Topic Models + Word Alignment = A Flexible Framework for Extracting Bilingual Dictionary from Comparable Corpus

Xiaodong Liu, Kevin Duh and Yuji Matsumoto
Graduate School of Information Science
Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara 630-0192, Japan
{xiaodong-l,kevinduh,matsu}@is.naist.jp

Abstract

We propose a flexible and effective framework for extracting a bilingual dictionary from comparable corpora. Our approach is based on a novel combination of topic modeling and word alignment techniques. Intuitively, our approach works by converting a comparable *document*-aligned corpus into a parallel *topic*-aligned corpus, then learning word alignments using co-occurrence statistics. This *topic*-aligned corpus is similar in structure to the *sentence*-aligned corpus frequently used in statistical machine translation, enabling us to exploit advances in word alignment research. Unlike many previous work, our framework does not require any language-specific knowledge for initialization. Furthermore, our framework attempts to handle polysemy by allowing multiple translation probability models for each word. On a large-scale Wikipedia corpus, we demonstrate that our framework reliably extracts high-precision translation pairs on a wide variety of comparable data conditions.

1 Introduction

A machine-readable bilingual dictionary plays a very important role in many natural language processing tasks. In machine translation (MT), dictionaries can help in the domain adaptation setting (Daume III and Jagarlamudi, 2011). In cross-lingual information retrieval (CLIR), dictionaries serve as efficient means for query translation (Resnik et al., 2011). Many other multilingual applications also rely on bilingual dictionaries as integral components.

One approach for building a bilingual dictionary resource uses parallel sentence-aligned corpora. This is often done in the context of Statistical MT, using word alignment algorithms such as the IBM models (Brown et al., 1993; Och and Ney, 2003). Unfortunately, parallel corpora may be scarce for certain language-pairs or domains of interest (e.g., medical and microblog).

Thus, the use of comparable corpora for bilingual dictionary extraction has become an active research topic (Haghighi et al., 2008; Vulić et al., 2011). Here, a comparable corpus is defined as collections of document pairs written in different languages but talking about the same topic (Koehn, 2010), such as interconnected Wikipedia articles. The challenge with bilingual dictionary extraction from comparable corpus is that existing word alignment methods developed for parallel corpus cannot be directly applied.

We believe there are several desiderata for bilingual dictionary extraction algorithms:

1. **Low Resource Requirement**: The approach should not rely on language-specific knowledge or a large scale seed lexicon.

2. **Polysemy Handling**: One should handle the fact that a word form may have multiple meanings, and such meanings may be translated differently.

3. **Scalability**: The approach should run efficiently an massively large-scale datasets.

Our framework addresses the above desired points by exploiting a novel combination of topic models and word alignment, as shown in Figure 1. Intuitively, our approach works by first converting a comparable *document*-aligned corpus into a par-

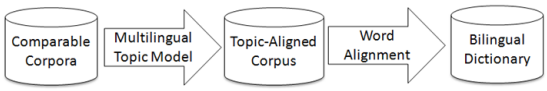

Figure 1: Proposed Framework

allel *topic*-aligned corpus, then apply word alignment methods to model co-occurence within topics. By employing topic models, we avoid the need for seed lexicon and operate purely in the realm of unsupervised learning. By using word alignment on topic model results, we can easily model polysemy and extract topic-dependent lexicons.

Specifically, let w^e be an English word and w^f be a French word. One can think of traditional bilingual dictionary extraction as obtaining (w^e, w^f) pairs in which the probability $p(w^e|w^f)$ or $p(w^f|w^e)$ is high. Our approach differs by modeling $p(w^e|w^f, t)$ or $p(w^f|w^e, t)$ instead, where t is a topic. The key intuition is that it is easier to tease out the translation of a polysemous word e given $p(w^f|w^e, t)$ rather than $p(w^f|w^e)$. A word may be polysemous, but given a topic, there is likely a one-to-one correspondence for the most appropriate translation. For example, under the simple model $p(w^f|w^e)$, the English word "free" may be translated into the Japanese word 自由 (as in free speech) or 無料 (as in free beer) with equal 0.5 probability; this low probability may cause both translation pairs to be rejected by the dictionary extraction algorithm. On the other hand, given $p(w^f|w^e, t)$, where t is "politics" or "shopping", we can allow high probabilities for both words depending on context.

Our contribution is summarized as follows:

- We propose a bilingual dictionary extraction framework that simultaneously achieves all three of the desiderata: low resource requirement, polysemy handling, and scalability. We are not aware of any previous works that address all three.
- Our framework is extremely flexible and simple-to-implement, consisting of a novel combination of existing topic modeling tools from machine learning and word alignment tools from machine translation.

2 Related Work

There is a plethora of research on bilingual lexicon extraction from comparable corpora, starting with seminal works of (Rapp, 1995; Fung and Lo, 1998). The main idea is to assume that translation pairs have similar contexts, i.e. the *distributional hypothesis*, so extraction consists of 3 steps: (1) identify context windows around words, (2) translate context words using a seed bilingual dictionary, and (3) extract pairs that have high resulting similarity. Methods differ in how the seed dictionary is acquired (Koehn and Knight, 2002; Déjean et al., 2002) and how similarity is defined (Fung and Cheung, 2004; Tamura et al., 2012). Projection-based approaches have also been proposed, though they can be shown to be related to the aforementioned distributional approaches (Gaussier et al., 2004); for example, Haghighi (2008) uses CCA to map vectors in different languages into the same latent space. Laroche (2010) presents a good summary.

Vulić et al. (2011) pioneered a new approach to bilingual dictionary extraction based on topic modeling approach which requires no seed dictionary. While our approach is motivated by (Vulić et al., 2011), we exploit the topic model in a very different way (explained in Section 4.2). They do not use word alignments like we do and thus cannot model polysemy. Further, their approach requires training topic models with a large number of topics, which may limit the scalability of the approach.

Recently, there has been much interest in multilingual topic models (MLTM) (Jagarlamudi and Daume, 2010; Mimno et al., 2009; Ni et al., 2009; Boyd-Graber and Blei, 2009). Many of these models give $p(t|e)$ and $p(t|f)$, but stop short of extracting a bilingual lexicon. Although topic models can group related e and f in the same topic cluster, the extraction of a high-precision dictionary requires additional effort. One of our contributions here is an effective way to do this extraction using word alignment methods.

3 System Components: Background

This section reviews MLTMs and Word Alignment, the main components of our framework. The knowledgeable readers may wish to skim this section for notation and move to Section 4, which describes our contribution.

3.1 Multilingual Topic Model

Any multilingual topic model may be used with our framework. We use the one by Mimno et

al. (2009), which extends the monolingual Latent Dirichlet Allocation model (Blei et al., 2003). Given a comparable corpus E in English and F in a foreign language, we assume that the document pair boundaries are known. For each document pair $d_i = [d_i^e, d_i^f]$ consisting of English document d_i^e and Foreign document d_i^f (where $i \in \{1, \ldots, D\}$, D is number of document pairs), we know that d_i^e and d_i^f talk about the same topics. While the monolingual topic model lets each document have its own so-called document-specific distribution over topics, the multilingual topic model assumes that documents in each tuple share the same topic prior (thus the comparable corpora assumption) and each topic consists of several language-specific word distributions. The generative story is shown in Algorithm 1.

for *each topic* k **do**
  for $l \in \{e, f\}$ **do**
    sample $\varphi_k^l \sim Dirichlet(\beta^l)$;
  end
end
for *each document pair* d_i **do**
  sample $\theta_i \sim Dirichlet(\alpha)$;
  for $l \in \{e, f\}$ **do**
    sample $z^l \sim Multinomial(\theta_i)$;
    for *each word* w^l *in* d_i^l **do**
      sample $w^l \sim p(w^l | z^l, \varphi^l)$;
    end
  end
end

Algorithm 1: Generative story for (Mimno et al., 2009). θ_i is the topic proportion of document pair d_i. Words w^l are drawn from language-specific distributions $p(w^l | z^l, \varphi^l)$, where language l indexes English e or Foreign f. Here pairs of language-specific topics φ^l are drawn from Dirichlet distributions with prior β^l.

3.2 Statistical Word Alignment

For a sentence-pair $(\boldsymbol{e}, \boldsymbol{f})$, let $\boldsymbol{e} = [w_1^e, w_2^e, \ldots w_{|\boldsymbol{e}|}^e]$ be the English sentence with $|\boldsymbol{e}|$ words and $\boldsymbol{f} = [w_1^f, w_2^f, \ldots w_{|\boldsymbol{f}|}^f]$ be the foreign sentence with $|\boldsymbol{f}|$ words. For notation, we will index English words by i and foreign words by j. The goal of word alignment is to find an alignment function $a : i \rightarrow j$ mapping words in $\boldsymbol{e}$ to words in $\boldsymbol{f}$ (and vice versa).

We will be using IBM Model 1 (Brown et al., 1993; Och and Ney, 2003), which proposes the following probabilistic model for alignment:

$$p(\boldsymbol{e}, a, |\boldsymbol{f}) \approx \prod_{i=1}^{|\boldsymbol{e}|} p(w_i^e | w_{a(i)}^f) \qquad (1)$$

Here, $p(w_i^e | w_{a(i)}^f)$ captures the translation probability of the English word at position i from the foreign word at position $j = a(i)$, where the actual alignment a is a hidden variable, and training can be done via EM. Although this model does not incorporate much linguistic knowledge, it enables us to *find correspondence between distinct objects from paired sets*. In machine translation, the distinct objects are words from different languages while the paired sets are sentence-aligned corpora. In our case, our distinct objects are also words from distinct languages but our pair sets will be *topic*-aligned corpora.

4 Proposed Framework for Bilingual Dictionary Extraction

The general idea of our proposed framework is sketched in Figure 1: First, we run a multilingual topic model to convert the comparable corpora to topic-aligned corpora. Second, we run a word alignment algorithm on the topic-aligned corpora in order to extract translation pairs. The innovation is in how this topic-aligned corpora is defined and constructed, the link between the two stages. We describe how this is done in Section 4.1 and show how existing approaches are subsumed in our general framework in Section 4.2.

4.1 Topic-Aligned Corpora

Suppose the original comparable corpus has D document pairs $[d_i^e, d_i^f]_{i=1,\ldots,D}$. We run a multilingual topic model with K topics, where K is user-defined (Section 3.1). The topic-aligned corpora is defined hierarchically as a *set of sets*: On the first level, we have a set of K topics, $\{t_1, \ldots, t_k, \ldots, t_K\}$. On the second level, for each topic t_k, we have a set of D "word collections" $\{C_{k,1}, \ldots, C_{k,i}, \ldots, C_{k,D}\}$. Each word collection $C_{k,i}$ represents the English and foreign words that occur simultaneously in topic t_k and document d_i.

For clarity, let us describe the topic-aligned corpora construction process step-by-step together with a flow chart in Figure 2:

1. Train a multilingual topic model.

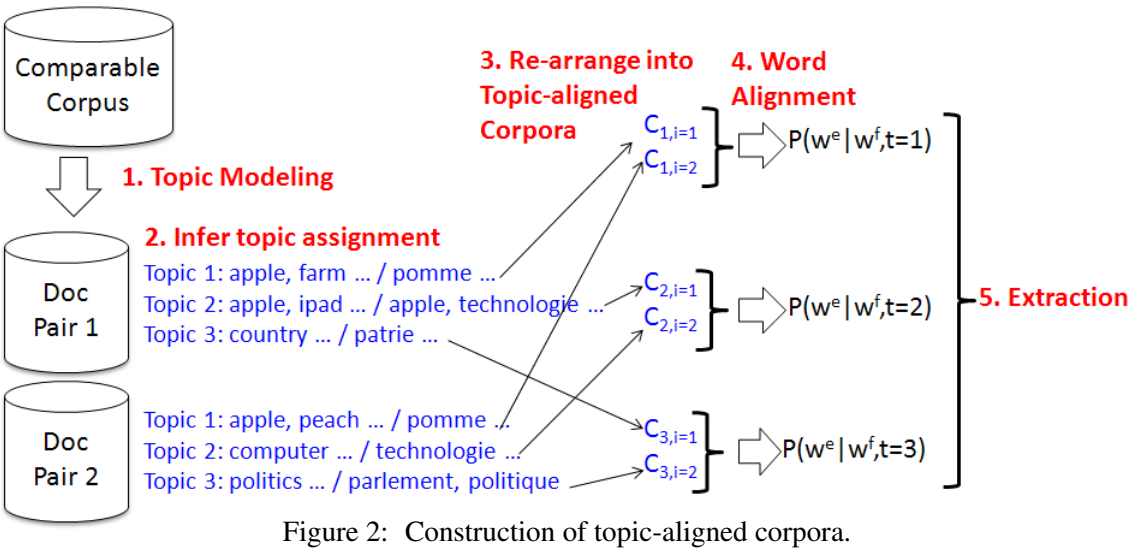

Figure 2: Construction of topic-aligned corpora.

2. Infer a topic assignment for each token in the comparable corpora, and generate a list of word collections $C_{k,i}$ occurring under a given topic.

3. Re-arrange the word collections such that $C_{k,i}$ belonging to the same topic are grouped together. This resulting set of sets is called topic-aligned corpora, since it represents word collections linked by the same topics.

4. For each topic t_k, we run IBM Model 1 on $\{C_{k,1}, \ldots, C_{k,i}, \ldots, C_{k,D}\}$. In analogy to statistical machine translation, we can think of this dataset as a parallel corpus of D "sentence pairs", where each "sentence pair" contains the English and foreign word tokens that co-occur under the same topic and the same document. Note that word alignment is run independently for each topic, resulting in K topic-dependent lexicons $p(w^e|w^f, t_k)$.

5. To extract a bilingual dictionary, we find pairs (w^e, w^f) with high probability under the model:

$$p(w^e|w^f) = \sum_k p(w^e|w^f, t_k)p(t_k|w^f) \quad (2)$$

The first term is the topic-dependent bilingual lexicon from Step 4; the second term is the topic posterior from the topic model in Step 1.

In practice, we will compute the probabilities of Equation 2 in both directions: $p(w^e|, w^f)$ as in Eq. 2 and $p(w^f|w^e) = \sum_k p(w^f|w^e, t_k)p(t_k|w^e)$. The bilingual dictionary can then be extracted based on a probabilities threshold or some bidirectional constraint. We choose to use a bidirectional constraint because it gives very high-precision dictionaries and avoid the need to tune probability thresholds. A pair $(\tilde{e}, \tilde{f})$ is extracted if the following holds:

$$\tilde{e} = \arg\max_e p(e|f = \tilde{f}); \tilde{f} = \arg\max_f p(f|e = \tilde{e}) \quad (3)$$

To summarize, the main innovation of our approach is that we allow for polysemy as topic-dependent translation explicitly in Equation 2, and use a novel combination of topic modeling and word alignment techniques to compute the term $p(w^e|w^f, t_k)$ in an unsupervised fashion.

4.2 Alternative Approaches

To the best of our knowledge, (Vulić et al., 2011) is the only work focuses on using topic models for bilingual lexicon extraction like ours, but they exploit the topic model results in a different way. Their "Cue Method" computes:

$$p(w^e|w^f) = \sum_k p(w^e|t_k)p(t_k|w^f) \quad (4)$$

This can be seen as a simplification of our Eq. 2, where Eq. 4 replaces $p(w^e|t_k, w^f)$ with the simpler $p(w^e|t_k)$. Another variant is the so-called Kullback-Liebler (KL) method, which scores translation pairs by $-\sum_k p(t_k|w^e) \log p(t_k|w^e)/p(t_k|w^f)$. In either case, their contribution is the use of topic-word distributions like $p(t_k|w^f)$ or $p(w^f|t_k)$ to compute translation probabilities.[1] Our formulation can be considered more general because we do not have the strong assumption that w^e is independent of

[1] A third variant uses TF-IDF weighting, but is conceptually similar and have similar results.

w^f given t_k, and focus on estimating $p(w^e|w^f, t_k)$ directly with word alignment methods.

5 Experimental Setup

5.1 Data Set

We perform experiments on the Kyoto Wiki Corpus[2]. We chose this corpus because it is a *parallel* corpus, where the Japanese edition of Wikipedia is translated manually into English sentence-by-sentence. This enables us to use standard word alignment methods to create a gold-standard lexicon for large-scale automatic evaluation.[3]

From this parallel data, we prepared several datasets at successively lower levels of comparability. As shown in Table 1, **Comp100%** is a comparable version of original parallel data, deleting all the sentence alignments but otherwise keeping all content on both Japanese and English sides. **Comp50%** and **Comp20%** are harder datasets that keep only 50% and 20% (respectively) of random English sentences per documents. We further use a *real* comparable corpus (**Wiki**)[4], which is prepared by crawling the online English editions of the corresponding Japanese articles in the Kyoto Wiki Corpus. The **Comp** datasets are controlled scenarios where all English content is guaranteed to have Japanese translations; no such guarantee exists in our **Wiki** data.

Dataset	#doc	#sent(e/j)	#voc(e/j)
Comp100%	14k	472k/472k	152k/116k
Comp50%	14k	236k/472k	100k/116k
Comp20%	14k	94k/472k	62k/116k
Wiki	3.6k	127k/163k	88k/61k

Table 1: Datasets: the number of document pairs (#doc), sentences (#sent) and vocabulary size (#voc) in English (e) and Japanese (j). For pre-processing, we did word segmentation on Japanese using Kytea (Neubig et al., 2011) and Porter stemming on English. A TF-IDF based stop-word lists of 1200 in each language is applied. #doc is smaller for **Wiki** because not all Japanese articles in **Comp100%** have English versions in Wikipedia during the crawl.

5.2 Experimental Results

1. How does the proposed framework compare to previous work?

We focus on comparing with previous topic-modeling approaches to bilingual lexicon extraction, namely (Vulić et al., 2011). The methods are:

- **Proposed**: The proposed method which exploits a combination of topic modeling and word alignment to incorporate topic-dependent translation probabilities (Eq. 2).
- **Cue**: From (Vulić et al., 2011), i.e. Eq. 4.
- **JS**: From (Vulić et al., 2011). Symmetrizing KL by Jensen-Shannon (JS) divergence improves results, so we report this variant.[5]

We also have a baseline that uses no topic models: **IBM-1** runs IBM Model 1 directly on the comparable dataset, assuming each document pair is a "sentence pair".

Figure 3 shows the ROC (Receiver Operating Characteristic) Curve on the **Wiki** dataset. The ROC curve lets us observe the change in Recall as we gradually accept more translation pairs as dictionary candidates. In particular, it measures the true positive rate (i.e. recall = $|\{Gold(e,f)\} \bigcap \{Extracted(e,f)\}|/\#Gold$) and false positive rate (fraction of false extractions over total number of extractions) at varying levels of thresholds. This is generated by first computing $p(e|f) + p(f|e)$ as the score for pair (e, f) for each method, then sorting the pairs by this score and successive try different thresholds.

The curve of the **Proposed** method dominates those of all other methods. It is also the best in Area-Under-Curve scores (Davis and Goadrich, 2006), which are 0.96, 0.90, 0.85 and 0.71, for **Proposed**, **IBM-1**, **Cue**, and **JS**, respectively.[6]

ROC is insightful if we are interested in comparing methods for all possible thresholds, but in practice we may desire a fixed operating point. Thus we apply the bidirectional heuristic of Eq.

[2] http://alaginrc.nict.go.jp/WikiCorpus/index_E.html

[3] We trained IBM Model 4 using GIZA++ for both directions $p(e|f)$ and $p(f|e)$. Then, we extract word pair $(\tilde{e}, \tilde{f})$ as a "gold standard" bilingual lexicon if it satisfies Eq. 3. Due to the large data size and the strict bidirectional requirement imposed by Eq. 3, these "gold standard" bilingual dictionary items are of high quality (94% precision by a manual check on 100 random items). Note sentence alignments are used only for creating this gold-standard.

[4] The English corresponding dataset, gold-standard and ML-LDA software used in our experiments are available at https://sites.google.com/site/buptxiaodong/home/resource

[5] Topic model hyperparameters for **Proposed**, **Cue**, and **JS** are $\alpha = 50/K$ and $\beta = 0.1$ following (Vulić et al., 2011).

[6] The Precision-Recall curve gives a similar conclusion. We do not show it here since the extremely low precision of **JS** makes the graph hard to visualize. Instead see Table 2.

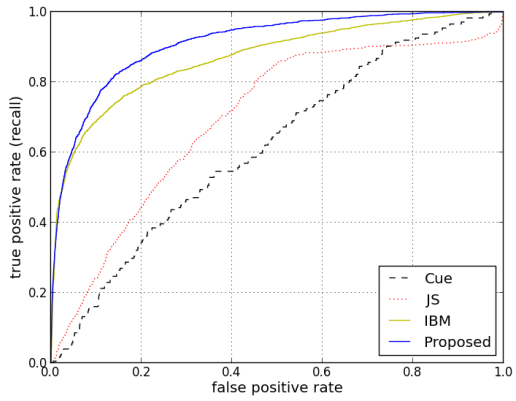

Figure 3: ROC curve on the **Wiki** dataset. Curves on upper-left is better. **Cue**, **JS**, **Proposed** all use K=400 topics. Note that **Proposed** is best.

K	Method	Prec	ManP	#Extracted
100	Cue	0.027	0.02	3800
	JS	0.013	0.01	3800
	Proposed	0.412	0.36	3800
400	Cue	0.059	0.02	2310
	JS	0.075	0.02	2310
	Proposed	**0.631**	**0.56**	2310
-	IBM-1	0.514	0.42	2310
-	IBM-1*	0.493	0.39	3714

Table 2: Precision on the **Wiki** dataset. K=number of topics. Precision (Prec) is defined as $\frac{|\{Gold(e,f)\} \bigcap \{Extracted(e,f)\}|}{\#Extracted}$. ManP is precision evaluated manually on 100 random items.

3 to extract a fixed set of lexicon for **Proposed**. For the other methods, we calibrated the thresholds to get the same number of extractions. Then we compare the precision, as shown in Table 2.

1. **Proposed** outperforms other methods, achieving 63% (automatic) precision and 56% (manual) precision.

2. The **JS** and **Cue** methods suffer from extremely poor precision. We found that this is due to insufficient number of topics, and is consistent with the results by (Vulić et al., 2011) which showed best results with $K >$ 2000. However, we could not train **JS/Cue** on such a large number of topics since it is computationally-demanding for a corpus as large as ours.[7] In this regard, the **Proposed** method is much more *scalable*, achieving good results with low K, satisfying one of original desiderata.[8]

3. **IBM-1** is doing surprisingly well, considering that it simply treats document pairs as sentence pairs. This may be due to some extent to the structure of the Kyoto Wiki dataset, which contains specialized topics (about Kyoto history, architecture, etc.), leading to a vocabulary-document co-occurrence matrix with sparse block-diagonal structure. Thus there may be enough statistics train **IBM-1** on documents.

[7]The experiments in (Vulić et al., 2011) has vocabulary size of 10k, compared to 150k in our experiments. We have attempted large $K \geq 1000$ but **Cue** did not finish after days.

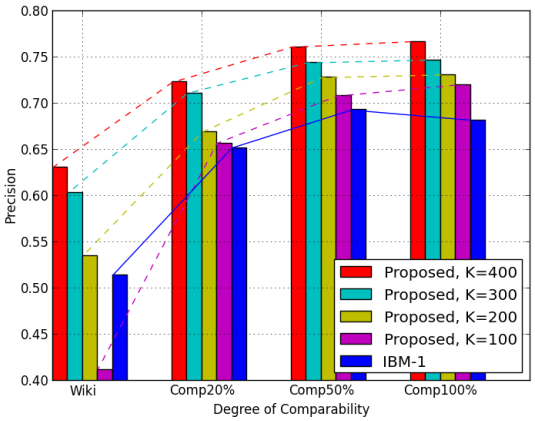

Figure 4: Robustness of method under different data conditions.

2. How does the proposed method perform under different degrees of "comparability"?

We next examined how our methods perform under different data conditions. Figure 4 plots the results in terms of Precision evaluated automatically. We observe that **Proposed (K=400)** is relatively stable, with a decrease of 14% Precision going from fully-comparable to real Wikipedia comparable corpora. The degradation for K=100 is much larger (31%) and therefore not recommended. We believe that robustness depends on K, because the

[8]We have a hypothesis as to why **Cue** and **JS** depend on large K. Eq. 2 is a valid expression for $p(w^e|w^f)$ that makes little assumptions. We can view Eq. 4 as simplifying the first term of Eq. 2 from $p(w^e|t_k, w^f)$ to $p(w^e|t_k)$. Both probability tables have the same output-space (w^e), so the same number of parameters is needed in reality to describe this distribution. By throwing out w^f, which has large cardinality, t_k needs to grow in cardinality to compensate for the loss of expressiveness.

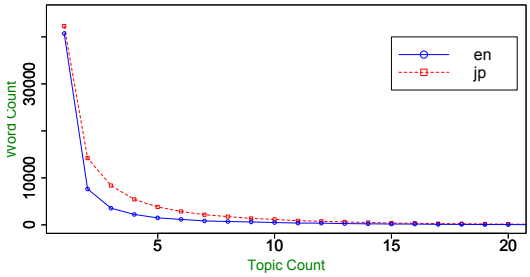

Figure 5: Power-law distribution of number of word types with X number of topics.

topic model of (Mimno et al., 2009) assumes one topic distribution per document pair. For low-levels of comparability, a small number of topics may not sufficiently model the differences in topical content. This suggests the use of hierarchical topic models (Haffari and Teh, 2009) or other variants in future work.

3. What are the statistical characteristics of topic-aligned corpora?

First, we show the word-topic distribution from multilingual topic modeling in the $K = 400$ scenario (first step of **Proposed**, **Cue**, and **JS**). For each word type w, we count the number of topics it may appear in, i.e. nonzero probabilities according to $p(w|t)$. Fig. 5 shows the number of word types that have x number of topics. This power-law is expected since we are modeling all words.[9]

Next we compute the statistics after constructing the topic-aligned corpora (Step 3 of Fig. 2). For each part of the topic-aligned corpora, we compute the ratio of distinct English word types vs. distinct Japanese word types. If the ratio is close to one, that means the partition into topic-aligned corpora effectively separates the skewed word-topic distribution of Fig 5. We found that the mean ratio averaged across topics is low at 1.721 (variance is 1.316), implying that within each topic, word alignment is relatively easy.

4. What kinds of errors are made?

We found that the proposed method makes several types of incorrect lexicon extractions. First, **Word Segmentation** "errors" on Japanese could make it impossible to find a proper English translation (e.g., 高市皇子 should translate to "Prince-Takechi" but system proposes "Takechi"). Second, an unrelated word pair (w^e, w^f) may be incorrectly placed in the same topic, leading to an **Incorrect Topic** error. Third, even if (w^e, w^f) intuitively belong to the same topic, they may not be direct translations; an extraction in this case would be a **Correct Topic, Incorrect Alignment** error (e.g. もんじゃ焼き, a particular panfried snack, is incorrectly translated as "panfry").

Table 3 shows the distribution of error types by a manual classification. **Incorrect Alignment** errors are most frequent, implying the topic models are doing a reasonable job of generating the *topic*-aligned corpus. The amount of **Incorrect Topic** is not trivial, though, so we would still imagine more advanced topic models to help. **Segmentation** errors are in general hard to solve, even with a better word segmenter, since in general one-to-one cross-lingual word correspondence is not consistent–we believe the solution is a system that naturally handles multi-word expressions (Baldwin, 2011).

Word Segmentation Error	14
Incorrect Topic	29
Correct Topic, Incorrect Alignment	40
Reason Unknown	7

Table 3: Counts of various error types.

5. What is the computation cost?

Timing results on a 2.4GHz Opteron CPU for various steps of **Proposed** and **Cue** are shown in Table 5. The proposed method is 5-8 times faster than **Cue**. For **Proposed**, computation time is dominated by topic modeling while GIZA++ on topic-aligned corpora is extremely fast. **Cue** additionally suffers from computational complexity in calculating Eq.4, especially when both $p(w^e|t_k)$ and $p(t_k|w^f)$ have high cardinality. In comparison, calculating Eq.2 is fast since $p(w^e|w^f, t_k)$ is in practice quite sparse.

6. What topic-dependent lexicons are learned and do they capture polysemy?

In our evaluation so far, we have only produced an one-to-one bilingual dictionary (due to the bidirectionality constraint of Eq.3). We have seen how topic-dependent translation models $p(w^f|w^e, t_k)$ is important in achieving good results. However, Eq.2 marginalizes over the topics so we do not know what topic-dependent lexicons are learned.

[9] This means that it is not possible to directly extract lexicon by taking the cross-product (w^f, w^e) of the top-n words in $p(w^f|t_k)$ and $p(w^e|t_k)$ for the same topic t_k, as suggested by (Mimno et al., 2009). When we attempted to do this, using top-2 words per $p(w^f|t_k)$ and $p(w^e|t_k)$, we could only obtain precision of 0.37 for 1600 extractions. This skewed distribution similarly explains the poor performance of **Cue**.

English	Japanese1(gloss), Japanese2(gloss)
interest	関心 (a sense of concern), 利息 (a charge of money borrowing)
count	数え(act of reciting numbers), 伯爵 (nobleman)
free	自由(as in "free" speech), 無料 (as in "free" beer)
blood	血縁(line of descent), 血 (the red fluid)
demand	需要(as noun), 要求(as verb)
draft	提案(as verb), 草稿 (as noun)
page	ページ (one leaf of e.g. a book), 侍童 (youthful attendant)
staff	スタッフ(general personel), 参謀 (as in political "chief of staff")
director	長官 (someone who controls), 理事 (board of directors) 監督 (movie director)
beach	浜(area of sand near water), 海水浴(leisure spot at beach)
actor	役者 (theatrical performer), 俳優 (movie actor)

Table 4: Examples of topic-dependent translations given by $p(w^f|w^e, t_k)$. The top portion shows examples of polysemous English words. The bottom shows examples where English is not decisively polysemous, but indeed has distinct translations in Japanese based on topic.

K	topic	giza	Eq.2	Eq.4	Prp	Cue
100	180	3	20	1440	203	1620
200	300	3	33	2310	336	2610
400	780	5	42	3320	827	4100

Table 5: Wall-clock times in minutes for Topic Modeling (topic), Word Alignment (giza), and $p(w^e|w^f)$ calculation. Overall time for **Proposed** (Prp) is topic+giza+Eq.2 and for **Cue** is topic+Eq.4.

Here, we explore the model $p(w^f|w^e, t_k)$ learned at Step 4 of Figure 2 to see whether it captures some of the polysemy phenomenon mentioned in the desiderata. It is not feasible to automatically evaluate topic-dependent dictionaries, since this requires "gold standard" of the form (e, f, t). Thus we cannot claim whether our method successfully extracts polysemous translations. Instead we will present some interesting examples found by our method. In Table 4, we look at potentially polysemous English words w^e, and list the highest-probability Japanese translations w^f conditioned on different t_k. We found many promising cases where the topic identification helps divide the different senses of the English word, leading to the correct Japanese translation achieving the highest probability.

6 Conclusion

We proposed an effective way to extract bilingual dictionaries by a novel combination of topic modeling and word alignment techniques. The key innovation is the conversion of a comparable *document*-aligned corpus into a parallel *topic*-aligned corpus, which allows word alignment techniques to learn topic-dependent translation models of the form $p(w^e|w^f, t_k)$. While this kind of topic-dependent translation has been proposed for the parallel corpus (Zhao and Xing, 2007), we are the first to enable it for comparable corpora. Our large-scale experiments demonstrated that the proposed framework outperforms existing baselines under both automatic metrics and manual evaluation. We further show that our topic-dependent translation models can capture some of the polysemy phenomenon important in dictionary construction. Future work includes:

1. Exploring other topic models (Haffari and Teh, 2009) and word alignment techniques (DeNero and Macherey, 2011; Mermer and Saraclar, 2011; Moore, 2004) in our framework.

2. Extract lexicon from massive multilingual collections. Mausum (2009) and Shezaf (2010) show that language pivots significantly improve the precision of distribution-based approaches. Since multilingual topic models can easily be trained on more than 3 languages, we expect it will give a big boost to our approach.

Acknowledgments

We thank Mamoru Komachi, Shuhei Kondo and the anonymous reviewers for valuable discussions and comments. Part of this research was executed under the Commissioned Research of National Institute of Information and Communications Technology (NICT), Japan.

References

Timothy Baldwin. 2011. Mwes and topic modelling: enhancing machine learning with linguistics. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, MWE '11, pages 1–1, Stroudsburg, PA, USA. Association for Computational Linguistics.

D. Blei, A. Ng, and M. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*.

Jordan Boyd-Graber and David M. Blei. 2009. Multilingual topic models for unaligned text. In *UAI*.

P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2).

Hal Daume III and Jagadeesh Jagarlamudi. 2011. Domain adaptation for machine translation by mining unseen words. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 407–412, Portland, Oregon, USA, June. Association for Computational Linguistics.

Jesse Davis and Mark Goadrich. 2006. The relationship between precision-recall and ROC curves. In *ICML*.

Hervé Déjean, Éric Gaussier, and Fatia Sadat. 2002. An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In *Proceedings of the 19th international conference on Computational linguistics - Volume 1*, COLING '02, pages 1–7.

John DeNero and Klaus Macherey. 2011. Model-based aligner combination using dual decomposition. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Pascale Fung and Percy Cheung. 2004. Mining verynon-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and em. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Pascale Fung and Yuen Yee Lo. 1998. Translating unknown words using nonparallel, comparable texts. In *COLING-ACL*.

Eric Gaussier, J.M. Renders, I. Matveeva, C. Goutte, and H. Dejean. 2004. A geometric view on bilingual lexicon extraction from comparable corpora. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 526–533, Barcelona, Spain, July.

Ghloamreza Haffari and Yee Whye Teh. 2009. Hierarchical dirichlet trees for information retrieval. In *NAACL*.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *Proceedings of ACL-08: HLT*, pages 771–779, Columbus, Ohio, June. Association for Computational Linguistics.

Jagadeesh Jagarlamudi and Hal Daume. 2010. Extracting multilingual topics from unaligned comparable corpora. In *ECIR*.

Philipp Koehn and Kevin Knight. 2002. Learning a translation lexicon from monolingual corpora. In *Proceedings of ACL Workshop on Unsupervised Lexical Acquisition*.

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1st edition.

Audrey Laroche and Philippe Langlais. 2010. Revisiting context-based projection methods for term-translation spotting in comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 617–625, Beijing, China, August. Coling 2010 Organizing Committee.

Mausam, Stephen Soderland, Oren Etzioni, Daniel S. Weld, Michael Skinner, and Jeff Bilmes. 2009. Compiling a massive, multilingual dictionary via probabilistic inference. In *ACL*.

Coskun Mermer and Murat Saraclar. 2011. Bayesian word alignment for statistical machine translation. In *ACL*.

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

Robert Moore. 2004. Improving IBM word alignment model 1. In *ACL*.

Graham Neubig, Yosuke Nakata, and Shinsuke Mori. 2011. Pointwise prediction for robust, adaptable japanese morphological analysis. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT) Short Paper Track*, pages 529–533, Portland, Oregon, USA, 6.

Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. 2009. Mining multilingual topics from wikipedia. In *WWW*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29(1):19–51, March.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*.

Philip Resnik, Douglas Oard, and Gina Levow. 2011. Improved cross-language retrieval using backoff translation. In *Proceedings of the First International Conference on Human Language Technology*.

Daphna Shezaf and Ari Rappoport. 2010. Bilingual lexicon generation using non-aligned signatures. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 98–107. Association for Computational Linguistics.

Akihiro Tamura, Taro Watanabe, and Eiichiro Sumita. 2012. 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*, pages 24–36, Jeju Island, Korea, July. Association for Computational Linguistics.

Ivan Vulić, Wim De Smet, and Marie-Francine Moens. 2011. 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*, pages 479–484, Portland, Oregon, USA, June. Association for Computational Linguistics.

Bing Zhao and Eric P. Xing. 2007. HM-BiTAM: Bilingual Topic Exploration, Word Alignment, and Translation. In *NIPS*.