A Re-examination of Machine Learning Approaches for Sentence-Level MT Evaluation

Joshua S. Albrecht and Rebecca Hwa
Department of Computer Science
University of Pittsburgh
{jsa8,hwa}@cs.pitt.edu

Abstract

Recent studies suggest that machine learning can be applied to develop good automatic evaluation metrics for machine translated sentences. This paper further analyzes aspects of learning that impact performance. We argue that previously proposed approaches of training a *Human-Likeness classifier* is not as well correlated with human judgments of translation quality, but that *regression-based learning* produces more reliable metrics. We demonstrate the feasibility of regression-based metrics through empirical analysis of learning curves and generalization studies and show that they can achieve higher correlations with human judgments than standard automatic metrics.

1 Introduction

As machine translation (MT) research advances, the importance of its evaluation also grows. Efficient evaluation methodologies are needed both for facilitating the system development cycle and for providing an unbiased comparison between systems. To this end, a number of automatic evaluation metrics have been proposed to approximate human judgments of MT output quality. Although studies have shown them to correlate with human judgments at the document level, they are not sensitive enough to provide reliable evaluations at the sentence level (Blatz et al., 2003). This suggests that current metrics do not fully reflect the set of criteria that people use in judging sentential translation quality.

A recent direction in the development of metrics for sentence-level evaluation is to apply machine learning to create an improved composite metric out of less indicative ones (Corston-Oliver et al., 2001; Kulesza and Shieber, 2004). Under the assumption that good machine translation will produce "human-like" sentences, classifiers are trained to predict whether a sentence is authored by a human or by a machine based on features of that sentence, which may be the sentence's scores from individual automatic evaluation metrics. The confidence of the classifier's prediction can then be interpreted as a judgment on the translation quality of the sentence. Thus, the composite metric is encoded in the confidence scores of the classification labels.

While the learning approach to metric design offers the promise of ease of combining multiple metrics and the potential for improved performance, several salient questions should be addressed more fully. First, is learning a "Human Likeness" classifier the most suitable approach for framing the MT-evaluation question? An alternative is regression, in which the composite metric is explicitly learned as a function that approximates humans' quantitative judgments, based on a set of human evaluated training sentences. Although regression has been considered on a small scale for a single system as confidence estimation (Quirk, 2004), this approach has not been studied as extensively due to scalability and generalization concerns. Second, how does the diversity of the model features impact the learned metric? Third, how well do learning-based metrics generalize beyond their training examples? In particular, how well can a metric that was developed based

Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 880–887,
Prague, Czech Republic, June 2007.

on one group of MT systems evaluate the translation qualities of new systems?

In this paper, we argue for the viability of a regression-based framework for sentence-level MT-evaluation. Through empirical studies, we first show that having an accurate Human-Likeness classifier does not necessarily imply having a good MT-evaluation metric. Second, we analyze the resource requirement for regression models for different sizes of feature sets through learning curves. Finally, we show that SVM-regression metrics generalize better than SVM-classification metrics in their evaluation of systems that are different from those in the training set (by languages and by years), and their correlations with human assessment are higher than standard automatic evaluation metrics.

2 MT Evaluation

Recent automatic evaluation metrics typically frame the evaluation problem as a comparison task: how *similar* is the machine-produced output to a set of human-produced reference translations for the same source text? However, as the notion of similarity is itself underspecified, several different families of metrics have been developed. First, similarity can be expressed in terms of string edit distances. In addition to the well-known word error rate (WER), more sophisticated modifications have been proposed (Tillmann et al., 1997; Snover et al., 2006; Leusch et al., 2006). Second, similarity can be expressed in terms of common word sequences. Since the introduction of BLEU (Papineni et al., 2002) the basic n-gram precision idea has been augmented in a number of ways. Metrics in the Rouge family allow for skip n-grams (Lin and Och, 2004a); Kauchak and Barzilay (2006) take paraphrasing into account; metrics such as METEOR (Banerjee and Lavie, 2005) and GTM (Melamed et al., 2003) calculate both recall and precision; METEOR is also similar to SIA (Liu and Gildea, 2006) in that word class information is used. Finally, researchers have begun to look for similarities at a deeper structural level. For example, Liu and Gildea (2005) developed the Sub-Tree Metric (STM) over constituent parse trees and the Head-Word Chain Metric (HWCM) over dependency parse trees.

With this wide array of metrics to choose from, MT developers need a way to evaluate them. One possibility is to examine whether the automatic metric ranks the human reference translations highly with respect to machine translations (Lin and Och, 2004b; Amigó et al., 2006). The reliability of a metric can also be more directly assessed by determining how well it correlates with human judgments of the same data. For instance, as a part of the recent NIST sponsored MT Evaluation, each translated sentence by participating systems is evaluated by two (non-reference) human judges on a five point scale for its *adequacy* (does the translation retain the meaning of the original source text?) and *fluency* (does the translation sound natural in the target language?). These human assessment data are an invaluable resource for measuring the reliability of automatic evaluation metrics. In this paper, we show that they are also informative in developing better metrics.

3 MT Evaluation with Machine Learning

A good automatic evaluation metric can be seen as a computational model that captures a human's decision process in making judgments about the adequacy and fluency of translation outputs. Inferring a cognitive model of human judgments is a challenging problem because the ultimate judgment encompasses a multitude of fine-grained decisions, and the decision process may differ slightly from person to person. The metrics cited in the previous section aim to capture certain aspects of human judgments. One way to combine these metrics in a uniform and principled manner is through a learning framework. The individual metrics participate as input features, from which the learning algorithm infers a composite metric that is optimized on training examples.

Reframing sentence-level translation evaluation as a classification task was first proposed by Corston-Oliver et al. (2001). Interestingly, instead of recasting the classification problem as a "Human Acceptability" test (distinguishing good translations outputs from bad one), they chose to develop a Human-Likeness classifier (distinguishing outputs seem human-produced from machine-produced ones) to avoid the necessity of obtaining manually labeled training examples. Later, Kulesza and Shieber (2004) noted that if a classifier provides a confidence score for its output, that value can be interpreted as a quantitative estimate of the input instance's translation quality. In particular, they trained an SVM classifier that makes its decisions based on a set of input features computed from the sentence to be evaluated; the distance between input feature vector and the separating hyperplane then serves as the evaluation score. The underlying assumption for both is that improving the accuracy of the classifier on the Human-Likeness test will also improve the implicit MT evaluation metric.

A more direct alternative to the classification approach is to learn via regression and explicitly optimize for a function (i.e. MT evaluation metric) that approximates human judgments in training examples. Kulesza and Shieber (2004) raised two main objections against regression for MT evaluations. One is that regression requires a large set of labeled training examples. Another is that regression may not generalize well over time, and re-training may become necessary, which would require collecting additional human assessment data. While these are legitimate concerns, we show through empirical studies (in Section 4.2) that the additional resource requirement is not impractically high, and that a regression-based metric has higher correlations with human judgments and generalizes better than a metric derived from a Human-Likeness classifier.

3.1 Relationship between Classification and Regression

Classification and regression are both processes of function approximation; they use training examples as sample instances to learn the mapping from inputs to the desired outputs. The major difference between classification and regression is that the function learned by a classifier is a set of decision boundaries by which to classify its inputs; thus its outputs are discrete. In contrast, a regression model learns a continuous function that directly maps an input to a continuous value. An MT evaluation metric is inherently a continuous function. Casting the task as a 2-way classification may be too coarse-grained. The Human-Likeness formulation of the problem introduces another layer of approximation by assuming equivalence between "Like Human-Produced" and "Well-formed" sentences. In Section 4.1, we show empirically that high accuracy in the Human-Likeness test does not necessarily entail good MT evaluation judgments.

3.2 Feature Representation

To ascertain the resource requirements for different model sizes, we considered two feature models. The smaller one uses the same nine features as Kulesza and Shieber, which were derived from BLEU and WER. The full model consists of 53 features: some are adapted from recently developed metrics; others are new features of our own. They fall into the following major categories[1]:

String-based metrics over references These include the nine Kulesza and Shieber features as well as precision, recall, and fragmentation, as calculated in METEOR; ROUGE-inspired features that are non-consecutive bigrams with a gap size of m, where $1 \leq m \leq 5$ (skip-m-bigram), and ROUGE-L (longest common subsequence).

Syntax-based metrics over references We unrolled HWCM into their individual chains of length c (where $2 \leq c \leq 4$); we modified STM so that it is computed over unlexicalized constituent parse trees as well as over dependency parse trees.

String-based metrics over corpus Features in this category are similar to those in *String-based metric over reference* except that a large English corpus is used as "reference" instead.

Syntax-based metrics over corpus A large dependency treebank is used as the "reference" instead of parsed human translations. In addition to adaptations of the *Syntax-based metrics over references*, we have also created features to verify the argument structures for certain syntactic categories.

4 Empirical Studies

In these studies, the learning models used for both classification and regression are support vector machines (SVM) with Gaussian kernels. All models are trained with SVM-Light (Joachims, 1999). Our primary experimental dataset is from NIST's 2003

[1]As feature engineering is not the primary focus of this paper, the features are briefly described here, but implementational details will be made available in a technical report.

Chinese MT Evaluations, in which the fluency and adequacy of 919 sentences produced by six MT systems are scored by two human judges on a 5-point scale[2]. Because the judges evaluate sentences according to their individual standards, the resulting scores may exhibit a biased distribution. We normalize human judges' scores following the process described by Blatz et al. (2003). The overall human assessment score for a translation output is the average of the sum of two judges' normalized fluency and adequacy scores. The full dataset ($6 \times 919 = 5514$ instances) is split into sets of training, heldout and test data. Heldout data is used for parameter tuning (i.e., the slack variable and the width of the Gaussian). When training classifiers, assessment scores are not used, and the training set is augmented with all available human reference translation sentences ($4 \times 919 = 3676$ instances) to serve as positive examples.

To judge the quality of a metric, we compute Spearman rank-correlation coefficient, which is a real number ranging from -1 (indicating perfect negative correlations) to +1 (indicating perfect positive correlations), between the metric's scores and the averaged human assessments on test sentences. We use Spearman instead of Pearson because it is a distribution-free test. To evaluate the relative reliability of different metrics, we use bootstrapping re-sampling and paired t-test to determine whether the difference between the metrics' correlation scores has statistical significance (at 99.8% confidence level)(Koehn, 2004). Each reported correlation rate is the average of 1000 trials; each trial consists of n sampled points, where n is the size of the test set. Unless explicitly noted, the qualitative differences between metrics we report are statistically significant. As a baseline comparison, we report the correlation rates of three standard automatic metrics: BLEU, METEOR, which incorporates recall and stemming, and HWCM, which uses syntax. BLEU is smoothed to be more appropriate for sentence-level evaluation (Lin and Och, 2004b), and the bigram versions of BLEU and HWCM are reported because they have higher correlations than when longer n-grams are included. This phenomenon has been previously observed by Liu and Gildea (2005).

[2]This corpus is available from the Linguistic Data Consortium as Multiple Translation Chinese Part 4.

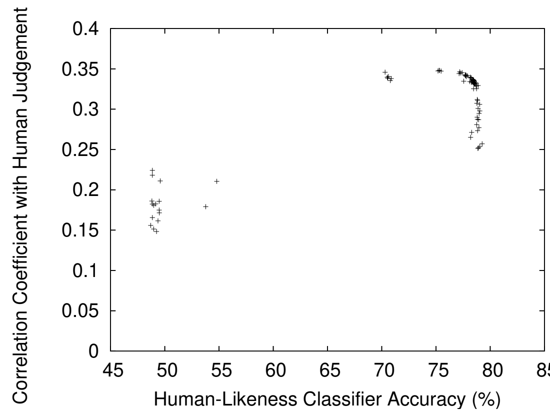


Figure 1: This scatter plot compares classifiers' accuracy with their corresponding metrics' correlations with human assessments

4.1 Relationship between Classification Accuracy and Quality of Evaluation Metric

A concern in using a metric derived from a Human-Likeness classifier is whether it would be predictive for MT evaluation. Kulesza and Shieber (2004) tried to demonstrate a positive correlation between the Human-Likeness classification task and the MT evaluation task empirically. They plotted the classification accuracy and evaluation reliability for a number of classifiers, which were generated as a part of a greedy search for kernel parameters and found some linear correlation between the two. This proof of concept is a little misleading, however, because the population of the sampled classifiers was biased toward those from the same neighborhood as the local optimal classifier (so accuracy and correlation may only exhibit linear relationship locally). Here, we perform a similar study except that we sampled the kernel parameter more uniformly (on a log scale). As Figure 1 confirms, having an accurate Human-Likeness classifier does not necessarily entail having a good MT evaluation metric. Although the two tasks do seem to be positively related, and in the limit there may be a system that is good at both tasks, one may improve classification without improving MT evaluation. For this set of heldout data, at the near 80% accuracy range, a derived metric might have an MT evaluation correlation coefficient anywhere between 0.25 (on par with unsmoothed BLEU, which is known to be unsuitable for sentence-level evaluation) and 0.35 (competitive with standard metrics).

4.2 Learning Curves

To investigate the feasibility of training regression models from assessment data that are currently available, we consider both a small and a large regression model. The smaller model consists of nine features (same as the set used by Kulesza and Shieber); the other uses the full set of 53 features as described in Section 3.2. The reliability of the trained metrics are compared with those developed from Human-Likeness classifiers. We follow a similar training and testing methodology as previous studies: we held out 1/6 of the assessment dataset for SVM parameter tuning; five-fold cross validation is performed with the remaining sentences. Although the metrics are evaluated on unseen test sentences, the sentences are produced by the same MT systems that produced the training sentences. In later experiments, we investigate generalizing to more distant MT systems.

Figure 2(a) shows the learning curves for the two regression models. As the graph indicates, even with a limited amount of human assessment data, regression models can be trained to be comparable to standard metrics (represented by METEOR in the graph). The small feature model is close to convergence after 1000 training examples[3]. The model with a more complex feature set does require more training data, but its correlation began to overtake METEOR after 2000 training examples. This study suggests that the start-up cost of building even a moderately complex regression model is not impossibly high.

Although we cannot directly compare the learning curves of the Human-Likeness classifiers to those of the regression models (since the classifier's training examples are automatically labeled), training examples for classifiers are not entirely free: human reference translations still must be developed for the source sentences. Figure 2(c) shows the learning curves for training Human-Likeness classifiers (in terms of improving a classifier's accuracy) using the same two feature sets, and Figure 2(b) shows the correlations of the metrics derived from the corresponding classifiers. The pair of graphs show, especially in the case of the larger feature set, that a large improvement in classification accuracy does not bring proportional improvement in its corresponding metrics's correlation; with an accuracy of near 90%, its correlation coefficient is 0.362, well below METEOR.

This experiment further confirms that judging Human-Likeness and judging Human-Acceptability are not tightly coupled. Earlier, we have shown in Figure 1 that different SVM parameterizations may result in classifiers with the same accuracy rate but different correlations rates. As a way to incorporate some assessment information into classification training, we modify the parameter tuning process so that SVM parameters are chosen to optimize for assessment correlations in the heldout data. By incurring this small amount of human assessed data, this parameter search improves the classifier's correlations: the metric using the smaller feature set increased from 0.423 to 0.431, and that of the larger set increased from 0.361 to 0.422.

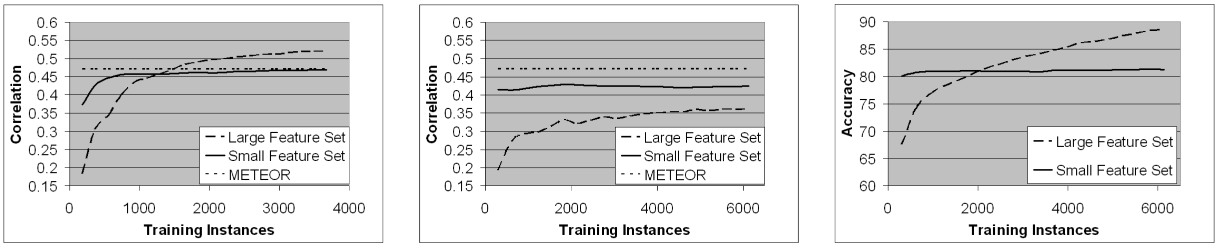
4.3 Generalization

We conducted two generalization studies. The first investigates how well the trained metrics evaluate systems from other years and systems developed for a different source language. The second study delves more deeply into how variations in the training examples affect a learned metric's ability to generalize to distant systems. The learning models for both experiments use the full feature set.

Cross-Year Generalization To test how well the learning-based metrics generalize to systems from different years, we trained both a regression-based metric (R03) and a classifier-based metric (C03) with the entire NIST 2003 Chinese dataset (using 20% of the data as heldout[4]). All metrics are then applied to three new datasets: NIST 2002 Chinese MT Evaluation (3 systems, 2634 sentences total), NIST 2003 Arabic MT Evaluation (2 systems, 1326 sentences total), and NIST 2004 Chinese MT Evaluation (10 systems, 4470 sentences total). The results

[3] The total number of labeled examples required is closer to 2000, since the heldout set uses 919 labeled examples.

[4] Here, too, we allowed the classifier's parameters to be tuned for correlation with human assessment on the heldout data rather than accuracy.



(a) (b) (c)

Figure 2: Learning curves: (a) correlations with human assessment using regression models; (b) correlations with human assessment using classifiers; (c) classifier accuracy on determining Human-Likeness.

Dataset	R03	C03	BLEU	MET.	HWCM
2002 Ara	**0.466**	0.384	0.423	0.431	0.424
2002 Chn	**0.309**	0.250	0.269	0.290	0.260
2004 Chn	**0.602**	0.566	0.588	0.563	0.546

Table 1: Correlations for cross-year generalization. Learning-based metrics are developed from NIST 2003 Chinese data. All metrics are tested on datasets from 2003 Arabic, 2002 Chinese and 2004 Chinese.

are summarized in Table 1. We see that R03 consistently has a better correlation rate than the other metrics.

At first, it may seem as if the difference between R03 and BLEU is not as pronounced for the 2004 dataset, calling to question whether a learned metric might become quickly out-dated, we argue that this is not the case. The 2004 dataset has many more participating systems, and they span a wider range of qualities. Thus, it is easier to achieve a high rank correlation on this dataset than previous years because most metrics can qualitatively discern that sentences from one MT system are better than those from another. In the next experiment, we examine the performance of R03 with respect to each MT system in the 2004 dataset and show that its correlation rate is higher for better MT systems.

Relationship between Training Examples and Generalization Table 2 shows the result of a generalization study similar to before, except that correlations are performed on each system. The rows order the test systems by their translation qualities from the best performing system (2004-Chn1, whose average human assessment score is 0.655 out of 1.0) to the worst (2004-Chn10, whose score is 0.255). In addition to the regression metric from the previous experiment (R03-all), we consider two more regression metrics trained from subsets of the 2003 dataset: R03-Bottom5 is trained from the subset that excludes the best 2003 MT system, and R03-Top5 is trained from the subset that excludes the worst 2003 MT system.

We first observe that on a per test-system basis, the regression-based metrics generally have better correlation rates than BLEU, and that the gap is as wide as what we have observed in the earlier cross-years studies. The one exception is when evaluating 2004-Chn8. None of the metrics seems to correlate very well with human judges on this system. Because the regression-based metric uses these individual metrics as features, its correlation also suffers.

During regression training, the metric is optimized to minimize the difference between its prediction and the human assessments of the training data. If the input feature vector of a test instance is in a very distant space from training examples, the chance for error is higher. As seen from the results, the learned metrics typically perform better when the training examples include sentences from higher-quality systems. Consider, for example, the differences between R03-all and R03-Top5 versus the differences between R03-all and R03-Bottom5. Both R03-Top5 and R03-Bottom5 differ from R03-all by one subset of training examples. Since R03-all's correlation rates are generally closer to R03-Top5 than to R03-Bottom5, we see that having seen extra training examples from a bad system is not as harmful as having not seen training examples from a good system. This is expected, since there are many ways to create bad translations, so seeing a partic-

	R03-all	R03-Bottom5	R03-Top5	BLEU	METEOR	HWCM
2004-Chn1	0.495	0.460	**0.518**	0.456	0.457	0.444
2004-Chn2	0.398	0.330	**0.440**	0.352	0.347	0.344
2004-Chn3	0.425	0.389	**0.459**	0.369	0.402	0.369
2004-Chn4	**0.432**	0.392	0.434	0.400	0.400	0.362
2004-Chn5	**0.452**	0.441	0.443	0.370	0.426	0.326
2004-Chn6	*0.405*	0.392	**0.406**	0.390	0.357	0.380
2004-Chn7	0.443	0.432	**0.448**	0.390	0.408	0.392
2004-Chn8	0.237	0.256	0.256	**0.265**	0.259	0.179
2004-Chn9	0.581	0.569	**0.591**	0.527	0.537	0.535
2004-Chn10	0.314	0.313	**0.354**	0.321	0.303	0.358
2004-all	0.602	0.567	**0.617**	0.588	0.563	0.546

Table 2: Metric correlations within each system. The columns specify which metric is used. The rows specify which MT system is under evaluation; they are ordered by human-judged system quality, from best to worst. For each evaluated MT system (row), the highest coefficient in bold font, and those that are statistically comparable to the highest are shown in italics.

ular type of bad translations from one system may not be very informative. In contrast, the neighborhood of good translations is much smaller, and is where all the systems are aiming for; thus, assessments of sentences from a good system can be much more informative.

4.4 Discussion

Experimental results confirm that learning from training examples that have been doubly approximated (class labels instead of ordinals, human-likeness instead of human-acceptability) does negatively impact the performance of the derived metrics. In particular, we showed that they do not generalize as well to new data as metrics trained from direct regression.

We see two lingering potential objections toward developing metrics with regression-learning. One is the concern that a system under evaluation might try to explicitly "game the metric[5]." This is a concern shared by all automatic evaluation metrics, and potential problems in stand-alone metrics have been analyzed (Callison-Burch et al., 2006). In a learning framework, potential pitfalls for individual metrics are ameliorated through a combination of evidences. That said, it is still prudent to defend against the potential of a system gaming a subset of the features. For example, our fluency-predictor features are not strong indicators of translation qualities by themselves. We want to avoid training a metric that assigns a higher than deserving score to a sentence that just happens to have many n-gram matches against the target-language reference corpus. This can be achieved by supplementing the current set of human assessed training examples with automatically assessed training examples, similar to the labeling process used in the Human-Likeness classification framework. For instance, as negative training examples, we can incorporate fluent sentences that are not adequate translations and assign them low overall assessment scores.

A second, related concern is that because the metric is trained on examples from current systems using currently relevant features, even though it generalizes well in the near term, it may not continue to be a good predictor in the distant future. While periodic retraining may be necessary, we see value in the flexibility of the learning framework, which allows for new features to be added. Moreover, adaptive learning methods may be applicable if a small sample of outputs of some representative translation systems is manually assessed periodically.

5 Conclusion

Human judgment of sentence-level translation quality depends on many criteria. Machine learning affords a unified framework to compose these criteria into a single metric. In this paper, we have demonstrated the viability of a regression approach to learning the composite metric. Our experimental results show that by training from some human as-

[5] Or, in a less adversarial setting, a system may be performing minimum error-rate training (Och, 2003)

sessments, regression methods result in metrics that have better correlations with human judgments even as the distribution of the tested population changes.

Acknowledgments

This work has been supported by NSF Grants IIS-0612791 and IIS-0710695. We would like to thank Regina Barzilay, Ric Crabbe, Dan Gildea, Alex Kulesza, Alon Lavie, and Matthew Stone as well as the anonymous reviewers for helpful comments and suggestions. We are also grateful to NIST for making their assessment data available to us.

References

Enrique Amigó, Jesús Giménez, Julio Gonzalo, and Lluís Màrquez. 2006. MT evaluation: Human-like vs. human acceptable. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, Sydney, Australia, July.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for MT evaluation with improved correlation with human judgments. In *ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, June.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2003. Confidence estimation for machine translation. Technical Report Natural Language Engineering Workshop Final Report, Johns Hopkins University.

Christopher Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *The Proceedings of the Thirteenth Conference of the European Chapter of the Association for Computational Linguistics*.

Simon Corston-Oliver, Michael Gamon, and Chris Brockett. 2001. A machine learning approach to the automatic evaluation of machine translation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, July.

Thorsten Joachims. 1999. Making large-scale SVM learning practical. In Bernhard Schöelkopf, Christopher Burges, and Alexander Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press.

David Kauchak and Regina Barzilay. 2006. Paraphrasing for automatic evaluation. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, New York City, USA, June.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-04)*.

Alex Kulesza and Stuart M. Shieber. 2004. A learning approach to improving sentence-level MT evaluation. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, Baltimore, MD, October.

Gregor Leusch, Nicola Ueffing, and Hermann Ney. 2006. CDER: Efficient MT evaluation using block movements. In *The Proceedings of the Thirteenth Conference of the European Chapter of the Association for Computational Linguistics*.

Chin-Yew Lin and Franz Josef Och. 2004a. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, July.

Chin-Yew Lin and Franz Josef Och. 2004b. Orange: a method for evaluating automatic evaluation metrics for machine translation. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, August.

Ding Liu and Daniel Gildea. 2005. Syntactic features for evaluation of machine translation. In *ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, June.

Ding Liu and Daniel Gildea. 2006. Stochastic iterative alignment for machine translation evaluation. In *Proceedings of the Joint Conference of the International Conference on Computational Linguistics and the Association for Computational Linguistics (COLING-ACL'2006) Poster Session*, July.

I. Dan Melamed, Ryan Green, and Joseph Turian. 2003. Precision and recall of machine translation. In *In Proceedings of the HLT-NAACL 2003: Short Papers*, pages 61–63, Edmonton, Alberta.

Franz Josef Och. 2003. Minimum error rate training for statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, PA.

Christopher Quirk. 2004. Training a sentence-level machine translation confidence measure. In *Proceedings of LREC 2004*.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 8th Conference of the Association for Machine Translation in the Americas (AMTA-2006)*.

Christoph Tillmann, Stephan Vogel, Hermann Ney, Hassan Sawaf, and Alex Zubiaga. 1997. Accelerated DP-based search for statistical translation. In *Proceedings of the 5th European Conference on Speech Communication and Technology (EuroSpeech '97)*.