Paraphrasing for Automatic Evaluation

David Kauchak
Department of Computer Science
University of California, San Diego
dkauchak@cs.ucsd.edu

Regina Barzilay
CSAIL
Massachusetts Institute of Technology
regina@csail.mit.edu

Abstract

This paper studies the impact of paraphrases on the accuracy of automatic evaluation. Given a reference sentence and a machine-generated sentence, we seek to find a paraphrase of the reference sentence that is closer in wording to the machine output than the original reference. We apply our paraphrasing method in the context of machine translation evaluation. Our experiments show that the use of a paraphrased synthetic reference refines the accuracy of automatic evaluation. We also found a strong connection between the quality of automatic paraphrases as judged by humans and their contribution to automatic evaluation.

1 Introduction

The use of automatic methods for evaluating machine-generated text is quickly becoming mainstream in natural language processing. The most notable examples in this category include measures such as BLEU and ROUGE which drive research in the machine translation and text summarization communities. These methods assess the quality of a machine-generated output by considering its similarity to a reference text written by a human. Ideally, the similarity would reflect the semantic proximity between the two. In practice, this comparison breaks down to n-gram overlap between the reference and the machine output.

1a. However, Israel's reply failed to completely clear the U.S. suspicions.
1b. However, Israeli answer unable to fully remove the doubts.

Table 1: A reference sentence and corresponding machine translation from the NIST 2004 MT evaluation.

Consider the human-written translation and the machine translation of the same Chinese sentence shown in Table 1. While the two translations convey the same meaning, they share only auxiliary words. Clearly, any measure based on word overlap will penalize a system for generating such a sentence. The question is whether such cases are common phenomena or infrequent exceptions. Empirical evidence supports the former. Analyzing 10,728 reference translation pairs[1] used in the NIST 2004 machine translation evaluation, we found that only 21 (less than 0.2%) of them are identical. Moreover, 60% of the pairs differ in at least 11 words. These statistics suggest that without accounting for paraphrases, automatic evaluation measures may never reach the accuracy of human evaluation.

As a solution to this problem, researchers use multiple references to refine automatic evaluation. Papineni et al. (2002) shows that expanding the number of references reduces the gap between automatic and human evaluation. However, very few human annotated sets are augmented with multiple references and those that are available are relatively

[1]Each pair included different translations of the same sentence, produced by two human translators.

small in size. Moreover, access to several references does not guarantee that the references will include the same words that appear in machine-generated sentences.

In this paper, we explore the use of paraphrasing methods for refinement of automatic evaluation techniques. Given a reference sentence and a machine-generated sentence, we seek to find a paraphrase of the reference sentence that is closer in wording to the machine output than the original reference. For instance, given the pair of sentences in Table 1, we automatically transform the reference sentence (1a.) into

> However, Israel's *answer* failed to completely *remove* the U.S. suspicions.

Thus, among many possible paraphrases of the reference, we are interested only in those that use words appearing in the system output. Our paraphrasing algorithm is based on the *substitute in context* strategy. First, the algorithm identifies pairs of words from the reference and the system output that could potentially form paraphrases. We select these candidates using existing lexico-semantic resources such as WordNet. Next, the algorithm tests whether the candidate paraphrase is admissible in the context of the reference sentence. Since even synonyms cannot be substituted in any context (Edmonds and Hirst, 2002), this filtering step is necessary. We predict whether a word is appropriate in a new context by analyzing its distributional properties in a large body of text. Finally, paraphrases that pass the filtering stage are used to rewrite the reference sentence.

We apply our paraphrasing method in the context of machine translation evaluation. Using this strategy, we generate a new sentence for every pair of human and machine translated sentences. This synthetic reference then replaces the original human reference in automatic evaluation.

The key findings of our work are as follows:
(1) **Automatically generated paraphrases improve the accuracy of the automatic evaluation methods.** Our experiments show that evaluation based on paraphrased references gives a better approximation of human judgments than evaluation that uses original references.
(2) **The quality of automatic paraphrases determines their contribution to automatic evaluation.** By analyzing several paraphrasing resources, we found that the accuracy and coverage of a paraphrasing method correlate with its utility for automatic MT evaluation.

Our results suggest that researchers may find it useful to augment standard measures such as BLEU and ROUGE with paraphrasing information thereby taking more semantic knowledge into account.

In the following section, we provide an overview of existing work on automatic paraphrasing. We then describe our paraphrasing algorithm and explain how it can be used in an automatic evaluation setting. Next, we present our experimental framework and data and conclude by presenting and discussing our results.

2 Related Work

Automatic Paraphrasing and Entailment Our work is closely related to research in automatic paraphrasing, in particular, to sentence level paraphrasing (Barzilay and Lee, 2003; Pang et al., 2003; Quirk et al., 2004). Most of these approaches learn paraphrases from a parallel or comparable monolingual corpora. Instances of such corpora include multiple English translations of the same source text written in a foreign language, and different news articles about the same event. For example, Pang et al. (2003) expand a set of reference translations using syntactic alignment, and generate new reference sentences that could be used in automatic evaluation.

Our approach differs from traditional work on automatic paraphrasing in goal and methodology. Unlike previous approaches, we are not aiming to produce *any* paraphrase of a given sentence since paraphrases induced from a parallel corpus do not necessarily produce a rewriting that makes a reference closer to the system output. Thus, we focus on words that appear in the system output and aim to determine whether they can be used to rewrite a reference sentence.

Our work also has interesting connections with research on automatic textual entailment (Dagan et al., 2005), where the goal is to determine whether a given sentence can be inferred from text. While we are not assessing an inference relation between a reference and a system output, the two tasks face similar challenges. Methods for entailment

recognition extensively rely on lexico-semantic resources (Haghighi et al., 2005; Harabagiu et al., 2001), and we believe that our method for contextual substitution can be beneficial in that context.

Automatic Evaluation Measures A variety of automatic evaluation methods have been recently proposed in the machine translation community (NIST, 2002; Melamed et al., 2003; Papineni et al., 2002). All these metrics compute n-gram overlap between a reference and a system output, but measure the overlap in different ways. Our method for reference paraphrasing can be combined with any of these metrics. In this paper, we report experiments with BLEU due to its wide use in the machine translation community.

Recently, researchers have explored additional knowledge sources that could enhance automatic evaluation. Examples of such knowledge sources include stemming and TF-IDF weighting (Babych and Hartley, 2004; Banerjee and Lavie, 2005). Our work complements these approaches: we focus on the impact of paraphrases, and study their contribution to the accuracy of automatic evaluation.

3 Methods

The input to our method consists of a reference sentence $R = r_1 \dots r_m$ and a system-generated sentence $W = w_1 \dots w_p$ whose words form the sets $\mathcal{R}$ and $\mathcal{W}$ respectively. The output of the model is a synthetic reference sentence S_{RW} that preserves the meaning of R and has maximal word overlap with W. We generate such a sentence by substituting words from R with contextually equivalent words from W.

Our algorithm first selects pairs of candidate word paraphrases, and then checks the likelihood of their substitution in the context of the reference sentence.

Candidate Selection We assume that words from the reference sentence that already occur in the system generated sentence should not be considered for substitution. Therefore, we focus on unmatched pairs of the form $\{(r, w) | r \in \mathcal{R} - \mathcal{W}, w \in \mathcal{W} - \mathcal{R}\}$. From this pool, we select candidate pairs whose members exhibit high semantic proximity. In our experiments we compute semantic similarity using WordNet, a large-scale lexico-semantic resource employed in many NLP applications for similar purposes. We consider a pair as a substitution candidate if its members are synonyms in WordNet.

2a. It is **hard** to believe that such tremendous changes have taken **place** for those people and lands that I have never stopped missing while living abroad.
2b. For someone born here but has been sentimentally attached to a foreign country far from **home**, it is **difficult** to believe this kind of changes.

Table 2: A reference sentence and a corresponding machine translation. Candidate paraphrases are in bold.

Applying this step to the two sentences in Table 2, we obtain two candidate pairs (**home**, **place**) and (**difficult**, **hard**).

Contextual Substitution The next step is to determine for each candidate pair (r_i, w_j) whether w_j is a valid substitution for r_i in the context of $r_1 \dots r_{i-1} \square r_{i+1} \dots r_m$. This filtering step is essential because synonyms are not universally substitutable[2]. Consider the candidate pair (**home**, **place**) from our example (see Table 2). Words **home** and **place** are paraphrases in the sense of "habitat", but in the reference sentence "**place**" occurs in a different sense, being part of the collocation "take place". In this case, the pair (**home**, **place**) cannot be used to rewrite the reference sentence.

We formulate contextual substitution as a binary classification task: given a context $r_1 \dots r_{i-1} \square r_{i+1} \dots r_m$, we aim to predict whether w_j can occur in this context at position i. For each candidate word w_j we train a classifier that models contextual preferences of w_j. To train such a classifier, we collect a large corpus of sentences that contain the word w_j and an equal number of randomly extracted sentences that do not contain this word. The former category forms positive instances, while the latter represents the negative. For the negative examples, a random position in a sentence is selected for extracting the context. This corpus is acquired automatically, and does not require any manual annotations.

[2]This can explain why previous attempts to use WordNet for generating sentence-level paraphrases (Barzilay and Lee, 2003; Quirk et al., 2004) were unsuccessful.

We represent context by n-grams and local collocations, features typically used in supervised word sense disambiguation. Both n-grams and collocations exclude the word w_j. An n-gram is a sequence of n adjacent words appearing in $r_1 \dots r_{i-1} \square r_{i+1} \dots r_m$. A local collocation also takes into account the position of an n-gram with respect to the target word. To compute local collocations for a word at position i, we extract all n-grams ($n = 1 \dots 4$) beginning at position $i - 2$ and ending at position $i + 2$. To make these position dependent, we prepend each of them with the length and starting position.

Once the classifier[3] for w_j is trained, we apply it to the context $r_1 \dots r_{i-1} \square r_{i+1} \dots r_m$. For positive predictions, we rewrite the string as $r_1 \dots r_{i-1} w_j r_{i+1} \dots r_m$. In this formulation, all substitutions are tested independently.

For the example from Table 2, only the pair (**difficult**, **hard**) passes this filter, and thus the system produces the following synthetic reference:

> For someone born here but has been sentimentally attached to a foreign country far from home, it is **hard** to believe this kind of changes.

The synthetic reference keeps the meaning of the original reference, but has a higher word overlap with the system output.

One of the implications of this design is the need to develop a large number of classifiers to test contextual substitutions. For each word to be inserted into a reference sentence, we need to train a separate classifier. In practice, this requirement is not a significant burden. The training is done off-line and only once, and testing for contextual substitution is instantaneous. Moreover, the first filtering step effectively reduces the number of potential candidates. For example, to apply this approach to the 71,520 sentence pairs from the MT evaluation set (described in Section 4.1.2), we had to train 2,380 classifiers.

We also discovered that the key to the success of this approach is the size of the corpus used for training contextual classifiers. We derived training corpora from the English Gigaword corpus, and the average size of a corpus for one classifier is 255,000 sentences. We do not attempt to substitute any words that have less that 10,000 appearances in the Gigaword corpus.

[3] In our experiments, we used the publicly available BoosTexter classifier (Schapire and Singer, 2000) for this task.

4 Experiments

Our primary goal is to investigate the impact of machine-generated paraphrases on the accuracy of automatic evaluation. We focus on automatic evaluation of machine translation due to the availability of human annotated data in that domain. The hypothesis is that by using a synthetic reference translation, automatic measures approximate better human evaluation. In section 4.2, we test this hypothesis by comparing the performance of BLEU scores with and without synthetic references.

Our secondary goal is to study the relationship between the quality of paraphrases and their contribution to the performance of automatic machine translation evaluation. In section 4.3, we present a manual evaluation of several paraphrasing methods and show a close connection between intrinsic and extrinsic assessments of these methods.

4.1 Experimental Set-Up

We begin by describing relevant background information, including the BLEU evaluation method, the test data set, and the alternative paraphrasing methods considered in our experiments.

4.1.1 BLEU

BLEU is the basic evaluation measure that we use in our experiments. It is the geometric average of the n-gram precisions of candidate sentences with respect to the corresponding reference sentences, times a brevity penalty. The BLEU score is computed as follows:

$$\begin{aligned} BLEU &= BP \cdot \sqrt[4]{\prod_{n=1}^{4} p_n} \\ BP &= min(1, e^{1-r/c}), \end{aligned}$$

where p_n is the n-gram precision, c is the cardinality of the set of candidate sentences and r is the size of the smallest set of reference sentences.

To augment BLEU evaluation with paraphrasing information, we substitute each reference with the corresponding synthetic reference.

4.1.2 Data

We use the Chinese portion of the 2004 NIST MT dataset. This portion contains 200 Chinese documents, subdivided into a total of 1788 segments. Each segment is translated by ten machine translation systems and by four human translators. A quarter of the machine-translated segments are scored by human evaluators on a one-to-five scale along two dimensions: adequacy and fluency. We use only adequacy scores, which measure how well content is preserved in the translation.

4.1.3 Alternative Paraphrasing Techniques

To investigate the effect of paraphrase quality on automatic evaluation, we consider two alternative paraphrasing resources: Latent Semantic Analysis (LSA), and Brown clustering (Brown et al., 1992). These techniques are widely used in NLP applications, including language modeling, information extraction, and dialogue processing (Haghighi et al., 2005; Serafin and Eugenio, 2004; Miller et al., 2004). Both techniques are based on distributional similarity. The Brown clustering is computed by considering mutual information between adjacent words. LSA is a dimensionality reduction technique that projects a word co-occurrence matrix to lower dimensions. This lower dimensional representation is then used with standard similarity measures to cluster the data. Two words are considered to be a paraphrase pair if they appear in the same cluster.

We construct 1000 clusters employing the Brown method on 112 million words from the North American New York Times corpus. We keep the top 20 most frequent words for each cluster as paraphrases. To generate LSA paraphrases, we used the Infomap software[4] on a 34 million word collection of articles from the American News Text corpus. We used the default parameter settings: a 20,000 word vocabulary, the 1000 most frequent words (minus a stoplist) for features, a 15 word context window on either side of a word, a 100 feature reduced representation, and the 20 most similar words as paraphrases.

While we experimented with several parameter settings for LSA and Brown methods, we do not claim that the selected settings are necessarily optimal. However, these methods present sensible comparison points for understanding the relationship between paraphrase quality and its impact on automatic evaluation.

Table 3 shows synthetic references produced by the different paraphrasing methods.

[4] http://infomap-nlp.sourceforge.net

Method	1 reference	2 references
BLEU	0.9657	0.9743
WordNet	0.9674	0.9763
ContextWN	**0.9677**	**0.9764**
LSA	0.9652	0.9736
Brown	0.9662	0.9744

Table 4: Pearson adequacy correlation scores for rewriting using one and two references, averaged over ten runs.

Method	vs. BLEU	vs. ContextWN
WordNet	◁◁	△△
ContextWN	◁◁	-
LSA	X	△△
Brown	◁◁	△

Table 5: Paired t-test significance for all methods compared to BLEU as well as our method for one reference. Two triangles indicates significant at the 99% confidence level, one triangle at the 95% confidence level and X not significant. Triangles point towards the better method.

4.2 Impact of Paraphrases on Machine Translation Evaluation

The standard way to analyze the performance of an evaluation metric in machine translation is to compute the Pearson correlation between the automatic metric and human scores (Papineni et al., 2002; Koehn, 2004; Lin and Och, 2004; Stent et al., 2005). Pearson correlation estimates how linearly dependent two sets of values are. The Pearson correlation values range from 1, when the scores are perfectly linearly correlated, to -1, in the case of inversely correlated scores.

To calculate the Pearson correlation, we create a document by concatenating 300 segments. This strategy is commonly used in MT evaluation, because of BLEU's well-known problems with documents of small size (Papineni et al., 2002; Koehn, 2004). For each of the ten MT system translations,

Reference:	The monthly magazine "Choices" has won the deep trust of the residents. The current Internet edition of "Choices" will give full play to its functions and will help consumers get quick access to market information.
System:	The public has a lot of faith in the "Choice" monthly magazine and the Council is now working on a web version. This will enhance the magazine's function and help consumer to acquire more up-to-date market information.
WordNet	The monthly magazine "Choices" has won the deep **faith** of the residents. The current Internet **version** of "Choices" will give full play to its functions and will help consumers **acquire** quick access to market information.
ContextWN	The monthly magazine "Choices" has won the deep trust of the residents. The current Internet **version** of "Choices" will give full play to its functions and will help consumers **acquire** quick access to market information.
LSA	The monthly magazine "**Choice**" has won the deep trust of the residents. The current **web** edition of "**Choice**" will give full play to its functions and will help **consumer** get quick access to market information.
Brown	The monthly magazine "Choices" has won the deep trust of the residents. The current Internet **version** of "Choices" will give full play to its functions and will help consumers get quick access to market information.

Table 3: Sample of paraphrasings produced by each method based on the corresponding system translation. Paraphrased words are in bold and filtered words underlined.

the evaluation metric score is calculated on the document and the corresponding human adequacy score is calculated as the average human score over the segments. The Pearson correlation is calculated over these ten pairs (Papineni et al., 2002; Stent et al., 2005). This process is repeated for ten different documents created by the same process. Finally, a paired t-test is calculated over these ten different correlation scores to compute statistical significance.

Table 4 shows Pearson correlation scores for BLEU and the four paraphrased augmentations, averaged over ten runs.[5] In all ten tests, our method based on contextual rewriting (ContextWN) improves the correlation with human scores over BLEU. Moreover, in nine out of ten tests ContextWN outperforms the method based on WordNet. The results of statistical significance testing are summarized in Table 5. All the paraphrasing methods except LSA, exhibit higher correlation with human scores than plain BLEU. Our method significantly outperforms BLEU, and all the other paraphrase-based metrics. This consistent improvement confirms the importance of contextual filtering.

[5]Depending on the experimental setup, correlation values can vary widely. Our scores fall within the range of previous researchers (Papineni et al., 2002; Lin and Och, 2004).

The third column in Table 4 shows that automatic paraphrasing continues to improve correlation scores even when two human references are paraphrased using our method.

4.3 Evaluation of Paraphrase Quality

In the last section, we saw significant variations in MT evaluation performance when different paraphrasing methods were used to generate a synthetic reference. In this section, we examine the correlation between the quality of automatically generated paraphrases and their contribution to automatic evaluation. We analyze how the substitution frequency and the accuracy of those substitutions contributes to a method's performance.

We compute the substitution frequency of an automatic paraphrasing method by counting the number of words it rewrites in a set of reference sentences. Table 6 shows the substitution frequency and the corresponding BLEU score. The substitution frequency varies greatly across different methods — LSA is by far the most prolific rewriter, while Brown produces very few substitutions. As expected, the more paraphrases identified, the higher the BLEU score for the method. However, this increase does

Method	Score	Substitutions
BLEU	0.0913	-
WordNet	0.0969	994
ContextWN	0.0962	742
LSA	0.992	2080
Brown	0.921	117

Table 6: Scores and the number of substitutions made for all 1788 segments, averaged over the different MT system translations

Method	Judge 1 accuracy	Judge 2 accuracy	Kappa
WordNet	63.5%	62.5%	0.74
ContextWN	**75%**	**76.0%**	0.69
LSA	30%	31.5%	0.73
Brown	56%	56%	0.72

Table 7: Accuracy scores by two human judges as well as the Kappa coefficient of agreement.

not translate into better evaluation performance. For instance, our contextual filtering method removes approximately a quarter of the paraphrases suggested by WordNet and yields a better evaluation measure. These results suggest that the substitution frequency cannot predict the utility value of the paraphrasing method.

Accuracy measures the correctness of the proposed substitutions in the context of a reference sentence. To evaluate the accuracy of different paraphrasing methods, we randomly extracted 200 paraphrasing examples from each method. A paraphrase example consists of a reference sentence, a reference word to be paraphrased and a proposed paraphrase of that reference (that actually occurred in a corresponding system translation). The judge was instructed to mark a substitution as correct only if the substitution was both semantically and grammatically correct in the context of the original reference sentence.

Paraphrases produced by the four methods were judged by two native English speakers. The pairs were presented in random order, and the judges were not told which system produced a given pair. We employ a commonly used measure, Kappa, to assess agreement between the judges. We found that

	negative	positive
filtered	40	27
non-filtered	33	100

Table 8: Confusion matrix for the context filtering method on a random sample of 200 examples labeled by the first judge.

on all the four sets the Kappa value was around 0.7, which corresponds to substantial agreement (Landis and Koch, 1977).

As Table 7 shows, the ranking between the accuracy of the different paraphrasing methods mirrors the ranking of the corresponding MT evaluation methods shown in Table 4. The paraphrasing method with the highest accuracy, ContextWN, contributes most significantly to the evaluation performance of BLEU. Interestingly, even methods with moderate accuracy, i.e. 63% for WordNet, have a positive influence on the BLEU metric. At the same time, poor paraphrasing accuracy, such as LSA with 30%, does hurt the performance of automatic evaluation.

To further understand the contribution of contextual filtering, we compare the substitutions made by WordNet and ContextWN on the same set of sentences. Among the 200 paraphrases proposed by WordNet, 73 (36.5%) were identified as incorrect by human judges. As the confusion matrix in Table 8 shows, 40 (54.5%) were eliminated during the filtering step. At the same time, the filtering erroneously eliminates 27 positive examples (21%). Even at this level of false negatives, the filtering has an overall positive effect.

5 Conclusion and Future Work

This paper presents a comprehensive study of the impact of paraphrases on the accuracy of automatic evaluation. We found a strong connection between the quality of automatic paraphrases as judged by humans and their contribution to automatic evaluation. These results have two important implications: (1) refining standard measures such as BLEU with paraphrase information moves the automatic evaluation closer to human evaluation and (2) applying paraphrases to MT evaluation provides a task-based assessment for paraphrasing accuracy.

We also introduce a novel paraphrasing method based on contextual substitution. By posing the paraphrasing problem as a discriminative task, we can incorporate a wide range of features that improve the paraphrasing accuracy. Our experiments show improvement of the accuracy of WordNet paraphrasing and we believe that this method can similarly benefit other approaches that use lexico-semantic resources to obtain paraphrases.

Our ultimate goal is to develop a contextual filtering method that does not require candidate selection based on a lexico-semantic resource. One source of possible improvement lies in exploring more powerful learning frameworks and more sophisticated linguistic representations. Incorporating syntactic dependencies and class-based features into the context representation could also increase the accuracy and the coverage of the method. Our current method only implements rewriting at the word level. In the future, we would like to incorporate substitutions at the level of phrases and syntactic trees.

Acknowledgments

The authors acknowledge the support of the National Science Foundation (Barzilay; CAREER grant IIS-0448168) and DARPA (Kauchak; grant HR0011-06-C-0023). Thanks to Michael Collins, Charles Elkan, Yoong Keok Lee, Philip Koehn, Igor Malioutov, Ben Snyder and the anonymous reviewers for helpful comments and suggestions. Any opinions, findings and conclusions expressed in this material are those of the author(s) and do not necessarily reflect the views of DARPA or NSF.

References

B. Babych, A. Hartley. 2004. Extending the BLEU evaluation method with frequency weightings. In *Proceedings of the ACL*, 621–628.

S. Banerjee, A. Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, 65–72.

R. Barzilay, L. Lee. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of NAACL-HLT*, 16–23.

P. F. Brown, P. V. deSouza, R. L. Mercer. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18:467–479.

I. Dagan, O. Glickman, B. Magnini, eds. 2005. *The PASCAL recognizing textual entailment challenge*, 2005.

P. Edmonds, G. Hirst. 2002. Near synonymy and lexical choice. *Computational Linguistics*, 28(2):105–144.

A. Haghighi, A. Ng, C. Manning. 2005. Robust textual inference via graph matching. In *Proceedings of NAACL-HLT*, 387–394.

S. Harabagiu, D. Moldovan, M. Pasca, R. Mihalcea, M. Surdeanu, R. Bunescu, R. Girju, V. Rus, P. Morarescu. 2001. The role of lexico-semantic feedback in open-domain textual question-answering. In *Proceedings of ACL*, 274–291.

P. Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP*, 388–395.

J. R. Landis, G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

C. Lin, F. Och. 2004. ORANGE: a method for evaluating automatic evaluation metrics for machine translation. In *Proceedings of COLING*, 501–507.

I. D. Melamed, R. Green, J. P. Turian. 2003. Precision and recall of machine translation. In *Proceedings of NAACL-HLT*, 61–63.

S. Miller, J. Guinness, A. Zamanian. 2004. Name tagging with word clusters and discriminative training. In *Proceedings of HLT-NAACL*, 337–342.

NIST. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics, 2002.

B. Pang, K. Knight, D. Marcu. 2003. Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. In *Proceedings of NAACL-HLT*, 102–209.

K. Papineni, S. Roukos, T. Ward, W. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the ACL*, 311–318.

C. Quirk, C. Brockett, W. Dolan. 2004. Monolingual machine translation for paraphrase generation. In *Proceedings of EMNLP*, 142–149.

R. E. Schapire, Y. Singer. 2000. Boostexter: A boosting-based system for text categorization. *Machine Learning*, 39(2/3):135–168.

R. Serafin, B. D. Eugenio. 2004. FLSA: Extending latent semantic analysis with features for dialogue act classification. In *Proceedings of the ACL*, 692–699.

A. Stent, M. Marge, M. Singhai. 2005. Evaluating evaluation methods for generation in the presense of variation. In *Proceedings of CICLING*, 341–351.