A Learning Approach to Improving Sentence-Level MT Evaluation

Alex Kulesza and Stuart M. Shieber
Division of Engineering and Applied Sciences
Harvard University
33 Oxford St.
Cambridge, MA 02138, USA
kulesza@post.harvard.edu
shieber@deas.harvard.edu

Abstract

The problem of evaluating machine translation (MT) systems is more challenging than it may first appear, as diverse translations can often be considered equally correct. The task is even more difficult when practical circumstances require that evaluation be done automatically over short texts, for instance, during incremental system development and error analysis. While several automatic metrics, such as BLEU, have been proposed and adopted for large-scale MT system discrimination, they all fail to achieve satisfactory levels of correlation with human judgments at the sentence level. Here, a new class of metrics based on machine learning is introduced. A novel method involving classifying translations as machine or human-produced rather than directly predicting numerical human judgments eliminates the need for labor-intensive user studies as a source of training data. The resulting metric, based on support vector machines, is shown to significantly improve upon current automatic metrics, increasing correlation with human judgments at the sentence level halfway toward that achieved by an independent human evaluator.

1 Introduction

The problem of evaluating machine translation systems has necessarily received significant attention alongside the development of MT systems themselves. As it specifically aims to address the needs of humans who communicate with language, MT is most naturally and effectively evaluated through the manual efforts of such users, and extensive research has been conducted in order to describe, test, and taxonomize various methods and considerations involved in doing so (Hovy et al. 2002; Van Slype 1979). However, human evaluation of MT systems, while providing the most direct and reliable assessment, has numerous drawbacks; primarily prohibitive are the time and expense involved in organizing and executing a user study. Subjects must be located, trained, presented with evaluation materials, and compensated; furthermore, to alleviate biases due to academic background or bilingual experience, a large number of subjects is usually preferable.

Therefore, while manual methods may remain viable for isolated, large-scale evaluations, researchers should ideally benefit from the ability to quickly and accurately assess their own systems repeatedly as new ideas are implemented; in these situations, automatic methods have the potential to expedite the development of successful MT systems by greatly reducing the resources required for evaluation. In addition, new statistical MT systems have successfully incorporated training

methods that directly optimize their results on automatic evaluation metrics (Och 2003). Having reliable metrics therefore allows MT systems to be not only tested but trained very rapidly.

In recent years, a number of automatic metrics have been proposed and used for evaluating the overall performance of MT systems. General-purpose measures like word error rate (WER) and position-independent word error rate (PER) (Tillman et al. 1997) rely on direct correspondence between the machine translation and a single human-produced reference. The more specialized metrics BLEU (Papineni et al. 2001) and NIST (Doddington 2002) consider the fraction of output n-grams that also appear in a *set* of human translations (n-gram precision), thereby acknowledging a greater diversity of acceptable MT results. The F-Measure has been proposed as a more comprehensible alternative for MT evaluation (Turian et al. 2003), and can be defined as a simple composite of unigram precision and recall. These metrics have demonstrated some success; the BLEU metric, for example, has been shown to correlate highly with the judgments of bilingual human evaluators (a correlation coefficient of 0.96) when averaged over a 500-sentence corpus (Papineni et al. 2001).

However, though useful for system discrimination, metrics correlating with human evaluations only over long texts (which tend to average out the "noise" of evaluation) are relatively useless for purposes of error analysis; they are too blunt to provide specific feedback. A metric that can accurately gauge the quality of short-text translations has the potential advantage of illuminating the individual sentences or phrases with which the MT system has the most trouble and consequently leading to improvements in its performance. Furthermore, applications such as confidence estimation rely directly on local automatic assessments (Foster et al. 2003). Since short-text evaluations can be averaged into evaluations of larger texts, a meaningful correlation for short texts is always preferable.

Experiments conducted at the 2003 Johns Hopkins Workshop on Speech and Language Engineering have suggested that the currently available automatic MT evaluation metrics provide insufficient correlation when considered at the level of an individual sentence (Foster et al. 2003). The Confidence Estimation team, after collecting human judgments of 633 single-sentence hypothesis translations on a scale from 1 to 5, reports that, although inter-judge correlation is disappointingly low at 0.46 (a sentiment shared by Turian et al. (2003) with respect to similar studies), the correlation between human judges and automatic metrics is significantly lower. Results here show that BLEU with three references exhibits a correlation of 0.25 on the same set of translations. None of the standard automatic metrics considered does better than 0.29, indicating significant room for improvement of automatic metrics at the sentence-level.

Others have also noted poor automatic metric performance, particularly with respect to short translations. Turian et al. (2003) report, based on experiments rank-correlating the outputs of BLEU, NIST, and the F-Measure with human judgments on translations from 1 to 100 sentences in length, that "even though human evaluation of MT is itself inconsistent and not very reliable, automatic MT evaluation measures are even less reliable and are still very far from being able to replace human judgment." (Turian et al. 2003) The Syntax for Statistical Machine Translation team from the 2003 Workshop further notes in its final report that the BLEU metric, used for training and evaluation of the team's MT system, seems insensitive to syntactic changes that should be noticeable to human judges at the sentence level (Och et al. 2003).

Here, a metric designed to evaluate single-sentence machine translations based on machine

learning is described and tested. A novel approach leads to a flexible classification-based metric and eliminates the need for expensive human evaluation data during training. The result significantly outperforms standard automatic metrics, improving correlation with human judgments to 0.38 at the sentence level.

2 Approach

It is at first tempting to suggest the use of general-purpose machine-learning methods as a means of directly approximating human evaluations. Given a large training set of human-evaluated hypothesis translations and reference counterparts, it should be possible to directly learn human evaluation scores as a function of some feature set generated over the input. However, this approach is problematic for at least two reasons. First, since learning methods are driven by training data, to successfully learn evaluation scores would require the initial development of a large set of human evaluations and, consequently, consume large amounts of time and money. If such a project were planned, it would need to involve a very carefully designed methodology in order to ensure that the evaluation data received is of the best possible quality and the greatest possible longevity. Given the extensive research history of human MT evaluation methods and practices (and the problematic lack of interannotator agreement already noted even in simple experiments), such a task might itself be very difficult or even prohibitive.

Second, even if such a gigantic resource could be created, it would be necessarily static, representing a fixed distribution of MT outputs on a fixed set of language pairs. With the constant evolution of MT systems, hypothesis translations judged in the project evaluations and subsequently used to generate metrics would eventually fail to adequately represent the population of MT outputs being considered. The training set would then no longer reflect the true distribution of hypothesis translations; under these conditions, machine-learning approaches are likely to fail.

It is crucial, then, that any machine-learning approach to automatic MT evaluation sustain the ability to be retrained and itself reevaluated with respect to modern and representative translation samples at regular intervals. In the limit, one would like to retrain whenever a new group of MT systems (or a new iteration of possibilities for a single system under development) is to be evaluated, but if doing so requires the process of human evaluation, then the advantage of the automatic evaluator is neutralized.

To provide the necessary flexibility, as well as to alleviate the problems inherent in producing human evaluations of MT outputs, an alternate training criterion is proposed. Instead of attempting to perform a direct regression on human evaluations of hypothesis translations, a simpler classification problem might be more feasible: has a given hypothesis translation been produced by a machine or by a human? This question has the distinct advantage of being answerable with existing information—training sets can be as large as the corpora for which we have reference translations. Furthermore, because human-produced reference translations can be fixed for a particular corpus, they remain a one-time startup cost and can be used repeatedly, regardless of any innovations in the design of MT systems. A metric predicated upon an ability to distinguish machine translations from human translations, therefore, can be retrained at will on new distributions of hypothesis MT outputs.

The classification criterion, of course, does not directly lend itself to the goal of improving correlation with human evaluators; however, evidence suggests that, at least currently, MT outputs and human translations are easily distinguished—and it goes without saying that human translations are superior (Foster et al. 2003). Indeed, if MT outputs were of such quality that they could not be distinguished from human translations, evaluation would no longer be necessary; by definition such an MT system would be successful. The proposed classification is therefore likely to correctly establish at least broad performance groups. Furthermore, by applying a method for classification that produces a continuous "confidence" measure, the classification criterion can actually induce an evaluator for which the output is viewed as some measure of certainty that a particular translation was produced by a human. If, in fact, the question of translation quality is closely related to the believability of a translation being human-produced, then such a system can reasonably expect to achieve success as an MT evaluation metric.

A learning approach also carries the great advantage of flexibility. In addressing problems such as the lack of sensitivity to syntactic features or the desire to ensure that evaluators continue to distinguish modern systems as they improve, new features can be designed and incorporated into a retrained model. Learning-based automatic metrics, therefore, have the potential to form a large class of customizable, infinitely adjustable evaluation solutions. Though the continually changing nature of such a metric implies that its scores will hold little absolute meaning, the advantages with respect to detailed error analysis and adaptability might be significant.

3 Implementation

To explore the approach outlined above, an automatic metric has been built by training a support vector machine classifier to distinguish machine translations from human translations and extracting a continuous measure of classification confidence. Performance of the new metric is evaluated by computing its correlation with a set of human judgments over single-sentence translations and comparing the result to those of the standard automatic metrics WER, PER, BLEU, and the F-Measure.

3.1 Obtaining a Metric from a Classifier

The support vector machine (SVM) is used here as a learning mechanism due to its useful conception of classification as a problem of linear separation in some complex feature space (Cristianini 2001). After training, the SVM produces a separator defined by a vector w, a constant b, and an inner product $\langle \cdot, \cdot \rangle$ in the feature space. The classification of a test example x is determined by the side of the separator on which it falls: $\mathrm{sign}(\langle w, x \rangle - b)$. In this case, the result determines whether the translation is thought to have been produced by a human or by a machine.

However, the separator also acts as an organization over the (feature) space of translations, defining the half-spaces of machine- and human-produced translations as well as a boundary subspace in which translations are equally human- and machine-like. By computing not only the side of the boundary on which a particular example falls but also the *distance* between that example and the boundary (removing the operator sign in the expression above), a measure of confidence is obtained.

Examples buried deep in the machine half-space are perhaps very clearly machine produced, while those close to the boundary might have qualities seen more commonly in translations produced by humans, even if they still fall into the "machine" class. Those machine-produced examples that appear in the human half-space, on the other hand, have successfully fooled the classifier, and are likely to be of very high quality. Signed distance from the separator is therefore used as the continuous output of the learning-based evaluator. The Torch3 machine learning library implementation of SVMs for classification is used here with Gaussian kernels (Collobert et al. 2002).

3.2 Features

In order to use the general purpose SVM for classifying translations, linguistic objects must be re-represented in numerical form with a vector of feature values. The features used here come directly from those considered by standard automatic metrics:

- Unmodified n-gram precisions for $1 \leq n \leq 5$, computed as the fraction of hypothesis n-grams appearing in any reference. (Only one hypothesis n-gram may match to any particular reference n-gram.) This feature corresponds to the basic unit of information used by BLEU and NIST.
- The minimum and maximum ratio of hypothesis length to reference length over the set of reference translations. This feature corresponds to the length penalties of BLEU and NIST.
- Word error rate, computed as the minimum edit distance between the hypothesis and any reference, where words are the units of deletion, insertion, and substitution.
- Position-independent word error rate, computed by removing the words in the shorter translation from those in the longer translation and returning the minimum size of the remaining set (over all references).

4 Results

The methods described produce an automatic machine translation evaluation metric that significantly outperforms current automatic metrics, as measured by the sentence level correlation of the metric outputs with human judgments for a test set of 633 machine-produced hypotheses. Additionally, results demonstrate that optimizing for the modified training criterion in fact strongly encourages correlation with the true evaluations, implying that successful machine learning approaches need not require expensive human judgments. Statistical significance measures over correlation coefficients given in the following sections are derived using Fisher's z' transformation.

4.1 Data Sets

Data for the training and testing of machine-learning based metrics are drawn from the alignment template statistical machine translation system that obtained the best results in the 2002 and 2003

Table 1: Data for figure 1

Metric	Correlation Coefficient
Human	0.4633
SVM	0.3771
WER	0.2909
F-Measure	0.2861
PER	0.2794
BLEU	0.2537

DARPA Chinese to English MT evaluations. Hypotheses are English translations of a single Chinese source sentence, produced either by the above MT system and drawn uniformly at random from a 100-best list or by a human translator, and references consist of three additional human translations of the same sentence. Care is taken to ensure that the hypothesis is never produced by the same translator as any of the references.

In total, 21,144 examples are used, half of which contain machine-produced hypotheses and half of which contain human-produced hypotheses. These examples are split approximately 2:1 into training and validation sets of 14,120 and 7,024, respectively; an equal number of human and machine hypotheses are maintained in each set. The machine-learning layer is optimized over the larger training set and then tested for classification accuracy on the validation set. The final assessment of any resulting metric, however, depends not on its classification abilities but on its continuous-output correlation with human judgments. For measuring this quality, a third data set is used—the test set—consisting of 633 hypothesis translations, all produced by the above MT system and evaluated by human judges on a scale from 1 to 5 during an experiment at the 2003 JHU Workshop. Ratings are percentile-normalized to the users reporting them in the manner described by Foster et al. (2003).

4.2 Improved Correlation with Human Judgments

Using a grid search over SVM hyper-parameters C, the trade-off between margin maximization and error minimization, and σ, the standard deviation of the Gaussian kernel, maximum validation performance is found at $C = 50$ and $\sigma = 10$. The resulting overall classification accuracy is 64.4%, with 58.7% accuracy classifying human translations and 70.0% accuracy classifying machine translations. Classification accuracy, however, does not reflect the success of the metric; instead the desired characteristic is a strong correlation with human judgments. For the selected model, the correlation coefficient with respect to human evaluations over the test set is 0.38.

Figure 1 shows the levels of correlation obtained by an independent human judge (the human metric), the SVM-based metric using optimal parameters, and various current automatic MT evaluation metrics. Numerical data are presented in table 1. The SVM-based metric outperforms all other automatic evaluators at 95% significance. The SVM-based metric fails to reach the performance level of a human judge, but makes up approximately half of the gap between the best previously known automatic metrics and the human evaluator.

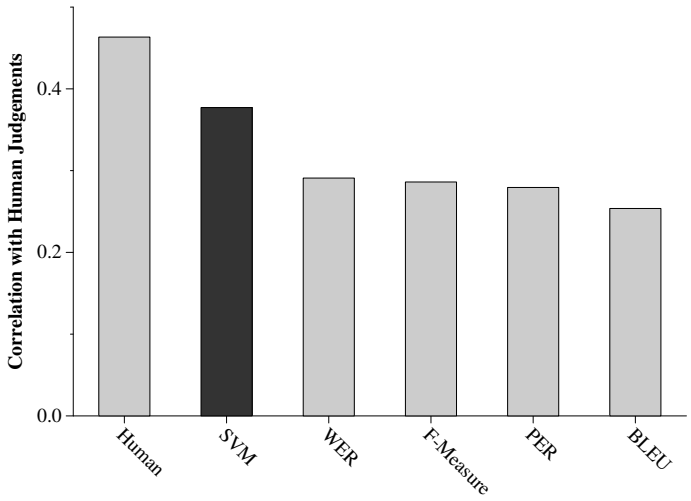

Figure 1: Correlation coefficients for the human metric, a machine-learned SVM metric, and four standard automatic metrics. The gap between the machine-learned metric and any other automatic metric is statistically significant at 95%, as is the gap between the human and machine-learned metrics. Performance differences among the four standard automatic metrics, however, are not significant.

4.3 Success of the Modified Criterion

In addition to producing an automatic MT evaluation metric that outperforms current automatic metrics, results show that the classification criterion is also generally successful as an approximation to the goal of correlation with human judgments. For a modified criterion to be successful, procedures that optimize for that criterion must be shown to also optimize for the desired or true criterion; in this case, knowing that improving the classification accuracy of a SVM model tends to improve its correlation with human judgments allows a metric designer to do without an expensive test set and simply seek to improve classification accuracy with faith that the desired effect will result. Figure 2 shows the strong empirical relationship observed between classification accuracy and human judgment correlation during the grid search over SVM parameters. Significant positive correlation is apparent (correlation coefficient of 0.855), indicating that tweaking SVM parameters to maximize classification accuracy tends to encourage correlation with human judgments as well.

This result justifies the somewhat blind model selection performed earlier. The optimal parameters exhibit not only the highest classification accuracy—observable using only the validation set, which is generated automatically—but also a human judgment correlation within 1.5% of the best achieved by any SVM-based metric. This kind of ability to choose a successful metric without relying on resource-intensive human judgments is the key to the approach: once the reliability of the modified criterion is established, new metrics can be implemented, optimized, and applied to modern MT systems without requiring any additional expensive data.

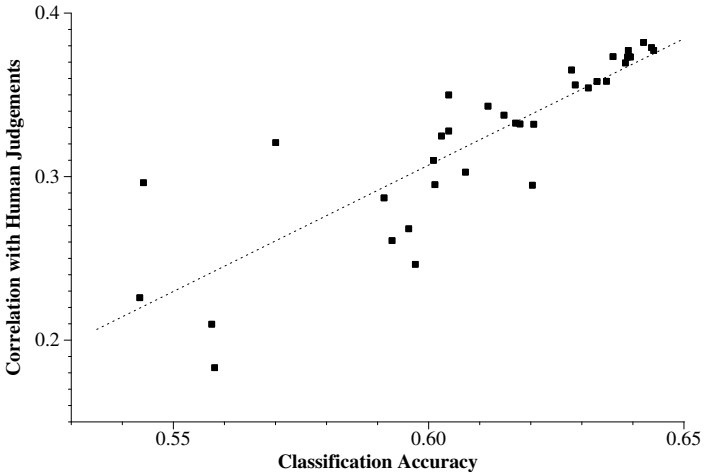

Figure 2: Automatic metric correlation with human judgments versus classification accuracy for a wide range of learning parameters. The results suggest that training for the simplified classification criterion induces the strong correlation necessary for a successful evaluation metric. The meta-correlation between classification accuracy and correlation results for SVM classifiers is 0.855, significant at 99%.

5 Conclusions

Applying machine learning to the problem of machine translation evaluation is primarily difficult because the desired human judgment targets are too expensive to make available in large quantities, particularly as the population of MT outputs may be changing constantly. The solution proposed here is to approximate human judgments with a binary decision variable that reports whether a translation was produced by a human or by a machine, thereby eliminating the need for user data collection. By utilizing measures of confidence, continuous judgments can be learned from this simple binary classification. Results indicate that this approach is effective, improving sentence-level correlation between metric scores and human judgments from 0.29 to 0.38, and that for a fixed set of input features, the classification criterion is strongly linked with such correlation. The method thus provides the opportunity to obtain improved correlation even without access to a human evaluated test set.

5.1 Future Work

Other related methods might be applied to the problem. Unsupervised learning might produce an organization of the data in which translations of a similar quality are located near each other; such a map could then be translated to a continuous metric with only a few human evaluations as reference points. More generally, active learning methods would allow a variable cost/performance trade-off,

using as much evaluation data as available while still taking advantage of a large unlabeled training set. These methods have the potential, however, to discover patterns in the data that are irrelevant to the task, and might end up grouping translations by some irrelevant characteristic. Whether such a characteristic drawn from the data would be any more or less informative than that of having been produced by a human or machine (effectively enforced here as the property of organization) is an open question.

A final aspect of machine learning approaches to MT evaluation that deserves attention is the degree of customization that they provide in addition to merely improved performance. While the focus here has been on improving correlation with human judgments, system designers might also find it useful to add features tuned specifically to issues of concern, thereby improving the metric's sensitivity to particular aspects of machine translations. For example, it has been observed that BLEU is not particularly sensitive to syntactic improvements (Och et al. 2003); a learning-based metric, on the other hand, could be modified specifically to take syntax into account, providing a much finer-grained error analysis than is currently possible.

6 Acknowledgements

Many thanks to the people of the 2003 Johns Hopkins Workshop on Speech and Language Engineering, especially George Foster, Simona Gandrabur, Cyril Goutte, and Franz Och. This material is based upon work supported by the National Science Foundation under Grant No. 0121285. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Collobert, Ronan, Samy Bengio, and Johnny Mariethoz. 2002. Torch: a modular machine learning software library. Tech. Rep. IDIAP-RR 02-46, IDIAP.

Cristianini, Nello. 2001. Support vector and kernel methods for learning. ICML Tutorial.

Doddington, George. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Human Language Technology: Notebook Proceedings*, 128–132. San Diego.

Foster, George, Simona Gandrabur, Cyril Goutte, Erin Fitzgerald, Alberto Sanchis, Nicola Ueffing, John Blatz, and Alex Kulesza. 2003. Confidence estimation for machine translation. Tech. Rep., Johns Hopkins Workshop on Speech and Language Engineering, Baltimore, MD.

Hovy, Eduard, Margaret King, and Andrei Popescu-Belis. 2002. Principles of context-based machine translation evaluation. *Machine Translation* 16:1–33.

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

Och, Franz Josef, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir Radev. 2003. Syntax for statistical machine translation. Tech. Rep., Johns Hopkins Workshop on Speech and Language Engineering, Baltimore, MD.

Papineni, Kishore A., Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: A method for automatic evaluation of machine translation. Tech. Rep. RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, NY.

Tillman, C., S. Vogel, H. Ney, H. Sawaf, and A. Zubiaga. 1997. Accelerated DP-based search for statistical translation. In *Proceedings of the 5th European Conference on Speech Communication and Technology (EuroSpeech '97)*, 2667–2670. Rhodes, Greece.

Turian, Joseph P., Luke Shen, and I. Dan Melamed. 2003. Evaluation of machine translation and its evaluation. In *Proceedings of* MT *Summit* IX, 23–28. New Orleans.

Van Slype, Georges. 1979. Critical methods for evaluating the quality of machine translation. Tech. Rep. BR-19142, European Commission Directorate General Scientific and Technical Information and Information Management, Bureau Marcel van Dijk.