Enriching SMT Training Data via Paraphrasing

Wei He[1], **Shiqi Zhao**[12], **Haifeng Wang**[2], **Ting Liu**[1]
[1]Research Center for Social Computing and Information
Retrieval, Harbin Institute of Technology
{whe,tliu}@ir.hit.edu.cn
[2]Baidu
{zhaoshiqi,wanghaifeng}@baidu.com

Abstract

This paper proposes a novel method to resolve the coverage problem of SMT system. The method generates paraphrases for source-side sentences of the bilingual parallel data, which are then paired with the target-side sentences to generate new parallel data. Within a statistical paraphrase generation framework, we employ an object function, named Sentence Novelty, to select paraphrases which having the most novel information to the bilingual training corpus of the SMT model. Meanwhile, the context is considered via a language model in the source language to ensure the fluency and accuracy of paraphrase substitution. Compared to a state-of-the-art phrase based SMT system (Moses), our method achieves an improvement of 1.66 points in terms of BLEU on a small training corpus which simulates a resource-poor environment, and 1.06 points on a training corpus of medium size.

1 Introduction

Current statistical machine translation (SMT) systems learn how to translate by analyzing bilingual parallel corpora. Generally speaking, high-quality translations can be produced when ample training data is available. Previous studies have indicated that the translation quality can be improved by 2 points of BLEU (Papineni et al., 2002) when the size of the parallel data is doubled (Koehn et al., 2003). However, for the so called *low density* language pairs that do not have large-scale parallel corpora, limited amount of training data usually leads to a problem of low coverage in that many phrases encountered at run-time have not been observed in the training data. According to Callison-Burch et al. (2006), for a training corpus containing 10,000 words, translations will have been learned for only 10% of the unigrams in the test set. For a training corpus containing 100,000 words this increases to 30%. This problem becomes more serious for higher-order n-grams, and for morphologically richer languages.

To overcome the coverage problem of SMT, besides the efforts of mining larger parallel corpora from various resources, some researchers have investigated to use paraphrasing approaches. The studies can be classified into two categories by the target of paraphrasing: (1) paraphrasing the input source sentences; (2) paraphrasing the training corpus. In the first category, the proposed approaches mainly focus on handling n-grams that are unknown to the SMT model. Callison-Burch et al. (2006) and Marton et al. (2009) paraphrase unknown terms in the input sentences using phrasal paraphrases extracted from bilingual and monolingual corpora. Mirkin et al. (2009) rewrite unknown terms with entailments and paraphrases acquired from WordNet. Onishi et al. (2010) and Du et al. (2010) build paraphrase lattices for input sentences and select the best translations using a lattice-based SMT decoder. In the second category of paraphrasing training corpus, Bond et al. (2008) and Nakov (2008) paraphrase the source side of training corpus using hand-crafted rules.

In this paper, we propose a method that enriches SMT training data using a statistical paraphrase generating (SPG) model. The method generates paraphrases for the source-side sen-

This work was partially done when the first author was visiting Baidu.

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 803–810,
Chiang Mai, Thailand, November 8 – 13, 2011. ©2011 AFNLP

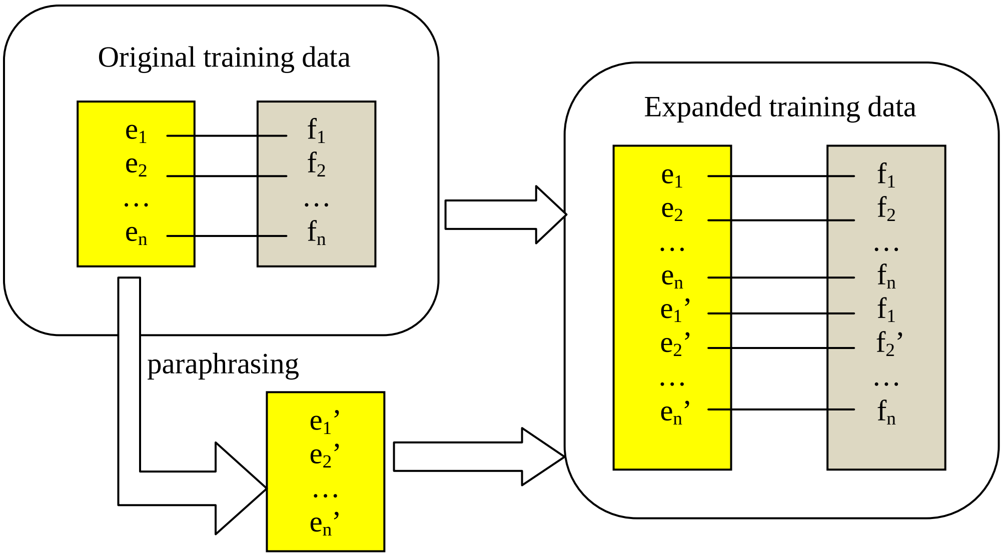

Figure 1. Sketch map of the paraphrasing based translation corpus expansion.

tences of the bilingual parallel data, which are then paired with the target-side sentences to generate new parallel data. The procedure is illustrated in Figure 1. The SPG framework can be considered as an application-specific source-to-source translating procedure (Zhao et al. 2009) which is similar to phrase based statistical machine translation. We employ an object function, named *Sentence Novelty*, to select paraphrases that introduce the most novel information to the bilingual training corpus. In our approach, the context of paraphrasing substitution is considered during generating paraphrasing sentences, which yields paraphrases with higher precision. Experimental results show that the performance of a state-of-the-art phrase based SMT system (Moses in this work) can be improved from 17.91 to 19.57 in terms of BLEU on a small training set, and from 25.46 to 26.52 on a training corpus of medium size. Results also indicate that our method gains a significant improvement over the method of Callison-Burch et al. (2006).

The rest of this paper is structured as follows. We review related work on improving SMT through paraphrasing in Section 2. The proposed statistical paraphrase generation model is described in Section 3. Section 4 presents our method of enlarging training data via paraphrasing. Section 5 and 6 present the experiments and results. We discuss our work in Section 7 and conclude the paper in Section 8.

2 Related Work

Previous studies on improving SMT through paraphrasing input sentences mainly focus on finding translations for unknown terms using phrasal paraphrases. In these methods, an unknown term can be paraphrased to a known term which has translations in the phrase table. Callison-Burch et al. (2006) acquire phrasal paraphrases from bilingual parallel corpora based on a pivot approach. The main idea is that phrases aligned with the same foreign phrase in a bilingual corpus may be paraphrases. The learned paraphrases are applied in a SMT system in the following manner. Suppose e_1 is an unknown source phrase, e_2 is a paraphrase of e_1, which can be translated as f in the phrase table, the method simply takes f as e_1's translation. A new phrase pair ($e_{1,}$ f) is added to the phrase table with an additional feature $h(f, e_1)$ to distinguish the original phrase pairs and the newly generated ones, which is defined as:

$$h(f, e_1) = \begin{cases} p(e_2|e_1) & \text{If phrase table entry } (f, e_1) \text{ is generated from } (f, e_2) \\ 1 & \text{Otherwise} \end{cases}$$

where $p(e_2|e_1)$ denotes the paraphrase probability.

Marton et al. (2009) propose a method similar to that of Callison-Burch et al. (2006). The only difference is that the paraphrases are extracted from monolingual corpora based on distributional hypothesis. Compared with bilingual corpora, it is easier to acquire monolingual corpora, especially for resource-poor languages.

Mirkin et al. (2009) utilize paraphrases and entailment rules, namely the synonyms and hyponyms from WordNet, to substitute unknown terms in source sentences. Some context models are also used for ranking and filtering the paraphrases and entailments before feeding them to the SMT engine.

Onishi et al. (2010) and Du et al. (2010) build paraphrase lattices for the input sentences. In this scenario, paraphrases are in fact competing with each other. All possible paraphrases are kept and finally selected by the SMT decoder.

Experimental results in these works have proved that the methods that paraphrase input sentences indeed improve SMT results by increasing coverage, especially on small training sets. However, the approaches have two problems. The first one is efficiency. All of the methods that improve SMT through paraphrasing input sentences can be considered as a two-stage procedure, i.e., collecting paraphrases for unknown terms and then translating. Obviously, low efficiency is the bottleneck for this kind of method, since it goes through decoding twice, one for paraphrasing and one for translating. The other problem is that the context is not considered during phrasal paraphrase substitution, which causes a low paraphrasing accuracy. Notice that many paraphrase substitutions are acceptable only in specific contexts. For example, *bank* and *shore* are paraphrases, but we can only substitute *bank* with *shore* in a context related to rivers. Without considering the paraphrase's context, the paraphrasing substitution has a relatively low accuracy, which limits the effect of these methods. The only exception is the work of Mirkin et al. (2009), which uses context models for ranking paraphrases. However, the generated paraphrases without paraphrasing probabilities are difficult to be incorporated with a statistical context model. As described in Mirkin et al. (2009), the main contribution of context models was to reduce the number of paraphrase candidates and improve the efficiency of the system. In contrast, in our method, all the work that enriches SMT with paraphrases is conducted in the training step, which avoids affecting the decoding procedure. Meanwhile, the context of paraphrase substitution is considered using a source language model in our method.

Other researches directly enlarge SMT training corpora based on paraphrase techniques. Nakov (2008) employs six rules for paraphrasing the training corpus. Here we list two rules as examples:

1. $[_{\text{NP}}$ **NP**$_1$ ***of*** **NP**$_2]$ → $[_{\text{NP}}$ **NP**$_2$ **gen NP**$_1]$

the lifting of the beef import ban → *the beef import ban's lifting*

2. **NP**$_{gen}$ → **NP**

Commissioner's statement→*Commissioner statement*

where: gen is a genitive marker: ' or 's; NP$_{gen}$ is an NP with an internal genitive marker.

Bond et al. (2008) use grammars to paraphrase the source side of training data, covering aspects like word order and minor lexical variations (tenses etc.) but not content words. The paraphrases are added to the source side of the corpus and the corresponding target sentences are duplicated.

The above-mentioned methods that expand training data via paraphrasing have two disadvantages: (1) hand-crafted paraphrasing rules are language-dependent; (2) to ensure the paraphrase accuracy, only some simple paraphrase rules are used. Our work should be classified into this category. But a clear difference is that our paraphrase generation method is a statistical one without any language specific feature, which (1) utilizes paraphrase resources extracted from large-scale corpora; (2) balances the accuracy and variation rate of paraphrases with a decoding algorithm that searches for the optimal path among all the paraphrasing candidates.

3 Paraphrase Generation

3.1 Paraphrasing Framework

We employ an application-driven statistical paraphrase generation framework which is proposed by Zhao et al. (2009). The framework is based on a log-linear model in which three submodels are defined, namely, a paraphrase model, a language model and a usability model, which control the adequacy, fluency and usability of the paraphrases, respectively.

Paraphrase generation is a decoding process similar to SMT. The input sentence S is first segmented into a sequence of I units $\bar{s}_1^I$, which are then paraphrased to a sequence of units $\bar{t}_1^I$. Let $(\bar{s}_i, \bar{t}_i)$ be a pair of paraphrase units, their paraphrase likelihood is computed using a score function $\varphi_{pm}(\bar{s}_i, \bar{t}_i)$. Thus the paraphrase score $p_{pm}(\bar{s}_1^I, \bar{t}_1^I)$ between S and T is decomposed into:

$$p_{pm}(\bar{s}_1^I, \bar{t}_1^I) = \prod_{i=1}^{I} \varphi_{pm}(\bar{s}_i, \bar{t}_i)^{\lambda_{pm}}$$

where λ_{pm} is the weight of the paraphrase model.

A four-gram language model is employed to ensure the fluency and eliminate the ambiguity of paraphrase. The language model based score for the paraphrase T is computed as:

$$p_{lm}(T) = \prod_{j=1}^{J} p(t_j \mid t_{j-3} t_{j-2} t_{j-1})^{\lambda_{lm}}$$

where J is the length of T, t_j is the j-th word of T, and λ_{lm} is the weight for the language model.

The usability model prefers paraphrase units that are more suitable for the application. The usability of T depends on paraphrase units it contains. We propose a specific usability model to enrich SMT training corpus, which is described in chapter 3.2.

3.2 Sentence Novelty Model

In this paper, we do not limit our method to handling unknown terms. Instead, our goal is to grub knowledge from paraphrases and enrich the translation corpora. Therefore, within the application-driven paraphrase generation framework, we propose a specific paraphrasing usability model, *sentence novelty*, for selecting paraphrases which contain the most novel n-grams to the translation model. Given a paraphrased sentence T, which consists of J words, the novelty function $Novel(TM,T,n,j)$ judges whether the occurrence of t_j generates a new n-gram to the translation model (TM) according to the prior n-1 words of t_j. Formally, the novel function for position j can be defined as:

$$Novel(TM,T,n,j) \begin{cases} 1 & \text{If } t_{j-n+1}\ldots t_j \text{ is a new n-gram to } TM \\ 0 & \text{otherwise} \end{cases}$$

Thus the novelty model for a paraphrased sentence T, considering the novelty of 1-gram to N-gram (N=4 in this work), is computed as:

$$p_{nm}(t) = \exp\Big(\sum_{j=1}^{J}\sum_{n=1}^{N} Novel(TM,t,n,j)\Big)^{\lambda_{nm}}$$

where λ_{nm} is the weight for the novelty model,

Now we can describe the complete formula of the SPG framework as:

$$\begin{aligned} p(T \mid S) = {} & \lambda_{pm}\sum_{i=1}^{I}\log\varphi_{pm}(\bar{s}_i,\bar{t}_i) \\ & + \lambda_{lm}\sum_{j=1}^{J}\log p(t_j \mid t_{j-3}t_{j-2}t_{j-1}) \\ & + \lambda_{nm}\sum_{j=1}^{J}\sum_{n=1}^{N} Novel(TM,T,n,j) \end{aligned}$$

4 Expanding SMT Training Corpus

4.1 Corpus Expansion

We enhance the SMT model by expanding the training corpus using paraphrases. Firstly, the sentence-level paraphrases are generated on the source side (English in our experiments) of the training bi-texts. Then the paraphrased sentences and the corresponding translations on the target side (Chinese in this work) which align with the original sentences compose new bilingual sentence pairs.

To grub knowledge from paraphrases as much as possible, we exploit two different strategies for paraphrase generation in the experiments: (1) generating 1-best paraphrase for every source sentence in the training corpus, and (2) generating k-best paraphrases for a source sentence and selecting m sentences from them which have the most novel n-grams. Thus we get two paraphrased bilingual corpora besides the original corpus. Sentence pairs generated by the two strategies are shown in Table 1. From the table, it can be seen that on the source side the 1-best paraphrase sentence has a relatively high quality, while the sentences selected from k-best paraphrase results have lower accuracy but higher coverage. On the target side, the original Chinese sentence is just copied to align with the generated paraphrase sentences.

4.2 Paraphrase Selecting Strategy

As mentioned above, in strategy (2) we selected m paraphrases in the generated top-k results, which have the most different n-grams. The reason of not using all the k-best results for improving SMT is that the top-k paraphrases generated for a sentence are generally very similar, if we train the SMT model on all these sentences, it would be quite time-consuming and much of the computation is vain. Therefore we propose an algorithm to select a subset from all the paraphrase sentences, which can cover most of the newly introduced information while dramatically reduce the numbers of paraphrases. The algorithm is described in Figure 2.

```
procedure SENTENCE_SELECTION
input: m, set S {k-best paraphrase sentences:$S_1$,…,$S_k$}
todo: select m sentences from set S
  M := {$S_1$}, remove $S_1$ from S
  while (|M| < m)
    MAX_DISTANCE := 0
    i-max := 0
    for $S_i$ := each sentences in S
       $A_i$ := AVERAGE_EDIT_DISTANCE($S_i$,M)
       if $A_i$ > MAX_DISTANCE
         MAX_DISTANCE := $A_i$
         i-max = i
    M := M ∪ {$S_{i-max}$}, remove $S_{i-max}$ from S
return M
```

Figure 2: The algorithm for paraphrase selection.

	Source sentences	Target sentences
original	Solving environmental problems is a big and urgent mission.	解决环境问题已经为刻不容缓的重大任务。
1-best	**The resolution of** environmental problems is a **large** and urgent **task**.	解决环境问题已经为刻不容缓的重大任务。
selected k-best	**The resolution of** environmental problems is a **large** and urgent **task**. **The solution** to environmental problems is **high** and urgent **task**. **The resolution of** environmental problems is a **major** urgent and mission. Solving environmental problems **are** a big and urgent **task**. …	解决环境问题已经为刻不容缓的重大任务。 解决环境问题已经为刻不容缓的重大任务。 解决环境问题已经为刻不容缓的重大任务。 解决环境问题已经为刻不容缓的重大任务。 …

Table 1: Examples of generated sentence pairs.

At the beginning of the algorithm, the selected sentence set *M* is empty. In each iteration, we select a sentence $S_{i\text{-}max}$ from the *k*-best paraphrase sentence set *S* and add it to *M*. In the selection of $S_{i\text{-}max}$, we calculate the average Edit Distance (ED) between each candidate sentence S_i and all sentences in *M* by the function of AVERAGE_EDIT_DISTANCE(S_i,*M*). The sentence most different (with the largest average ED) from the already selected sentences in *M* is selected. In the algorithm, the edit distance among the sentences in *M* has been considered as the optimization objective function, which ensures the sentences with most novel n-grams are selected.

4.3 Model Integrating

After corpus expansion, we get three bilingual corpora for SMT training: (1) the original corpus, (2) corpus of 1-best paraphrases on the source side and original translations on the target side, (3) corpus of selected *m* paraphrases on the source side and original translations on the target side. Notice that the reliability of the three corpora is in a descending order. The original corpus which is produced by human can be considered as golden standard corpus. The quality of corpus consisting of 1-best paraphrases (*1-PARA* corpus in the following context) is lower than the original corpus. The corpus of *m* paraphrase sentences which were selected from the *k*-best paraphrase results by the algorithm described in 4.2 (namely, *M-PARA* corpus) may be the most noisy.

Considering the different reliabilities of these corpora, simply merging them into a new corpus and train a translation model is not the optimal solution. Therefore, within a phrase-based SMT framework, we train three phrase tables from these corpora, and then integrate these phrase tables with different weights. The integration is a procedure of linear interpolation which can be described in the following formula:

$$PT = \sum_{i=1}^{n} \lambda_i PT_i$$

where λ_n is weight of PT_n, which is set up empirically.

We first merge the phrase tables trained from the original corpus and *1-PARA* corpus, and get a new phrase table (Original + *1-PARA*). Then we integrate the Original + *1-PARA* phrase table with the phrase table trained from the *M-PARA* corpus and get another phrase table (Original + *1-PARA* + *M-PARA)*. The effectiveness of these enriched phrase tables is tested in the experimental section.

5 Experimental Setup

5.1 Paraphrase Resources

The paraphrase generating framework we used is not limited to a certain type of paraphrase resource. Any paraphrase resources with paraphrasing probability can be integrated into the framework. We simply choose phrasal paraphrases acquired from the Europarl corpus using Callison-Burch's paraphrase extracting toolkit[1]. The toolkit supports extraction of both phrasal paraphrases and syntactically constrained paraphrases. In this paper, we only use the phrasal paraphrase extracting part of the toolkit which extracts paraphrases from bilingual corpus using

[1] http://cs.jhu.edu/~ccb/howto-extract-paraphrases.html

pivot method (Colin Bannard and Chris Callison-Burch. 2005).

We extract phrasal paraphrases for all n-grams ($n \leq 6$) in the source sentences in the training data. Some operations are performed on the extracted phrasal paraphrases to ensure the accuracy: (1) paraphrases with score < .03 are filtered out, (2) paraphrases consisting of nothing but stop-words are removed.

5.2 SMT Data

For the baseline system, we trained on the Sinorama and FBIS corpora (LDC2005T10 and LDC2003E14). After tokenization and filtering, this bilingual corpus contained 319,694 lines (7.9M tokens on Chinese side and 9.2M tokens on English side). We trained a 4-gram language model on the Chinese side of the bi-text. Then we randomly selected 29,000 lines form the bi-text, and constructed a reduced training corpus to simulate a resource-poor language. We tested the system using the English-Chinese NIST MT 2008 evaluation set. The test set contains 1859 English sentences, each of which has four human references for automatic evaluation. For development, we used the Chinese-English NIST MT 2005 evaluation set, taking one of the English references as source, and the Chinese source as a single reference translation. All the Chinese sentences in the training corpora were segmented with the word segmentation tool from Language Technology Platform (LTP)[2]. We used two metrics, BLEU[3] and TER[4] (Snover et al., 2005), for automatic evaluation. Following the evaluation standard of NIST, the system translations and references were split into Chinese characters in automatic evaluation.

5.3 Translation Model

We used Moses, a state-of-the-art phrase-based SMT model (Koehn et al., 2007), in decoding. In Moses, the generated translation hypotheses are scored mainly based on a translation model, a language model, and a reordering model. These components are deemed as features and combined within a log-linear framework:

$$e^{*} = aug \max_{e} \{ \sum_{i=1}^{n} \lambda_i h_i(e, f) \}$$

where $h_i(e, f)$ is a feature function with λ_i as the weight. The feature weights can be trained with Minimum Error Rate Training (MERT) (Och, 2003) on the development set using BLEU as the objective function.

	29k	**Full**
Ori.	324k	3603k
+1-PARA	507k(+56%)	4514k(+25%)
+1-PARA+M-PARA	878k(+171%)	8359k(+132%)

Table 2: Number of phrase pairs in different phrase tables.

Set	**Model**	**BLEU-4**	**TER**
29k	Baseline	17.91	66.83
	CB	18.75	66.68
	Ori.+1-PARA	19.26***	65.98
	Ori.+1-PARA +M-PARA	**19.57*** **	**65.88**
Full	Baseline	25.46	62.36
	Callison-Burch	25.76	61.62
	Ori.+1-PARA	**26.52*** **	**61.36**
	Ori.+1-PARA +M-PARA	26.33***	61.47

Table 3: Experimental results: "***" means that the method performs significantly better than both the baseline and Callison-Burch with $\rho < 0.01$, using Koehn's (2004) pair-wise bootstrap test for BLEU with 95% confidence interval.

	1gram	2gram	3gram	4gram
Baseline	50.4%	17.9%	3.9%	0.6%
+1-PARA	54.2%	21.5%	5.0%	0.9%
+1-PA.+M-PA.	56.3%	24.7%	6.2%	1.1%
CB	56.8%	26.3%	6.4%	1.5%

Table 4: Coverage rate of phrase tables trained from 29k training set.

6 Results and Analysis

After corpus expansion and model integrating, the size of the original phrase table is increased. The number of phrase pairs in the original PT and the merged PTs which are extracted from 29k and full corpora are shown in Table 2.

As can be seen, the number of phrase pairs is significantly increased after corpus expansion. Specifically, the sizes of the augmented phrase tables are increased by 56% and 171% for the 29k set, 25% and 132% for the full set, which prove that our sentence novelty model has made considerable contributions to the enrichment of phrase tables.

We evaluated the effectiveness of the enriched phrase tables in translation. In order to conduct a direct comparison with the existing techniques, we took Moses trained with the original bi-text

[2] http://ir.hit.edu.cn/ltp/
[3] ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v13a.pl
[4] http://www.umiacs.umd.edu/~snover/terp/

1	Source sentence	cyber experts said the investigations in india will take time
	Baseline result	**网络**版**专家说，在印度将**把**时间调查。**
	Our result	**网络**版**专家说，在 印度的调查需要 时间。**
2	Source sentence	it happens again and again .
	Baseline result	它再**次**与再**次发生。**
	Our result	**发生这 一次 又一次。**
3	Source sentence	people just talk about cars and stuff .
	Baseline result	**人们只谈车和材料。**
	Our result	**人只谈**公园**和材料。**

Table 5: Translation examples on 29k-bitext systems. The n-grams that match the references are highlighted in bold. Here *our result* refers to the system of Original+1-PARA+M-PARA.

as a baseline. We also developed another comparison system by re-implementing the method proposed by Callison-Burch et al. (2006) (CB for short hereafter), which used the same paraphrase resource as described above to paraphrase unknown phrases (up to 6-gram) of the test sentences. The feature weights for the model of Callison-Burch et al. (2006) were trained with MERT on the development set.

The evaluation results are shown in Table 3. We can see that our method outperforms both the baseline and CB models under both evaluation metrics. On the 29k subset, the improved model using only 1-best paraphrases has a significant 1.35 BLEU points gain over its baseline, and the model integrated with both 1-best and m-best paraphrases has a further improvement of 1.66 points. On the full set, the model augmented by 1-best paraphrases achieves the best performance, which gained 1.06 BLEU points over the baseline; the model augmented by 1-best and m-best selected paraphrases got an improvement of 0.87 points. A possible reason is that for small training data, the increment of coverage is more important for improving the translation model; while on a larger training corpus, the accuracy of paraphrases plays a more important role.

Notice that the training set, test set and development set in this work are the same as (Marton et al. 2009), which reported a negative result on the full training set. In contrast, our method outperformed the baseline on both small and full training sets.

We further compared the coverage rate of different models on the test set. The result of 29k training set is shown in Table 4. It can be seen that the coverage of 1-gram, 2-grams, 3-grams and 4-grams on the test set is increased by our Ori.+1-PARA. And the method Ori.+1-PARA+ M-PARA has further improved the coverage. It is not surprising that the phrase table of Callison-Burch (CB) has the best coverage on the test set among all the compared models, since their method targets on paraphrasing unknown terms of the test set at run-time. While our method does not target on unknown terms, our goal is enriching the knowledge of SMT system using paraphrases with novel information. Therefore given a specific test set, the unknown terms covered by our method are just a subset of CB's method. However, on such a subset, our method gains significant improvement in translation quality over CB's approach. A possible reason that can explain this is that CB's method only considers the paraphrase probability which controls the adequacy, while in our method, the adequacy, fluency and novelty of the generated paraphrase sentences are well balanced by the SPG framework which can produce paraphrases of better quality.

7 Discussion

We have shown that the SPG framework with an object function of sentence novelty can improve the performance of SMT on both training corpus of small and medium size. Although the experiments are performed on a resource-rich language pair, i.e. English-to-Chinese, the method is portable to other language pairs because our approach is language-independent. No language-specific features are used in the SPG framework. Our proposed method has another advantage of not relying on certain paraphrase resources, and therefore can use any type of training data for paraphrasing. This advantage is important for those resource-poor language pairs.

We further examine the translation results of the baseline and our method. Some examples are shown in Table 5. In row 1 and row 2, the baseline results are improved by our method mainly in the translation of *will take time* and *again and* *again*. The phrases can only be translated word by word in the baseline model. But in our augmented model, the phrases can match complete translation phrases which are extracted from the paraphrase-expanded training data. The paraphrase quality remains an issue with this method. A negative example is shown in row 3, which is caused by a wrong paraphrase substitution *cars* → *park*.

8 Conclusions and Future Work

This paper proposes a novel method for enriching SMT training data by paraphrasing the source-side sentences of the bilingual parallel data through a statistical paraphrase generation framework. Within the framework, a paraphrase model and a language model in the source language are employed to ensure the accuracy of paraphrase. And a proposed object function, named sentence novelty, is used to select paraphrases which have the most novel information for SMT system. Experimental results demonstrate that our method significantly improves the baseline by 1.66 and 1.06 on small and medium size training corpora in terms of BLEU. We have also proved in experiments that our method significantly outperforms the model proposed by Callison-Burch et al. (2006). In the future work, we will plan to test the effectiveness of our method on a large-scale corpus.

Acknowledgement

This work was supported by National Natural Science Foundation of China (NSFC) via grant 61073126, 61133012, 60803093.

References

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of ACL*.

Francis Bond, Eric Nichols, Darren Scott Appling, and Michael Paul. 2008. Improving Statistical Machine Translation by Paraphrasing the Training Data. In *Proceedings of the IWSLT*, pages 150–157.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved Statistical Machine Translation Using Paraphrases. In *Proceedings of NAACL*, pages 17-24.

Jinhua Du, Jie Jiang, Andy Way. 2010. Facilitating Translation Using Source Language Paraphrase Lattices. In *Proceedings of EMNLP*, pages 420-429.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proceedings of HLT/NAACL*, pages 48–54

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP*, pages 388-395.

Philipp Koehn, Hieu Hoang, Alexandra 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 ACL Demo and Poster Sessions*, pages 177–180.

Yuval Marton, Chris Callison-Burch, and Philip Resnik. 2009. Improved Statistical Machine Translation Using Monolingually-Dervied Paraphrases. In *Proceedings of EMNLP*, pages 381-390.

Shachar Mirkin, Lucia Specia, Nicola Cancedda, Ido Dagan, Marc Dymetman, Idan Szpektor. 2009. Source-Language Entailment Modeling for Translation Unknown Terms. In *Proceedings of ACL*, pages 791-799.

Preslav Nakov. 2008. Improved Statistical Machine Translation Using Monolingual Paraphrases. In *Proceedings of ECAI*, pages 338-342.

Fanz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of ACL*, pages 160-167.

Takashi Onishi, Masao Utiyama, Eiichiro Sumita. 2010. Paraphrase Lattice for Statistical Machine Translation. In *Proceedings of ACL*, pages 1-5.

Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL*, pages 311-318.

Matthew Snover, Bonnie J. Dorr, Richard Schwartz, John Makhoul, Linnea Micciulla, and Ralph Weischedel. 2005. A study of translation error rate with targeted human annotation. *Technical Report LAMP-TR-126, CS-TR-4755, UMIACS-TR-2005-58*, University of Maryland, July, 2005.

Shiqi Zhao, Xiang Lan, Ting Liu, and Sheng Li. 2009. Application-driven Statistical Paraphrase Generation. *In Proceedings of ACL*, pages 834-842.