The Significance of Recall in Automatic Metrics for MT Evaluation

Alon Lavie and Kenji Sagae and Shyamsundar Jayaraman

Language Technologies Institute
Carnegie Mellon University
{alavie,sagae,shyamj}@cs.cmu.edu

Abstract. Recent research has shown that a balanced harmonic mean (F1 measure) of unigram precision and recall outperforms the widely used BLEU and NIST metrics for Machine Translation evaluation in terms of correlation with human judgments of translation quality. We show that significantly better correlations can be achieved by placing more weight on recall than on precision. While this may seem unexpected, since BLEU and NIST focus on n-gram precision and disregard recall, our experiments show that correlation with human judgments is highest when almost all of the weight is assigned to recall. We also show that stemming is significantly beneficial not just to simpler unigram precision and recall based metrics, but also to BLEU and NIST.

1 Introduction

Automatic Metrics for machine translation (MT) evaluation have been receiving significant attention in the past two years, since IBM's BLEU metric was proposed and made available [1]. BLEU and the closely related NIST metric [2] have been extensively used for comparative evaluation of the various MT systems developed under the DARPA TIDES research program, as well as by other MT researchers. Several other automatic metrics for MT evaluation have been proposed since the early 1990s. These include various formulations of measures of "edit distance" between an MT-produced output and a reference translation [3] [4], and similar measures such as "word error rate" and "position-independent word error rate" [5], [6].

The utility and attractiveness of automatic metrics for MT evaluation has been widely recognized by the MT community. Evaluating an MT system using such automatic metrics is much faster, easier and cheaper compared to human evaluations, which require trained bilingual evaluators. In addition to their utility for comparing the performance of different systems on a common translation task, automatic metrics can be applied on a frequent and ongoing basis during system development, in order to guide the development of the system based on concrete performance improvements.

In this paper, we present a comparison between the widely used BLEU and NIST metrics, and a set of easily computable metrics based on unigram precision and recall. Using several empirical evaluation methods that have been proposed

in the recent literature as concrete means to assess the level of correlation of automatic metrics and human judgments, we show that higher correlations can be obtained with fairly simple and straightforward metrics. While recent researchers [7] [8] have shown that a balanced combination of precision and recall (F1 measure) has improved correlation with human judgments compared to BLEU and NIST, we claim that even better correlations can be obtained by assigning more weight to recall than to precision. In fact, our experiments show that the best correlations are achieved when recall is assigned almost all the weight. Previous work by Lin and Hovy [9] has shown that a recall-based automatic metric for evaluating summaries outperforms the BLEU metric on that task. Our results show that this is also the case for evaluation of MT. We also demonstrate that stemming both MT-output and reference strings prior to their comparison, which allows different morphological variants of a word to be considered as "matches", significantly further improves the performance of the metrics.

We describe the metrics used in our evaluation in Section 2. We also discuss certain characteristics of the BLEU and NIST metrics that may account for the advantage of metrics based on unigram recall. Our evaluation methodology and the data used for our experimentation are described in section 3. Our experiments and their results are described in section 4. Future directions and extensions of this work are discussed in section 5.

2 Evaluation Metrics

The metrics used in our evaluations, in addition to BLEU and NIST, are based on explicit word-to-word matches between the translation being evaluated and each of one or more reference translations. If more than a single reference translation is available, the translation is matched with each reference *independently*, and the best-scoring match is selected. While this does not allow us to simultaneously match different portions of the translation with different references, it supports the use of recall as a component in scoring each possible match. For each metric, including BLEU and NIST, we examine the case where matching requires that the matched word in the translation and reference be identical (the standard behavior of BLEU and NIST), and the case where stemming is applied to both strings prior to the matching[1]. In the second case, we stem both translation and references prior to matching and then require identity on stems. We plan to experiment in the future with less strict matching schemes that will consider matching synonymous words (with some cost), as described in section 5.

2.1 BLEU and NIST

The main principle behind IBM's BLEU metric [1] is the measurement of the overlap in unigrams (single words) and higher order n-grams of words, between a

[1] We include BLEU and NIST in our evaluations on stemmed data, but since neither one includes stemming as part of the metric, the resulting BLEU-stemmed and NIST-stemmed scores are not truly BLEU and NIST scores. They serve to illustrate the effectiveness of stemming in MT evaluation.

translation being evaluated and a set of one or more reference translations. The main component of BLEU is n-gram precision: the proportion of the matched n-grams out of the total number of n-grams in the evaluated translation. Precision is calculated separately for each n-gram order, and the precisions are combined via a geometric averaging. BLEU does not take recall into account directly. Recall – the proportion of the matched n-grams out of the total number of n-grams in the reference translation, is extremely important for assessing the quality of MT output, as it reflects to what degree the translation covers the entire content of the translated sentence. BLEU does not use recall because the notion of recall is unclear when simultaneously matching against multiple reference translations (rather than a single reference). To compensate for recall, BLEU uses a *Brevity Penalty*, which penalizes translations for being "too short". The NIST metric is conceptually similar to BLEU in most aspects, including the weaknesses discussed below:

- **The Lack of Recall:** We believe that the brevity penalty in BLEU does not adequately compensate for the lack of recall. Our experimental results strongly support this claim.
- **Lack of Explicit Word-matching Between Translation and Reference:** N-gram counts don't require an explicit word-to-word matching, but this can result in counting incorrect "matches", particularly for common function words. A more advanced metric that we are currently developing (see section 4.3) uses the explicit word-matching to assess the grammatical coherence of the translation.
- **Use of Geometric Averaging of N-grams:** Geometric averaging results in a score of "zero" whenever one of the component n-gram scores is zero. Consequently, BLEU scores at the sentence level can be meaningless. While BLEU was intended to be used only for aggregate counts over an entire test-set (and not at the sentence level), a metric that exhibits high levels of correlation with human judgments at the sentence level would be highly desirable. In experiments we conducted, a modified version of BLEU that uses equal-weight arithmetic averaging of n-gram scores was found to have better correlation with human judgments at both the sentence and system level.

2.2 Metrics Based on Unigram Precision and Recall

The following metrics were used in our evaluations:

1. **Unigram Precision:** As mentioned before, we consider only exact one-to-one matches between words. Precision is calculated as follows:

$$P = \frac{m}{w_t}$$

where m is the number of words in the translation that match words in the reference translation, and w_t is the number of words in the translation. This may be interpreted as *the fraction of the words in the translation that are present in the reference translation.*

2. **Unigram Precision with Stemming:** Same as above, but the translation and references are stemmed before precision is computed.
3. **Unigram Recall:** As with precision, only exact one-to-one word matches are considered. Recall is calculated as follows:
$$P = \frac{m}{w_r}$$
where m is the number of matching words, and w_r is the number of words in the reference translation. This may be interpreted as *the fraction of words in the reference that appear in the translation.*
4. **Unigram Recall with Stemming:** Same as above, but the translation and references are stemmed before recall is computed.
5. F_1**:** The harmonic mean [10] of precision and recall. F_1 is computed as follows:
$$F_1 = \frac{2PR}{P+R}$$
6. F_1 **with Stemming:** Same as above, but using the stemmed version of both precision and recall.
7. **Fmean:** This is similar to F_1, but recall is weighted nine times more heavily than precision. The precise amount by which recall outweighs precision is less important than the fact that most of the weight is placed on recall. The balance used here was estimated using a development set of translations and references (we also report results on a large test set that was not used in any way to determine any parameters in any of the metrics). Fmean is calculated as follows:
$$Fmean = \frac{10PR}{9P+R}$$

3 Evaluating MT Evaluation Metrics

3.1 Data

We evaluated the metrics described in section 2 and compared their performances with BLEU and NIST on two large data sets: the DARPA/TIDES 2002 and 2003 Chinese-to-English MT Evaluation sets. The data in both cases consists of approximately 900 sentences with four reference translations each. Both evaluations had corresponding human assessments, with two human judges evaluating each translated sentence. The human judges assign an Adequacy Score and a Fluency Score to each sentence. Each score ranges from one to five (with one being the poorest grade and five the highest). The adequacy and fluency scores of the two judges for each sentence are averaged together, and an overall average adequacy and average fluency score is calculated for each evaluated system. The total human score for each system is the sum of the average adequacy and average fluency scores, and can range from two to ten. The data from the 2002 evaluation contains system output and human evaluation scores for seven systems. The 2003 data includes system output and human evaluation scores for six systems. The 2002 set was used in determining the weights of precision and recall in the Fmean metric.

3.2 Evaluation Methodology

Our goal in the evaluation of the MT scoring metrics is to effectively quantify how well each metric correlates with human judgments of MT quality. Several different experimental methods have been proposed and used in recent work by various researchers. In our experiments reported here, we use two methods of assessment:

1. **Correlation of Automatic Metric Scores and Human Scores at the System-level:** We plot the automatic metric score assigned to each tested system against the average total human score assigned to the system, and calculate a correlation coefficient between the metric scores and the human scores. Melamed et al [7], [8] suggest using the Spearman rank correlation coefficient as an appropriate measure for this type of correlation experiment. The rank correlation coefficient abstracts away from the absolute scores and measures the extent to which the two scores (human and automatic) similarly rank the systems. We feel that this rank correlation is not a sufficiently sensitive evaluation criterion, since even poor automatic metrics are capable of correctly ranking systems that are very different in quality. We therefore opted to evaluate the correlation using the Pearson correlation coefficient, which takes into account the distances of the data points from an optimal regression curve. This method has been used by various other researchers [6] and also in the official DARPA/TIDES evaluations.
2. **Correlation of Score Differentials between Pairs of Systems:** For each pair of systems we calculate the differentials between the systems for both the human score and the metric score. We then plot these differentials and calculate a Pearson correlation coefficient between the differentials. This method was suggested by Coughlin [11]. It provides significantly more data points for establishing correlation between the MT metric and the human scores. It makes the reasonable assumption that the differentials of automatic metric and human scores should highly correlate. This assumption is reasonable if both human scores and metric scores are linear in nature, which is generally true for the metrics we compare here.

As mentioned before, the values presented in this paper are Pearson's correlation coefficients, and consequently they range from -1 to 1, with 1 representing a very strong association between the automatic score and the human score. Thus the different metrics are assessed primarily by looking at which metric has a higher correlation coefficient in each scenario.

In order to validate the statistical significance of the differences in the scores, we apply a commonly used bootstrapping sampling technique [12] to estimate the variability over the test set, and establish confidence intervals for each of the system scores and the correlation coefficients.

Table 1. Correlation coefficients with human judgments for each metric on the DARPA/TIDES 2002 Chinese data set

Metric	Pearson's Coefficient	Confidence Interval
NIST	0.603	+/- 0.049
NIST-stem	**0.740**	+/- 0.043
BLEU	0.461	+/- 0.058
BLEU-stem	0.528	+/- 0.061
P	0.175	+/- 0.052
P-stem	0.257	+/- 0.065
R	0.615	+/- 0.042
R-stem	**0.757**	+/- 0.042
F1	0.425	+/- 0.047
F1-stem	0.564	+/- 0.052
Fmean	0.585	+/- 0.043
Fmean-stem	**0.733**	+/- 0.044

4 Metric Evaluation

4.1 Correlation of Automatic Metric Scores and Human Scores at the System-level

We first compare the various metrics in terms of the correlation they have with total human scores at the system level. For each metric, we plot the metric and total human scores assigned to each system and calculate the correlation coefficient between the two scores. Tables 1 and 2 summarize the results for the various metrics on the 2002 and 2003 data sets. All metrics show much higher levels of correlation with human judgments on the 2003 data, compared with the 2002 data. The 2002 data exhibits several anomalies that have been identified and discussed by several other researchers [13]. Three of the 2002 systems have output that contains significantly higher amounts of "noise" (non ascii characters) and upper-cased words, which are detrimental to the automatic metrics. The variability within the 2002 set is also much higher than within the 2003 set, as reflected by the confidence intervals of the various metrics.

The levels of correlation of the different metrics are quite consistent across both 2002 and 2003 data sets. Unigram-recall and F-mean have significantly higher levels of correlation than BLEU and NIST. Unigram-precision, on the other hand, has a poor level of correlation. The performance of F1 is inferior to F-mean on the 2002 data. On the 2003 data, F1 is inferior to Fmean, but stemmed F1 is about equivalent to Fmean. Stemming improves correlations for all metrics on the 2002 data. On the 2003 data, stemming improves correlation on all metrics except for recall and Fmean, where the correlation coefficients are already so high that stemming no longer has a statistically significant effect. Recall, Fmean and NIST also exhibit more stability than the other metrics, as reflected by the confidence intervals.

Table 2. Correlation coefficients with human judgments for each metric on the DARPA/TIDES 2003 Chinese data set

Metric	Pearson's Coefficient	Confidence Interval
NIST	0.892	+/- 0.013
NIST-stem	0.915	+/- 0.010
BLEU	0.817	+/- 0.021
BLEU-stem	0.843	+/- 0.018
P	0.683	+/- 0.041
P-stem	0.752	+/- 0.041
R	**0.961**	+/- 0.011
R-stem	**0.940**	+/- 0.014
F1	0.909	+/- 0.025
F1-stem	**0.948**	+/- 0.014
Fmean	**0.959**	+/- 0.012
Fmean-stem	**0.952**	+/- 0.013

4.2 Correlation of Score Differentials between Pairs of Systems

We next calculated the score differentials for each pair of systems that were evaluated and assessed the correlation between the automatic score differentials and the human score differentials. The results of this evaluation are summarized in Tables 3 and 4. The results of the system pair differential correlation experiments are very consistent with the system-level correlation results. Once again, Unigram-recall and F-mean have significantly higher levels of correlation than BLEU and NIST. The effects of stemming are somewhat less pronounced in this evaluation.

4.3 Discussion

It is clear from these results that unigram-recall has a very strong correlation with human assessment of MT quality, and stemming often strengthens this correlation. This follows the intuitive notion that MT system output should contain as much of the system output should contain as much of the meaning of the input as possible. It is perhaps surprising that unigram-precision, on the other hand, has such low correlation. It is still important, however, to factor precision into the final score assigned to a system, to prevent systems that output very long translations from receiving inflated scores (as an extreme example, a system that outputs every word in its vocabulary for every translation would consistently score very high in unigram recall, regardless of the quality of the translation). Our Fmean metric is effective in combining precision and recall. Because recall is weighted heavily, the Fmean scores have high correlations. For both data sets tested, recall and Fmean performed equally well (differences were statistically insignificant), even though precision performs much worse. Because we use a weighted harmonic mean, where precision and recall are multiplied, low

Table 3. Correlation coefficients for pairwise system comparisons on the DARPA/TIDES 2002 Chinese data set

Metric	Pearson's Coefficient	Confidence Interval
NIST	0.679	+/- 0.042
NIST-stem	0.774	+/- 0.041
BLEU	0.498	+/- 0.054
BLEU-stem	0.559	+/- 0.058
P	0.298	+/- 0.051
P-stem	0.325	+/- 0.064
R	0.743	+/- 0.032
R-stem	**0.845**	+/- 0.029
F1	0.549	+/- 0.042
F1-stem	0.643	+/- 0.046
Fmean	0.711	+/- 0.033
Fmean-stem	**0.818**	+/- 0.032

levels of precision properly penalize the Fmean score (thus disallowing the case of a system scoring high simply by outputting many words).

One feature of BLEU and NIST that is not included in simple unigram-based metrics is the approximate notion of word order or grammatical coherence achieved by the use of higher-level n-grams. We have begun development of a new metric that combines the Fmean score with an explicit measure of grammatical coherence. This metric, METEOR (Metric for Evaluation of Translation with Explicit word Ordering), performs a maximal-cardinality match between translations and references, and uses the match to compute a coherence-based penalty. This computation is done by assessing the extent to which the matched words between translation and reference constitute well ordered coherent "chunks". Preliminary experiments with METEOR have yielded promising results, achieving similar levels of correlation (but so far not statistically significantly superior) as compared to the simpler measures of Fmean and recall.

5 Current and Future Work

We are currently in the process of enhancing the METEOR metric in several directions:

Expanding the Matching between Translation and References: Our experiments indicate that stemming already significantly improves the quality of the metric by expanding the matching. We plan to experiment with further expanding the matching to include synonymous words, by using information from synsets in WordNet. Since the reliability of such matches is likely to be somewhat reduced, we will consider assigning such matches a lower confidence that will be taken into account within score computations.

Table 4. Correlation coefficients for pairwise system comparisons on the DARPA/TIDES 2003 Chinese data set

Metric	Pearson's Coefficient	Confidence Interval
NIST	0.886	+/- 0.017
NIST-stem	0.924	+/- 0.013
BLEU	0.758	+/- 0.027
BLEU-stem	0.793	+/- 0.025
P	0.573	+/- 0.053
P-stem	0.666	+/- 0.058
R	**0.954**	+/- 0.014
R-stem	0.923	+/- 0.018
F1	0.881	+/- 0.024
F1-stem	**0.950**	+/- 0.017
Fmean	**0.954**	+/- 0.015
Fmean-stem	0.940	+/- 0.017

Combining Precision, Recall and Sort Penalty: Results so far indicate that recall plays the most important role in obtaining high-levels of correlation with human judgments. We are currently exploring alternative ways for combining the components of precision, recall and a coherence penalty with the goal of optimizing correlation with human judgments, and exploring whether an optimized combination of these factors on one data set is also persistent in performance across different data sets.

The Utility of Multiple Reference Translations: The metrics described use multiple reference translations in a weak way: we compare the translation with each reference separately and select the reference with the best match. This was necessary in order to incorporate recall in our metric, which we have shown to be highly advantageous. We are in the process of quantifying the utility of multiple reference translations across the metrics by measuring the correlation improvements as a function of the number of reference translations. We will then consider exploring ways in which to improve our matching against multiple references. Recent work by Pang, Knight and Marcu [14] provides the mechanism for producing semantically meaningful additional "synthetic" references from a small set of real references. We plan to explore whether using such synthetic references can improve the performance of our metric.

Matched Words are not Created Equally: Our current metrics treats all matched words between a system translation and a reference equally. It is safe to assume, however, that matching semantically important words should carry significantly more weight than the matching of function words. We plan to explore schemes for assigning different weights to matched words, and investigate if such schemes can further improve the sensitivity of the metric and its correlation with human judgments of MT quality.

Acknowledgments

This research was funded in part by NSF grant number IIS-0121631.

References

1. Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, PA, July.
2. Doddington, George. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In *Proceedings of the Second Conference on Human Language Technology (HLT-2002)*. San Diego, CA. pp. 128–132.
3. K.-Y. Su, M.-W. Wu, and J.-S. Chang. 1992. A New Quantitative Quality Measure for Machine Translation Systems. In *Proceedings of the fifteenth International Conference on Computational Linguistics (COLING-92)*. Nantes, France. pp. 433–439.
4. Y. Akiba, K. Imamura, and E. Sumita. 2001. Using Multiple Edit Distances to Automatically Rank Machine Translation Output. In *Proceedings of MT Summit VIII*. Santiago de Compostela, Spain. pp. 15–20.
5. S. Niessen, F. J. Och, G. Leusch, and H. Ney. 2000. An Evaluation Tool for Machine Translation: Fast Evaluation for Machine Translation Research. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000)*. Athens, Greece. pp. 39–45.
6. Gregor Leusch, Nicola Ueffing and Herman Ney. 2003. String-to-String Distance Measure with Applications to Machine Translation Evaluation. In *Proceedings of MT Summit IX*. New Orleans, LA. Sept. 2003. pp. 240–247.
7. I. Dan Melamed, R. Green and J. Turian. 2003. Precision and Recall of Machine Translation. In *Proceedings of HLT-NAACL 2003*. Edmonton, Canada. May 2003. Short Papers: pp. 61–63.
8. Joseph P. Turian, Luke Shen and I. Dan Melamed. 2003. Evaluation of Machine Translation and its Evaluation. In *Proceedings of MT Summit IX*. New Orleans, LA. Sept. 2003. pp. 386–393.
9. Chin-Yew Lin and Eduard Hovy. 2003. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In *Proceedings of HLT-NAACL 2003*. Edmonton, Canada. May 2003. pp. 71–78.
10. C. van Rijsbergen. 1979. Information Retrieval. Butterworths. London, England. 2nd Edition.
11. Deborah Coughlin. 2003. Correlating Automated and Human Assessments of Machine Translation Quality. In *Proceedings of MT Summit IX*. New Orleans, LA. Sept. 2003. pp. 63–70.
12. Bradley Efron and Robert Tibshirani. 1986. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science, 1(1)*. pp. 54–77.
13. George Doddington. 2003. Automatic Evaluation of Language Translation using N-gram Co-occurrence Statistics. Presentation at DARPA/TIDES 2003 MT Workshop. NIST, Gathersberg, MD. July 2003.
14. Bo Pang, Kevin Knight and Daniel Marcu. 2003. Syntax-based Alignment of Multiple Translations: Extracting Paraphrases and Generating New Sentences. In *Proceedings of HLT-NAACL 2003*. Edmonton, Canada. May 2003. pp. 102–109.