Pivot-Based Topic Models for Low-Resource Lexicon Extraction

John Richardson[†] **Toshiaki Nakazawa**[‡] **Sadao Kurohashi**[†]
[†]Graduate School of Informatics, Kyoto University, Kyoto 606-8501
[‡]Japan Science and Technology Agency, Kawaguchi-shi, Saitama 332-0012
john@nlp.ist.i.kyoto-u.ac.jp, nakazawa@pa.jst.jp, kuro@i.kyoto-u.ac.jp

Abstract

This paper proposes a range of solutions to the challenges of extracting large and high-quality bilingual lexicons for low-resource language pairs. In such scenarios there is often no parallel or even comparable data available. We design three effective pivot-based approaches inspired by the state-of-the-art technique of bilingual topic modelling, extending previous work to take advantage of trilingual data. The proposed models are shown to outperform traditional methods significantly and can be adapted based upon the nature of available training data. We demonstrate the accuracy of these pivot-based approaches in a realistic scenario generating an Icelandic-Korean lexicon from Wikipedia.

1 Introduction

Data-driven approaches to natural language processing have been shown to be greatly effective, and the case of bilingual lexicon extraction is no exception. Recent advances in this area have enabled the construction of large, high-quality bilingual lexicons, requiring less parallel data by making use of comparable corpora.

While such comparable corpora are readily available for many language pairs, particularly when one of those languages is English, previous direct approaches fail when there is no such data available. For many language pairs there simply does not exist comparable (and even less so parallel) data. Even for languages with a large volume of available parallel data, most corpora cover only limited domains.

There are two natural methods to deal with this problem: constructing or mining new data for the direct approach, and finding new ways to make better use of what data is already available. For an example of the construction of comparable corpora, see Zhu et al. (2013). We take the second approach and design pivot-based models for bilingual lexicon extraction. The major advantage of using a pivot language is that it is possible to take advantage of the large volume of comparable data sharing a common language such as English.

In this paper we develop pivot-based approaches to make use of modern bilingual lexicon extraction methods that can be trained on comparable corpora. We present a selection of efficient algorithms using the framework of topic modelling (Blei et al., 2003). Topic modelling has been a popular approach for bilingual lexicon extraction, however its use as a pivot model has yet to be explored. The use of topic models as a semantic similarity measure is a scalable method for low-resource languages because document-aligned comparable pivot training data (such as for English and a low-resource language) is growing ever more widely available. Examples of such sources are Wikipedia, multilingual newspaper articles and mined Web data.

While there have been many studies on bilingual lexicon extraction, there has been little focus on the important problem of resource construction for low-resource language pairs. We

29th Pacific Asia Conference on Language, Information and Computation pages 369 - 377
Shanghai, China, October 30 - November 1, 2015

present a variety of solutions to this problem, demonstrating their application to a practical scenario, and compare their effectiveness to mainstream approaches.

2 Related Work

The use of pivot models has been a common theme in the development of Natural Language Processing systems that deal with low-resource languages. In the field of Machine Translation, pivot models can be used in both decoding and the construction of parallel training data. Utiyama and Isahara (2007) give a comparison of possible methods for integrating a pivot language into phrase-based SMT systems.

Bilingual lexicon extraction has had a long history of using pivot languages. Tanaka and Umemura (1994) build a pivot lexicon by combining bilingual dictionaries, and more recently there have been attempts to extract lexicons or paraphrase patterns (Zhao et al., 2008) from bilingual corpora. A common problem with the use of a pivot language is associated noise, leading to a number of studies aiming to improve pivot lexicons, such as by using cross-lingual cooccurrences (Tanaka and Iwasaki, 1996) and 'non-aligned signatures' (Shezaf and Rappoport, 2010), a form of word context similarity.

Bilingual lexicon mining from non-parallel data has seen much popularity in recent years. Studies have considered a variety of methods such as canonical correlation analysis (Haghighi et al., 2008) and label propagation (Tamura et al., 2012). We use the method of bilingual topic modelling (Vulić et al., 2011), which has been recently applied to a variety of fields such as transliteration mining (Richardson et al., 2013).

3 Model Details

We consider the task of translating a source word s from language S to a target word t from language T. The baseline model is a direct approach using S-T training data. After describing the baseline model (bilingual LDA), we introduce three novel methods of taking advantage of data including a pivot language P, such as S-P + P-T and S-P-T data.

3.1 Baseline: Bilingual LDA

We begin with a baseline non-pivot lexicon extraction model $M_{ST} : S \times T \rightarrow \mathbb{R}$ that gives a similarity score to a source-target word pair (using S-T training data).

The non-pivot lexicon extraction model M_{ST} makes use of a bilingual topic similarity measure. We elected to use bilingual topic models rather than the more intuitive method of comparing monolingual context vectors (Rapp, 1995) as we believe topic modelling is more suitable for processing uncommon language pairs. This is because a bilingual seed lexicon is required for methods that learn a mapping between source and target vector spaces, such as Haghighi et al. (2008), in order to match cross-language word pairs. This data is unlikely to be available in sufficient quantity for low-resource language pairs, however comparable documents can be found from sources such as Wikipedia.

We base our implementation on the state-of-the-art system of Vulić et al. (2011) for comparison. This method uses the bilingual Latent Dirichlet Allocation (BiLDA) algorithm (Mimno et al., 2009), an extension of monolingual LDA (Blei et al., 2003). Monolingual LDA takes as its input a set of monolingual documents and generates a word-topic distribution ϕ classifying words appearing in these documents into semantically similar topics. Bilingual LDA extends this by considering pairs of comparable documents in each of two languages, and outputs a pair of word-topic distributions ϕ and ψ, one for each input language. The graphical model for polylingual LDA is illustrated in Figure 1.

In order to apply bilingual topic models to a lexicon extraction task, we must construct an effective word similarity measure for translation candidates. This can be achieved by a variety of methods comparing the similarity of K-dimensional word-topic vectors. We use the simple and well-studied cosine similarity measure (as defined below) to measure the similarity between topic distribution vectors ψ_{k,w_e} and ϕ_{k,w_f} for translation candidates w_e and w_f.

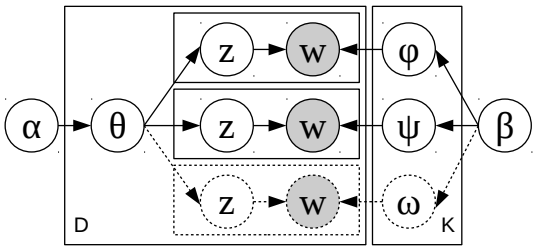

Figure 1: Graphical model for polylingual LDA with K topics, D document pairs and hyper-parameters α and β. Bilingual LDA is shown with solid lines and trilingual LDA adds the dotted lines. Topics for each document are sampled from the common distribution θ, and the two (three) languages have word-topic distributions ϕ, ψ (and ω). For further details of the LDA formulation see Blei et al. (2003).

$$Cos(w_e, w_f) = \frac{\sum_{k=1}^{K} \psi_{k,w_e} \phi_{k,w_f}}{\sqrt{\sum_{k=1}^{K} \psi_{k,w_e}^2} \sqrt{\sum_{k=1}^{K} \phi_{k,w_f}^2}} \quad (1)$$

3.2 Trilingual LDA Model

A simple yet interesting extension to applying bilingual LDA to source-target data is training trilingual LDA on a set of source-pivot-target language documents. Although in practice there may not exist such a large quantity of available trilingual data, we show in our experiments that this method is able to outperform the bilingual case even when there is a smaller volume of available trilingual data.

An advantage of this approach is that we can expect the additional (pivot) language to provide an additional point of reference, stabilizing the topic-document distribution. We show that this leads to a considerable reduction in noise, improving the translation accuracy.

The mathematical formulation is a natural extension of the bilingual case. We generate a triple of word-topic distributions ϕ, ψ and ω and a shared document-topic distribution θ using the same method as described above for bilingual LDA. The model is trained on triples of aligned comparable documents.

3.3 Pivot Model

In this section we consider an efficient method to construct a pivot model $M_{SP,PT} : S \times T \to \mathbb{R}$ (using S-P and P-T training data) that builds upon the non-pivot models M_{SP} and M_{PT}, which are built with the baseline (bilingual LDA) approach. The generation of a target word $t \in T$ is modelled as the two-step translation of a source word $s \in S$ to a pivot word $p \in P$ and then this p into T. We assume that for any translation candidate pair s, t:

$$M_{SP,PT}(s,t) = \max_{p \in P} M_{SP}(s,p) M_{PT}(p,t) \quad (2)$$

We would now like to generate the n-best distinct translations, however the size of the search space has increased to $|P||T|$ compared to $|T|$ for the non-pivot model.

The natural method for searching this space is to score every pivot translation $s \to p_i$ with M_{SP} ($|P|$ scoring operations) and then for each p_i to score every target translation $p_i \to t_j$ with M_{PT} ($|P||T|$ scoring operations). These scores are then multiplied together and sorted to generate an n-best list. As we have no further information about M it is not possible to reduce the complexity of this search without making some approximations.

We use a faster, approximate algorithm that greatly reduces the number of scoring operations required by using a beam search. The scoring operation, i.e. calculating $M(s,t)$, is the most time consuming step and therefore the most important to be avoided. Using a beam width b, the top-b pivot candidates $p_1, ..., p_b \in P$ for s are first generated, requiring $|P|$ scoring operations as we have no way to sort the p in advance. Then for each p_i, we generate the top-b target candidates $t_{i,1}, ..., t_{i,b}$ for the translation of p_i into T. This step requires only $b|T|$ scoring operations.[1]

There will be some search errors with this method and therefore b should be increased if a very accurate n-best list is required. The

[1]This can be further reduced to $b'|T|$ where $b' \leq b$ by keeping track of the final top-n list of translations t^*. This allows us to discard p_i for which $M_{SP}(s,p_i) \leq M_{SP,PT}(s,t_n^*)$, as we have $M_{PT}(p_i,t) \leq 1$.

approximate algorithm collapses into the exact method as b increases. If there are many s to translate, it would be possible to cache the M_{PT}, further improving the performance.

See Figure 2 for an illustration of our search algorithm.

3.4 'Box' Model

For many low-resource language pairs there does not exist source-target or trilingual data and therefore the pivot model is the only available option. However this is not always the case. For comparison we create one further model, the 'box' model, using all available data.

The 'box' model uses source-pivot, pivot-target, source-target and source-pivot-target data. The data is combined by creating (source, pivot, target) triples for each document. For each language L, if there is a version of the document written in L, we add it to the triple, otherwise we insert an empty string. We liken this method to packing boxes, one per document for each language, with whatever data is available. These triples are then used to train a trilingual topic model as in Section 3.2.

This approach has the advantages of avoiding noise and search errors that can be introduced by the pivot model in Section 3.3, however it relies on the availability of sufficient training data. When such data is not available we are still able to use the pivot model.

4 Experiments

In this section we consider a task where we wish to extract a Korean-Icelandic (KO-IS) and Icelandic-Korean (IS-KO) lexicon from comparable Wikipedia documents using English (EN) as a pivot language. This is a realistic scenario in which we have a sufficient quantity of aligned pivot-source and pivot-target document pairs but considerably less source-target data. We chose this language pair to demonstrate the effectiveness of our model on both low-resource and distant language pairs. English was the most natural pivot language for this task, however in some cases it might be preferable to use a different language.

The topic models were all trained on document-aligned Wikipedia data. We extracted these documents from mid-2013 Wikipedia XML dumps and they were aligned using Wikipedia 'langlinks'. The distribution of aligned document pairs including combinations of these three languages is shown in Table 1.

EN	IS	KO	Documents
✓	✓	?	22K
✓	?	✓	140K
?	✓	✓	14K
✓	✓	✓	14K
2+ languages			190K

Table 1: Number of aligned documents for each language combination. ✓ means 'included', ? means 'possibly included'. The last row shows the number of documents containing at least 2 languages.

Note that there is considerably less IS-KO data than for either EN-IS or EN-KO (only 60% of EN-IS, 10% of EN-KO). In fact the majority of trilingual data covers the same documents as the IS-KO subset, as the documents with IS and KO data very commonly also have an English version.

While it is true that there does exist some IS-KO data in Wikipedia that could be used directly to build an IS-KO lexicon, we show that there is not enough to extract translation pairs with high accuracy. Furthermore, we also show that the proposed pivot model in Section 3.3 functions well without requiring any of this data.

4.1 Settings

We used an in-house English lemmatizer and tokenizer to prepare the English data. Icelandic data was processed with IceNLP (Loftsson and Rögnvaldsson, 2007) and Korean analyzed with HanNanum (Park et al., 2010). For each language we extracted the most frequent 100K nouns for our experiments, a vocabulary size over 10 times larger than in previous work (Vulić et al., 2011).

The test data consisted of $N = 200$ (EN, KO, IS) translation triples. These were created by randomly selecting 200 nouns from our English

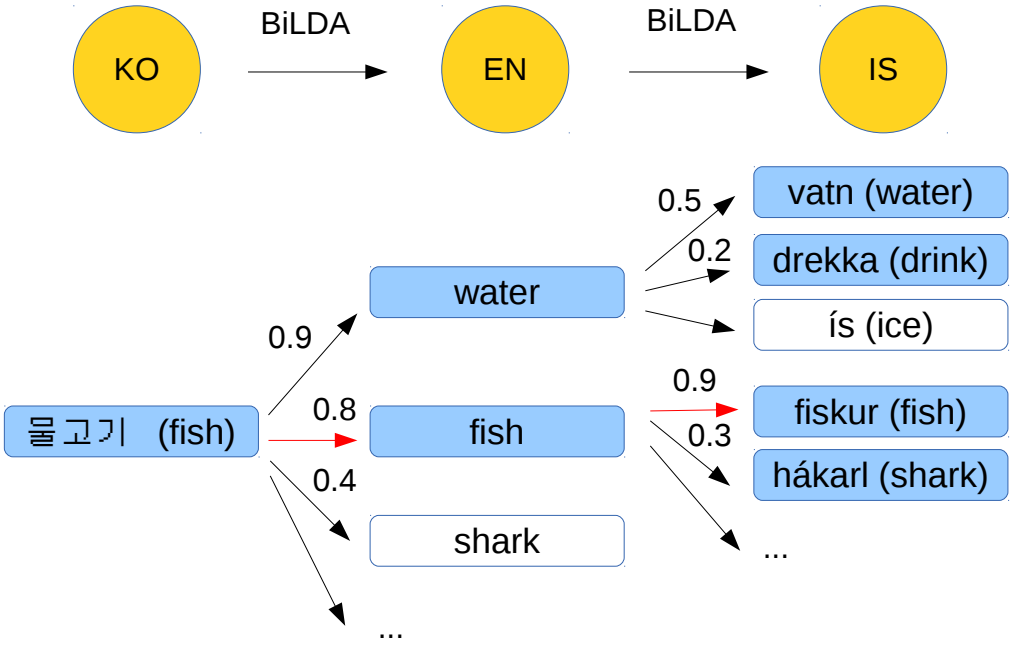

Figure 2: An illustration of the beam search algorithm using $b = 2$. The numbers shown are example similarity scores and the red arrows show the optimal path.

Wikipedia vocabulary and translating these by hand into Korean and Icelandic. For comparison the same test data was used for all experiments.

We used the PolyLDA++ tool (Richardson et al., 2013) to generate multilingual topic models. The training was run over 1000 iterations using $K = 2000$ topics. We set the LDA hyperparameters as $\alpha = 50/K$ and $\beta = 0.01$, which are the settings used most commonly in previous work on topic modelling.

The models were evaluated by generating an n-best list of translations for each word in the test set. The following statistics were then measured for the extracted lexicon, where $rank_i$ was the rank given to the correct translation in the n-best list (∞ if not in n-best list). We used $n = 10$. We also used $b = 10$ for the search beam width.

- Top-1 accuracy:

$$\frac{1}{N}\sum_{i=1}^{N}\delta_{rank_i,1} \quad (3)$$

- Mean Reciprocal Rank (MRR):

$$\frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank_i} \quad (4)$$

Lang Pair	Method	Top-1	MRR
IS-KO	baseline	0.265	0.334
	pivot	**0.310**	**0.365**
KO-IS	baseline	0.220	0.286
	pivot	**0.240**	**0.321**

Table 2: Results of direct/pivot comparison experiment.

4.2 Comparison between Direct and Pivot Model

Before applying the proposed pivot-based approaches to a realistic lexicon extraction scenario, we first verified the effectiveness of the pivot model in Section 3.3 using a controlled data set.

We consider a task where we have a corpus of aligned triples of (EN, KO, IS) documents. Our data contained 14K triples (see Table 1) with a combined vocabulary size of 30K nouns. The experiment is to test the effectiveness of using the KO-IS data directly (baseline) with the non-pivot model M_{ST} against using the pivot model $M_{SP,PT}$ with only KO-EN and EN-IS data.

This is designed to be a fair comparison as we have the same number of documents in the pivot

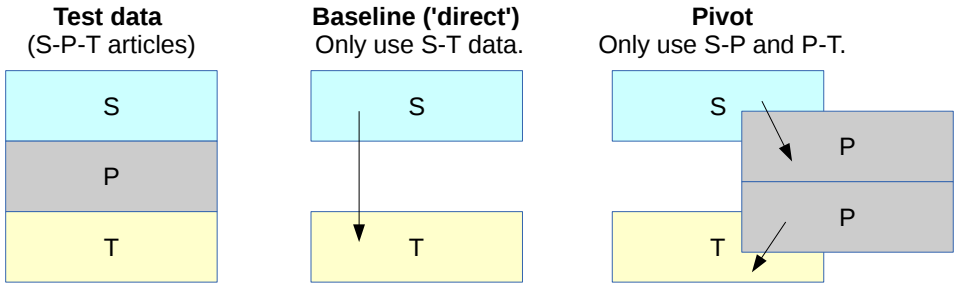

Figure 3: Training data used for direct/pivot comparison experiment.

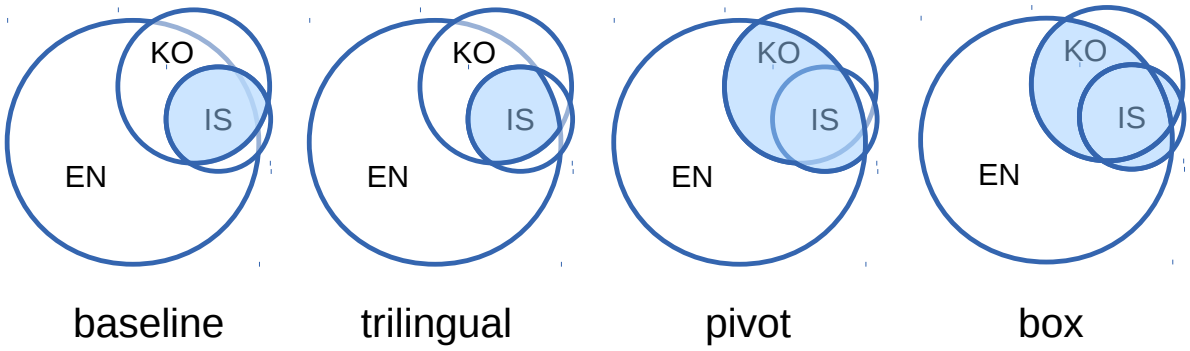

Figure 4: Subsets of Wikipedia data required for each method.

and non-pivot training sets. The organization of the training data is shown in Figure 3.

Table 2 shows the experimental results. These results show that when the same amount of data is available the pivot model is even more effective than using the source-target data directly.

In fact the scores are higher for the pivot model and we believe there could be two reasons for this. Despite the same number of documents being used, the English articles are on average longer than their Icelandic and Korean counterparts and this could improve the effectiveness of training.

It is also possible that many of the Icelandic and Korean articles were produced by partially or fully translating their corresponding English pages. This would lead to a tighter similarity in the models containing the pivot language.

4.3 Lexicon Extraction Experiment

We now turn to the main experiment, in which we consider the task of extracting a bilingual lexicon from Wikipedia for a low-resource language pair (IS-KO and KO-IS). In order to demonstrate the practical application of the proposed model, we use all the available data in Wikipedia, combining pivot and non-pivot models.

- The baseline score ('baseline') is calculated for the non-pivot model M_{ST} using only KO-IS data. This emulates the current state-of-the-art non-pivot lexicon extraction algorithm, which is only able to use the KO-IS data and model for direct translation. See Section 3.1.
- The trilingual score ('trilingual') is the accuracy of our model trained using a trilingual topic model on trilingual (KO-EN-IS)

data, which in practice is the most difficult to obtain. See Section 3.2.

- The pivot score ('pivot') is evaluated for the proposed pivot model $M_{SP,PT}$, able to make use of the KO-EN and EN-IS data. See Section 3.3.

- The score ('box'), using all possible data, is constructed by combining baseline (KO-IS), pivot (EN-KO, EN-IS) and trilingual (EN-KO-IS) data. See Section 3.4.

Figure 4 shows the data that is required (and was used) for each method. The results of the experiment are shown in Table 3.

Lang Pair	Method	Top-1	MRR
IS-KO	baseline	0.255	0.324
	trilingual	0.350	0.428
	pivot	0.380	0.459
	box	**0.420**	**0.495**
KO-IS	baseline	0.230	0.296
	trilingual	0.315	0.392
	pivot	0.305	0.398
	box	**0.390**	**0.475**

Table 3: Results of lexicon extraction experiment.

5 Analysis and Discussion

It can be seen from the results that all three proposed models considerably outperform the baseline. This demonstrates that these approaches are able to improve the quality of extracted lexicons for low-resource language pairs by making use of pivot language data, giving a large accuracy improvement over previous work.

An interesting observation is that the trilingual model is able to greatly improve upon the baseline even though it uses less training data. It is probable that the addition of the additional language (English) has helped to reduce the noise in the Korean-Icelandic model by stabilizing the document-topic distribution.

The pivot approach further improves on this by making use of the relatively large volume of EN-KO and EN-IS data. Furthermore, the pivot model score is not far from the most effective method 'box', which requires all the data, some of which is difficult in general to obtain (trilingual and KO-IS data). This shows that the pivot model is still able to compete with a model trained directly on source-target data.

Candidate	Meaning	Score
결혼	marriage	0.875
남편	husband	0.796
아내	wife	0.756
약혼	engagement	0.732
결혼식	wedding	0.726

Table 4: An example of a good translation: 'hjóna-band' (marriage).

Candidate	Meaning	Score
스튜어트	Stewart	0.355
주장	claim	0.327
반증	disproof	0.301
논란	controversy	0.296
증언	testimony	0.289

Table 5: An example of a bad translation: 'tilgangur' (purpose).

The most effective method was the 'box' approach and this is perhaps to be expected as it was able to make use of the largest volume of data. For relatively high-resource language pairs this method is likely to be the most effective as more data is available, however the pivot model becomes the only available option as the source-target data becomes sparse. When the necessary data is available, the 'box' approach can improve upon the pivot model.

Tables 4 and 5 give examples of successful and incorrect translations using the pivot model. The model can be seen to perform more effectively on words with a concrete meaning (Table 4) and less so on abstract concepts (Table 5), which often have more variation in their represention across languages. Analysis of the n-best lists revealed a tendency for clumping of pivot words. As in the example in Table 6, the same pivot word was often used to generate groups of consecutive target language words. This how-

Rank	Pivot p	Target t	$M_{SP}(s,p)$	$M_{PT}(p,t)$	Score
1	feminism	나혜석 (Na Hyeseok)	0.902	0.969	0.873
2	feminism	여자 (woman)	0.902	0.967	0.871
3	feminism	여성 (female)	0.902	0.907	0.818
...	...	...	...	...	...
9	feminism	여학교 (girls' school)	0.902	0.517	0.466
10	wife	아내 (wife)	0.315	0.914	0.288

Table 6: Analysis of translation for 'kona' (woman), showing high clumping.

Rank	Pivot p	Target t	$M_{SP}(s,p)$	$M_{PT}(p,t)$	Score
1	world	세계 (world)	0.712	0.851	0.606
2	world	월드 (world)	0.712	0.619	0.441
3	cosmos	창조 (creation)	0.278	0.965	0.268
4	cosmos	만물 (all things)	0.278	0.928	0.258
5	universe	우주론 (cosmology)	0.225	0.973	0.219
6	universe	빅뱅 (big bang)	0.225	0.965	0.217

Table 7: Analysis of translation for 'heimur' (world), showing less clumping.

ever seems not to reduce the quality of the output, as we did not notice any significant change in the MRR scores when adding the restriction that only one target word could be generated from any pivot word. An example with less clumping is shown in Table 7.

6 Conclusion and Future Work

In this paper we have presented three novel pivot-based approaches for bilingual lexicon extraction with low-resource language pairs. The proposed models are able to generate a high-quality lexicon for language pairs with no direct source-target training data, and we have shown that each model considerably outperforms a state-of-the-art non-pivot baseline. With a variety of approaches it is possible to select an appropriate method based on the size and nature of available training data.

There is much still to explore in the area of the construction of lexicons for low-resource language pairs. A possible extension to the proposed model is to use a larger pivot base, of not just one but of multiple pivot languages acting as a form of interlingua, similar to the idea in Dabre et al. (2014). This could improve the quality of the model in cases where there is not such a clear choice for an appropriate pivot language.

Another possibility for improvement is removing the assumption that there is an appropriate pivot word, using instead a direct mapping between the word-topic vector spaces for source-pivot and pivot-target topic models.

In the future we would like to use the proposed method to improve machine translation by extracting a large lexicon and applying it to a low-resource translation task.

Acknowledgments

We would like to thank the reviewers for their instructive comments. The first author is supported by a Japanese Government Scholarship (MEXT).

References

David Blei, Andrew Ng and Michael Jordan. 2003. Latent Dirichlet Allocation. In *The Journal of Machine Learning Research*, Volume 3.

Raj Dabre, Fabien Cromieres, Sadao Kurohashi and Pushpak Bhattacharyya. 2014. Leveraging Small Multilingual Corpora for SMT Using Many Pivot Languages. In *NAACL 2014*.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick and Dan Klein. 2008. Learning Bilingual Lexicons from Monolingual Corpora. In *ACL 2008*.

Hrafn Loftsson and Eiríkur Rögnvaldsson. 2007. IceNLP: A Natural Language Processing Toolkit for Icelandic. In *Interspeech 2007*.

David Mimno, Hanna Wallach, Jason Naradowsky, David Smith and Andrew McCallum. 2009. Polylingual topic models. In *EMNLP 2009*.

Park, S., Choi, D., Kim, E., and Choi, K.-S. 2010. A plug-in component-based Korean morphological analyzer. In *HCLT 2010*.

Reinhard Rapp. 1995. Identifying Word Translations in Non-Parallel Texts. In *ACL 1995*.

John Richardson, Toshiaki Nakazawa and Sadao Kurohashi. 2013. Robust Transliteration Mining from Comparable Corpora with Bilingual Topic Models. In *IJCNLP 2013*.

Daphna Shezaf and Ari Rappoport. 2010. Bilingual Lexicon Generation Using Non-Aligned Signatures. In *ACL 2010*.

Akihiro Tamura, Taro Watanabe and Eiichiro Sumita. 2012. Bilingual Lexicon Extraction from Comparable Corpora Using Label Propagation. In *EMNLP-CoNLL 2012*.

Kumiko Tanaka and Hideya Iwasaki. 1996. Extraction of lexical translations from non-aligned corpora. In *Conference on Computational Linguistics 1996*.

Kumiko Tanaka and Kyoji Umemura. 1994. Construction of a bilingual dictionary intermediated by a third language. In *Conference on Computational Linguistics 1994*.

Masao Utiyama and Hitoshi Isahara. 2007. Comparison of Pivot Methods for Phrase-based Statistical Machine Translation. In *NAACL 2007*.

Ivan Vulić, Wim De Smet and Marie-Francine Moens. 2011. Identifying Word Translations from Comparable Corpora Using Latent Topic Models. In *ACL 2011*.

Shiqi Zhao, Haifeng Wang, Ting Liu and Sheng Li. 2008. Pivot Approach for Extracting Paraphrase Patterns from Bilingual Corpora. In *ACL 2008*.

Zede Zhu, Miao Li, Lei Chen, Zhenxin Yang. 2013. Building Comparable Corpora Based on Bilingual LDA Model. In *ACL 2013*.