proposed) does not demonstrate significant difference over Baseline ($680k$). The reason for this could be that the ratio of the number of extracted lexicons to the number of lexicons in the Baseline dictionary is much smaller than the other settings, leading to a smaller ratio of newly extracted sentences that does not result in a significant difference in MT. Baseline ($5k$) + lexicon (proposed) and Baseline ($10k$) + lexicon (proposed) do not show good performance compared with Baseline ($0k$) and Baseline ($680k$) for the same reasons. The performance of Baseline ($0k$) + lexicon (proposed) is only slightly lower than that of Baseline ($680k$), indicating that only a small number of seed parallel sentences are required for the proposed method to show good performance (e.g., $5k$ sentences for training the classifier). This is the main advantage of the proposed method compared with previous methods that require a large number of seed parallel sentences (several hundreds of thousands to millions (Zhao and Vogel 2002; Munteanu and Marcu 2005; Tillmann 2009; Smith et al. 2010; Abdul-Rauf and Schwenk 2011; Stefanescu et al. 2012; Stefanescu and Ion 2013; Ling et al. 2013)).

Focusing on the difference in the number of dictionary entries between the Baseline and Baseline + lexicon systems, we can see that there are only a few overlaps between the extracted lexicons and Baseline dictionary, even when we use all parallel sentences ($680k$) in the seed parallel corpus for lexicon generation. The reason for this is the domain difference between the seed parallel corpus and Wikipedia data. The proposed method can extract in-domain lexicons from comparable corpora; therefore, it does not require an in-domain seed parallel corpus, which is another advantage of the proposed method.

Figure 7 shows examples of sentences additionally extracted by combining the extracted bilingual lexicons with Baseline ($10k$). The Baseline system cannot extract these sentence pairs due to the low word overlap between them based on the Baseline generated dictionary. Combining the extracted bilingual lexicons increases the word overlap, thereby resulting in these sentences being extracted. Based on our investigation, approximately two-thirds of the additionally extracted sentences are truly parallel sentences. The remaining erroneous parallel sentences are extracted due to the noise contained in the extracted bilingual lexicons. Example 3 in Figure 7 shows an erroneous parallel sentence pair that is extracted due to the noisy lexicons "州 (state), 西部 (west)" and "路易斯安那州 (Louisiana), オレゴン (Oregon)." One possible way to address this problem is further discarding noisy lexicon pairs by setting stricter filtering threshold; however, this may decrease the coverage of the lexicon.

Table 4 shows the effect of parallel sentences used for training the classifier on bilingual lexicon extraction based parallel sentence extraction. The experiments were conducted depending on Baseline ($0k$) + lexicon (proposed), owing to its good performance (Table 3). The $5k$ Seed demonstrates the best performance, and $2.5k$ Seed + $2.5k$ Extraction outperforms $5k$ Extraction. $5k$ Extraction significantly decreases the number of extracted sentences. We suspect that there are two reasons for this. First, the selected sentences have high classification probabilities; thus, they tend to have large word overlap depending on the lexicon, which differs from the other extracted sentences and educes the likelihood they will be extracted. Second, although we selected sentences with the highest classification probabilities, they still contain noise. The $2.5k$ Seed + $2.5k$ Extraction combines two different sets of sentences, resulting in a greater number of sentences being extracted compared with $5k$ Extraction. To make the comparison fairer, we also lowered the classification probability threshold, thereby making the number of extracted sentences the same as that of $5k$ Seed (labeled "$5k$ Extraction (same)" and "$2.5k$ Seed + $2.5k$ Extraction (same)").[31]

[31] For $5k$ Extraction (same) the threshold was 0.32; for $2.5k$ Seed + $2.5k$ Extraction (same) the threshold was 0.28.

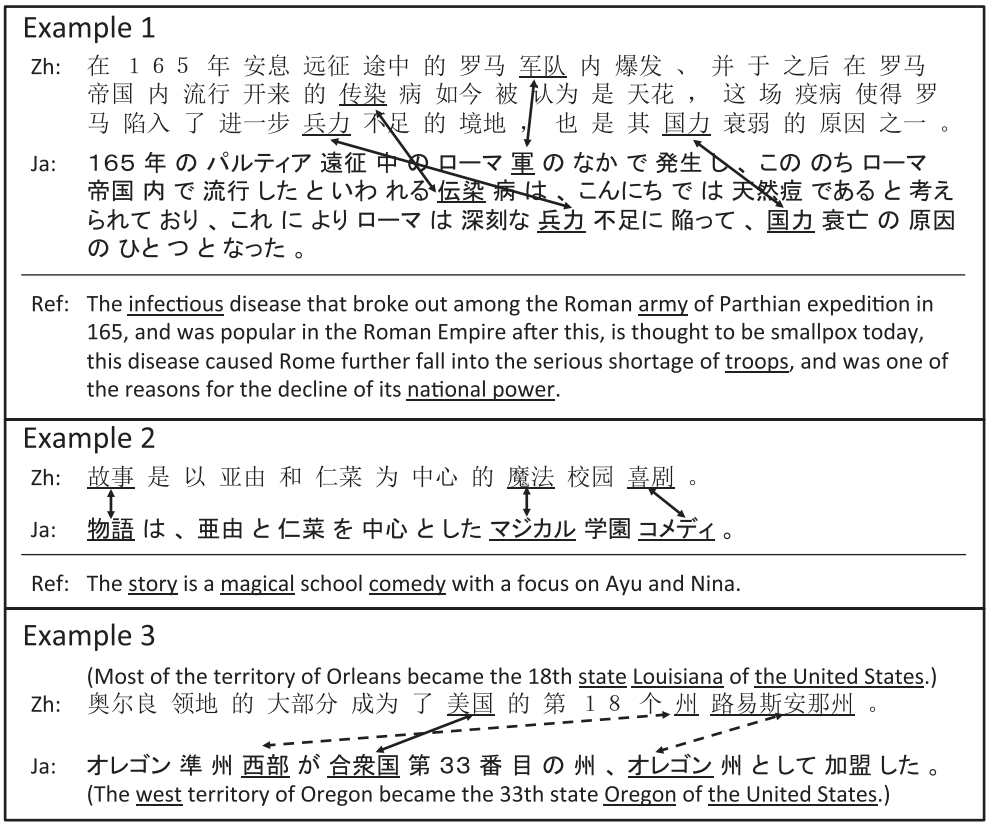

Fig. 7 Example sentences additionally extracted by combining the extracted bilingual lexicons with the Baseline (example 1 and 2 are truly parallel sentences; example 3 is an erroneous parallel sentence pair). The lexicon pairs that do not exist in the Baseline generated dictionary but were extracted by the proposed bilingual lexicon extraction method are linked (correct lexicon pairs are linked with solid lines; incorrect lexicon pairs are linked with dashed lines).

Although the OOV word rates become the same level, they also show worse translation results compared with the results obtained with $5k$ Seed. This is because a greater number of noisy sentences are produced after lowering the threshold.

5 Conclusion

Extracting parallel sentences from comparable corpora is an effective way to solve the scarceness of parallel corpora that SMT suffers. Parallel sentence extraction relies highly on bilingual lexicons that are also very scarce. We proposed an unsupervised bilingual lexicon extraction

based parallel sentence extraction system. We first extract bilingual lexicons from comparable corpora, and then extract parallel sentences using the extracted lexicons. Our bilingual lexicon extraction method is based on a combination of TMBM and CBM in an iterative process. Our parallel sentence extraction method uses a binary classifier for parallel sentence identification. The extracted bilingual lexicons are used to calculate the word overlap features for the classifier. Experiments conducted on Wikipedia data have verified the effectiveness of the proposed system and methods.

In this study, we only performed bilingual lexicon extraction based parallel sentence extraction experiments with the Japanese-Chinese language pair. In future, we plan to perform experiments with other language pairs such as Chinese-English and Japanese-English. Moreover, we only conducted experiments using Wikipedia data. The proposed system is expected to work well with other comparable corpora wherein article alignment is required beforehand, such as bilingual news articles, social media, and the Web. We also plan to perform experiments on such comparable corpora to construct a large parallel corpus for various domains.

Acknowledgement

This work was supported by the Japan Society for the Promotion of Science (JSPS) Grant-in-Aid for JSPS Fellows. We thank the anonymous reviewers for their valuable comments.

Reference

Abdul-Rauf, S. and Schwenk, H. (2011). "Parallel Sentence Generation from Comparable Corpora for Improved SMT." *Machine Translation*, **25** (4), pp. 341–375.

Adafre, S. F. and de Rijke, M. (2006). "Finding Similar Sentences across Multiple Languages in Wikipedia." In *Proceedings of the Workshop on NEW TEXT Wikis and Blogs and Other Dynamic Text Sources*, pp. 62–69.

Aker, A., Paramita, M., Pinnis, M., and Gaizauskas, R. (2014). "Bilingual Dictionaries for All EU Languages." In *Proceedings of LREC 2014*, pp. 2839–2845. ACL Anthology Identifier: L14-1623.

Andrade, D., Nasukawa, T., and Tsujii, J. (2010). "Robust Measurement and Comparison of Context Similarity for Finding Translation Pairs." In *Proceedings of COLING 2010*, pp. 19–27.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). "Latent Dirichlet Allocation." *Journal of Machine Learning Research*, **3**, pp. 993–1022.

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., and Mercer, R. L. (1993). "The Mathematics of Statistical Machine Translation: Parameter Estimation." *Association for Computational Linguistics*, **19** (2), pp. 263–312.

Chang, C.-C. and Lin, C.-J. (2011). "LIBSVM: A Library for Support Vector Machines." *ACM Transactions on Intelligent Systems and Technology*, **2**, pp. 27:1–27:27.

Chu, C., Nakazawa, T., Kawahara, D., and Kurohashi, S. (2012). "Exploiting Shared Chinese Characters in Chinese Word Segmentation Optimization for Chinese–Japanese Machine Translation." In *Proceedings of EAMT 2012*, pp. 35–42.

Chu, C., Nakazawa, T., and Kurohashi, S. (2014a). "Constructing a Chinese–Japanese Parallel Corpus from Wikipedia." In *Proceedings of LREC 2014*, pp. 642–647.

Chu, C., Nakazawa, T., and Kurohashi, S. (2014b). "Iterative Bilingual Lexicon Extraction from Comparable Corpora with Topical and Contextual Knowledge." In *Proceedings of CICLing 2014*, pp. 8404:2:296–309.

Do, T. N. D., Besacier, L., and Castelli, E. (2010). "A Fully Unsupervised Approach for Mining Parallel Data from Comparable Corpora." In *Proceedings of EAMT 2010*.

Fung, P. (1995). "Compiling Bilingual Lexicon Entries from a Non-Parallel English-Chinese Corpus." In *Proceedings of the 3rd Annual Workshop on Very Large Corpora*, pp. 173–183.

Fung, P. and Cheung, P. (2004). "Multi-level Bootstrapping For Extracting Parallel Sentences From a Quasi-Comparable Corpus." In *Proceedings of Coling 2004*, pp. 1051–1057.

Fung, P. and Yee, L. Y. (1998). "An IR Approach for Translating New Words from Nonparallel, Comparable Texts." In *Proceedings of ACL-COLING 1998*, pp. 414–420.

Garera, N., Callison-Burch, C., and Yarowsky, D. (2009). "Improving Translation Lexicon Induction from Monolingual Corpora via Dependency Contexts and Part-of-Speech Equivalences." In *Proceedings of CoNLL 2009*, pp. 129–137.

Gaussier, E., Renders, J., Matveeva, I., Goutte, C., and Dejean, H. (2004). "A Geometric View on Bilingual Lexicon Extraction from Comparable Corpora." In *Proceedings of ACL 2004*, pp. 526–533.

Haghighi, A., Liang, P., Berg-Kirkpatrick, T., and Klein, D. (2008). "Learning Bilingual Lexicons from Monolingual Corpora." In *Proceedings of ACL-HLT 2008*, pp. 771–779.

Hong, G., Li, C.-H., Zhou, M., and Rim, H.-C. (2010). "An Empirical Study on Web Mining of Parallel Data." In *Proceedings of COLING 2010*, pp. 474–482.

Jiang, L., Yang, S., Zhou, M., Liu, X., and Zhu, Q. (2009). "Mining Bilingual Data from the Web with Adaptively Learnt Patterns." In *Proceedings of ACLI-JCNLP 2009*, pp. 870–878.

Koehn, P. (2004). "Statistical Significance Tests for Machine Translation Evaluation." In Lin,

D. and Wu, D. (Eds.), *Proceedings of EMNLP 2004*, pp. 388–395.

Koehn, P. (2010). *Statistical Machine Translation* (1st edition). Cambridge University Press.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). "Moses: Open Source Toolkit for Statistical Machine Translation." In *Proceedings of ACL 2007*, pp. 177–180.

Koehn, P. and Knight, K. (2002). "Learning a Translation Lexicon from Monolingual Corpora." In *Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition*, pp. 9–16.

Kurohashi, S., Nakamura, T., Matsumoto, Y., and Nagao, M. (1994). "Improvements of Japanese morphological analyzer JUMAN." In *Proceedings of the International Workshop on Sharable Natural Language*, pp. 22–28.

Laroche, A. and Langlais, P. (2010). "Revisiting Context-based Projection Methods for Term-Translation Spotting in Comparable Corpora." In *Proceedings of COLING 2010*, pp. 617–625.

Ling, W., Xiang, G., Dyer, C., Black, A., and Trancoso, I. (2013). "Microblogs as Parallel Corpora." In *Proceedings of ACL 2013*, pp. 176–186.

Liu, X., Duh, K., and Matsumoto, Y. (2013). "Topic Models + Word Alignment = A Flexible Framework for Extracting Bilingual Dictionary from Comparable Corpus." In *Proceedings of CoNLL 2013*, pp. 212–221.

Lu, B., Jiang, T., Chow, K., and Tsou, B. K. (2010). "Building a Large English-Chinese Parallel Corpus from Comparable Patents and its Experimental Application to SMT." In *Proceedings of BUCC 2010*, pp. 42–49.

Mimno, D., Wallach, H. M., Naradowsky, J., Smith, D. A., and McCallum, A. (2009). "Polylingual Topic Models." In *Proceedings of EMNLP 2009*, pp. 880–889.

Morin, E. and Prochasson, E. (2011). "Bilingual Lexicon Extraction from Comparable Corpora Enhanced with Parallel Corpora." In *Proceedings of BUCC 2011*, pp. 27–34.

Munteanu, D. S. and Marcu, D. (2005). "Improving Machine Translation Performance by Exploiting Non-Parallel Corpora." *Computational Linguistics*, **31** (4), pp. 477–504.

Och, F. J. (2003). "Minimum Error Rate Training in Statistical Machine Translation." In *Proceedings of ACL 2003*, pp. 160–167.

Och, F. J. and Ney, H. (2003). "A Systematic Comparison of Various Statistical Alignment Models." *Computational Linguistics*, **29** (1), pp. 19–51.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). "Bleu: A Method for Automatic Evaluation of Machine Translation." In *Proceedings of ACL 2002*, pp. 311–318.

Prochasson, E. and Fung, P. (2011). "Rare Word Translation Extraction from Aligned Comparable Documents." In *Proceedings of ACL-HLT 2011*, pp. 1327–1335.

Qian, L., Wang, H., Zhou, G., and Zhu, Q. (2012). "Bilingual Lexicon Construction from Comparable Corpora via Dependency Mapping." In *Proceedings of COLING 2012*, pp. 2275–2290.

Rapp, R. (1995). "Identifying Word Translations in Non-Parallel Texts." In *Proceedings of ACL 1995*, pp. 320–322.

Rapp, R. (1999). "Automatic Identification of Word Translations from Unrelated English and German Corpora." In *Proceedings of ACL 1999*, pp. 519–526.

Resnik, P. and Smith, N. A. (2003). "The Web As a Parallel Corpus." *Computational Linguistics*, **29** (3), pp. 349–380.

Richardson, J., Nakazawa, T., and Kurohashi, S. (2013). "Robust Transliteration Mining from Comparable Corpora with Bilingual Topic Models." In *Proceedings of IJCNLP 2013*, pp. 261–269.

Smith, J. R., Quirk, C., and Toutanova, K. (2010). "Extracting Parallel Sentences from Comparable Corpora using Document Level Alignment." In *Proceedings of NAACL-HLT 2010*, pp. 403–411.

Stefanescu, D. and Ion, R. (2013). "Parallel-Wiki: A Collection of Parallel Sentences Extracted from Wikipedia." In *Proceedings of CICLing 2013*, pp. 117–128.

Stefanescu, D., Ion, R., and Hunsicker, S. (2012). "Hybrid Parallel Sentence Mining from Comparable Corpora." In *Proceedings of EAMT 2012*, pp. 137–144.

Tamura, A., Watanabe, T., and Sumita, E. (2012). "Bilingual Lexicon Extraction from Comparable Corpora Using Label Propagation." In *Proceedings of EMNLP-CoNLL 2012*, pp. 24–36.

Tillmann, C. (2009). "A Beam-Search Extraction Algorithm for Comparable Data." In *Proceedings of ACL-IJCNLP 2009*, pp. 225–228.

Tsuruoka, Y., Miyao, Y., and Kazama, J. (2011). "Learning with Lookahead: Can History-Based Models Rival Globally Optimized Models?" In *Proceedings of CoNLL 2011*, pp. 238–246.

Uszkoreit, J., Ponte, J., Popat, A., and Dubiner, M. (2010). "Large Scale Parallel Document Mining for Machine Translation." In *Proceedings of COLING 2010*, pp. 1101–1109.

Utiyama, M. and Isahara, H. (2003). "Reliable Measures for Aligning Japanese-English News Articles and Sentences." In *Proceedings of ACL 2003*, pp. 72–79.

Utiyama, M. and Isahara, H. (2007). "A Japanese-English Patent Parallel Corpus." In *Proceedings of MT Summit XI*, pp. 475–482.

Voorhees, E. M. (1999). "The TREC-8 Question Answering Track Report." In *Proceedings of the 8th Text Retrieval Conference (TREC-8)*, pp. 77–82.

Vulić, I., De Smet, W., and Moens, M.-F. (2011). "Identifying Word Translations from Comparable Corpora Using Latent Topic Models." In *Proceedings of ACL-HLT 2011*, pp. 479–484.

Vulić, I. and Moens, M.-F. (2012). "Detecting Highly Confident Word Translations from Comparable Corpora without Any Prior Knowledge." In *Proceedings of EACL 2012*, pp. 449–459.

Vulić, I. and Moens, M.-F. (2013a). "Cross-Lingual Semantic Similarity of Words as the Similarity of Their Semantic Word Responses." In *Proceedings of NAACL-HLT 2013*, pp. 106–116.

Vulić, I. and Moens, M.-F. (2013b). "A Study on Bootstrapping Bilingual Vector Spaces from Non-Parallel Data (and Nothing Else)." In *Proceedings of EMNLP 2013*, pp. 1613–1624.

Yu, K. and Tsujii, J. (2009). "Extracting Bilingual Dictionary from Comparable Corpora with Dependency Heterogeneity." In *Proceedings of NAACL-HLT 2009*, pp. 121–124.

Zhao, B. and Vogel, S. (2002). "Adaptive Parallel Sentences Mining from Web Bilingual News Collections." In *Proceedings of the 2002 IEEE International Conference on Data Mining*, pp. 745–748.

Chenhui Chu: received his B.S. in Software Engineering from Chongqing University in 2008. He received his M.S. and Ph.D. in Informatics from Kyoto University in 2012 and 2015, respectively. He is currently a researcher at the Japan Science and Technology Agency. His research interests include natural language processing, particularly machine translation.

Toshiaki Nakazawa: received his B.S. in Information and Communication Engineering and M.S. in Information Science and Technology from the University of Tokyo in 2005 and 2007, respectively. He obtained his Ph.D. in Informatics from Kyoto University in 2010. He is currently a researcher at the Japan Science and Technology Agency. His research interests center on natural language processing, particularly machine translation.

Sadao Kurohashi: received his B.S., M.S., and Ph.D. in Electrical Engineering from Kyoto University in 1989, 1991, and 1994, respectively. In 1994, he was a visiting researcher at the Institute for Research in Cognitive Science, University of Pennsylvania. He is currently a Professor at the Graduate School of Informatics, Kyoto University. His research interests include natural language processing, knowledge acquisition/representation, and information retrieval. He received a 10th anniversary best paper award from the Journal of Natural

Language Processing in 2004, a Funai IT promotion award in 2009, and an IBM faculty award in 2009.

(Received October 9, 2014)
(Revised December 30, 2014)
(Rerevised April 7, 2015; May 15, 2015)
(Accepted June 3, 2015)