Improve SMT Quality with Automatically Extracted Paraphrase Rules

Wei He[1], **Hua Wu**[2], **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
{wu_hua,wanghaifeng}@baidu.com

Abstract

We propose a novel approach to improve SMT via paraphrase rules which are automatically extracted from the bilingual training data. Without using extra paraphrase resources, we acquire the rules by comparing the source side of the parallel corpus with the target-to-source translations of the target side. Besides the word and phrase paraphrases, the acquired paraphrase rules mainly cover the structured paraphrases on the sentence level. These rules are employed to enrich the SMT inputs for translation quality improvement. The experimental results show that our proposed approach achieves significant improvements of 1.6~3.6 points of BLEU in the oral domain and 0.5~1 points in the news domain.

1 Introduction

The translation quality of the SMT system is highly related to the coverage of translation models. However, no matter how much data is used for training, it is still impossible to completely cover the unlimited input sentences. This problem is more serious for online SMT systems in real-world applications. Naturally, a solution to the coverage problem is to bridge the gaps between the input sentences and the translation models, either from the input side, which targets on rewriting the input sentences to the MT-favored expressions, or from the side of translation models, which tries to enrich the translation models to cover more expressions.

In recent years, paraphrasing has been proven useful for improving SMT quality. The proposed methods can be classified into two categories according to the paraphrase targets: (1) enrich translation models to cover more bilingual expressions; (2) paraphrase the input sentences to reduce OOVs or generate multiple inputs. In the first category, He et al. (2011), Bond et al. (2008) and Nakov (2008) enriched the SMT models via paraphrasing the training corpora. Kuhn et al. (2010) and Max (2010) used paraphrases to smooth translation models. For the second category, previous studies mainly focus on finding translations for unknown terms using phrasal paraphrases. 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 OOVs with entailments and paraphrases acquired from WordNet. Onishi et al. (2010) and Du et al. (2010) use phrasal paraphrases to build a word lattice to get multiple input candidates. In the above methods, only word or phrasal paraphrases are used for input sentence rewriting. No structured paraphrases on the sentence level have been investigated. However, the information in the sentence level is very important for disambiguation. For example, we can only substitute *play* with *drama* in a context related to *stage* or *theatre*. Phrasal paraphrase substitutions can hardly solve such kind of problems.

In this paper, we propose a method that rewrites

This work was done when the first author was visiting Baidu.
*Correspondence author: tliu@ir.hit.edu.cn

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 979–987,
Jeju, Republic of Korea, 8-14 July 2012.

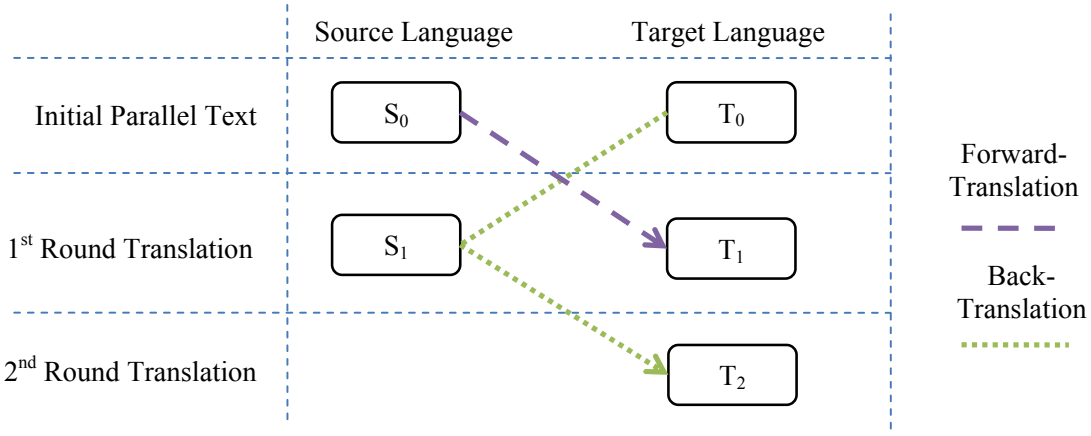

Figure 1: Procedure of Forward-Translation and Back-Translation.

the input sentences of the SMT system using automatically extracted paraphrase rules which can capture structures on sentence level in addition to paraphrases on the word or phrase level. Without extra paraphrase resources, a novel approach is proposed to acquire paraphrase rules from the bilingual training corpus based on the results of Forward-Translation and Back-Translation. The rules target on rewriting the input sentences to an MT-favored expression to ensure a better translation. The paraphrase rules cover all kinds of paraphrases on the word, phrase and sentence levels, enabling structure reordering, word or phrase insertion, deletion and substitution. The experimental results show that our proposed approach achieves significant improvements of 1.6~3.6 points of BLEU in the oral domain and 0.5~1 points in the news domain.

The remainder of the paper is organized as follows: Section 2 makes a comparison between the Forward-Translation and Back-Translation. Section 3 introduces our methods that extract paraphrase rules from the bilingual corpus of SMT. Section 4 describes the strategies for constructing word lattice with paraphrase rules. The experimental results and some discussions are presented in Section 5 and Section 6. Section 7 compares our work to the previous researches. Finally, Section 8 concludes the paper and suggests directions for future work.

2 Forward-Translation vs. Back-Translation

The Back-Translation method is mainly used for automatic MT evaluation (Rapp 2009). This approach is very helpful when no target language reference is available. The only requirement is that the MT system needs to be bidirectional. The procedure includes translating a text into certain foreign language with the MT system (Forward-Translation), and translating it back into the original language with the same system (Back-Translation). Finally the translation quality of Back-Translation is evaluated by using the original source texts as references.

Sun et al. (2010) reported an interesting phenomenon: given a bilingual text, the Back-Translation results of the target sentences is better than the Forward-Translation results of the source sentences. Clearly, let (S_0, T_0) be the initial pair of bilingual text. A source-to-target translation system *SYS_ST* and a target-to-source translation system *SYS_TS* are trained using the bilingual corpus. $FT(\cdot)$ is a Forward-Translation function, and $BT(\cdot)$ is a function of Back-Translation which can be deduced with two rounds of translations: $BT(s) = SYS\_TS(SYS\_ST(S))$. In the first round of translation, S_0 and T_0 are fed into *SYS_ST* and *SYS_TS*, and we get T_1 and S_1 as translation results. In the second round, we translate S_1 back into the target side with *SYS_ST*, and get the translation T_2. The procedure is illustrated in Figure 1, which can also formally be described as:

1. $T_1 = FT(S_0) = SYS\_ST(S_0)$.
2. $T_2 = BT(T_0)$, which can be decomposed into two steps: $S_1 = SYS\_TS(T_0)$, $T_2 = SYS\_ST(S_1)$.

Using T_0 as reference, an interesting result is reported in Sun et al. (2010) that T_2 achieves a higher score than T_1 in automatic MT evaluation. This outcome is important because T_2 is translated

No.	LHS	RHS
1	乘坐/ride X_1 公共汽车/bus	乘坐/ride X_1 巴士/bus
2	在/at X_1 处/location 向左拐/turn left	向左拐/turn left 在/at X_1 处/location
3	把/NULL X_1 给/give 我/me	给/give 我/me X_1
4	从/from X_1 到/to X_2 要/need 多长/how long 时间/time	要/need 花/spend 多长/how long 时间/time 从/from X_1 到/to X_2

Table 1: Examples of Chinese Paraphrase rules, together with English translations for every word

from a machine-generated text S_1, but T_1 is translated from a human-write text S_0. Why the machine-generated text results in a better translation than the human-write text? Two possible reasons may explain this phenomenon: (1) in the first round of translation $T_0 \rightarrow S_1$, some target word orders are reserved due to the reordering failure, and these reserved orders lead to a better result in the second round of translation; (2) the text generated by an MT system is more likely to be matched by the reversed but homologous MT system.

Note that all the texts of S_0, S_1, S_2, T_0 and T_1 are sentence aligned because the initial parallel corpus (S_0, T_0) is aligned in the sentence level. The aligned sentence pairs in (S_0, S_1) can be considered as paraphrases. Since S_1 has some MT-favored structures which may result in a better translation, an intuitive idea is whether we can learn these structures by comparing S_1 with S_0. This is the main assumption of this paper. Taking (S_0, S_1) as paraphrase resource, we propose a method that automatically extracts paraphrase rules to capture the MT-favored structures.

3 Extraction of Paraphrase Rules

3.1 Definition of Paraphrase Rules

We define a paraphrase rule as follows:

1. A paraphrase rule consists of two parts, left-hand-side (LHS) and right-hand-side (RHS). Both of LHS and RHS consist of non-terminals (slot) and terminals (words).
2. LHS must start/end with a terminal.
3. There must be at least one terminal between two non-terminals in LHS.

A paraphrase rule in the format of:

LHS → RHS

which means the words matched by LHS can be paraphrased to RHS. Taking Chinese as a case study, some examples of paraphrase rules are shown in Table 1.

3.2 Selecting Paraphrase Sentence Pairs

Following the methods in Section 2, the initial bilingual corpus is (S_0, T_0). We train a source-to-target PBMT system (*SYS_ST*) and a target-to-source PBMT system (*SYS_TS*) on the parallel corpus. Then a Forward-Translation is performed on S_0 using *SYS_ST*, and a Back-Translation is performed on T_0 using *SYS_TS* and *SYS_ST*. As mentioned above, the detailed procedure is: $T_1 = SYS\_ST(S_0)$, $S_1 = SYS\_TS(T_0)$, $T_2 = SYS\_ST(S_1)$. Finally we compute BLEU (Papineni et al. 2002) score for every sentence in T_2 and T_1, using the corresponding sentence in T_0 as reference. If the sentence in T_2 has a higher BLEU score than the aligned sentence in T_1, the corresponding sentences in S_0 and S_1 are selected as candidate paraphrase sentence pairs, which are used in the following steps of paraphrase extractions.

3.3 Word Alignments Filtering

We can construct word alignment between S_0 and S_1 through T_0. On the initial corpus of (S_0, T_0), we conduct word alignment with Giza++ (Och and Ney, 2000) in both directions and then apply the grow-diag-final heuristic (Koehn et al., 2005) for symmetrization. Because S_1 is generated by feeding T_0 into the PBMT system *SYS_TS*, the word alignment between T_0 and S_1 can be acquired from the verbose information of the decoder.

The word alignments of S_0 and S_1 contain noises which are produced by either wrong alignment of GIZA++ or translation errors of *SYS_TS*. To ensure the alignment quality, we use some heuristics to filter the alignment between S_0 and S_1:

1. If two identical words are aligned in S_0 and S_1, then remove all the other links to the two words.

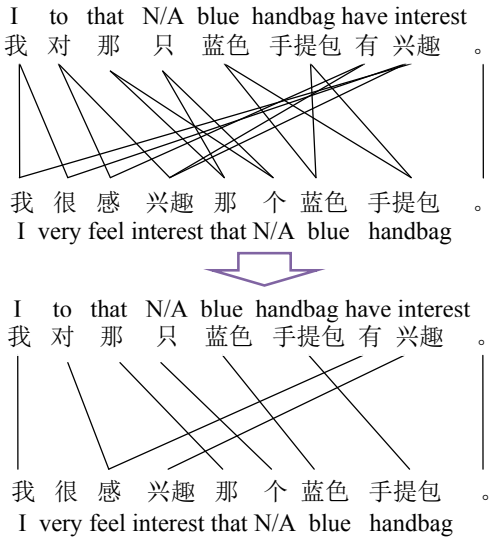

Figure 2: Example for Word Alignment Filtration

2. Stop words (including some function words and punctuations) can only be aligned to either stop words or null.

Figure 2 illustrates an example of using the heuristics to filter alignment.

3.4 Extracting Paraphrase Rules

From the word-aligned sentence pairs, we then extract a set of rules that are consistent with the word alignments. We use the rule extracting methods of Chiang (2005). Take the sentence pair in Figure 2 as an example, two initial phrase pairs PP_1 = "那 只 蓝色 手提包 ||| 那 个 蓝色 手提包" and PP_2 = "对 那 只 蓝色 手提包 有 兴趣 ||| 很 感 兴趣 那 个 蓝色 手提包" are identified, and PP_1 is contained by PP_2, then we could form the rule:

对 X_1 有 兴趣 → 很 感 兴趣 X_1
to have interest very feel interest

4 Paraphrasing the Input Sentences

The extracted paraphrase rules aim to rewrite the input sentences to an MT-favored form which may lead to a better translation. However, it is risky to directly replace the input sentence with a paraphrased sentence, since the errors in automatic paraphrase substitution may jeopardize the translation result seriously. To avoid such damage, for a given input sentence, we first transform all paraphrase rules that match the input sentences to phrasal paraphrases, and then build a word lattice

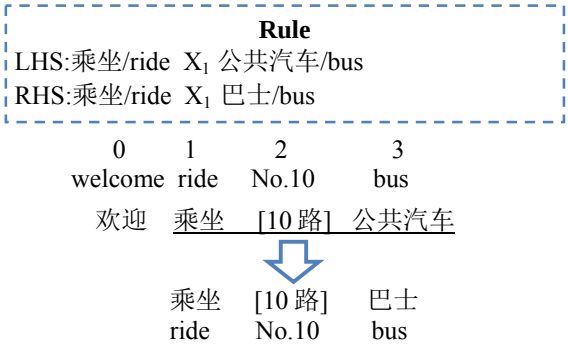

Figure 3: Example for Applying Paraphrase Rules

for SMT decoder using the phrasal paraphrases. In this case, the decoder can search for the best result among all the possible paths.

The input sentences are first segmented into sub-sentences by punctuations. Then for each sub-sentence, the matched paraphrase rules are ranked according to: (1) the number of matched words; (2) the frequency of the paraphrase rule in the training data. Actually, the ranking strategy tends to select paraphrase rules that have more matched words (therefore less ambiguity) and higher frequency (therefore more reliable).

4.1 Applying Paraphrase Rules

Given an input sentence S and a paraphrase rule R $<R_{LHS}, R_{RHS}>$, if S matches R_{LHS}, then the matched part can be replaced by R_{RHS}. An example for applying the paraphrase rules is illustrated in Figure 3.

From Figure 3, we can see that the words of position 1~3 are replaced to "乘坐 10 路 巴士". Actually, only the words at position 3 and 4 are paraphrased to the word "巴士", other words are left unchanged. Therefore, we can use a triple, *<MIN_RP_TEXT, COVER_START, COVER_LEN>* (<巴士, 3, 1> in this example) to denote the paraphrase rule, which means the minimal text to replace is "巴士", and the paraphrasing starts at position 3 and covers 1 words.

In this manner, all the paraphrase rules matched for a certain sentence can be converted to the format of *<MIN_RP_TEXT, COVER_START, COVER_LEN>*, which can also be considered as phrasal paraphrases. Then the methods of building phrasal paraphrases into word lattice for SMT inputs can be used in our approaches.

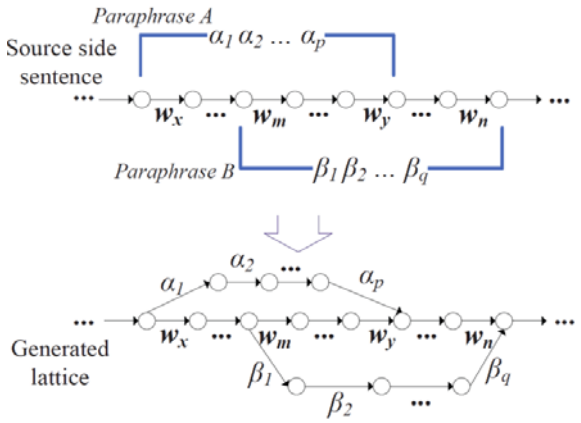

Figure 4: An example of lattice-based paraphrases for an input sentence.

4.2 Construction of Paraphrase Lattice

Given an input sentence, all the matched paraphrase rules are converted to phrasal paraphrases first. Then we build the phrasal paraphrases into word lattices using the methods proposed by Du et al. (2010). The construction process takes advantage of the correspondence between detected phrasal paraphrases and positions of the original words in the input sentence, and then creates extra edges in the lattices to allow the decoder to consider paths involving the paraphrase words. An example is illustrated in Figure 4: given a sequence of words $\{w_1,\ldots,w_N\}$ as the input, two phrases $\alpha = \{\alpha_1,\ldots\alpha_p\}$ and $\beta = \{\beta_1,\ldots, \beta_q\}$ are detected as paraphrases for $P_1 = \{w_x,\ldots, w_y\}$ $(1 \le x \le y \le N)$ and $P_2 = \{w_m,\ldots,w_n\}$ $(1 \le m \le n \le N)$ respectively. The following steps are taken to transform them into word lattices:

1. Transform the original source sentence into word lattice. N + 1 nodes (θ_k, $0 \le k \le N$) are created, and N edges labeled with w_i $(1 \le i \le N)$ are generated to connect them sequentially.
2. Generate extra nodes and edges for each of the paraphrases. Taking α as an example, firstly, $p - 1$ nodes are created, and then p edges labeled with α_j $(1 \le j \le p)$ are generated to connect node θ_{x-1}, p-1 nodes and θ_{y-1}.

Via step 2, word lattices are generated by adding new nodes and edges coming from paraphrases.

4.3 Weight Estimation

The weights of new edges in the lattices are estimated by an empirical method base on ranking positions. Following Du et al. (2010), supposing that $E = \{e_1,\ldots,e_k\}$ are a set of new edges constructed from k paraphrase rules, which are sorted in a descending order. Then the weight for an edge e_i is calculated as:

$$w(e_i) = \frac{1}{k+i} \quad (1 \le i \le k)$$

where k is a predefined tradeoff parameter between decoding speed and the number of potential paraphrases being considered.

5 Experiments

5.1 Experimental Data

In our experiments, we used Moses (Koehn et al., 2007) as the baseline system which can support lattice decoding. The alignment was obtained using GIZA++ (Och and Ney, 2003) and then we symmetrized the word alignment using the grow-diag-final heuristic. Parameters were tuned using Minimum Error Rate Training (Och, 2003). To comprehensively evaluate the proposed methods in different domains, two groups of experiments were carried out, namely, the oral group (G_{oral}) and the news group (G_{news}). The experiments were conducted in both Chinese-English and English-Chinese directions for the oral group, and Chinese-English direction for the news group. The English sentences were all tokenized and lowercased, and the Chinese sentences were segmented into words by Language Technology Platform (LTP)[1]. We used SRILM[2] for the training of language models (5-gram in all the experiments). The metrics for automatic evaluation were BLEU[3] and TER[4] (Snover et al., 2005).

The detailed statistics of the training data in G_{oral} are showed in Table 2. For the bilingual corpus, we used the BTEC and PIVOT data of IWSLT 2008, HIT corpus[5] and other Chinese LDC (CLDC)

[1] http://ir.hit.edu.cn/ltp/
[2] http://www.speech.sri.com/projects/srilm/
[3] ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v13a.pl
[4] http://www.umiacs.umd.edu/~snover/terp/
[5] The HIT corpus contains the CLDC Olympic corpus (2004-863-008) and the other HIT corpora available at http://mitlab.hit.edu.cn/index.php/resources/29-the-resource/111-share-bilingual-corpus.html.

Corpus	#Sen. pairs	#Ch. words	#En words
BETC	19,972	174k	190k
PIVOT	20,000	162k	196k
HIT	80,868	788k	850k
CLDC	190,447	1,167k	1,898k
Tanaka	149,207	-	1,375k

Table 2: Statistics of training data in G_{oral}

	Corpus	#Sen.	#Ref.
develop	CSTAR03 test set	506	16
	IWSLT06 dev set	489	7
test	IWSLT04 test set	500	16
	IWSLT05 test set	506	16
	IWSLT06 test set	500	7
	IWSLT07 test set	489	6

Table 3: Statistics of test/develop sets in G_{oral}

	Corpus	#Sen.	#Ref.
develop	NIST 2002	878	10
	NIST 2005	1,082	4
test	NIST 2004	1,788	5
	NIST 2006	1,664	4
	NIST 2008	1,357	4

Table 4: Statistics of test/develop sets in G_{news}

corpora, including the Chinese-English Sentence Aligned Bilingual Corpus (CLDC-LAC-2003-004) and the Chinese-English Parallel Corpora (CLDC-LAC-2003-006). We trained a Chinese language model for the E-C translation on the Chinese part of the bi-text. For the English language model of C-E translation, an extra corpus named Tanaka was used besides the English part of the bilingual corpora. For testing and developing, we used six Chinese-English development corpora of IWSLT 2008. The statistics are shown in Table 3.

In detail, we chose CSTAR03-test and IWSLT06-dev as the development set; and used IWSLT04-test, IWSLT05-test, IWSLT06-dev and IWSLT07-test for testing. For English-Chinese evaluation, we used IWSLT English-Chinese MT evaluation 2005 as the test set. Due to the lacking of development set, we did not tune parameters on English-Chinese side, instead, we just used the default parameters of Moses.

In the experiments of the news group, we used the Sinorama and FBIS corpora (LDC2005T10 and LDC2003E14) for bilingual corpus. After tokenization and filtering, this bilingual corpus contained 319,694 sentence pairs (7.9M tokens on Chinese side and 9.2M tokens on English side). We trained a 5-gram language model on the English side of the bi-text. The system was tested using the Chinese-English MT evaluation sets of NIST 2004, NIST 2006 and NIST 2008. For development, we used the Chinese-English MT evaluation sets of NIST 2002 and NIST 2005. Table 4 shows the statistics of test/development sets used in the news group.

5.2 Results

We extract both Chinese and English rules in G_{oral}, and Chinese paraphrase rules in G_{news} by comparing the results of Forward-Translation and Back-Translation as described in Section 3. During the extraction, some heuristics are used to ensure the quality of paraphrase rules. Take the extraction of Chinese paraphrase rules in G_{oral} as a case study. Suppose (C_0, E_0) are the initial bilingual corpus of G_{oral}. A Chinese-English and an English-Chinese MT system are trained on (C_0, E_0). We perform Back-Translation on E_0 ($E_0 \xrightarrow{E\ to\ C} C_1 \xrightarrow{C\ to\ E} E_2$), and Forward-Translation on C_0 ($C_0 \xrightarrow{C\ to\ E} E_1$). Suppose e_{1i} and e_{2i} are two aligned sentences in E_1 and E_2, c_{0i} and c_{1i} are the corresponding sentences in C_0 and C_1. (c_{0i}, c_{1i}) are selected for the extraction of paraphrase rules if two conditions are satisfied: (1) BLEU(e_{2i}) – BLEU(e_{1i}) > θ_1, and (2) BLEU(e_{2i}) > θ_2, where BLEU(·) is a function for computing BLEU score; θ_1 and θ_2 are thresholds for balancing the rules number and the quality of paraphrase rules. In our experiment, θ_1 and θ_2 are empirically set to 0.1 and 0.3.

As a result, we extract 912,625 Chinese and 1,116,375 English paraphrase rules for G_{oral}, and for G_{news} the number of Chinese paraphrase rules is 2,877,960. Then we use the extracted paraphrase rules to improve SMT by building word lattices for the input sentences.

The Chinese-English experimental results of G_{oral} and G_{news} are shown in Table 5 and Table 6, respectively. It can be seen that our method outperforms the baselines in both oral and news domains. Our system gains significant improvements of 1.6~3.6 points of BLEU in the oral domain, and 0.5~1 points of BLEU in the news domain. Figure 5 shows the effect of considered paraphrases (k) in the step of building

Model	BLEU				TER			
	iwslt 04	iwslt 05	iwslt 06	iwslt 07	iwslt 04	iwslt 05	iwslt 06	iwslt 07
baseline	0.5353	0.5887	0.2765	0.3977	0.3279	0.2874	0.5559	0.4390
para. improved	**0.5712**	**0.6107**	**0.2924**	**0.4193**	**0.3055**	**0.2722**	**0.5374**	**0.4217**

Table 5: Experimental results of G_{oral} in Chinese-English direction

Model	BLEU			TER		
	nist 04	nist 06	nist 08	nist 04	nist 06	nist 08
baseline	0.2795	0.2389	0.1933	0.6554	0.6515	0.6652
para. improved	**0.2891**	**0.2485**	**0.1978**	**0.6451**	**0.6407**	**0.6582**

Table 6: Experimental results of G_{news} in Chinese-English direction

model	IWSLT 2005	
	BLEU	TER
baseline	0.4644	0.4164
para. improved	**0.4853**	**0.3883**

Table 7: Experimental results of G_{oral} in English-Chinese direction

trans. / para.	improve	comparable	worsen	total
correct	36	20	4	60
incorrect	1	9	14	24

Table 8: Human analysis of the paraphrasing results in IWSLT 2007 CE translation

word lattices. The result of English-Chinese experiments in G_{oral} is shown in Table 7.

6 Discussion

We make a detailed analysis on the Chinese-English translation results that are affected by our paraphrase rules. The aim is to investigate what kinds of paraphrases have been captured in the rules. Firstly the input path is recovered from the translation results according to the tracing information of the decoder, and therefore we can examine which path is selected by the SMT decoder from the paraphrase lattice. A human annotator is asked to judge whether the recovered paraphrase sentence keeps the same meaning as the original input. Then the annotator compares the baseline translation with the translations proposed by our approach. The analysis is carried out on the IWSLT 2007 Chinese-English test set, 84 out of 489 input sentences have been affected by paraphrases, and the statistic of human evaluation is shown in Table 8.

It can be seen in Table 8 that the paraphrases achieve a relatively high accuracy, 60 (71.4%)

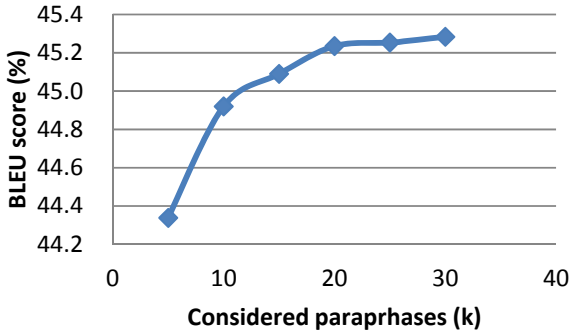

Figure 5: Effect of considered paraphrases (k) on BLEU score

paraphrased sentences retain the same meaning, and the other 24 (28.6%) are incorrect. Among the 60 correct paraphrases, 36 sentences finally result in an improved translation. We further analyze these paraphrases and the translation results to investigate what kinds of transformation finally lead to the translation quality improvement. The paraphrase variations can be classified into four categories:

(1) Reordering: The original source sentences are reordered to be similar to the order of the target language.

(2) Word substitution: A phrase with multi-word translations is replaced by a phrase with a single-word translation.

(3) Recovering omitted words: Ellipsis occurs frequently in spoken language. Recovering the omitted words often leads to a better translation.

(4) Removing redundant words: Mostly, translating redundant words may confuse the SMT system and would be unnecessary. Removing redundant words can mitigate this problem.

Cate.	Num	Original Sentence/Translation	Paraphrased Sentence/Translation
(1)	11	**香烟/cigarette 可以/can 免税/duty-free 带 /take 多少/how much** 支/N/A ? what a cigarette can i take duty-free ?	**多少/how much 香烟/cigarettes 可以/can 免税 /duty-free 带/take** 支/N/A ? how many cigarettes can i take duty-free one ?
(2)	18	你/you 有/have **多久/how long 的/N/A** 教学/teaching 经验/experience ? you have how long teaching experience ?	你/you 有/have **多少/how much** 教学/teaching 经验/experience ? how much teaching experience you have ?
(3)	10	需要/need 押金/deposit 吗/N/A ? you need a deposit ?	**你/you** 需要/need 押金/deposit 吗/N/A ? do you need a deposit ?
(4)	4	戒指/ring 掉/fall 进/into 洗脸池/washbasin **里/in** 了/N/A 。 ring off into the washbasin is in .	戒指/ring 掉/fall 进/into 洗脸池/washbasin 了 /N/A 。 ring off into the washbasin .

Table 9: Examples for classification of paraphrase rules

Four examples for category (1), (2), (3) and (4) are shown in Table 9, respectively. The numbers in the second column indicates the number of the sentences affected by the rules, among the 36 sentences with improved paraphrasing and translation. A sentence can be classified into multiple categories. Except category (2), the other three categories cannot be detected by the previous approaches, which verify our statement that our rules can capture structured paraphrases on the sentence level in addition to the paraphrases on the word or phrase level.

Not all the paraphrased results are correct. Sometimes an ill paraphrased sentence can produce better translations. Take the first line of Table 9 as an example, the paraphrased sentence "多少/How many 香烟/cigarettes 可以/can 免税/duty-free 带 /take 支/NULL" is not a fluent Chinese sentence, however, the rearranged word order is closer to English, which finally results in a much better translation.

7 Related Work

Previous studies on improving SMT using paraphrase rules focus on hand-crafted rules. Nakov (2008) employs six rules for paraphrasing the training corpus. 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.

A disadvantage for hand-crafted paraphrase rules is that it is language dependent. In contrast, our method that automatically extracted paraphrase rules from bilingual corpus is flexible and suitable for any language pairs.

Our work is similar to Sun et al. (2010). Both tried to capture the MT-favored structures from bilingual corpus. However, a clear difference is that Sun et al. (2010) captures the structures implicitly by training an MT system on (S_0, S_1) and "translates" the SMT input to an MT-favored expression. Actually, the rewriting process is considered as a black box in Sun et al. (2010). In this paper, the MT-favored expressions are captured explicitly by automatically extracted paraphrase rules. The advantages of the paraphrase rules are: (1) Our method can explicitly capture the structure information in the sentence level, enabling global reordering, which is impossible in Sun et al. (2010). (2) For each rule, we can control its quality automatically or manually.

8 Conclusion

In this paper, we propose a novel method for extracting paraphrase rules by comparing the source side of bilingual corpus to the target-to-source translation of the target side. The acquired paraphrase rules are employed to enrich the SMT inputs, which target on rewriting the input sentences to an MT-favored form. The paraphrase rules cover all kinds of paraphrases on the word, phrase and sentence levels, enabling structure reordering, word or phrase insertion, deletion and substitution. Experimental results show that the paraphrase rules can improve SMT quality in both the oral and news domains. The manual investigation on oral translation results indicate that the paraphrase rules capture four kinds of MT-favored transformation to ensure translation quality improvement.

Acknowledgement

This work was supported by National Natural Science Foundation of China (NSFC) (61073126, 61133012), 863 High Technology Program (2011AA01A207).

References

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.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In Proceedings of ACL, pages 263–270.

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

Wei He, Shiqi Zhao, Haifeng Wang and Ting Liu. 2011. Enriching SMT Training Data via Paraphrasing. In Proceedings of IJCNLP, pages 803-810.

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, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. 2005. Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation. In Proceedings of IWSLT.

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.

Roland Kuhn, Boxing Chen, George Foster and Evan Stratford. 2010. Phrase Clustering for Smoothing TM Probabilities-or, How to Extract Paraphrases from Phrase Tables. In Proceedings of COLING, pages 608–616.

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

Aurélien Max. 2010. Example-Based Paraphrasing for Improved Phrase-Based Statistical Machine TranslationIn Proceedings of EMNLP, pages 656-666.

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.

Franz Josef Och and Hermann Ney. 2000. Improved Statistical Alignment Models. In Proceedings of ACL, pages 440-447.

Franz 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.

Reinhard Rapp. 2009. The Back-translation Score: Automatic MT Evaluation at the Sentence Level without Reference Translations. In Proceedings of ACL-IJCNLP, pages 133-136.

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.

Yanli Sun, Sharon O'Brien, Minako O'Hagan and Fred Hollowood. 2010. A Novel Statistical Pre-Processing Model for Rule-Based Machine Translation System. In Proceedings of EAMT, 8pp.