A Generative Model for Extracting Parallel Fragments from Comparable Documents

Somayeh Bakhshaei[1], Shahram Khadivi[2*] and Reza Safabakhsh[1]
[1]Computer Engineering Department, Amirkabir University of Technology, Tehran, Iran
[2]eBay Inc., Aachen, Germany
bakhshaei@aut.ac.ir, skhadivi@ebay.com, safa@aut.ac.ir

Abstract

Although parallel corpora are essential language resources for many NLP tasks, they are rare or even not available for many language pairs. Instead, comparable corpora are widely available and contain parallel fragments of information that can be used applications like statistical machine translations. In this research, we propose a generative LDA based model for extracting parallel fragments from comparable documents without using any initial parallel data or bilingual lexicon. The experimental results show significant improvement if the extracted sentence fragments generated by the proposed method are used in addition to an existing parallel corpus in an SMT task. According to human judgment, the accuracy of the proposed method for an English-Persian task is about 66%. Also, the OOV rate for the same task is reduced by 28%.

1 Introduction

Parallel corpora are essential for many applications like statistical machine translation (SMT). Even resource rich language pairs in terms of parallel corpora always need more data, since languages evolve and diversify over time. Comparable corpora are considered as a widely available language resource that contains notably large amount of parallel sentence fragments. However, mining these fragments is a challenging task, and therefore many different approaches were proposed to extract parallel sentences, parallel fragments, or parallel lexicon. It has been shown in the previous works that extracting parallel sentences from comparable corpora usually results in a noisy parallel corpus (Munteanu & Marcu, 2006). Since comparable documents rarely contain exact parallel sentences, instead they contain a good amount of parallel subsentences or fragments. Thus, it is better to search for parallel fragments instead of parallel sentences.

Recent research works in fragment extraction have shown significant improvements in SMT quality, if parallel fragments are also used in the training phase (Chiao & Zweigenbaum, 2002; Déjean, et al., 2002; Fung & McKeown, 1997; Fung & Yee, 1998; Gupta, et al., 2013; Otero, P. G, 2007; Rapp, R., 1999; Saralegui, et al. 2008) . In this work, we also focus on extracting parallel fragments from comparable corpora. Our proposed approach is a generative model based on latent Dirichlet allocation (LDA) principles (Blei & Jordan, 2003).

In our proposed generative model, we assume there are parallel topics as hidden variables that model the parallel fragments in a comparable document corpus. We define parallel fragments as a sequence of occurrence of one of these parallel topics. This sequence occurs densely on a pair of comparable documents. It is possible to consider more than one topic in the structure of topic sequence but in this work we have limited it to one for simplicity and lower computational complexities. Considering more topics in the structure of a sequence that produces parallel fragments is suggested as our future work.

The rest of the paper is organized as follows. Section 2 describes the related works. Section 3 describes the generative process for producing comparable documents. The model architecture is described in section 4 with a graphical model. Section 5 describes the data, tools and resources

[*] This work has been done when Shahram Khadivi was with Amirkabir University of Technology.

Proceedings of the Eighth Workshop on Building and Using Comparable Corpora, pages 43–51,
Beijing, China, July 30, 2015.

used for this work and then the experiments and evaluation results are presented. Section 6 concludes and presents avenues for future works.

2 Related works

Comparable corpora are useful resources for many research fields of NLP. Also, SMT as one of the major problems of the NLP field can benefit from comparable corpora. Previous researches have suggested different approaches for extracting parallel information from comparable corpora. The main approaches are categorized as: **Lexicon Induction, Wikipedia based, Bridge Language, Graph based, Bootstrapping** and **EM**.

Works that are reported for **Lexicon Induction** are almost focused on extracting words from comparable corpora. These works use different methods that we categorize as: Seed based, model based and graph based methods.

The aim of the **Seed based Lexicon Induction** approach is expanding an initial parallel seed. Most of these researches use the context vector idea (Fung & Yee, 1998; Irvine & Callison-Burch, 2013; Rapp, 1995; Rapp, R., 1999). Gaussier, et al. (2004) proposes a geometric model for finding the synonym words in the space of the context vectors. Garera, et al. (2009) defines context vectors on the dependency tree rather than using adjacency. Some works use specific features for describing words like temporal co-occurrences (Schafer & Yarowsky, 2002), linguistic features (Kholy, et al., 2013; Koehn & Knight, 2002), and web based visual similarity features (Bergsma & Van Durme, 2011; Fiser & Ljubesic, 2011). The suggested features are almost efficient for similar or closely related languages but not all of the language pairs.

The **Model based Lexicon Induction** approach contains works that suggest a model for extracting parallel words. (Daumé III & Jagarlamudi, 2011; Haghighi, et al., 2008) use a generative model based on Canonical Correlation Analysis (CCA) (Hardoon, et al., 2004). They assume that by mapping words to a feature space, similar words are located in a subspace which is called the latent space of common concepts. Although their model is strong, they have defined it based on orthographical features (in addition to context vectors) that reduce the efficiency of the model for nonrelated languages. Diab & Finch (2000) also defines a matching function on similar words of languages. They assume that for two synonyms with close distributional profiles, the distributional profile of their corresponding translation should also be correlated in a comparable corpus. The optimization phase of the model that is based on gradient descent is very complex and time complexity is the biggest challenge of this model facing big data. The experiment is restricted to highly frequent words. Quirk, et al. (2007) also proposes a generative model. Their model is a developed version of IBM 1, 2 models. Although these are generative models for extracting parallel fragments, they completely differ from our model. Our model is based on the LDA model and we define a simpler but more efficient model with an accurate probabilistic distribution for parallel fragments in comparable corpora.

Wikipedia as a multilingual encyclopedia is a rich source of multilingual comparable corpora. There are lots of works reported in the **Wikipedia based** researches (Otero & López, 2010). Otero & López (2010) download the entire Wikipedia for any two languages, makes the "CorpusPedia", and then extracts information from this corpus. However, in recent works it is shown that only a small ad-hoc corpus containing Wikipedia articles can be beneficial for an existing MT system (Pal, et al., 2014). Although the Wikipedia based approach is a successful method for producing parallel information, the limitation of Wikipedia articles for most of the language pairs is a big problem.

The methods of Cross-lingual Information Retrieval are widely used for mining comparable corpora. The **Bridge language** idea is specially used for extracting parallel information between languages (Gispert & Mario, 2006; Kumar, et al., 2007; Mann & Yarowsky, 2001; Wu & Wang, 2007). Some papers use multiple languages for pivoting (Soderland, et al., 2009). The big problem of this approach is its unavoidable noisy output. Thus some other papers use a two-step version of this model for solving the problem. They first produce output and then refine it by removing its noise (Shezaf & Rappoport, 2010; Kaji, et al., 2008).

A wide range of researches are using a **Graph** for extracting parallel information from comparable corpora. Laws, et al., (2010) make a graph on the source (src) and target (trg) words (nodes are considered as src/trg words) and finds the similar nodes using the SimRank idea (Jeh & Widom, 2002). Some works define an optimization problem for finding the similarity on the edges of the graph of src and trg words (Muthukrishnan, et al., 2011). Razmara, et al.,

(2013) and Saluja & Navrátil, (2013) use graphs for solving the out-of-vocabulary (OOV) error in MT. Razmara, et al. (2013) make the nodes of the graph on phrases in addition to words. Minkov & Cohen (2012) use words and their stems for his graph nodes, and also the dependency tree for preserving the structure of words in source and target sentences. Some other works use the simple but efficient **EM** algorithm for producing a bilingual lexicon (Koehn & Knight, 2000).

A wide range of bootstraps are applied for extracting bilingual information from comparable corpora. Two-level approaches starts with (Munteanu & Marcu, 2006) that changes a sentence to a signal, based on LLR score and then uses a filter for extracting parallel fragments. This approach is continued in the latter works (Xiang, et al., 2013). Chu, et al. (2013) use the similar idea on quasi–comparable corpora. Klementiev, et al. (2012) use a heuristic approach for making context vectors directly on parallel phrases instead of parallel words. (Aker & Gaizauskas, 2012; Hewavitharana & Vogel, 2013) define a classifier for extracting parallel fragments.

3 LDA Based Generative Model

For extracting parallel fragments we use the LDA concept (Blei & Jordan, 2003). The base of our model is a bilingual topic model. Bilingual topic models were studied in previous works. Multilingual topic models similar to this work were presented in (Ni, et al., 2009) and (Mimno, et al., 2009). However, their models are polylingual topic models that are trained on words and our model is the extended version of this type of models but with additional capability of producing parallel fragments. In (Boyd-Graber J. a., 2009) a bilingual topic model is presented. The model is trained on a pair of src and trg words which are prepared by a matching function while training topic models. Another proposed model is (Boyd-Graber & P. Resnik, 2010) that is a customized version of LDA for sentimental analysis.

We infer topics as distributions over words as usual in topic model but the model is biased to a specific distribution of topics over words of documents. We assume that a pair of comparable documents is made of a topic distribution. We define topics over words but only the topics that are proper for producing parallel fragments are chosen. Therefore we limit them to ones that produce a dense bilingual sequence of source and target words in a comparable document pair. We use a definite function for controlling the topics and producing parallel fragments; this function is called $m()$. Function m accepts pairs of fragments, $\langle f^s, f^t \rangle$, if Conditions (1) satisfies and rejects them otherwise. The graphical presentation of proposed model is depicted in Figure 1. Model variables and relations are also shown in the figure. Here, we have used a known variable m.

Each pair of comparable documents will be generated with the generative process of Table 1. In this process β^s, β^t and α are hyper-parameters of the Dirichlet distributions. Topic distribution ϕ^t and ϕ^s is drawn from $Dir(\beta^t)$ & $Dir(\beta^s)$ respectively. First a sample distribution $\theta \sim Dir(\alpha)$ is drawn for both source and target document. Then each word of the comparable document pair is drawn from a multinomial distribution parameterized with θ, $z \sim Mult(\theta)$. Source and target words are generated from the respective topic distribution: $w^* \sim \phi^*_{w^*|z}$.

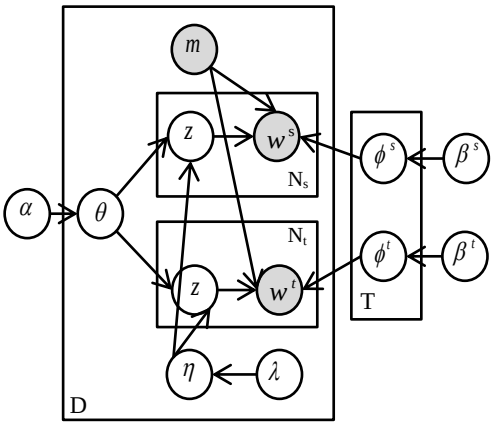

Figure 1: Graphical model for extracting parallel fragments. Function m determines if the assigned sequence of topics to the observed fragments satisfies Conditions (1). If so, the parallel pair of source and target fragments will be produced.

According to the given definition for parallel fragments in section 1, we produce a dense sequence of topics. In fact, by a dense sequence of a topic we mean a sub sentence of source and target document with limited length in which most of its words come from one topic distribution. For controlling these sub-sentences, we define the following conditions:

1. Length of the fragments is limited,
2. At least 50% of the words of a valid fragment come from one specific topic.

We refer to these conditions as Conditions (1) in the rest of the paper.

1. *Sample source & target topic distributions*:
   $\phi^s \sim Dir(\beta^s)$, $\phi^t \sim Dir(\beta^t)$,
2. *For each document pair* $D =< S, T >$,
   a) *Sample a topic distribution*: $\theta \sim Dir(\alpha)$,
   b) *Sample the length of parallel fragments*:
   $\eta \sim Poisson(\lambda)$,
   c) *Produce* η *random indexes from* S & T,
   d) *Sample a topic* $z' \sim Mult(\theta)$,
   e) *Sample each of the chosen indexes*:
   $w^* \sim \phi_{w^*|z'}$,
   f) *Sample the rest of indexes*:
   $z \sim Mult(\theta), w^* \sim \phi_{w^*|z}$,
   g) *If sequence of chosen topics satisfies conditions*(1):
   produce parallel fragments.

Table 1: Generative model for producing comparable document corpus.

4 Inference

Given a corpus of comparable documents our goal is to infer the unknown parameters of the model. According to Figure 2 we infer topics ϕ^t and ϕ^s, distribution of parallel topics on the source and target documents, θ and topic assignment z.

We use a collapsed Gibbs sampler (Neal, 2000) for sampling the latent variable of topic assignment z. We use two sets *I* and *J*. These two sets are random indexes chosen from source and target word indexes of the source and target documents, respectively:

$$\{I\}_1^{\eta} = Rand(N_{d,s})$$

$$\{J\}_1^{\eta} = Rand(N_{d,t})$$

The size of these two sets is defined based on the maximum length of parallel fragments in each document pair. The maximum length of parallel fragments, η, is randomly sampled from a Poisson distribution, $Poisson(\lambda)$:

$$\eta \sim Poisson(\lambda)$$

The words that appear in the indexes of sets I and J are respectively shown as *w(I)* and *w(J)*. Words of these two sets are made from one topic and build the dense sequence of words. We set $N_k^{w(I)}$ *and* $N_k^{w(J)}$ as the number of assignment of topic k to the source and target words, $w(I)$ *and* $w(J)$, occurring in the words indexes of the sets *I* and *J*. Also $N_{I,\neg i}^k$ *and* $N_{J,\neg i}^k$ are the number of times topic *k* occurs in the indexes defined in *I* and *J* sets in the source and target documents.

$$p(z_i = k \mid z_{-i}, I, J, \phi^s, \phi^t) = \mathrm{p}(I \mid z_i = k, \phi^s, z_{-i})\, p(\phi^s)\, \mathrm{p}(J \mid z_i = k, \phi^t, z_{-i})\, p(\phi^t)$$

$$\approx \frac{N_{k,\neg i}^{w(I)} + \beta^s}{\sum_n N_{k,\neg i}^n + \beta^s} \frac{N_{I,\neg i}^k + \beta^s}{\sum_n N_{I,\neg i}^n + \beta^s} \times \frac{N_{k,\neg i}^{w(J)} + \beta^t}{\sum_n N_{k,\neg i}^n + \beta^t} \frac{N_{J,\neg i}^k + \beta^t}{\sum_n N_{J,\neg i}^n + \beta^t}$$

We assume source and target words that are located in the current index i, w_i^s *and* w_i^t, are a member of $w(I)$ *and* $w(J)$ respectively, but while we are generating a word index outside *I* and *J*, then $N_k^{w(I)}$ *and* $N_k^{w(J)}$ changes to $N_k^{w_i}$ *and* $N_k^{w_j}$.

Finally function $m()$ produces parallel fragments $< f^s, f^t >$ only if they are consistent with Conditions (1) defined in Section 3.

Corpus		#Documents	#Words
Raw_ccNews	en	194K	47M
	fa	194K	42M
Refined_ccNews	en	97K	29M
	fa	97K	23M

Table 2: Statistics of used comparable corpora. The number of documents and running words is reported for each side of the corpus.

	Side	#Fragments	#Words
Extracted parallel fragments	en	75K	416K
	fa	75K	448K

Table 3: Statistic of extracted parallel fragments.

5 Experimental Setup

We have two strategies for evaluating our model. In the first step we try to measure the quality of extracted fragments from comparable documents. In the other scenario we evaluate the quality of the extracted parallel fragment by evaluating the quality of the SMT system equipped with this extra information.

5.1 Data

The data we use is a corpus of comparable documents, ccNews. The languages of these data are Farsi (fa) and English (en). The domain of these documents is News gathered between years 2007

#	src/trg	Worse parallel fragment samples	Error type
1	En	permanent security council members	Type 1.1 in target fragment.
	Fa	شورای دایم سازمان امنیت	
	En CT	permanent security council	
2	En	nobel peace prize winner	Type 1.2 in target fragment.
	Fa	برنده هندی جایزه صلح نوبل	
	En CT	indian nobel peace prize winner	
3	En	official irna news agency	Type 2.
	Fa	خبرگزاری نیمه رسمی فارس	
	En CT	official fars news agency	
4	En	eu foreign	Type 1.1 in source fragment.
	Fa	مسئول سیاست خارجی اتحادیه اروپا	
	En CT	eu foreign policy chief	
5	En	unanimously adopted the resolution imposing sanctions	Type 3.
	Fa	شورای امنیت سازمان ملل روز شنبه به اتفاق آرا	
	En CT	the un security council on saturday unanimously	

Table 4: Some worse parallel fragments produced by our model are recognized by manually checking the model output. The errors are highlighted and the correct translation of English part for the extracted Farsi fragment is written in EnCT row.

to 2010. The raw version of this corpus (Raw_ccNews) has about 193K documents and about 47M and 42M words, respectively in en and fa sides. We did some refinement on the corpus and the result is named Refined_ccNews corpus, as seen in Table 2. We removed repeated documents and also pairs of documents with incompatible ratio of words are removed.

The incompatibility of words ratio is defined as the proportion of words of one side to the other side. This ratio is set to be in the interval [0.5, 2]. That is:

$$0.5 \le \frac{\#\,words\ of\ \ source\ side\ document}{\#\,words\ of\ \ t\arg et\ side\ document} \le 2$$

The full information of the corpus is reported in Table 2.

5.2 Topic Model Parameters

In the experiments the hyper-parameter of the model are manually set to β^s, β^t=0.8 and $\alpha = 1$. And the number of topics in the models is set to T=800. The side effect of the training model is a parallel topic model. These topics are those that have common words with the source and target side of at least one comparable document pair. The iteration of Gibbs sampling is set to 1000.

The parallel fragments of the last iteration produced by $m()$ function are reported as the final result.

5.3 Results Analysis

The statistic of extracted parallel fragments is reported in Table 3. On average, 75K parallel fragments are extracted from 97K comparable documents. These numbers show that the model just produces high confidence samples and ignores most of them.

Evaluation Strategy 1 - According to our knowledge there is no criterion to automatically evaluate the quality of extracted data. Thus for evaluating the quality of the results we use human judgment. We asked a human translator familiar with both Farsi and English languages to check the quality of the parallelized fragments and mark the pairs that are wrongly parallelized and to write down a definition of the occurred error.

The results of manually checking the extracted fragments are shown in Table 4. In this table we have reported some of the worst errors of the model.

According to human judge, we recognized some specific types of error in the model output. These errors are categorized into three types:

1. Wrong boundaries for parallel fragments,
   1.1. Tighter boundaries that lead to incomplete phrases,
   1.2. Wider boundaries that lead to additional wrong tokens in the start/end of parallel fragments.
2. Same class words that are not the exact translation of each other.

3. Completely wrong samples,

Type 1 error is related to the samples in which boundaries are not correctly chosen by the model. This type is separated into two sub parts for tighter or wider boundaries which respectively ignores or adds some key tokens to the parallel fragments which leads to error.

Type 2 errors are produced because of using co-class words instead of synonyms. This is because the model intentionally groups words based on co-occurrence instead of considering meaning which it has inherited from the LDA base of the model (the model is actually a topic model and this is a usual behavior of topic models). This bug of the model can be considered as future works for improving the model accuracy.

At the end, the reason for Type 3 errors is not obviously known. These samples are produced because of the inner noises of the model. We guess these are the unavoidable noises of comparable documents that are extended to the model output.

According to this classification of errors, the proportion of each error type is computed. The results are reported in Table 5. These are the proportion of each type observed in a set of 400 random fragments which is evaluated by human translator. The most observed error is related to type 1. Thus the human evaluation suggests 66% accuracy for the model output.

Evaluation Strategy 2 – In the second step, for evaluating the model output, we consider the effect of these extracted data in the quality of an existing SMT system. For this aim, at first we train a base line system on a parallel corpus. Our corpus is the Mizan parallel corpus[2]. The domain of this corpus is literature. For challenging the translation system, we used an out-of-domain test. Our test is selected from the news domain.

The standard phrase-based decoder that we use for training models is the Moses system (Koehn, et al., 2007) in which we use default values for all of the decoder parameters. We also use a 4-gram language model trained using SRILM (Stolcke, 2002) with Kneser-Ney smoothing (Kneser & Ney, 1995). To tune the decoder's feature weights with minimum error rate (Och, 2003), we use a development (dev) set of 1000 single-reference sentences, and we evaluate the models performance on a test set of 1032 multiple-references sentences. For more information on the data see Table 6. Domain of the dev set and training corpus is literature while the test set domain is news.

As it is seen in Table 7, different approaches are proposed for how to use parallel fragments for improving the baseline system. Description of the models is explained in the follow.

Baseline - This is an SMT system that is trained on main corpus (Mizan). The BLEU score of the baseline system is 10.41% on dev and 8.01% on test set. The OOV error in this system is 3509 and 768 on test and dev sets respectively.

Baseline+ParallelFragments - In this system we directly add the parallel fragments to our main corpus and train a new system. The BLEU score improvement is about 0.27% and 0.22% respectively on test and dev sets. OOV error reduces too.

Baseline+ParallelFragments (Giza weightes) - This approach is the same as **Baseline+ParallelFragments** but we use the weighted corpus for Giza alignment. The weight of main corpus and parallel fragments is set to 10 and 1 respectively.

BaseLine+PT_ParallelFragments - In this approach we combine the phrase tables of baseline and the system trained on parallel fragments. Actually because of the difference domain of main corpus and parallel fragments, it is expected that combining these two resources harm the quality of the baseline system. So, we use the phrase table which is trained on parallel fragments as the back off for the phrase table of the baseline system. The results show significant improvement in this case. The BLEU score improves by about 1% on test set and OOV error is decreased by 28%.

Thus, the results shown in Table 7 reveals that the extracted parallel fragments can improve the quality of the translation output.

Error Type	P
Type 1	33%
Type 2	0.04%
Type 3	0.02%

Table 5: Analysis of model output base of error types recognized by human translator judgment.

[2] Supreme Council of Information and Communication Technology. (2013). Mizan English-Persian Parallel Corpus. Tehran, I.R. Iran. Retrieved from http://dadegan.ir/catalog/mizan.

6 Conclusion

In this paper we have proposed a generative LDA based model for extracting parallel fragments from comparable corpora. The main contribution of the proposed model is that it is developed for extracting parallel fragments from comparable documents corpus without the need to any parallel data such as initial seed or dictionary.

We have evaluated the output of the model by using a human translator judgment and also by using the extracted data for expanding the training data set of a SMT system. Results of the augmented system show improvement of the output quality.

The result of human judgment categorizes the dominant errors of the model to three types. Most errors are related to the wrong recognized boundaries by the model. We have considered the refinement of these kinds of errors as our future works. We have also shown that the model is able to reduce the OOV error.

			#Line	#Words	Domain
Train	en		1021323	13636292	Literature
	fa		1021323	13686642	
Test	en		1032	28112	News
	fa	1	1032	30451	
		2	1032	33725	
		3	1032	33128	
		4	1032	32417	
Dev	en		1000	23055	Literature
	fa		1000	26351	

Table 6: Statistic of Train, Test and Dev set for making the SMT system.

SMT system	Test		Dev	
	BLEU (%)	OOV	BLEU (%)	OOV
Baseline	10.41	3509	8.01	768
+ParallelFragments	10.68	2459	8.22	737
+ ParallelFragments (Giza weighted)	10.53	2460	8.23	737
+PT_ParallelFragments	**11.46**	2530	8.14	734

Table 7: Results of trained SMT systems.

References

Aker, A., & Gaizauskas, Y. F. (2012). Automatic bilingual phrase extraction from comparable corpora. *Proceedings of COLING 2012: Posters* (pp. 23–32). COLING 2012, Mumbai.

Bergsma, S., & Van Durme, B. (2011). Learning bilingual lexicons using the visual similarity of labeled web images. *In IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, (pp. Vol. 22, No. 3, p. 1764).

Blei, D. M., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3, 993-1022.

Boyd-Graber, J. a. (2009). Multilingual topic models for unaligned text. *In Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence* (pp. 75-82). AUAI Press.

Boyd-Graber, J., & P. Resnik. (2010). Holistic sentiment analysis across languages: Multilingual supervised latent Dirichlet allocation. *In Proceedings of the 2010 Conference on EMNLP*, (pp. 45-55).

Chiao, Y. C., & Zweigenbaum, P. (2002). Looking for candidate translational equivalents in specialized, comparable corpora. *In Proceedings of the 19th international conference on Computational linguistics-Volume 2*, (pp. 1-5).

Chu, C., Nakazawa, T., & Kurohashi, S. (2013). Accurate Parallel Fragment Extraction from Quasi–Comparable Corpora using Alignment Model and Translation Lexicon. *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, (pp. 1144--1150).

Daumé III, H., & Jagarlamudi, J. (2011). Domain adaptation for machine translation by mining unseen words. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2.* Association for Computational Linguistics.

Déjean, H., Gaussier, É., & Sadat, F. (2002). Bilingual terminology extraction: an approach based on a multilingual thesaurus applicable to comparable corpora. *In Proceedings of the 19th International Conference on Computational Linguistics COLING*.

Diab, M., & Finch, S. (2000). *A statistical word-level translation model for comparable corpora*. IN PROCEEDINGS OF THE CONFERENCE ON CONTENT-BASED MULTIMEDIA INFORMATION ACCESS (RIAO).

Fiser, D., & Ljubesic, N. (2011). Bilingual lexicon extraction from comparable corpora for closely related languages. *In RANLP*, 125-131.

Fung, P., & McKeown, K. (1997). Finding terminology translations from non-parallel corpora. *In Proceedings of the 5th Annual Workshop on Very Large Corpora*, (pp. 192-202).

Fung, P., & Yee, L. Y. (1998). An ir approach for translating new words from nonparallel, comparable texts. *In Proceedings of the 36th Annual Meeting of the Association and 17th International Conference on Computational Linguistics - Volume 1, ACL '98* (pp. 414-420). Association for Computational Linguistics.

Garera, N., Callison-Burch, C., & Yarowsky, D. (2009). Improving translation lexicon induction from monolingual corpora via dependency contexts and part-of-speech equivalences. *Proceedings of the Thirteenth Conference on Computational Natural Languag Learning.* Association for Computational Linguistic.

Gaussier, E., Renders, J. M., Matveeva, I., Goutte, C., & Déjean, H. (2004). A geometric view on bilingual lexicon extraction from comparable corpora. *In Proceedings of the 42nd Annual Meeting on ACL*, (p. 526).

Gispert, A. d., & Mario, B. (2006). Catalan-english statistical machine translation without parallel corpus: bridging through spanish. *in Proc. of 5th International Conference on Language Resources and Evaluation (LREC).* Genoa, Italy.

Gupta, R., Pal, S., & Bandyopadhyay, S. (2013). Improving mt system using extracted parallel fragments of text from comparable corpora. *In proceedings of 6th workshop of BUCC, ACL, Sofia, Bulgaria*, (pp. 69-76).

Haghighi, A., Liang, P., Berg-Kirkpatrick, T., & Klein, D. (2008). Learning Bilingual Lexicons from Monolingual Corpora. *In ACL, Vol. 2008*, (pp. 771-779).

Hardoon, D. R., Szedmak, S., & Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural Computation 16.12*, 2639-2664.

Hewavitharana, S., & Vogel, S. (2013). Extracting parallel phrases from comparable data. *Building and Using Comparable Corpora. Springer Berlin Heidelberg*, 191-204.

Irvine, A., & Callison-Burch, C. (2013). Supervised bilingual lexicon induction with multiple monolingual signals. *Proceedings of NAACL-HLT.*

Jeh, G., & Widom, J. (2002). Simrank: A measure of structural-context similarity. *In KDD '02*, pages 538–543.

Kaji, H., Tamamura, S., & Erdenebat, D. (2008). Automatic construction of a japanese-chinese dictionary via english. *In LREC.*

Kholy, E., Nizar Habash, A., Leusch, G., Matusov, E., & Sawaf, H. (2013). Language independent connectivity strength features for phrase pivot statistical machine translation. In Proc. of ACL, vol. 13.

Klementiev, A., Irvine, A., Callison-Burch, C., & Yarowsky, D. (2012). Toward statistical machine translation without parallel corpora. *In Proceedings of the 13th Conference of the European Chapter of the ACL*, (pp. 130-140).

Kneser, R., & Ney, H. (1995). Improved backing-off for m-gram language modeling. *In Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on Vol. 1,* (pp. 181-184). IEEE.

Koehn, P., & Knight, K. (2000). Estimating word translation probabilities from unrelated monolingual corpora using the EM algorithm. *AAAI/IAAI.*

Koehn, P., & Knight, K. (2002). Learning a translation lexicon from monolingual corpora. *Proceedings of the ACL-02 workshop on Unsupervised lexical acquisition-Volume 9.* Association for Computational Linguistics.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., et al. (2007). Moses: Open source toolkit for statistical machine translation. *In Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, (pp. 177-180).

Kumar, S., Och, F. J., & Macherey, W. (2007). Improving Word Alignment with Bridge Languages. *EMNLP-CoNLL.*

Laws, F., Michelbacher, L., Dorow, B., Scheible, C., Heid, U., & Schütze, H. (2010). A linguistically grounded graph model for bilingual lexicon extraction. *In Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (pp. 614-622). Association for Computational Linguistics.

Mann, G. S., & Yarowsky, D. (2001). Multipath translation lexicon induction via bridge languages. *In NAACL.*

Mimno, D., Wallach, H. M., Naradowsky, J., Smith, D. A., & McCallum, A. (2009). Polylingual topic models. *In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2*, (pp. 880-889).

Minkov, E., & Cohen, W. W. (2012). Graph based similarity measures for synonym extraction from parsed text. *Workshop Proceedings of TextGraphs-7 on Graph-based Methods for Natural Language Processing.* Association for Computational Linguistics.

Munteanu, D. S., & Marcu, D. (2006). Extracting parallel sub-sentential fragments from non-parallel corpora. *In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL* , (pp. 81-88).

Muthukrishnan, P., Radev, D., & Mei, Q. (2011). Simultaneous similarity learning and feature-weight learning for document clustering.". *Proceedings of textgraphs-6: Graph-based methods for natural language processing.* Association for Computational.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics, 9(2),* 249-265.

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

Och, F. J. (2003). Minimum error rate training in statistical machine translation. *Association for Computational Linguistics.* Proceedings of the

41st Annual Meeting on Association for Computational Linguistics-Volume 1.

Otero, P. G. (2007). Learning bilingual lexicons from comparable english and spanish corpora. *Proceedings of MT Summit xI*, (pp. 191-198).

Otero, P. G., & López, I. G. (2010). Wikipedia as multilingual source of comparable corpora. *In Proceedings of the 3rd Workshop on BUCC, LREC*, (pp. 21-25).

Pal, S., Pakray, P., & Naskar, S. K. (2014). Automatic Building and Using Parallel Resources for SMT from Comparable Corpora. *In Proceedings of the 3rd Workshop on Hybrid Approaches to Translation (HyTra)@ EACL*, (pp. 48-57).

Quirk, C., Udupa, R., & Menezes, A. (2007). Generative models of noisy translations with applications to parallel fragment extraction. *Proceedings of the Machine Translation Summit XI*, (pp. 377-384).

Rapp, R. (1995). Identifying word translationsin non-parallel texts. In Proceedings of Association for Computational Linguistics.

Rapp, R. (1999). Automatic identification of word translations from unrelated english and german corpora. *In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, ACL '99* (pp. 519-526). Association for Computational Linguistics.

Razmara, M., Siahbani, M., Haffari, G., & Sarkar, A. (2013). Graph propagation for paraphrasing out-of-vocabulary words in statistical machine translation. InProceedings of the Conference of the Association for Computational Linguistics (ACL).

Saluja, A., & Navrátil, J. (2013). Graph-Based Unsupervised Learning of Word Similarities Using Heterogeneous Feature Types. *Graph-Based Methods for Natural Language Processin*.

Saralegui, X. I., San Vicente, I., & Gurrutx-aga, A. (2008). Automatic generation of bilingual lexicons from comparable corpora in a popular science domain. *In LREC 2008 workshop on BUCC.*

Schafer, C., & Yarowsky, D. (2002). Iducing translation lexicons via diverse similarity measures and bridge languages. *In proceedings of the 6th conference on Natural language learning - Volume 20, COLING-02* (pp. 1-7). Stroudsburg, PA, USA: Association for Computational Linguistics.

Shezaf, D., & Rappoport, A. (2010). Bilingual lexicon generation using non-aligned signatures. *In Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 98-107). Association for Computational Linguistics.

Soderland, S., Etzioni, O., Weld, D. S., Skinner, M., & Bilmes, J. (2009). 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*, (pp. 262-270).

Stolcke, A. (2002). SRILM-an extensible language modeling toolkit. *In INTERSPEECH.* .

Wu, H., & Wang, H. (2007). Pivot language approach for phrase-based statistical machine translation. *Machine Translation 21.3*, 165-181.

Xiang, L., Zhou, Y., & Zon, C. (2013). An Efficient Framework to Extract Parallel Units from Comparable Data. *Natural Language Processing and Chinese Computing Springer Berlin Heidelberg*, 151-163.