NAIST-IS-DD1161205

Doctoral Dissertation

Bilingual Dictionary Extraction via Multilingual Topic Models

Xiaodong Liu

March 13, 2015

Department of Information Processing
Graduate School of Information Science
Nara Institute of Science and Technology

A Doctoral Dissertation
submitted to Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of ENGINEERING

Xiaodong Liu

Thesis Committee:

Professor Yuji Matsumoto	(Supervisor)
Professor Satoshi Nakamura	(Co-supervisor)
Associate Professor Masashi Shimbo	(Co-supervisor)
Assistant Professor Kevin Duh	(Co-supervisor)
Assistant Professor Hiroyuki Shindo	(Co-supervisor)

Bilingual Dictionary Extraction via Multilingual Topic Models [*]

Xiaodong Liu

Abstract

A machine readable bilingual dictionary plays a crucial role in many natural language processing tasks, such as statistical machine translation and cross-language information retrieval. In this thesis, we propose a framework for extracting a bilingual dictionary from comparable corpora by exploiting a novel combination of topic modeling and word aligners, such as the IBM models. Using a multilingual topic model, we first convert a comparable *document*-aligned corpus into a parallel *topic*-aligned corpus. This novel *topic*-aligned corpus is similar in structure to the *sentence*-aligned corpus frequently employed in statistical machine translation and allows us to extract a bilingual dictionary using a word alignment model.

The main advantages of our framework are that (1) no seed dictionary is necessary for bootstrapping the process, and (2) multilingual comparable corpora in more than two languages can also be exploited. In our experiments on a large-scale Wikipedia dataset, we demonstrate that our approach can extract higher precision dictionaries compared to previous approaches, and that our method improves further as we add more languages to the dataset.

Keywords:

Comparable Corpus, Multilingual Topic Model, Bilingual Dictionary

[*] Doctoral Dissertation, Department of Information Processing, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD1161205, March 13, 2015.

多言語トピックモデルを用いた対訳辞書の抽出 *

劉曉東

内容梗概

機械翻訳、情報抽出などの自然言語処理のタスクでは、対訳辞書は極めて重要な役割を果たしている。本研究では、トピックモデル及びIBMなどの単語アライナーを組み合わせ、対照コーパスから対訳辞書を抽出するフレームワークを提案する。多言語トピックモデルを利用することにより、まずドキュメント整列の対照コーパスをトピック整列の対訳コーパスに転換する。このトピック整列のコーパスは、機械翻訳によく使われる対訳コーパスに構造的に類似するため、単語整列モデルを利用し、対訳辞書を抽出することが可能となる。

このフレームワークの利点として：（１）プロセスのブートストラップとして、シード辞書は不必要である；（２）二つの言語以上の多言語対照コーパスも同様に利用できる。ウィキペディアのデータセットを使っての大規模実験では、提案手法が先行研究より高精度の辞書を抽出できることが実証された。それに、より多くの言語をデータセットに追加すれば、本手法の性能はより一層高くなる。

キーワード

対照コーパス、対訳辞書、多言語トピックモデル

[*] 奈良先端科学技術大学院大学 情報科学研究科 情報処理学専攻 博士論文, NAIST-IS-DD1161205, 2015年3月13日.

Acknowledgments

I am eternally grateful to my adviser, Professor Yuji Matsumoto, for continuous help and support throughout my PhD program in Nara Institute of Science and Technology (NAIST). I feel very honored to be your student and a member of Matsumoto-Ken.

I also would like to thank my supervisor, Kevin Duh. More than anyone, he taught me how to do research, how to prepare experiments, how to write papers and how to show the results to other researchers. He also encouraged me to talk with other researchers outside of lab. Furthermore, he always put me back to the right research directions and keeps me focus on the main problems.

Many thanks also to the other numbers of my committee, especially, Professor Satoshi Nakamura, Associate Professor Masashi Shimbo and Assistant Professor Hiroyuki Shindo for insightful suggestions and comments. I also would like to thank Mamoru Komachi, a previous assistant professor in NAIST, for his help and support, when I was a new member of Matsumoto-Ken. Special thanks to Yuko Kitawaga and Gakuseika staffs of NAIST.

Outside of NAIST, I also would like to thank my mentor Tomoya Iwakura, in Fujitsu, for exciting my interest in new research directions (e.g., structure learning). It was my honor to work with my mentors Jianfeng Gao, Xiaolong Li, Byungki Byun and other members of Deep Learning for Text Processing in MSR and Ranking&Intent Group in Bing. They were so helpful in teaching me about the most exciting research in the world, deep learning, ranking, query understanding and other techniques related to distributed learning. Specially, thanks to Fuji Ren, Xiaojie Wang and Caixia Yuan for recommending me to study overseas.

I would like to thank all my friends, who have accompanied me along the way. Finally, my utmost thanks to my family, brothers and sisters. I dedicate my thesis to them.

Contents

Acknowledgments **v**

1 Introduction **1**

1.1 Motivation 1

1.2 Contribution 4

1.3 Outline of Dissertation 5

2 Preliminaries **7**

2.1 Topic Models 7

2.1.1 Latent Dirichlet Allocation 7

2.1.2 Inference 10

2.2 Statistical Alignment Model 13

2.2.1 IBM Model 1 14

2.3 Summary 15

3 Bilingual Dictionary from Comparable Corpus via Topic Models **19**

3.1 Introduction 19

3.2 Related Work 21

3.2.1 Context Vector Approach 21

3.2.2 Projection-based Approach 23

3.2.3 Parallel Corpora based Approach 24

3.3 Proposed Framework for Bilingual Dictionary Extraction 25

3.3.1 Multilingual Topic Model 25

3.3.2 Topic-Aligned Corpora 27

3.3.3 Alternative Approaches 29

3.4 Experimental Setup 30

3.4.1 Data Set 30

3.4.2 Experimental Results 31

3.5 Summary 42

4 Enhancing Bilingual Dictionary Extraction via Multilingual Comparable Corpora 43

4.1 Introduction 43

4.2 Related Work 46

4.2.1 Pivot-Language in Lexicon Extraction and Machine Translation 47

4.2.2 Multilingual Word Representation Learning 47

4.3 Multilingual Topic Model 48

4.3.1 Generative Process 49

4.3.2 Inference 49

4.4 Experiments 51

4.4.1 Experiment Setting 52

4.4.2 Lexical Extraction Results: Comparison with Baselines 53

4.4.3 Lexicon Extraction Results: Additional Languages 54

4.4.4 Practical Issues: Model selection and Run-time 57

4.4.5 Detailed Analyses of Results 59

4.5 Summary 60

5 Hybrid Systems: Boosting for Context Vector Approach 61

5.1 Proposed approach 61

5.2 Experiment 62

5.3 Summary 64

6 Conclusion and Future Work 65

6.1 Conclusion 65

6.2 Future Work 65

Bibliography 67

List of Figures

1.1 An example of a *parallel* corpus. 2
1.2 An instance of comparable corpus. 3

2.1 Graphic model representation of LDA. 9

3.1 The proposed framework for a bilingual dictionary extraction. 20
3.2 An example of context vector approach. 22
3.3 An example of projection-based approach. 23
3.4 An example of Parallel Corpora based Approach. 24
3.5 Graphical representation of multilingual topic model. 26
3.6 Construction of topic-aligned corpora. 28
3.7 Per-word Likelihood by number of topics. 32
3.8 ROC curve on the **Wiki** dataset. 35
3.9 Robustness of method under different data conditions. 37
3.10 Power-law distribution of number of word types with X number of topics. 39
3.11 Comparison of different word alignment tools. 40

4.1 The rates of inter-linked different languages. 44
4.2 The *full-connected* comparable corpora. 44
4.3 The *partially-connected* comparable corpora. 45
4.4 The proposed framework for English-Japanese dictionaries extraction from multilingual comparable corpora. 46
4.5 Graphical representation of partially-connected multilingual topic model. 50
4.6 Effect of using additional languages: Precision-vs-threshold curve. . . 55
4.7 Effect of using additional languages: Number of extraction-vs-threshold curve. 56
4.8 Precision by number of topics (K). 57
4.9 Per-word Likelihood by number of topics. 58
4.10 Training time of MLTM by dataset and number of topics. 59

5.1 Hybrid system for bilingual dictionary extraction. 62
5.2 Comparison of different seeds for the context vector approach. 63
5.3 Comparison of different similarity metrics. 64

List of Tables

3.1 Statistics of KyotoWiki Dataset. 31
3.2 Examples of topic-word distribution part 1. 33
3.3 Examples of topic-word distribution part 2. 34
3.4 Precision on the **Wiki** dataset. 36
3.5 Counts of various error types. 40
3.6 Examples of topic-dependent translations given by $p(w^f|w^e,t_k)$. 41
3.7 Running time of the proposed framework. 41

4.1 Statistics of our multilingual **Wiki** crawl dataset. 52
4.2 Statistics of stop-words in different languages. 53
4.3 Comparison with baselines using the Japanese-English part of **Wiki** dataset. 54
4.4 Comparison of proposed method using additional languages. 55
4.5 Running time analysis. 60
4.6 Error analysis on multilingual case. 60

Chapter 1

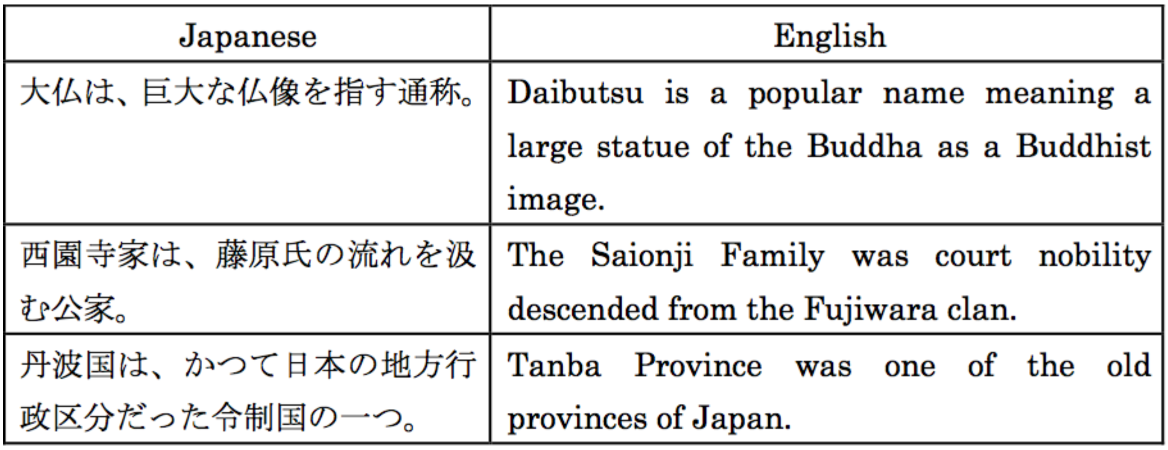
Introduction

This chapter sets the topic of the dissertation. Section 1.1 discusses the background and the motivation for studying bilingual lexicon extraction from *comparable* corpora. Then, Section 1.2 summarizes the contributions of this thesis. Finally, the outline of the dissertation is provided in Section 1.3.

1.1 Motivation

A machine-readable bilingual dictionary plays a very important role in many natural language processing tasks. In Statistical Machine Translation (SMT), dictionaries can help in the domain adaptation setting [13]. In Cross-Lingual Information Retrieval (CLIR), dictionaries serve as efficient means for query translation [52]. Many other multilingual applications also rely on bilingual dictionaries as integral components. For example, Volkova et al., [60] used a bilingual dictionary to analyze multilingual sentiment in social media; Zhang et al,. [64] incorporated a Chinese-English dictionary into a probabilistic topic model to explore bilingual latent topics in Chinese and English texts.

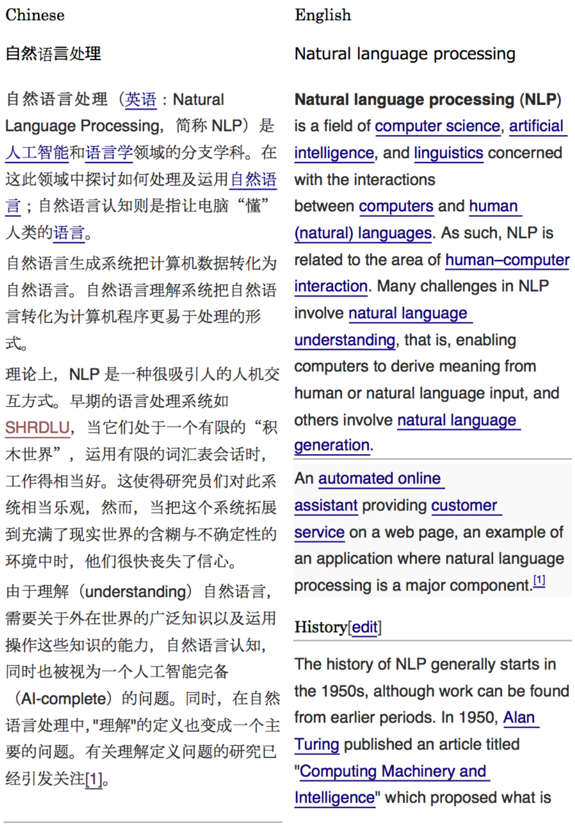
In the last two decades, researchers have focused on building bilingual dictionaries either from *parallel* corpora or *comparable* corpora. A *parallel* corpus usually refers to a collection of texts, which consist of sentence-to-sentence translation between two languages, as shown in Figure 1.1. On the other hand, a *comparable* corpus is defined as a collection of document pairs written in different languages, but talking about the same topic [31], such as interconnected Wikipedia articles, as shown Figure 1.2.



Japanese	English
大仏は、巨大な仏像を指す通称。	Daibutsu is a popular name meaning a large statue of the Buddha as a Buddhist image.
西園寺家は、藤原氏の流れを汲む公家。	The Saionji Family was court nobility descended from the Fujiwara clan.
丹波国は、かつて日本の地方行政区分だった令制国の一つ。	Tanba Province was one of the old provinces of Japan.

Figure 1.1: An example of a *parallel* corpus. Note that each line denotes a Japanese-English pair, which can be directly translated into each other.

One typical approach for building a bilingual dictionary resource uses *parallel corpora*. This is often done in the context of SMT, using word alignment algorithms such as the IBM models [12, 48]. Unfortunately, parallel corpora may be scarce for certain language-pairs or domains of interest (e.g., medical and microblog). On the other hand, comparable corpora are more abundant than parallel corpora. Thus, the use of comparable corpora for bilingual dictionary extraction has become an active research topic [24, 61, 39]. The challenge with bilingual dictionary extraction from comparable corpus is that existing word alignment methods developed for parallel corpus cannot be directly applied because of assumptions of sentence alignment. Some researchers have tried a framework by 1) converting comparable corpora into parallel corpora and then 2) using word alignment models [55]. However, such kind of models rely on large seed dictionaries and complicated features, e.g., orthographic features and translation features derived from SMT model. On the other hand, the standard approach (also known as context vector approach) usually needs large seed dictionaries to boost new translation pairs under the *distributional hypothesis* [51, 19, 57]. However, such kind of large seed dictionaries are difficult to obtain. Moreover, polysemy words, which have multiple meanings or senses, e.g., *bank* referring to either "the land alongside to a river or lake" or "a financial establishment", may be translated into different words in foreign languages. Such a problem, which is always ignored, plagues the accuracy of bilingual lexicons.



Chinese

自然语言处理

自然语言处理（英语：Natural Language Processing，简称 NLP）是人工智能和语言学领域的分支学科。在这此领域中探讨如何处理及运用自然语言；自然语言认知则是指让电脑“懂”人类的语言。

自然语言生成系统把计算机数据转化为自然语言。自然语言理解系统把自然语言转化为计算机程序更易于处理的形式。

理论上，NLP 是一种很吸引人的人机交互方式。早期的语言处理系统如 SHRDLU，当它们处于一个有限的“积木世界”，运用有限的词汇表会话时，工作得相当好。这使得研究员们对此系统相当乐观，然而，当把这个系统拓展到充满了现实世界的含糊与不确定性的环境中时，他们很快丧失了信心。

由于理解（understanding）自然语言，需要关于外在世界的广泛知识以及运用操作这些知识的能力，自然语言认知，同时也被视为一个人工智能完备（AI-complete）的问题。同时，在自然语言处理中，"理解"的定义也变成一个主要的问题。有关理解定义问题的研究已经引发关注[1]。

English

Natural language processing

Natural language processing (NLP) is a field of computer science, artificial intelligence, and linguistics concerned with the interactions between computers and human (natural) languages. As such, NLP is related to the area of human–computer interaction. Many challenges in NLP involve natural language understanding, that is, enabling computers to derive meaning from human or natural language input, and others involve natural language generation.

An automated online assistant providing customer service on a web page, an example of an application where natural language processing is a major component.[1]

History[edit]

The history of NLP generally starts in the 1950s, although work can be found from earlier periods. In 1950, Alan Turing published an article titled "Computing Machinery and Intelligence" which proposed what is

Figure 1.2: An instance of interconnected Wikipedia articles (a.k.a comparable corpora). Note that the Chinese-English document pair cannot be translated into each other directly, but both articles are about the same story.

The motivation of this thesis is that we focus on developing a serial of algorithms to extract the bilingual dictionary from *comparable* corpora. Furthermore, to facilitate further work in the area, we release the extracted dictionaries by using our framework. Last, we believe there are several desiderata for bilingual dictionary extraction algorithms:

- **Low Resource Requirement**: The model should only use cheaper corpora, e.g., one should rely on *comparable corpora* instead of *parallel corpora*, and should not rely on any language-specific knowledge, e.g., suffix or prefix of specific language properties, or seed lexicons.
- **Polysemy Handling**: One should handle the fact that a word form may have

multiple meanings, known as *polysemy* phenomenon, e.g., the English word "free" may refer to the sense of "given without charge" that could be translated into Japanese as "無料", or, the sense of "not restrained" that can be translated into Japanese as "自由". Such kind of multi-meaning phenomenon, which is very common in many languages, may plague the accuracy of bilingual extraction systems and related multilingual processing systems, e.g., statistical machine translation and cross-language information retrieval systems.

- **Flexibility**: The approach should be very flexible to encoding additional information if available. For example, additional languages for comparable corpora can be encoded in our framework to enhance the accuracy of bilingual dictionary extraction, as shown in Chapter 4.

- **Scalability**: The approach should run efficiently on massively large-scale datasets. Precisely, our model is 70 times faster than the baseline models and work very well on small number of topics, as shown Chapter 3 and Chapter 4.

1.2 Contribution

Our contribution is summarized as follows:

- We propose a bilingual dictionary extraction framework that simultaneously achieves all the desiderata introduced in the previous section: low resource requirement, polysemy handling, flexibility and scalability.

- 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.

- We further propose a hybrid system for bilingual dictionary extraction by: 1) firstly, extracting seeds using the previous framework and then 2) boosting the context-vector based approaches.

- To facilitate further work in this area, all preprocessed data, extracted lexicons and topic modeling code are available at https://bitbucket.org/allenLao/topic-modeling-gibbs.

1.3 Outline of Dissertation

The outline of this dissertation is as follows.

In Chapter 2, we introduce some preliminaries, which include latent semantic models, such as Latent Dirichlet Allocation (LDA) and word alignment models (a.k.a IBM models), the main components of our framework.

In Chapter 3, we introduce our proposed framework, which is a novel combination of Multilingual Topic Model (MLTM) with the word alignment model. The basic idea is to

1. Segment the comparable corpora into *Topic-Aligned* corpora, which is the key component of our framework, via multilingual topic models.

2. Extract a bilingual lexicon using word alignment models.

We show that the proposed bilingual dictionary extraction framework is effective compared to other baseline systems. Furthermore, we show how our framework disambiguate polysemy in terms of topic specific translations, which is a translation probability, $p(w_f|w_e,t_k)$, given a topic, t_k and a source word w_e.

In Chapter 4, we extend the multilingual topic model to handle the *partial-aligned* comparable corpora, such as Wikipedia articles. Additionally, multilingual comparable corpora in more than two languages, which is becoming increasingly prevalent due to the spread of the multilingual web, are used to enhance the performance of our proposed framework, as introduced in Chapter 3. We show how we can improve the extraction of Japanese-English dictionaries using comparable data not only from Japanese and English, but also from other languages such as Chinese and French.

Chapter 5 shows an extension of our framework to boost the context vector approach by:

1. Extracting bilingual lexicons using our proposed models.

2. Using such lexicons as seeds to boost the context vector model.

Note that all our models and approaches do not use any extra lexicons or dictionaries, and all are in an unsupervised regime.

Finally, we give a summarization for our dissertation in Chapter 6. We discuss several open questions in bilingual lexicon extractions and future works as well.

Chapter 2

Preliminaries

2.1 Topic Models

In the last decade, a variety of probabilistic topic models, such as Probabilistic Latent Semantic Indexing (PLSI) [27] and Latent Dirichlet Allocation (LDA) [9], have been used to analyze the content of documents and the meaning of words. The basic idea behind those models is that a document is a mixture of topics, where a topic is a probability distribution over words. A topic model is a *generative model* for documents: it specifies a simple probabilistic procedure by which documents can be generated. In the following of this section, we mainly describe the basic topic model, LDA.

2.1.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation, proposed by Blei et al. [9], is an extension of PLSI [27] with the Dirichlet distribution prior to prevent the over-fitting problem. The aim of LDA is to explore the latent structure of semantic concepts (a.k.a topics), based on the co-occurrence between words, in a document collection. The experiments show that such learned topics can capture synonymy and polysemy, which are common phenomena in linguistics. Thus, LDA is widely used in Natural language processing systems [39, 36, 38], and information retrieval systems [46, 20]. To facilitate the understanding of LDA, we introduce the common notations as follows:

1. M: number of documents to generate (constant).

2. K: number of topics or mixture components (constant).

3. V: number of vocabularies in the corpus (constant).

4. α and η: hyperparameters of the Dirichlet distribution for the topic proportion and the topic-word distribution.

5. θ_m: parameters for the topic distribution of a given document d_m, $p(z|d_m)$. Note that θ_m is a vector of variables with K dimension. Thus, the topic distribution parameters for the whole corpus, which includes M documents, can be notated as an M by K matrix, Θ.

6. β_k: topic-word distribution of the mixture component of topic k, $p(w|z = k)$. Note that β_k is a vector of variables with V dimension. Thus, the global topic-word distribution for the whole corpus can be denoted by an K by V matrix, B.

7. N: length of a document, here modeled with a Poisson distribution with constant parameter [9].

8. $z_{m,n}$: index of topic for the n-th word in the m-th document.

9. $w_{m,n}$: the n-th word in the m-th document.

10. d_m: m-th document in the corpus.

11. D or W: the corpus with M documents, $\{d_m\}_{m=0}^{M}$.

Conventionally, LDA is represented as a directed graph in terms of the graphic model as shown in Figure 2.1. The basic idea is that documents are represented as random mixtures over topics, where a topic is a distribution over words. Its generative story is shown in Algorithm 1:

For each document d_m, a K-dimensional Dirichlet random variable θ_m can take values in the $(k-1)$-simplex, and has the following probability density on this simplex:

$$p(\theta_m|\alpha) = \frac{\Gamma(\sum_{i=k}^{K} \alpha_k)}{\prod_{k}^{K} \Gamma(\alpha_k)} \theta_1^{\alpha_1 - 1} ... \theta_K^{\alpha_K - 1} \tag{2.1}$$

On the other hand, for each topic k, its topic word distribution has a V-dimensional Dirichlet over the vocabulary V as:

$$p(\beta_k|\eta) = \frac{\Gamma(\sum_{i=v}^{V} \alpha_v)}{\prod_{v}^{V} \Gamma(\eta_v)} \beta_{k,1}^{\eta_1 - 1} ... \beta_{k,V}^{\eta_V - 1} \tag{2.2}$$

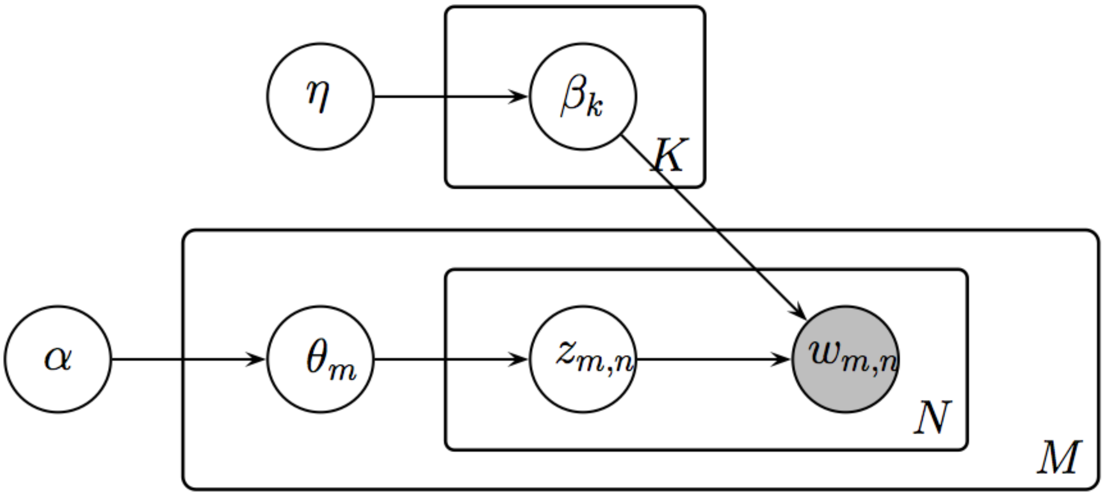


Figure 2.1: Graphic model representation of LDA. The boxes are ”plates” representing relocates. The outer plate represents documents, while the inner plate represents the repeated choice of the topics and words with in a document. The gray node $w_{m,n}$ denotes observations (a.k.a words in a document). M, N denote the number of documents in the corpus and the number of words in a specific document, respectively.

Given the hyperparameters α and η, the joint distribution of all hidden variables and observed document, denoted as a sequence of words w_m, is given by the following:

$$p(w_m, z, \theta_m, B|\alpha, \eta) = \prod_{k=1}^{K} p(\beta_k|\eta) \prod_{n=1}^{N} p(w_{m,n}|\beta_{z_{m,n}})p(z_{m,n}|\theta_m)p(\theta_m|\alpha) \tag{2.3}$$

Marginalizing all hidden variables, the likelihood of the given document, denoted as a sequence of words w_m, is computed by:

$$p(w_m|\alpha, \eta) = \iint p(\beta|\eta)p(\theta_m|\alpha) \cdot \prod_{n=1}^{N} p(w_{m,n}|\theta_m, \beta)d\beta d\theta_m \tag{2.4}$$

Finally, the likelihood of the whole corpus $W = \{w_m\}_{m=1}^{M}$ is determined by the product of the likelihood of all the independent documents as:

$$p(W|\alpha, \eta) = \prod_{m=1}^{M} p(w_m|\alpha, \eta) \tag{2.5}$$

Algorithm 1: Generative story for LDA.

```
for each topic $k$ do
    sample $\beta_k \sim Dirichlet(\eta)$
end
for each document $d_m$ in corpus do
    sample $\theta_m \sim Dirichlet(\alpha)$
    for each word $w_{m,n}$ in $d_m$ do
        sample $z \sim Multinomial(\theta_m)$
        sample $w_{m,n} \sim p(w_{m,n}|z,\beta)$
    end
end
```

2.1.2 Inference

The central computational problem for LDA is approximating the posterior given a document. Exact inference for LDA is generally intractable. Therefore, the approximate inference methods are always selected. For simplicity, we will only develop a collapsed Gibbs sampling [46, 25, 42], a type of Markov Chain Monte Carlo (MCMC), to estimate the posterior given a document in this thesis.

MCMC is a procedure for obtaining samples from complicated probability distributions, allowing a Markov chain to converge to the target distribution and then draw samples from the Markov chain. Each state of variable being sampled is assigned to a value, and transitions between states follow a simple rule. Here, the next state is reached by sequentially sampling all variables from their distributions when conditioned on the current values of all other variables and data. In the case of LDA, we only sample a topic assignment, $z_{m,n}$, given each word, $w_{m,n}$. For simplicity, we use an index $i = (m,n)$ to denote a word index or a topic index. Thus, a document is notated by a word vector, $w = \{w_i = t, w_{\neg i}\}$ and corresponding topic states is denoted by $z = \{z_i = k, z_{\neg i}\}$. Note that we use $w_{\neg i}$ to indicate the word vector excluding the word w_i and $z_{\neg i}$ indicates a topic vector excluding the i-th states. Thus, the conditional posterior for $z_i = k$ given the current states of topic assignments, z, words, w, and hyper parameters α and η, is computed by:

$$p(z_i = k|z_{\neg i}, w, \alpha, \eta) = \frac{p(w, z_i|\alpha, \eta)}{p(w, z_{\neg i}|\alpha, \eta)} \tag{2.6}$$

The problem becomes how to compute $p(w,z|\alpha,\eta)$[1]. From Figure 2.1, we further factorize the $p(w,z|\alpha,\eta)$ as follows:

$$p(w,z|\alpha,\eta) = p(w|z,\eta)p(z|\alpha) \tag{2.7}$$

Note that given variables topic index z and hyperparameter η, w and α are independent of each other according to the probability graph theories.

The first term of Eq 2.7, $p(w|z,\eta)$, is derived from a multinomial on the observed word, with the Dirichlet priori η. It can be further factorized as:

$$p(w|z,\eta) = \int p(w|z,\beta)p(\beta|\eta)d\beta. \tag{2.8}$$

The multinomial distribution and Dirichlet distribution is a pair of the *conjugate* distribution for each other. Therefore, the Eq 2.8 is also a Dirichlet distribution and can be computed by:

$$p(w|z,\eta) = \prod_{z=k}^{K} \frac{\Delta(n_z+\eta)}{\Delta(\eta)},\ n_z = \{n_z^{(v)}\}_{v=1}^{V}. \tag{2.9}$$

Where $\Delta(\eta)$ is computed by $\frac{\prod_{v=1}^{V}\Gamma(\eta_v)}{\Gamma(\sum_{v=1}^{V}\eta_v)}$[2] and $\Delta(n_z+\eta)$ is computed by $\frac{\prod_{v=1}^{V}\Gamma(\eta_v+n_z^{(v)})}{\Gamma(\sum_{v=1}^{V}\eta_v+n_z^{(v)})}$. $n_z^{(v)}$ denotes the number of time that a word, v, has been assigned to the topic z.

Similarly, we compete the second term of Eq 2.7 as:

$$\begin{aligned} p(z|\alpha) &= \int p(z|\theta)p(\theta|\alpha)d\theta \\ &= \prod_{m=1}^{M} \frac{\Delta(n_m+\alpha)}{\Delta(\alpha)},\ n_m = \{n_m^{(k)}\}_{k=1}^{K}. \end{aligned} \tag{2.10}$$

Where $\Delta(\alpha)$ is computed by $\frac{\prod_{k=1}^{K}\Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K}\alpha_k)}$ and $\Delta(n_m+\alpha)$ is computed by $\frac{\prod_{k=1}^{K}\Gamma(\alpha_k+n_m^{(k)})}{\Gamma(\sum_{k=1}^{K}\alpha_k+n_m^{(k)})}$. $n_m^{(k)}$ indicates the number of times that topic k has been assigned with a word of document m.

By putting the two terms together, Eq 2.7 becomes:

$$p(w,z|\alpha,\eta) = \prod_{z=k}^{K} \frac{\Delta(n_z+\eta)}{\Delta(\eta)} \prod_{m=1}^{M} \frac{\Delta(n_m+\alpha)}{\Delta(\alpha)}. \tag{2.11}$$

[1] The term $p(w,z_{\neg i}|\alpha,\eta)$ can be computed similarly. For simplicity, we omit the subscript i.

[2] The gamma function is defined for all complex numbers except the negative integers and zero as: $\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}dx$.

Note that this formula has the same structure as Eq 2.7. Similarly, $p(w, z_{\neg i}|\alpha, \eta)$ is computed:

$$p(w, z_{\neg i}|\alpha, \eta) = \prod_{z=k}^{K} \frac{\Delta(n_{z,\neg i} + \eta)}{\Delta(\eta)} \prod_{m=1}^{M} \frac{\Delta(n_{m,\neg i} + \alpha)}{\Delta(\alpha)}. \tag{2.12}$$

The update equation from which the Gibbs sampler draws from the hidden variable, Eq 2.6, yields:

$$\begin{aligned} p(z_i = k|z_{\neg i}, w, \alpha, \eta) &= \frac{p(w, z_i|\alpha, \eta)}{p(w, z_{\neg i}|\alpha, \eta)} \\ &\propto \frac{\Delta(n_z + \eta)}{\Delta(\eta)} \frac{\Delta(n_m + \alpha)}{\Delta(\alpha)} \frac{\Delta(\eta)}{\Delta(n_{z,\neg i} + \eta)} \frac{\Delta(\alpha)}{\Delta(n_{m,\neg i} + \alpha)} \\ &= \frac{\Delta(n_z + \eta)}{\Delta(n_{z,\neg i} + \eta)} \frac{\Delta(n_m + \alpha)}{\Delta(n_{m,\neg i} + \alpha)} \\ &= \frac{\prod_{v=1}^{V} \Gamma(\eta_v + n_z^{(v)})}{\Gamma(\sum_{v=1}^{V} \eta_v + n_z^{(v)})} \frac{\Gamma(\sum_{v=1}^{V} \eta_v + n_{z,\neg i}^{(v)})}{\prod_{v=1}^{V} \Gamma(\eta_v + n_{z,\neg i}^{(v)})} \\ &\cdot \frac{\prod_{t=1}^{K} \Gamma(\alpha_t + n_m^{(t)})}{\Gamma(\sum_{t=1}^{K} \alpha_t + n_m^{(t)})} \frac{\Gamma(\sum_{t=1}^{K} \alpha_t + n_{m,\neg i}^{(t)})}{\prod_{t=1}^{K} \Gamma(\alpha_t + n_{m,\neg i}^{(t)})} \\ &= \frac{n_{z=k,\neg i}^{(v)} + \eta_v}{\sum_{v=1}^{V} n_{z=k,\neg i}^{(v)} + \eta_v} \cdot \frac{n_{m,\neg i}^{(k)} + \alpha_k}{\sum_{t=1}^{K} n_{m,\neg i}^{(t)} + \alpha_t} \\ &= \underbrace{\frac{n_{z=k,\neg i}^{(v)} + \eta_v}{\sum_{v=1}^{V} n_{z=k,\neg i}^{(v)} + \eta_v}}_{\text{global topic word distribution}} \cdot \underbrace{\frac{n_{m,\neg i}^{(k)} + \alpha_k}{N_m + K\alpha_t - 1}}_{\text{local topic proportion}}. \end{aligned} \tag{2.13}$$

where the counts $n_{\cdot,\neg i}^{(\cdot)}$ indicate that the token i is excluded from the corresponding document or topic. To derive the gamma function, we use the fact $\Gamma(t+1) = t\Gamma(t)$, in other words, $\Gamma(t+1)/\Gamma(t) = t$. Note that from Eq 2.13, we observe that given hyperparamters α and η, words w in a document and its corresponding topic states z, the probability of a word w_i is governed by the global topic word distribution and the local topic proportion of a document. In the following derivatives of the other topic models, we will directly use this fact.

Finally, we obtain the variables, θ and β, which we are interested in, based on the state of the Markov chain w and z. Since the Dirichlet distribution is a conjugate distribution of the multinomial distribution, the probability of θ_n of the component $z =$

k, given the m-th document, w_m, its corresponding topic states, z, and hyperparameters α, is computed by:

$$\begin{aligned} p(\theta_{m,k}|w_m, z, \alpha) &= Dirichlet(\theta_m|n_m + \alpha) \\ &= \frac{n_m^{(k)} + \alpha_k}{\sum_{k=1}^{K} n_m^{(k)} + \alpha_k}. \end{aligned} \tag{2.14}$$

Similarly, the probability of β of the component $z = k$ and vocabulary v, given all the documents, w, its corresponding topic states, z, and hyperparameters η, is computed by:

$$\begin{aligned} p(\beta_k|w, z, \eta) &= Dirichlet(\beta_k|n_k + \eta) \\ &= \frac{n_k^{(v)} + \eta_v}{\sum_{v=1}^{V} n_k^{(v)} + \eta_v}. \end{aligned} \tag{2.15}$$

2.2 Statistical Alignment Model

The task of statistical machine translation is that given a foreign sentence, f, we seek an English sentence, e, that maximize probability $p(e|f)$,

$$\arg\max_e p(e|f). \tag{2.16}$$

Applying Bayes' rule, Eq 2.16 is equal to

$$\begin{aligned} \arg\max_e p(e|f) &= \arg\max_e \frac{p(f|e)p(e)}{p(f)} \\ &= \arg\max_e p(f|e)p(e) \end{aligned} \tag{2.17}$$

where $p(f|e)$ is referred as the translation model, which evaluates how likely an English sentence, e, is translated into a foreign sentence, f. The second term $p(e)$ is referred to as the language model, which is the chance that someone would say the English sentence e. This model is also known as ”the Noisy Channel Model” [31, 2, 40].

In the following sections, we focus on the translation model, $p(f|e)$, and omit the language model $p(e)$. Since it is difficult to compute $p(f|e)$ directly in the sentence level, it is always decomposed into words or phrases for different purposes.

Assume that e has l_e words and f has l_f words, thus there are $l_e * l_f$ different connections that can be drawn between them (each of the l_f foreign words can be aligned to any of the l_e English words). Conventionally, the connections of foreign words between the English words are modeled by an alignment function, $a(e, f)$, which is why it is referred as "alignment models".

Now, we factorize the $p(f|e)$ in terms of the conditional probability $p(f,a|e)$ as:

$$p(f|e) = \sum_{a} p(f,a|e) \tag{2.18}$$

where a is an alignment function to connect English words, e, with foreign words, f. Without loss of generality, the English sentence is denoted as $e = e_1^{l_e} = e_1 e_2 ... e_{l_e}$ with l_e words, and the foreign sentence is denoted as $f = f_1^{l_f} = f_1 f_2 ... f_{l_f}$, with l_f words. The alignment function, a, is defined as $a_1^{l_f} = a_1 a_2 ... a_{l_f}$, each of which has value between 0 and l such that if the word in position j of the foreign sentence is aligned to the word in position i of the English sentence, then $a_j = i$, and if it is not aligned to any English word, then $a_j = 0$. Further, $p(f,a|e)$ is written as:

$$p(f,a|e) = p(l_f|e) \prod_{j=1}^{l_f} p(a_j|a_1^{j-1}, f_1^{j-1}, l_f, e) p(f_j|a_1^j, f_1^{j-1}, l_f, e) \tag{2.19}$$

Since we only use the simplest statistical word alignment model, a.k.a IBM Model 1 [40], which will be introduced in the following sections.

2.2.1 IBM Model 1

In IBM model 1, we assume that $p(l_f|e)$ is independent of e and l_f, in other words, the length of a foreign sentence is independent of the given English sentence; the probability of the current state of the alignment function, $p(a_j|a_1^{j-1}, f_1^{j-1}, l_f, e)$, depends only on the length of the English sentence, l_e, which is equal to $\frac{1}{l_e+1}$; and that the translation probability of current foreign word f_j, $p(f_j|a_1^j, f_1^{j-1}, l_f, e)$, depends only on e_{a_j}, which is $p(f_j|a_1^j, f_1^{j-1}, l_f, e) = t(f_j|e_{a_j})$. Based on such assumptions, the Eq 2.19 is simplified as:

$$p(f,a|e) = \frac{\varepsilon}{(l_e+1)^{l_f}} \prod_{j=1}^{l_f} t(f_j|e_{a_j}). \tag{2.20}$$

Note that the alignment is determined by specifying the values of a_j for 0 to l_f, each of which can take any value from 0 to l_e. Thus, Eq 2.18 can be rewritten as:

$$\begin{aligned} p(f|e) &= \frac{\varepsilon}{(l_e+1)^{l_f}} \sum_{a_1=0}^{l_e} \cdots \sum_{a_{l_f}=0}^{l_f} \prod_{j=1}^{l_f} t(f_j|e_{a_j}) \\ &= \frac{\varepsilon}{(l_e+1)^{l_f}} \prod_{j=1}^{l_f} \sum_{a_j=0}^{l_e} t(f_j|e_{a_j}) \end{aligned} \tag{2.21}$$

The probabilities of alignment given the English and foreign sentences are computed by using Bayes' rule:

$$\begin{aligned} p(a|e,f) &= \frac{p(f,a|e)}{p(f|e)} \\ &= \prod_{j=1}^{l_f} \frac{t(f_j,e_{a_j})}{\sum_{a_j=0}^{l_e} t(f_j,e_{a_j})} \end{aligned} \tag{2.22}$$

To estimate the parameters of IBM model 1, Expectation Maximization (EM) algorithm shown in Algorithm 2, is used. EM algorithm is an iterative method for finding the maximum likelihood of $p(f|e)$ as Eq 2.21. Since it is in the probability space, given an English word e, the translation probability to a foreign word is constrained to $\sum_f t(f|e) = 1$. This problem can be done in terms of Lagrange multipliers as

$$L(t,\lambda) = \prod_{j=1}^{l_f} \frac{t(f_j,e_{a_j})}{\sum_{a_j=0}^{l_e} t(f_j,e_{a_j})} - \lambda_e(\sum_f t(f|e) - 1) \tag{2.23}$$

Taking partial derivatives of $L(t,\lambda)$ with respect to λ and t, respectively, and setting them to zero, we solve the equations and obtain the following equation:

$$t(f_j|e_{a_j}) = \frac{\sum_{(e,f)} c(f_j|e_{a_j};e,f)}{\sum_e \sum_{(e,f)} c(f_j|e_{a_j};e,f)}. \tag{2.24}$$

where the function c is a count function that collects evidence from a sentence pair (e,f) that a particular English word, e_{a_j}, is translated into a foreign word f_j.

2.3 Summary

In this chapter, we briefly introduced two key components to understand the proposed framework for bilingual lexicon extraction. More precisely, we reviewed one of

Algorithm 2: EM learning algorithm for IBM Model 1.

```
Input: sentence pairs (e, f)
Output: translation probabilities, $t(f|e)$
initialize $t(f|e)$ uniformly
while not converged do
    for each word e in English Vocabulary $V_e$ do
        for each word f in foreign vocabulary $V_f$ do
            $count(f|e) = 0$
        end
        $total(e) = 0$
    end
    for all sentence pairs (e,f) do
        for all words f in f do
            $total(f) = 0$
            for all words e in e do
                $total(f) += t(f|e)$
            end
        end
    end
    for all words f in f do
        for all words e in e do
            $count(f|e) += \frac{t(f|e)}{total(f)}$
            $total(e) += \frac{t(f|e)}{total(f)}$
        end
    end
    for all English words e do
        for all foreign words f do
            $t(f|e) = \frac{count(f|e)}{total(f)}$ (Eq 2.24)
        end
    end
end
[1cm]
```

the most popular probability models, LDA, which is used as a building block for many applications. Then, we briefly reviewed the word alignment model, especially IBM1.

Chapter 3

Bilingual Dictionary from Comparable Corpus via Topic Models

In this chapter, we propose a flexible framework for bilingual dictionary extraction via topic models and word alignment models. It is organized as follows: In Section 3.1, we introduce the background of bilingual dictionary extraction from comparable corpora. Then, related works are reviewed in Section 3.2. Our framework is proposed in Section 3.3. In Section 3.4, we conduct a serial of experiments to evaluate our systems. At last, we summarize this chapter.

3.1 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 [13]. In cross-lingual information retrieval (CLIR), dictionaries serve as efficient means for query translation [52]. Many other multi-lingual applications also rely on bilingual dictionaries as integral components.

Conversional approaches for building a bilingual dictionary resource use parallel corpora. This is often done in the context of Statistical MT, using word alignment algorithms such as the IBM models [12, 48]. 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 [24, 61]. Here, a comparable corpus is defined as collections of document pairs written in different languages but talking about the same topic [31],

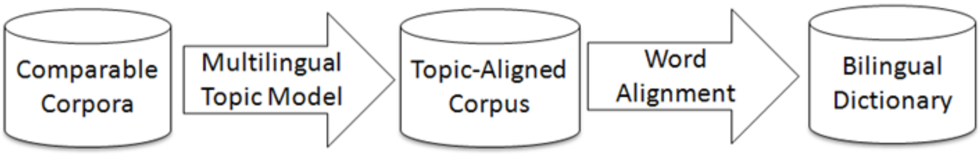


Figure 3.1: The proposed framework for a bilingual dictionary extraction. Multilingual topic model is used for converting a document-aligned comparable corpus to topic-aligned corpora. Given a topic, word alignment models are used to model co-occurrence across languages.

such as interconnected Wikipedia articles. The challenge with bilingual dictionary extraction from comparable corpora is that existing word alignment methods developed for parallel corpus cannot be directly applied to bilingual dictionary extraction from comparable corpora.

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 3.1. Intuitively, our approach works by first converting a comparable *document*-aligned corpus into a parallel *topic*-aligned corpus, then apply word alignment methods to model co-occurence within topics. By employing topic models, we avoid the need for a 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.

3.2 Related Work

The numerous works on bilingual lexicon from comparable corpora can be divided into two broad categories: context vector approaches (Section 3.2.1) and projection-based approaches (Section 3.2.2).

3.2.1 Context Vector Approach

The context vector approach, starting with seminal works of [51, 19], is built on the assumption that a word and its corresponding translation tend to appear in similar contexts across languages, also known as the *distributional hypothesis*. A typical context vector approach for the bilingual dictionary extraction consists of three steps, as shown in Figure 3.2:

1. Represent contexts of a word using an existing seed dictionary. This ranges from simple representations based on bag-of-words [51, 19] or TF-IDF of words in a context window [51], to more elaborate representations such as dependency trees [3].

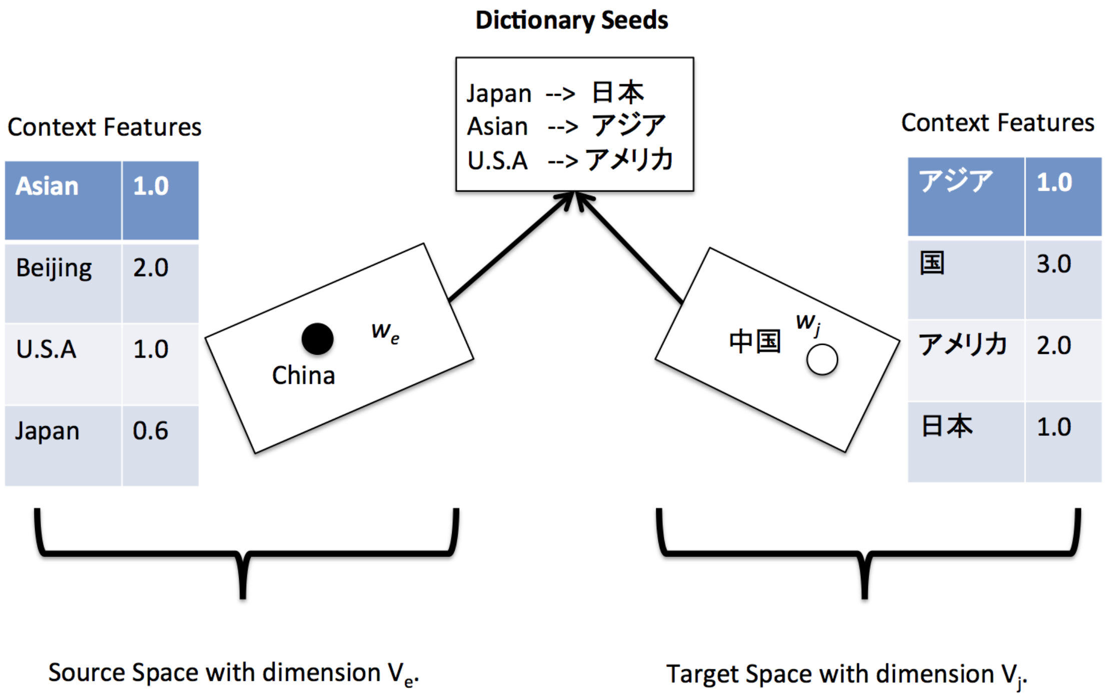


Figure 3.2: An example of context vector approach. Each word in the source language (English), e.g., *China* is represented as a vector of the context features, w_e=[Asian:1.0, Beijing:2.0, U.S.A:1.0, Japan:0.6], with V_e dimension, and each word in the target language (Japanese), e.g. 中国 is represented as a vector of the context features, w_j=[アジア:1.0, 国:3.0, アメリカ:2.0, 日本:1.0], with V_j dimension. The similarity, e.g, cosine similarity, between the two words, *China* and 中国, via an existing dictionary is: $score(w_e, w_j) = cosine(w_e, w_j) = \frac{1.0*1.0+1.0*2.0+0.6*1.0}{\sqrt{1.0^2+1.0^2+0.6^2}\sqrt{1.0^2+2.0^2+1.0^2}} = 0.957$. Note that words which are not in the seeds dictionary are not used to compute similarity score.

2. Measure similarity/distance between words in this common space, e.g., using cosine similarity [32, 19] or Manhattan distance [51].

3. Extract word pairs with high similarity.

Methods differ in how the seed dictionary is acquired [32, 16] and how similarity is defined [18, 57]. It is important to note that all these methods critically rely on a seed dictionary to ensure that word in different languages are represented in the same space. To alleviate the dependence on the size of the seed dictionary, Tamura et al., [57] have used an unsupervised label propagation method to improve robustness.

3.2.2 Projection-based Approach

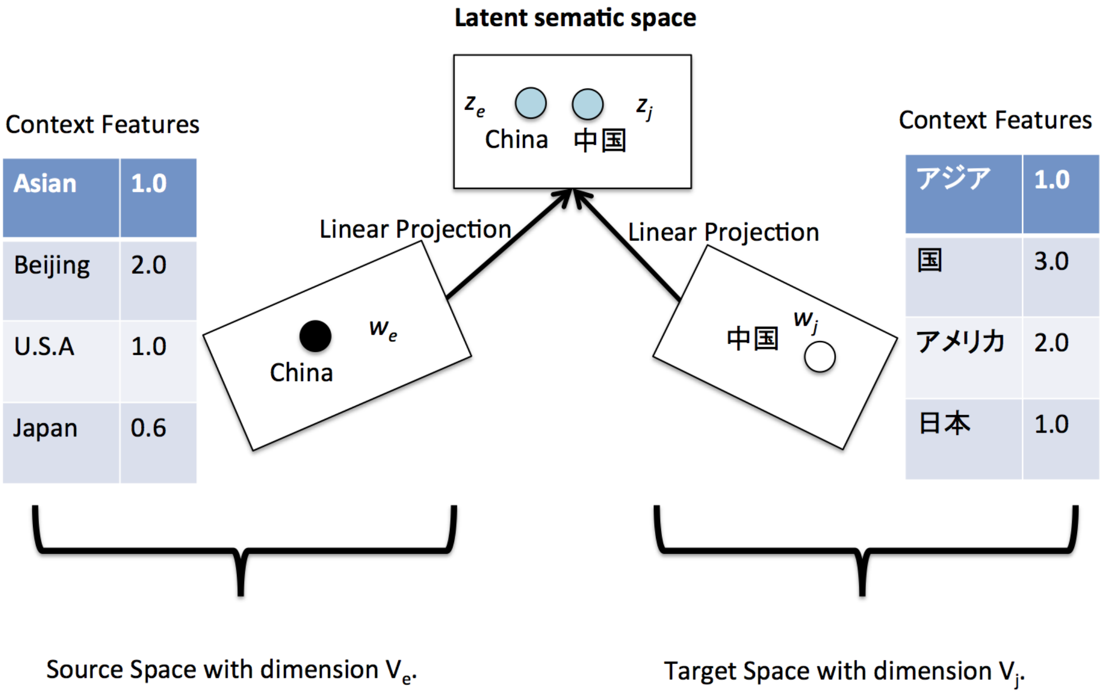


Figure 3.3: An example of projection-based approach. Usually, it is difficult to compute similarity between a source word, w_e, and a target word, w_j, directly, since they are in different spaces, which have different dimensions. The projection-based approach 1) first maps both source words with V_e dimension and target with V_j dimension into a latent semantic space with S, $z_e = M_e w_e$ and $z_j = M_j w_j$, where matrices M_e (V_e by S) and M_j (V_j by S) can be optimized by an EM algorithm; 2) then compute the similarly, i.e., cosine similarity, between a source word and a target word in semantic space, $score(w_e, w_j) = cosine(w_e, w_j) = cosine(M_e w_e, M_j w_j) = cosine(z_e, z_j)$.

Projection-based approaches have also been proposed, though they can be shown to be related to the aforementioned distributional approaches [21]; for example, Haghighi [24] uses canonical correlation analysis (CCA) to map vectors in different languages into the same latent space. Laroche [35] presents a good summary for the projection-based approaches. A typical example of projection-based approach is shown in Figure 3.3.

Vulić et al. [61] pioneered a new approach to bilingual dictionary extraction. The main idea is: firstly, map words in different languages into the same semantic space using multilingual topic models; and then, several statistical measures, such as Kullback-Leibler divergence, are used to compute similarity between words in cross-languages; finally, extract word pairs with high resulting probability. This method is totally unsupervised learning style and do not require any seed dictionary.

Our approach is motivated by [61]. However, we exploit the topic model in a very different way (explained in Section 3.3.3). They do not use word alignments like we do, and as a result their approach requires training topic models with a large number of topics, which may limit the scalability of the approach. Further, we explore extensions of multilingual topic models in more than two languages.

Recently, there has been much interest in multilingual topic models (MLTM) [29, 42, 47, 11]. 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.

3.2.3 Parallel Corpora based Approach

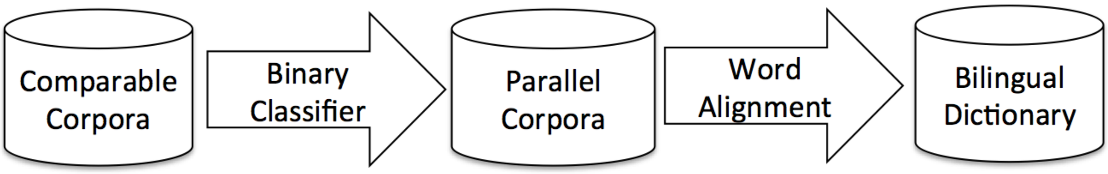


Figure 3.4: The framework of parallel corpora based approach.

There are a few researches, which extracted bilingual dictionaries from comparable corpora by 1) converting comparable into parallel corpora and then 2) constructing bilingual dictionaries via word alignment models [44, 55], as shown in Figure 3.4. Methods differ in how the parallel corpora were constructed. Munteanu and Marcu [44] employed bilingual suffix trees to build parallel corpora, which worked well on language pairs having similar word order, e.g., English-French, however, it can not be extended language pairs having different language order, e.g, English-Japanese. On other hands, Smith, Quirk and Toutanova [55] constructed parallel corpora relying on a large amount of manually designed features, such as seed dictionaries features and orthographic features, which are difficult to extend to other languages pairs.

Our multilingual topic model framework differs in that we construct topic-aligned corpora, in which each topic-aligned sentence is semantic related, rather than a fragment of parallel sentence. Furthermore, our framework do not rely on any language-specific features or seed dictionaries.

3.3 Proposed Framework for Bilingual Dictionary Extraction

The general idea of our proposed framework is sketched in Figure 3.1: 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 comparable corpora and parallel corpora. We describe how this is done in Section 3.3.2 and show how existing approaches are subsumed in our general framework in Section 3.3.3.

3.3.1 Multilingual Topic Model

Any multilingual topic model may be used with our framework. We use the one by Mimno et al. [42], which extends the monolingual Latent Dirichlet Allocation model [9]. 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 topic. 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 3 and corresponding graphical representation is shown in Fig 3.5.

In this paper, we develop a collapsed Gibbs sampling [25, 42, 47], a type of MCMC, to estimate the posterior given a tuple of documents. Concretely, given a tuple of

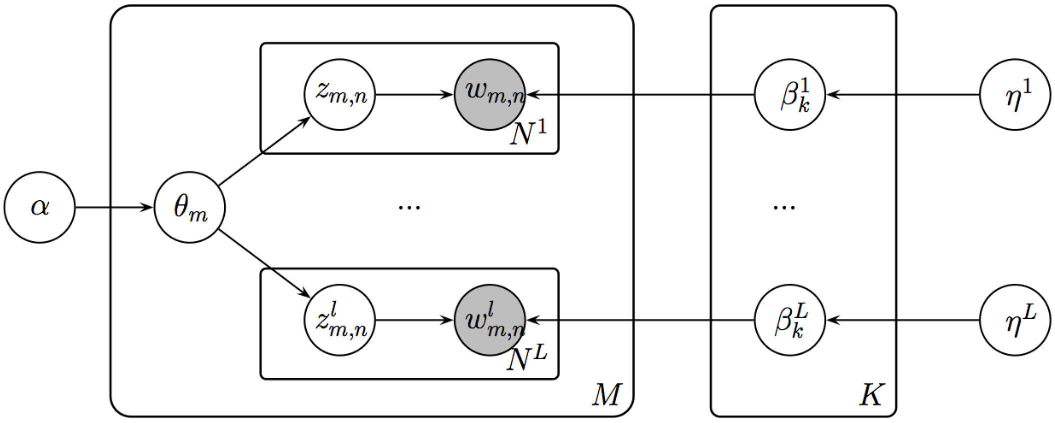


Figure 3.5: Graphical representation of multilingual topic model.

documents m, the possibility of the topic k of the i word in the language l yields:

$$p(z_i^l = k|\vec{w^l}, \vec{z^l}_{\neg i}, \beta^1, ..., \beta^L, \alpha) \propto \frac{n_{l,k,\neg i}^{(v)} + \eta^l}{\sum_{v'=1}^{V^l} n_{l,k,\neg i}^{(v')} + \eta^l \cdot V^l} \cdot \left(\sum_{l'=1}^{L} (n_{l',m}^{(k)})_{\neg l,i} + \alpha\right) \tag{3.1}$$

Here, the document in language l denotes $\vec{w^l} = \{w_i^l = v, w_{\neg i}^l\}$ with the corresponding topic states $\vec{z^l} = \{z_i^l = k, \vec{z^l}_{\neg i}\}$; the counts $n_{l,k,\neg i}^{(v)}$ indicate that the token i is excluded from the corresponding document l in the tuple; the counts $(n_{l',m}^{(k)})_{\neg l,i}$ denote that the token i is excluded from the corresponding topic k when $l = l'$ is held in the tuple; V^l denotes vocabulary in language l.

Finally, we compute the multinomial parameter sets of Θ and B:

$$\beta_{k,v}^l = \frac{n_{l,k,}^{(v)} + \eta^l}{\sum_{v'=1}^{V^l} n_{l,k}^{(v')} + \eta^l \cdot V^l} \tag{3.2}$$

$$\theta_{m,k} = \frac{\sum_{l'=1}^{L} n_{l',m}^{(k)} + \alpha}{\sum_{k'} \sum_{l'=1}^{L} y_m^{l'} \cdot n_{l',m}^{(k')}} \tag{3.3}$$

Direchlet hyperparameters α and η can be optimized by a simple and stable fixed-point iteration for a maximum likelihood estimator as [43].

Algorithm 3: Generative story for the multilingual topic model [42]. 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.

```
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
```

3.3.2 Topic-Aligned Corpora

Suppose the original comparable corpus has D document pairs $[d_i^e, d_i^f]_{i=1,...,D}$. We run a multilingual topic model with K topics, where K is user-defined (Section 3.3.1). The topic-aligned corpora is defined hierarchically as a *set of sets*: On the first level, we have a set of K topics, $\{t_1, \dots, t_k, \dots, t_K\}$. On the second level, for each topic t_k, we have a set of D “word collections“ $\{C_{k,1}, \dots, C_{k,i}, \dots, C_{k,D}\}$. 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 3.6:

1). Train a multilingual topic model (Section 3.3.1).

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.

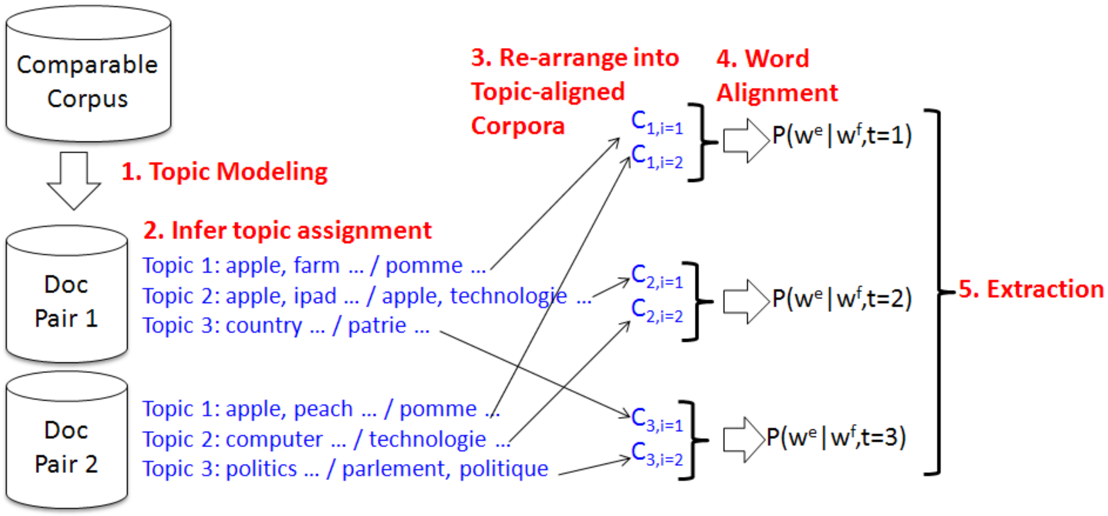

Comparable Corpus
1. Topic Modeling
Doc Pair 1
2. Infer topic assignment
Topic 1: apple, farm ... / pomme ...
Topic 2: apple, ipad ... / apple, technologie ...
Topic 3: country ... / patrie ...
Doc Pair 2
Topic 1: apple, peach ... / pomme ...
Topic 2: computer ... / technologie ...
Topic 3: politics ... / parlement, politique
3. Re-arrange into Topic-aligned Corpora
C1,i=1
C1,i=2
C2,i=1
C2,i=2
C3,i=1
C3,i=2
4. Word Alignment
P(we|wf,t=1)
P(we|wf,t=2)
P(we|wf,t=3)
5. Extraction

Figure 3.6: Construction of topic-aligned corpora.

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 [12, 48] 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 in 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:

$$p(w^e|w^f) = \sum_k p(w^e|w^f,t_k)p(t_k|w^f) \tag{3.4}$$

The first term is the topic-dependent bilingual lexicon from Step 4; the second term is estimated as follows using topic model parameters:

$$p(t_k|w^f) = \frac{p(t_k,w^f)}{\sum_k p(t_k,w^f)} \propto p(w^f|t_k)p(t_k) \tag{3.5}$$

If we assume that $p(t_k)$ is the uniform distribution over topics, $p(t_k|w^f)$ is defined as:

$$p(t_k|w^f) \propto p(w^f|t_k) \tag{3.6}$$

Here, $p(w^f|t_k)$ is the topic posterior from the topic model in Step 1.

In practice, we compute the probabilities of Equation 3.4 in both directions: $p(w^e|w^f)$ as in Eq. 3.4 and $p(w^f|w^e) = \sum_k p(w^f|w^e,t_k)p(t_k|w^e)$. Subsequently, several options are conceivable for extracting bilingual lexicon: Option (a) is to set a threshold δ and extract all pairs $(\tilde{e},\tilde{f})$ with $p(w^f = \tilde{f}|w^e = \tilde{e}) + p(w^e = \tilde{e}|w^f = \tilde{f}) > \delta$. Option (b) is to set thresholds δ_1 and δ_2, and extract lexicons based on the following bidirectional constraint that a pair $(\tilde{e},\tilde{f})$ is extracted only if:

$$\begin{aligned} p(w^e = \tilde{e}|w^f = \tilde{f}) &> \delta_1 \\ p(w^f = \tilde{f}|w^e = \tilde{e}) &> \delta_2 \end{aligned} \tag{3.7}$$

We show results from both options in our experiments. Option (a) is useful for generating a ranked list and computing precision-recall curves since δ can be adjusted to allow for different number of extracted pairs. Option (b) gives very high precision extractions since it takes the intersection from both $p(w^f|w^e)$ and $p(w^e|w^f)$; however, it is not easy to tune δ_1 and δ_2 to extract a given number of pairs, since the intersection is not known beforehand. In our experiments, we "set" δ_1 and δ_2 to retrieve only one candidate translation per model, extracting a pair $(\tilde{e},\tilde{f})$ if the following holds:

$$\begin{aligned} \tilde{e} &= \arg\max_{w^e} p(w^e|w^f = \tilde{f}) \\ \tilde{f} &= \arg\max_{w^f} p(w^f|w^e = \tilde{e}). \end{aligned} \tag{3.8}$$

3.3.3 Alternative Approaches

To the best of our knowledge, [61] is the only work that 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:

$$p(w^e|w^f) = \sum_k p(w^e|t_k)p(t_k|w^f) \tag{3.9}$$

This can be seen as a simplification of our Eq. 3.4, where Eq. 3.9 replaces $p(w^e|w^f,t_k)$ with the simpler $p(w^e|t_k)$. This is a strong assumption which essentially claims that the topic distribution t_k summarizes all information about w^f for predicting w^e. Our formulation can be considered more realistic because we do not have the assumption that w^e is independent of w^f given t_k; we model $p(w^e|w^f,t_k)$ directly and estimate its parameters with word alignment methods.

Another variant proposed by [61] is the so-called Kullback-Leibler (KL) method. It scores translation pairs by:

$$\begin{aligned} KL(w^e,w^f) &= Divergence(p(t_k|w^e)||p(t_k|w^f)) \\ &= -\sum_k p(t_k|w^e)\log p(t_k|w^e)/p(t_k|w^f) \end{aligned} \tag{3.10}$$

The information content is the same as the Cue Method (Eq. 3.9); it is simply a different scoring equation. In our experiment, we find that a symmetric version of KL, known as Jensen Shannon Divergence, gave better results:

$$JS(w^e,w^f) = \frac{1}{2}KL(w^e,w^{ef}) + \frac{1}{2}KL(w^f,w^{ef}) \tag{3.11}$$

where w^{ef} denotes the average of the *word-topic* distributions of both e and f, i.e. $w^{ef} = [p(t|w^e)+p(t|w^f)]/2$.[1]

3.4 Experimental Setup

3.4.1 Data Set

We perform experiments on the KyotoWiki 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. 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. 3.8. Due to the large data size and the strict bidirectional requirement imposed by Eq. 3.8, 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.

[1] The third and final variant by [61], TF-ITF, performs poorly and is not reported.

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

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 3.1: 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 [45] 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.

From this parallel data, we prepared several datasets at successively lower levels of comparability. As shown in Table 3.1, **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**)[3], 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.

3.4.2 Experimental Results

1. What is the best topic number *K* and what multilingual topic model can learn?

Due to the manually setting the topic number K requirement, it is important question to select the best K for the experiments. Our strategy is using the per-word log-likelihood on the hold-out data (dev-data) set for model selection. Per-word log-likelihood, which is widely used in the machine learning and statistics community, is defined as the geometric mean of the inverse marginal probability of each word in the held-out (dev) set of documents D_{dev}:

[3] 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

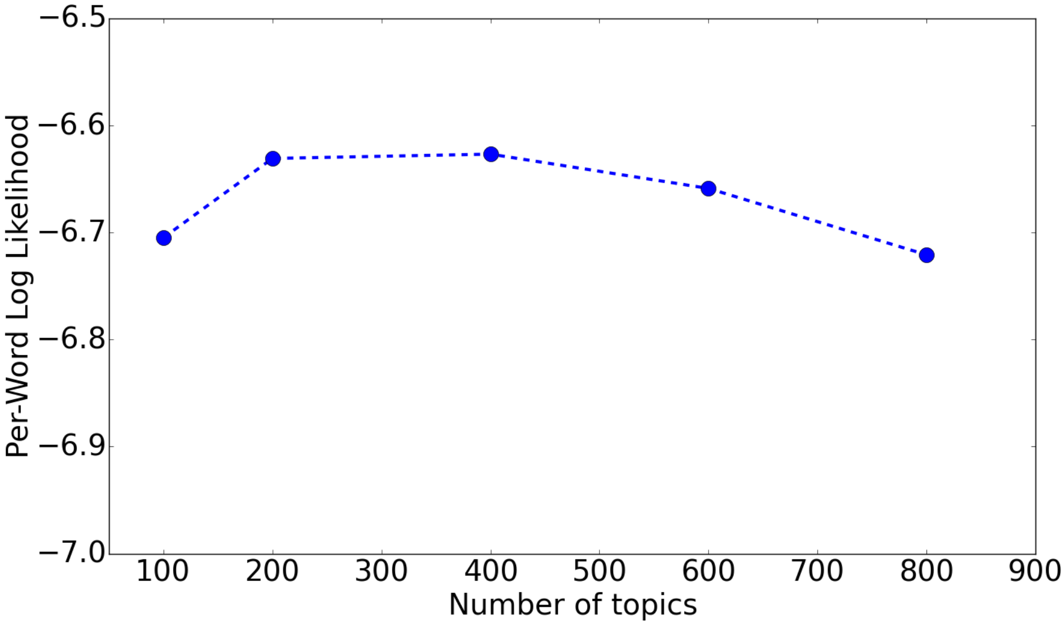


Figure 3.7: Per-word Likelihood by number of topics. Note that it is a reasonable unsupervised metric for model selection and a higher per-word log-likelihood score indicates better performance.

$$likelihood_{pw} = \frac{\sum_{t \in D_{dev}} \log p(t|D_{train})}{\sum_{t \in D_{dev}} n_t} \tag{3.12}$$

Here, n_t denotes the number of words for the tth tuple of documents in dev corpus. Following [58], we estimate $p(t|D_{train}) = \prod_l p(w_t^l|D_{train}) = \prod_l \prod_{w \in w^l} \sum_k \theta_{t,k} \beta_{k,w}^l$. The hidden variables $\theta_{t,k}$ and $\beta_{k,w}^l$ can be computed as Eq 4.2 and Eq 4.3. A higher per-word log-likelihood score indicates better performance [9, 26, 58].

We print the top 20 words of four randomly selected topics to validate the multilingual topic model. Table 3.2 and Table 3.3 summarize the those examples, i.e., topic 5 is more related the religion which associate words (i.e., English words: *shinto, god, religion* and Japanese words: 神道, 信仰, 祀る, 宗教) and topic 18 is more related the food which associate words (i.e., English words: *dish, curri, rice, meat* and Japanese words: 料理, 食べる, 食, 肉). All the words in the same topic are coherent to each other in both English and Japanese, which shows the power of multilingual topic model.

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

We focus on comparing with previous topic-modeling approaches to bilingual lexicon

Topic 5				Topic 10			
Japanese		English		Japanese		English	
word	probability	word	probability	word	probability	word	probability
神道	0.04891	shinto	0.03313	絵	0.04055	print	0.04464
信仰	0.03777	god	0.03025	浮世絵	0.03094	geisha	0.03712
祀る	0.02023	spirit	0.02920	画	0.01983	edo	0.02472
宗教	0.01404	kami	0.02367	着色	0.01542	artist	0.01922
魂	0.01098	altar	0.02279	北斎	0.01514	ukiyo	0.01887
巫女	0.01006	ritual	0.02134	所蔵	0.01500	popular	0.01302
霊	0.00963	religion	0.01982	美人	0.01444	seri	0.01144
祭り	0.00813	worship	0.01516	歌川	0.01411	woodblock	0.01127
祭祀	0.00753	sacr	0.01492	芸妓	0.00924	district	0.01005
神事	0.00748	shrine	0.01396	絹本	0.00910	hokusai	0.00978
民俗	0.00721	believ	0.01252	美術館	0.00886	hiroshig	0.00961
神楽	0.00721	ceremoni	0.01003	斎	0.00750	women	0.00865
塚	0.00683	ancient	0.00995	舞妓	0.00741	entertain	0.00865
民間	0.00640	offer	0.00995	天保	0.00708	publish	0.00839
習合	0.00619	practic	0.00963	枚	0.00699	utagawa	0.00795
祭る	0.00560	perform	0.00891	国芳	0.00661	produc	0.00777
自然	0.00543	anim	0.00883	太夫	0.00652	art	0.00743
儀礼	0.00538	deiti	0.00883	広重	0.00642	maiko	0.00716
祭	0.00533	divin	0.00875	遊女	0.00600	genr	0.00690
依る	0.00527	word	0.00818	英	0.00595	gion	0.00681

Table 3.2: Examples of topic-word distribution part 1. Note that we only give top 20 words for each topic pairs (Japanese and English). Here topic number K is set to 400.

extraction, namely [61]. Note that topic model hyperparameters for **Proposed**, **Cue**, and **JS** are $\alpha = 50/K$ and $\beta = 0.01$ following [61]. The methods are:

- **Proposed**: The proposed method which exploits a combination of topic modeling and word alignment to incorporate topic-dependent translation probabilities (Eq. 3.4).

- **Cue**: From [61], i.e. Eq. 3.9.

Topic 18				Topic 22			
Japanese		English		Japanese		English	
word	probability	word	probability	word	probability	word	probability
料理	0.07008	dish	0.03497	建築	0.04109	build	0.04369
食べる	0.02059	curri	0.02410	建物	0.01949	palac	0.03868
食	0.01696	meat	0.02226	屋根	0.01549	roof	0.02638
肉	0.01524	rice	0.02180	柱	0.01335	style	0.02285
店	0.01481	sauc	0.02048	造	0.01289	built	0.02226
調理	0.01441	cuisin	0.01956	瓦	0.01103	tile	0.01578
丼	0.01183	cook	0.01864	様式	0.00899	architectur	0.01572
カレー	0.01154	food	0.01660	壁	0.00868	construct	0.01454
卵	0.00903	fri	0.01594	階	0.00857	hall	0.01319
焼く	0.00899	veget	0.01383	建てる	0.00846	structur	0.01219
焼き	0.00767	ingredi	0.01337	設ける	0.00818	gate	0.01189
鍋	0.00764	fish	0.01093	棟	0.00815	wall	0.01148
野菜	0.00760	restaur	0.00975	メートル	0.00769	hous	0.01131
具	0.00744	chicken	0.00935	発掘	0.00730	locat	0.01125
ソース	0.00678	egg	0.00915	東西	0.00695	resid	0.01048
魚	0.00668	tempura	0.00876	施設	0.00678	design	0.01030
揚げる	0.00661	popular	0.00836	調査	0.00650	heian	0.01030
食材	0.00661	prepar	0.00836	構造	0.00601	floor	0.01019
天ぷら	0.00658	usual	0.00823	敷地	0.00597	room	0.01019
カツ	0.00635	soup	0.00817	内部	0.00583	north	0.00889

Table 3.3: Examples of topic-word distribution part 2. Note that we only give top 20 words for each topic pairs (Japanese and English). Here topic number K is set to 400.

- **JS**: From [61]. Symmetrizing KL by Jensen-Shannon (JS) divergence improves results, so we report this variant.

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 3.8 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 pos-

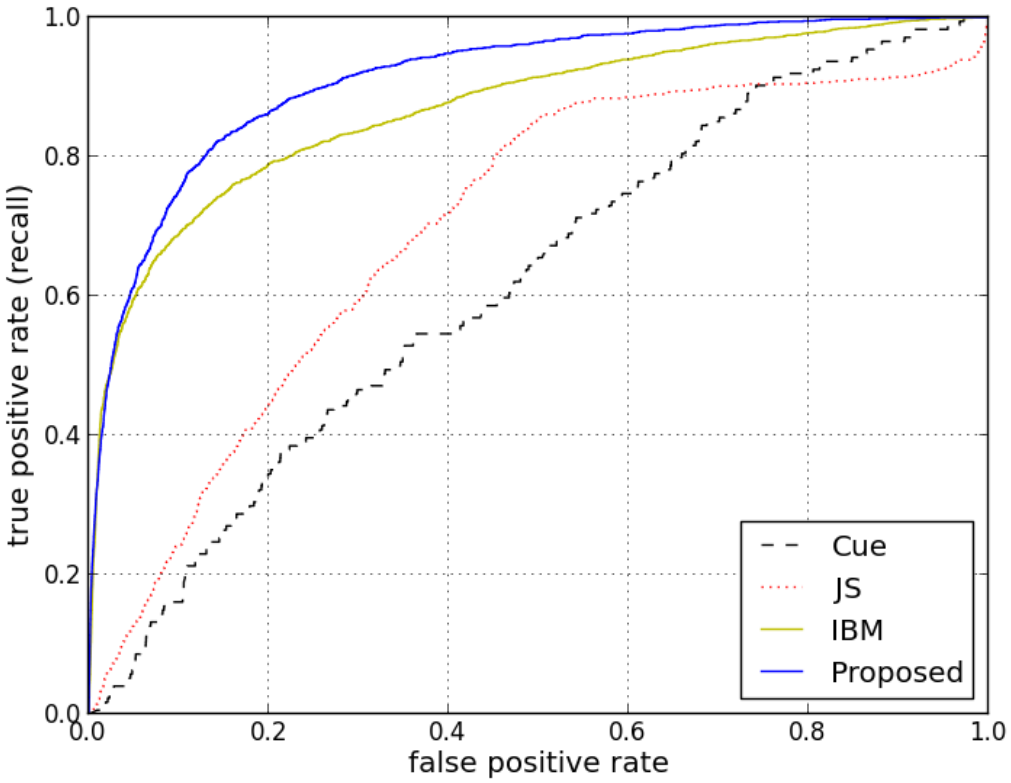


Figure 3.8: ROC curve on the **Wiki** dataset. Curves on upper-left is better. **Cue**, **JS**, **Proposed** all use K=400 topics. Note that **Proposed** is best.

itive 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 [14], which are 0.96, 0.90, 0.85 and 0.71, for **Proposed**, **IBM-1**, **Cue**, and **JS**, respectively.[4]

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 bidirec-

[4] 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 3.4.

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 3.4: 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.

tional heuristic of Eq. 3.7 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 3.4.

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 [61] 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. The experiments in [61] has vocabulary size of 10k, compared to 150k in our experiments. We also tried large $K \geq 1000$, however **Cue** still obtains a bad result (10.4% with $K = 2000$). Furthermore, we also use per-word log-likelihood, a typical metric for model selection in machine learning community, as shown in Figure 3.7. This figure shows the best topic number is $K = 400$. In this regard, the **Proposed** method is much more *scalable* and *reasonable*, achieving good results with low K, satisfying one of original desiderata. We have a hypothesis as to why **Cue** and **JS** depend on large K. Eq. 3.4 is a valid expression for $p(w^e|w^f)$ that makes little assumptions. We can view Eq. 3.9 as simplifying the first term of Eq. 3.4 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.

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.

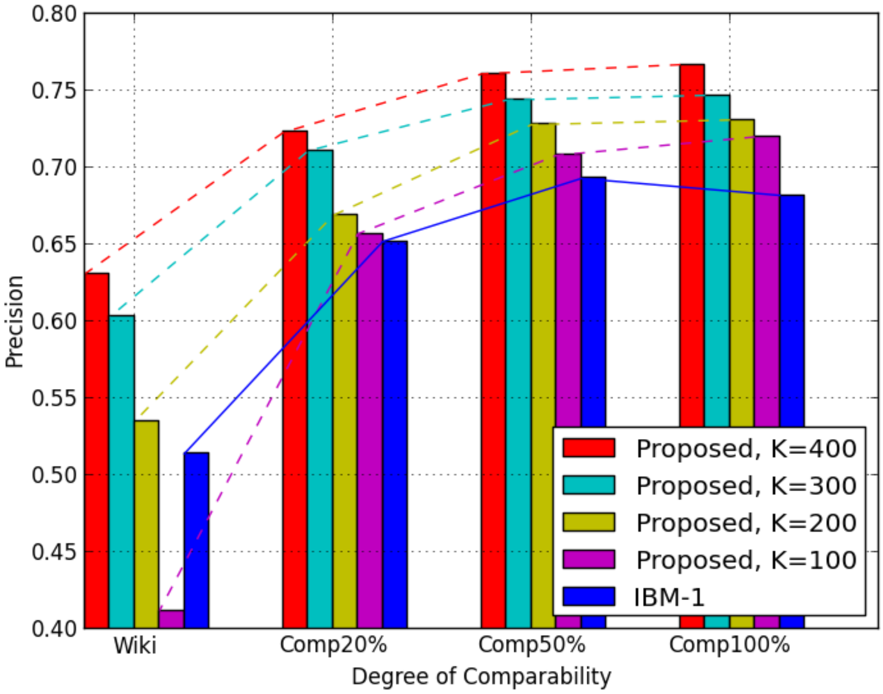


Figure 3.9: Robustness of method under different data conditions.

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

We next examined how our methods perform under different data conditions. Figure

3.9 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 topic model of [42] 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 [23] or other variants in future work.

4. 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. 3.10 shows the number of word types that have x number of topics. This power-law is expected since we are modeling all words. 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 [42]. 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**.

Next we compute the statistics after constructing the topic-aligned corpora (Step 3 of Fig. 3.6). 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 1, that means the partition into topic-aligned corpora effectively separates the skewed word-topic distribution of Fig 3.10. 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.

5. 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 in-

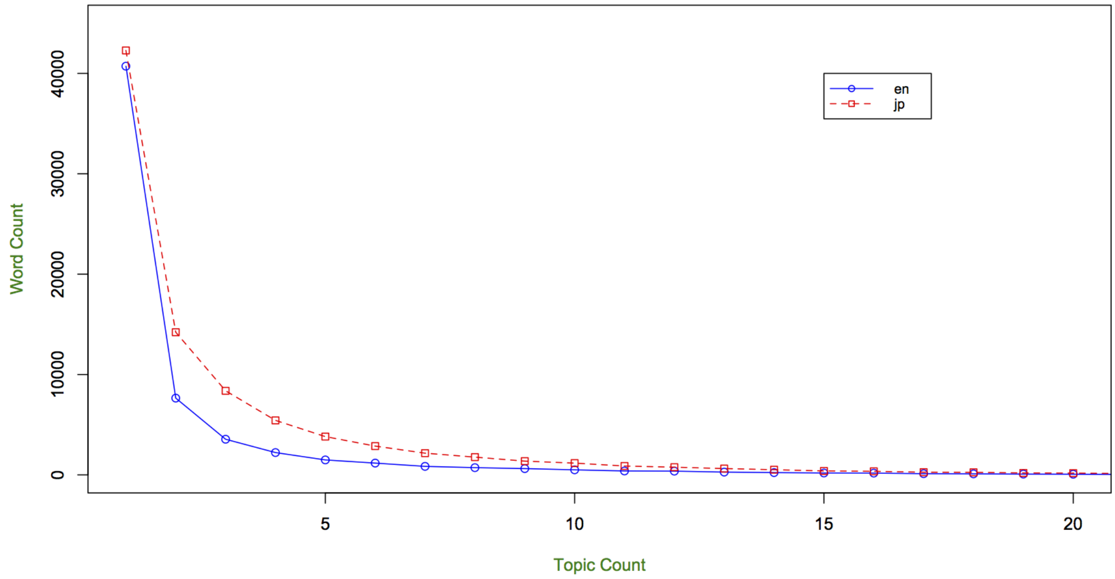


Figure 3.10: Power-law distribution of number of word types with X number of topics.

correctly 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 3.5 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 [6].

Since word alignment errors were frequent, we conducted an additional experiment to compare several popular word alignment methods in statistical machine translation as follows:

1. **Giza-vb:** a modification of IBM Models training using Variational Bayes EM learning [53]

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

Table 3.5: Counts of various error types.

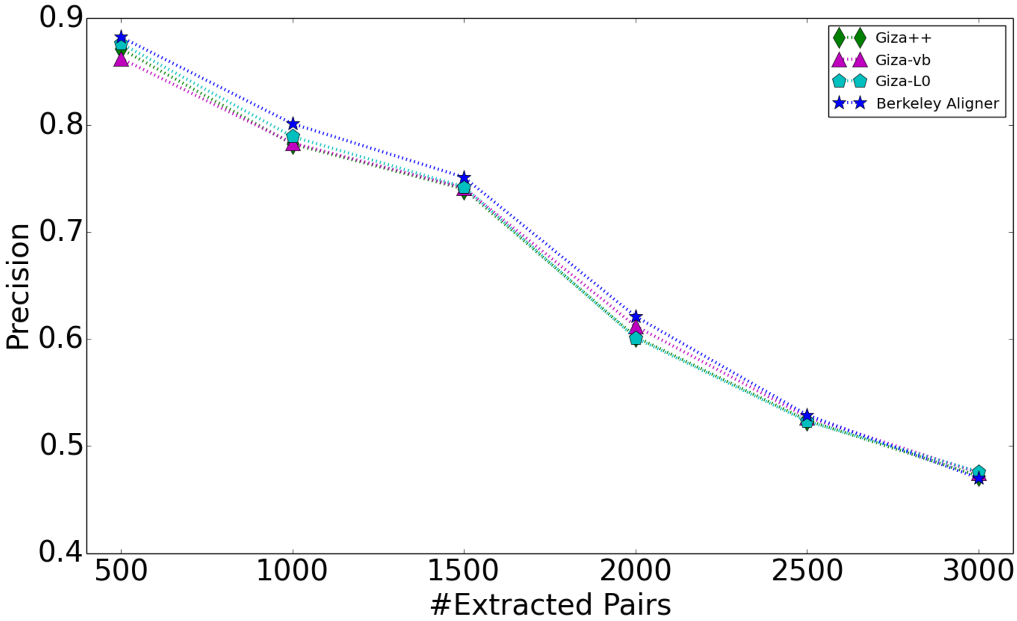


Figure 3.11: Comparison of different word alignment tools: Precision-vs-#Extracted pairs curve. The Dirichlet parameter α is set to 0.01 in Giza-vb; parameters of Giva-L0 are set to default and bi-direction joint IBM1 setting is used in Berkeley Aligner.

2. **Giza-L0:** a modification of IBM Models to generate sparser alignments using approximate L0-norm optimization [59]

3. **Berkeley Aligner:** a symmetrical aligner, which enforces agreement in bi-direction word alignment [37]

Figure 3.11 shows the lexicon extraction results using different alignment tools. While the differences are in general not very large, we observe that the Berkeley Aligner appears slightly better than all other IBM Model variants, implying that bi-directional constraints may be helpful in this kind of topic-aligned data.

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 3.6: 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.

6. What is the computation cost?

K	topic	giza	Eq.3.4	Eq.3.9	Prp	Cue
100	180	3	20	1440	203	1620
200	300	3	33	2310	336	2610
400	780	5	42	3320	827	4100

Table 3.7: 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.3.4 and for **Cue** is topic+Eq.3.9.

Timing results on a 2.4GHz Opteron CPU for various steps of **Proposed** and **Cue** are shown in Table 3.7. 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.3.9, especially when both $p(w^e|t_k)$ and $p(t_k|w^f)$ have high cardinality. In comparison, calculating Eq.3.4 is fast since $p(w^e|w^f,t_k)$ is in practice quite sparse.

3.5 Summary

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 [65], 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. Furthermore, we showed that our topic-dependent translation models can capture some of the polysemy phenomenon important in dictionary construction.

Chapter 4

Enhancing Bilingual Dictionary Extraction via Multilingual Comparable Corpora

In this chapter, we use multilingual comparable corpora (more then two languages) to enhance the accuracy of bilingual dictionary extraction. Although multilingual dictionaries can be built by using the previous proposed framework in Chapter 3, we focus on how additional languages help to improve the accuracy of bilingual (English-Japanese) lexicon extraction. This chapter is organized as follows: In Section 4.1, we introduce the background of bilingual lexicon extraction from multilingual comparable corpora. Then related works are reviewed in Section 4.2. In Section 4.3, we propose a multilingual topic model to encode more information. Next, we conduct a serial of experiments to evaluate our proposed systems in Section 4.4. At last, we summarize this chapter.

4.1 Introduction

Due to the scarcity of parallel corpora for certain language-pairs or domains of interest (e.g., medical and microblog), extracting bilingual dictionaries from comparable corpora, introduced in Chapter 3, arouses many attention in recent years [24, 61, 39]. Such kind of researches are divided into two broad categories: context vector approaches and projection-based methods. One of the biggest problems for those approaches is that they all need either big seed dictionaries or *fully-connected* comparable corpora, which are refered to a collection of tuples of documents containing all the different languages, as shown in Figure 4.2.

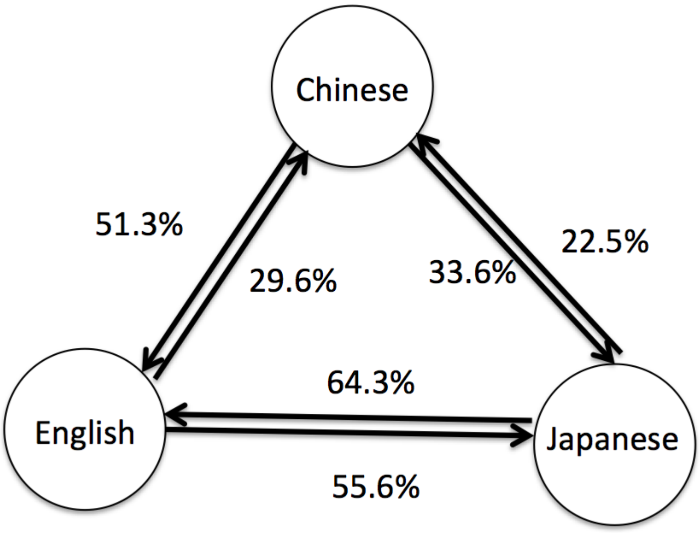


Figure 4.1: The rates of inter-linked different languages (Chinese, English and Japanese) in Wikipedia [5]. Note that more than half of documents do not have interlink to other languages on average.

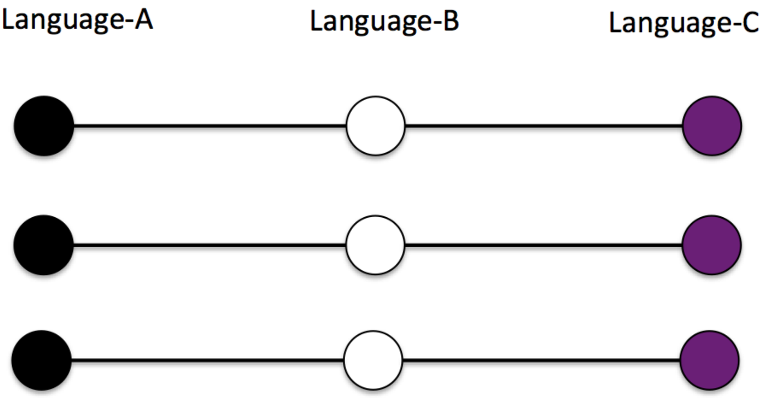


Figure 4.2: The *full-connected* comparable corpora in three different languages, e.g., Chinese, English and Japanese.

In fact, the *fully-connected* comparable corpora are still difficult to obtain, e.g., the rate of interlinked documents between different languages in Wikipedia is very low, as shown in Figure 4.1 [5]. It implies that most of information, which is provided by monolingual document or partially connected document pairs, is wasted by *standard* multilingual topic models [61, 39]. Furthermore, it is believed that a small proportion of mono-language documents may help estimate true word topic distributions and reveal the relationship between topics across languages. Thus, we propose a new multilingual topic model for *partially-connected* comparable corpus that maximizes the information usage of the data. Here, *partially-connected* comparable corpora are referred a collation of tuples of documents, where a tuple does not necessarily contains documents in all languages, as shown in Figure 4.3.

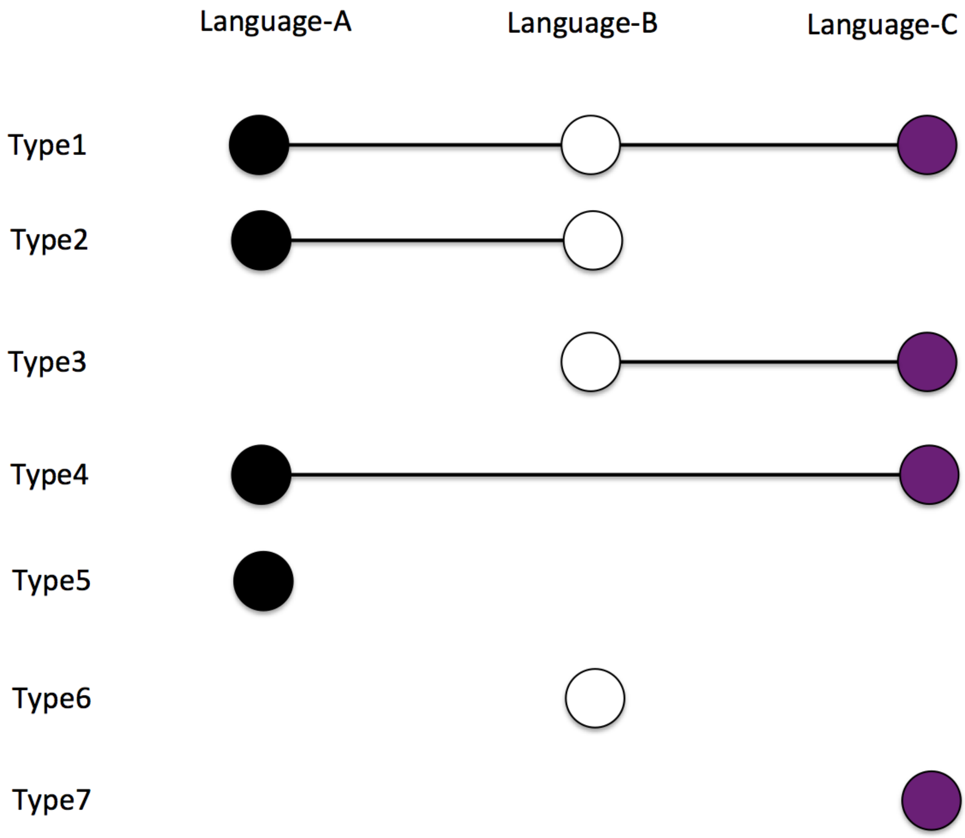


Figure 4.3: The *partially-connected* comparable corpora in three different languages, e.g., Chinese, English and Japanese.

In this section, we extend the standard multilingual topic model, which is the key component of the proposed framework, to exploit *partially-connected* comparable corpora in more than two languages which are *partially-connected*. Such kind of corpus is easier to obtain in the real world, such as Wikipedia corpora. We show how we can improve the extraction of Japanese-English dictionaries using comparable data not only from Japanese and English, but also from other languages such as Chinese and French, in our previous framework. Furthermore, it shows the flexibility of our proposed framework, which a novel combination of topic models and word alignment models. Note that due to using the multilingual comparable corpora, we can build multilingual dictionaries by using the proposed framework. However, we only focus on how addition languages help to improve the accuracy of finding English and Japanese translation pairs, as shown in Figure 4.4.

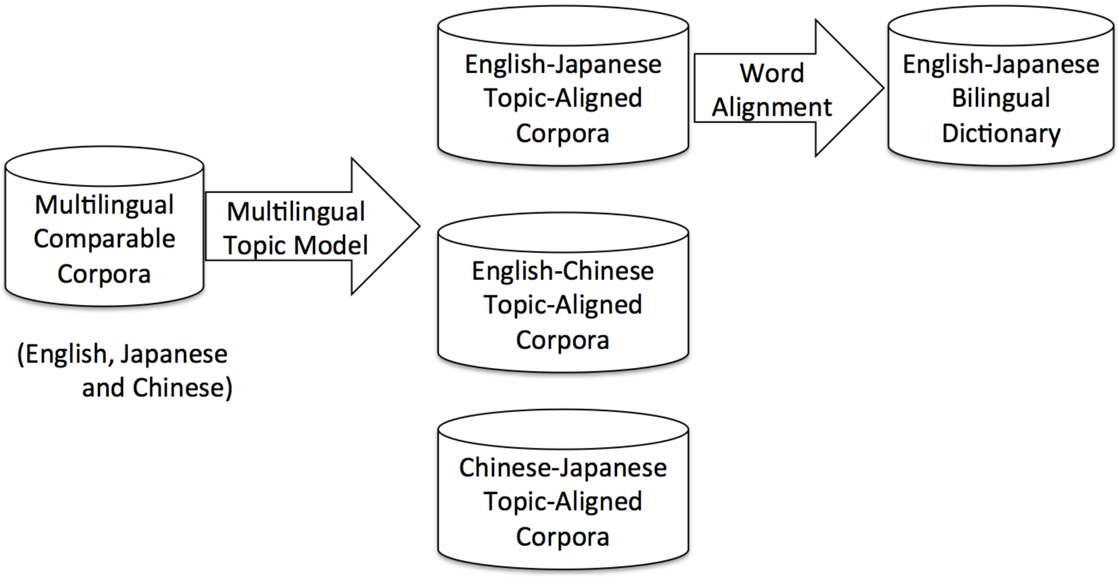

Multilingual Comparable Corpora
(English, Japanese and Chinese)
Multilingual Topic Model
English-Japanese Topic-Aligned Corpora
English-Chinese Topic-Aligned Corpora
Chinese-Japanese Topic-Aligned Corpora
Word Alignment
English-Japanese Bilingual Dictionary

Figure 4.4: The proposed framework for English-Japanese dictionaries extraction from multilingual comparable corpora (English, Japanese and Chinese).

4.2 Related Work

The numerous works on bilingual lexicon from comparable corpora can be divided into two broad categories: context vector approaches and projection-based approaches, which were introduced in Section 3.2. We also briefly touch upon research on pivot languages in (Section 4.2.1) and multilingual word representation learning (Section

4.2.2); to the best of our knowledge, these are promising approaches but have not yet been employed for bilingual lexicon extraction.

4.2.1 Pivot-Language in Lexicon Extraction and Machine Translation

Multilingual corpora in more than two languages have been exploited to various degrees. This is often done using a pivot language: for example, given a Japanese-English dictionary and an English-Chinese dictionary, one can exploit transitive properties with English serving as the pivot to find Japanese-Chinese translations. When available, such multilingual information has been shown to improve the quality of bilingual lexicons [34, 54, 1].

This pivot language idea has proven beneficial in applications such as cross-lingual information retrieval [22], and machine translation [50, 62]. It is also used for bootstrapping the construction of WordNet for low resource languages [10], and for directly creating multilingual lexical resources [56, 15, 41].

Our multilingual topic model approach differs in that there is no concept of pivot: data in all languages are treated equally. In this respect, extension to many languages is straightforward, as long as computation efficiency issues can be solved.

4.2.2 Multilingual Word Representation Learning

Multilingual word representation learning, which is an extension of monolingual word representation learning, is a set of deep learning algorithms that enables new ways to do cross-lingual processing. It works by mapping words in different languages into the same low-dimensional space in order to capture syntactic and semantic similarities across languages. For instance, Klementiev et al. [30] proposed training bilingual word representations by jointly training monolingual neural language models together with a regularizer that enforces seed translations to have similar representations. Chandar et al. [49] proposed a novel autoencoder algorithm for learning bilingual word representations; importantly, their algorithm only depends on bag-of-words representations of aligned sentences, and does not rely on word alignments.

These works focus on cross-lingual classification tasks, but conceivably their results could be adapted to our comparable lexicon extraction task. For example, the vectors might be used as seed within context-based models like [51]. Alternatively, these vectors could be used within our framework to generate something like the topic-aligned corpora; i.e. words with similar vectors are grouped together and given to our word aligner, analogously to how we group words with the same topic together. We believe the use of vector representations is an interesting area of future work.

4.3 Multilingual Topic Model

We adopt the Multilingual Topic Model (MLTM) proposed by Ni et al. [47] and Mimno et al. [42], which extends the monolingual Latent Dirichlet Allocation model [9]. MLTM learns word-topic distributions and topic-document distributions from comparable corpora. In the original works, it is assumed that each document tuple t_m in the comparable document is *fully-connected*; for example, if we have a quad-lingual comparable corpus consisting of Chinese (c), English (e), French (f), and Japanese (j), it is assumed that all document tuples in the collection contains documents in all four languages, i.e. $t_m = [d_m^c, d_m^e, d_m^f, d_m^j]\ \forall m$, where m indexes tuples in the collection and d_m^c represents a Chinese document, d_m^e represents an English document, etc.

However, such fully-connected comparable corpora are rare in practice. Taking the entire Wikipedia as an example, Arai et al., [5] showed that among all Japanese documents, only 64.4% have links to English entries and only 22.5% have links to Chinese entries. In our own Wikipedia crawl (described in Section 4.4.1), we find that within our target set of 14k Japanese documents, the proportion of linked English, Chinese, and French documents is only around 20-30%; if we restrict to tuples that contain documents in all four languages, this number drops to 12%. In some cases, this is because the *interlanguage-link* information is missing; but in most cases, this disparity is largely the result of different human editors contributing independently in different languages [17].

We assume that a tuple does not necessarily contain documents in all languages and call such comparable corpora *partially-connected*. In the following, we extend the MLTM of [47, 42] to handle partially-connected corpora with a maximum of L languages per tuple.

4.3.1 Generative Process

The generative process of our proposed partially-connected multilingual topic model is this: First, we define our comparable corpus as a collection of M tuples in different languages, i.e. $t_m = [d_m^1, ..., d_m^l, ..., d_m^L]$ with $m \in \{1, ..., M\}$. Given a tuple of documents $t_m = [d_m^1, ..., d_m^l, ..., d_m^L]$, there is a corresponding auxiliary variable $y_m = [y_m^l, ..., y_m^l, ..., y_m^L]$, which is a L-dimensional binary vector that indicates the presence or absence of documents in the language l in tuple m. The value of l-th of vector m, $y_m^l \in \{0, 1\}$, where 1 indicates presence and 0 indicates absence. For example, a tuple of documents which may contain Chinese, English and Japanese can be represented by a 3-dimensional binary vector: $[1, 1, 1]$ denotes that it contains all of the three languages; while $[0, 1, 0]$ denotes that it only contains English document. It is not difficult to see that it is very flexible to encode the relationship of a tuple of document for the multilingual topic model.

The generative story is shown in Algorithm 4 and a graphical representation is shown in Figure 4.5.

Here, language-specific topic word distributions β^l are drawn from symmetric Dirichlet distributions with prior η^l; θ_m is the topic proportion of a tuple of documents t_m drawn from symmetric Dirichlet distribution with prior α; z^l are topic indices in language l; words w^l are drawn from language-specific distributions $p(w^l | z^l, \beta^l, y_m^l)$, where $l \in \{1, ..., L\}$.

4.3.2 Inference

The central computational problem for partial multilingual topic model is approximating the posterior given a tuple of documents. In general, it is hard to estimate the posterior of a Bayesian model using exact inference methods. Therefore, approximate inference algorithms are always selected to deal with these kind of models. One of the approximate inference methods is based on Markov Chain Monte Carlo (MCMC) [42, 7], a sampling approach. The basic idea of MCMC is that first, a Markov chain is constructed; and then its stationary distribution, which is the posterior of interest, is computed.

In this paper, we develop a collapsed Gibbs sampling [25, 42, 47], a type of MCMC, to estimate the posterior given a tuple of documents. Concretely, given a tuple of

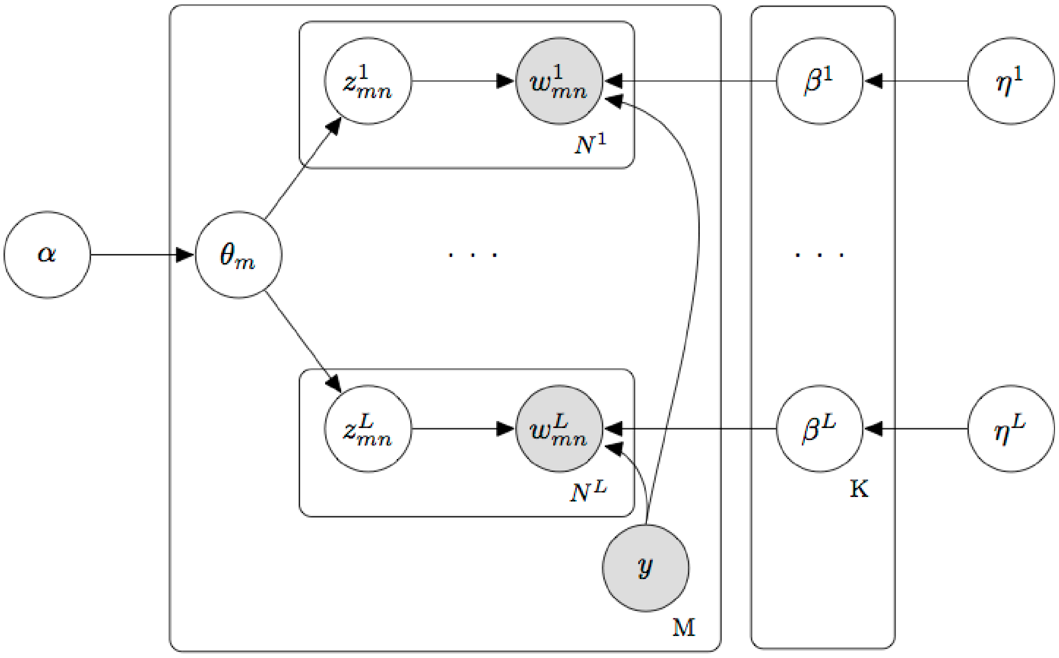


Figure 4.5: Graphical representation of partially-connected multilingual topic model.

documents m, the possibility of the topic k of the i word in the language l yields:

$$p(z_i^l = k|\vec{w^l}, \vec{z^l}_{\neg i}, \beta^1, ..., \beta^L, \alpha) \propto \frac{n_{l,k,\neg i}^{(v)} + \eta^l}{\sum_{v'=1}^{V^l} n_{l,k,\neg i}^{(v')} + \eta^l \cdot V^l} \cdot (\sum_{l'=1}^{L} y_m^{l'} \cdot (n_{l',m}^{(k)})_{\neg l,i} + \alpha) \quad (4.1)$$

Here, the document in language l denotes $\vec{w^l} = \{w_i^l = v, w_{\neg i}^l\}$ with the corresponding topic states $\vec{z^l} = \{z_i^l = k, \vec{z^l}_{\neg i}\}$; the counts $n_{l,k,\neg i}^{(v)}$ indicate that the token i is excluded from the corresponding document l in the tuple; the counts $(n_{l',m}^{(k)})_{\neg l,i}$ denote that the token i is excluded from the corresponding topic k when $l = l'$ is held in the tuple; V^l denotes vocabulary in language l.

Finally, we compute the multinomial parameter sets of Θ and B:

$$\beta_{k,v}^l = \frac{n_{l,k,}^{(v)} + \eta^l}{\sum_{v'=1}^{V^l} n_{l,k}^{(v')} + \eta^l \cdot V^l} \quad (4.2)$$

$$\theta_{m,k} = \frac{\sum_{l'=1}^{L} y_m^{l'} \cdot n_{l',m}^{(k)} + \alpha}{\sum_{k'} \sum_{l'=1}^{L} y_m^{l'} \cdot n_{l',m}^{(k')}} \quad (4.3)$$

Direchlet hyperparameters α and η can be optimized by a simple and stable fixed-point iteration for a maximum likelihood estimator as [43].

Algorithm 4: Generative story for partially-connected multilingual topic model

```
for each topic $k$ do
    for $l \in \{1, ..., L\}$ do
        sample $\varphi_k^l \sim Dirichlet(\beta^l)$
    end
end
for each tuple $t_m$ in corpus do
    sample $y_m$
    for each document pair $t_m$ do
        sample $\theta_m \sim Dirichlet(\alpha)$
        for $l \in \{1, ..., L\}$ do
            if $y_m^l == 1$ then
                sample $z^l \sim Multinomial(\theta_m)$
                for each word $w^l$ in $d_i^l$ do
                    sample $w^l \sim p(w^l | z^l, \varphi^l, y_m^l)$
                end
            end
        end
    end
end
```

4.4 Experiments

First, we describe our experiment setup in Section 4.4.1. Section 4.4.2 compares our method with previous works, and Section 4.4.3 shows how our method improves given additional languages in the comparable data. Section 4.4.4 discusses practical issues such as hyper-parameter selection and run-time, while Section 4.4.5 provides with detailed analyses of the results. Finally, Section 5 demonstrates how our approach can be used to provide high-precision dictionaries to bootstrap existing context vector methods.

Wiki	#document	#vocabulary	$\#(j \cap e)$	$\#(j \cap e \cap c)$	$\#\ (j \cap e \cap c \cap f)$
Japanese (j)	14,033	40k	-	-	-
English (e)	4,087	20k	4,087	-	-
Chinese (c)	3,494	23k	-	2,338	-
French (f)	2,871	12k	-	-	1,719

Table 4.1: Statistics of our multilingual **Wiki** crawl dataset. Here, $\#(l \cap l')$ denotes the number of "fully-connected" document tuples by intersecting languages l and l'.

4.4.1 Experiment Setting

We perform experiments based on the Kyoto Wiki Corpus[1]. We choose Kyoto Wiki Corpus because it is a *parallel* corpus, where the Japanese edition of Wikipedia is translated manually into English sentence-by-sentence (14k document pairs, 472k sentences). This enables us to use standard word alignment methods to create a "gold-standard" lexicon for large-scale automatic evaluation. First, we ran IBM Model 4 on this parallel corpus. Then we extracted 166k the $(\tilde{e}, \tilde{f})$ pairs based on the strict bidirectional requirement of Eq. 3.7, with threshod $\delta_1 = \delta_2 = 0.3$. We refer to these pairs as a "gold standard" bilingual lexicon. Due to the large data size and the strict bidirectional requirement, these "gold standard" bilingual dictionary items are of high quality (92% precision by a manual check on 500 random items). Note that sentence alignments are used only for creating this gold-standard and are not used in subsequence experiments.

To evaluate the proposed framework, we use a *real* comparable corpus crawled from Wikipedia (denoted as **Wiki**). We keep the Japanese side of the original Kyoto Wiki Corpus, but crawl the online English, Chinese and French editions by following the inter-language links from the Japanese page. The statistics of the crawl are shown in Table 4.1. Observe that this is a partially-connected comparable corpus: the number of corresponding articles in English, Chinese, and French is much smaller, consisting of only 20-30% of the original Japanese. The number of fully-connected tuples in all four languages is only 12%, as seen in the intersection $(j \cap e \cap c \cap f)$.

For pre-processing, we did word segmentation on Japanese and Chinese using Kytea [45]; Porter stemming on English and French using NLTK Version 3.0[2]. Finally, we remove the 2,000 rarest words and stop-words from each language: English, Chinese,

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

[2] www.nltk.org/api/nltk.stem.html

Language	Japanese	English	Chinese	French
Number of stop-words	44	571	125	463

Table 4.2: Statistics of stop-words in different languages.

French and Japanese as shown in Table 4.2. To facilitate future work in this area, our stop-word lists for these four languages are released at:
https://bitbucket.org/allenLao/stopwords.

4.4.2 Lexical Extraction Results: Comparison with Baselines

We begin by comparing with previous topic-modeling approaches to bilingual lexicon extraction, namely [61]. Using the automatically-created ”gold-standard” lexicon, we evaluate methods by Precision, defined as $\frac{|\{Gold(e,f)\}\bigcap\{Extracted(e,f)\}|}{\#Extracted}$.

Table 4.3 shows the precisions of our proposed method, compared with the baseline **Cue** and **JS** methods from [61]. All these methods first run our MLTM with $K = 400$ topics[3] on the partially-connected Japanese-English **Wiki** dataset, which consists of $14,033+4,087 = 18,120$ documents.

Our method extracts a total of 1,457 pairs using the bidirectional constraint in Eq. 3.8. This achieved a precision of 0.742. For comparison, we adjusted the threshold δ (Option (a) discussed in Section 3.3.2), such that the **Cue** (Eq. 3.9) and JS (Eq. 3.11) methods give roughly the same number of extracted pairs as our proposed method. The resulting precision of Cue and **JS** are very poor, at 0.073 and 0.091, respectively. Vulic [61] reports that a large number of topics is necessary for good results, so we re-ran the baselines with $K = 2,000$, the suggested value in [61]. Despite the long run-time of MLTM for large K, the precision only increased to 0.104 and 0.123 for **Cue** and **JS**, respectively.

It can be seen that our proposed method is much more effective at extracting bilingual lexicon, in particular in large-vocabulary datasets (The vocabulary size in [61] is $7k$ and $9k$ in Italian and English respectively). We have a hypothesis as to why **Cue** and **JS** depend on large K. Eq. 3.4 is a valid expression for $p(w^e|w^f)$ that makes little assumptions. We can view Eq. 3.9 as simplifying the first term of Eq. 3.4 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

[3] MLTM hyperparameters are $\alpha = 50/K$ and $\beta = 0.01$ following [61].

System	Precision	#Extracted
Proposed (K=400)	0.742	1,457
Cue (K=400)	0.073	1,400
JS (K=400)	0.091	1,400
Cue (K=2,000)	0.104	1,400
JS (K=2,000)	0.123	1,400
IBM-1	0.521	1,400

Table 4.3: Comparison with baselines using the Japanese-English part of **Wiki** dataset.

throwing out w^f, which has large cardinality, t_k needs to grow in cardinality to compensate for the loss of expressiveness.

As an additional baseline, we directly run IBM Model 1 on the fully-connected Japanese-English comparable corpora, treating each document pair as ”sentence pair”. This **IBM-1** baseline does not employ MLTM and the score of a pair (e, f) is defined as the average lexical probabilities obtained from IBM Model 1 in both directions. Interestingly, this baseline achieves a precision of 0.52, better than **Cue** and **JS**. But our proposed method still performs better, implying that the combination of existing word alignment models and MLTM attains good synergy.

4.4.3 Lexicon Extraction Results: Additional Languages

We now examine the effects of adding additional languages on Japanese-English lexicon extraction. Table 4.4 shows how precision improved as we add Chinese (3,494 comparable documents in addition to the original 18,120 Japanese-English corpora), as well as both Chinese and French (3,494+2,871=6,365 comparable documents). Since the number of extractions changes (because probability value changes affects the bidirectional constraint of Eq. 3.8), we also manually evaluated precision (ManualPrec) on a fixed random set of 100 pairs.

From Table 4.4, we see that adding Chinese documents improves the (automatic) precision from 0.742 to 0.761. Adding both Chinese and French documents further improves results, with (automatic) precision gaining 3% (0.742 $\rightarrow$ 0.774) and manual precision gaining 9% (0.62 $\rightarrow$ 0.71). We observe these improvements because adding more languages and data improves the estimation of the MLTM. Specifically, in our

System	Precision	ManualPrec	#Extracted
Full-Japanese-English	0.612	0.51	1,745
Japanese-English (= **Proposed** in Table 4.3)	0.742	0.62	1,457
Japanese-English-Chinese	0.761	0.64	1,365
Japanese-English-Chinese-French	0.774	0.71	1,372

Table 4.4: Comparison of proposed method using additional languages in the **Wiki** dataset. $K = 400$ and MLTM hyperparameters are same as described in Section 4.4.2.

bilingual extraction equation (Eq 3.4), more data can directly improve the estimation of the topic distribution $p(t_k|w^f)$; further, more data may also indirectly improve the estimation of the the topic-dependent bilingual lexicon $p(w^e|w^f, t_k)$ via better posterior inference results for input into the word alignment step. Note that the word alignment part is the same for the various systems in Table 4.4, so improvements come from better MLTM.

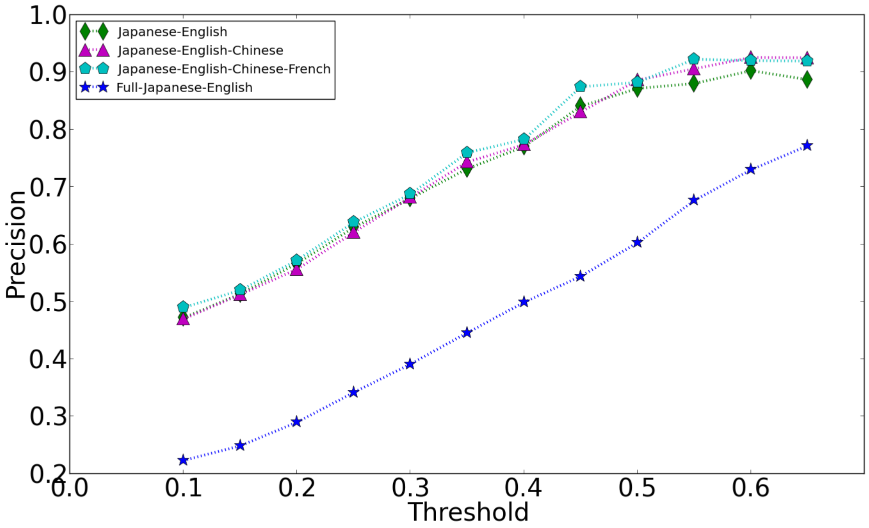


Figure 4.6: Effect of using additional languages: Precision-vs-threshold curve.

We also show results using only the fully-connected Japanese-English comparable corpus (Full-Japanese-English). This system only runs MLTM on 4,087 document pairs, and as a result the precision is lower than the partially-connected case (0.612

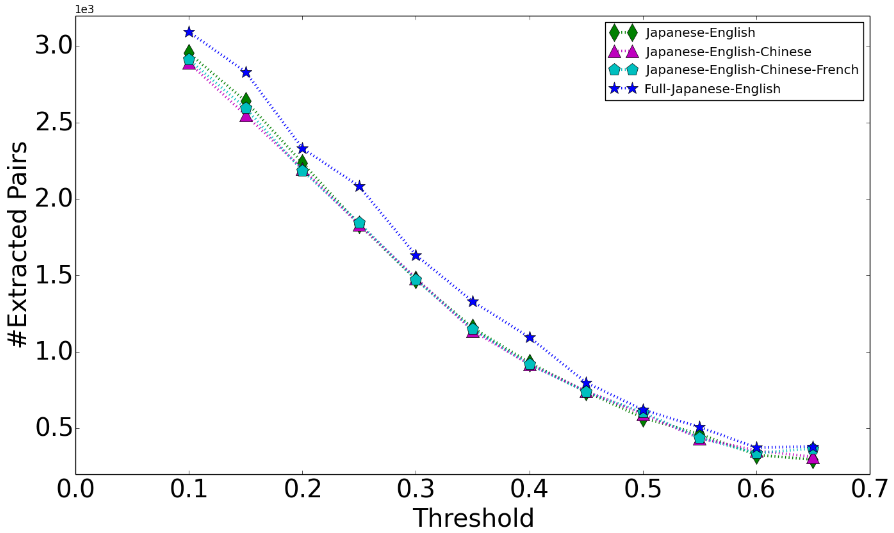


Figure 4.7: Effect of using additional languages: Number of extraction-vs-threshold curve.

vs 0.742). This demonstrates that our MLTM is effective in exploiting monolingual documents in estimating its parameters.

Finally, we also compare the systems not by using the bidirectional constraint, but by varying the threshold δ (as discussed in Option (a) at the end of Section 3.3.2. Figure 4.6 shows how precision varies as we lower the threshold. Figure 4.7 plots the number of extracted pairs vs. threshold on the same data. We observe a large overall gain in precision regardless of threshold as we move from fully-connected to partially-connected data, which corroborates with the results in Table 4.4. The number of extractions are roughly similar for the various partially-connected systems, while fully-connected has slightly larger number (but lower precision). The differences between the various systems using partially-connected corpora does not seem very large. But this is not surprising, given the large amount of monolingual Japanese documents (140,033) in our dataset compared to additional Chinese and French documents (around 3,000). Nevertheless, we do observe that the Japanese-English-Chinese-French system does indeed have the best precision curve.

4.4.4 Practical Issues: Model selection and Run-time

The most important parameter in our approach is the number of topics K in MLTM. As K goes to one, the proposed approach becomes equivalent to running word alignment directly on comparable documents, treating each document pair as a ”sentence pair.” As K increases, the topic-aligned corpora become more fine-grained and the lexicon extraction precision improves. However, if K is too large, then each word collection $C_{k,i}$ in the topic-aligned corpora becomes too small; and if the topic model incorrectly assigns translation pairs to different topics, it becomes impossible to extract it in subsequent word alignment step.

First, we show how precision varies with different values of K in Figure 4.8. We observe that for low values of K (e.g. 100, 200), the precision is relatively low around 0.4-0.6. The best precision is achieved with $K = 400$, followed closely by $K = 600$ and $K = 800$, all in the 0.7-0.8 range.

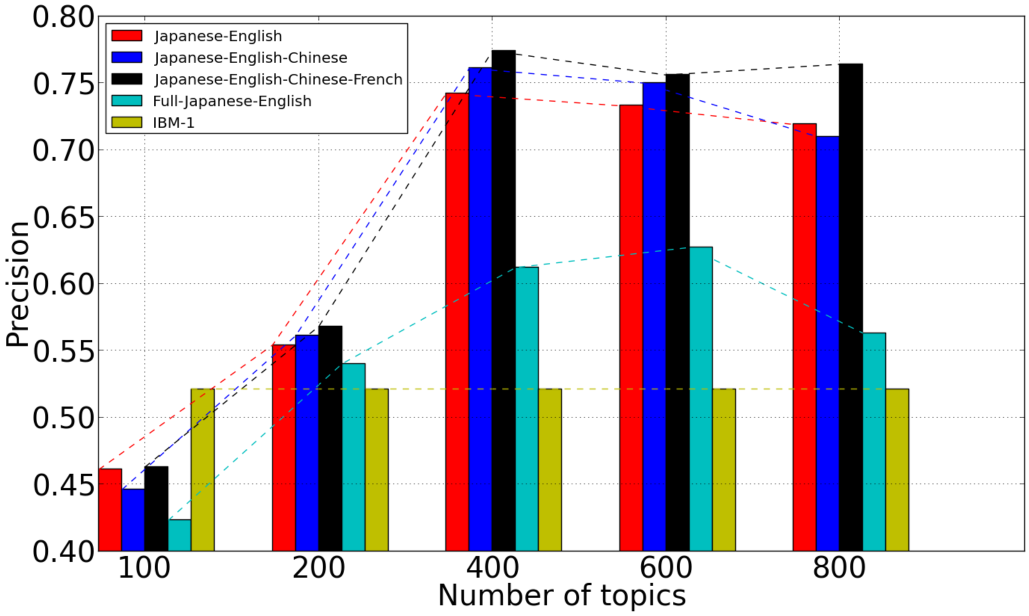


Figure 4.8: Precision by number of topics (K).

While it is expected that results vary somewhat by K, the important question is whether the best K can be selected a priori in an unsupervised manner. Now we show that the per-word log-likelihood on the held-out data set is effective for model selection. Per-word log-likelihood, which is widely used in the machine learning and statistics community, is defined as the geometric mean of the inverse marginal probability of each word in the held-out (dev) set of documents D_{dev}, as Eq 3.12. A higher per-word log-likelihood score indicates better performance [9, 26, 58].

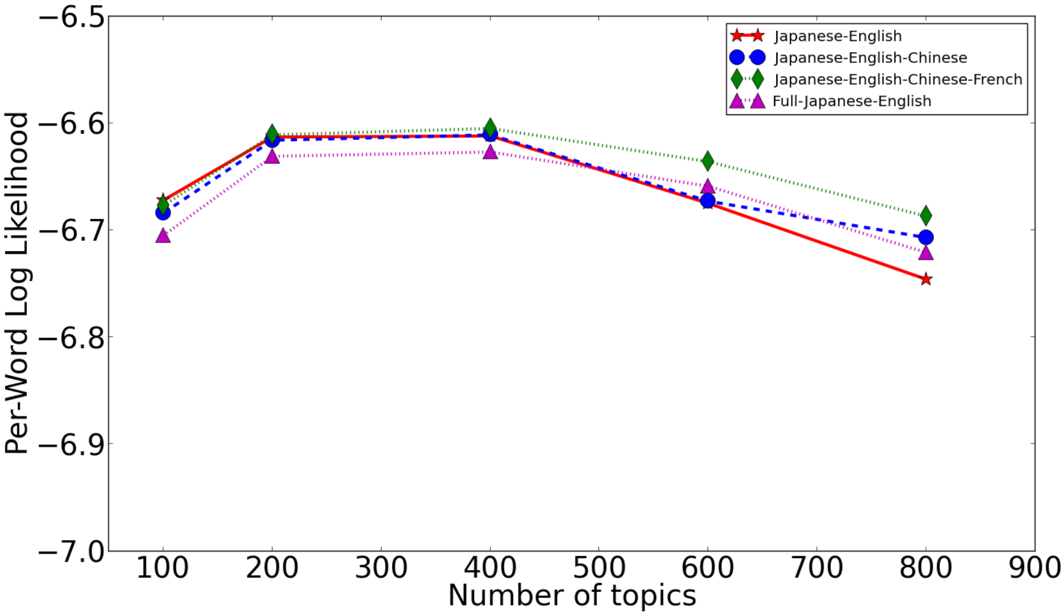


Figure 4.9: Per-word Likelihood by number of topics. Note that this figure correlates with Figure 4.8, suggesting per-word likelihood is a reasonable unsupervised metric for model selection.

Figure 4.9 summaries our results for the model selection, plotting the per-word likelihoods of a 100-tuple held-out dev set. We observe that per-word likelihood successively picks out $K = 400$ as the best model for various setups, which generally corresponds to the best precision results in Figure 4.8.

Finally, we show the run-time of MLTM on a 2.4GHz Opteron CPU for varying K in Figure 4.10. As expected, run-time increases with K: on datasets as large as ours, training with $K = 400$ takes approximately 4 days, and $K = 800$ takes 8 days. Time complexity of MLTM is $O(NK\sum_{m=1}^{M}\sum_{l=1}^{L} w_m^l)$, where N indcates number of iterations; K, M, L denotes the number of the topics, size of corpus and numbers of the languages; w_m^l denotes the number of words in tuple m written in language l.

The overall time for various systems is shown in Table 4.5. First, note that MLTM time dominates the overall time for all systems, so the training time does not differ

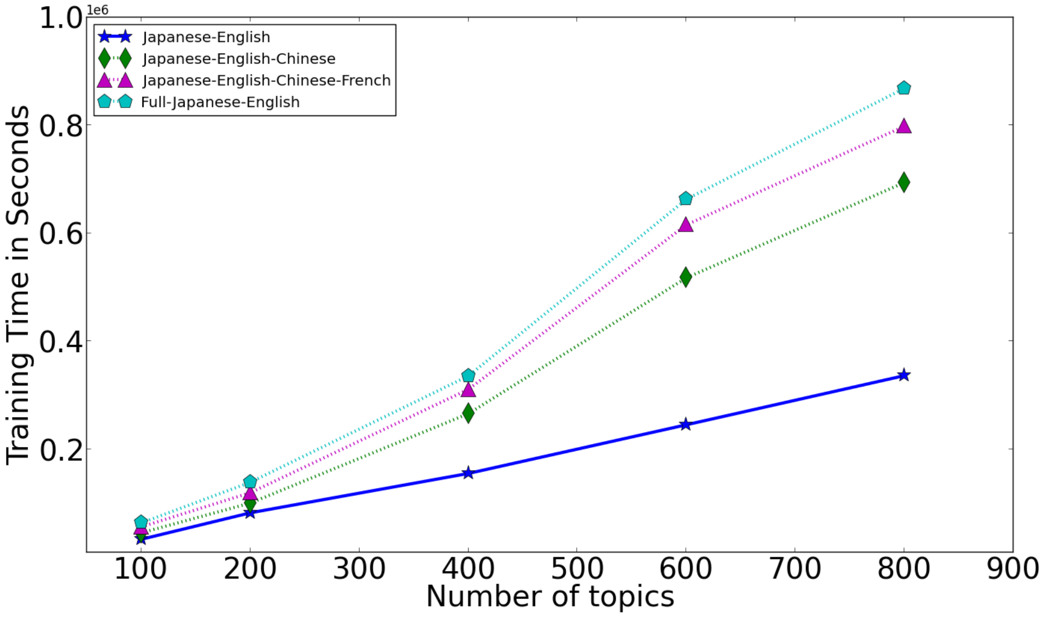


Figure 4.10: Training time of MLTM by dataset and number of topics.

much among methods if we use the same number of topics in MLTM; but in practice **Proposed** requires fewer number of topics, so it is much faster to train. Second, assuming the same number of topics, the breakdown of training time show that **Proposed** is still relatively fast because both GIZA++ and Eq. 3.4 are fast. Comparing Eq. 3.9 of **Cue** to Eq. 3.4 of **Proposed**, we see that both need to compute $p(t_k|w^f)$, but the $\sum_k$ in Eq. 3.4 tends to be faster because $p(w^e|w^f,t_k)$ in Eq. 3.4 tends to be sparse while $p(w^e|t_k)$ in Eq. 3.9 is dense.

4.4.5 Detailed Analyses of Results

Some examples of how the proposed multilingual model reduces translation errors are shown in Table 4.6. Taking "music" as an example, if only use the English-Japanese corpus, we erroneously find that "歌唱" (to sing) has high translation probability; this is understandable, though, because the words are roughly in the same topic. However, with additional language data (Chinese, French), the topic distributions becomes more precise, so the error disappears and the correct translation "音楽" (music) is left with higher probability.

#Topic	MLTM (En-Ja)	Proposed(Giza++/Eq 3.4)	JS(Eq 3.11)	Cue (Eq 3.9)
100	3.4e4	472 + 50	631	361
200	8.2e4	491 + 99	1,312	720
400	1.6e5	608 + 202	2,345	1,031
600	2.4e5	815 + 279	3,729	1,424
800	3.3e5	830 + 371	4,216	2,043

Table 4.5: Wall-clock times in seconds for Word Alignment (giza), and $p(w^e|w^f)$ calculation. Overall time for **Proposed** is the training time of MLTM, word alignment (Giza++) plus Eq.3.4; for **Cue** it is the training time of MLTM plus Eq.3.9; for **JS** it training time of MLTM plus Eq.3.11. Here, Eq.3.9 and Eq.3.11 are computed in parallel with 100 threads. **MLTM (En-Ja)** denotes the training time of multilingual topic model on English and Japanese corpus. The training time of multilingual topics for different settings is shown in Figure 4.10.

English Words	English-Japanese	+Chinese	+Chinese+French
music	音楽 [music](0.323) 歌唱 [sing] (0.203)	音楽 [music](0.442)	音楽 [music] (0.445)
ikoma	生駒 [ikoma] (0.497) 石切 [ishikiri] (0.205) 近鉄 [train] (0.272)	生駒 [ikoma] (0.574) 近鉄 [train] (0.252)	生駒 [ikoma] (0.619) 近鉄 [train] (0.254)
yoshino	吉野 [yoshino] (0.389)	吉野 [yoshino] (0.603)	吉野 [yoshino](0.373) 吉野山 (0.204)

Table 4.6: Error analysis on multilingual case. The words colored in red indicate translation errors; the words in [*] are the corresponding translation; the numbers in (*) are the translation probabilities.

4.5 Summary

We extend our framework by 1) proposing a novel multilingual topic model to handle *partially-connected* corpora; 2) using additional multilingual comparable corpora. On large-scale experiments, we show improvements in the precision of our Japanese-English lexicon as we include more languages, i.e. Chinese and French, to the comparable corpora.

Chapter 5

Hybrid Systems: Boosting for Context Vector Approach

In this chapter, we propose a hybrid system, which works as follows: 1) extracting the bilingual seeds by the proposed framework, as Figure 3.1, and then 2) boosting the context vector approaches. The rest of this chapter is organized as follows: in Section 5.1, we introduce our proposed method and then we describe the experiments and analysis in Section 5.2.

5.1 Proposed approach

Most researches on bilingual dictionaries extraction either focus on the *context-vector* methods [51, 19, 32, 16] or *projection-based* approaches [21]. Here, we propose a simple system, which combines two methods. Figure 5.1 shows the framework of our hybrid system, which includes two parts: 1) the top part is the same as the topic model + word alignment model approach introduced in Chapter 3 and the bottom is a context-vector method. Differing with other context-vector methods, our system does not use a "gold" dictionary.

There are many ways to compute the similarity between a source word, w_e, and a target word, w_f. We will only introduce several metrics commonly used in context-vector approaches. One of the popular similarity metric is **cosine** similarity measure, defined as

$$score(\mathbf{w_e}, \mathbf{w_f}) = sim_{cosine}(\mathbf{w_e}, \mathbf{w_f}) = \frac{\sum_i^N w_{(e,i)} \times w_{(f,i)}}{\sqrt{\sum_i^N w_{(e,i)}^2}\sqrt{\sum_i^N w_{(f,i)}^2}}. \tag{5.1}$$

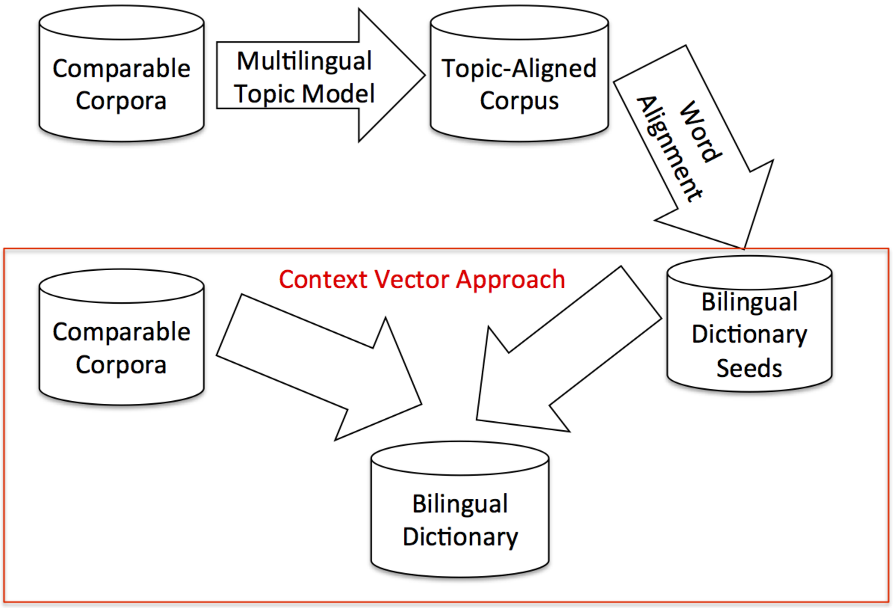


Figure 5.1: Hybrid system for bilingual dictionary extraction. Note that the top part is the same as the proposed framework, shown in Figure 3.1 and the bottom part is a context-vector method, which is similar to Figure 3.2.

The **Jaccard** similarity, which was originally designed for binary vectors, is also used widely in context vector approaches [35]. It is defined as:

$$score(\mathbf{w_e},\mathbf{w_f}) = sim_{jaccard}(\mathbf{w_e},\mathbf{w_f}) = \frac{\sum_i^N \min(w_{(e,i)},w_{(f,i)})}{\sum_i^N \max(w_{(e,i)},w_{(f,i)})}. \tag{5.2}$$

Alternatively, we can use the **Dice** measure, which is defined as

$$score(\mathbf{w_e},\mathbf{w_f}) = sim_{dice}(\mathbf{w_e},\mathbf{w_f}) = \frac{2 \times \sum_i^N \min(w_{(e,i)}w_{(f,i)})}{\sum_i^N (w_{(e,i)}+w_{(f,i)})}. \tag{5.3}$$

5.2 Experiment

The data used in our experiments are the same as the data described in Section 4.4.1. We evaluate this hybrid approach as follows:

- First, the 1,457 high precision dictionary seeds extracted by our proposed method in Table 4.3 are used as seed for the context vector approach of [51]. This hybrid system is called **WikiSeeds** and the resulting precision is reported in Table 4.4.

- For comparison, we run the context vector approach, the same context vector method as [51] with different amounts of "gold seeds". The purpose is to observe how many manual gold seed translations are necessary to attain the extraction result of our purely unsupervised **WikiSeeds** system. In particular, for a fair comparison, for cases under 1457 seeds (**GoldSeeds500** and **GoldSeeds1000**), we randomly sample 500 and 1,000 unique Japanese vocabularies in **WikiSeeds** and look-up their corresponding English translation in the "gold standard" lexicon described in Section 4.4.1. For cases above 1,457 seeds (**GoldSeeds1500** and **GoldSeeds3000**), we use all the gold standard lexicon associated with the 1,457 vocabulary, with additional translation pairs randomly sampled from the gold standard lexicon.1

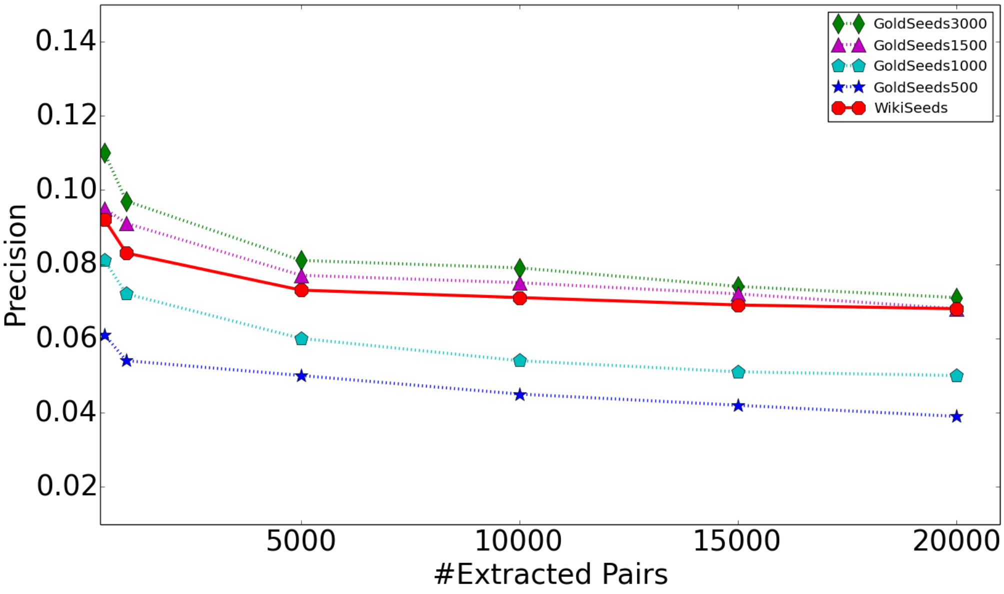


Figure 5.2: Comparison of different seeds for the context vector approach: Precision-vs-#Extracted pairs curve. Note that **GoldSeeds#** denotes the size of "gold seeds". Note that we use the cosine similarity as Eq 5.1.

From Figure 5.2, we find that **WikiSeeds** outperforms both 500 and 1000 gold seeds in precision across all numbers of extracted pairs. As expected, a roughly equal number of gold seeds (1500) outperforms **WikiSeeds**, but the differences are not large. Such observations imply that our extracted seeds can be used in context vector approach when there are no large seeds existing in certain language pairs. The code of the context vector approach is released at: `https://bitbucket.org/allenLao/context_based_model_for_dic/src`.

Further, we evaluate the effect of different similarity metrics to the accuracy of lexicons extraction. Here, we adopt the extracted seeds, **WikiSeeds**, and use the same context vector approach of [51] with different similarity metrics (**cosine**, **Jaccard** and **Dice**). Figure 5.3 summarizes the experiments of similarity measure comparison. We observe that **cosine** and **Jaccard** measures are very competitive, and outperform **Dice**.

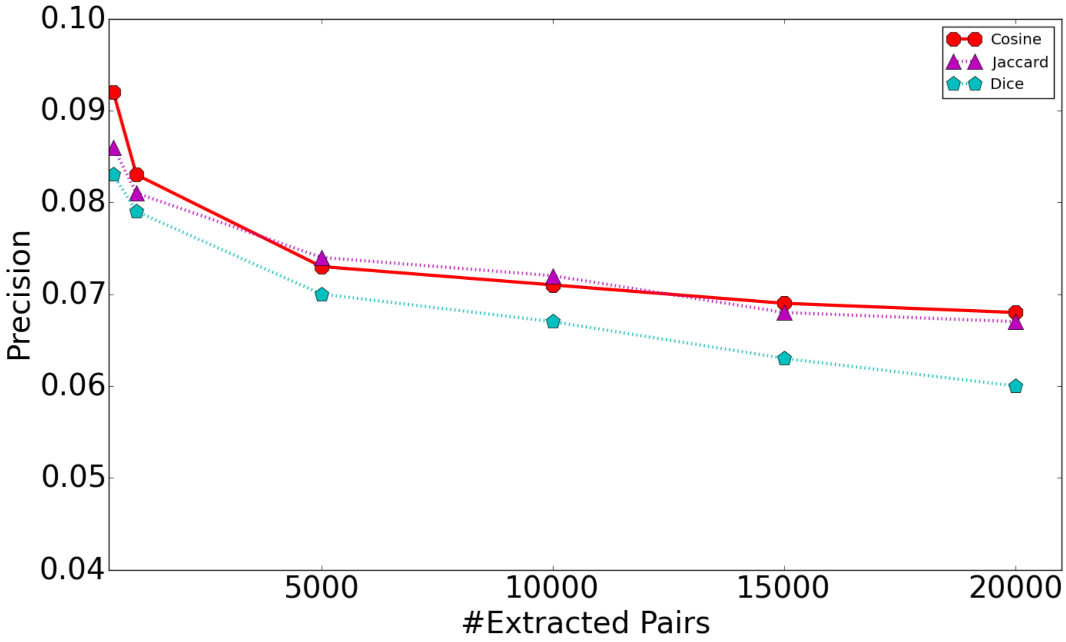


Figure 5.3: Comparison of different similarity metrics (cosine, Jaccard and Dice). Note that all the experiments are are used the extracted seeds, **WikiSeeds**.

5.3 Summary

In this chapter, we propose a hybrid system for bilingual dictionary extraction. We show that the context vector model by using the automatically extracted lexicons achieves similar results as by using the "gold" seeds. Furthermore, we show that the cosine and the Jaccard similarity metrics work better that the Dice similarity measure.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

In this thesis, 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 $p(w^e|w^f, t_k)$. The main advantages of our approach are that (1) it does not require any bilingual seed dictionary, and (2) it can effectively exploit comparable corpora consists of documents in more than two languages.

Our large-scale experiments demonstrate that the proposed framework outperforms existing baselines under both automatic metrics and manual evaluation. Further, we show improvements in the precision of our Japanese-English lexicon as we include more languages, i.e. Chinese and French, to the comparable corpora. Last, we use the extracted seeds to boost the context vector based model and show that such kind of "noisy" seeds are as good as "gold" seeds. To facilitate further work in this area, all preprocessed data and topic modeling code are available at
https://bitbucket.org/allenLao/topic-modeling-gibbs.

6.2 Future Work

In this section, I will discuss some open challenges and future work as follows.

- **Scalability**: Both Table 3.7 and Figure 4.10 show that multilingual topic models take most of time in our framework, if the topic number K is big. It prevents our framework from handling a very large-scale corpora, such as the whole Wikipedia. The problem can be solved by either using distributed learning algorithms [63], e.g., Zhao et al., training the topic model in the framework of MapReduce, or stochastic online learning for topic models [26].

- **Seeds**: While our framework is purely unsupervised in the sense that it requires no seed dictionary, we can imagine several interesting extensions if such a seed dictionary is available. First, the seeds could be used as a prior for the multilingual topic model, for instance by employing the Dirichlet tree prior of [4, 28]. Second, the seed translation could also be incorporated into the word alignment step (as supervised alignments) to improve performance of the topic-dependent translations, $p(w^f|w^e,t_k)$. In general, the modularity of our method makes it relatively flexible to incorporate additional resources and knowledge into the lexicon extraction process.

- **Bilingual Phrase Lexicon**: To our best knowledge, there are not any researches which focus on the bilingual phrase lexicon extraction, while it is very important in Machine translation community. First, the phrase lexicons can be extracted at the preprocessing step using a chunker, e.g., [33], which is not available for many languages. Second, we can solve such problem in post-multilingual topic model as [8], which is a purely unsupervised learning approach and can be easily adapted to other languages.

Bibliography

[1] A. Aker, M. Paramita, M. Pinnis, and R. Gaizauskas. Bilingual dictionaries for all eu languages. In N. C. C. Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis eds., *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA).

[2] Y. Al-Onaizan, J. Curin, M. Jahr, K. Knight, J. Lafferty, D. Melamed, F.-J. Och, D. Purdy, N. A. Smith, and D. Yarowsky. Statistical machine translation. In *Final Report, JHU Summer Workshop*, Vol. 30, 1999.

[3] D. Andrade, T. Matsuzaki, Jun 'ichi. Effective use of dependency structure for bilingual lexicon creation. In *Computational Linguistics and Intelligent Text Processing*, pp. 80–92. Springer, 2011.

[4] D. Andrzejewski, X. Zhu, and M. Craven. Incorporating domain knowledge into topic modeling via dirichlet forest priors. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 25–32. ACM, 2009.

[5] Y. Arai, T. Fukuhara, H. Masuda, and H. Nakagawa. Analyzing interlanguage links of wikipedias. In *Proceedings of the Wikimania Conference*, 2008.

[6] T. Baldwin. 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, pp. 1–1, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[7] D. M. Blei and M. I. Jordan. Variational inference for dirichlet process mixtures. *Bayesian analysis*, 1(1):121–143, 2006.

[8] D. M. Blei and J. D. Lafferty. Visualizing topics with multi-word expressions. *stat*, 1050:6, 2009.

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[10] F. Bond, H. Isahara, K. Kanzaki, and K. Uchimoto. Bootstrapping a wordnet using multiple existing wordnets. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, 2008.

[11] J. Boyd-Graber and D. M. Blei. Multilingual topic models for unaligned text. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 75–82. AUAI Press, 2009.

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

[13] H. Daume III and J. Jagarlamudi. 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*, pp. 407–412, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[14] J. Davis and M. Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning (ICML)*, pp. 233–240. ACM, 2006.

[15] G. de Melo and G. Weikum. Towards a universal wordnet by learning from combined evidence. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*, pp. 513–522, New York, NY, USA, 2009. ACM.

[16] H. Déjean, E. Gaussier, and F. Sadat. 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, pp. 1–7, 2002.

[17] K. Duh, C.-M. A. Yeung, T. Iwata, and M. Nagata. Managing information disparity in multilingual document collections. *ACM Trans. Speech Lang. Process.*, 10(1):1:1–1:28, Mar. 2013.

[18] P. Fung and P. Cheung. 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 (EMNLP)*, 2004.

[19] P. Fung and L. Y. Yee. An ir approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, pp. 414–420. Association for Computational Linguistics, 1998.

[20] J. Gao, K. Toutanova, and W.-t. Yih. Clickthrough-based latent semantic models for web search. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pp. 675–684. ACM, 2011.

[21] E. Gaussier, J. Renders, I. Matveeva, C. Goutte, and H. Dejean. 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*, pp. 526–533, Barcelona, Spain, July 2004.

[22] T. Gollins and M. Sanderson. Improving cross language retrieval with triangulated translation. In *Proceedings of the 24th ACM Conference of the Special Interest Group in Information Retrieval (SIGIR)*, 2001.

[23] G. Haffari and Y. W. Teh. Hierarchical dirichlet trees for information retrieval. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 173–181. Association for Computational Linguistics, 2009.

[24] A. Haghighi, P. Liang, T. Berg-Kirkpatrick, and D. Klein. Learning bilingual lexicons from monolingual corpora. In *Proceedings of ACL-08: HLT*, pp. 771–779, Columbus, Ohio, June 2008. Association for Computational Linguistics.

[25] G. Heinrich. Parameter estimation for text analysis, 2004.

[26] M. Hoffman, F. R. Bach, and D. M. Blei. Online learning for latent dirichlet allocation. In *advances in neural information processing systems (NIPS)*, pp. 856–864, 2010.

[27] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 50–57. ACM, 1999.

[28] Y. Hu, K. Zhai, V. Eidelman, and J. Boyd-Graber. Polylingual tree-based topic models for translation domain adaptation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1166–1176, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[29] J. Jagarlamudi and H. Daumé III. Extracting multilingual topics from unaligned comparable corpora. In *Advances in Information Retrieval*, pp. 444–456. Springer, 2010.

[30] A. Klementiev, I. Titov, and B. Bhattarai. Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*, pp. 1459–1474, Mumbai, India, December 2012. The COLING 2012 Organizing Committee.

[31] P. Koehn. *Statistical Machine Translation.* Cambridge University Press, New York, NY, USA, 1st edition, 2010.

[32] P. Koehn and K. Knight. Learning a translation lexicon from monolingual corpora. In *Proceedings of ACL Workshop on Unsupervised Lexical Acquisition*, 2002.

[33] T. Kudo and Y. Matsumoto. Chunking with support vector machines. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pp. 1–8. Association for Computational Linguistics, 2001.

[34] H.-s. Kwon, H.-w. Seo, and J.-h. Kim. Bilingual lexicon extraction via pivot language and word alignment tool. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pp. 11–15, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

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

[36] W. Li and A. McCallum. Semi-supervised sequence modeling with syntactic topic models. In *Proceedings of the National Conference on Artificial Intelligence*, p. 813. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2005.

[37] P. Liang, T. Ben, and K. Dan. Alignment by agreement. In *Proceedings of HLT-NAACL*, pp. 104–111, New York City, USA, June 2006. Association for Computational Linguistics.

[38] P. Liang, S. Petrov, M. I. Jordan, and D. Klein. The infinite pcfg using hierarchical dirichlet processes. In *EMNLP-CoNLL*, pp. 688–697. Citeseer, 2007.

[39] X. Liu, K. Duh, and Y. Matsumoto. Topic models + word alignment = a flexible framework for extracting bilingual dictionary from comparable corpus. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pp. 212–221, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[40] A. Lopez. Statistical machine translation. *ACM Computing Surveys (CSUR)*, 40(3):8, 2008.

[41] B. Magnini, C. Strapparava, F. Ciravegna, and E. Pianta. Multilingual lexical knowledge bases: Applied wordnet prospects. In *Proceedings of the International Workshop on "The Future of the Dictionary"*, 1994.

[42] D. Mimno, H. M. Wallach, J. Naradowsky, D. A. Smith, and A. 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.

[43] T. Minka. Estimating a dirichlet distribution, 2000.

[44] D. S. Munteanu and D. Marcu. Processing comparable corpora with bilingual suffix trees. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pp. 289–295. Association for Computational Linguistics, 2002.

[45] G. Neubig, Y. Nakata, and S. Mori. Pointwise prediction for robust, adaptable japanese morphological analysis. In *The 49th Annual Meeting of the Associa-*

tion for Computational Linguistics: Human Language Technologies (ACL-HLT) Short Paper Track, pp. 529–533, Portland, Oregon, USA, 6 2011.

[46] X. Ni, J.-T. Sun, J. Hu, and Z. Chen. Mining multilingual topics from wikipedia. In *Proceedings of the 18th international conference on World wide web*, pp. 1155–1156. ACM, 2009.

[47] X. Ni, J.-T. Sun, J. Hu, and Z. Chen. Mining multilingual topics from wikipedia. In *Proceedings of the 18th international conference on World wide web*, pp. 1155–1156. ACM, 2009.

[48] F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29(1):19–51, Mar. 2003.

[49] S. C. A. P, S. Lauly, H. Larochelle, M. M. Khapra, B. Ravindran, V. C. Raykar, and A. Saha. An autoencoder approach to learning bilingual word representations. *CoRR*, abs/1402.1454, 2014.

[50] M. Paul, H. Yamamoto, E. Sumita, and S. Nakamura. On the importance of pivot language selection for statistical machine translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pp. 221–224. Association for Computational Linguistics, 2009.

[51] R. Rapp. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 1995.

[52] P. Resnik, D. Oard, and G. Levow. Improved cross-language retrieval using backoff translation. In *Proceedings of the first international conference on Human language technology research*, pp. 1–3. Association for Computational Linguistics, 2001.

[53] D. Riley and D. Gildea. Improving the performance of giza++ using variational bayes. 2010.

[54] F. Sadat, H. Dejean, and E. Gaussier. A combination of models for bilingual lexicon extraction from comparable corpora. In *Proceedings of the Seminaire Papillon*, 2002.

[55] J. R. Smith, C. Quirk, and K. Toutanova. Extracting parallel sentences from comparable corpora using document level alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 403–411. Association for Computational Linguistics, 2010.

[56] S. Soderland, O. Etzioni, D. S. Weld, M. Skinner, J. Bilmes, et al. Compiling a massive, multilingual dictionary via probabilistic inference. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pp. 262–270. Association for Computational Linguistics, 2009.

[57] A. Tamura, T. Watanabe, and E. Sumita. 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*, pp. 24–36, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

[58] Y. W. Teh, D. Newman, and M. Welling. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In *Advances in neural information processing systems (NIPS)*, pp. 1353–1360, 2006.

[59] A. Vaswani, L. Huang, and D. Chiang. Smaller alignment models for better translations: unsupervised word alignment with the l 0-norm. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pp. 311–319. Association for Computational Linguistics, 2012.

[60] S. Volkova, T. Wilson, and D. Yarowsky. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 18–21, 2013.

[61] I. Vulić, W. De Smet, and M.-F. Moens. 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*, pp. 479–484, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[62] H. Wu and H. Wang. Revisiting pivot language approach for machine translation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 154–162, Suntec, Singapore, August 2009. Association for Computational Linguistics.

[63] K. Zhai, J. Boyd-Graber, N. Asadi, and M. L. Alkhouja. Mr. lda: A flexible large scale topic modeling package using variational inference in mapreduce. In *Proceedings of the 21st international conference on World Wide Web*, pp. 879–888. ACM, 2012.

[64] D. Zhang, Q. Mei, and C. Zhai. Cross-lingual latent topic extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1128–1137. Association for Computational Linguistics, 2010.

[65] B. Zhao and E. P. Xing. Hm-bitam: Bilingual topic exploration, word alignment, and translation. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1689–1696, 2008.

List of Publications

Journal Papers

- Xiaodong Liu, Kevin Duh and Yuji Matsumoto, Multilingual Topic Models for Bilingual Dictionary Extraction, *ACM Transactions on Asian Language Information Processing (TALIP)*, Accepted by Nov. 2014 and waiting for publish (**Chapter 4 and Chapter 5**).

- Xiaodong Liu, Fei Cheng, Kevin Duh and Yuji Matsumoto, A Hybrid Ranking Approach to Chinese Spelling Check, *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, Accepted by Mar. 2015 and waiting for publish.

International Conferences (refereed)

- Xiaodong Liu, Kevin Duh and Yuji Matsumoto, Topic Models + Word Alignment = A Flexible Framework for Extracting Bilingual Dictionary from Comparable Corpus, *Proceedings of the Seventeenth Conference on Computational Natural Language Learning (CoNLL)*, 212-221, Aug 2013 (**Chapter 3**).

- Xiaodong Liu, Kevin Duh, Tomoya Iwakura and Yuji Matsumoto, Learning Character Representations for Chinese Word Segmentation, *NIPS Workshop on Modern Machine Learning and Natural Language Processing*, Nov 2014.

- Xiaodong Liu, A Novel Joint Model of Word Alignment and Hierarchical Dirichlet Process for Statistical Machine Learning, *9th Conference on Bayesian Nonparametrics*, Poster, 2013.

- Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng and Kevin Duh, Representation Learning Using Multi-Task Deep Neural Networks for Semantic Classification and Information Retrieval, *2015 the North American Chapter of Association for Computational Linguistics-Human Language Technologies (NAACL-HLT 2015)*, Denver, Colorado, USA (to appear).

Non-refereed Publications

- Xiaodong Liu, Fei Cheng, Yanyan Luo, Kevin Duh and Yuji Matsumoto, A Hybrid Chinese Spelling Correction System Using Language Model and Statistical Machine Translation with Reranking, *Proceedings of Seventh SIGHAN Workshop on Chinese Language Processing*, 54-58, 2013.

- Xiaodong Liu, Kevin Duh and Yuji Matsumoto, A Flexible Framework for Extracting Bilingual Dictionary from Comparable Corpus Without Language- Specific Knowledge , 情報処理学会研究報告 第209回自然言語処理研究会, Vol.2012-NL-209, No.14, pp.1-8, Nov 2012.