Improving Pivot-Based Statistical Machine Translation by Pivoting the Co-occurrence Count of Phrase Pairs

Xiaoning Zhu[1*], Zhongjun He[2], Hua Wu[2], Conghui Zhu[1], Haifeng Wang[2], and Tiejun Zhao[1]
Harbin Institute of Technology, Harbin, China[1]
{xnzhu,chzhu,tjzhao}@mtlab.hit.edu.cn
Baidu Inc., Beijing, China[2]
{hezhongjun,wu_hua,wanghaifeng}@baidu.com

Abstract

To overcome the scarceness of bilingual corpora for some language pairs in machine translation, pivot-based SMT uses pivot language as a "bridge" to generate source-target translation from source-pivot and pivot-target translation. One of the key issues is to estimate the probabilities for the generated phrase pairs. In this paper, we present a novel approach to calculate the translation probability by pivoting the co-occurrence count of source-pivot and pivot-target phrase pairs. Experimental results on Europarl data and web data show that our method leads to significant improvements over the baseline systems.

1 Introduction

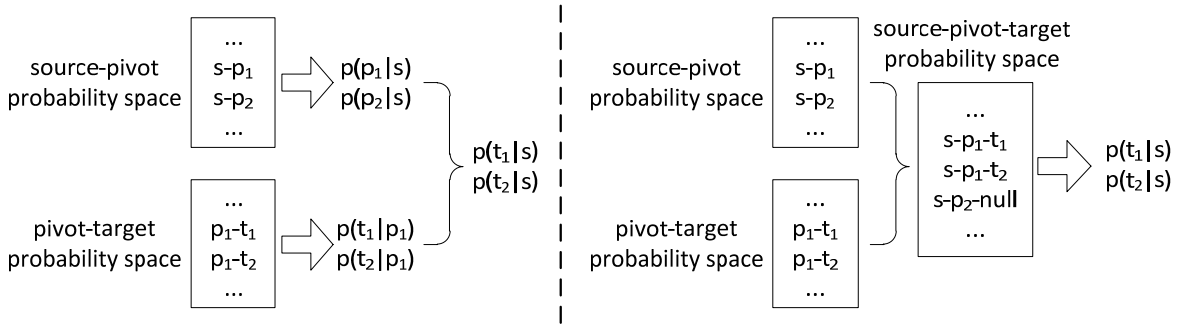
Statistical Machine Translation (SMT) relies on large bilingual parallel data to produce high quality translation results. Unfortunately, for some language pairs, large bilingual corpora are not readily available. To alleviate the parallel data scarceness, a conventional solution is to introduce a “bridge” language (named pivot language) to connect the source and target language (de Gispert and Marino, 2006; Utiyama and Isahara, 2007; Wu and Wang, 2007; Bertoldi et al., 2008; Paul et al., 2011; El Kholy et al., 2013; Zahabi et al., 2013), where there are large amounts of source-pivot and pivot-target parallel corpora.

Among various pivot-based approaches, the triangulation method (Cohn and Lapata, 2007; Wu and Wang, 2007) is a representative work in pivot-based machine translation. The approach proposes to build a source-target phrase table by merging the source-pivot and pivot-target phrase table. One of the key issues in this method is to estimate the translation probabilities for the generated source-target phrase pairs. Conventionally, the probabilities are estimated by multiplying the posterior probabilities of source-pivot and pivot-target phrase pairs. However, it has been shown that the generated probabilities are not accurate enough (Cui et al., 2013). One possible reason may lie in the non-uniformity of the probability space. Through Figure 1. (a), we can see that the probability distributions of source-pivot and pivot-target language are calculated separately, and the source-target probability distributions are induced from the source-pivot and pivot-target probability distributions. Because of the absence of the pivot language (e.g., p2 is in source-pivot probability space but not in pivot-target one), the induced source-target probability distribution is not complete, which will result in inaccurate probabilities.

To solve this problem, we propose a novel approach that utilizes the co-occurrence count of source-target phrase pairs to estimate phrase translation probabilities more precisely. Different from the triangulation method, which merges the source-pivot and pivot-target phrase pairs after training the translation model, we propose to merge the source-pivot and pivot-target phrase pairs immediately after the phrase extraction step, and estimate the co-occurrence count of the source-pivot-target phrase pairs. Finally, we compute the translation probabilities according to the estimated co-occurrence counts, using the standard training method in phrase-based SMT (Koehn et al., 2003). As Figure 1. (b) shows, the

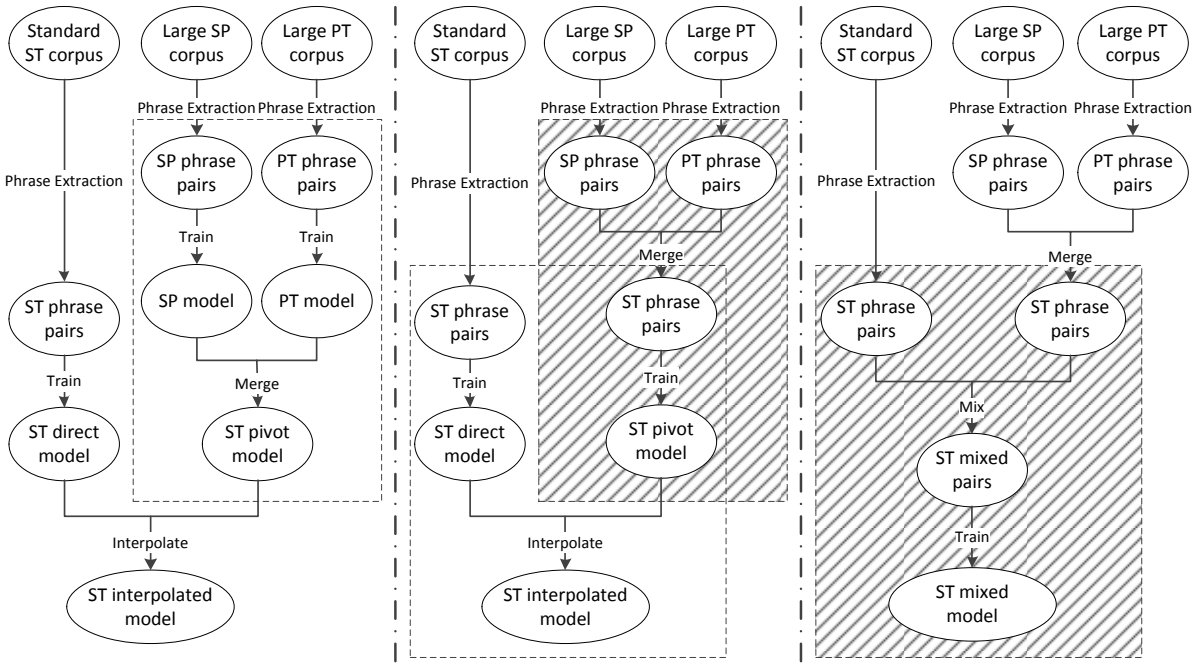
[*] This work was done when the first author was visiting Baidu.

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1665–1675,
October 25-29, 2014, Doha, Qatar.



(a) the *triangulation* method (b) the *co-occurrence count* method

Figure 1: An example of probability space evolution in pivot translation.



(a) the *triangulation* method (b) the *co-occurrence count* method (c) the *mixed* model

Figure 2: Framework of the triangulation method, the co-occurrence count method and the mixed model. The shaded box in (b) denotes difference between the co-occurrence count method and the triangulation method. The shaded box in (c) denotes the difference between the interpolation model and the mixed model.

source-target probability distributions are calculated in a complete probability space. Thus, it will be more accurate than the traditional triangulation method. Figure 2. (a) and (b) show the difference between the triangulation method and our co-occurrence count method.

Furthermore, it is common that a small standard bilingual corpus can be available between the source and target language. The direct translation model trained with the standard bilingual corpus exceeds in translation performance, but its weakness lies in low phrase coverage. However, the pivot model has characteristics characters. Thus, it is important to combine the direct and pivot translation model to compensate mutually and further improve the translation performance. To deal with this problem, we propose a mixed model by merging the phrase pairs extracted by pivot-based method and the phrase pairs extracted from the standard bilingual corpus. Note that, this is different from the conventional interpolation method, which interpolates the direct and pivot translation model. See Figure 2. (b) and (c) for further illustration.

The remainder of this paper is organized as follows. In Section 2, we describe the related work. We introduce the co-occurrence count method in Section 3, and the mixed model in Section 4. In Section 5 and Section 6, we describe and analyze the experiments. Section 7 gives a conclusion of the paper.

2 Related Work

Several methods have been proposed for pivot-based translation. Typically, they can be classified into 3 kinds as follows:

Transfer Method: The transfer method (Utiyama and Isahara, 2007; Wang et al., 2008; Costa-jussà et al., 2011) connects two translation systems: a source-pivot MT system and a pivot-target MT system. Given a source sentence, (1) the source-pivot MT system translates it into the pivot language, (2) and the pivot-target MT system translates the pivot sentence into the target sentence. During each step (source to pivot and pivot to target), multiple translation outputs will be generated, thus a minimum Bayes-risk system combination method is often used to select the optimal sentence (González-Rubio et al., 2011; Duh et al., 2011). The problem with the transfer method is that it needs to decode twice. On one hand, the time cost is doubled; on the other hand, the translation error of the source-pivot translation system will be transferred to the pivot-target translation.

Synthetic Method: It aims to create a synthetic source-target corpus by: (1) translate the pivot part in source-pivot corpus into target language with a pivot-target model; (2) translate the pivot part in pivot-target corpus into source language with a pivot-source model; (3) combine the source sentences with translated target sentences or/and combine the target sentences with translated source sentences (Utiyama et al., 2008; Wu and Wang, 2009). However, it is difficult to build a high quality translation system with a corpus created by a machine translation system.

Triangulation Method: The triangulation method obtains source-target phrase table by merging source-pivot and pivot-target phrase table entries with identical pivot language phrases and multiplying corresponding posterior probabilities (Wu and Wang, 2007; Cohn and Lapata, 2007), which has been shown to work better than the other pivot approaches (Utiyama and Isahara, 2007). A problem of this approach is that the probability space of the source-target phrase pairs is non-uniformity due to the mismatching of the pivot phrase.

3 Our Approach

In this section, we will introduce our method for learning a source-target phrase translation model with a pivot language as a bridge. We extract the co-occurrence count of phrase pairs for each language pair with a source-pivot and a pivot-target corpus. Then we generate the source-target phrase pairs with induced co-occurrence information. Finally, we compute translation probabilities using the standard phrase-based SMT training method.

3.1 Phrase Translation Probabilities

Following the standard phrase extraction method (Koehn et al., 2003), we can extract phrase pairs $(\bar{s}, \bar{p})$ and $(\bar{p}, \bar{t})$ from the corresponding word-aligned source-pivot and pivot-target training corpus, where $\bar{s}$, $\bar{p}$ and $\bar{t}$ denotes the phrase in source, pivot and target language respectively.

Formally, given the co-occurrence count $c(\bar{s}, \bar{p})$ and $c(\bar{p}, \bar{t})$, we can estimate $c(\bar{s}, \bar{t})$ by Equation 1:

$$c(\bar{s}, \bar{t}) = \sum_{\bar{p}} g(c(\bar{s}, \bar{p}), c(\bar{p}, \bar{t})) \quad (1)$$

where $g(\cdot)$ is a function to merge the co-occurrences count $c(\bar{s}, \bar{p})$ and $c(\bar{p}, \bar{t})$. We propose four calculation methods for function $g(\cdot)$.

Given the co-occurrence count $c(\bar{s}, \bar{p})$ and $c(\bar{p}, \bar{t})$, we first need to induce the co-occurrence count $c(\bar{s}, \bar{p}, \bar{t})$. The $c(\bar{s}, \bar{p}, \bar{t})$ is counted when the source phrase, pivot phrase and target phrase occurred together, thus we can infer that $c(\bar{s}, \bar{p}, \bar{t})$ is smaller than $c(\bar{s}, \bar{p})$ and $c(\bar{p}, \bar{t})$. In this circumstance, we consider that $c(\bar{s}, \bar{p}, \bar{t})$ is approximately equal to the minimum value of $c(\bar{s}, \bar{p})$ and $c(\bar{p}, \bar{t})$, as shown in Equation 2.

$$c(\bar{s}, \bar{p}, \bar{t}) \approx \sum_{\bar{p}} \min(c(\bar{s}, \bar{p}), c(\bar{p}, \bar{t})) \quad (2)$$

Because the co-occurrence count of source-target phrase pairs needs the existence of pivot phrase $\bar{p}$, we intuitively believe that the co-occurrence count $c(\bar{s}, \bar{t})$ is equal to the co-occurrence count $c(\bar{s}, \bar{p}, \bar{t})$. Under this assumption, we can obtain the co-occurrence count $c(\bar{s}, \bar{t})$ as shown in Equation 3. Furthermore, to testify our assumption, we also try the *maximum* value (Equation 4) to infer the co-occurrence count of $(\bar{s}, \bar{t})$ phrase pair.

$$c(\bar{s},\bar{t}) = \sum_{\bar{p}} \min(c(\bar{s},\bar{p}), c(\bar{p},\bar{t})) \quad (3)$$

$$c(\bar{s},\bar{t}) = \sum_{\bar{p}} \max(c(\bar{s},\bar{p}), c(\bar{p},\bar{t})) \quad (4)$$

In addition, if source-pivot and pivot-target parallel corpus greatly differ in quantities, then the *minimum* function would likely just take the counts from the smaller corpus. To deal with the problem of the imbalance of the parallel corpora, we also try the *arithmetic mean* (Equation 5) and *geometric mean* (Equation 6) function to infer the co-occurrence count of source-target phrase pairs.

$$c(\bar{s},\bar{t}) = \sum_{\bar{p}} (c(\bar{s},\bar{p}) + c(\bar{p},\bar{t}))/2 \quad (5)$$

$$c(\bar{s},\bar{t}) = \sum_{\bar{p}} \sqrt{c(\bar{s},\bar{p}) \times c(\bar{p},\bar{t})} \quad (6)$$

When the co-occurrence count of source-target language is calculated, we can estimate the phrase translation probabilities with the following Equation 7 and Equation 8.

$$\phi(\bar{s}|\bar{t}) = \frac{c(\bar{s},\bar{t})}{\sum_{\bar{s}} c(\bar{s},\bar{t})} \quad (7)$$

$$\varphi(\bar{t}|\bar{s}) = \frac{c(\bar{s},\bar{t})}{\sum_{\bar{t}} c(\bar{s},\bar{t})} \quad (8)$$

3.2 Lexical Weight

Given a phrase pair $(\bar{s},\bar{t})$ and a word alignment a between the source word positions $i = 1,\cdots,n$ and the target word positions $j = 0,\cdots,m$, the lexical weight of phrase pair $(\bar{s},\bar{t})$ can be calculated by the following Equation 9 (Koehn et al., 2003).

$$p_\omega(\bar{s}|\bar{t},a) = \prod_{i=1}^{n} \frac{1}{|\{j|(i,j)\in a\}|} \sum_{\forall(i,j)\in a} \omega(s_i|t_j) \quad (9)$$

The lexical translation probability distribution $\omega(s|t)$ between source word *s* and target word *t* can be estimated with Equation 10.

$$\omega(s|t) = \frac{c(s,t)}{\sum_{s'} c(s',t)} \quad (10)$$

To compute the lexical weight for a phrase pair $(\bar{s},\bar{t})$ generated by $(\bar{s},\bar{p})$ and $(\bar{p},\bar{t})$, we need the alignment information a, which can be obtained as Equation 11 shows.

$$a = \{(s,t)|\exists p\colon (s,p)\in a_1 \& (p,t)\in a_2\} \quad (11)$$

where a_1 and a_2 indicate the word alignment information in the phrase pair $(\bar{s},\bar{p})$ and $(\bar{p},\bar{t})$ respectively.

4 Integrate with Direct Translation

If a standard source-target bilingual corpus is available, we can train a direct translation model. Thus we can integrate the direct model and the pivot model to obtain further improvements. We propose a mixed model by merging the co-occurrence count in direct translation and pivot translation. Besides, we also employ an interpolated model (Wu and Wang, 2007) by merging the direct translation model and pivot translation model using a linear interpolation.

4.1 Mixed Model

Given n pivot languages, the co-occurrence count can be estimated using the method described in Section 3.1. Then the co-occurrence count and the lexical weight of the mixed model can be estimated with the following Equation 12 and 13.

$$c(s,t) = \sum_{i=0}^{n} c_i(s,t) \quad (12)$$

$$p_\omega(\bar{s}|\bar{t},a) = \sum_{i=0}^{n} \alpha_i\, p_{\omega,i}(\bar{s}|\bar{t},a) \quad (13)$$

where $c_0(s,t)$ and $p_{\omega,0}(\bar{s}|\bar{t},a)$ are the co-occurrence count and lexical weight in the direct translation model respectively. $c_i(s,t)$ and $p_{\omega,i}(\bar{s}|\bar{t},a)$ denote the co-occurrence count and lexical weight in the pivot translation model. α_i is the interpolation coefficient, requiring $\sum_{i=0}^{n} \alpha_i = 1$.

4.2 Interpolated Model

Following Wu and Wang (2007), the interpolated model can be modelled with Equation 14.

$$\phi(\bar{s}|\bar{t}) = \sum_{i=0}^{n} \beta_i \phi_i(\bar{s}|\bar{t}) \quad (14)$$

where $\phi_0(\bar{s}|\bar{t})$ is the phrase translation probability in direct translation model; $\phi_i(\bar{s}|\bar{t})$ is the phrase translation probability in pivot translation model. The lexical weight is obtained with Equation 13. β_i is the interpolation coefficient, requiring $\sum_{i=0}^{n} \beta_i = 1$.

Language Pairs	Sentence Pairs	Source Words	Target Words
de-en	1.9M	48.5M	50.9M
es-en	1.9M	54M	51.7M
fr-en	2M	58.1M	52.4M

Table 1: Training data of Europarl corpus

System	BLEU%					
	de-es	de-fr	es-de	es-fr	fr-de	fr-es
Baseline	27.04	23.01	20.65	33.84	20.87	38.31
Minimum	**27.93***	**23.94***	**21.52***	**35.38***	**21.48***	**39.62***
Maximum	25.70	21.59	20.26	32.58	20.50	37.30
Arithmetic mean	26.01	22.24	20.13	33.38	20.37	37.37
Geometric mean	27.31	**23.49***	**21.10***	**34.76***	**21.15***	**39.19***

Table 2: Comparison of different merging methods on in-domain test set. * indicates the results are significantly better than the baseline ($p<0.05$).

System	BLEU%					
	de-es	de-fr	es-de	es-fr	fr-de	fr-es
Baseline	15.34	13.52	11.47	21.99	12.19	25.00
Minimum	**15.77***	**14.08***	**11.99***	**23.90***	**12.55***	**27.05***
Maximum	13.41	11.83	10.17	20.48	10.83	22.75
Arithmetic mean	13.96	12.10	10.57	21.07	11.30	23.70
Geometric mean	15.09	13.30	11.52	**23.32***	**12.46***	**26.22***

Table 3: Comparison of different merging methods on out-of-domain test set.

5 Experiments on Europarl Corpus

Our first experiment is carried out on Europarl[1] corpus, which is a multi-lingual corpus including 21 European languages (Koehn, 2005). In our work, we perform translations among French (fr), German (de) and Spanish (es). Due to the richness of available language resources, we choose English (en) as the pivot language. Table 1 summarized the statistics of training data. For the language model, the same monolingual data extracted from the Europarl are used.

The word alignment is obtained by GIZA++ (Och and Ney, 2000) and the heuristics "grow-diag-final" refinement rule (Koehn et al., 2003). Our translation system is an in-house phrase-based system analogous to Moses (Koehn et al., 2007). The baseline system is the triangulation method (Wu and Wang, 2007), including an interpolated model which linearly interpolate the direct and pivot translation model.

We use WMT08[2] as our test data, which contains 2000 in-domain sentences and 2051 out-of-domain sentences with single reference. The translation results are evaluated by case-insensitive BLEU-4 metric (Papineni et al., 2002). The statistical significance tests using 95% confidence interval are measured with paired bootstrap resampling (Koehn, 2004).

5.1 Results

We compare 4 merging methods with the baseline system. The results are shown in Table 2 and Table 3. We find that the *minimum* method outperforms the others, achieving significant improvements over the baseline on all translation directions. The absolute improvements range from 0.61 (fr-de) to 1.54 (es-fr) in BLEU% score on in-domain test data, and range from 0.36 (fr-de) to 2.05 (fr-es) in BLEU% score on out-of-domain test data. This indicates that our method is effective and robust in general.

[1] http://www.statmt.org/europarl

[2] http://www.statmt.org/wmt08/shared-task.html

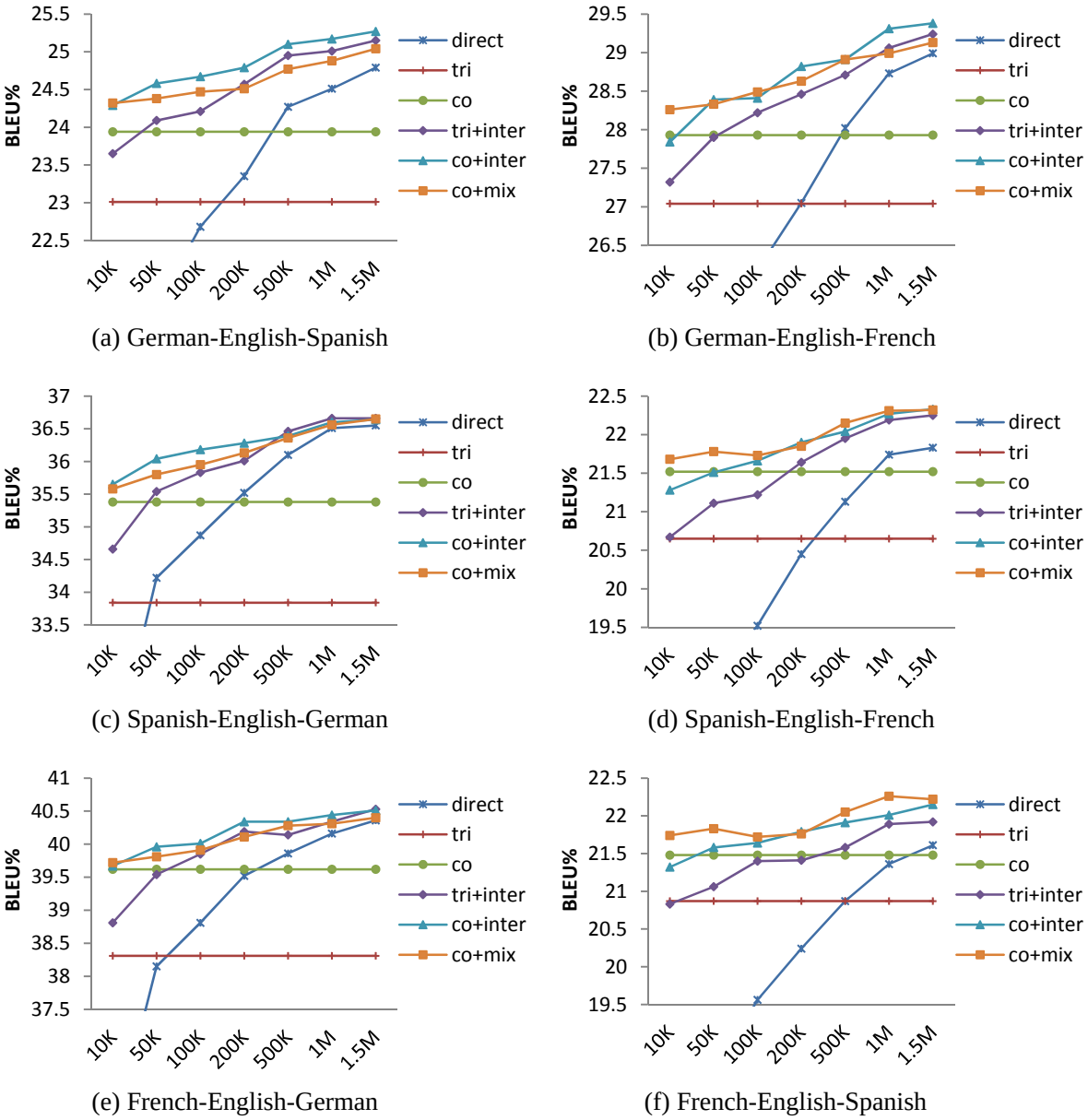


Figure 3: Comparisons of pivot-based methods on different scales of source-target standard corpora. (*direct*: direct model; *tri*: triangulation model; *co*: co-occurrence count model; *tri+inter*: triangulation model interpolated with direct model ; *co+inter*: co-occurrence count model interpolated with direct model; *co+mix*: mixed model). X-axis represents the scale of the standard training data.

The *geometric mean* method also achieves improvement, but not as significant as the *minimum* method. However, the *maximum* and the *arithmetic mean* methods show a decrement in BLEU scores. This reminds us that how to choose a proper merging function for the co-occurrence count is a key problem. In the future, we will explore more sophisticated method to merge co-occurrence count.

5.2 Analysis

The pivot-based translation is suitable for the scenario that there exists large amount of source-pivot and pivot-target bilingual corpora and only a little source-target bilingual data. Thus, we randomly select 10K, 50K, 100K, 200K, 500K, 1M, 1.5M sentence pairs from the source-target bilingual corpora to simulate the lack of source-target data. With these corpora, we train several direct translation models with different scales of bilingual data. We interpolate each direct translation model with the pivot model (both triangulation method and co-occurrence count method) to obtain the interpolated model respectively. We also mix the direct model and pivot model using the method described in Section 4.1. Following

Wu and Wang (2007), we set $\alpha_0 = 0.9$, $\alpha_1 = 0.1$, $\beta_0 = 0.9$ and $\beta_1 = 0.1$ empirically. The experiments are carried out on 6 translation directions: German-Spanish, German-French, Spanish-German, Spanish-French, French-German and French-Spanish. The results are shown in Figure 3. We only list the results on in-domain test sets. The trend of the results on out-of domain test sets is similar with in-domain test sets.

The results are explained as follows:

(1) Comparison of Pivot Translation and Direct Translation

The pivot translation models are better than the direct translation models trained on a small source-target bilingual corpus. With the increment of source-target corpus, the direct model first outperforms the triangulation model and then outperforms the co-occurrence count model consecutively.

Taking Spanish-English-French translation as an example, the co-occurrence count model achieves BLEU% scores of 35.38, which is close to the direct translation model trained with 200K source-target bilingual data. Compared with the co-occurrence count model, the triangulation model only achieves BLEU% scores of 33.84, which is close to the direct translation model trained with 50K source-target bilingual data.

(2) Comparison of Different Interpolated Models

For the pivot model trained by triangulation method and co-occurrence count method, we interpolate them with the direct translation model trained with different scales of bilingual data. Figure 3 shows the translation results of the different interpolated models. For all the translation directions, our co-occurrence count method interpolated with the direct model is better than the triangulation model interpolated with the direct model.

The two interpolated model are all better than the direct translation model. With the increment of the source-target training corpus, the gap becomes smaller. This indicates that the pivot model and its affiliated interpolated model are suitable for language pairs with small bilingual data. Even if the scale of source-pivot and pivot-target corpora is close to the scale of source-target bilingual corpora, the pivot translation model can help the direct translation model to improve the translation performance. Take Spanish-English-French translation as an issue, when the scale of Spanish-French parallel data is 1.5M sentences pairs, which is close to the Spanish-English and English-French parallel data, the performance of *co+mix* model is still outperforms the *direct* translation model.

(3) Comparison of Interpolated Model and Mixed Model

When only a small source-target bilingual corpus is available, the mix model outperforms the interpolated model. With the increasing of source-target corpus, the mix model is close to the interpolated model or worse than the interpolated model. This indicates that the mix model has a better performance when the source-target corpus is small which is close to the realistic scenario.

5.3 Integrate the Co-occurrence Count Model and Triangulation Model

Experimental results in the previous section show that, our co-occurrence count models generally outperform the baseline system. In this section, we carry out experiments that integrates co-occurrence count model into the triangulation model.

For French-English-German translation, we apply a linear interpolation method to integrate the co-occurrence count model into triangulation model following the method described in Section 4.2. We set α as the interpolation coefficient of triangulation model and $1 - \alpha$ as the interpolation coefficient of co-occurrence count model respectively. The experiments take 9 values for interpolation coefficient, from 0.1 to 0.9. The results are shown in Figure 4.

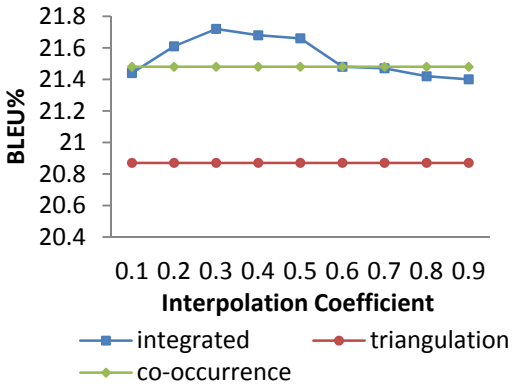


Figure 4: Results of integrating the co-occurrence count model and the triangulation model.

When using interpolation coefficient ranging from 0.2 to 0.7, the integrated models outperform the triangulation and the co-occurrence count model. However, for the other intervals, the inte-

Language Pairs	Sentence Pairs	Source Words	Target Words
zh-en-1	1M	18.1M	17.7M
zh-en-2	2M	36.2M	35.5M
zh-en-3	3M	54.2M	53.2M
zh-en-4	4M	72.3M	70.9M
zh-en-5	5M	90.4M	88.6M
en-jp	1M	9.2M	11.1M

Table 4: Training data of web corpus

System	BLEU%				
	zh-en-jp-1*	zh-en-jp-2	zh-en-jp-3	zh-en-jp-4	zh-en-jp-5
Baseline	29.07	29.39	29.44	29.67	29.80
Minimum	**31.13***	**31.28***	**31.43***	**31.62***	**32.02***
Maximum	28.88	29.01	29.12	29.37	29.59
Arithmetic mean	29.08	29.36	29.51	29.79	30.01
Geometric mean	**30.77***	**31.30***	**31.75***	**32.07***	**32.34***

Table 5: Comparison of different merging methods on the imbalanced web data. (zh-en-jp-1 means the translation system is trained with zh-en-1 as source-pivot corpus and en-jp as pivot-target corpus, and so on.)

grated models perform slightly lower than the co-occurrence count model, but still show better results than the triangulation model. The trend of the curve infers that the integrated model synthesizes the contributions of co-occurrence count model and triangulation model. Additionally, it also indicates that, the choice of the interpolation coefficient affects the translation performances.

6 Experiments on Web Data

The experimental on Europarl is artificial, as the training data for directly translating between source and target language actually exists in the original data sets. Thus, we conducted several experiments on a more realistic scenario: translating Chinese (zh) to Japanese (jp) via English (en) with web crawled data.

As mentioned in Section 3.1, the source-pivot and pivot-target parallel corpora can be imbalanced in quantities. If one parallel corpus was much larger than another, then *minimum* heuristic function would likely just take the counts from the smaller corpus.

In order to analyze this issue, we manually set up imbalanced corpora. For source-pivot parallel corpora, we randomly select 1M, 2M, 3M, 4M and 5M Chinese-English sentence pairs. On the other hand, we randomly select 1M English-Japanese sentence pairs as pivot-target parallel corpora. The training data of Chinese-English and English-Japanese language pairs are summarized in Table 4. For the Chinese-Japanese direct corpus, we randomly select 5K, 10K, 20K, 30K, 40K, 50K, 60K, 70K, 80K, 90K and 100K sentence pairs to simulate the lack of bilingual data. We built a 1K in-house test set with four references. For Japanese language model training, we used the monolingual part of English-Japanese corpus.

Table 5 shows the results of different co-occurrence count merging methods. First, the *minimum* method and the *geometric mean* method outperform the other two merging methods and the baseline system with different training corpus. When the scale of source-pivot and pivot-target corpus is roughly balanced (zh-en-jp-1), the *minimum* method achieves an absolute improvement of 2.06 percentages points on BLEU over the baseline, which is also better than the other merging methods. While, with the growth of source-pivot corpus, the gap between source-pivot corpus and pivot-target corpus becomes bigger. In this circumstance, the *geometric mean* method becomes better than the *minimum* method. Compared to the *minimum* method, the *geometric mean* method considers both the source-pivot and the pivot-target corpus, which may lead to a better result in the case of imbalanced training corpus.

Furthermore, with the imbalanced corpus zh-en-jp-5, we compared the translation performance of our co-occurrence count model (with *geometric mean* merging method), triangulation model, interpolated model, mixed model and the direct translation models. Figure 5 summarized the results.

The co-occurrence count model can achieve an absolute improvement of 2.54 percentages points on BLEU over the baseline. The triangulation method outperforms the direct translation when only 5K sentence pairs are available. Meanwhile, the number is 10K when using the co-occurrence count method. The co-occurrence count models interpolated with the direct model significantly outperform the other models.

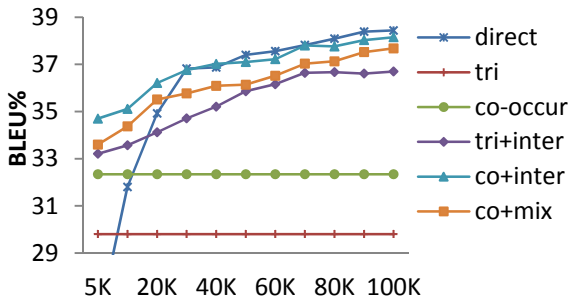


Figure 5: Results on Chinese-Japanese Web Data. X-axis represents the scale of the standard training data.

In this experiment, the training data contains parallel sentences on various domains. And the training corpora (Chinese-English and English-Japanese) are typically very different, since they are obtained on the web. It indicates that our co-occurrence count method is robust in the realistic scenario.

7 Conclusion

This paper proposed a novel approach for pivot-based SMT by pivoting the co-occurrence count of phrase pairs. Different from the triangulation method merging the source-pivot and pivot-target language after training the translation model, our method merges the source-pivot and pivot-target language after extracting the phrase pairs, thus the computing for phrase translation probabilities is under the uniform probability space. The experimental results on Europarl data and web data show significant improvements over the baseline systems. We also proposed a mixed model to combine the direct translation and pivot translation, and the experimental results show that the mixed model has a better performance when the source-target corpus is small which is close to the realistic scenario.

A key problem in the approach is how to learn the co-occurrence count. In this paper, we use the *minimum* function on balanced corpora and the *geometric mean* function on imbalanced corpora to estimate the co-occurrence count intuitively. In the future, we plan to explore more effective approaches.

Acknowledgments

We would like to thank Yiming Cui for insightful discussions, and three anonymous reviewers for many invaluable comments and suggestions to improve our paper. This work is supported by National Natural Science Foundation of China (61100093), and the State Key Development Program for Basic Research of China (973 Program, 2014CB340505).

Reference

Nicola Bertoldi, Madalina Barbaiani, Marcello Federico, and Roldano Cattoni. 2008. Phrase-Based statistical machine translation with Pivot Languages. In *Proceedings of the 5th International Workshop on Spoken Language Translation (IWSLT)*, pages 143-149.

Trevor Cohn and Mirella Lapata. 2007. Machine Translation by Triangulation: Make Effective Use of Multi-Parallel Corpora. In *Proceedings of 45th Annual Meeting of the Association for Computational Linguistics*, pages 828-735.

Marta R. Costa-jussà, Carlos Henríquez, and Rafael E. Banchs. 2011. Enhancing Scarce-Resource Language Translation through Pivot Combinations. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 1361-1365.

Yiming Cui, Conghui Zhu, Xiaoning Zhu, Tiejun Zhao and Dequan Zheng. 2013. Phrase Table Combination Deficiency Analyses in Pivot-based SMT. In *Proceedings of 18th International Conference on Application of Natural Language to Information Systems*, pages 355-358.

Adria de Gispert and Jose B. Marino. 2006. Catalan-English statistical machine translation without parallel corpus: bridging through Spanish. In *Proceedings of 5th International Conference on Language Resources and Evaluation (LREC)*, pages 65-68.

Kevin Duh, Katsuhito Sudoh, Xianchao Wu, Hajime Tsukada and Masaaki Nagata. 2011. Generalized Minimum Bayes Risk System Combination. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 1356-1360.

Ahmed El Kholy, Nizar Habash, Gregor Leusch, Evgeny Matusov and Hassan Sawaf. 2013. Language Independent Connectivity Strength Features for Phrase Pivot Statistical Machine Translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 412-418.

Ahmed El Kholy, Nizar Habash, Gregor Leusch, Evgeny Matusov and Hassan Sawaf. 2013. Selective Combination of Pivot and Direct Statistical Machine Translation Models. In *Proceedings of the 6th International Joint Conference on Natural Language Processing*, pages 1174-1180.

Jesús González-Rubio, Alfons Juan and Francisco Casacuberta. 2011. Minimum Bayes-risk System Combination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 1268-1277.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *HLT-NAACL: Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127-133.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 388-395.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of MT Summit X*, pages 79-86.

Philipp Koehn, Hieu Hoang, Alexanda Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, demonstration session*, pages 177-180.

Philipp Koehn, Alexandra Birch, and Ralf Steinberger. 2009. 462 Machine Translation Systems for Europe. In *Proceedings of the MT Summit XII*.

Gregor Leusch, Aurélien Max, Josep Maria Crego and Hermann Ney. 2010. Multi-Pivot Translation by System Combination. In *Proceedings of the 7th International Workshop on Spoken Language Translation*, pages 299-306.

Franz Josef Och and Hermann Ney. 2000. A comparison of alignment models for statistical machine translation. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 1086-1090.

Michael Paul, Andrew Finch, Paul R. Dixon and Eiichiro Sumita. 2011. Dialect Translation: Integrating Bayesian Co-segmentation Models with Pivot-based SMT. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1-9.

Michael Paul and Eiichiro Sumita. 2011. Translation Quality Indicators for Pivot-based Statistical MT. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 811-818.

Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computation Linguistics*, pages 311-319.

Rie Tanaka, Yohei Murakami and Toru Ishida. 2009. Context-Based Approach for Pivot Translation Services. In the Twenty-first International Conference on Artificial Intelligence, pages 1555-1561.

Jörg Tiedemann. 2012. Character-Based Pivot Translation for Under-Resourced Languages and Domains. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 141-151.

Masatoshi Tsuchiya, Ayu Purwarianti, Toshiyuki Wakita and Seiichi Nakagawa. 2007. Expanding Indonesian-Japanese Small Translation Dictionary Using a Pivot Language. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 197-200.

Takashi Tsunakawa, Naoaki Okazaki and Jun'ichi Tsujii. 2010. Building a Bilingual Lexicon Using Phrase-based Statistical Machine Translation via a Pivot Language. In

Proceedings of the 22th International Conference on Computational Linguistics (Coling), pages 127-130.

Masao Utiyama and Hitoshi Isahara. 2007. A Comparison of Pivot Methods for Phrase-Based Statistical Machine Translation. In *Proceedings of Human Language Technology: the Conference of the North American Chapter of the Association for Computational Linguistics*, pages 484-491.

Masao Utiyama, Andrew Finch, Hideo Okuma, Michael Paul, Hailong Cao, Hirofumi Yamamoto, Keiji Yasuda,and Eiichiro Sumita. 2008. The NICT/ATR speech Translation System for IWSLT 2008. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 77-84.

Haifeng Wang, Hua Wu, Xiaoguang Hu, Zhanyi Liu, Jianfeng Li, Dengjun Ren, and Zhengyu Niu. 2008. The TCH Machine Translation System for IWSLT 2008. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 124-131.

Hua Wu and Haifeng Wang. 2007. Pivot Language Approach for Phrase-Based Statistical Machine Translation. In *Proceedings of 45th Annual Meeting of the Association for Computational Linguistics*, pages 856-863.

Hua Wu and Haifeng Wang. 2009. Revisiting Pivot Language Approach for Machine Translation. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th IJCNLP of the AFNLP*, pages 154-162.

Samira Tofighi Zahabi, Somayeh Bakhshaei and Shahram Khadivi. Using Context Vectors in Improving a Machine Translation System with Bridge Language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 318-322.