Regression for Sentence-Level MT Evaluation with Pseudo References

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

Abstract

Many automatic evaluation metrics for machine translation (MT) rely on making comparisons to human translations, a resource that may not always be available. We present a method for developing sentence-level MT evaluation metrics that do not directly rely on human reference translations. Our metrics are developed using regression learning and are based on a set of weaker indicators of fluency and adequacy (*pseudo references*). Experimental results suggest that they rival standard reference-based metrics in terms of correlations with human judgments on new test instances.

1 Introduction

Automatic assessment of translation quality is a challenging problem because the evaluation task, at its core, is based on subjective human judgments. Reference-based metrics such as BLEU (Papineni et al., 2002) have rephrased this subjective task as a somewhat more objective question: how closely does the translation resemble sentences that are known to be good translations for the same source? This approach requires the participation of human translators, who provide the "gold standard" reference sentences. However, keeping humans in the evaluation loop represents a significant expenditure both in terms of time and resources; therefore it is worthwhile to explore ways of reducing the degree of human involvement.

To this end, Gamon et al. (2005) proposed a learning-based evaluation metric that does not compare against reference translations. Under a learning framework, the input (i.e., the sentence to be evaluated) is represented as a set of *features*. These are measurements that can be extracted from the input sentence (and may be individual metrics themselves). The learning algorithm combines the features to form a model (a composite evaluation metric) that produces the final score for the input. Without human references, the features in the model proposed by Gamon et al. were primarily language model features and linguistic indicators that could be directly derived from the input sentence alone. Although their initial results were not competitive with standard reference-based metrics, their studies suggested that a referenceless metric may still provide useful information about translation fluency. However, a potential pitfall is that systems might "game the metric" by producing fluent outputs that are not adequate translations of the source.

This paper proposes an alternative approach to evaluate MT outputs without comparing against human references. While our metrics are also trained, our model consists of different features and is trained under a different learning regime. Crucially, our model includes features that capture some notions of adequacy by comparing the input against *pseudo references*: sentences from other MT systems (such as commercial off-the-shelf systems or open sourced research systems). To improve fluency judgments, the model also includes features that compare the input against target-language "references" such as large text corpora and treebanks.

Unlike human translations used by standard reference-based metrics, pseudo references are not

Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 296–303,
Prague, Czech Republic, June 2007.

"gold standards" and can be worse than the sentences being evaluated; therefore, these "references" in-and-of themselves are not necessarily informative enough for MT evaluation. The main insight of our approach is that through regression, the trained metrics can make more nuanced comparisons between the input and pseudo references. More specifically, our regression objective is to infer a function that maps a feature vector (which measures an input's similarity to the pseudo references) to a score that indicates the quality of the input. This is achieved by optimizing the model's output to correlate against a set of training examples, which are translation sentences labeled with quantitative assessments of their quality by human judges. Although this approach does incur some human effort, it is primarily for the development of training data, which, ideally, can be amortized over a long period of time.

To determine the feasibility of the proposed approach, we conducted empirical studies that compare our trained metrics against standard reference-based metrics. We report three main findings. First, pseudo references are informative comparison points. Experimental results suggest that a regression-trained metric that compares against pseudo references can have higher correlations with human judgments than applying standard metrics with multiple human references. Second, the learning model that uses both adequacy and fluency features performed the best, with adequacy being the more important factor. Third, when the pseudo references are multiple MT systems, the regression-trained metric is predictive even when the input is from a better MT system than those providing the references. We conjecture that comparing MT outputs against other imperfect translations allows for a more nuanced discrimination of quality.

2 Background and Related Work

For a formally organized event, such as the annual MT Evaluation sponsored by National Institute of Standard and Technology (NIST MT Eval), it may be worthwhile to recruit multiple human translators to translate a few hundred sentences for evaluation references. However, there are situations in which multiple human references are not practically available (e.g., the source may be of a large quantity, and no human translation exists). One such instance is translation quality assurance, in which one wishes to identify poor outputs in a large body of machine translated text automatically for human to post-edit. Another instance is in day-to-day MT research and development, where new test set with multiple references are also hard to come by. One could work with previous datasets from events such as the NIST MT Evals, but there is a danger of over-fitting. One also could extract a single reference from parallel corpora, although it is known that automatic metrics are more reliable when comparing against multiple references.

The aim of this work is to develop a trainable automatic metric for evaluation without human references. This can be seen as a form of confidence estimation on MT outputs (Blatz et al., 2003; Ueffing et al., 2003; Quirk, 2004). The main distinction is that confidence estimation is typically performed with a particular system in mind, and may rely on system-internal information in estimation. In this study, we draw on only system-independent indicators so that the resulting metric may be more generally applied. This allows us to have a clearer picture of the contributing factors as they interact with different types of MT systems.

Also relevant is previous work that applied machine learning approaches to MT evaluation, both with human references (Corston-Oliver et al., 2001; Kulesza and Shieber, 2004; Albrecht and Hwa, 2007; Liu and Gildea, 2007) and without (Gamon et al., 2005). One motivation for the learning approach is the ease of combining multiple criteria. Literature in translation evaluation reports a myriad of criteria that people use in their judgments, but it is not clear how these factors should be combined mathematically. Machine learning offers a principled and unified framework to induce a computational model of human's decision process. Disparate indicators can be encoded as one or more input features, and the learning algorithm tries to find a mapping from input features to a score that quantifies the input's quality by optimizing the model to match human judgments on training examples. The framework is attractive because its objective directly captures the goal of MT evaluation: how would a user rate the quality of these translations?

This work differs from previous approaches in two aspects. One is the representation of the model; our model treats the metric as a distance measure even though there are no human references. Another is the training of the model. More so than when human references are available, regression is central to the success of the approach, as it determines how much we can trust the distance measures against each pseudo reference system.

While our model does not use human references directly, its features are adapted from the following distance-based metrics. The well-known BLEU (Papineni et al., 2002) is based on the number of common n-grams between the translation hypothesis and human reference translations of the same sentence. Metrics such as ROUGE, Head Word Chain (HWC), METEOR, and other recently proposed methods all offer different ways of comparing machine and human translations. ROUGE utilizes 'skip n-grams', which allow for matches of sequences of words that are not necessarily adjacent (Lin and Och, 2004a). METEOR uses the Porter stemmer and synonym-matching via WordNet to calculate recall and precision more accurately (Banerjee and Lavie, 2005). The HWC metrics compare dependency and constituency trees for both reference and machine translations (Liu and Gildea, 2005).

3 MT Evaluation with Pseudo References using Regression

Reference-based metrics are typically thought of as measurements of "similarity to good translations" because human translations are used as references, but in more general terms, they are distance measurements between two sentences. The distance between a translation hypothesis and an imperfect reference is still somewhat informative. As a toy example, consider a one-dimensional line segment. A distance from the end-point uniquely determines the position of a point. When the reference location is anywhere else on the line segment, a relative distance to the reference does not uniquely specify a location on the line segment. However, the position of a point can be uniquely determined if we are given its relative distances to two reference locations.

The problem space for MT evaluation, though more complex, is not dissimilar to the toy scenario. There are two main differences. First, we do not know the actual distance function – this is what we are trying to learn. The distance functions we have at our disposal are all heuristic approximations to the true translational distance function. Second, unlike human references, whose quality value is assumed to be maximum, the quality of a pseudo reference sentence is not known. In fact, prior to training, we do not even know the quality of the reference systems. Although the direct way to calibrate a reference system is to evaluate *its* outputs, this is not practically ideal, since human judgments would be needed each time we wish to incorporate a new reference system. Our proposed alternative is to calibrate the reference systems against an existing set of human judgments for a range of outputs from different MT systems. That is, if many of the reference system's outputs are similar to those MT outputs that received low assessments, we conclude this reference system may not be of high quality. Thus, if a new translation is found to be similar with this reference system's output, it is more likely for the new translation to also be bad.

Both issues of combining evidences from heuristic distances and calibrating the quality of pseudo reference systems can be addressed by a probabilistic learning model. In particular, we use regression because its problem formulation fits naturally with the objective of MT evaluations. In regression learning, we are interested in approximating a function f that maps a multi-dimensional input vector, $\mathbf{x}$, to a continuous real value, y, such that the error over a set of m training examples, $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$, is minimized according to a loss function.

In the context of MT evaluation, y is the "true" quantitative measure of translation quality for an input sentence[1]. The function f represents a mathematical model of human judgments of translations; an input sentence is represented as a feature vector, $\mathbf{x}$, which contains the information that can be extracted from the input sentence (possibly including comparisons against some reference sentences) that are relevant to computing y. Determining the set of relevant features for this modeling is on-going re-

[1]Perhaps even more so than grammaticality judgments, there is variability in people's judgments of translation quality. However, like grammaticality judgments, people do share some similarities in their judgments at a coarse-grained level. Ideally, what we refer to as the true value of translational quality should reflect the consensus judgments of all people.

search. In this work, we consider some of the more widely used metrics as features. Our full feature vector consists of $r \times 18$ adequacy features, where r is the number of reference systems used, and 26 fluency features:

Adequacy features: These include features derived from BLEU (e.g., n-gram precision, where $1 \leq n \leq 5$, length ratios), PER, WER, features derived from METEOR (precision, recall, fragmentation), and ROUGE-related features (non-consecutive bigrams with a gap size of g, where $1 \leq g \leq 5$ and longest common subsequence).

Fluency features: We consider both string-level features such as computing n-gram precision against a target-language corpus as well as several syntax-based features. We parse each input sentence into a dependency tree and compared aspects of it against a large target-language dependency treebank. In addition to adapting the idea of Head Word Chains (Liu and Gildea, 2005), we also compared the input sentence's argument structures against the treebank for certain syntactic categories.

Due to the large feature space to explore, we chose to work with support vector regression as the learning algorithm. As its loss function, support vector regression uses an ϵ-insensitive error function, which allows for errors within a margin of a small positive value, ϵ, to be considered as having zero error (cf. Bishop (2006), pp.339-344). Like its classification counterpart, this is a kernel-based algorithm that finds sparse solutions so that scores for new test instances are efficiently computed based on a subset of the most informative training examples. In this work, Gaussian kernels are used.

The cost of regression learning is that it requires training examples that are manually assessed by human judges. However, compared to the cost of creating new references whenever new (test) sentences are evaluated, the effort of creating human assessment training data is a limited (ideally, one-time) cost. Moreover, there is already a sizable collection of human assessed data for a range of MT systems through multiple years of the NIST MT Eval efforts. Our experiments suggest that there is enough assessed data to train the proposed regression model.

Aside from reducing the cost of developing human reference translations, the proposed metric also provides an alternative perspective on automatic MT evaluation that may be informative in its own right. We conjecture that a metric that compares inputs against a diverse population of differently imperfect sentences may be more discriminative in judging translation systems than solely comparing against gold standards. That is, two sentences may be considered equally bad from the perspective of a gold standard, but subtle differences between them may become more prominent if they are compared against sentences in their peer group.

4 Experiments

We conducted experiments to determine the feasibility of the proposed approach and to address the following questions: (1) How informative are pseudo references in-and-of themselves? Does varying the number and/or the quality of the references have an impact on the metrics? (2) What are the contributions of the adequacy features versus the fluency features to the learning-based metric? (3) How do the quality and distribution of the training examples, together with the quality of the pseudo references, impact the metric training? (4) Do these factors impact the metric's ability in assessing sentences produced within a single MT system? How does that system's quality affect metric performance?

4.1 Data preparation and Experimental Setup

The implementation of support vector regression used for these experiments is SVM-Light (Joachims, 1999). We performed all experiments using the 2004 NIST Chinese MT Eval dataset. It consists of 447 source sentences that were translated by four human translators as well as ten MT systems. Each machine translated sentence was evaluated by two human judges for their fluency and adequacy on a 5-point scale[2]. To remove the bias in the distributions of scores between different judges, we follow the normalization procedure described by Blatz et al. (2003). The two judge's total scores (i.e., sum of the normalized fluency and adequacy scores) are then averaged.

[2]The NIST human judges use human reference translations when making assessments; however, our approach is generally applicable when the judges are bilingual speakers who compare source sentences with translation outputs.

We chose to work with this NIST dataset because it contains numerous systems that span over a range of performance levels (see Table 1 for a ranking of the systems and their averaged human assessment scores). This allows us to have control over the variability of the experiments while answering the questions we posed above (such as the quality of the systems providing the pseudo references, the quality of MT systems being evaluated, and the diversity over the distribution of training examples).

Specifically, we reserved four systems (MT2, MT5, MT6, and MT9) for the role of pseudo references. Sentences produced by the remaining six systems are used as evaluative data. This set includes the best and worst systems so that we can see how well the metrics performs on sentences that are better (or worse) than the pseudo references. Metrics that require no learning are directly applied onto all sentences of the evaluative set. For the learning-based metrics, we perform six-fold cross validation on the evaluative dataset. Each fold consists of sentences from one MT system. In a round robin fashion, each fold serves as the test set while the other five are used for training and heldout. Thus, the trained models have seen neither the test instances nor other instances from the MT system that produced them.

A metric is evaluated based on its Spearman rank correlation coefficient between the scores it gave to the evaluative dataset and human assessments for the same data. The correlation coefficient is a real number between -1, indicating perfect negative correlations, and +1, indicating perfect positive correlations. To compare the relative quality of different metrics, we apply bootstrapping re-sampling on the data, and then use paired t-test to determine the statistical significance of the correlation differences (Koehn, 2004). For the results we report, unless explicitly mentioned, all stated comparisons are statistically significant with 99.8% confidence. We include two standard reference-based metrics, BLEU and METEOR, as baseline comparisons. BLEU is smoothed (Lin and Och, 2004b), and it considers only matching up to bigrams because this has higher correlations with human judgments than when higher-ordered n-grams are included.

SysID	Human-assessment score
MT1	0.661
MT2	**0.626**
MT3	0.586
MT4	0.578
MT5	**0.537**
MT6	**0.530**
MT7	0.530
MT8	0.375
MT9	**0.332**
MT10	0.243

Table 1: The human-judged quality of ten participating systems in the NIST 2004 Chinese MT Evaluation. We used four systems as references (highlighted in boldface) and the data from the remaining six for training and evaluation.

4.2 Pseudo Reference Variations vs. Metrics

We first compare different metrics' performance on the six-system evaluative dataset under different configurations of human and/or pseudo references. For the case when only one human reference is used, the reference was chosen at random from the 2004 NIST Eval dataset[3]. The correlation results on the evaluative dataset are summarized in Table 2.

Some trends are as expected: comparing within a metric, having four references is better than having just one; having human references is better than an equal number of system references; having a high quality system as reference is better than one with low quality. Perhaps more surprising is the consistent trend that metrics do significantly better with four MT references than with one human reference, and they do almost as well as using four human references. The results show that pseudo references are informative, as standard metrics were able to make use of the pseudo references and achieve higher correlations than judging from fluency alone. However, higher correlations are achieved when learning with regression, suggesting that the trained metrics are better at interpreting comparisons against pseudo references.

Comparing within each reference configuration, the regression-trained metric that includes both ad-

[3]One reviewer asked about the quality this human's translations. Although we were not given official rankings of the human references, we compared each person against the other three using MT evaluation metrics and found this particular translator to rank third, though the quality of all four are significantly higher than even the best MT systems.

equacy and fluency features always has the highest correlations. If the metric consists of only adequacy features, its performance degrades with the decreasing quality of the references. At another extreme, a metric based only on fluency features has an overall correlation rate of 0.459, which is lower than most correlations reported in Table 2. This confirms the importance of modeling adequacy; even a single mid-quality MT system may be an informative pseudo reference. Finally, we note that a regression-trained metric with the full features set that compares against 4 pseudo references has a higher correlation than BLEU with four human references. These results suggest that the feedback from the human assessed training examples was able to help the learning algorithm to combine different features to form a better composite metric.

4.3 Sentence-Level Evaluation on Single Systems

To explore the interaction between the quality of the reference MT systems and that of the test MT systems, we further study the following pseudo reference configurations: all four systems, a high-quality system with a medium quality system, two systems of medium-quality, one medium with one poor system, and only the high-quality system. For each pseudo reference configuration, we consider three metrics: BLEU, METEOR, and the regression-trained metric (using the full feature set). Each metric evaluates sentences from four test systems of varying quality: the best system in the dataset (MT1), the worst in the set (MT10), and two mid-ranged systems (MT4 and MT7). The correlation coefficients are summarized in Table 3. Each row specifies a metric/reference-type combination; each column specifies an MT system being evaluated (using sentences from all other systems as training examples). The fluency-only metric and standard metrics using four human references are baselines.

The overall trends at the dataset level generally also hold for the per-system comparisons. With the exception of the evaluation of MT10, regression-based metrics always has higher correlations than standard metrics that use the same reference configuration (comparing correlation coefficients within each cell). When the best MT reference system (MT2) is included as pseudo references, regression-based metrics are typically better than or not statistically different from standard applications of BLEU and METEOR with 4 human references. Using the two mid-quality MT systems as references (MT5 and MT6), regression metrics yield correlations that are only slightly lower than standard metrics with human references. These results support our conjecture that comparing against multiple systems is informative.

The poorer performances of the regression-based metrics on MT10 point out an asymmetry in the learning approach. The regression model aims to learn a function that approximates human judgments of translated sentences through training examples. In the space of all possible MT outputs, the neighborhood of good translations is much smaller than that of bad translations. Thus, as long as the regression models sees some examples of sentences with high assessment scores during training, it should have a much better estimation of the characteristics of good translations. This idea is supported by the experimental data. Consider the scenario of evaluating MT1 while using two mid-quality MT systems as references. Although the reference systems are not as high quality as the system under evaluation, and although the training examples shown to the regression model were also generated by systems whose overall quality was rated lower, the trained metric was reasonably good at ranking sentences produced by MT1. In contrast, the task of evaluating sentences from MT10 is more difficult for the learning approach, perhaps because it is sufficiently different from all training and reference systems. Correlations might be improved with additional reference systems.

4.4 Discussions

The design of these experiments aims to simulate practical situations to use our proposed metrics. For the more frequently encountered language pairs, it should be possible to find at least two mid-quality (or better) MT systems to serve as pseudo references. For example, one might use commercial off-the-shelf systems, some of which are free over the web. For less commonly used languages, one might use open source research systems (Al-Onaizan et al., 1999; Burbank et al., 2005).

Datasets from formal evaluation events such as

Ref type and #	Ref Sys.	BLEU-S(2)	METEOR	Regr (adq. only)	Regr (full)
4 Humans	all humans	0.628	0.591	0.588	**0.644**
1 Human	HRef #3	0.536	0.512	0.487	0.597
4 Systems	all MTRefs	0.614	0.583	0.584	*0.632*
2 Systems	Best 2 MTRefs	0.603	0.577	0.573	0.620
	Mid 2 MTRefs	0.579	0.555	0.528	0.608
	Worst 2 MTRefs	0.541	0.508	0.467	0.581
1 System	Best MTRef	0.576	0.559	0.534	0.596
	Mid MTRef (MT5)	0.538	0.528	0.474	0.577
	Worst MTRef	0.371	0.329	0.151	0.495

Table 2: Comparisons of metrics (columns) using different types of references (rows). The full regression-trained metric has the highest correlation (shown in boldface) when four human references are used; it has the second highest correlation rate (shown in italic) when four MT system references are used instead. A regression-trained metric with only fluency features has a correlation coefficient of 0.459.

Ref Type	Metric	MT-1	MT-4	MT-7	MT-10
No ref	Regr.	0.367	0.316	0.301	-0.045
4 human refs	Regr.	**0.538***	**0.473***	**0.459***	0.247
	BLEU-S(2)	0.466	0.419	0.397	0.321*
	METEOR	0.464	0.418	0.410	0.312
4 MTRefs	Regr.	**0.498**	**0.429**	**0.421**	0.243
	BLEU-S(2)	0.386	0.349	0.404	0.240
	METEOR	0.445	0.354	0.333	0.243
Best 2 MTRefs	Regr.	**0.492**	*0.418*	0.403	0.201
	BLEU-S(2)	0.391	0.330	0.394	0.268
	METEOR	0.430	0.333	0.327	0.267
Mid 2 MTRefs	Regr.	0.450	0.413	0.388	0.219
	BLEU-S(2)	0.362	0.314	0.310	0.282
	METEOR	0.391	0.315	0.284	0.274
Worst 2 MTRefs	Regr.	0.430	0.386	0.365	0.158
	BLEU-S(2)	0.320	0.298	0.316	0.223
	METEOR	0.351	0.306	0.302	0.228
Best MTRef	Regr.	0.461	0.401	**0.414**	0.122
	BLEU-S(2)	0.371	0.330	0.380	0.242
	METEOR	0.375	0.318	0.392	0.283

Table 3: Correlation comparisons of metrics by test systems. For each test system (columns) the overall highest correlations is distinguished by an asterisk (*); correlations higher than *standard metrics using human-references* are highlighted in boldface; those that are statistically comparable to them are italicized.

NIST MT Evals, which contains human assessed MT outputs for a variety of systems, can be used for training examples. Alternatively, one might directly recruit human judges to assess sample sentences from the system(s) to be evaluated. This should result in better correlations than what we reported here, since the human assessed training examples will be more similar to the test instances than the setup in our experiments.

In developing new MT systems, pseudo references may supplement the single human reference translations that could be extracted from a parallel text. Using the same setup as Exp. 1 (see Table 2), adding pseudo references does improve correlations. Adding four pseudo references to the single human reference raises the correlation coefficient to 0.650 (from 0.597) for the regression metric. Adding them to four human references results in a correlation coefficient of 0.660 (from 0.644)[4].

5 Conclusion

In this paper, we have presented a method for developing sentence-level MT evaluation metrics without using human references. We showed that by learning from human assessed training examples,

[4]BLEU with four human references has a correlation of 0.628. Adding four pseudo references increases BLEU to 0.650.

the regression-trained metric can evaluate an input sentence by comparing it against multiple machine-generated pseudo references and other target language resources. Our experimental results suggest that the resulting metrics are robust even when the sentences under evaluation are from a system of higher quality than the systems serving as references. We observe that regression metrics that use multiple pseudo references often have comparable or higher correlation rates with human judgments than standard reference-based metrics. Our study suggests that in conjunction with regression training, multiple imperfect references may be as informative as gold-standard references.

Acknowledgments

This work has been supported by NSF Grants IIS-0612791 and IIS-0710695. We would like to thank Ric Crabbe, Dan Gildea, Alon Lavie, Stuart Shieber, and Noah Smith and the anonymous reviewers for their suggestions. We are also grateful to NIST for making their assessment data available to us.

References

Yaser Al-Onaizan, Jan Curin, Michael Jahr, Kevin Knight, John Lafferty, I. Dan Melamed, Franz-Josef Och, David Purdy, Noah A. Smith, and David Yarowsky. 1999. Statistical machine translation. Technical report, JHU. citeseer.nj.nec.com/al-onaizan99statistical.html.

Joshua S. Albrecht and Rebecca Hwa. 2007. A re-examination of machine learning approaches for sentence-level MT evaluation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-2007)*.

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.

Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer Verlag.

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.

Andrea Burbank, Marine Carpuat, Stephen Clark, Markus Dreyer, Declan Groves Pamela. Fox, Keith Hall, Mary Hearne, I. Dan Melamed, Yihai Shen, Andy Way, Ben Wellington, and Dekai Wu. 2005. Final report of the 2005 language engineering workshop on statistical machine translation by parsing. Technical Report Natural Language Engineering Workshop Final Report, "JHU".

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.

Michael Gamon, Anthony Aue, and Martine Smets. 2005. Sentence-level MT evaluation without reference translations: Beyond language modeling. In *European Association for Machine Translation (EAMT)*, May.

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.

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.

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. 2007. Source-language features and maximum correlation training for machine translation evaluation. In *Proceedings of the HLT/NAACL-2007*, April.

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*.

Nicola Ueffing, Klaus Macherey, and Hermann Ney. 2003. Confidence measures for statistical machine translation. In *Machine Translation Summit IX*, pages 394–401, September.