Leveraging Multiple MT Engines for Paraphrase Generation

Shiqi Zhao†‡, Haifeng Wang†, Xiang Lan‡, and Ting Liu‡
†Baidu Inc.
‡HIT Center for Information Retrieval, Harbin Institute of Technology
{zhaoshiqi, wanghaifeng}@baidu.com,
{xlan, tliu}@ir.hit.edu.cn

Abstract

This paper proposes a method that leverages multiple machine translation (MT) engines for paraphrase generation (PG). The method includes two stages. Firstly, we use a multi-pivot approach to acquire a set of candidate paraphrases for a source sentence S. Then, we employ two kinds of techniques, namely the selection-based technique and the decoding-based technique, to produce a best paraphrase T for S using the candidates acquired in the first stage. Experimental results show that: (1) The multi-pivot approach is effective for obtaining plenty of valuable candidate paraphrases. (2) Both the selection-based and decoding-based techniques can make good use of the candidates and produce high-quality paraphrases. Moreover, these two techniques are complementary. (3) The proposed method outperforms a state-of-the-art paraphrase generation approach.

1 Introduction

This paper addresses the problem of paraphrase generation (PG), which seeks to generate paraphrases for sentences. PG is important in many natural language processing (NLP) applications. For example, in machine translation (MT), a sentence can be paraphrased so as to make it more translatable (Zhang and Yamamoto, 2002; Callison-Burch et al., 2006). In question answering (QA), a question can be paraphrased to improve the coverage of answer extraction (Duboue and Chu-Carroll, 2006; Riezler et al., 2007). In natural language generation (NLG), paraphrasing can help to increase the expressive power of the NLG systems (Iordanskaja et al., 1991).

In this paper, we propose a novel PG method. For an English sentence S, the method first acquires a set of candidate paraphrases with a multi-pivot approach, which uses MT engines to automatically translate S into multiple pivot languages and then translate them back into English. Furthermore, the method employs two kinds of techniques to produce a best paraphrase T for S using the candidates, i.e., the selection-based and decoding-based techniques. The former selects a best paraphrase from the candidates based on Minimum Bayes Risk (MBR), while the latter trains a MT model using the candidates and generates paraphrases with a MT decoder.

We evaluate our method on a set of 1182 English sentences. The results show that: (1) although the candidate paraphrases acquired by MT engines are noisy, they provide good raw materials for further paraphrase generation; (2) the selection-based technique is effective, which results in the best performance; (3) the decoding-based technique is promising, which can generate paraphrases that are different from the candidates; (4) both the selection-based and decoding-based techniques outperform a state-of-the-art approach SPG (Zhao et al., 2009).

2 Related Work

2.1 Methods for Paraphrase Generation

MT-based method is the mainstream method on PG. It regards PG as a monolingual machine translation problem, i.e., "translating" a sentence S into another sentence T in the same language.

Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 1326–1334,
Beijing, August 2010

Quirk et al. (2004) first presented the MT-based method. They trained a statistical MT (SMT) model on a monolingual parallel corpus extracted from comparable news articles and applied the model to generate paraphrases. Their work shows that SMT techniques can be extended to PG. However, its usefulness is limited by the scarcity of monolingual parallel data.

To overcome the data sparseness problem, Zhao et al. (2008a) improved the MT-based PG method by training the paraphrase model using multiple resources, including monolingual parallel corpora, monolingual comparable corpora, bilingual parallel corpora, etc. Their results show that bilingual parallel corpora are the most useful among the exploited resources. Zhao et al. (2009) further improved the method by introducing a *usability sub-model* into the paraphrase model so as to generate varied paraphrases for different applications.

The main disadvantage of the MT-based method is that its performance heavily depends on the fine-grained paraphrases, such as paraphrase phrases and patterns, which provide paraphrase options in decoding. Hence one has to first extract fine-grained paraphrases from various corpora with different methods (Zhao et al., 2008a; Zhao et al., 2009), which is difficult and time-consuming.

In addition to the MT-based method, researchers have also investigated other methods for paraphrase generation, such as the pattern-based methods (Barzilay and Lee, 2003; Pang et al., 2003), thesaurus-based methods (Bolshakov and Gelbukh, 2004; Kauchak and Barzilay, 2006), and NLG-based methods (Kozlowski et al., 2003; Power and Scott, 2005).

2.2 Pivot Approach for Paraphrasing

Bannard and Callison-Burch (2005) introduced the pivot approach to extracting paraphrase phrases from bilingual parallel corpora. Their basic assumption is that two English phrases aligned with the same phrase in a foreign language (also called a pivot language) are potential paraphrases. Zhao et al. (2008b) extended the approach and used it to extract paraphrase patterns. Both of the above works have proved the effectiveness of the pivot approach in paraphrase extraction.

Pivot approach can also be used in paraphrase generation. It generates paraphrases by translating sentences from a source language to one (single-pivot) or more (multi-pivot) pivot languages and then translating them back to the source language. Duboue et al. (2006) first proposed the multi-pivot approach for paraphrase generation, which was specially designed for question expansion in QA. In addition, Max (2009) presented a single-pivot approach for generating sub-sentential paraphrases. A clear difference between our method and the above works is that we propose selection-based and decoding-based techniques to generate high-quality paraphrases using the candidates yielded from the pivot approach.

3 Multi-pivot Approach for Acquiring Candidate Paraphrases

A single-pivot PG approach paraphrases a sentence S by translating it into a pivot language PL with a MT engine MT_1 and then translating it back into the source language with MT_2. In this paper, a single-pivot PG system is represented as a triple (MT_1, PL, MT_2). A multi-pivot PG system is made up of a set of single-pivot systems with various pivot languages and MT engines. Given m pivot languages and n MT engines, we can build a multi-pivot PG system consisting of N ($N \leq n * m * n$) single-pivot ones, where $N = n * m * n$ iff all the n MT engines can perform bidirectional translation between the source and each pivot language.

In this work, we experiment with 6 pivot languages (Table 1) and 3 MT engines (Table 2) in the multi-pivot approach. All the 3 MT engines are off-the-shelf systems, in which Google and Microsoft translators are SMT engines, while Systran translator is a rule-based MT engine. Each MT engine can translate English to all the 6 pivot languages and back to English. We thereby construct a multi-pivot PG system consisting of 54 (3*6*3) single-pivot systems.

The advantages of the multi-pivot PG approach lie in two aspects. First, it effectively makes use of the vast bilingual data and translation rules underlying the MT engines. Second, the approach is simple, which just sends sentences to the online MT engines and gets the translations back.

Source Sentence	he said there will be major cuts in the salaries of high-level civil servants .
(GG, G, MS)	he said there **are significant** cuts in the salaries of high-level **officials** .
(GG, F, GG)	he said there will be **significant** cuts in the salaries of *top civil level* .
(MS, C, MS)	he said **that** there will be *a major senior civil service pay cut* .
(MS, F, ST)	he said there will be **great** cuts in the **wages** of the *high level civils servant* .
(ST, G, GG)	he said **that** there **are** major cuts in the salaries of **senior government officials** .

Table 3: Examples of candidate paraphrases obtained using the multi-pivot approach.

1	French (**F**)	4	Italian (**I**)
2	German (**G**)	5	Portuguese (**P**)
3	Spanish (**S**)	6	Chinese (**C**)

Table 1: Pivot languages used in the approach.

1	Google Translate (**GG**) (translate.google.com)
2	Microsoft Translator (**MS**) (www.microsofttranslator.com)
3	Systran Online Translation (**ST**) (www.systransoft.com)

Table 2: MT engines utilized in the approach.

4 Producing High-quality Paraphrases using the Candidates

Table 3 shows some examples of candidate paraphrases for a sentence. As can be seen, the candidates do provide some correct and useful paraphrase substitutes (in bold) for the source sentence. However, they also contain quite a few errors (in italic) due to the limited translation quality of the MT engines. The problem is even worse when the source sentences get longer and more complicated. Therefore, we need to combine the outputs of the multiple single-pivot PG systems and produce high-quality paraphrases out of them. To this end, we investigate two techniques, namely, the selection-based and decoding-based techniques.

4.1 Selection-based Technique

Given a source sentence S along with a set D of candidate paraphrases $\{T_1, T_2, ..., T_i, ...T_N\}$, the goal of the selection-based technique is to select the best paraphrase $\hat{T}_i$ for S from D. The paraphrase selection technique we propose is based on Minimum Bayes Risk (MBR). In detail, the MBR based technique first measures the quality of each candidate paraphrase $T_i \in D$ in terms of Bayes risk (BR), and then selects the one with the minimum BR as the best paraphrase. In detail, given S, a candidate $T_i \in D$, a reference paraphrase T[1], and a loss function $L(T, T_i)$ that measures the quality of T_i relative to T, we define the Bayes risk as follows:

$$BR(T_i) = E_{P(T,S)}[L(T, T_i)], \qquad (1)$$

where the expectation is taken under the true distribution $P(T, S)$ of the paraphrases. According to (Bickel and Doksum, 1977), the candidate paraphrase that minimizes the Bayes risk can be found as follows:

$$\hat{T}_i = \arg\min_{T_i \in D} \sum_{T \in \mathcal{T}} L(T, T_i) P(T|S), \qquad (2)$$

where $\mathcal{T}$ represents the space of reference paraphrases. In practice, however, the collection of reference paraphrases is not available. We thus construct a set $D' = D \bigcup \{S\}$ to approximate $\mathcal{T}$[2]. In addition, we cannot estimate $P(T|S)$ in Equation (2), either. Therefore, we make a simplification by assigning a constant c to $P(T|S)$ for each $T \in D'$, which can then be removed:

$$\hat{T}_i = \arg\min_{T_i \in D} \sum_{T \in D'} L(T, T_i). \qquad (3)$$

Equation (3) can be further rewritten using a gain function $G(T, T_i)$ instead of the loss function:

[1] Here we assume that we have the collection of all possible paraphrases of S, which are used as references.

[2] The source sentence S is included in D' based on the consideration that a sentence is allowed to keep unchanged during paraphrasing.

$$\hat{T}_i = \arg\max_{T_i \in D} \sum_{T \in D'} G(T, T_i). \qquad (4)$$

We define the gain function based on BLEU: $G(T, T_i) = BLEU(T, T_i)$. BLEU is a widely used metric in the automatic evaluation of MT (Papineni et al., 2002). It measures the similarity of two sentences by counting the overlapping n-grams (n=1,2,3,4 in our experiments):

$$BLEU(T, T_i) = BP \cdot \exp(\sum_{n=1}^{4} w_n \log p_n(T, T_i)),$$

where $p_n(T, T_i)$ is the n-gram precision of T_i and $w_n = 1/4$. BP (≤ 1) is a brevity penalty that penalizes T_i if it is shorter than T.

In summary, for each sentence S, the MBR based technique selects a paraphrase that is the most similar to all candidates and the source sentence. The underlying assumption is that correct paraphrase substitutes should be common among the candidates, while errors committed by the single-pivot PG systems should be all different. We denote this approach as **S-1** hereafter.

Approaches for comparison. In the experiments, we also design another two paraphrase selection approaches S-2 and S-3 for comparison with S-1.

S-2: S-2 selects the best single-pivot PG system from all the 54 ones. The selection is also based on MBR and BLEU. For each single-pivot PG system, we sum up its gain function values over a set of source sentences (i.e., $\sum_S \sum_{T_S \in D'_S} G(T_S, T_{Si})$). Then we select the one with the maximum gain value as the best single-pivot system. In our experiments, the selected best single-pivot PG system is (ST, P, GG), the candidate paraphrases acquired by which are then returned as the best paraphrases in S-2.

S-3: S-3 is a simple baseline, which just randomly selects a paraphrase from the 54 candidates for each source sentence S.

4.2 Decoding-based Technique

The selection-based technique introduced above has an inherent limitation that it can only select a paraphrase from the candidates. That is to say, it can never produce a perfect paraphrase if all the candidates have some tiny flaws. To solve this problem, we propose the decoding-based technique, which trains a MT model using the candidate paraphrases of each source sentence S and generates a new paraphrase T for S with a MT decoder.

major cuts	high-level civil servants
significant cuts	senior officials
major cuts*	high-level officials
important cuts	senior civil servants
big cuts	
great cuts	

Table 4: Extracted phrase pairs. (*This is called a *self-paraphrase* of the source phrase, which is generated when a phrase keeps unchanged in some of the candidate paraphrases.)

In this work, we implement the decoding-based technique using Giza++ (Och and Ney, 2000) and Moses (Hoang and Koehn, 2008), both of which are commonly used SMT tools. For a sentence S, we first construct a set of parallel sentences by pairing S with each of its candidate paraphrases: $\{(S,T_1),(S,T_2),...,(S,T_N)\}$ ($N = 54$). We then run word alignment on the set using Giza++ and extract aligned phrase pairs as described in (Koehn, 2004). Here we only keep the phrase pairs that are aligned $\geq$3 times on the set, so as to filter errors brought by the noisy sentence pairs. The extracted phrase pairs are stored in a phrase table. Table 4 shows some extracted phrase pairs.

Note that Giza++ is sensitive to the data size. Hence it is interesting to examine if the alignment can be improved by augmenting the parallel sentence pairs. To this end, we have tried augmenting the parallel set for each sentence S by pairing any two candidate paraphrases. In this manner, C_N^2 sentence pairs are augmented for each S. We conduct word alignment using the $(N + C_N^2)$ sentence pairs and extract aligned phrases from the original N pairs. However, we have not found clear improvement after observing the results. Therefore, we do not adopt the augmentation strategy in our experiments.

Using the extracted phrasal paraphrases, we conduct decoding for the sentence S with Moses, which is based on a log-linear model. The default setting of Moses is used, except that the distortion model for phrase reordering is turned off[3]. The language model in Moses is trained using a 9 GB English corpus. We denote the above approach as **D-1** in what follows.

Approach for comparison. The main advantage of the decoding-based technique is that it allows us to customize the paraphrases for different requirements through tailoring the phrase table or tuning the model parameters. As a case study, this paper shows how to generate paraphrases with varied *paraphrase rates*[4].

D-2: The extracted phrasal paraphrases (including self-paraphrases) are stored in a phrase table, in which each phrase pair has 4 scores measuring their alignment confidence (Koehn et al., 2003). Our basic idea is to control the paraphrase rate by tuning the scores of the self-paraphrases. We thus extend D-1 to D-2, which assigns a weight λ ($\lambda > 0$) to the scores of the self-paraphrase pairs. Obviously, if we set $\lambda < 1$, the self-paraphrases will be penalized and the decoder will prefer to generate a paraphrase with more changes. If we set $\lambda > 1$, the decoder will tend to generate a paraphrase that is more similar to the source sentence. In our experiments, we set $\lambda = 0.1$ in D-2.

5 Experimental Setup

Our test sentences are extracted from the parallel reference translations of a Chinese-to-English MT evaluation[5], in which each Chinese sentence c has 4 English reference translations, namely e_1, e_2, e_3, and e_4. We use e_1 as a test sentence to paraphrase and e_2, e_3, e_4 as human paraphrases of e_1 for comparison with the automatically generated paraphrases. We process the test set by manually filtering ill-formed sentences, such as the ungrammatical or incomplete ones. 1182 out of 1357 test sentences are retained after filtering. Statistics show that about half of the test sentences are from news and the other half are from essays. The average length of the test sentences is 34.12 (words).

Score	Adequacy	Fluency
5	All	Flawless English
4	Most	Good English
3	Much	Non-native English
2	Little	Disfluent English
1	None	Incomprehensible

Table 5: Five point scale for human evaluation.

Manual evaluation is used in this work. A paraphrase T of a sentence S is manually scored based on a five point scale, which measures both the "adequacy" (i.e., how much of the meaning of S is preserved in T) and "fluency" of T (See Table 5). The five point scale used here is similar to that in the human evaluation of MT (Callison-Burch et al., 2007). In MT, adequacy and fluency are evaluated separately. However, we find that there is a high correlation between the two aspects, which makes it difficult to separate them. Thus we combine them in this paper.

We compare our method with a state-of-the-art approach SPG[6] (Zhao et al., 2009), which is a statistical approach specially designed for PG. The approach first collects a large volume of fine-grained paraphrase resources, including paraphrase phrases, patterns, and collocations, from various corpora using different methods. Then it generates paraphrases using these resources with a statistical model[7].

6 Experimental Results

We evaluate six approaches, i.e., S-1, S-2, S-3, D-1, D-2 and SPG, in the experiments. Each approach generates a 1-best paraphrase for a test sentence S. We randomize the order of the 6 paraphrases of each S to avoid bias of the raters.

[3] We conduct monotone decoding as previous work (Quirk et al., 2004; Zhao et al., 2008a, Zhao et al., 2009).

[4] The paraphrase rate reflects how different a paraphrase is from the source sentence.

[5] 2008 NIST Open Machine Translation Evaluation: Chinese to English Task.

[6] SPG: Statistical Paraphrase Generation.

[7] We ran SPG under the setting of baseline-2 as described in (Zhao et al., 2009).

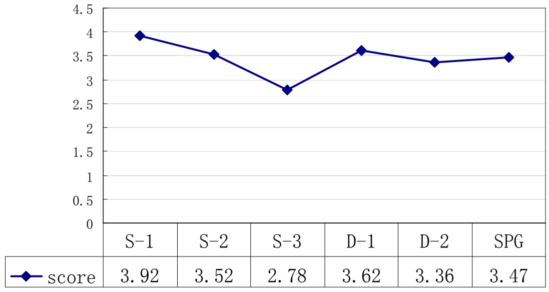


Figure 1: Evaluation results of the approaches.

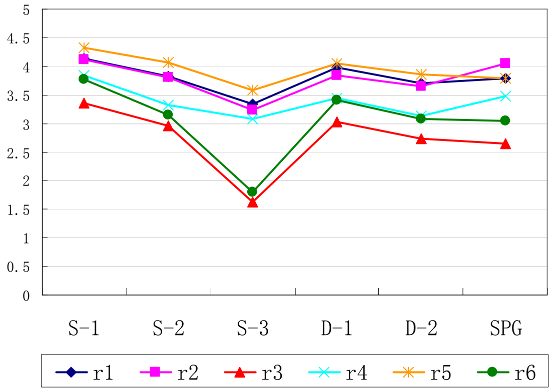


Figure 2: Evaluation results from each rater.

6.1 Human Evaluation Results

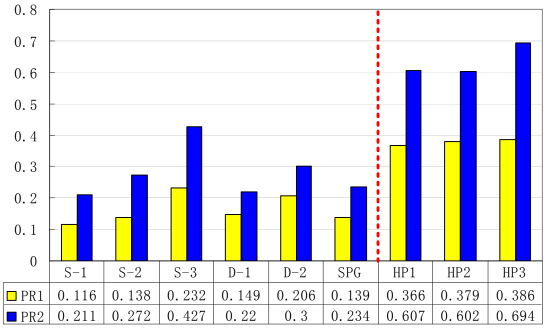
We have 6 raters in the evaluation, all of whom are postgraduate students. In particular, 3 raters major in English, while the other 3 major in computer science. Each rater scores the paraphrases of 1/6 test sentences, whose results are then combined to form the final scoring result. The average scores of the six approaches are shown in Figure 1. We can find that among the selection-based approaches, the performance of S-3 is the worst, which indicates that randomly selecting a paraphrase from the candidates works badly. S-2 performs much better than S-3, suggesting that the quality of the paraphrases acquired with the best single-pivot PG system are much higher than the randomly selected ones. S-1 performs the best in all the six approaches, which demonstrates the effectiveness of the MBR-based selection technique. Additionally, the fact that S-1 evidently outperforms S-2 suggests that it is necessary to extend a single-pivot approach to a multi-pivot one.

To get a deeper insight of S-1, we randomly sample 100 test sentences and manually score all of their candidates. We find that S-1 successfully picks out a paraphrase with the highest score for 72 test sentences. We further analyze the remaining 28 sentences for which S-1 fails and find that the failures are mainly due to the BLEU-based gain function. For example, S-1 sometimes selects paraphrases that have correct phrases but incorrect phrase orders, since BLEU is weak in evaluating phrase orders and sentence structures. In the next step we shall improve the gain function by investigating other features besides BLEU.

In the decoding-based approaches, D-1 ranks the second in the six approaches only behind S-1. We will further improve D-1 in the future rather than simply use Moses in decoding with the default setting. However, the value of D-1 lies in that it enables us to break down the candidates and generate new paraphrases flexibly. The performance decreases when we extend D-1 to D-2 to achieve a larger paraphrase rate. This is mainly because more errors are brought in when more parts of a sentence are paraphrased.

We can also find from Figure 1 that S-1, S-2, and D-1 all get higher scores than SPG, which shows that our method outperforms this state-of-the-art approach. This is more important if we consider that our method is lightweight, which makes no effort to collect fine-grained paraphrase resources beforehand. After observing the results, we believe that the outperformance of our method can be mainly ascribed to the selection-based and decoding-based techniques, since we avoid many errors by voting among the candidates. For instance, an ambiguous phrase may be incorrectly paraphrased by some of the single-pivot PG systems or the SPG approach. However, our method may obtain the correct paraphrase through statistics over all candidates and selecting the most credible one.

The human evaluation of paraphrases is subjective. Hence it is necessary to examine the coherence among the raters. The scoring results from the six raters are depicted in Figure 2. As it can be seen, they show similar trends though the raters have different degrees of strictness.



	S-1	S-2	S-3	D-1	D-2	SPG	HP1	HP2	HP3
PR1	0.116	0.138	0.232	0.149	0.206	0.139	0.366	0.379	0.386
PR2	0.211	0.272	0.427	0.22	0.3	0.234	0.607	0.602	0.694

Figure 3: Paraphrase rates of the approaches.

6.2 Paraphrase Rate

Human evaluation assesses the quality of paraphrases. However, the paraphrase rates cannot be reflected. A paraphrase that is totally transformed from the source sentence and another that is almost unchanged may get the same score. Therefore, we propose two strategies, i.e., PR1 and PR2, to compute the paraphrase rate:

$$PR1(T) = 1 - \frac{OL(S,T)}{L(S)}; \;\; PR2(T) = \frac{ED(S,T)}{L(S)}.$$

Here, PR1 is defined based on word overlapping rate, in which $OL(S,T)$ denotes the number of overlapping words between a paraphrase T and its source sentence S, $L(S)$ denotes the number of words in S. PR2 is defined based on edit distance, in which $ED(S,T)$ denotes the edit distance between T and S. Obviously, PR1 only measures the percentage of words that are changed from S to T, whereas PR2 further takes word order changes into consideration. It should be noted that PR1 and PR2 not only count the correct changes between S and T, but also count the incorrect ones. We compute the paraphrase rate for each of the six approaches by averaging the paraphrase rates over the whole test set. The results are shown in the left part of Figure 3.

On the whole, the paraphrase rates of the approaches are not high. In particular, we can see that the paraphrase rate of D-2 is clearly higher than D-1, which is in line with our intention of designing D-2. We can also see that the paraphrase rate of S-3 is the highest among the approaches. We find it is mainly because the paraphrases generated with S-3 contain quite a lot of errors, which contribute most of the changes.

7 Analysis

7.1 Effectiveness of the Proposed Method

Our analysis starts from the candidate paraphrases acquired with the multi-pivot approach. Actually, the results of S-3 reflect the average quality of the candidate paraphrases. A score of 2.78 (See Figure 1) indicates that the candidates are unacceptable according to the human evaluation metrics. This is in line with our expectation that the automatically acquired paraphrases through a two-way translation are noisy. However, the results of S-1 and D-1 demonstrate that, using the selection-based and decoding-based techniques, we can produce paraphrases of good quality. Especially, S-1 gets a score of nearly 4, which suggests that the paraphrases are pretty good according to our metrics. Moreover, our method outperforms SPG built on pre-extracted fine-grained paraphrases. It shows that our method makes good use of the paraphrase knowledge from the large volume of bilingual data underlying the multiple MT engines.

7.2 How to Choose Pivot Languages and MT Engines in the Multi-pivot Approach

In our experiments, besides the six pivot languages used in the multi-pivot system, we have also tried another five pivot languages, including Arabic, Japanese, Korean, Russian, and Dutch. They are finally abandoned since we find that they perform badly. Our experience on choosing pivot languages is that: (1) a pivot language should be a language whose translation quality can be well guaranteed by the MT engines; (2) it is better to choose a pivot language similar to the source language (e.g., French - English), which is easier to translate; (3) the translation quality of a pivot language should not vary a lot among the MT engines. On the other hand, it is better to choose MT engines built on diverse models and corpora, which can provide different paraphrase options. We plan to employ a syntax-based MT engine in our further experiments besides the currently used phrase-based SMT and rule-based MT engines.

S	he said there will be major cuts in the salaries of high-level civil servants .
S-1	he said **that** there will be **significant** cuts in the salaries of **senior officials** .
S-2	he said there will be major cuts in salaries of *civil servants high level* .
S-3	he said **that** there will be **significant** cuts in the salaries of **senior officials** .
D-1	he said **,** there will be **significant** cuts in salaries of **senior** civil servants .
D-2	he said **,** there will be **significant** cuts in salaries of **senior officials** .
SPG	he said **that** there will be *the main* cuts in the **wages** of high-level civil servants .
HP1	he said there will be a **big salary cut for high-level government employees** .
HP2	he said **salaries of senior public servants would be slashed** .
HP3	he **claimed to implement huge salary cut to senior** civil servants .

Table 6: Comparing the automatically generated paraphrases with the human paraphrases.

7.3 Comparing the Selection-based and Decoding-based Techniques

It is necessary to compare the paraphrases generated via the selection-based and decoding-based techniques. As stated above, the selection-based technique can only select a paraphrase from the candidates, while the decoding-based technique can generate a paraphrase different from all candidates. In our experiments, we find that for about 90% test sentences, the paraphrases generated by the decoding-based approach D-1 are outside the candidates. In particular, we compare the paraphrases generated by S-1 and D-1 and find that, for about 40% test sentences, S-1 gets higher scores than D-1, while for another 21% test sentences, D-1 gets higher scores than S-1[8]. This indicates that the selection-based and decoding-based techniques are complementary. In addition, we find examples in which the decoding-based technique can generate a perfect paraphrase for the source sentence, even if all the candidate paraphrases have obvious errors. This also shows that the decoding-based technique is promising.

7.4 Comparing Automatically Generated Paraphrases with Human Paraphrases

We also analyze the characteristics of the generated paraphrases and compare them with the human paraphrases (i.e., the other 3 reference translations in the MT evaluation, see Section 5, which are denoted as HP1, HP2, and HP3). We find that, compared with the automatically generated paraphrases, the human paraphrases are more complicated, which involve not only phrase replacements, but also structure reformulations and even inferences. Their paraphrase rates are also much higher, which can be seen in the right part of Figure 3. We show the automatic and human paraphrases for the example sentence of this paper in Table 6. To narrow the gap between the automatic and human paraphrases, it is necessary to learn structural paraphrase knowledge from the candidates in the future work.

8 Conclusions and Future Work

We put forward an effective method for paraphrase generation, which has the following contributions. First, it acquires a rich fund of paraphrase knowledge through the use of multiple MT engines and pivot languages. Second, it presents a MBR-based technique that effectively selects high-quality paraphrases from the noisy candidates. Third, it proposes a decoding-based technique, which can generate paraphrases that are different from the candidates. Experimental results show that the proposed method outperforms a state-of-the-art approach SPG.

In the future work, we plan to improve the selection-based and decoding-based techniques. We will try some standard system combination strategies, like confusion networks and consensus decoding. In addition, we will refine our evaluation metrics. In the current experiments, paraphrase correctness (adequacy and fluency) and paraphrase rate are evaluated separately, which seem to be incompatible. We plan to combine them together and propose a uniform metric.

[8] For the rest 39%, S-1 and D-1 get identical scores.

References

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with Bilingual Parallel Corpora. In *Proceedings of ACL*, pages 597-604.

Regina Barzilay and Lillian Lee. 2003. Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment. In *Proceedings of HLT-NAACL*, pages 16-23.

Peter J. Bickel and Kjell A. Doksum. 1977. *Mathematical Statistics: Basic Ideas and Selected Topics*. Holden-Day Inc., Oakland, CA, USA.

Igor A. Bolshakov and Alexander Gelbukh. 2004. Synonymous Paraphrasing Using WordNet and Internet. In *Proceedings of NLDB*, pages 312-323.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz and Josh Schroeder. 2007. (Meta-) Evaluation of Machine Translation. In *Proceedings of ACL-2007 Workshop on Statistical Machine Translation*, pages 136-158.

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

Pablo Ariel Duboue and Jennifer Chu-Carroll. 2006. Answering the Question You Wish They Had Asked: The Impact of Paraphrasing for Question Answering. In *Proceedings of HLT-NAACL*, pages 33-36.

Hieu Hoang and Philipp Koehn. 2008. Design of the Moses Decoder for Statistical Machine Translation. In *Proceedings of ACL Workshop on Software engineering, testing, and quality assurance for NLP*, pages 58-65.

Lidija Iordanskaja, Richard Kittredge, and Alain Polguère. 1991. Lexical Selection and Paraphrase in a Meaning-Text Generation Model. In Cécile L. Paris, William R. Swartout, and William C. Mann (Eds.): Natural Language Generation in Artificial Intelligence and Computational Linguistics, pages 293-312.

David Kauchak and Regina Barzilay. 2006. Paraphrasing for Automatic Evaluation. In *Proceedings of HLT-NAACL*, pages 455-462.

Philipp Koehn. 2004. Pharaoh: a Beam Search Decoder for Phrase-Based Statistical Machine Translation Models: User Manual and Description for Version 1.2.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proceedings of HLT-NAACL*, pages 127-133.

Raymond Kozlowski, Kathleen F. McCoy, and K. Vijay-Shanker. 2003. Generation of single-sentence paraphrases from predicate/argument structure using lexico-grammatical resources. In *Proceedings of IWP*, pages 1-8.

Aurélien Max. 2009. Sub-sentential Paraphrasing by Contextual Pivot Translation. In *Proceedings of the 2009 Workshop on Applied Textual Inference, ACL-IJCNLP 2009*, pages 18-26.

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

Bo Pang, Kevin Knight, and Daniel Marcu. 2003. Syntax-based Alignment of Multiple Translations: Extracting Paraphrases and Generating New Sentences. In *Proceedings of HLT-NAACL*, pages 102-109.

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.

Richard Power and Donia Scott. 2005. Automatic generation of large-scale paraphrases. In *Proceedings of IWP*, pages 73-79.

Chris Quirk, Chris Brockett, and William Dolan. 2004. Monolingual Machine Translation for Paraphrase Generation. In *Proceedings of EMNLP*, pages 142-149.

Stefan Riezler, Alexander Vasserman, Ioannis Tsochantaridis, Vibhu Mittal and Yi Liu. 2007. Statistical Machine Translation for Query Expansion in Answer Retrieval. In *Proceedings of ACL*, pages 464-471.

Yujie Zhang and Kazuhide Yamamoto. 2002. Paraphrasing of Chinese Utterances. In *Proceedings of COLING*, pages 1163-1169.

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

Shiqi Zhao, Cheng Niu, Ming Zhou, Ting Liu, and Sheng Li. 2008a. Combining Multiple Resources to Improve SMT-based Paraphrasing Model. In *Proceedings of ACL-08:HLT*, pages 1021-1029.

Shiqi Zhao, Haifeng Wang, Ting Liu, and Sheng Li. 2008b. Pivot Approach for Extracting Paraphrase Patterns from Bilingual Corpora. In *Proceedings of ACL-08:HLT*, pages 780-788.