Revisiting Pivot Language Approach for Machine Translation

Hua Wu and **Haifeng Wang**
Toshiba (China) Research and Development Center
5/F., Tower W2, Oriental Plaza, Beijing, 100738, China
{wuhua, wanghaifeng}@rdc.toshiba.com.cn

Abstract

This paper revisits the pivot language approach for machine translation. First, we investigate three different methods for pivot translation. Then we employ a hybrid method combining RBMT and SMT systems to fill up the data gap for pivot translation, where the source-pivot and pivot-target corpora are independent. Experimental results on spoken language translation show that this hybrid method significantly improves the translation quality, which outperforms the method using a source-target corpus of the same size. In addition, we propose a system combination approach to select better translations from those produced by various pivot translation methods. This method regards system combination as a translation evaluation problem and formalizes it with a regression learning model. Experimental results indicate that our method achieves consistent and significant improvement over individual translation outputs.

1 Introduction

Current statistical machine translation (SMT) systems rely on large parallel and monolingual training corpora to produce translations of relatively higher quality. Unfortunately, large quantities of parallel data are not readily available for some languages pairs, therefore limiting the potential use of current SMT systems. In particular, for speech translation, the translation task often focuses on a specific domain such as the travel domain. It is especially difficult to obtain such a domain-specific corpus for some language pairs such as Chinese to Spanish translation.

To circumvent the data bottleneck, some researchers have investigated to use a pivot language approach (Cohn and Lapata, 2007; Utiyama and Isahara, 2007; Wu and Wang 2007; Bertoldi et al., 2008). This approach introduces a third language, named the pivot language, for which there exist large source-pivot and pivot-target bilingual corpora. A pivot task was also designed for spoken language translation in the evaluation campaign of IWSLT 2008 (Paul, 2008), where English is used as a pivot language for Chinese to Spanish translation.

Three different pivot strategies have been investigated in the literature. The first is based on phrase table multiplication (Cohn and Lapata 2007; Wu and Wang, 2007). It multiples corresponding translation probabilities and lexical weights in source-pivot and pivot-target translation models to induce a new source-target phrase table. We name it the *triangulation* method. The second is the sentence translation strategy, which first translates the source sentence to the pivot sentence, and then to the target sentence (Utiyama and Isahara, 2007; Khalilov et al., 2008). We name it the *transfer* method. The third is to use existing models to build a synthetic source-target corpus, from which a source-target model can be trained (Bertoldi et al., 2008). For example, we can obtain a source-pivot corpus by translating the pivot sentence in the source-pivot corpus into the target language with pivot-target translation models. We name it the *synthetic* method.

The working condition with the pivot language approach is that the source-pivot and pivot-target parallel corpora are independent, in the sense that they are not derived from the same set of sentences, namely independently sourced corpora. Thus, some linguistic phenomena in the source-pivot corpus will lost if they do not exist in the pivot-target corpus, and vice versa. In order to fill up this data gap, we make use of rule-based machine translation (RBMT) systems to translate the pivot sentences in the source-pivot or pivot-target corpus into target or source sentences. As a result, we can build a synthetic multilingual corpus, which can be used to improve the translation quality. The idea of using RBMT systems to improve the translation quality of SMT sysems has been explored in Hu et al. (2007). Here, we re-examine the hybrid method to fill up the data gap for pivot translation.

Although previous studies proposed several pivot translation methods, there are no studies to combine different pivot methods for translation quality improvement. In this paper, we first compare the individual pivot methods and then investigate to improve pivot translation quality by combining the outputs produced by different systems. We propose to regard system combination as a translation evaluation problem. For translations from one of the systems, this method uses the outputs from other translation systems as pseudo references. A regression learning method is used to infer a function that maps a feature vector (which measures the similarity of a translation to the pseudo references) to a score that indicates the quality of the translation. Scores are first generated independently for each translation, then the translations are ranked by their respective scores. The candidate with the highest score is selected as the final translation. This is achieved by optimizing the regression learning model's output to correlate against a set of training examples, where the source sentences are provided with several reference translations, instead of manually labeling the translations produced by various systems with quantitative assessments as described in (Albrecht and Hwa, 2007; Duh, 2008). The advantage of our method is that we do not need to manually label the translations produced by each translation system, therefore enabling our method suitable for translation selection among any systems without additional manual work.

We conducted experiments for spoken language translation on the pivot task in the IWSLT 2008 evaluation campaign, where Chinese sentences in travel domain need to be translated into Spanish, with English as the pivot language. Experimental results show that (1) the performances of the three pivot methods are comparable when only SMT systems are used. However, the triangulation method and the transfer method significantly outperform the synthetic method when RBMT systems are used to improve the translation quality; (2) The hybrid method combining SMT and RBMT system for pivot translation greatly improves the translation quality. And this translation quality is higher than that of those produced by the system trained with a real Chinese-Spanish corpus; (3) Our sentence-level translation selection method consistently and significantly improves the translation quality over individual translation outputs in all of our experiments.

Section 2 briefly introduces the three pivot translation methods. Section 3 presents the hybrid method combining SMT and RBMT systems. Section 4 describes the translation selection method. Experimental results are presented in Section 5, followed by a discussion in Section 6. The last section draws conclusions.

2 Pivot Methods for Phrase-based SMT

2.1 Triangulation Method

Following the method described in Wu and Wang (2007), we train the source-pivot and pivot-target translation models using the source-pivot and pivot-target corpora, respectively. Based on these two models, we induce a source-target translation model, in which two important elements need to be induced: phrase translation probability and lexical weight.

Phrase Translation Probability We induce the phrase translation probability by assuming the independence between the source and target phrases when given the pivot phrase.

$$\phi(\bar{s}|\bar{t}) = \sum_{\bar{p}} \phi(\bar{s}|\bar{p})\phi(\bar{p}|\bar{t}) \qquad (1)$$

Where $\bar{s}$, $\bar{p}$ and $\bar{t}$ represent the phrases in the languages L_s, L_p and L_t, respectively.

Lexical Weight According to the method described in Koehn et al. (2003), there are two important elements in the lexical weight: word alignment information a in a phrase pair $(\bar{s}, \bar{t})$ and lexical translation probability $w(s|t)$.

Let a_1 and a_2 represent the word alignment information inside the phrase pairs $(\bar{s}, \bar{p})$ and $(\bar{p}, \bar{t})$ respectively, then the alignment information inside $(\bar{s}, \bar{t})$ can be obtained as shown in Eq. (2).

$$a = \{(s,t)|\exists p : (s,p) \in a_1 \ \& \ (p,t) \in a_2\} \qquad (2)$$

Based on the the induced word alignment information, we estimate the co-occurring frequencies of word pairs directly from the induced phrase pairs. Then we estimate the lexical translation probability as shown in Eq. (3).

$$w(s|t) = \frac{count(s,t)}{\sum_{s'} count(s',t)} \quad (3)$$

Where $count(s,t)$ represents the co-occurring frequency of the word pair (s,t).

2.2 Transfer Method

The transfer method first translates from the source language to the pivot language using a source-pivot model, and then from the pivot language to the target language using a pivot-target model. Given a source sentence s, we can translate it into n pivot sentences $p_1, p_2, ..., p_n$ using a source-pivot translation system. Each p_i can be translated into m target sentences $t_{i1}, t_{i2}, ..., t_{im}$. We rescore all the $n \times m$ candidates using both the source-pivot and pivot-target translation scores following the method described in Utiyama and Isahara (2007). If we use h^{fp} and h^{pt} to denote the features in the source-pivot and pivot-target systems, respectively, we get the optimal target translation according to the following formula.

$$\hat{t} = \underset{t}{\operatorname{argmax}} \sum_{k=1}^{L} (\lambda_k^{sp} h_k^{sp}(s,p) + \lambda_k^{pt} h_k^{pt}(p,t)) \quad (4)$$

Where L is the number of features used in SMT systems. λ^{sp} and λ^{pt} are feature weights set by performing minimum error rate training as described in Och (2003).

2.3 Synthetic Method

There are two possible methods to obtain a source-target corpus using the source-pivot and pivot-target corpora. One is to obtain target translations for the source sentences in the source-pivot corpus. This can be achieved by translating the pivot sentences in source-pivot corpus to target sentences with the pivot-target SMT system. The other is to obtain source translations for the target sentences in the pivot-target corpus using the pivot-source SMT system. And we can combine these two source-target corpora to produced a final synthetic corpus.

Given a pivot sentence, we can translate it into n source or target sentences. These n translations together with their source or target sentences are used to create a synthetic bilingual corpus. Then we build a source-target translation model using this corpus.

3 Using RBMT Systems for Pivot Translation

Since the source-pivot and pivot-target parallel corpora are independent, the pivot sentences in the two corpora are distinct from each other. Thus, some linguistic phenomena in the source-pivot corpus will lost if they do not exist in the pivot-target corpus, and vice versa. Here we use RBMT systems to fill up this data gap. For many source-target language pairs, the commercial pivot-source and/or pivot-target RBMT systems are available on markets. For example, for Chinese to Spanish translation, English to Chinese and English to Spanish RBMT systems are available.

With the RBMT systems, we can create a synthetic multilingual source-pivot-target corpus by translating the pivot sentences in the pivot-source or pivot-target corpus. The source-target pairs extracted from this synthetic multilingual corpus can be used to build a source-target translation model. Another way to use the synthetic multilingual corpus is to add the source-pivot or pivot-target sentence pairs in this corpus to the training data to rebuild the source-pivot or pivot-target SMT model. The rebuilt models can be applied to the triangulation method and the transfer method as described in Section 2.

Moreover, the RBMT systems can also be used to enlarge the size of bilingual training data. Since it is easy to obtain monolingual corpora than bilingual corpora, we use RBMT systems to translate the available monolingual corpora to obtain synthetic bilingual corpus, which are added to the training data to improve the performance of SMT systems. Even if no monolingual corpus is available, we can also use RBMT systems to translate the sentences in the bilingual corpus to obtain alternative translations. For example, we can use source-pivot RBMT systems to provide alternative translations for the source sentences in the source-pivot corpus.

In addition to translating training data, the source-pivot RBMT system can be used to translate the test set into the pivot language, which can be further translated into the target language with the pivot-target RBMT system. The translated test set can be added to the training data to further improve translation quality. The advantage of this method is that the RBMT system can provide translations for sentences in the test set and cover some out-of-vocabulary words in the test set that are uncovered by the training data. It can also change the distribution of some phrase pairs and reinforce some phrase pairs relative to the test set.

4 Translation Selection

We propose a method to select the optimal translation from those produced by various translation systems. We regard sentence-level translation selection as a machine translation (MT) evaluation problem and formalize this problem with a regression learning model. For each translation, this method uses the outputs from other translation systems as pseudo references. The regression objective is to infer a function that maps a feature vector (which measures the similarity of a translation from one system to the pseudo references) to a score that indicates the quality of the translation. Scores are first generated independently for each translation, then the translations are ranked by their respective scores. The candidate with the highest score is selected.

The similar ideas have been explored in previous studies. Albrecht and Hwa (2007) proposed a method to evaluate MT outputs with pseudo references using support vector regression as the learner to evaluate translations. Duh (2008) proposed a ranking method to compare the translations proposed by several systems. These two methods require quantitative quality assessments by human judges for the translations produced by various systems in the training set. When we apply such methods to translation selection, the relative values of the scores assigned by the subject systems are important. In different data conditions, the relative values of the scores assigned by the subject systems may change. In order to train a reliable learner, we need to prepare a balanced training set, where the translations produced by different systems under different conditions are required to be manually evaluated. In extreme cases, we need to relabel the training data to obtain better performance. In this paper, we modify the method in Albrecht and Hwa (2007) to only prepare human reference translations for the training examples, and then evaluate the translations produced by the subject systems against the references using BLEU score (Papineni et al., 2002). We use smoothed sentence-level BLEU score to replace the human assessments, where we use additive smoothing to avoid zero BLEU scores when we calculate the n-gram precisions. In this case, we can easily retrain the learner under different conditions, therefore enabling our method to be applied to sentence-level translation selection from any sets of translation systems without any additional human work.

ID	Description
1-4	n-gram precisions against pseudo references ($1 \leq n \leq 4$)
5-6	PER and WER
7-8	precision, recall, fragmentation from METEOR (Lavie and Agarwal, 2007)
9-12	precisions and recalls of non-consecutive bigrams with a gap size of m ($1 \leq m \leq 2$)
13-14	longest common subsequences
15-19	n-gram precision against a target corpus ($1 \leq n \leq 5$)

Table 1: Feature sets for regression learning

In regression learning, we infer a function f that maps a multi-dimensional input vector $\mathbf{x}$ to a continuous real value y, such that the error over a set of m training examples, $(\mathbf{x_1}, y_1), (\mathbf{x_2}, y_2), ..., (\mathbf{x_m}, y_m)$, is minimized according to a loss function. In the context of translation selection, y is assigned as the smoothed BLEU score. The function f represents a mathematic model of the automatic evaluation metrics. The input sentence is represented as a feature vector $\mathbf{x}$, which are extracted from the input sentence and the comparisons against the pseudo references. We use the features as shown in Table 1.

5 Experiments

5.1 Data

We performed experiments on spoken language translation for the pivot task of IWSLT 2008. This task translates Chinese to Spanish using English as the pivot language. Table 2 describes the data used for model training in this paper, including the BTEC (Basic Travel Expression Corpus) Chinese-English (CE) corpus and the BTEC English-Spanish (ES) corpus provided by IWSLT 2008 organizers, the HIT olympic CE corpus (2004-863-008)[1] and the Europarl ES corpus[2]. There are two kinds of BTEC CE corpus: BTEC CE1 and

[1] http://www.chineseldc.org/EN/purchasing.htm

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

Corpus	Size	SW	TW
BTEC CE1	20,000	164K	182K
BTEC CE2	18,972	177K	182K
HIT CE	51,791	490K	502K
BTEC ES	19,972	182K	185K
Europarl ES	400,000	8,485K	8,219K

Table 2: Training data. SW and TW represent source words and target words, respectively.

BTEC CE2. BTEC CE1 was distributed for the pivot task in IWSLT 2008 while BTEC CE2 was for the BTEC CE task, which is parallel to the BTEC ES corpus. For Chinese-English translation, we mainly used BTEC CE1 corpus. We used the BTEC CE2 corpus and the HIT Olympic corpus for comparison experiments only. We used the English parts of the BTEC CE1 corpus, the BTEC ES corpus, and the HIT Olympic corpus (if involved) to train a 5-gram English language model (LM) with interpolated Kneser-Ney smoothing. For English-Spanish translation, we selected 400k sentence pairs from the Europarl corpus that are close to the English parts of both the BTEC CE corpus and the BTEC ES corpus. Then we built a Spanish LM by interpolating an out-of-domain LM trained on the Spanish part of this selected corpus with the in-domain LM trained with the BTEC corpus.

For Chinese-English-Spanish translation, we used the development set (devset3) released for the pivot task as the test set, which contains 506 source sentences, with 7 reference translations in English and Spanish. To be capable of tuning parameters on our systems, we created a development set of 1,000 sentences taken from the training sets, with 3 reference translations in both English and Spanish. This development set is also used to train the regression learning model.

5.2 Systems and Evaluation Method

We used two commercial RBMT systems in our experiments: System A for Chinese-English bidirectional translation and System B for English-Chinese and English-Spanish translation. For phrase-based SMT translation, we used the Moses decoder (Koehn et al., 2007) and its support training scripts. We ran the decoder with its default settings and then used Moses' implementation of minimum error rate training (Och, 2003) to tune the feature weights on the development set.

To select translation among outputs produced by different pivot translation systems, we used SVM-light (Joachins, 1999) to perform support vector regression with the linear kernel.

Translation quality was evaluated using both the BLEU score proposed by Papineni et al. (2002) and also the modified BLEU (BLEU-Fix) score[3] used in the IWSLT 2008 evaluation campaign, where the brevity calculation is modified to use closest reference length instead of shortest reference length.

5.3 Results by Using SMT Systems

We conducted the pivot translation experiments using the BTEC CE1 and BTEC ES described in Section 5.1. We used the three methods described in Section 2 for pivot translation. For the transfer method, we selected the optimal translations among 10×10 candidates. For the synthetic method, we used the ES translation model to translate the English part of the CE corpus to Spanish to construct a synthetic corpus. And we also used the BTEC CE1 corpus to build a EC translation model to translate the English part of ES corpus into Chinese. Then we combined these two synthetic corpora to build a Chinese-Spanish translation model. In our experiments, only 1-best Chinese or Spanish translation was used since using n-best results did not greatly improve the translation quality. We used the method described in Section 4 to select translations from the translations produced by the three systems. For each system, we used three different alignment heuristics (grow, grow-diag, grow-diag-final[4]) to obtain the final alignment results, and then constructed three different phrase tables. Thus, for each system, we can get three different translations for each input. These different translations can serve as pseudo references for the outputs of other systems. In our case, for each sentence, we have 6 pseudo reference translations. In addition, we found out that the *grow* heuristic performed the best for all the systems. Thus, for an individual system, we used the translation results produced using the *grow* alignment heuristic.

The translation results are shown in Table 3. ASR and CRR represent different input conditions, namely the result of automatic speech recog-

[3] https://www.slc.atr.jp/Corpus/IWSLT08/eval/IWSLT08_auto_eval.tgz

[4] A description of the alignment heuristics can be found at http://www.statmt.org/jhuws/?n=FactoredTraining.Training Parameters

Method	BLEU	BLEU-Fix
Triangulation	33.70/27.46	31.59/25.02
Transfer	33.52/28.34	31.36/26.20
Synthetic	34.35/27.21	32.00/26.07
Combination	38.14/29.32	34.76/27.39

Table 3: CRR/ASR translation results by using SMT systems

nition and correct recognition result, respectively. Here, we used the 1-best ASR result. From the translation results, it can be seen that three methods achieved comparable translation quality on both ASR and CRR inputs, with the translation results on CRR inputs are much better than those on ASR inputs because of the errors in the ASR inputs. The results also show that our translation selection method is very effective, which achieved absolute improvements of about 4 and 1 BLEU scores on CRR and ASR inputs, respectively.

5.4 Results by Using both RBMT and SMT Systems

In order to fill up the data gap as discussed in Section 3, we used the RBMT System A to translate the English sentences in the ES corpus into Chinese. As described in Section 3, this corpus can be used by the three pivot translation methods. First, the synthetic Chinese-Spanish corpus can be combined with those produced by the EC and ES SMT systems, which were used in the synthetic method. Second, the synthetic Chinese-English corpus can be added into the BTEC CE1 corpus to build the CE translation model. In this way, the intersected English phrases in the CE corpus and ES corpus becomes more, which enables the Chinese-Spanish translation model induced using the triangulation method to cover more phrase pairs. For the transfer method, the CE translation quality can be also improved, which would result in the improvement of the Spanish translation quality.

The translation results are shown in the columns under ”EC RBMT” in Table 4. As compared with those in Table 3, the translation quality was greatly improved, with absolute improvements of at least 5.1 and 3.9 BLEU scores on CRR and ASR inputs for system combination results. The above results indicate that RBMT systems indeed can be used to fill up the data gap for pivot translation.

In our experiments, we also used a CE RBMT system to enlarge the size of training data by pro-

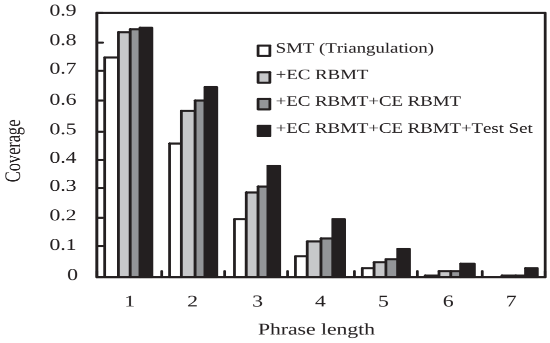

Figure 1: Coverage on test source phrases

viding alternative English translations for the Chinese part of the CE corpus. The translation results are shown in the columns under “+CE RBMT” in Table 4. From the translation results, it can be seen that, enlarging the size of training data with RBMT systems can further improve the translation quality.

In addition to translating the training data, the CE RBMT system can be also used to translate the test set into English, which can be further translated into Spanish with the ES RBMT system B.[5][6] The translated test set can be further added to the training data to improve translation quality. The columns under “+Test Set” in Table 4 describes the translation results. The results show that translating the test set using RBMT systems greatly improved the translation result, with further improvements of about 2 and 1.5 BLEU scores on CRR and ASR inputs, respectively.

The results also indicate that both the triangulation method and the transfer method greatly outperformed the synthetic method when we combined both RBMT and SMT systems in our experiments. Further analysis shows that the synthetic method contributed little to system combination. The selection results are almost the same as those selected from the translations produced by the triangulation and transfer methods.

In order to further analyze the translation results, we evaluated the above systems by examining the coverage of the phrase tables over the test phrases. We took the triangulation method as a case study, the results of which are shown in Fig-

[5] Although using the ES RBMT system B to translate the training data did not improve the translation quality, it improved the translation quality by translating the test set.

[6] The RBMT systems achieved a BLEU score of 24.36 on the test set.

	EC RBMT		+ CE RBMT		+ Test Set	
Method	**BLEU**	**BLEU-Fix**	**BLEU**	**BLEU-Fix**	**BLEU**	**BLEU-Fix**
Triangulation	40.69/31.02	37.99/29.15	41.59/31.43	39.39/29.95	44.71/32.60	42.37/31.14
Transfer	42.06/31.72	39.73/29.35	43.40/33.05	40.73/30.06	45.91/34.52	42.86/31.92
Synthetic	39.10/29.73	37.26/28.45	39.90/30.00	37.90/28.66	41.16/31.30	37.99/29.36
Combination	43.21/33.23	40.58/31.17	45.09/34.10	42.88/31.73	47.06/35.62	44.94/32.99

Table 4: CRR/ASR translation results by using RBMT and SMT systems

Method	**BLEU**	**BLEU-Fix**
Triangulation	45.64/33.15	42.11/31.11
Transfer	47.18/34.56	43.61/32.17
Combination	48.42/36.42	45.42/33.52

Table 5: CRR/ASR translation results by using additional monolingual corpora

Method	**Sys. A**	**Sys. B**	**Sys. A+B**
Triangulation	40.69	39.28	41.01
Transfer	42.06	39.57	43.03
Synthetic	39.10	38.24	39.26
Combination	43.21	40.59	44.27

Table 6: CRR translation results (BLEU scores) by using different RBMT systems

ure 1. It can be seen that using RBMT systems to translate the training and/or test data can cover more source phrases in the test set, which results in translation quality improvement.

5.5 Results by Using Monolingual Corpus

In addition to translating the limited bilingual corpus, we also translated additional monolingual corpus to further enlarge the size of the training data. We assume that it is easier to obtain a monolingual pivot corpus than to obtain a monolingual source or target corpus. Thus, we translated the English part of the HIT Olympic corpus into Chinese and Spanish using EC and ES RBMT systems. The generated synthetic corpus was added to the training data to train EC and ES SMT systems. Here, we used the synthetic CE Olympic corpus to train a model, which was interpolated with the CE model trained with both the BTEC CE1 corpus and the synthetic BTEC corpus to obtain an interpolated CE translation model. Similarly, we obtained an interpolated ES translation model. Table 5 describes the translation results.[7] The results indicate that translating monolingual corpus using the RBMT system further improved the translation quality as compared with those in Table 4.

6 Discussion

6.1 Effects of Different RBMT Systems

In this section, we compare the effects of two commercial RBMT systems with different translation accuracy on spoken language translation. The goals are (1) to investigate whether a RBMT system can improve pivot translation quality even if its translation accuracy is not high, and (2) to compare the effects of RBMT system with different translation accuracy on pivot translation. Besides the EC RBMT system A used in the above section, we also used the EC RBMT system B for this experiment.

We used the two systems to translate the test set from English to Chinese, and then evaluated the translation quality against Chinese references obtained from the IWSLT 2008 evaluation campaign. The BLEU scores are 43.90 and 29.77 for System A and System B, respectively. This shows that the translation quality of System B on spoken language corpus is much lower than that of System A. Then we applied these two different RBMT systems to translate the English part of the BTEC ES corpus into Chinese as described in Section 5.4. The translation results on CRR inputs are shown in Table 6.[8] We replicated some of the results in Table 4 for the convenience of comparison. The results indicate that the higher the translation accuracy of the RBMT system is, the better the pivot translation is. If we compare the results with those only using SMT systems as described in Table 3, the translation quality was greatly improved by at least 3 BLEU scores, even if the translation ac-

[7] Here we excluded the synthetic method since it greatly falls behind the other two methods.

[8] We omitted the ASR translation results since the trends are the same as those for CRR inputs. And we only showed BLEU scores since the trend for BLEU-Fix scores is similar.

Method	Multilingual	+ BTEC CE1
Triangulation	41.86/39.55	42.41/39.55
Transfer	42.46/39.09	43.84/40.34
Standard	42.21/40.23	42.21/40.23
Combination	43.75/40.34	44.68/41.14

Table 7: CRR translation results by using multilingual corpus. "/" separates the BLEU and BLEU-fix scores.

curacy of System B is not so high. Combining two RBMT systems further improved the translation quality, which indicates that the two systems complement each other.

6.2 Results by Using Multilingual Corpus

In this section, we compare the translation results by using a multilingual corpus with those by using independently sourced corpora. BTEC CE2 and BTEC ES are from the same source sentences, which can be taken as a multilingual corpus. The two corpora were employed to build CE and ES SMT models, which were used in the triangulation method and the transfer method. We also extracted the Chinese-Spanish (CS) corpus to build a standard CS translation system, which is denoted as *Standard*. The comparison results are shown in Table 7. The translation quality produced by the systems using a multilingual corpus is much higher than that produced by using independently sourced corpora as described in Table 3, with an absolute improvement of about 5.6 BLEU scores. If we used the EC RBMT system, the translation quality of those in Table 4 is comparable to that by using the multilingual corpus, which indicates that our method using RBMT systems to fill up the data gap is effective. The results also indicate that our translation selection method for pivot translation outperforms the method using only a real source-target corpus.

For comparison purpose, we added BTEC CE1 into the training data. The translation quality was improved by only 1 BLEU score. This again proves that our method to fill up the data gap is more effective than that to increase the size of the independently sourced corpus.

6.3 Comparison with Related Work

In IWSLT 2008, the best result for the pivot task is achieved by Wang et al. (2008). In order to compare the results, we added the bilingual HIT Olympic corpus into the CE training data.[9] We also compared our translation selection method with that proposed in (Wang et al., 2008) that is based on the target sentence average length (TSAL). The translation results are shown in Table 8. "Wang" represents the results in Wang et al. (2008). "TSAL" represents the translation selection method proposed in Wang et al. (2008), which is applied to our experiment. From the results, it can be seen that our method outperforms the best system in IWSLT 2008 and that our translation selection method outperforms the method based on target sentence average length.

	Ours	Wang	TSAL
BLEU	49.57	-	48.25
BLEU-Fix	46.74	45.10	45.27

Table 8: Comparison with related work

7 Conclusion

In this paper, we have compared three different pivot translation methods for spoken language translation. Experimental results indicated that the triangulation method and the transfer method generally outperform the synthetic method. Then we showed that the hybrid method combining RBMT and SMT systems can be used to fill up the data gap between the source-pivot and pivot-target corpora. By translating the pivot sentences in independent corpora, the hybrid method can produce translations whose quality is higher than those produced by the method using a source-target corpus of the same size. We also showed that even if the translation quality of the RBMT system is low, it still greatly improved the translation quality.

In addition, we proposed a system combination method to select better translations from outputs produced by different pivot methods. This method is developed through regression learning, where only a small size of training examples with reference translations are required. Experimental results indicate that this method can consistently and significantly improve translation quality over individual translation outputs. And our system outperforms the best system for the pivot task in the IWSLT 2008 evaluation campaign.

[9] We used about 70k sentence pairs for CE model training, while Wang et al. (2008) used about 100k sentence pairs, a CE translation dictionary and more monolingual corpora for model training.

References

Joshua S. Albrecht and Rebecca Hwa. 2007. Regression for Sentence-Level MT Evaluation with Pseudo References. In *Proceedings of the 45th Annual Meeting of the Accosiation of Computational Linguistics*, pages 296–303.

Nicola Bertoldi, Madalina Barbaiani, Marcello Federico, and Roldano Cattoni. 2008. Phrase-Based Statistical Machine Translation with Pivot Languages. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 143-149.

Tevor Cohn and Mirella Lapata. 2007. Machine Translation by Triangulation: Making Effective Use of Multi-Parallel Corpora. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 348–355.

Kevin Duh. 2008. Ranking vs. Regression in Machine Translation Evaluation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 191–194.

Xiaoguang Hu, Haifeng Wang, and Hua Wu. 2007. Using RBMT Systems to Produce Bilingual Corpus for SMT. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 287–295.

Thorsten Joachims. 1999. Making Large-Scale SVM Learning Practical. In Bernhard Schöelkopf, Christopher Burges, and Alexander Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press.

Maxim Khalilov, Marta R. Costa-Jussà, Carlos A. Henríquez, José A.R. Fonollosa, Adolfo Hernández, José B. Mariño, Rafael E. Banchs, Chen Boxing, Min Zhang, Aiti Aw, and Haizhou Li. 2008. The TALP & I2R SMT Systems for IWSLT 2008. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 116–123.

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, 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 Associa-tion for Computational Linguistics, demonstration session*, pages 177–180.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Proceedings of Workshop on Statistical Machine Translation at the 45th Annual Meeting of the Association of Computational Linguistics*, pages 228–231.

Franz J. Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167.

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 Computational Linguistics*, pages 311–318.

Michael Paul. 2008. Overview of the IWSLT 2008 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 1–17.

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.

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.