Iterative Bilingual Lexicon Extraction from Comparable Corpora Using Topic Model and Context Based Methods

Chenhui Chu, Toshiaki Nakazawa, Sadao Kurohashi

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

Abstract

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) (Brown et al., 1993; Koehn et al., 2007) and dictionary based cross–language information retrieval (CLIR) (Pirkola et al., 2001). 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) (Vulić et al., 2011) and context based method (CBM) (Rapp, 1999). Both methods are based on the Distributional Hypothesis (Harris, 1954), 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 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 (Vulić et al., 2011). 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 (Vulić et al., 2011).

2. 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

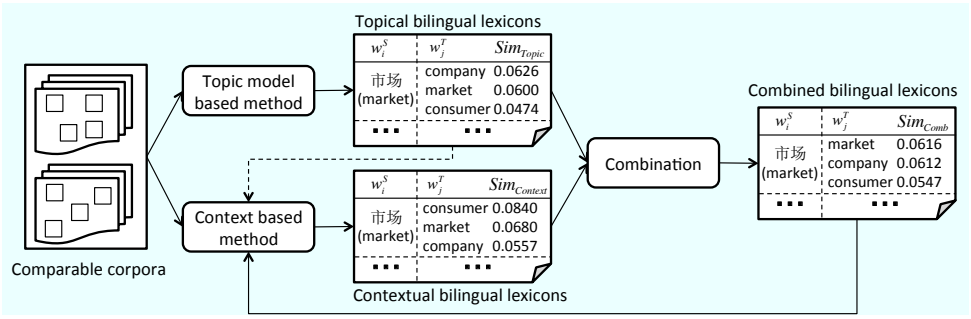

Figure 1: Bilingual lexicon extraction system.

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 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 2.1., 2.2. and 2.3. respectively.

2.1. 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 (Mimno et al., 2009), which is an extension of the standard LDA model (Blei et al., 2003). 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 (Vulić et al., 2011), 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

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

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:

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

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:

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

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

2.2. 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. 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. The number of dimensions of the contructed vector is equal to the vocabulary size. We further reweight each component in the vector by multiplying by the *IDF* score following (Garera et al., 2009), which is defined as follows:

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

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 2.1. as seed dictionary. Note that the topical bilingual lexicons are noisy especially for the rare words (Vulić and Moens, 2012). However, since they provide comprehensible and useful contextual information in the target language for the source word (Vulić et al., 2011), 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.

2.3. Combination

TMBM measures the distributional similarity of two word 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:

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

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.

3. Experiments

We evaluated our proposed method on Chinese–English and Japanese–English Wikipedia data.

3.1. Training Data

We created the training data according to the following steps. We downloaded Chinese[1] (20120921), Japanese[2] (20120916) and English[3] (20121001) Wikipedia database dumps. 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. 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 Japanese and 188,480 English nouns.

3.2. Settings

For BiLDA topic model training, we used the implementation PolyLDA++ by Richardson et al. (Richardson et al., 2013)[4]. We set the hyper–parameters $\alpha = 50/K, \beta = 0.01$ following (Vulić et al., 2011), 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. (Vulić

[1] http://dumps.wikimedia.org/zhwiki
[2] http://dumps.wikimedia.org/jawiki
[3] http://dumps.wikimedia.org/enwiki
[4] https://bitbucket.org/trickytoforget/polylda

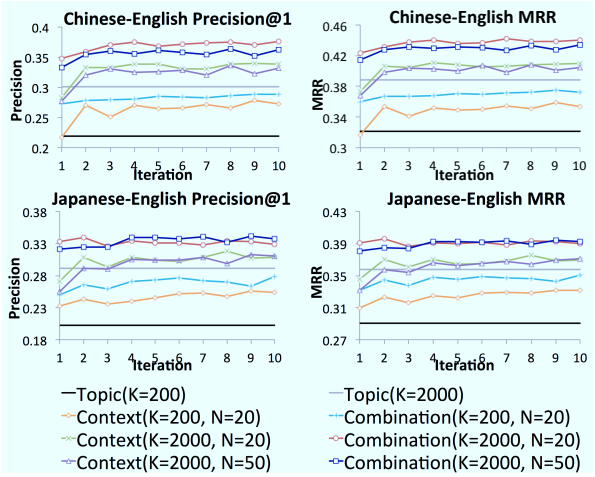

Figure 2: Results for Chinese–English and Japanese–English on the test sets.

et al., 2011)[5], 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.

3.3. Evaluation Criterion

We manually created Chinese–English and Japanese–English test sets for the most 1,000 frequent source words in the training data with the help of Google Translate[6]. Following (Vulić et al., 2011), we evaluated the accuracy using Precision@1 and Mean Reciprocal Rank (MRR) (Voorhees, 1999).

3.4. Results

The results for the Chinese–English and Japanese–English test sets are shown in Figure 2, where "Topic" denotes the lexicons extracted only using TMBM described in Section 2.1., "Context" denotes the lexicons extracted only using CBM method described in Section 2.2., "Combination" denotes the lexicons after applying the combination method described in Section 2.3., "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 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.

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

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

4. Conclusion

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.

5. References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

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

Nikesh Garera, Chris Callison-Burch, and David Yarowsky. 2009. Improving translation lexicon induction from monolingual corpora via dependency contexts and part-of-speech equivalences. In *Proceedings of CoNLL 2009*, pages 129–137.

Zellig S. Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL 2007*, pages 177–180.

David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Proceedings of EMNLP 2009*, pages 880–889.

Ari Pirkola, Turid Hedlund, Heikki Keskustalo, and Kalervo Jä rvelin. 2001. Dictionary-based cross-language information retrieval: Problems, methods, and research findings. *Information Retrieval*, 4:209–230.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of ACL 1999*, pages 519–526.

John Richardson, Toshiaki Nakazawa, and Sadao Kurohashi. 2013. Robust transliteration mining from comparable corpora with bilingual topic models. In *Proceedings of IJCNLP 2013*, pages 261–269.

Ellen M. Voorhees. 1999. The TREC-8 question answering track report. In *Proceedings of the Eighth TExt Retrieval Conference (TREC-8)*, pages 77–82.

Ivan Vulić and Marie-Francine Moens. 2012. Detecting highly confident word translations from comparable corpora without any prior knowledge. In *Proceedings of EACL 2012*, pages 449–459.

Ivan Vulić, Wim De Smet, and Marie-Francine Moens. 2011. Identifying word translations from comparable corpora using latent topic models. In *Proceedings of ACL-HLT 2011*, pages 479–484.