PEM: A Paraphrase Evaluation Metric Exploiting Parallel Texts

Chang Liu[1] and **Daniel Dahlmeier**[2] and **Hwee Tou Ng**[1,2]
[1]Department of Computer Science, National University of Singapore
[2]NUS Graduate School for Integrative Sciences and Engineering
{liuchan1,danielhe,nght}@comp.nus.edu.sg

Abstract

We present PEM, the first fully automatic metric to evaluate the quality of paraphrases, and consequently, that of paraphrase generation systems. Our metric is based on three criteria: *adequacy*, *fluency*, and *lexical dissimilarity*. The key component in our metric is a robust and shallow semantic similarity measure based on pivot language N-grams that allows us to approximate adequacy independently of lexical similarity. Human evaluation shows that PEM achieves high correlation with human judgments.

1 Introduction

In recent years, there has been an increasing interest in the task of paraphrase generation (PG) (Barzilay and Lee, 2003; Pang et al., 2003; Quirk et al., 2004; Bannard and Callison-Burch, 2005; Kauchak and Barzilay, 2006; Zhao et al., 2008; Zhao et al., 2009). At the same time, the task has seen applications such as machine translation (MT) (Callison-Burch et al., 2006; Madnani et al., 2007; Madnani et al., 2008), MT evaluation (Kauchak and Barzilay, 2006; Zhou et al., 2006a; Owczarzak et al., 2006), summary evaluation (Zhou et al., 2006b), and question answering (Duboue and Chu-Carroll, 2006).

Despite the research activities, we see two major problems in the field. First, there is currently no consensus on what attributes characterize a good paraphrase. As a result, works on the application of paraphrases tend to build their own PG system in view of the immediate needs instead of using an existing system.

Second, and as a consequence, no automatic evaluation metric exists for paraphrases. Most works in this area resort to ad hoc manual evaluations, such as the percentage of "yes" judgments to the question of "is the meaning preserved". This type of evaluation is incomprehensive, expensive, and non-comparable between different studies, making progress hard to judge.

In this work we address both problems. We propose a set of three criteria for good paraphrases: *adequacy*, *fluency*, and *lexical dissimilarity*. Considering that paraphrase evaluation is a very subjective task with no rigid definition, we conduct experiments with human judges to show that humans generally have a consistent intuition for good paraphrases, and that the three criteria are good indicators.

Based on these criteria, we construct PEM (Paraphrase Evaluation Metric), a fully automatic evaluation metric for PG systems. PEM takes as input the original sentence R and its paraphrase candidate P, and outputs a single numeric score b estimating the quality of P as a paraphrase of R. PG systems can be compared based on the average scores of their output paraphrases. To the best of our knowledge, this is the first automatic metric that gives an objective and unambiguous ranking of different PG systems, which serves as a benchmark of progress in the field of PG.

The main difficulty of deriving PEM is to measure semantic closeness without relying on lexical level similarity. To this end, we propose *bag of pivot language N-grams* (*BPNG*) as a robust, broad-coverage, and knowledge-lean semantic representation for natural language sentences. Most importantly, BPNG does not depend on lexical or syntactic similarity, allowing us to address the conflicting requirements of paraphrase evaluation. The only linguistic re-

Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 923–932,
MIT, Massachusetts, USA, 9-11 October 2010.

source required to evaluate BPNG is a parallel text of the target language and an arbitrary other language, known as the pivot language.

We highlight that paraphrase evaluation and paraphrase recognition (Heilman and Smith, 2010; Das and Smith, 2009; Wan et al., 2006; Qiu et al., 2006) are related yet distinct tasks. Consider two sentences S_1 and S_2 that are the same except for the substitution of a single synonym. A paraphrase recognition system should assign them a very high score, but a paraphrase evaluation system would assign a relatively low one. Indeed, the latter is often a better indicator of how useful a PG system potentially is for the applications of PG described earlier.

The rest of the paper is organized as follows. We survey other automatic evaluation metrics in natural language processing (NLP) in Section 2. We define the task of paraphrase evaluation in Section 3 and develop our metric in Section 4. We conduct a human evaluation and analyze the results in Section 5. The correlation of PEM with human judgments is studied in Section 6. Finally, we discuss our findings and future work in Section 7 and conclude in Section 8.

2 Related work

The most well-known automatic evaluation metric in NLP is BLEU (Papineni et al., 2002) for MT, based on N-gram matching precisions. The simplicity of BLEU lends well to MT techniques that directly optimize the evaluation metric.

The weakness of BLEU is that it operates purely at the lexical surface level. Later works attempt to take more syntactic and semantic features into consideration (see (Callison-Burch et al., 2009) for an overview). The whole spectrum of NLP resources has found application in machine translation evaluation, including POS tags, constituent and dependency parses, WordNet (Fellbaum, 1998), semantic roles, textual entailment features, and more. Many of these metrics have been shown to correlate better with human judges than BLEU (Chan and Ng, 2008; Liu et al., 2010). Interestingly, few MT evaluation metrics exploit parallel texts as a source of information, when statistical MT is centered almost entirely around mining parallel texts.

Compared to these MT evaluation metrics, our method focuses on addressing the unique requirement of paraphrase evaluation: that lexical closeness does not necessarily entail goodness, contrary to the basis of MT evaluation.

Inspired by the success of automatic MT evaluation, Lin (2004) and Hovy et al. (2006) propose automatic metrics for summary evaluation. The former is entirely lexical based, whereas the latter also exploits constituent and dependency parses, and semantic features derived from WordNet.

The only prior attempt to devise an automatic evaluation metric for paraphrases that we are aware of is ParaMetric (Callison-Burch et al., 2008), which compares the collection of paraphrases discovered by automatic paraphrasing algorithms against a manual gold standard collected over the same sentences. The recall and precision of several current paraphrase generation systems are evaluated. ParaMetric does not attempt to propose a single metric to correlate well with human judgments. Rather, it consists of a few indirect and partial measures of the quality of PG systems.

3 Task definition

The first step in defining a paraphrase evaluation metric is to define a good paraphrase. Merriam-Webster dictionary gives the following definition: *a restatement of a text, passage, or work giving the meaning in another form*. We identify two key points in this definition: (1) that the meaning is preserved, and (2) that the lexical form is different. To which we add a third, that the paraphrase must be fluent.

The first and last point are similar to MT evaluation, where *adequacy* and *fluency* have been established as the standard criteria. In paraphrase evaluation, we have one more: *lexical dissimilarity*. Although lexical dissimilarity is seemingly the easiest to judge automatically among the three, it poses an interesting challenge to automatic evaluation metrics, as overlap with the reference has been the basis of almost all evaluation metrics. That is, while MT evaluation and paraphrase evaluation are conceptually closely related, the latter actually highlights the deficiencies of the former, namely that in most automatic evaluations, semantic equivalence is underrepresented and substituted by lexical and syntactic

equivalence.

The task of paraphrase evaluation is then defined as follows: Given an original sentence R and a paraphrase candidate P, output a numeric score b estimating the quality of P as a paraphrase of R by considering adequacy, fluency, and lexical dissimilarity. In this study, we use a scale of 1 to 5 (inclusive) for b, although that can be transformed linearly into any range desired.

We observe here that the overall assessment b is not a linear combination of the three measures. In particular, a high dissimilarity score is meaningless by itself. It could simply be that the paraphrase is unrelated to the source sentence, or is incoherent. However, when accompanied by high adequacy and fluency scores, it differentiates the mediocre paraphrases from the good ones.

4 Paraphrase Evaluation Metric (PEM)

In this section we devise our metric according to the three proposed evaluation criteria, namely adequacy, fluency, and dissimilarity. The main challenge is to measure the adequacy, or semantic similarity, completely independent of any lexical similarity. We address this problem in Sections 4.1 to 4.3. The remaining two criteria are addressed in Section 4.4, and we describe the final combined metric PEM in Section 4.5.

4.1 Phrase-level semantic representation

Without loss of generality, suppose we are to evaluate English paraphrases, and have been supplied many sentence-aligned parallel texts of French and English as an additional resource. We can then align the parallel texts at word level automatically using well-known algorithms such as GIZA++ (Och and Ney, 2003) or the Berkeley aligner (Liang et al., 2006; Haghighi et al., 2009).

To measure adequacy without relying on lexical similarity, we make the key observation that the aligned French texts can act as a proxy of the semantics to a fragment of an English text. If two English phrases are often mapped to the same French phrase, they can be considered similar in meaning. Similar observations have been made by previous researchers (Wu and Zhou, 2003; Bannard and Callison-Burch, 2005; Callison-Burch et al., 2006; Snover et al., 2009). We can treat the distribution of aligned French phrases as a semantic representation of the English phrase. The semantic distance between two English phrases can then be measured by their degree of overlap in this representation.

In this work, we use the widely-used phrase extraction heuristic in (Koehn et al., 2003) to extract phrase pairs from parallel texts into a phrase table[1]. The phrases extracted do not necessarily correspond to the speakers' intuition. Rather, they are units whose boundaries are preserved during translation. However, the distinction does not affect our work.

4.2 Segmenting a sentence into phrases

Having established a way to measure the similarity of two English phrases, we now extend the concept to sentences. Here we discuss how to segment an English sentence (the original or the paraphrase) into phrases.

From the phrase table, we know the frequencies of all the phrases and we approximate the probability of a phrase p by:

$$Pr(p) = \frac{N(p)}{\sum_{p'} N(p')} \quad (1)$$

$N(\cdot)$ is the count of a phrase in the phrase table, and the denominator is a constant for all p. We define the likelihood of segmenting a sentence S into a sequence of phrases $(p_1, p_2, \ldots, p_n)$ by:

$$Pr(p_1, p_2, \ldots, p_n | S) = \frac{1}{Z(S)} \prod_{i=1}^{n} Pr(p_i) \quad (2)$$

where $Z(S)$ is a normalizing constant. The best segmentation of S according to Equation 2 can be calculated efficiently using a dynamic programming algorithm. Note that $Z(S)$ does not need to be calculated, as it is the same for all different segmentations of S. The formula has a strong preference for longer phrases, since every $Pr(p_i)$ has a large denominator.

Many sentences are impossible to segment into known phrases, including all those containing out-of-vocabulary words. We therefore allow any single word w to be considered as a phrase, and if $N(w) = 0$, we use $N(w) = 0.5$ instead.

[1]The same heuristic is used in the popular MT package Moses.

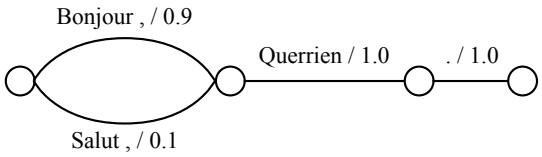

Figure 1: A confusion network in the pivot language

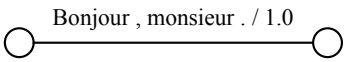

Figure 2: A degenerated confusion network in the pivot language

4.3 Sentence-level semantic representation

Simply merging the phrase-level semantic representations is insufficient to produce a sensible sentence-level semantic representation. For example, assume the English sentence *Morning , sir .* is segmented as a single phrase, because the following phrase pair is found in the phrase table:

En: Morning , sir .

Fr: Bonjour , monsieur .

However, another English sentence *Hello , Querrien .* has an out-of-vocabulary word *Querrien* and consequently the most probable segmentation is found to be "*Hello ,* ||| *Querrien* ||| .":

En: Hello ,

Fr: Bonjour , (Pr(Bonjour ,|Hello ,) $= 0.9$)

Fr: Salut , (Pr(Salut ,|Hello ,) $= 0.1$)

En: Querrien

Fr: Querrien

En: .

Fr: .

A naive comparison of the bags of French phrases aligned to *Morning , sir .* and *Hello , Querrien .* depicted above would conclude that the two sentences are completely unrelated, as their bags of aligned French phrases are completely disjoint. We tackle this problem by constructing a confusion network representation of the French phrases, as shown in Figures 1 and 2. The confusion network is formed by first joining the different French translations of every English phrase in parallel, and then joining these segments in series.

The confusion network is a compact representation of an exponentially large number of (likely malformed) weighted French sentences. We can easily enumerate the N-grams from the confusion network representation and collect the statistics for this ensemble of French sentences efficiently. In this work, we consider N up to 4. The N-grams for *Hello , Querrien .* are:

1-grams: Bonjour (0.9), Salut (0.1), *comma* (1.0), Querrien (1.0), *period* (1.0).

2-grams: Bonjour *comma* (0.9), Salut *comma* (0.1), *comma* Querrien (1.0), Querrien *period* (1.0).

3-grams: Bonjour *comma* Querrien (0.9), Salut *comma* Querrien (0.1), *comma* Querrien *period* (1.0).

4-grams: Bonjour *comma* Querrien *period* (0.9), Salut *comma* Querrien *period* (0.1).

We call this representation of an English sentence a *bag of pivot language N-grams* (BPNG), where French is the pivot language in our illustrating example. We can extract the BPNG of *Morning , sir .* analogously:

1-grams: Bonjour (1.0), *comma* (1.0), monsieur (1.0), *period* (1.0).

2-grams: Bonjour *comma* (1.0), *comma* monsieur (1.0), monsieur *period* (1.0).

3-grams: Bonjour *comma* monsieur (1.0), *comma* monsieur *period* (1.0).

4-grams: Bonjour *comma* monsieur *period* (1.0).

The BPNG of *Hello , Querrien.* can now be compared sensibly with that of the sentence *Morning , sir .* We use the F_1 agreement between the two BPNGs as a measure of the semantic similarity. The F_1 agreement is defined as

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The precision and the recall for an original sentence R and a paraphrase P is defined as follows. Let French N-gram $g \in \text{BPNG}(R) \cup \text{BPNG}(P)$, and $W_R(g)$ and $W_P(g)$ be the weights of g in the BPNG of R and P respectively, then

$$\text{Precision} = \frac{\sum_g \min(W_R(g), W_P(g))}{\sum_g W_P(g)}$$

$$\text{Recall} = \frac{\sum_g \min(W_R(g), W_P(g))}{\sum_g W_R(g)}$$

In our example, the numerators for both the precision and the recall are $0.9 + 1 + 1 + 0.9$, for the N-grams Bonjour, *comma*, *period*, and Bonjour *comma*

respectively. The denominators for both terms are 10.0. Consequently, F_1 = Precision = Recall = 0.38, and we conclude that the two sentences are 38% similar. We call the resulting metric the *pivot language* F_1. Note that since F_1 is symmetric with respect to the precision and the recall, our metric is unaffected whether we consider *Morning, sir.* as the paraphrase of *Hello, Querrien* . or the other way round.

An actual example from our corpus is:

Reference sihanouk ||| put forth ||| this proposal ||| in ||| a statement ||| made ||| yesterday ||| .

Paraphrase shihanuk ||| put forward ||| this proposal ||| in his ||| yesterday ||| 's statement ||| .

The ||| sign denotes phrase segmentation as described earlier. Our semantic representation successfully recognizes that *put forth* and *put forward* are paraphrases of each other, based on their similar Chinese translation statistics (*ti2 chu1* in Chinese).

4.4 Fluency and dissimilarity

We measure the fluency of a paraphrase by a normalized language model score P_n, defined by

$$P_n = \frac{\log Pr(S)}{\text{length}(S)}$$

where $Pr(S)$ is the sentence probability predicted by a standard 4-gram language model.

We measure dissimilarity between two English sentences using the *target language* F_1, where we collect the bag of all N-grams up to 4-grams from each English (referred to as the target language) sentence. The target language F_1 is then defined as the F_1 agreement of the two bags of N-grams, analogous to the definition of the pivot language F_1. The target language F_1 correlates positively with the similarity of the two sentences, or equivalently, negatively with the dissimilarity of the two sentences.

4.5 The metric

To produce the final PEM metric, we combine the three component automatic metrics, pivot language F_1, normalized language model, and target language F_1, which measure adequacy, fluency, and dissimilarity respectively.

As discussed previously, a linear combination of the three component metrics is insufficient. We turn to support vector machine (SVM) regression with the radial basis function (RBF) kernel. The RBF is a simple and expressive function, commonly used to introduce non-linearity into large margin classifications and regressions.

$$\text{RBF}(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2}$$

We use the implementation in SVM^{light} (Joachims, 1999). The SVM is to be trained on a set of human-judged paraphrase pairs, where the three component automatic metrics are fit to the human overall assessment. After training, the model can then be used to evaluate new paraphrase pairs in a fully automatic fashion.

5 Human evaluation

To validate our definition of paraphrase evaluation and the PEM method, we conduct an experiment to evaluate paraphrase qualities manually, which allows us to judge whether paraphrase evaluation according to our definition is an inherently coherent and well-defined problem. The evaluation also allows us to establish an upper bound for the paraphrase evaluation task, and to validate the contribution of the three proposed criteria to the overall paraphrase score.

5.1 Evaluation setup

We use the Multiple-Translation Chinese Corpus (MTC)[2] as a source of paraphrases. The MTC corpus consists of Chinese news articles (993 sentences in total) and multiple sentence-aligned English translations. We select one human translation as the original text. Two other human translations and two automatic machine translations serve as paraphrases of the original sentences. We refer to the two human translations and the two MT systems as paraphrase systems *human1, human2, machine1,* and *machine2*.

We employ three human judges to manually assess the quality of 300 original sentences paired with each of the four paraphrases. Therefore, each judge assesses 1,200 paraphrase pairs in total. The

[2]LDC Catalog No.: LDC2002T01

judgment for each paraphrase pair consists of four scores, each given on a five-point scale:

- Adequacy (*Is the meaning preserved adequately?*)
- Fluency (*Is the paraphrase fluent English?*)
- Lexical Dissimilarity (*How much has the paraphrase changed the original sentence?*)
- Overall score

The instructions given to the judges for the overall score were as follows.

> *A good paraphrase should convey the same meaning as the original sentence, while being as different as possible on the surface form and being fluent and grammatical English. With respect to this definition, give an overall score from 5 (perfect) to 1 (unacceptable) for this paraphrase.*

The paraphrases are presented to the judges in a random order and without any information as to which paraphrase system produced the paraphrase.

In addition to the four paraphrase systems mentioned above, for each original English sentence, we add three more artificially constructed paraphrases with pre-determined "human" judgment scores: (1) the original sentence itself, with adequacy 5, fluency 5, dissimilarity 1, and overall score 2; (2) a random sentence drawn from the same domain, with adequacy 1, fluency 5, dissimilarity 5, and overall score 1; and (3) a random sentence generated by a unigram language model, with adequacy 1, fluency 1, dissimilarity 5, and overall score 1. These artificial paraphrases serve as controls in our evaluation. Our final data set therefore consists of 2,100 paraphrase pairs with judgments on 4 different criteria.

5.2 Inter-judge correlation

The first step in our evaluation is to investigate the correlation between the human judges. We use Pearson's correlation coefficient, a common measure of the linear dependence between two random variables.

We investigate inter-judge correlation at the sentence and at the system level. At the sentence level, we construct three vectors, each containing the 1,200 sentence level judgments from one judge for the overall score. The pair-wise correlations between these three vectors are then taken. Note that we exclude the three artificial control paraphrase systems from consideration, as that would inflate the correlation. At the system level, we construct three vectors each of size four, containing the average scores given by one judge to each of the four paraphrase systems human1, human2, machine1, and machine2. The correlations are then taken in the same fashion.

	Sentence Level		System Level	
	Judge A	Judge B	Judge A	Judge B
Judge B	0.6406	-	0.9962	-
Judge C	0.6717	0.5993	0.9995	0.9943

Table 1: Inter-judge correlation for overall paraphrase score

	Sentence Level	System Level
Adequacy	0.7635	0.7616
Fluency	0.3736	0.3351
Dissimilarity	-0.3737	-0.3937
Dissimilarity (A,F$\geq$4)	0.8881	0.9956

Table 2: Correlation of paraphrase criteria with overall score

The results are listed in Table 1. The inter-judge correlation is between 0.60 and 0.67 at the sentence level and above 0.99 at the system level. These correlation scores can be considered very high when compared to similar results reported in MT evaluations, e.g., Blatz et al. (2003). The high correlation confirms that our evaluation task is well defined.

Having confirmed that human judgments correlate strongly, we combine the scores of the three judges by taking their arithmetic mean. Together with the three artificial control paraphrase systems, they form the human reference evaluation which we use for the remainder of the experiments.

5.3 Adequacy, fluency, and dissimilarity

In this section, we empirically validate the importance of our three proposed criteria: adequacy, fluency, and lexical dissimilarity. This can be done by measuring the correlation of each criterion with the overall score. The system and sentence level correlations are shown in Table 2.

We can see a positive correlation of adequacy and

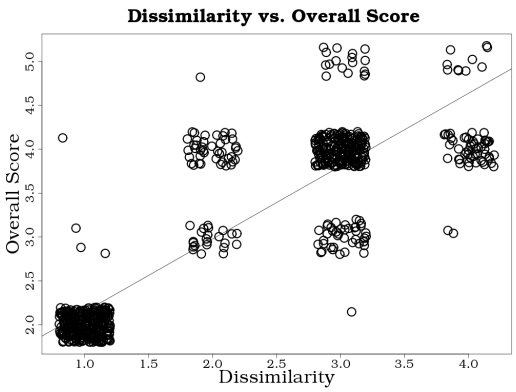

Figure 3: Scatter plot of dissimilarity vs. overall score for paraphrases with high adequacy and fluency.

fluency with the overall score, and the correlation with adequacy is particularly strong. Thus, higher adequacy and to a lesser degree higher fluency indicate higher paraphrase quality to the human judges.

On the other hand, dissimilarity is found to have a negative correlation with the overall score. This can be explained by the fact that the two human translations usually have much higher similarity with the reference translation, and at the same time are scored as better paraphrases. This effect dominates a simple linear fitting of the paraphrase score vs. the dissimilarity, resulting in the counter intuitive negative correlation. We note that a high dissimilarity alone tells us little about the quality of the paraphrase. Rather, we expect dissimilarity to be a differentiator between the mediocre and good paraphrases.

To test this hypothesis, we select the subset of paraphrase pairs that receive adequacy and fluency scores of at least four and again measure the correlation of the dissimilarity and the overall score. The result is tabulated in the last row of Table 2 and shows a strong correlation. Figure 3 shows a scatter plot of the same result[3].

The empirical results presented so far confirm that paraphrase evaluation is a well-defined task permitting consistent subjective judgments, and that adequacy, fluency, and dissimilarity are suitable criteria for paraphrase quality.

[3]We automatically add jitter (small amounts of noise) for ease of presentation.

6 PEM vs. human evaluation

In the last section, we have shown that the three proposed criteria are good indicators of paraphrase quality. In this section, we investigate how well PEM can predict the overall paraphrase quality from the three automatic metrics (pivot language F_1, normalized language model, and target language F_1), designed to match the three evaluation criteria. We describe the experimental setup in Section 6.1, before we show the results in Section 6.2.

6.1 Experimental setup

We build the phrase table used to evaluate the pivot language F_1 from the FBIS Chinese-English corpus, consisting of about 250,000 Chinese sentences, each with a single English translation. The paraphrases are taken from the MTC corpus in the same way as the human experiment described in Section 5.1. Both FBIS and MTC are in the Chinese newswire domain.

We stem all English words in both data sets with the Porter stemmer (Porter, 1980). We use the maximum entropy segmenter of (Low et al., 2005) to segment the Chinese part of the FBIS corpus. Subsequently, word level Chinese-English alignments are generated using the Berkeley aligner (Liang et al., 2006; Haghighi et al., 2009) with five iterations of training. Phrases are then extracted with the widely-used heuristic in Koehn et al. (2003). We extract phrases of up to four words in length.

Bags of Chinese pivot language N-grams are extracted for all paraphrase pairs as described in Section 4.3. For computational efficiency, we consider only edges of the confusion network with probabilities higher than 0.1, and only N-grams with probabilities higher than 0.01 in the bag of N-grams. We collect N-grams up to length four.

The language model used to judge fluency is trained on the English side of the FBIS parallel text. We use SRILM (Stolcke, 2002) to build a 4-gram model with the default parameters.

The PEM SVM regression is trained on the paraphrase pairs for the first 200 original English sentences and tested on the paraphrase pairs of the remaining 100 original English sentences. Thus, there are 1,400 instances for training and 700 instances for testing. For each instance, we calculate the values

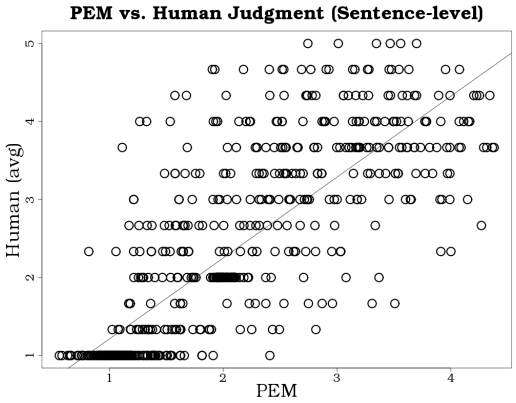

Figure 4: Scatter plot of PEM vs. human judgment (overall score) at the sentence level

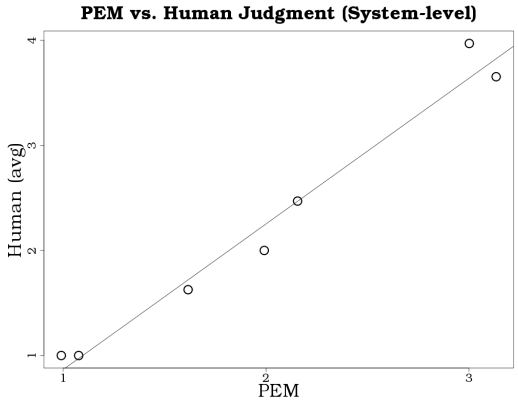

Figure 5: Scatter plot of PEM vs. human judgment (overall score) at the system level

of pivot language F_1, normalized language model score, and target language F_1. These values serve as the input features to the SVM regression and the target value is the human assessment of the overall score, on a scale of 1 to 5.

6.2 Results

As in the human evaluation, we investigate the correlation of the PEM scores with the human judgments at the sentence and at the system level. Figure 4 shows the sentence level PEM scores plotted against the human overall scores, where each human overall score is the arithmetic mean of the scores given by the three judges. The Pearson correlation between the automatic PEM scores and the human judgments is 0.8073. This is substantially higher than the sentence level correlation of MT metrics like BLEU. For example, the highest sentence level Pearson correlation by any metric in the MetricsMATR 2008 competition (Przybocki et al., 2009) was 0.6855 by METEOR-v0.6; BLEU achieved a correlation of 0.4513.

	Sentence Level	System Level
PEM vs. Human Avg.	0.8073	0.9867
PEM vs. Judge A	0.5777	0.9757
PEM vs. Judge B	0.5281	0.9892
PEM vs. Judge C	0.5231	0.9718

Table 3: Correlation of PEM with human judgment (overall score)

Figure 5 shows the system level PEM scores plotted against the human scores. The Pearson correlation between PEM scores and the human scores at the system level is 0.9867.

We also calculate the Pearson correlation between PEM and each individual human judge. Here, we exclude the three artificial control paraphrase systems from the data, to make the results comparable to the inter-judge correlation presented in Section 5.2. The correlation is between 0.52 and 0.57 at the sentence level and between 0.97 and 0.98 at the system level. As we would expect, the correlation between PEM and a human judge is not as high as the correlation between two human judges, but PEM still shows a strong and consistent correlation with all three judges. The results are summarized in Table 3.

7 Discussion and future work

The paraphrases that we use in this study are not actual machine generated paraphrases. Instead, the English paraphrases are multiple translations of the same Chinese source sentence. Our seven "paraphrase systems" are two human translators, two machine translation systems, and three artificially created extreme scenarios. The reason for using multiple translations is that we could not find any PG system that can paraphrase a whole input sentence and is publicly available. We intend to obtain and evaluate paraphrases generated from real PG systems and compare their performances in a follow-up study.

Our method models paraphrasing up to the phrase level. Unfortunately, it makes no provisions for syn-

tactic paraphrasing at the sentence level, which is probably a much greater challenge, and the literature offers few successes to draw inspirations from. We hope to be able to partially address this deficiency in future work.

The only external linguistic resource required by PEM is a parallel text of the target language and another arbitrary language. While we only use Chinese-English parallel text in this study, other language pairs need to be explored too. Another alternative is to collect parallel texts against multiple foreign languages, e.g., using Europarl (Koehn, 2005). We leave this for future work.

Our evaluation method does not require human-generated references like in MT evaluation. Therefore, we can easily formulate a paraphrase generator by directly optimizing the PEM metric, although solving it is not trivial:

$$\text{paraphrase}(R) = \arg\max_{P} \text{PEM}(P, R)$$

where R is the original sentence and P is the paraphrase.

Finally, the PEM metric, in particular the semantic representation BPNG, can be useful in many other contexts, such as MT evaluation, summary evaluation, and paraphrase recognition. To facilitate future research, we will package and release PEM under an open source license at `http://nlp.comp.nus.edu.sg/software`.

8 Conclusion

We proposed PEM, a novel automatic metric for paraphrase evaluation based on adequacy, fluency, and lexical dissimilarity. The key component in our metric is a novel technique to measure the semantic similarity of two sentences through their N-gram overlap in an aligned foreign language text. We conducted an extensive human evaluation of paraphrase quality which shows that our proposed metric achieves high correlation with human judgments. To the best of our knowledge, PEM is the first automatic metric for paraphrase evaluation.

Acknowledgments

This research was done for CSIDM Project No. CSIDM-200804 partially funded by a grant from the National Research Foundation (NRF) administered by the Media Development Authority (MDA) of Singapore.

References

C. Bannard and C. Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proc. of ACL*.

R. Barzilay and L. Lee. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proc. of HLT-NAACL*.

J. Blatz, E. Fitzgerald, G. Foster, S. Gandrabur, C. Goutte, A. Kulesza, A. Sanchis, and N. Ueffing. 2003. Confidence estimation for machine translation. Technical report, CLSP Workshop Johns Hopkins University.

C. Callison-Burch, P. Koehn, and M. Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proc. of HLT-NAACL*.

C. Callison-Burch, T. Cohn, and M. Lapata. 2008. ParaMetric: An automatic evaluation metric for paraphrasing. In *Proc. of COLING*.

C. Callison-Burch, P. Koehn, C. Monz, and J. Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of WMT*.

Y.S. Chan and H.T. Ng. 2008. MAXSIM: A maximum similarity metric for machine translation evaluation. In *Proc. of ACL-08: HLT*.

D. Das and N.A. Smith. 2009. Paraphrase identification as probabilistic quasi-synchronous recognition. In *Proc. of ACL-IJCNLP*.

P. Duboue and J. Chu-Carroll. 2006. Answering the question you wish they had asked: The impact of paraphrasing for question answering. In *Proc. of HLT-NAACL Companion Volume: Short Papers*.

C. Fellbaum, editor. 1998. *WordNet: An electronic lexical database*. MIT Press, Cambridge, MA.

A. Haghighi, J. Blitzer, J. DeNero, and D. Klein. 2009. Better word alignments with supervised ITG models. In *Proc. of ACL-IJCNLP*.

M. Heilman and N.A. Smith. 2010. Tree edit models for recognizing textual entailments, paraphrases, and answers to questions. In *Proc. of NAACL*.

E. Hovy, C.Y. Lin, L. Zhou, and J. Fukumoto. 2006. Automated summarization evaluation with basic elements. In *Proc. of LREC*.

T. Joachims. 1999. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press.

D. Kauchak and R. Barzilay. 2006. Paraphrasing for automatic evaluation. In *Proc. of HLT-NAACL*.

P. Koehn, F.J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proc. of HLT-NAACL*.

P. Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*, volume 5.

P. Liang, B. Taskar, and D. Klein. 2006. Alignment by agreement. In *Proc. of HLT-NAACL*.

C.Y. Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proc. of the ACL-04 Workshop on Text Summarization Branches Out*.

C. Liu, D. Dahlmeier, and H.T. Ng. 2010. TESLA: translation evaluation of sentences with linear-programming-based analysis. In *Proc. of WMT*.

J.K. Low, H.T. Ng, and W. Guo. 2005. A maximum entropy approach to Chinese word segmentation. In *Proc. of the 4th SIGHAN Workshop*.

N. Madnani, N.F. Ayan, P. Resnik, and B.J. Dorr. 2007. Using paraphrases for parameter tuning in statistical machine translation. In *Proc. of WMT*.

N. Madnani, P. Resnik, B.J. Dorr, and R. Schwartz. 2008. Are multiple reference translations necessary? Investigating the value of paraphrased reference translations in parameter optimization. In *Proc. of AMTA*.

F.J. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1).

K. Owczarzak, D. Groves, J. Van Genabith, and A. Way. 2006. Contextual bitext-derived paraphrases in automatic MT evaluation. In *Proc. of WMT*.

B. Pang, K. Knight, and D. Marcu. 2003. Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. In *Proc. of HLT-NAACL*.

K. Papineni, S. Roukos, T. Ward, and W.J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. of ACL*.

M. Porter. 1980. An algorithm for suffix stripping. *Program*, 40(3).

M. Przybocki, K. Peterson, S. Bronsart, and G Sanders. 2009. Evaluating machine translation with LFG dependencies. *Machine Translation*, 23(2).

L. Qiu, M.Y. Kan, and T.S. Chua. 2006. Paraphrase recognition via dissimilarity significance classification. In *Proc. of EMNLP*.

C. Quirk, C. Brockett, and W. Dolan. 2004. Monolingual machine translation for paraphrase generation. In *Proc. of EMNLP*.

M. Snover, N. Madnani, B. Dorr, and R. Schwartz. 2009. Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. In *Proc. of WMT*.

A. Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proc. of ICSLP*.

S. Wan, M. Dras, R. Dale, and C Paris. 2006. Using dependency-based features to take the 'para-farce' out of paraphrase. In *Proc. of ALTW 2006*.

H. Wu and M. Zhou. 2003. Synonymous collocation extraction using translation information. In *Proc. of ACL*.

S.Q. Zhao, C. Niu, M. Zhou, T. Liu, and S. Li. 2008. Combining multiple resources to improve SMT-based paraphrasing model. In *Proc. of ACL-08: HLT*.

S.Q. Zhao, X. Lan, T. Liu, and S. Li. 2009. Application-driven statistical paraphrase generation. In *Proc. of ACL-IJCNLP*.

L. Zhou, C.Y. Lin, and E. Hovy. 2006a. Re-evaluating machine translation results with paraphrase support. In *Proc. of EMNLP*.

L. Zhou, C.Y. Lin, D.S. Munteanu, and E. Hovy. 2006b. ParaEval: Using paraphrases to evaluate summaries automatically. In *Proc. of HLT-NAACL*.