Pivot, Box and Trilingual: Lexicon Extraction for Low-Resource Language Pairs with Extended Topic Models

John Richardson† Toshiaki Nakazawa‡ Sadao Kurohashi†
†Graduate School of Informatics, Kyoto University
‡Japan Science and Technology Agency
john@nlp.ist.i.kyoto-u.ac.jp, nakazawa@pa.jst.jp, kuro@i.kyoto-u.ac.jp

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. While training data is readily available for many language pairs, many existing approaches fail for languages for which there simply does not exist parallel data.

While there have been many studies on bilingual lexicon extraction, there has been little focus on the important problem of accommodating low-resource language pairs. We present a variety of solutions to this problem, demonstrating their application to a practical scenario, and compare their effectiveness to mainstream approaches.

In this paper we develop pivot-based approaches for bilingual lexicon extraction using the framework of topic modelling [1]. Topic modelling has been a popular approach for bilingual lexicon extraction, however its use as a pivot model has yet to be explored.

2 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.

2.1 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 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. [2], 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. [4] for comparison. This method uses the bilingual Latent Dirichlet Allocation (BiLDA) algorithm [3], an extension of monolingual LDA [1]. 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.

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

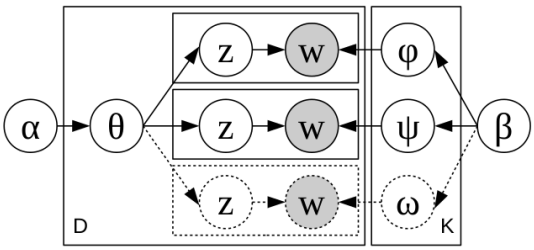

Figure 1: Graphical model for polylingual LDA with K topics, D document pairs and hyper-parameters α and β. The w and z denote words and topics respectively. Bilingual LDA is shown with solid lines and trilingual LDA adds the dotted lines. Topics for each document are sampled from the common distribution θ, and the two (three) languages have word-topic distributions ϕ, ψ (and ω).

2.2 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.

2.3 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:

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

It would also be possible to consider a sum over all pivot words, however we found that this approach was less successful due to noise introduced by irrelevant pivot words.

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 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.

2.4 '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

[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 $M_{PT}(p_i,t) \leq 1$.

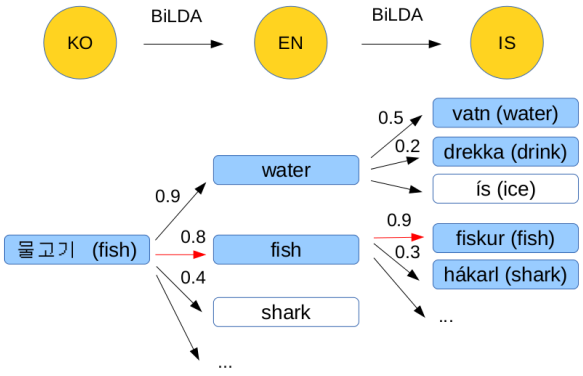

Figure 2: Pivot model: an illustration of the beam search algorithm using $b = 2$.

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 2.2.

This approach has the advantages of avoiding noise and search errors that can be introduced by the pivot model in Section 2.3, 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.

3 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.

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.

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.

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.

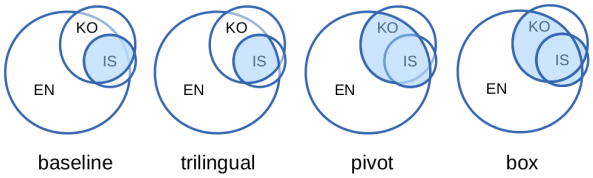

Figure 3: Subsets of Wikipedia data required for each method.

3.1 Settings

For each language we extracted the most frequent 100K nouns for our experiments, a vocabulary size over 10 times larger than in previous work [4]. The test data consisted of $N = 200$ (EN, KO, IS) translation triples. These were created by randomly selecting 200 nouns from our English Wikipedia vocabulary and translating these by hand into Korean and Icelandic. For comparison the same test data was used for all experiments. The test data contained only one correct translation for each word.

We used the PolyLDA++ tool to generate multilingual topic models. The training was run over 1000 iterations using $K = 2000$ topics and hyperparameters set as $\alpha = 50/K$ and $\beta = 0.01$.

The models were evaluated by generating a 10-best list of translations for each word in the test set. The top-1 accuracy and mean reciprocal rank (MRR) were then measured for the extracted lexicon.

3.2 Lexicon Extraction Experiment

Our experiments 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.

Figure 3 shows the data that is required (and was used) for each method. The results of the experiment are shown in Table 2.

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 2: Results of lexicon extraction experiment.

4 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.

The trilingual model is able to improve upon the baseline. It could be supposed 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 uses all the data, some of which is difficult to obtain. This shows that the pivot model can compete with a model trained directly on source-target data.

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 3 and 4 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 3) and less so on abstract concepts (Table 4), which often have more variation in their represention across languages.

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

Table 3: An example of a good translation: 'hjónaband' (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 4: An example of a bad translation: 'tilgangur' (purpose).

5 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.

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. This could improve the quality of the model in cases where there is not such a clear choice for an appropriate pivot language.

References

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, 2003.

[2] Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. Learning bilingual lexicons from monolingual corpora. In *ACL*, Vol. 2008, pp. 771–779, 2008.

[3] David Mimno, Hanna M Wallach, Jason Naradowsky, David A Smith, and Andrew McCallum. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pp. 880–889. Association for Computational Linguistics, 2009.

[4] Ivan Vulić, Wim De Smet, and Marie-Francine Moens. Identifying word translations from comparable corpora using latent topic models. In *ACL (Short Papers)*, pp. 479–484, 2011.