Journal of Zhejiang University-SCIENCE C (Computers & Electronics)
ISSN 1869-1951 (Print); ISSN 1869-196X (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Topic-aware pivot language approach for statistical machine translation*

Jin-song SU[†1,2], Xiao-dong SHI[3], Yan-zhou HUANG[3], Yang LIU[4], Qing-qiang WU[1,2], Yi-dong CHEN[3], Huai-lin DONG[1]

([1]*Software School, Xiamen University, Xiamen 361005, China*)

([2]*Center for Digital Media Computing, Xiamen University, Xiamen 361005, China*)

([3]*Cognitive Science Department, Xiamen University, Xiamen 361005, China*)

([4]*Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China*)

[†]E-mail: jssu@xmu.edu.cn

Received Aug. 4, 2013; Revision accepted Nov. 7, 2013; Crosschecked Feb. 19, 2014

Abstract: The pivot language approach for statistical machine translation (SMT) is a good method to break the resource bottleneck for certain language pairs. However, in the implementation of conventional approaches, pivot-side context information is far from fully utilized, resulting in erroneous estimations of translation probabilities. In this study, we propose two topic-aware pivot language approaches to use different levels of pivot-side context. The first method takes advantage of document-level context by assuming that the bridged phrase pairs should be similar in the document-level topic distributions. The second method focuses on the effect of local context. Central to this approach are that the phrase sense can be reflected by local context in the form of probabilistic topics, and that bridged phrase pairs should be compatible in the latent sense distributions. Then, we build an interpolated model bringing the above methods together to further enhance the system performance. Experimental results on French-Spanish and French-German translations using English as the pivot language demonstrate the effectiveness of topic-based context in pivot-based SMT.

Key words: Natural language processing, Pivot-based statistical machine translation, Topical context information
doi:10.1631/jzus.C1300208 **Document code:** A **CLC number:** TP391.1

1 Introduction

Recently, statistical machine translation (SMT) has obtained rapid development with more and more novel translation models being proposed and put into practice. Typically, bilingual data has an important influence on SMT system performance. The more data is used to train the translation model, the better SMT system we will obtain. However, it may not be easy to set up a large-scale bilingual corpus for many resource-poor language pairs. Therefore, how to break the bottleneck of training data is always a research focus in SMT.

To solve this problem, most researchers have focused on how to collect more sentence pairs. They either obtain more bilingual sentences by information retrieval technology (Hildebrand *et al.*, 2005) or convert the monolingual sentences into synthetic parallel ones by self-training (Ueffing *et al.*, 2007; Bertoldi and Federico, 2009). However, the bilingual corpus is scarce for many language pairs. In addition, the quality of the synthetic parallel sentences is not

* Project supported by the National High-Tech R&D Program of China (No. 2012BAH14F03), the National Natural Science Foundation of China (Nos. 61005052 and 61303082), the Research Fund for the Doctoral Program of Higher Education of China (No. 20120121120046), the Natural Science Foundation of Fujian Province of China (No. 2011J01360), and the Fundamental Research Funds for the Central Universities, China (No. 2010121068)

guaranteed. So, these methods may not be suitable for the translation task of resource-poor language pairs.

From a different perspective, some researchers have investigated how to use a pivot language method (Cohn and Lapata, 2007; Utiyama and Isahara, 2007; Wu and Wang, 2007; Bertoldi *et al.*, 2008; Tanaka *et al.*, 2009). Even if there is no available source-target bilingual corpus, this method can build a source-target translation model by bringing in a pivot language, for which there exist large scale source-pivot and pivot-target bilingual corpora. However, the conventional approach simply bridges both sides of the source-target phrase pair using the pivot phrase. Many pivot phrases are likely to have different meanings depending on a specific context, and this may result in erroneous estimations of source-target translation probabilities. For example, in a French-Spanish translation task using English as the pivot language, the French phrase 'banque' (a financial organization) and the Spanish phrase 'ribereño' (the border of a river) are respectively aligned to the English phrase 'bank'. Using the conventional approach, the phrase pair '(banque, ribereño)' is often induced, although their meanings are completely different. In fact, Wu and Wang (2007) have noticed this phenomenon and tried two methods to solve this problem: one estimates the lexical translation probability based on the co-occurrence frequency of the word pair in the induced phrase pairs; the other is embedded with the cross-language word similarity. However, such methods still have some limitations. First, they incorporate the context information at the word level rather than at the phrase level, while most SMT systems conduct translation by using sequences of phrases. Second, they exploit only the surface context while a larger scale context is totally ignored. Third, the cross-language word similarity is calculated by using a vector space model (VSM), which is prone to suffer from data sparseness. Therefore, we believe that the pivot-side context is far from being fully utilized.

In this study, we first propose two approaches to improve conventional pivot-based SMT with topic-based pivot-side context, and then build an interpolated model exploiting different levels of context to further enhance pivot-based SMT. Although based on the triangulation method proposed by Wu and Wang (2007), our methods can overcome the data sparsity, as well as capture the previously ignored context information. Specifically, we make the following contributions:

1. Exploiting document-level context: In the first method, we deal mainly with the effect of document-level context, which is ignored in conventional pivot-based SMT. Assuming that the bridged phrase pairs should be similar in the document-level topic distributions, we introduce the pivot-based document-level topic as a hidden variable in the implementation of conventional phrase table multiplication. For example, the French-English phrase pair (banque, bank) often occurs in the document about a finance topic, while the English-Spanish phrase pair (bank, ribereño) appears in the document related to geography or other topics, so the translation probability of the induced phrase pair (banque, ribereño) will decrease because they belong to different topics.

2. Overcoming data sparsity in the conventional representation of local context: In the second method, we focus on the effect of local context. Taking advantage of the topic model, the proposed method can overcome data sparsity in conventional representation. Assuming that the pivot words found in a corpus share a global set of latent senses, we employ a probabilistic model to induce the latent sense distribution of each phrase pair from its pivot-side context words, and discourage the bridged phrase pairs without compatible sense distributions. Also, in the example mentioned above, the proposed method will identify that the French-English phrase pair (banque, bank) has a different sense from the English-Spanish phrase pair (bank, ribereño), since the former is incompatible with the context words of the latter such as river and boat.

3. Combining different levels of context: In general, the two proposed methods apply different levels of context to improve pivot-based SMT. Finally, we build an interpolated model to combine the advantages of these two methods, aiming to further enhance the system performance.

We evaluate the proposed methods on the French-Spanish and French-German translation data sets. Experiments show that the methods significantly outperform the conventional pivot language approaches.

2 Pivot-based SMT by triangulation

The conventional pivot language approach proposed by Wu and Wang (2007) builds a source-target translation model through utilizing a phrase table multiplication. Specifically, their method consists mainly of two parts, phrase translation probability and lexical weight.

The phrase translation probability measures the co-occurrence frequency of a phrase pair. Assuming the independence between the source and target phrases when given the pivot phrase, Wu and Wang (2007) induced a source-target phrase pair $(\tilde{f},\tilde{e})$ from the source-pivot and pivot-target pairs $(\tilde{f},\tilde{p})$ and $(\tilde{p},\tilde{e})$, and calculated its phrase probability $\phi(\tilde{e}|\tilde{f})$ as follows (due to the limit of space, here we omit the computation of the phrase translation probability $\phi(\tilde{f}|\tilde{e})$ which can be calculated in a similar way):

$$\begin{aligned}\phi(\tilde{e}|\tilde{f}) &= \sum_{\tilde{p}} \phi(\tilde{e},\tilde{p}|\tilde{f}) \\ &= \sum_{\tilde{p}} \phi(\tilde{e}|\tilde{p},\tilde{f}) \cdot \phi(\tilde{p}|\tilde{f}) \\ &= \sum_{\tilde{p}} \phi(\tilde{e}|\tilde{p}) \cdot \phi(\tilde{p}|\tilde{f}).\end{aligned} \tag{1}$$

The lexical weight is used to validate the quality of the phrase pair by checking how well its words are translated to each other. For the lexical weight, two important elements should be considered: the alignment of the induced source-target phrase pair and lexical translation probability.

Given the word alignment information a_{fp} and a_{pe} inside phrase pairs $(\tilde{f},\tilde{p})$ and $(\tilde{p},\tilde{e})$, the alignment information a_{fe} inside phrase pair $(\tilde{f},\tilde{e})$ can be derived in the following way:

$$a_{fe} = \{(f,e)|\exists p : (f,p) \in a_{fp} \ \& \ (p,e) \in a_{pe}\}. \tag{2}$$

Then, Wu and Wang (2007) proposed a phrase method to estimate the lexical translation probability. They first collected the co-occurrence frequencies of word pairs according to the alignment information of the induced phrase pair, and then adopted the maximum likelihood estimation (MLE) method to calculate the lexical translation probability $w(e|f)$:

$$\text{count}(f,e) = \sum_{k=1}^{K} \phi_k(\tilde{e}|\tilde{f}) \sum_i \delta(f,\tilde{f}_i)\delta(e,\tilde{e}_{a_i}), \tag{3}$$

$$w(e|f) = \frac{\text{count}(f,e)}{\sum_{e'} \text{count}(f,e')}. \tag{4}$$

Herein K denotes the number of the induced phrase pairs. $\delta(x,y) = 1$ if $x = y$; otherwise, $\delta(x,y) = 0$. $\text{count}(f,e)$ represents the co-occurrence frequency of the word pair (f,e) in all induced phrase pairs.

Wu and Wang (2007) also tried to introduce cross-language similarity to adjust the lexical translation probability:

$$w(e|f) = \sum_p w(e|p) \cdot w(p|f) \cdot \text{sim}(f,e;p), \tag{5}$$

where $\text{sim}(f,e;p)$ is the cross-language similarity. Note that the phrase method performs better than the others according to the experimental results reported in Wu and Wang (2007).

Finally, with the derived word alignment information a_{fe} and lexical translation probability $w(e|f)$, the lexical weight $p_w(\tilde{e}|\tilde{f})$ for the induced phrase pair $(\tilde{f},\tilde{e})$ is calculated as in the conventional method (Koehn *et al.*, 2003). If there exist multiple alignments for the phrase pair, we keep only the one with the maximum lexical weight.

3 Topic-aware pivot-based SMT

In this section, we first briefly review the principle of latent Dirichlet allocation (LDA) (Blei *et al.*, 2003), which is the basis of our work. Then, we propose two topic-aware pivot language approaches to utilize different levels of context on the pivot side. Finally, we build an interpolated model to bring different pivot language methods together.

3.1 Latent Dirichlet allocation

Recently, topic models have been rapidly developed with many models being proposed and widely applied. Among these models, LDA is the most common one currently in use. Compared with other models, LDA has better performance (Blei *et al.*, 2003). Besides, it is a generative model with hyper parameters, which can be used to infer the topic distributions of unseen documents. Therefore, in this work, we use it rather than other models to mine the latent topics. Next, we give a brief description of LDA.

During the modeling process, LDA regards each document as a mixture proportion of various topics,

and generates each word by multinomial distribution under topics. Currently, there are mainly two methods which can be used to estimate parameters and conduct inference for LDA, variational inference and Gibbs sampling. In this work, we use the latter to train an LDA model because of its simplicity and widespread use. In the generation process of each document, the posterior document-topic distribution is sampled by LDA first. Then, for each word in the document, it samples a topic index from the document-topic distribution and samples the word conditioned on the topic index according to the topic-word distribution.

By LDA, the latent topics hidden in a collection of documents can be easily discovered in an unsupervised fashion. Specifically, we can obtain two types of parameters: one is topic-word distribution representing each topic as a distribution over words; the other is the posterior topic distribution of each document. Collecting the context words within the windows of different sizes to form the training documents, we can easily obtain different levels of context for each phrase pair, which are represented in the form of dimensionality reduction and play an important role in the proposed method.

3.2 Phrase table multiplication using pivot document-level topics as hidden variables

In the first method, we deal mainly with the effect of document-level context, and assume that the bridged phrase pairs should be similar in the document-level topic distributions. Similar to the conventional method, we build a source-target translation model in a way of phrase table multiplication. However, instead of inducing a source-target phrase pair by only a pivot phrase, the proposed method uses the pivot phrase with the document-level topic as a bridge.

Specifically, we use the pivot documents of source-pivot and pivot-target bilingual corpora to train a topic model, where the document-topic distribution can be used to represent the document-level context of phrase pairs on the pivot side. Then, we introduce the pivot topic as a hidden variable, and decompose the source-to-target phrase translation probability $\phi(\tilde{e}|\tilde{f})$ as follows:

$$\begin{aligned}\phi(\tilde{e}|\tilde{f}) &= \sum_{\tilde{p}}\sum_{t_p}\phi(\tilde{e},\tilde{p},t_p|\tilde{f})\\ &= \sum_{\tilde{p}}\sum_{t_p}\phi(\tilde{e}|\tilde{p},t_p,\tilde{f})\cdot\phi(\tilde{p},t_p|\tilde{f})\\ &= \sum_{\tilde{p}}\sum_{t_p}\phi(\tilde{e}|\tilde{p},t_p)\cdot\phi(\tilde{p},t_p|\tilde{f}),\end{aligned} \tag{6}$$

where $\phi(\tilde{e}|\tilde{p},t_p)$ is the probability of translating $\tilde{p}$ into $\tilde{e}$ under the pivot topic t_p, and $\phi(\tilde{p},t_p|\tilde{f})$ is the probability of translating $\tilde{f}$ into $\tilde{p}$ with topic t_p.

Following Su *et al.* (2012) and Xiao *et al.* (2012), we assume that in one document, all the phrases have the same topic distribution as the one of the document they belong to. Thus, we can use MLE to solve $\phi(\tilde{e}|\tilde{p},t_p)$:

$$\phi(\tilde{e}|\tilde{p},t_p)=\frac{\sum\limits_{d\in C_{pe}}\mathrm{count}_d(\tilde{p},\tilde{e})\cdot p(t_p|d)}{\sum\limits_{\tilde{e}'}\sum\limits_{d\in C_{pe}}\mathrm{count}_d(\tilde{p},\tilde{e}')\cdot p(t_p|d)}, \tag{7}$$

where C_{pe} is the pivot-target bilingual corpus, and $\mathrm{count}_d(\tilde{p},\tilde{e})$ denotes the number of phrase pairs $(\tilde{p},\tilde{e})$ in document d. In a similar way, we compute $\phi(\tilde{p},t_p|\tilde{f})$ as follows:

$$\phi(\tilde{p},t_p|\tilde{f})=\frac{\sum\limits_{d\in C_{fp}}\mathrm{count}_d(\tilde{f},\tilde{p})\cdot p(t_p|d)}{\sum\limits_{\tilde{p}'}\sum\limits_{d\in C_{fp}}\mathrm{count}_d(\tilde{f},\tilde{p}')}, \tag{8}$$

where C_{fp} is the source-pivot bilingual corpus.

Finally, with the induced word alignments and phrase translation probabilities, we also resort to the conventional method shown in Section 2 to calculate the lexical weights of the induced phrase pairs.

3.3 Translation probability embedded with the topic-based sense similarity

Different from the method mentioned above, the proposed second method focuses on the effect of local context. Assuming that the meaning of each phrase pair is reflected by its pivot-side context words, we adjust the induced phrase translation probabilities with the sense similarity based on their context words.

In this respect, the conventional methods mostly adopt the vector-based model to compute the meaning similarity because it is unsupervised and easy to implement. The main drawback of a vector-based

model, however, is that it is unable to overcome the context sparse problem, which is especially serious for the phrases with low frequency. To solve this problem, we first assume that the pivot words found in source-pivot and pivot-target corpora share a global set of latent senses $Z=\{z_n|n=1,2,\cdots,N\}$ (to avoid confusion with the previous document-level topic, we use different notations to denote the latent sense). Under this representation, the meaning of pivot word $\tilde{p}_k$ is simplified into a distribution over latent senses:

$$\text{sense}(\tilde{p}_k)=(p(z_1|\tilde{p}_k),p(z_2|\tilde{p}_k),\cdots,p(z_N|\tilde{p}_k)), \quad (9)$$

and the meaning of pivot phrase $\tilde{p}$ can be represented as the average distributions of the non-stop words it contains:

$$\text{sense}(\tilde{p})=(p(z_1|\tilde{p}),p(z_2|\tilde{p}),\cdots,p(z_N|\tilde{p})), \quad (10)$$

$$p(z_i|\tilde{p})=\sum_{k=1}^{|\tilde{p}|}\frac{p(z_i|\tilde{p}_k)}{|\tilde{p}|}, \quad (11)$$

where $|\tilde{p}|$ denotes the word number of $\tilde{p}$.

This representation is essentially a means of reducing the dimensionality of the original vector-based one. Thus, the key to using this representation is inducing the latent senses. Following Dinu and Lapata (2010), we also apply LDA to induce the latent senses. We collect the pivot-side context words within a symmetric window of a fixed size, forming a document for each pivot word. Using these documents as training data, we adopt the Gibbs sampling inference (Griffiths and Steyvers, 2004) to train the LDA model. Note that this model is different from the previous document-level topic models. After model training, the set of latent senses is represented in the form of probabilistic topics in LDA. Formally, we can easily obtain two parameters: θ gives the sense distribution of each pivot word, and ϕ embodies each sense as the generation probabilities of context words. These two parameters can be used to infer the posterior distribution of unseen pivot words in the latent sense space.

Our goal is to measure the sense similarity between phrase pairs based on their pivot-side context words. Therefore, for the pivot word $\tilde{p}_w$ belonging to phrase pairs $(\tilde{f},\tilde{p})$ and $(\tilde{p},\tilde{e})$, we also gather its pivot-side context words within a fixed width window to form a document. However, unlike the model training, only when $\tilde{f}$ co-occurs with $\tilde{p}$, will the context words be collected from the source-pivot corpus. Using the above topic model, we infer this document and then represent the meaning of $\tilde{p}_w$ with the resulting posterior topic distribution. After solving the sense distributions of all pivot non-stop words $\tilde{p}$ contains, we obtain the pivot-side sense distribution of $(\tilde{f},\tilde{p})$ by Eqs. (10) and (11). To avoid confusion, here we use $\text{sense}_{\tilde{f}}(\tilde{p})$ to represent this distribution.

Similarly, we can easily obtain the pivot-side sense distribution of $(\tilde{p},\tilde{e})$, which we denote with $\text{sense}_{\tilde{e}}(\tilde{p})$. Following Wu and Wang (2007), we use cosine distance to measure the sense similarity between $(\tilde{f},\tilde{p})$ and $(\tilde{p},\tilde{e})$:

$$\text{sim}(\tilde{f},\tilde{e};\tilde{p})=\cos(\text{sense}_{\tilde{f}}(\tilde{p}),\text{sense}_{\tilde{e}}(\tilde{p})). \quad (12)$$

And we embed the translation probability $\phi(\tilde{e}|\tilde{f})$ with the above sense similarity:

$$\phi(\tilde{e}|\tilde{f})=\frac{\sum_{\tilde{p}}\phi(\tilde{e}|\tilde{p})\cdot\phi(\tilde{p}|\tilde{f})\cdot\text{sim}(\tilde{f},\tilde{e};\tilde{p})}{\sum_{\tilde{e}'}\sum_{\tilde{p}}\phi(\tilde{e}'|\tilde{p})\cdot\phi(\tilde{p}|\tilde{f})\cdot\text{sim}(\tilde{f},\tilde{e}';\tilde{p})}. \quad (13)$$

Finally, we follow the conventional method shown in Section 2 to compute the lexical weights of the induced phrase pairs.

3.4 Combination of different methods

In general, the methods mentioned above emphasize the context information of different levels. So, there is an interesting question about whether the pivot-based SMT can be further improved if we use these methods simultaneously.

To answer the above question, we build an interpolated model by performing linear interpolation on different pivot models. Specifically, in the final model, the phrase translation probability $\phi(\tilde{e}|\tilde{f})$ and lexical weight $p_w(\tilde{e}|\tilde{f})$ are estimated as

$$\phi(\tilde{e}|\tilde{f})=\sum_i\alpha_i\cdot\phi_i(\tilde{e}|\tilde{f}),\ \sum_i\alpha_i=1, \quad (14)$$

$$p_w(\tilde{e}|\tilde{f})=\sum_i\beta_i\cdot p_{w,i}(\tilde{e}|\tilde{f}),\ \sum_i\beta_i=1, \quad (15)$$

where $\phi_i(\tilde{e}|\tilde{f})$ and $p_{w,i}(\tilde{e}|\tilde{f})$ denote the phrase translation probability and lexical weight of the ith pivot model, respectively, and α_i and β_i are the corresponding interpolation coefficients, respectively.

4 Experiment

We evaluate the proposed methods on the French-to-Spanish translation task using English as the pivot language. After a brief description of the experimental setup, we report and discuss the system performance under different conditions.

4.1 Experimental setup

To comprehensively investigate the generality of the proposed methods, we carry out experiments on two data sets. In the experiment with the first data set, the training data comes from the French-English and English-Spanish parts of the Europarl corpus (http://www.statmt.org/europarl). We select the same development and test sets used in the experiments of Wu and Wang (2007). In the experiment with the second data set, the training data is from the OPUS corpus (http://opus.lingfil.uu.se/) and is not limited to a specific domain. The development and test sets are also extracted from the French-Spanish part of the OPUS corpus. Each sentence in the sets is with a single reference. To avoid confusion, we name the above two data sets WMT and OPUS data sets, respectively. Tables 1 and 2 show the statistics of the various data sets.

Table 1 Data sets of the WMT experiment

Data set	n_{d}	n_{s}	n_{sw}	n_{tw}
F2E train	3424	1M	30.2M	27.2M
E2S train	3465	1M	27.3M	28.4M
Dev	–	2000	67 295	60 628
In-Test	–	2000	68 103	61 866
Out-Test	–	1064	32 849	29 864

F2E: French-to-English; E2S: English-Spanish; Dev: development set; In-Test: in-domain test set; Out-Test: out-domain test set. n_{d}: number of documents; n_{s}: number of sentences; n_{sw}: number of source words; n_{tw}: number of target words

As for the training corpora, we use the GIZA++ toolkit (Och and Ney, 2003) with the heuristics 'grow-diag-final-and' to generate two word-aligned corpora, where the bilingual phrases with a maximum length of five are extracted. For the language model in the WMT experiment, we directly use the 3-gram language model provided by the shared task of the NAACL/HLP 2006 Workshop on SMT. In the OPUS experiment, we use SRILM toolkits (Stolcke, 2002) to train one 4-gram language model on the target part of the English-Spanish OPUS corpus (34.6M sentences with 282.8M words). To represent different levels of context using the probability distributions over topics, we adopt the GibbsLDA++ toolkit (http://gibbslda.sourceforge.net/) to train two topic models: one is a document-level topic model, using the pivot documents as training data; the other is a local topic model, collecting the context words within a symmetric window of size 10 to form a training document for each pivot word. In this process, we empirically set the parameters as follows: hyper-parameter $\alpha = 2/\mathrm{topic\_num}$, hyper-parameter $\beta = 0.01$, and the number of Gibbs sampling iterations iters = 500. As for the topic model training and inference, on our server with 128 GB RAM and eight cores of 2.93 GHz CPU, the longest time of our model is about 138 h (the longest time is recorded when training the local context model with the topic number topic_num = 500). Since the training can be done offline, we believe that the training time is not critical to the practical use of our system.

Table 2 Data sets of the OPUS experiment

Data set	Genre	n_{d}	n_{s}	n_{sw}	n_{tw}
F2E train	ECB	953	135.2k	4.5M	4.0M
	KDE4	853	68.8k	1.4M	1.2M
	Subtitle	654	201.4k	1.7M	1.9M
	JRC	5649	200.0k	6.9M	6.4M
	WMT	707	200.7k	6.1M	5.5M
E2S train	ECB	842	77.0k	2.2M	2.5M
	KDE4	892	75.7k	1.4M	1.5M
	Subtitle	636	201.3k	2.1M	1.7M
	JRC	5725	200.0k	6.2M	7.0M
	WMT	718	201.1k	5.3M	5.7M
Dev	Mixed	–	2000	63 145	60 739
Test	Mixed	–	2000	63 736	61 651

ECB: European Central Bank corpus; KDE4: KDE location files; Subtitle: the corpus from opensubtitles.org; JRC: JRC-Acquis corpus; WMT: the corpus provided by the workshop on SMT; Mixed: mixed-domain data set. n_{d}: number of documents; n_{s}: number of sentences; n_{sw}: number of source words; n_{tw}: number of target words

In our experiments, we use MOSES (http://www.statmt.org/moses/), a famous open-source machine translation system, as the experimental decoder. During decoding, we set the table-limit as 50, the stack-size as 100, and perform minimum-error-rate training (MERT) (Och, 2003) to tune the feature weights of the log-linear model. For the translation quality, we use case-insensitive BLEU-4 (Papineni *et al.*, 2002) and METEOR (Denkowski and Lavie, 2011) metrics to evaluate the translation results, and finally conduct paired bootstrap sampling

(Koehn, 2004) to test the significance in score differences. To alleviate the impact of the instability of MERT, we run it three times for each experiment and present the average BLEU and METEOR scores on the three runs following the suggestion by Clark *et al.* (2011).

4.2 Results and analyses

For clarity, in the following sections we name the proposed two topic-aware pivot language approaches (Sections 3.2 and 3.3) document-context and local-context methods, respectively.

4.2.1 Effect of different methods

In the first group of experiments, we investigate mainly the effectiveness of different topic-aware pivot language approaches.

1. Comparative methods: In addition to the conventional triangulation method, which was proposed by Wu and Wang (2007) and shown in Section 2, we compare our approach with two other methods widely applied in pivot-based SMT:

(1) The transfer method (de Gispert and Mariño, 2006; Utiyama and Isahara, 2007; Khalilov *et al.*, 2008) is implemented at the sentence level. This method first translates the source sentence to the pivot sentence, and then to the target sentence.

(2) The synthetic method (Bertoldi *et al.*, 2008; Schwenk, 2008; Huck and Ney, 2012) obtains an additional source-target corpus by translating the pivot sentences in source-pivot or pivot-target corpora into the target or source ones.

Finally, we use an interpolation method to see if the effects caused by different levels of context are complementary. Here we also compare our results with those of the log-linear combination method which considers the translation probabilities of different pivot models as independent features of the log-linear model for SMT.

2. Parameters: The topic number is an important parameter in the proposed methods. In the case of the document-context method, we try different topic numbers to train topic models: from 20 to 100 with an increment of 10 each time. In the case of the local-context method, we experiment with different topic numbers from 100 to 500 with increment size 100. In this process, we determine the optional topic numbers by maximizing the BLEU scores of the development sets. For the coefficients α_i and β_j (Eqs. (14) and (15)) in the interpolation method, we tune these weights on the two development sets: $\alpha_0 = 0.9, \alpha_1 = 0.1, \beta_0 = 0.9, \beta_1 = 0.1$ for the WMT experiment and $\alpha_0 = 0.8, \alpha_1 = 0.2, \beta_0 = 0.7, \beta_1 = 0.3$ for the OPUS experiment. Because the previous experimental results showed that the utilization of document-level context is more effective than local context, we assign greater values to the weights of the model with document-level context, which are denoted by α_0 and β_0. Specifically, we try different α_0 and β_0 respectively from 0.5 to 0.9 with an increment of 0.1 each time. We find that the final optional weights produce slightly better performances than other values on two data sets.

Table 3 reports the experimental results in the WMT data set. On the in-domain test set, the BLEU and METEOR scores of the baseline are 33.72 and 51.23, respectively. In system performance, the transfer method is slightly inferior to the baseline, while the synthetic one is similar to it. By utilizing topic-based context, both of the proposed two methods improve the system performance to different extensions. In spite of slight improvements on the development set (in the WMT experiment, the BLEU scores of the development set using triangulation, document-context, local-context, and log-linear methods are 44.15, 45.44, 44.78, and 45.62, respectively), the log-linear method fails to achieve the same improvement, or even degrades performance on the test sets. In contrast, the interpolation method outperforms the baseline and individual pivot language approaches. Specifically, the BLEU and METEOR scores of the system are 34.58 and 52.05, 0.86 and 0.82 points higher than the baseline, respectively. These differences are statistically significant at $P < 0.01$ using the significance test tool developed by Zhang *et al.* (2004). For this result, we speculate that the combination method further reduces the pivot phrase disambiguation by exploiting different levels of context. However, the log-linear method overfits the development set with four extra features; thus, it does not obtain the same effect on the test sets.

The above results verify that the pivot-side context is helpful for pivot-based SMT on the in-domain test set. However, regardless of the method used, we do not obtain stable improvement on the out-domain test set. This may be because the probabilis-

Table 3 Experimental results in the WMT data set

Method	Number of topics	BLEU		METEOR	
		WMT In-Test	WMT Out-Test	WMT In-Test	WMT Out-Test
Triangulation(baseline)	–	33.72	18.73	51.23	40.68
Transfer	–	33.36	18.05	50.77	40.13
Synthetic	–	34.00	18.63	51.00	40.93
Document-context	40	34.50	18.48	51.82	40.49
Local-context	400	34.26	18.92	51.55	41.00
Log-linear	–	34.47	18.80	51.60	40.71
Interpolation	–	**34.58**	19.07	**52.05**	40.92

tic distribution of pivot-side context is not identical to the out-domain data, thus leading to no positive effect on the translation system. Therefore, we will consider only the in-domain test set in the following WMT experiments.

Table 4 shows the experimental results in the OPUS data set. These results are similar to those in the previous experiment. The BLEU and METEOR scores of the baseline are 38.67 and 53.03, respectively. The transfer method is slightly inferior to the baseline; by contrast, the synthetic method performs slightly better than the baseline. Then, using the proposed two methods to capture different levels of context, we bring different levels of improvements to the system performance. When we use the interpolation method, the system achieves the best performance. The BLEU and Meteor scores under this condition are 39.77 and 54.05, achieving 1.1 and 1.02 improvements over the baseline respectively, both of which are significant at $P < 0.01$ by using paired bootstrap sampling.

Table 4 Experimental results in the OPUS data set

Method	Number of topics	BLEU	METEOR
Triangulation(baseline)	–	38.67	53.03
Transfer	–	38.19	52.55
Synthetic	–	39.00	53.38
Document-context	80	39.58	53.82
Local-context	400	39.30	53.44
Log-linear	–	39.61	53.60
Interpolation	–	**39.77**	**54.05**

4.2.2 Effect of different representations on the local-context method

As described in Section 3.3, we adopt the topic-based representation rather than the VSM to compute the sense similarity, so we compare these representations and study their effects on the local-context method. For the implementation of VSM, we extract the most frequent 5000 pivot words as context words, and adopt the approach of Chen *et al.* (2010) to set the feature weights of the VSM.

Table 5 gives the experimental results of the local-context method using different representations. On two test sets, the local-context method using VSM improves the performance by 0.38 and 0.43 BLEU points, 0.09 and 0.36 Meteor points over the baseline, respectively. Furthermore, if we replace VSM with topic-based representation, the two improvements increase to 0.54 and 0.63 BLEU points, and 0.32 and 0.41 Meteor points, respectively. This echoes the promising results in Dinu and Lapata (2010) and embodies the advantage of the topic-based representation over VSM in sense similarity computation.

Table 5 Experimental results of the local-context method using different representations

Method	WMT In-Test		OPUS test	
	BLEU	METEOR	BLEU	METEOR
Baseline	33.72	51.23	38.67	53.03
VSM	34.10	51.32	39.10	53.39
LDA	34.26	51.55	39.30	53.44

4.2.3 Comparison with translation models trained from the source-target parallel corpus

To obtain a more detailed analysis of the above experimental results, we compare our models with those trained from direct source-target parallel corpora in the following aspects: (1) the performances of the translation system; (2) the distributions of the translation probability.

In this experiment, we use additional source-target training data to directly build source-target

translation models. For the WMT experiment, the training data is from the WMT French-Spanish corpus and contains 1.68M parallel sentences. For the OPUS experiment, we use a part of the OPUS French-Spanish corpus as the training data, which consists of 796k parallel sentences in roughly the same proportion to the French-English training data shown in Table 2.

1. System performance: To conduct the experiments, we extract different numbers of parallel sentences from the additional source-target parallel corpora: from 10k to 70k with an increment of 10k each time. Using each of these corpora, we train a direct translation model.

Table 6 reports the experiment results. Whether on the WMT data set or the OPUS data set, the proposed interpolation approach achieves improvements over the direct translation model trained with 50k sentence pairs. Specifically, on two test sets, the BLEU scores acquired by the direct translation model are 33.89 and 38.90 respectively, and the proposed interpolation approach achieves absolute improvements of 0.69 and 0.87, respectively. The METEOR scores of the direct translation model on two test sets are 51.01 and 53.57, respectively; by contrast, the interpolation method achieves significant improvements of 1.04 and 0.48, respectively. When the direct parallel corpus is increased, direct translation models quickly outperform our model. This demonstrates that the proposed method is suitable for the language pairs with small-scale training data available.

2. Probability distribution: Meanwhile, we investigate the effect of topic-based context on the pivot-based SMT from another perspective. For each source phrase, we compare its distributions in three models: (1) the translation model built from the direct parallel corpus; (2) our pivot-based translation model; (3) the conventional pivot-based translation model by the triangulation method. Here the distribution of a source phrase means its probabilities translated into different target phrases.

Different from the previous experiment, we use an entire source-target parallel corporus to train two direct translation models. Note that because the above models are built from the corpora of different sizes, we focus only on the probabilities of the target phrases occurring in all models. To speed up the computation, here we concern only the candidate source phrases, which are used to translate the development and test sets.

With different distributions of a given source phrase, we compute two Kullback-Leibler distances: one is the distance of distributions between the direct translation model and our pivot-based translation model, and the other is the distance of distributions between the direct translation model and conventional pivot-based translation model. Here, we think the direct source-target translation model can reflect the translation probability distributions of phrases better in reality. Thus, if there are more source phrases with less distance in our pivot-based translation model than the conventional model, we believe the proposed approach can obtain better estimation of the translation probability.

Table 7 shows the results. Compared with the baseline, the proposed topic-aware pivot language approaches enable more source phrases in the translation probability distributions closer to that of the source-target model. These results robustly demonstrate the effectiveness of pivot-side context from another angle. In particular, under the interpolation condition, the maximum number of better source phrases is obtained, and this indicates that different

Table 6 Comparison with the direct translation model

Model	Data size	BLEU		METEOR	
		WMT In-Test	OPUS test	WMT In-Test	OPUS test
Direct model	10k	29.68	32.38	48.78	49.24
	20k	30.96	34.61	49.40	50.59
	30k	31.63	35.92	49.85	51.59
	40k	32.02	37.45	50.26	52.62
	50k	**33.89**	**38.90**	**51.01**	**53.57**
	60k	35.22	40.77	52.33	55.02
	70k	36.87	42.35	53.59	56.22
Interpolation		**34.58**	**39.77**	**52.05**	**54.05**

levels of context are complementary.

Table 7 Number of source phrases whose probability distributions are closer to the ones in reality

vs. Triangulation	Number of source phrases	
	WMT	OPUS
Candidate source phrases	191 527	131 816
Document-context	142 006	96 753
Local-context (LDA)	128 021	79 224
Local-context (VSM)	118 717	74 385
Interpolation	147 923	101 276

4.2.4 Experiments on French-to-German translation

The translation results of the previous experiments indicate the effectiveness of the proposed methods. To investigate the effectiveness of the proposed methods by using independently sourced parallel corpora, here we conduct French-German translation using English as the pivot language. Our training and test data sets are still from the WMT and OPUS corpora. We build two translation models: one for the WMT test set using the WMT French-English and the OPUS English-German parallel sentences; the other for the OPUS test set using the OPUS French-English and the WMT English-German parallel sentences. Tables 8 and 9 show the statistics of data sets used in this section.

In the experiment with the WMT test set, the development set and 3-gram language model we use are also from the shared task of NAACL/HLP 2006 Workshop on SMT. In the experiment with the OPUS test set, the development set is extracted from the French-German part of the OPUS corpus. As for language model training, we use the SRILM toolkits (Stolcke, 2002) to train a 4-gram language model on the target part of the English-German OPUS corpus (17.6M sentences with 102.8M words).

During training, we adopt the same method to set the parameters of topic models. Besides, we tune the interpolated weights on the development sets. To be specific, we set the following interpolated weights: $\alpha_0 = 0.7, \alpha_1 = 0.3, \beta_0 = 0.8, \beta_1 = 0.2$ for the experiment with the WMT test set and $\alpha_0 = 0.7, \alpha_1 = 0.3, \beta_0 = 0.9, \beta_1 = 0.1$ for the experiment with the OPUS test set.

The translation results are shown in Tables 10 and 11, respectively. The final results are quite similar to the previous ones in the French-Spanish experiments. In most cases, the interpolated model significantly outperforms the other models. Overall, the interpolated model obtains 0.71 and 0.9 BLEU points, 0.73 and 1.14 Meteor points better than the baseline model in the two test sets, respectively, and all of these improvements are also significant at $P < 0.01$ by paired bootstrap sampling.

Table 8 Data sets of the WMT experiment

Data set	Genre	n_{d}	n_{s}	n_{sw}	n_{tw}
F2E train	WMT	3424	1M	30.2M	27.2M
E2G train	ECB	1098	96.8k	2.8M	2.57M
	KDE4	1102	74.37k	1.44M	1.37M
	Subtitle	73	48.75k	0.5M	0.45M
	JRC	9708	400k	13.1M	11.7M
	WMT	192	401k	11.3M	10.6M
Dev	WMT	–	2000	67 295	55 147
Test	WMT	–	2000	68 103	55 546

n_{d}: number of documents; n_{s}: number of sentences; n_{sw}: number of source words; n_{tw}: number of target words

Table 9 Data sets of the OPUS experiment

Data set	Genre	n_{d}	n_{s}	n_{sw}	n_{tw}
F2E train	ECB	953	135.2k	4.5M	4M
	KDE4	853	68.8k	1.4M	1.2M
	Subtitle	654	201.4k	1.7M	1.9M
	JRC	5649	200.0k	6.9M	6.4M
	WMT	707	200.7k	6.1M	5.5M
E2G train	WMT	3589	1M	27.6M	26.2M
Dev	Mixed	–	2000	82 775	75 589
Test	Mixed	–	2000	82 305	75 647

n_{d}: number of documents; n_{s}: number of sentences; n_{sw}: number of source words; n_{tw}: number of target words

Table 10 Experimental results in the WMT data set

Method	Number of topics	BLEU	METEOR
Triangulation (baseline)	–	14.11	27.96
Transfer	–	13.80	27.66
Synthetic	–	14.13	28.25
Document-context	60	14.71	28.54
Local-context	300	14.50	28.51
Log-linear	–	14.71	28.66
Interpolation	–	**14.82**	**28.69**

5 Related works

The most common application of the pivot language approach is in the establishment of a translation model. In this respect, the related works can be classified into the following three kinds. The first

Table 11 Experimental results in the OPUS data set

Method	Number of topics	BLEU	METEOR
Triangulation (baseline)	–	12.32	22.67
Transfer	–	12.30	22.38
Synthetic	–	12.67	22.98
Document-context	70	13.10	23.54
Local-context	400	12.88	23.14
Log-linear	–	13.26	23.72
Interpolation	–	**13.22**	**23.81**

is the triangulation method, which builds a source-target translation model in the way of phrase table multiplication (Cohn and Lapata, 2007; Wu and Wang, 2007). The second is named the transfer method (de Gispert and Mariño, 2006; Utiyama and Isahara, 2007; Khalilov *et al.*, 2008), which translates the source sentence to the pivot sentence first, and then to the target one. The third is the synthetic method (Bertoldi *et al.*, 2008; Schwenk, 2008; Huck and Ney, 2012), which creates a source-target corpus by translating the pivot sentence in the source-pivot corpus into the target language with pivot-target translation models. Along this line, much research compared these methods and explored the impacts of various factors on the overall performance of pivot-based SMT (Habash and Hu, 2009; Paul *et al.*, 2009; Wu and Wang, 2009; Costa-Jussà *et al.*, 2011).

Meanwhile, many researchers continued the study of pivot-based SMT from different perspectives. In the research field of word alignment, Borin (2000) first used multilingual corpora to increase alignment coverage. More researchers applied pivot-based technology to optimize the parameters of the statistical word alignment model (Filali and Bilmes, 2005; Wang *et al.*, 2006; Kumar *et al.*, 2007). Besides, Callison-Burch *et al.* (2006) used pivot languages for paraphrase extraction to handle the unseen phrases. Crego *et al.* (2010) presented a framework based on pivot language to conduct lexical adaptation.

Different from the above-mentioned research work, the proposed methods incorporate the pivot-side context into pivot-based SMT based on probabilistic topics. Our work is inspired by the following research work: one is context-based SMT (Zhao and Xing, 2006; 2007; Tam *et al.*, 2007; He *et al.*, 2008; Mauser *et al.*, 2009; Shen *et al.*, 2009; Chen *et al.*, 2010; Gong *et al.*, 2011; Ruiz and Federico, 2011; Su *et al.*, 2012; Xiao *et al.*, 2012), which has shown the effectiveness of different levels of context in SMT; the other is topic-based context similarity (Dinu and Lapata, 2010). The most similar one to ours is probably the context-based approach for pivot translation services (Tanaka *et al.*, 2009). Tanaka *et al.* (2009) proposed context-based coordination to maintain the consistency of word meaning during pivot translation services, while our work extends the conventional pivot-based SMT to a topic-aware one, using different levels of context in different ways.

6 Conclusions and future work

In this study, we have proposed methods to incorporate different levels of context into pivot-based SMT. Experimental results show that the proposed methods significantly outperform the conventional approach. Further improvement is achieved by using an interpolation method to bring different topic-aware pivot language approaches together.

Intuitively, the more training data we use, the better topic model we will obtain. Therefore, we will study the effect of additional monolingual data on the proposed methods in the future. Furthermore, we will explore the applications of the proposed methods in other natural language processing tasks, such as paraphrasing.

References

Bertoldi, N., Federico, M., 2009. Domain adaptation for statistical machine translation with monolingual resources. Proc. 4th Workshop on Statistical Machine Translation, p.182-189. [doi:10.3115/1626431.1626468]

Bertoldi, N., Barbaiani, M., Federico, M., *et al.*, 2008. Phrase-based statistical machine translation with pivot languages. Proc. Int. Workshop on Spoken Language Translation, p.143-149.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, **3**:993-1022.

Borin, L., 2000. You'll take the high road and I'll take the low road: using a third language to improve bilingual word alignment. Proc. 18th Conf. on Computational Linguistics, p.97-103. [doi:10.3115/990820.990835]

Callison-Burch, C., Koehn, P., Osborne, M., 2006. Improved statistical machine translation using paraphrases. Proc. Main Conf. on Human Language Technology Conf. of the North American Chapter of the Association of Computational Linguistics, p.17-24. [doi:10.3115/1220835.1220838]

Chen, B.X., Foster, G., Kuhn, R., 2010. Bilingual sense similarity for statistical machine translation. Proc. 48th

Annual Meeting of the Association for Computational Linguistics, p.834-843.

Clark, J.H., Dyer, C., Lavie, A., *et al.*, 2011. Better hypothesis testing for statistical machine translation: controlling for optimizer instability. Proc. 49th Annual Meeting of the Association for Computational Linguistics, p.176-181.

Cohn, T., Lapata, M., 2007. Machine translation by triangulation: making effective use of multi-parallel corpora. Proc. 45th Annual Meeting of the Association for Computational Linguistics, p.728-735.

Costa-Jussà, M.R., Henríquez, C., Banchs, R.E., 2011. Enhancing scarce-resource language translation through pivot combinations. Proc. 5th Int. Joint Conf. on Natural Language Processing, p.1361-1365.

Crego, J.M., Max, A., Yvon, F., 2010. Local lexical adaptation in machine translation through triangulation: SMT helping SMT. Proc. 23rd Int. Conf. on Computational Linguistics, p.232-240.

de Gispert, A., Mariño, J.B., 2006. Catalan-English statistical machine translation without parallel corpus: bridging through Spanish. Proc. 5th Int. Conf. on Language Resources and Evaluation, p.65-68.

Denkowski, M., Lavie, A., 2011. Meteor 1.3: automatic metric for reliable optimization and evaluation of machine translation systems. Proc. 6th Workshop on Statistical Machine Translation, p.85-91.

Dinu, G., Lapata, M., 2010. Measuring distributional similarity in context. Proc. Conf. on Empirical Methods in Natural Language Processing, p.1162-1172.

Filali, K., Bilmes, J., 2005. Leveraging multiple languages to improve statistical MT word alignments. Proc. IEEE Automatic Speech Recognition and Understanding Workshop, p.92-97.

Gong, Z.X., Zhou, G.D., Li, L.Y., 2011. Improve SMT with source-side "topic-document" distributions. Proc. 13th Machine Translation Summit, p.496-502.

Griffiths, T.L., Steyvers, M., 2004. Finding scientific topics. PNAS, p.90-95.

Habash, N., Hu, J., 2009. Improving Arabic-Chinese statistical machine translation using English as pivot language. Proc. 4th Workshop on Statistical Machine Translation, p.173-181.

He, Z.J., Liu, Q., Lin, S.X., 2008. Improving statistical machine translation using lexicalized rule selection. Proc. 22nd Int. Conf. on Computational Linguistics, p.321-328.

Hildebrand, A.S., Eck, M., Vogel, S., *et al.*, 2005. Adaptation of the translation model for statistical machine translation based on information retrieval. EAMT 10th Annual Conf., p.133-142.

Huck, M., Ney, H., 2012. Pivot lightly-supervised training for statistical machine translation. Proc. 10th Conf. of the Association for Machine Translation in the Americas, p.50-57.

Khalilov, M., Costa-Jussà, M.R., Henríquez, C.A., *et al.*, 2008. The TALP&I2R SMT sytstems for IWSLT 2008. Proc. Int. Workshop on Spoken Language Translation, p.116-123.

Koehn, P., 2004. Statistical significance tests for machine translation evaluation. Proc. Conf. on Empirical Methods in Natural Language Processing, p.388-395.

Koehn, P., Och, F.J., Marcu, D., 2003. Statistical phrase-based translation. Proc. Conf. of the North American Chapter of the Association for Computational Linguistics, p.48-54. [doi:10.3115/1073445.1073462]

Kumar, S., Och, F.J., Macherey, W., 2007. Improving word alignment with bridge languages. Proc. Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, p.42-50.

Mauser, A., Hasan, S., Ney, H., 2009. Extending statistical machine translation with discriminative and trigger-based lexicon models. Proc. Conf. on Empirical Methods in Natural Language Processing, p.210-218.

Och, F.J., 2003. Minimum error rate training in statistical machine translation. Proc. 41st Annual Meeting on Association for Computational Linguistics, p.160-167. [doi:10.3115/1075096.1075117]

Och, F.J., Ney, H., 2003. A systematic comparison of various statistical alignment models. *Comput. Linguist.*, **29**(1):19-51. [doi:10.1162/089120103321337421]

Papineni, K., Roukos, S., Ward, T., *et al.*, 2002. BLEU: a method for automatic evaluation of machine translation. Proc. 40th Annual Meeting on Association for Computational Linguistics, p.311-318. [doi:10.3115/1073083.1073135]

Paul, M., Yamamoto, H., Sumita, E., *et al.*, 2009. On the importance of pivot language selection for statistical machine translation. Proc. Annual Conf. of the North American Chapter of the Association for Computational Linguistics, p.221-224.

Ruiz, N., Federico, M., 2011. Topic adaptation for lecture translation through bilingual latent semantic models. Proc. 6th Workshop on Statistical Machine Translation, p.294-302.

Schwenk, H., 2008. Investigations on large-scale lightly-supervised training for statistical machine translation. Proc. Int. Workshop on Spoken Language Translation, p.182-189.

Shen, L.B., Xu, J.X., Zhang, B., *et al.*, 2009. Effective use of linguistic and contextual information for statistical machine translation. Proc. Conf. on Empirical Methods in Natural Language Processing, p.72-80.

Stolcke, A., 2002. SRILM - an extensible language modeling toolkit. Proc. 7th Int. Conf. on Spoken Language Processing, p.901-904.

Su, J.S., Wu, H., Wang, H.F., *et al.*, 2012. Translation model adaptation for statistical machine translation with monolingual topic information. Proc. 50th Annual Meeting of the Association for Computational Linguistics, p.459-468.

Tam, Y.C., Lane, I., Schultz, T., 2007. Bilingual LSA-based adaptation for statistical machine translation. *Mach. Transl.*, **21**(4):187-207. [doi:10.1007/s10590-008-9045-2]

Tanaka, R., Murakami, Y., Ishida, T., 2009. Context-based approach for pivot translation services. Proc. 21st Int. Joint Conf. on Artificial Intelligence, p.1555-1561.

Ueffing, N., Haffari, G., Sarkar, A., 2007. Semi-supervised model adaptation for statistical machine translation. *Mach. Transl.*, **21**(2):77-94. [doi:10.1007/s10590-008-9036-3]

Utiyama, M., Isahara, H., 2007. A comparison of pivot methods for phrase-based statistical machine translation. Proc. Annual Conf. of the North American Chapter of the Association for Computational Linguistics, p.484-491.

Wang, H.F., Wu, H., Liu, Z.Y., 2006. Word alignment for languages with scarce resources using bilingual corpora of other language pairs. Proc. 21st Int. Conf. on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, p.874-881.

Wu, H., Wang, H.F., 2007. Pivot language approach for phrase-based statistical machine translation. *Mach. Transl.*, **21**(3):165-181. [doi:10.1007/s10590-008-9041-6]

Wu, H., Wang, H.F., 2009. Revisiting pivot language approach for machine translation. Proc. Joint Conf. of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th Int. Joint Conf. on Natural Language Processing, p.154-162.

Xiao, X.Y., Xiong, D.Y., Zhang, M., *et al.*, 2012. A topic similarity model for hierarchical phrase-based translation. Proc. 50th Annual Meeting of the Association for Computational Linguistics, p.750-758.

Zhang, Y., Vogel, S., Waibel, A., 2004. Interpreting BLEU/NIST scores: how much improvement do we need to have a better system? Proc. 4th Int. Conf. on Language Resources and Evaluation, p.2051-2054.

Zhao, B., Xing, E.P., 2006. BiTAM: bilingual topic AdMixture models for word alignment. Proc. 21st Int. Conf. on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, p.969-976.

Zhao, B., Xing, E.P., 2007. HM-BiTAM: bilingual topic exploration, word alignment, and translation. Proc. Advances in Neural Information Processing Systems, p.1689-1696.