Iterative Bilingual Lexicon Extraction from Comparable Corpora with Topical and Contextual Knowledge

Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi

Graduate School of Informatics, Kyoto University, Kyoto, Japan
{chu,nakazawa}@nlp.ist.i.kyoto-u.ac.jp, kuro@i.kyoto-u.ac.jp

Abstract. In the literature, two main categories of methods have been proposed for bilingual lexicon extraction from comparable corpora, namely topic model and context based methods. In this paper, we present a bilingual lexicon extraction system that is based on a novel combination of these two methods in an iterative process. Our system does not rely on any prior knowledge and the performance can be iteratively improved. To the best of our knowledge, this is the first study that iteratively exploits both topical and contextual knowledge for bilingual lexicon extraction. Experiments conduct on Chinese–English and Japanese–English Wikipedia data show that our proposed method performs significantly better than a state–of–the–art method that only uses topical knowledge.

1 Introduction

Bilingual lexicons are important for many bilingual natural language processing (NLP) tasks, such as statistical machine translation (SMT) [1, 2] and dictionary based cross–language information retrieval (CLIR) [3]. Since manual construction of bilingual lexicons is expensive and time–consuming, automatic construction is desirable. Mining bilingual lexicons from parallel corpora is a possible method. However, it is only feasible for a few language pairs and domains, because parallel corpora remain a scarce resource. As comparable corpora are far more widely available than parallel corpora, extracting bilingual lexicons from comparable corpora is an attractive research field.

In the literature, two main categories of methods have been proposed for bilingual lexicon extraction from comparable corpora, namely topic model based method (TMBM) [4] and context based method (CBM) [5]. Both methods are based on the Distributional Hypothesis [6], stating that words with similar meaning have similar distributions across languages. TMBM measures the similarity of two words on cross–lingual topical distributions, while CBM measures the similarity on contextual distributions across languages.

In this paper, we present a bilingual lexicon extraction system that is based on a novel combination of TMBM and CBM. The motivation is that a combination of these two methods can exploit both topical and contextual knowledge to measure the distributional similarity of two words, making bilingual lexicon

A. Gelbukh (Ed.): CICLing 2014, Part II, LNCS 8404, pp. 296–309, 2014.

extraction more reliable and accurate than only using one knowledge source. The key points for the combination are as follows:

- TMBM can extract bilingual lexicons from comparable corpora without any prior knowledge. The extracted lexicons are semantically related and provide comprehensible and useful contextual information in the target language for the source word [4]. Therefore, it is effective to use the lexicons extracted by TMBM as a seed dictionary, which is required for CBM.
- The lexicons extracted by CBM can be combined with the lexicons extracted by TMBM to further improve the accuracy.
- The combined lexicons again can be used as the seed dictionary for CBM. Therefore the accuracy of the lexicons can be iteratively improved.

Our system not only maintains the advantage of TMBM that does not require any prior knowledge, but also can iteratively improve the accuracy of bilingual lexicon extraction through combination CBM. To the best of our knowledge, this is the first study that iteratively exploits both topical and contextual knowledge for bilingual lexicon extraction. Experimental results on Chinese–English and Japanese–English Wikipedia data show that our proposed method performs significantly better than the method only using topical knowledge [4].

2 Related Work

2.1 Topic Model Based Methods (TMBM)

TMBM uses the Distributional Hypothesis on topics, stating that two words are potential translation candidates if they are often present in the same cross–lingual topics and not observed in other cross–lingual topics [4]. It 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 because it does not require any prior knowledge.

Vulić et al. [4] first propose this method. Later, Vulić and Moens [7] extend this method to detect highly confident word translations by a symmetrization process and the one-to-one constraints, and demonstrate a way to build a high quality seed dictionary using both BiLDA and cognates. Liu et al. [8] develop this method by converting document–aligned comparable corpora into a parallel topic–aligned corpus using BiLDA topic models, and identify word translations with the help of word alignment. Richardson et al. [9] exploit this method in the task of transliteration.

Our study differs from previous studies in using a novel combination of TMBM and CBM.

2.2 Context Based Methods (CBM)

From the pioneering work of [10, 11], various studies have been conducted on CBM for extracting bilingual lexicons from comparable corpora. CBM is based

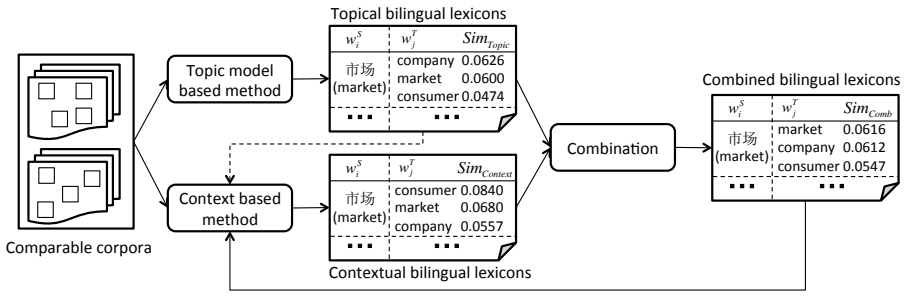

Fig. 1. Bilingual lexicon extraction system

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 treats a candidate with higher similarity score as a more confident translation. Previous studies use different definitions of context, such as window–based context [11, 5, 12–15], sentence–based context [16] and syntax–based context [17, 18] etc. Previous studies also use different measures to compute the similarity between the vectors, such as cosine similarity [16, 17, 14, 15], Euclidean distance [11, 18], city–block metric [5] and Spearman rank order [12] etc.

Basically, 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 this study. In previous studies, a seed dictionary is usually manually created [5, 17], and sometimes complemented by bilingual lexicons extracted from a parallel corpus [16, 15] or the Web [14]. In addition, some studies try to create a seed dictionary using cognates [12, 13], 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 do not require a seed dictionary [10, 11, 18]. However, these studies show lower accuracy compared to the conventional methods using a seed dictionary.

Our study differs from previous studies in using a seed dictionary automatically acquired without any prior knowledge, which is learned from comparable corpora in an unsupervised way.

3 Proposed Method

The overview of our proposed bilingual lexicon extraction system is presented in Figure 1. We first apply TMBM to obtain bilingual lexicons from comparable corpora, 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 $Sim_{Topic}(w_i^S, w_j^T)$. Then using the topical bilingual lexicons as an initial seed dictionary, we apply CBM to obtain

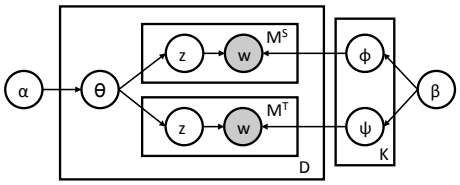

Fig. 2. The BiLDA topic model

bilingual lexicons, which we call 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 done 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. Therefore, we iteratively use the combined bilingual lexicons as the seed dictionary for CBM and conduct combination, to improve the contextual bilingual lexicons and further improve the combined bilingual lexicons.

Our system not only maintains the advantage of TMBM that does not require any prior knowledge, but also can iteratively improve the accuracy by a novel combination with CBM. Details of TMBM, CBM and combination method will be described in Section 3.1, 3.2 and 3.3 respectively.

3.1 Topic Model Based Method (TMBM)

In this section, we describe TMBM to calculate the topical similarity score $Sim_{Topic}(w_i^S, w_j^T)$.

We train a BiLDA topic model presented in [19], which is an extension of the standard LDA model [20]. Figure 2 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 sampled from θ in conjugation 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 can calculate the similarity of word–topic distributions to identify word translations. For similarity calculation, we use the *TI+Cue* measure presented in [4], which shows the best performance for identifying word translations in their study. *TI+Cue* measure is a linear combination of the *TI* and *Cue* measures, defined as follows:

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

TI and *Cue* measures interpret and exploit the word–topic distributions in different ways, thus combining the two 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 *TF–ITF* (term frequency – inverse topic frequency) score. *TF–ITF* score is computed in a word–topic space, which is similar to *TF–IDF* (term frequency – inverse document frequency) score that is computed in a word–document space. *TF* measures the importance of a word w_i within a particular topic z_k, while *ITF* of a word w_i measures the importance of w_i across all topics. Let $n_k^{(w_i)}$ be the number of times the word w_i is associated with the topic z_k, W denotes the vocabulary and K denotes the number of topics, then

$$TF_{i,k} = \frac{n_k^{(w_i)}}{\sum_{w_j \in W} n_k^{(w_j)}}, ITF_i = log \frac{K}{1 + |k : n_k^{(w_i)} > 0|} \tag{2}$$

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, T^j be the target vector for a target word w_j^T, then cosine similarity is defined as follows:

$$Cos(w_i^S, w_j^T) = \frac{\sum_{k=1}^{K} S_k^i \times T_k^j}{\sqrt{\sum_{k=1}^{K} (S_k^i)^2} \times \sqrt{\sum_{k=1}^{K} (T_k^j)^2}} \tag{3}$$

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:

$$P(w_j^T | w_i^S) = \sum_{k=1}^{K} \psi_{k,j} \frac{\phi_{k,i}}{Norm_\phi} \tag{4}$$

where $Norm_\phi$ denotes the normalization factor given by $Norm_\phi = \sum_{k=1}^{K} \phi_{k,i}$ for a word w_i.

3.2 Context Based Method (CBM)

In this section, we describe CBM to calculate the contextual similarity score $Sim_{Context}(w_i^S, w_j^T)$.

We use window–based context, and leave the comparison of using different definitions of context as future work. Given a word, we count all its immediate context words, with a window size of 4 (2 preceding words and 2 following words). We build a context by collecting the counts in a bag of words fashion, namely we do not distinguish the positions that the context words appear in. The number of dimensions of the constructed vector is equal to the vocabulary size. We further reweight each component in the vector by multiplying by the *IDF* score following [17], which is defined as follows:

$$IDF(t, D) = log \frac{|D|}{1 + |\{d \in D : t \in d\}|} \tag{5}$$

where $|D|$ is the total number of documents in the corpus, and $|\{d \in D : t \in d\}|$ denotes number of documents where 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 cosine similarity (Equation 3).

As initial, we use the topical bilingual lexicons extracted in Section 3.1 as seed dictionary. Note that the topical bilingual lexicons are noisy especially for the rare words [7]. However, since they provide comprehensible and useful contextual information in the target language for the source word [4], 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 will be improved. Therefore, we further use the combined lexicons as seed dictionary for CBM, which will produce better contextual bilingual lexicons. Again, we combine the better contextual bilingual lexicons to the topical bilingual lexicons. By repeating these steps, both the contextual bilingual lexicons and the combined bilingual lexicons will be iteratively improved.

Applying CBM and combination one time is defined as one iteration. At iteration 1, the topical bilingual lexicons are used as seed dictionary for CBM. From the second iteration, the combined lexicons are used as seed dictionary. In all iterations, we produce a seed dictionary for all the source words in the vocabulary, and use the Top 1 candidate to project the source context vector to the target language. We stop the iteration when the predefined number of iterations have been done.

3.3 Combination

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

$$Sim_{Comb}(w_i^S, w_j^T) = \gamma Sim_{Topic}(w_i^S, w_j^T) + (1-\gamma) Sim_{Context}(w_i^S, w_j^T) \quad (6)$$

To reduce computational complexity, we only keep the Top–N translation candidates for a source word during all the steps in our system. We first produce a Top–N candidate list for a source word using TMBM. Then we apply CBM to calculate the similarity only for the candidates in the list. Finally, we conduct combination. Therefore, the combination process is a kind of re–ranking of the 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:

$$Sim_{Norm}(w_i^S, w_j^T) = \frac{Sim(w_i^S, w_j^T)}{\sum_{n=1}^{N} Sim(w_i^S, w_n^T)} \quad (7)$$

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

4 Experiments

We evaluated our proposed method on Chinese–English and Japanese–English Wikipedia data. For people who want to reproduce the results reported in the this paper, we released a software that contains all the required code and data at `http://www.CICLing.org/2014/data/24`.

Note that Wikipedia is a special type of comparable corpora, because document alignment is established via interlanguage links. For many other types of comparable corpora, it is necessary to perform document alignment as an initial step. Many methods have been proposed for document alignment in the literature, such as IR–based [21, 22], feature–based [23] and topic–based [24] methods. After document alignment, our proposed method can be applied to any type of comparable corpora.

4.1 Experimental Data

We created the experimental data according to the following steps. We downloaded Chinese[1] (20120921), Japanese[2] (20120916) and English[3] (20121001) Wikipedia database dumps. We used an open–source Python script[4] to extract and clean the text from the dumps. Since 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[5]. We aligned the articles on the same topic in Chinese–English and Japanese–English Wikipedia via the interlanguage links. From the aligned articles, we selected 10,000 Chinese–English and Japanese–English pairs as our training corpora.

We preprocessed the Chinese and Japanese corpora using a tool proposed by Chu et al. [25] and JUMAN [26] respectively for segmentation and Part–of–Speech (POS) tagging. The English corpora were POS tagged using Lookahead POS Tagger [27]. To reduce data sparsity, we kept only lemmatized noun forms. The vocabularies of the Chinese–English data contain 112,682 Chinese and 179,058 English nouns. The vocabularies of the Japanese–English data contain 47,911[6] Japanese and 188,480 English nouns.

[1] `http://dumps.wikimedia.org/zhwiki`
[2] `http://dumps.wikimedia.org/jawiki`
[3] `http://dumps.wikimedia.org/enwiki`
[4] `http://code.google.com/p/recommend-2011/source/browse/Ass4/WikiExtractor.py`
[5] `http://svn.wikimedia.org/svnroot/mediawiki/branches/REL1_12/phase3/includes/ZhConversion.php`
[6] The vocabulary size of Japanese is smaller than that of Chinese and English, because we kept only common, sahen and proper nouns, place, person and organization names among all sub POS tags of noun in JUMAN.

4.2 Experimental Settings

For BiLDA topic model training, we used the implementation PolyLDA++ by Richardson et al. [9][7]. We set the hyper–parameters $\alpha = 50/K, \beta = 0.01$ following [4], where K denotes the number of topics. We trained the BiLDA topic model using Gibbs sampling with $1,000$ iterations. For the combined *TI+Cue* method, we used the toolkit BLETM obtained from Vulić et al. [4][8], where we set the linear interpolation parameter $\lambda = 0.1$ following their study. For our proposed method, we empirically set the linear interpolation parameter $\gamma = 0.8$, and conducted 20 iterations.

4.3 Evaluation Criterion

We manually created Chinese–English and Japanese–English test sets for the most 1,000 frequent source words in the experimental data with the help of Google Translate[9]. Following [4], we evaluated the accuracy using the following two metrics:

- Precision@1: Percentage of words where the Top 1 word from the list of translation candidates is the correct one.
- Mean Reciprocal Rank (MRR) [28]: Let w be a source word, $rank_w$ denotes the rank of its correct translation within the list of translation candidates, V denotes the set of words used for evaluation. Then MRR is defined as:

$$MRR = \frac{1}{|V|} \sum_{w \in V} \frac{1}{rank_w} \tag{8}$$

  Note that we only used the Top 20 candidates from the ranked list for calculating MRR.

4.4 Results

The results for the Chinese–English and Japanese–English test sets are shown in Figure 3, where "Topic" denotes the lexicons extracted only using TMBM described in Section 3.1, "Context" denotes the lexicons extracted only using CBM method described in Section 3.2, "Combination" denotes the lexicons after applying the combination method described in Section 3.3, "K" denotes the number of topics and "N" denotes the number of translation candidates for a word we compared in our experiments.

In general, we can see that our proposed method can significantly improve the accuracy in both Precision@1 and MRR metrics compared to "Topic". "Context" outperforms "Topic", which verifies the effectiveness of using the lexicons extracted by TMBM as seed dictionary for CBM. "Combination" performs better

[7] `https://bitbucket.org/trickytoforget/polylda`
[8] `http://people.cs.kuleuven.be/~ivan.vulic/software/BLETMv1.0wExamples.zip`
[9] `http://translate.google.com`

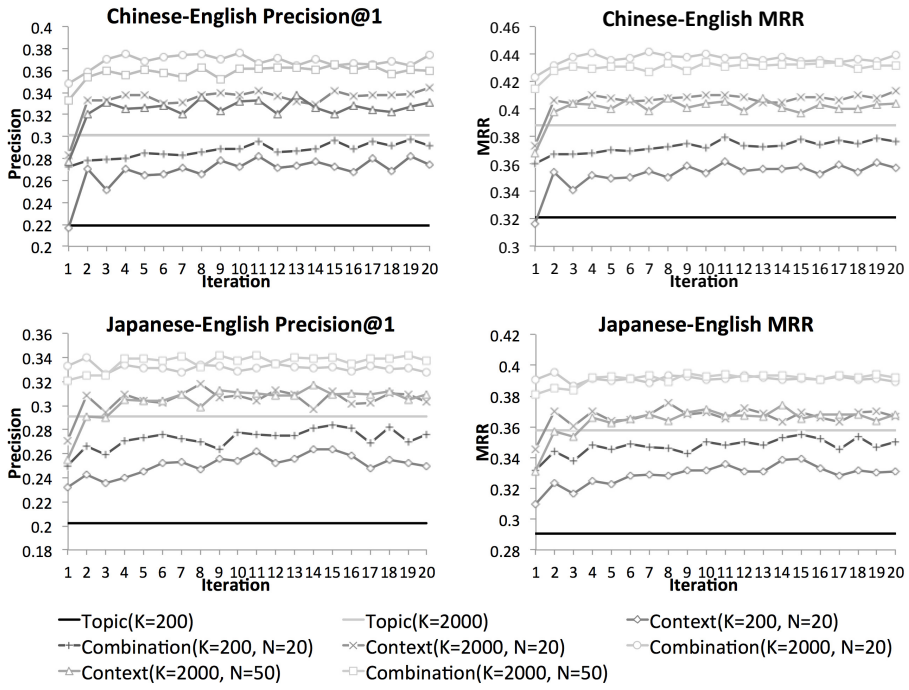

Fig. 3. Results for Chinese–English and Japanese–English on the test sets

than both "Topic" and "Context", 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. Detailed analysis for the results will be given in Section 5.

5 Discussion

5.1 Why Are Our "Topic" Scores Lower Than [4]?

The "Topic" scores are lower than the ones in [4], which are over 0.6 when $K = 2000$. The main reason is that the experimental data we used is much more sparse. Our vocabulary size is from tens of thousands to hundreds of thousands (see Section 4.1), while in [4] it is only several thousands (7,160 Italian and 9,166 English nouns). Moreover, the number of document pairs we used for training is less than [4], which is 10,000 compared to 18,898 pairs.

Another reason is the evaluation method. It may underestimate simply because of the incompleteness of our test set (e.g. our system successfully finds the correct translation "vehicle" for the Chinese word "车", but our test set only contains "car" as the correct translation.).

5.2 How Does the Proposed Method Perform on Different Language Pairs?

Our proposed method is language–independent, which is also indicated by the experimental results on two different language pairs of Chinese–English and Japanese–English. In Figure 3, we can see that although the "Topic" scores and the absolute values of improvement by our proposed method on Chinese–English and Japanese–English are different because of the different characteristics of the data, the improvement curves are similar.

5.3 How Many Iterations Are Required?

In our experiments, we conducted 20 iterations. The accuracy improves significantly in the first few iterations, and after that the performance becomes stable (see Figure 3). We suspect the reason is that there is an upper bound for our proposed method. After several iterations, the performance nearly reaches that upper bound, making it difficult to be further improved, thus the performance becomes stable. The iteration number at which the performance becomes stable depends on the particular experimental settings. Therefore, we may conclude that several iterations are enough to achieve a significant improvement and the performance at each respective iteration depends heavily on the experimental settings.

5.4 How Does the Number of Topics Affect the Performance?

According to [4], the number of topics can significantly affect the performance of the "Topic" system. In our experiments, we compared 2,000 topics that show the best performance in [4], to a small number of topics 200. Similar to [4], using 2,000 topics is significantly better than 200 topics for the "Topic" lexicons.

For the affect on the improvement by our proposed method, the improvements over "Topic" are smaller on 2,000 topics than the ones on 200 topics for both "Context" and "Combination". We suspect the reason is that the absolute values of improvement on the seed dictionary cannot lead to the same level of improvement for CBM. At iteration 1, the improvement of the "Topic" scores cannot fully reflect on the "Context" scores. Thus, the "Context" scores are lower than the "Topic" scores for 2,000 topics, while they are similar to or higher than the "Topic" scores for 200 topics (see Figure 3). The performance at iteration 1 impacts the overall improvement performance for the future iterations.

5.5 How Does the Number of Candidates Affect the Performance?

In our experiments, we measured the difference using 20 and 50 translation candidates for each word. The results show that using more candidates slightly decreases the performance (see Figure 3). 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 leads to

Table 1. Improved examples of "研究↔research" (left) and "施↔facility" (right)

Candidate	Sim_{Topic}	$Sim_{Context}$	Sim_{Comb}	Candidate	Sim_{Topic}	$Sim_{Context}$	Sim_{Comb}
research	0.0530	**0.2176**	**0.0859**	facility	0.0561	0.1127	**0.0674**
scientist	0.0525	0.1163	0.0653	center	0.0525	**0.1135**	0.0647
science	**0.0558**	0.0761	0.0599	building	0.0568	0.0933	0.0641
theory	0.0509	0.0879	0.0583	landmark	**0.0571**	0.0578	0.0572
journal	0.0501	0.0793	0.0559	plan	0.0460	0.1007	0.0570

more noisy pairs. According to our investigation on Precision@N of the two settings, the difference is quite small. For Chinese–English: Precision@20=0.5620, Precision@50=0.5780, while for Japanese–English: Precision@20=0.4930, Precision@50=0.5030. Therefore, we suspect the decrease is because the negative effect outweighs the positive. Furthermore, using more candidates will increase the computational complexity. Therefore, we believe a small number of candidates such as 20 is appropriate for our proposed method.

5.6 What Kind of Lexicons Are Improved?

Although TMBM has the advantage of finding topic related translations, it lacks of the ability to distinguish candidates that have highly similar word–topic distributions to the source word. This weakness can be solved with CBM.

Table 1 (left) shows an improved example of the Chinese word "研究↔research". All the candidates identified by "Topic" are strongly related to the topic of academia. The differences among the Sim_{Topic} scores are quite small, because of the high similarities of the word–topic distributions between these candidates and the source word, and "Topic" fails to find the correct translation. However, the differences in contextual similarities between the candidates and the source word are quite explicit. With the help of $Sim_{Context}$ scores, our proposed method finds the correct translation. Based on our investigation on the improved lexicons, most improvements belong to this type, where the Sim_{Topic} scores are similar, while the $Sim_{Context}$ scores are easy to distinguish.

Table 1 (right) shows an improved example of the Japanese word "施↔facility". The Sim_{Topic} scores are similar to the ones in the example on the left side of Table 1 that are not quite distinguishable, and "Topic" fails to find the correct translation. The difference is that CBM also fails to find the correct translation, and the Top 2 $Sim_{Context}$ scores are quite similar. The combination of the two methods successfully finds the correct translation, although this could be by chance. Based on our investigation, a small number of improvements belong to this type, where both Sim_{Topic} and $Sim_{Context}$ scores are not distinguishable.

5.7 What Kind of Errors Are Made?

As described in Section 5.5, for nearly half of the words in the test sets, the correct translation is not included in the Top N candidate list produced by TMBM.

We investigated these words and found several types of errors. The majority of errors are caused by unsuccessful identification despite topic alignment being correct (e.g. Japanese word "手↔player" is translated as "team"). Some errors are caused by unsuccessful topic alignment between the source and target words (e.g. Japanese word "置↔establishment" is translated as "kumagaya" which is a Japanese city name). There are also errors caused by words that do not clearly fit into one topic (e.g. Chinese word "爵士↔jazz/sir" may belong to either a musical or social topic). The remaining errors are due to English compound nouns. There are several pairs that contain English compound nouns in our test sets (e.g. "香港↔Hong Kong" in Chinese–English, and "ソ↔soviet union" in Japanese–English). Currently, our system cannot deal with compound nouns, and we leave it as future work for this study.

There are still some errors for words with their correct translation included in the Top N candidate list produced by TMBM, although our proposed method significantly improves the accuracy. Based on our investigation, most errors happen in the case that either the "Topic" or "Context" gives a significantly lower score to the correct translation than the scores given to the incorrect translations, while the other gives the highest or almost highest score to the correct translation. In this case, a simple linear combination of the two scores is not discriminative enough, and incorporating both scores as features in a machine learning way may be more effective.

6 Conclusion and Future Work

In this paper, we presented a bilingual lexicon extraction system exploiting both topical and contextual knowledge. Our system is based on a novel combination of TMBM and CBM, which does not rely on any prior knowledge and can be iteratively improved. Experiments conducted on Chinese–English and Japanese–English Wikipedia data verified the effectiveness of our system for bilingual lexicon extraction from comparable corpora.

As future work, firstly, we plan to compare different definitions of context for CBM. Secondly, we plan to conduct experiments on other comparable corpora rather than Wikipedia, where document alignment is required beforehand. Finally, we plan to extend our system to handle compound nouns, rare words and polysemy.

Acknowledgments. The first author is supported by Hattori International Scholarship Foundation[10]. We also thank the anonymous reviewers for their valuable comments.

References

1. Brown, P.F., Della Pietra, S.A., Della Pietra, V.J., Mercer, R.L.: The mathematics of statistical machine translation: Parameter estimation. Association for Computational Linguistics 19, 263–312 (1993)

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

2. 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., Herbst, E.: Moses: Open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, Prague, Czech Republic, pp. 177–180. Association for Computational Linguistics (2007)
3. Pirkola, A., Hedlund, T., Keskustalo, H., Järvelin, K.: Dictionary-based cross-language information retrieval: Problems, methods, and research findings. Information Retrieval 4, 209–230 (2001)
4. Vulić, I., De Smet, W., Moens, M.F.: Identifying word translations from comparable corpora using latent topic models. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA, pp. 479–484. Association for Computational Linguistics (2011)
5. Rapp, R.: Automatic identification of word translations from unrelated english and german corpora. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, College Park, Maryland, USA, pp. 519–526. Association for Computational Linguistics (1999)
6. Harris, Z.S.: Distributional structure. Word 10, 146–162 (1954)
7. Vulić, I., Moens, M.F.: Detecting highly confident word translations from comparable corpora without any prior knowledge. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, pp. 449–459. Association for Computational Linguistics (2012)
8. Liu, X., Duh, K., Matsumoto, Y.: Topic models + word alignment = a flexible framework for extracting bilingual dictionary from comparable corpus. In: Proceedings of the Seventeenth Conference on Computational Natural Language Learning, Sofia, Bulgaria, pp. 212–221. Association for Computational Linguistics (2013)
9. Richardson, J., Nakazawa, T., Kurohashi, S.: Robust transliteration mining from comparable corpora with bilingual topic models. In: Proceedings of the Sixth International Joint Conference on Natural Language Processing, Nagoya, Japan, pp. 261–269. Asian Federation of Natural Language Processing (2013)
10. Rapp, R.: Identifying word translations in non-parallel texts. In: Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, Cambridge, Massachusetts, USA, pp. 320–322. Association for Computational Linguistics (1995)
11. Fung, P.: 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 (1995)
12. Koehn, P., Knight, K.: Learning a translation lexicon from monolingual corpora. In: Proceedings of the ACL 2002 Workshop on Unsupervised Lexical Acquisition, Philadelphia, Pennsylvania, USA, pp. 9–16. Association for Computational Linguistics (2002)
13. Haghighi, A., Liang, P., Berg-Kirkpatrick, T., Klein, D.: Learning bilingual lexicons from monolingual corpora. In: Proceedings of ACL 2008, HLT, Columbus, Ohio, pp. 771–779. Association for Computational Linguistics (2008)
14. Prochasson, E., Fung, P.: Rare word translation extraction from aligned comparable documents. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA, pp. 1327–1335. Association for Computational Linguistics (2011)
15. Tamura, A., Watanabe, T., Sumita, E.: Bilingual lexicon extraction from comparable corpora using label propagation. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Jeju Island, Korea, pp. 24–36. Association for Computational Linguistics (2012)

16. Fung, P., Yee, L.Y.: An ir approach for translating new words from nonparallel, comparable texts. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Montreal, Quebec, Canada, vol. 1, pp. 414–420. Association for Computational Linguistics (1998)
17. Garera, N., Callison-Burch, C., Yarowsky, D.: Improving translation lexicon induction from monolingual corpora via dependency contexts and part-of-speech equivalences. In: Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009), Boulder, Colorado, pp. 129–137. Association for Computational Linguistics (2009)
18. Yu, K., Tsujii, J.: Extracting bilingual dictionary from comparable corpora with dependency heterogeneity. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Boulder, Colorado. Companion Volume: Short Papers, pp. 121–124. Association for Computational Linguistics (2009)
19. Mimno, D., Wallach, H.M., Naradowsky, J., Smith, D.A., McCallum, A.: Polylingual topic models. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, pp. 880–889. Association for Computational Linguistics (2009)
20. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
21. Utiyama, M., Isahara, H.: Reliable measures for aligning japanese-english news articles and sentences. In: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, Japan, pp. 72–79. Association for Computational Linguistics (2003)
22. Munteanu, D.S., Marcu, D.: Improving machine translation performance by exploiting non-parallel corpora. Computational Linguistics 31, 477–504 (2005)
23. Vu, T., Aw, A.T., Zhang, M.: Feature-based method for document alignment in comparable news corpora. In: Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), Athens, Greece, pp. 843–851. Association for Computational Linguistics (2009)
24. Zhu, Z., Li, M., Chen, L., Yang, Z.: Building comparable corpora based on bilingual lda model. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria. Short Papers, vol. 2, pp. 278–282. Association for Computational Linguistics (2013)
25. Chu, C., Nakazawa, T., Kawahara, D., Kurohashi, S.: Exploiting shared Chinese characters in Chinese word segmentation optimization for Chinese-Japanese machine translation. In: Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT 2012), Trento, Italy, pp. 35–42 (2012)
26. Kurohashi, S., Nakamura, T., Matsumoto, Y., Nagao, M.: Improvements of Japanese morphological analyzer JUMAN. In: Proceedings of the International Workshop on Sharable Natural Language, pp. 22–28 (1994)
27. Tsuruoka, Y., Miyao, Y., Kazama, J.: Learning with lookahead: Can history-based models rival globally optimized models? In: Proceedings of the Fifteenth Conference on Computational Natural Language Learning, Portland, Oregon, USA, pp. 238–246. Association for Computational Linguistics (2011)
28. Voorhees, E.M.: The TREC-8 question answering track report. In: Proceedings of the Eighth TExt Retrieval Conference (TREC-8), pp. 77–82 (1999)