The Role of Pseudo References in MT Evaluation

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

Abstract

Previous studies have shown automatic evaluation metrics to be more reliable when compared against many human translations. However, multiple human references may not always be available. It is more common to have only a single human reference (extracted from parallel texts) or no reference at all. Our earlier work suggested that one way to address this problem is to train a metric to evaluate a sentence by comparing it against *pseudo references*, or imperfect "references" produced by off-the-shelf MT systems. In this paper, we further examine the approach both in terms of the training methodology and in terms of the role of the human and pseudo references. Our expanded experiments show that the approach generalizes well across multiple years and different source languages.

1 Introduction

Standard automatic metrics are *reference-based*; that is, they compare system-produced translations against human-translated references produced for the same source. Since there is usually no single best way to translate a sentence, each MT output should be compared against many references. On the other hand, creating multiple human references is itself a costly process. For many naturally occurring datasets (e.g., parallel corpora) only a single reference is readily available.

The focus of this work is on developing automatic metrics for sentence-level evaluation with *at most one human reference*. One way to supplement the single human reference is to use *pseudo references*, or sentences produced by off-the-shelf MT systems, as stand-ins for human references. However, since pseudo references may be imperfect translations themselves, the comparisons cannot be fully trusted. Previously, we have taken a learning-based approach to develop a composite metric that combines measurements taken from multiple pseudo references (Albrecht and Hwa, 2007). Experimental results suggested the approach to be promising; but those studies did not consider how well the metric might generalize across multiple years and different languages. In this paper, we investigate the applicability of the pseudo-reference metrics under these more general conditions.

Using the WMT06 Workshop shared-task results (Koehn and Monz, 2006) as training examples, we train a metric that evaluates new sentences by comparing them against pseudo references produced by three off-the-shelf MT systems. We apply the learned metric to sentences from the WMT07 shared-task (Callison-Burch et al., 2007b) and compare the metric's predictions against human judgments. We find that additional pseudo references improve correlations for automatic metrics.

2 Background

The ideal evaluation metric reports an accurate distance between an input instance and its gold standard, but even when comparing against imperfect standards, the measured distances may still convey some useful information – they may help to triangulate the input's position relative to the true gold standard.

In the context of sentence-level MT evaluations,

Proceedings of the Third Workshop on Statistical Machine Translation, pages 187–190,
Columbus, June 2008.

the challenges are two-fold. First, the ideal quantitative distance function between a translation hypothesis and the proper translations is not known; current automatic evaluation metrics produce approximations to the true translational distance. Second, although we may know the qualitative goodness of the MT systems that generate the pseudo references, we do not know how imperfect the pseudo references are. These uncertainties make it harder to establish the true distance between the input hypothesis and the (unobserved) acceptable gold standard translations.

In order to combine evidence from these uncertain observations, we take a learning-based approach. Each hypothesis sentence is compared with multiple pseudo references using multiple metrics. Representing the measurements as a set of input features and using human-assessed MT sentences as training examples, we train a function that is optimized to correlate the features with the human assessments in the training examples. Specifically, for each input sentence, we compute a set of 18 kinds of reference-based measurements for each pseudo reference as well as 26 monolingual fluency measurements. The full set of measurements then serves as the input feature vector into the function, which is trained via support vector regression. The learned function can then be used as an evaluation metric itself: it takes the measurements of a new sentence as input and returns a composite score for that sentence.

The approach is considered successful if the metric's predictions on new test sentences correlate well with quantitative human assessments. Like other learned models, the metric is expected to perform better on data that are more similar to the training instances. Therefore, a natural question that arises with a metric developed in this manner is: how well does it generalize?

3 Research Questions

To better understand the capability of metrics that compare against pseudo-references, we consider the following aspects:

The role of learning Standard reference-based metrics can also use pseudo references; however, they would treat the imperfect references as gold standard. In contrast, the learning process aims to determine how much each comparison with a pseudo reference might be trusted. To observe the role of learning, we compare trained metrics against standard reference-based metrics, all using pseudo references.

The amount vs. types of training data The success of any learned model depends on its training experiences. We study the trade-off between the size of the training set and the specificity of the training data. We perform experiments comparing a metric trained from a large pool of heterogeneous training examples that include translated sentences from multiple languages and individual metrics trained from particular source languages.

The role of a single human reference Previous studies have shown the importance of comparing against multiple references. The approach in this paper attempts to approximate multiple human references with machine-produced sentences. Is a single trust-worthy translation more useful than multiple imperfect translations? To answer this question, we compare three different reference settings: using just a single human reference, using just the three pseudo references, and using all four references.

4 Experimental Setup

For the experiments reported in this paper, we used human-evaluated MT sentences from past shared-tasks of the WMT 2006 and WMT 2007. The data consists of outputs from German-English, Spanish-English, and French-English MT systems. The outputs are translations from two corpora: *Europarl* and *news commentary*. System outputs have been evaluated by human judges on a 5-point scale (Callison-Burch et al., 2007a). We have normalized scores to reduce biases from different judges (Blatz et al., 2003).

We experimented with using four different subsets of the WMT2006 data as training examples: only German-English, only Spanish-English, only French-English, all 06 data. The metrics are trained using support vector regression with a Gaussian kernel as implemented in the SVM-Light package (Joachims, 1999). The SVM parameters are tuned via grid-search on development data, 20% of the full training set that has been reserved for this purpose.

We used three MT systems to generate pseudo references: Systran[1], GoogleMT [2], and Moses (Koehn et al., 2007). We chose these three systems because they are widely accessible and because they take relatively different approaches. Moreover, although they have not all been human-evaluated in the past WMT shared tasks, they are well-known for producing good translations.

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.

Two standard reference-based metrics, BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005), are used for comparisons. BLEU is smoothed (Lin and Och, 2004), and it considers only matching up to bigrams because this has higher correlations with human judgments than when higher-ordered n-grams are included.

5 Results

The full experimental comparisons are summarized in Table 1. Each cell shows the correlation coefficient between the human judgments and a metric (column) that uses a particular kind of references (row) for some evaluation data set (block row).

The role of learning With the exception of the German-English data, the learned metrics had higher correlations with human judges than the baselines, which used standard metrics with a single human reference. On the other hand, results suggest that pseudo references often also improve correlations for standard metrics. This may seem counter-intuitive because we can easily think of cases in which pseudo references hurt standard metrics (e.g., use poor outputs as pseudo references). We hypothesize that because the pseudo references came from high-quality MT systems and because standard metrics are based on simple word matches, the chances for bad judgments (input words matched against pseudo reference, but both are wrong) are relatively small compared to chances for good judgments. We further hypothesize that the learned metrics would be robust against the qualities of the pseudo reference MT systems.

The amount vs. types of training data Comparing the three metrics trained from single language datasets against the metric trained from all of WMT06 dataset, we see that the learning process benefitted from the larger quantity of training examples. It may be the case that the MT systems for the three language pairs are at a similar stage of maturity such that the training instances are mutually helpful.

The role of a single human reference Our results reinforce previous findings that metrics are more reliable when they have access to more than a single human reference. Our experimental data suggests that a single human reference often may not be as reliable as using three pseudo references alone. Finally, the best correlations are achieved by using both human and pseudo references.

6 Conclusion

We have presented an empirical study on automatic metrics for sentence-level MT evaluation with at most one human reference. We show that pseudo references from off-the-shelf MT systems can be used to augment the single human reference. Because they are imperfect, it is important to weigh the trustworthiness of these references through a training phase. The metric seems robust even when the applied to sentences from different systems of a later year. These results suggest that multiple imperfect translations make informative comparison points in supplement to human references.

Acknowledgments

This work has been supported by NSF Grants IIS-0612791.

[1]Available from `http://www.systransoft.com/`. We note that Systran is also a participating system under evaluation. Although Sys-Test will be deemed to be identical to Sys-Ref, it will not automatically receive a high score because the measurement is weighted by whether Sys-Ref was reliable during training. Furthermore, measurements between Sys-Test and other pseudo-references will provide alternative evidences for the metric to consider.

[2]`http://www.google.com/language_tools/`

Eval. Data	Ref Type	METEOR	BLEU	SVM(de06)	SVM(es06)	SVM(fr06)	SVM(wmt06)
de	1HR	0.458	*0.471*				
europarl	3PR	0.521*	0.527*	0.422	0.403	0.480*	0.467
07	1HR+3PR	0.535*	0.547*	0.471	0.480*	0.477*	0.523*
de	1HR	0.290	*0.333*				
news	3PR	0.400*	0.400*	0.262	0.279	0.261	0.261
07	1HR+3PR	0.432*	0.417*	0.298	0.321	0.269	0.330
es	1HR	0.377	*0.412*				
europarl	3PR	0.453*	0.483*	0.336	0.453*	0.432*	0.456*
07	1HR+3PR	0.491*	0.503*	0.405	0.513*	0.483*	0.510*
es	1HR	0.317	*0.332*				
news	3PR	0.320	0.317	0.393*	0.381*	0.426*	0.426*
07	1HR+3PR	0.353*	0.325	0.429*	0.427*	0.380*	0.486*
fr	1HR	*0.265*	0.246				
europarl	3PR	0.196	0.285*	0.270*	0.284*	0.355*	0.366*
07	1HR+3PR	0.221	0.290*	0.277*	0.324*	0.304*	0.381*
fr	1HR	0.226	*0.280*				
news	3PR	0.356*	0.383*	0.237	0.252	0.355*	0.373*
07	1HR+3PR	0.374*	0.394*	0.272	0.339*	0.319*	0.388*

Table 1: Correlation comparisons of metrics (columns) using different references (row): a single human reference (1HR), 3 pseudo references (3PR), or all (1HR+3PR). The type of training used for the regression-trained metrics are specified in parentheses. For each evaluated corpus, correlations higher than *standard metric using one human reference* are marked by an asterisk(*).

References

Joshua S. Albrecht and Rebecca Hwa. 2007. Regression for sentence-level MT evaluation with pseudo references. 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.

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.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007a. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic, June. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Cameron Shaw Fordyce, and Christof Monz, editors. 2007b. *Proceedings of the Second Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Prague, Czech Republic, June.

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 and Christof Monz, editors. 2006. *Proceedings on the Workshop on Statistical Machine Translation*. Association for Computational Linguistics, New York City, June.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. *Proceedings of ACL, Demonstration Session*.

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

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.