Joint Learning of a Dual SMT System for Paraphrase Generation

Hong Sun*
School of Computer Science and Technology
Tianjin University
kaspersky@tju.edu.cn

Ming Zhou
Microsoft Research Asia
mingzhou@microsoft.com

Abstract

SMT has been used in paraphrase generation by translating a source sentence into another (pivot) language and then back into the source. The resulting sentences can be used as candidate paraphrases of the source sentence. Existing work that uses two independently trained SMT systems cannot directly optimize the paraphrase results. Paraphrase criteria especially the paraphrase rate is not able to be ensured in that way. In this paper, we propose a joint learning method of two SMT systems to optimize the process of paraphrase generation. In addition, a revised BLEU score (called $iBLEU$) which measures the adequacy and diversity of the generated paraphrase sentence is proposed for tuning parameters in SMT systems. Our experiments on NIST 2008 testing data with automatic evaluation as well as human judgments suggest that the proposed method is able to enhance the paraphrase quality by adjusting between semantic equivalency and surface dissimilarity.

1 Introduction

Paraphrasing (at word, phrase, and sentence levels) is a procedure for generating alternative expressions with an identical or similar meaning to the original text. Paraphrasing technology has been applied in many NLP applications, such as machine translation (MT), question answering (QA), and natural language generation (NLG).

As paraphrasing can be viewed as a translation process between the original expression (as input) and the paraphrase results (as output), both in the same language, statistical machine translation (SMT) has been used for this task. Quirk et al. (2004) build a monolingual translation system using a corpus of sentence pairs extracted from news articles describing same events. Zhao et al. (2008a) enrich this approach by adding multiple resources (e.g., thesaurus) and further extend the method by generating different paraphrase in different applications (Zhao et al., 2009). Performance of the monolingual MT-based method in paraphrase generation is limited by the large-scale paraphrase corpus it relies on as the corpus is not readily available (Zhao et al., 2010).

In contrast, bilingual parallel data is in abundance and has been used in extracting paraphrase (Bannard and Callison-Burch, 2005; Zhao et al., 2008b; Callison-Burch, 2008; Kok and Brockett, 2010; Kuhn et al., 2010; Ganitkevitch et al., 2011). Thus researchers leverage bilingual parallel data for this task and apply two SMT systems (dual SMT system) to translate the original sentences into another pivot language and then translate them back into the original language. For question expansion, Duboué and Chu-Carroll (2006) paraphrase the questions with multiple MT engines and select the best paraphrase result considering cosine distance, length, etc. Max (2009) generates paraphrase for a given segment by forcing the segment being translated independently in both of the translation processes. Context features are added into the SMT system to improve translation correctness against polysemous. To reduce

[1] This work has been done while the author was visiting Microsoft Research Asia.

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 38–42,
Jeju, Republic of Korea, 8-14 July 2012.

the noise introduced by machine translation, Zhao et al. (2010) propose combining the results of multiple machine translation engines' by performing MBR (Minimum Bayes Risk) (Kumar and Byrne, 2004) decoding on the N-best translation candidates.

The work presented in this paper belongs to the pivot language method for paraphrase generation. Previous work employs two separately trained SMT systems the parameters of which are tuned for SMT scheme and therefore cannot directly optimize the paraphrase purposes, for example, optimize the diversity against the input. Another problem comes from the contradiction between two criteria in paraphrase generation: adequacy measuring the semantic equivalency and paraphrase rate measuring the surface dissimilarity. As they are incompatible (Zhao and Wang, 2010), the question arises how to adapt between them to fit different application scenarios. To address these issues, in this paper, we propose a joint learning method of two SMT systems for paraphrase generation. The jointly-learned dual SMT system: (1) Adapts the SMT systems so that they are tuned specifically for paraphrase generation purposes, e.g., to increase the dissimilarity; (2) Employs a revised BLEU score (named $iBLEU$, as it's an input-aware BLEU metric) that measures adequacy and dissimilarity of the paraphrase results at the same time. We test our method on NIST 2008 testing data. With both automatic and human evaluations, the results show that the proposed method effectively balance between adequacy and dissimilarity.

2 Paraphrasing with a Dual SMT System

We focus on sentence level paraphrasing and leverage homogeneous machine translation systems for this task bi-directionally. Generating sentential paraphrase with the SMT system is done by first translating a source sentence into another pivot language, and then back into the source. Here, we call these two procedures a dual SMT system. Given an English sentence e_s, there could be n candidate translations in another language F, each translation could have m candidates $\{e^{'}\}$ which may contain potential paraphrases for e_s. Our task is to locate the candidate that best fit in the demands of paraphrasing.

2.1 Joint Inference of Dual SMT System

During the translation process, it is needed to select a translation from the hypothesis based on the quality of the candidates. Each candidate's quality can be expressed by log-linear model considering different SMT features such as translation model and language model.

When generating the paraphrase results for each source sentence e_s, the selection of the best paraphrase candidate e'^* from $e' \in C$ is performed by:

$$\begin{aligned} e'^*(e_s, \{f\}, \lambda^M) \quad &= \\ \arg\max_{e' \in C, f \in \{f\}} \quad &\sum_{m=1}^{M} \lambda_m h_m(e'|f) t(e', f) \quad (1) \end{aligned}$$

where $\{f\}$ is the set of sentences in pivot language translated from e_s, h_m is the m_{th} feature value and λ_m is the corresponding weight. t is an indicator function equals to 1 when e' is translated from f and 0 otherwise.

The parameter weight vector λ is trained by MERT (Minimum Error Rate Training) (Och, 2003). MERT integrates the automatic evaluation metrics into the training process to achieve optimal end-to-end performance. In the joint inference method, the feature vector of each e' comes from two parts: vector of translating e_s to $\{f\}$ and vector of translating $\{f\}$ to e', the two vectors are jointly learned at the same time:

$$(\lambda_1^*, \lambda_2^*) = \arg\max_{(\lambda_1, \lambda_2)} \sum_{s=1}^{S} G(r_s, e'^*(e_s, \{f\}, \lambda_1, \lambda_2)) \quad (2)$$

where G is the automatic evaluation metric for paraphrasing. S is the development set for training the parameters and for each source sentence several human translations r_s are listed as references.

2.2 Paraphrase Evaluation Metrics

The joint inference method with MERT enables the dual SMT system to be optimized towards the quality of paraphrasing results. Different application scenarios of paraphrase have different demands on the paraphrasing results and up to now, the widely mentioned criteria include (Zhao et al., 2009; Zhao et al., 2010; Liu et al., 2010; Chen and Dolan, 2011; Metzler et al., 2011): Semantic adequacy, fluency and dissimilarity. However, as pointed out by (Chen and Dolan, 2011), there is the lack of automatic metric that is capable to measure all the three criteria in paraphrase generation. Two issues are also raised in (Zhao and Wang, 2010) about using automatic metrics: paraphrase changes less gets larger BLEU score and the evaluations of paraphrase quality and rate tend to be incompatible.

To address the above problems, we propose a metric for tuning parameters and evaluating the quality of each candidate paraphrase c :

$$\begin{aligned} iBLEU(s, r_s, c) &= \alpha BLEU(c, r_s) \\ &- (1-\alpha) BLEU(c, s) \quad (3) \end{aligned}$$

where s is the input sentence, r_s represents the reference paraphrases. $BLEU(c, r_s)$ captures the semantic equivalency between the candidates and the references (Finch et al. (2005) have shown the capability for measuring semantic equivalency using BLEU score); $BLEU(c, s)$ is the BLEU score computed between the candidate and the source sentence to measure the dissimilarity. α is a parameter taking balance between adequacy and dissimilarity, smaller α value indicates larger punishment on self-paraphrase. Fluency is not explicitly presented because there is high correlation between fluency and adequacy (Zhao et al., 2010) and SMT has already taken this into consideration. By using $iBLEU$, we aim at adapting paraphrasing performance to different application needs by adjusting α value.

3 Experiments and Results

3.1 Experiment Setup

For English sentence paraphrasing task, we utilize Chinese as the pivot language, our experiments are built on English and Chinese bi-directional translation. We use 2003 NIST Open Machine Translation Evaluation data (NIST 2003) as development data (containing 919 sentences) for MERT and test the performance on NIST 2008 data set (containing 1357 sentences). NIST Chinese-to-English evaluation data offers four English human translations for every Chinese sentence. For each sentence pair, we choose one English sentence e_1 as source and use the three left sentences e_2, e_3 and e_4 as references.

The English-Chinese and Chinese-English systems are built on bilingual parallel corpus containing 497,862 sentences. Language model is trained on 2,007,955 sentences for Chinese and 8,681,899 sentences for English. We adopt a phrase based MT system of Chiang (2007). 10-best lists are used in both of the translation processes.

Joint learning	BLEU	Self-BLEU	$iBLEU$
No Joint	27.16	35.42	/
$\alpha = 1$	30.75	53.51	30.75
$\alpha = 0.9$	28.28	48.08	20.64
$\alpha = 0.8$	27.39	35.64	14.78
$\alpha = 0.7$	23.27	26.30	8.39

Table 1: $iBLEU$ Score Results(NIST 2008)

	Adequacy (0/1/2)	Fluency (0/1/2)	Variety (0/1/2)	Overall (0/1/2)
No Joint	30/82/88	22/83/95	25/117/58	23/127/50
$\alpha = 1$	33/53/114	15/80/105	62/127/11	16/128/56
$\alpha = 0.9$	31/77/92	16/93/91	23/157/20	20/119/61
$\alpha = 0.8$	31/78/91	19/91/90	20/123/57	19/121/60
$\alpha = 0.7$	35/105/60	32/101/67	9/108/83	35/107/58

Table 2: Human Evaluation Label Distribution

3.2 Paraphrase Evaluation Results

The results of paraphrasing are illustrated in Table 1. We show the BLEU score (computed against references) to measure the adequacy and self-BLEU (computed against source sentence) to evaluate the dissimilarity (lower is better). By "No Joint", it means two independently trained SMT systems are employed in translating sentences from English to Chinese and then back into English. This result is listed to indicate the performance when we do not involve joint learning to control the quality of paraphrase results. For joint learning, results of α from 0.7 to 1 are listed.

From the results we can see that, when the value of α decreases to address more penalty on self-paraphrase, the self-BLEU score rapidly decays while the consequence effect is that BLEU score computed against references also drops seriously. When α drops under 0.6 we observe the sentences become completely incomprehensible (this is the reason why we leave out showing the results of α under 0.7). The best balance is achieved when α is between 0.7 and 0.9, where both of the sentence quality and variety are relatively preserved. As α value is manually defined and not specially tuned, the exper-

Source	Torrential rains hit western india , 43 people dead
No Joint	**Rainstorms in** western india , 43 **deaths**
Joint($\alpha = 1$)	**Rainstorms** hit western india , 43 people dead
Joint($\alpha = 0.9$)	**Rainstorms** hit western india 43 people dead
Joint($\alpha = 0.8$)	**Heavy rain in** western india , 43 **dead**
Joint($\alpha = 0.7$)	**Heavy rain in** western india , 43 **killed**

Table 3: Example of the Paraphrase Results

iments only achieve comparable results with no joint learning when α equals 0.8. However, the results show that our method is able to effectively control the self-paraphrase rate and lower down the score of self-BLEU, this is done by both of the process of joint learning and introducing the metric of $iBLEU$ to avoid trivial self-paraphrase. It is not capable with no joint learning or with the traditional BLEU score does not take self-paraphrase into consideration.

Human evaluation results are shown in Table 2. We randomly choose 100 sentences from testing data. For each setting, two annotators are asked to give scores about semantic adequacy, fluency, variety and overall quality. The scales are 0 (meaning changed; incomprehensible; almost same; cannot be used), 1 (almost same meaning; little flaws; containing different words; may be useful) and 2 (same meaning; good sentence; different sentential form; could be used). The agreements between the annotators on these scores are 0.87, 0.74, 0.79 and 0.69 respectively. From the results we can see that human evaluations are quite consistent with the automatic evaluation, where higher BLEU scores correspond to larger number of good adequacy and fluency labels, and higher self-BLEU results tend to get lower human evaluations over dissimilarity.

In our observation, we found that adequacy and fluency are relatively easy to be kept especially for short sentences. In contrast, dissimilarity is not easy to achieve. This is because the translation tables are used bi-directionally so lots of source sentences' fragments present in the paraphrasing results.

We show an example of the paraphrase results under different settings. All the results' sentential forms are not changed comparing with the input sentence and also well-formed. This is due to the short length of the source sentence. Also, with smaller value of α, more variations show up in the paraphrase results.

4 Discussion

4.1 SMT Systems and Pivot Languages

We have test our method by using homogeneous SMT systems and a single pivot language. As the method highly depends on machine translation, a natural question arises to what is the impact when using different pivots or SMT systems. The joint learning method works by combining both of the processes to concentrate on the final objective so it is not affected by the selection of language or SMT model.

In addition, our method is not limited to a homogeneous SMT model or a single pivot language. As long as the models' translation candidates can be scored with a log-linear model, the joint learning process can tune the parameters at the same time. When dealing with multiple pivot languages or heterogeneous SMT systems, our method will take effect by optimizing parameters from both the forward and backward translation processes, together with the final combination feature vector, to get optimal paraphrase results.

4.2 Effect of $iBLEU$

$iBLEU$ plays a key role in our method. The first part of $iBLEU$, which is the traditional BLEU score, helps to ensure the quality of the machine translation results. Further, it also helps to keep the semantic equivalency. These two roles unify the goals of optimizing translation and paraphrase adequacy in the training process.

Another contribution from $iBLEU$ is its ability to balance between adequacy and dissimilarity as the two aspects in paraphrasing are incompatible (Zhao and Wang, 2010). This is not difficult to explain because when we change many words, the meaning and the sentence quality are hard to preserve. As the paraphrasing task is not self-contained and will be employed by different applications, the two measures should be given different priorities based on the application scenario. For example, for a query expansion task in QA that requires higher recall, variety should be considered first. Lower α value is preferred but should be kept in a certain range as significant change may lead to the loss of constraints presented in the original sentence. The advantage of the proposed method is reflected in its ability to adapt to different application requirements by adjusting the value of α in a reasonable range.

5 Conclusion

We propose a joint learning method for pivot language-based paraphrase generation. The jointly learned dual SMT system which combines the training processes of two SMT systems in paraphrase generation, enables optimization of the final paraphrase quality. Furthermore, a revised BLEU score that balances between paraphrase adequacy and dissimilarity is proposed in our training process. In the future, we plan to go a step further to see whether we can enhance dissimilarity with penalizing phrase tables used in both of the translation processes.

References

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

Chris Callison-Burch. 2008. Syntactic constraints on paraphrases extracted from parallel corpora. In *EMNLP*, pages 196–205.

David Chen and William B. Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *ACL*, pages 190–200.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

Pablo Ariel Duboué and Jennifer Chu-Carroll. 2006. Answering the question you wish they had asked: The impact of paraphrasing for question answering. In *HLT-NAACL*.

Andrew Finch, Young-Sook Hwang, and Eiichiro Sumita. 2005. Using machine translation evaluation techniques to determine sentence-level semantic equivalence. In *In IWP2005*.

Juri Ganitkevitch, Chris Callison-Burch, Courtney Napoles, and Benjamin Van Durme. 2011. Learning sentential paraphrases from bilingual parallel corpora for text-to-text generation. In *EMNLP*, pages 1168–1179.

Stanley Kok and Chris Brockett. 2010. Hitting the right paraphrases in good time. In *HLT-NAACL*, pages 145–153.

Roland Kuhn, Boxing Chen, George F. Foster, and Evan Stratford. 2010. Phrase clustering for smoothing tm probabilities - or, how to extract paraphrases from phrase tables. In *COLING*, pages 608–616.

Shankar Kumar and William J. Byrne. 2004. Minimum bayes-risk decoding for statistical machine translation. In *HLT-NAACL*, pages 169–176.

Chang Liu, Daniel Dahlmeier, and Hwee Tou Ng. 2010. Pem: A paraphrase evaluation metric exploiting parallel texts. In *EMNLP*, pages 923–932.

Aurelien Max. 2009. Sub-sentential paraphrasing by contextual pivot translation. In *Proceedings of the 2009 Workshop on Applied Textual Inference, ACLI-JCNLP*, pages 18–26.

Donald Metzler, Eduard H. Hovy, and Chunliang Zhang. 2011. An empirical evaluation of data-driven paraphrase generation techniques. In *ACL (Short Papers)*, pages 546–551.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *ACL*, pages 160–167.

Chris Quirk, Chris Brockett, and William B. Dolan. 2004. Monolingual machine translation for paraphrase generation. In *EMNLP*, pages 142–149.

Shiqi Zhao and Haifeng Wang. 2010. Paraphrases and applications. In *COLING (Tutorials)*, pages 1–87.

Shiqi Zhao, Cheng Niu, Ming Zhou, Ting Liu, and Sheng Li. 2008a. Combining multiple resources to improve smt-based paraphrasing model. In *ACL*, pages 1021–1029.

Shiqi Zhao, Haifeng Wang, Ting Liu, and Sheng Li. 2008b. Pivot approach for extracting paraphrase patterns from bilingual corpora. In *ACL*, pages 780–788.

Shiqi Zhao, Xiang Lan, Ting Liu, and Sheng Li. 2009. Application-driven statistical paraphrase generation. In *ACL/AFNLP*, pages 834–842.

Shiqi Zhao, Haifeng Wang, Xiang Lan, and Ting Liu. 2010. Leveraging multiple mt engines for paraphrase generation. In *COLING*, pages 1326–1334.