

METEOR, M-BLEU and M-TER: Flexible Matching and Parameter Tuning for High-Correlation with Human Judgments of Machine Translation Quality

Abhaya Agarwal and Alon Lavie

Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, 15213, USA
{abhayaa,alavie}@cs.cmu.edu

Abstract

We describe our submission to the NIST Metrics for Machine Translation Challenge consisting of 4 metrics - two versions of METEOR, M-BLEU and M-TER. We first give a brief description of METEOR. That is followed by description of M-BLEU and M-TER, enhanced versions of two other widely used metrics BLEU and TER, which extend the exact word matching used in these metrics with the flexible matching based on stemming and Wordnet in METEOR.

1 Introduction

METEOR, initially proposed and released in 2004 (Lavie et al., 2004) was explicitly designed to improve correlation with human judgments of MT quality at the segment level. Previous publications on METEOR (Lavie et al., 2004; Banerjee and Lavie, 2005; Lavie and Agarwal, 2007) have described the details underlying the metric and have extensively compared its performance with BLEU and several other MT evaluation metrics. In (Lavie and Agarwal, 2007), we described the process of tuning free parameters within the metric to optimize the correlation with human judgments and the extension of the metric for evaluating translations in languages other than English.

This paper provides a brief technical description of METEOR and the process of re-tuning the metric for improving correlation with the human judgments of translation quality. Also, in order to establish the usefulness of the flexible matching based on stemming and Wordnet, we extend two other widely used metrics, BLEU and TER, which use exact word matching, with the matcher module of METEOR.

2 The METEOR Metric

METEOR evaluates a translation by computing a score based on explicit word-to-word matches be-

tween the translation and a given reference translation. If more than one reference translation is available, the translation is scored against each reference independently, and the best scoring pair is used. Given a pair of strings to be compared, METEOR creates a *word alignment* between the two strings. An alignment is mapping between words, such that every word in each string maps to at most *one* word in the other string. This alignment is incrementally produced by a sequence of word-mapping modules. The “exact” module maps two words if they are exactly the same. The “porter stem” module maps two words if they are the same after they are stemmed using the Porter stemmer. The “WN synonymy” module maps two words if they are considered synonyms, based on the fact that they both belong to the same “synset” in WordNet.

The word-mapping modules initially identify all possible word matches between the pair of strings. We then identify the largest subset of these word mappings such that the resulting set constitutes an alignment as defined above. If more than one maximal cardinality alignment is found, METEOR selects the alignment for which the word order in the two strings is most similar (the mapping that has the least number of “crossing” unigram mappings). The order in which the modules are run reflects word-matching preferences. The default ordering is to first apply the “exact” mapping module, followed by “porter stemming” and then “WN synonymy”.

Once a final alignment has been produced between the system translation and the reference translation, the METEOR score for this pairing is computed as follows. Based on the number of mapped unigrams found between the two strings (m), the total number of unigrams in the translation (t) and the total number of unigrams in the reference (r), we calculate unigram precision $P = m/t$ and unigram recall $R = m/r$. We then compute a parametrized har-

monic mean of P and R (van Rijsbergen, 1979):

$$F_{mean} = \frac{P \cdot R}{\alpha \cdot P + (1 - \alpha) \cdot R}$$

Precision, recall and Fmean are based on single-word matches. To take into account the extent to which the matched unigrams in the two strings are in the same word order, METEOR computes a penalty for a given alignment as follows. First, the sequence of matched unigrams between the two strings is divided into the fewest possible number of “chunks” such that the matched unigrams in each chunk are adjacent (in both strings) and in identical word order. The number of chunks (ch) and the number of matches (m) is then used to calculate a fragmentation fraction: $frag = ch/m$. The penalty is then computed as:

$$Pen = \gamma \cdot frag^\beta$$

The value of γ determines the maximum penalty ($0 \leq \gamma \leq 1$). The value of β determines the functional relation between fragmentation and the penalty. Finally, the METEOR score for the alignment between the two strings is calculated as:

$$score = (1 - Pen) \cdot F_{mean}$$

2.1 Tuning the Free Parameters

The free parameters in the metric, α , β and γ are tuned to achieve maximum correlation with the human judgments as described in (Lavie and Agarwal, 2007).

The first step is to define a suitable measure of correlation between the particular kind of human judgment in hand and the metric assigned scores. For adequacy and fluency scores on a discreet scale, Pearson’s and Spearman’s coefficient of correlation are two popular choices. For ranking judgments, we use average of Spearman correlation across all the sentences as described in (Agarwal and Lavie, 2008).

The parametrs are tuned by doing a exhaustive grid search in the feasible ranges of parameter values, looking for parameters that maximize the correlation over the training data.

We used the MT06 dataset provided as development data for tuning the parameters of Meteor. The two versions of Meteor were submitted - one for optimizing Pearson’s coefficient of correlation with adequacy judgments and another for maximizing the correct number of binary preference judgments.

The re-tuned parameter values are shown in Table 1.

	Adequacy	Ranking
α	0.8	0.95
β	2.5	.5
γ	0.4	0.5

Table 1: Optimal Values of Tuned Parameters for different criteria

2.2 Availability

Current and earlier versions of METEOR are available as free download from <http://www.cs.cmu.edu/~alavie/METEOR/>.

3 Extending BLEU and TER with Flexible Matching

Many widely used metrics like BLEU (Papineni et al., 2002) and TER (Snover et al., 2006) are based on measuring string level similarity between the reference translation and translation hypothesis, just like METEOR . Most of them, however, depend on finding exact matches between the words in two strings. Many researchers (Banerjee and Lavie, 2005; Liu and Gildea, 2006), have observed consistent gains by using more flexible matching criteria. In the following experiments, we extend the BLEU and TER metrics to use the stemming and Wordnet based word mapping modules from METEOR .

Given a translation hypothesis and reference pair, we first align them using the word mapping modules from METEOR . We then rewrite the reference translation by replacing the matched words with the corresponding words in the translation hypothesis. We now compute BLEU and TER with these new references without changing anything inside the metrics.

In our earlier experiments, M-BLEU and M-TER have shown some improvements over the original metrics in various languages (Agarwal and Lavie, 2008).

4 Availability

A standalone matcher to generate the re-written references is included in the METEOR package. For full M-BLEU and M-TER code, please contact the authors.

5 Conclusions

In this paper, we described METEOR and the process of re-tuning the parameters to better correlate with human rankings of translation hypotheses. We also presented enhanced BLEU and TER that use the flexible word matching module from Meteor. The new version of METEOR is available on our website at: <http://www.cs.cmu.edu/~alavie/METEOR/> . This

release also includes the flexible word matcher module which can be used to extend any metric with the flexible matching.

Acknowledgments

The work reported in this paper was supported by NSF Grant IIS-0534932.

References

- Abhaya Agarwal and Alon Lavie. 2008. METEOR, M-BLEU and M-TER: Evaluation Metric for High Correlation with Human Rankings of Machine Translation Output. In *Proceedings of the Third ACL Workshop on Statistical Machine Translation*, Columbus, USA, June.
- Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June.
- Alon Lavie and Abhaya Agarwal. 2007. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Proceedings of the Second ACL Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic, June.
- Alon Lavie, Kenji Sagae, and Shyamsundar Jayaraman. 2004. The Significance of Recall in Automatic Metrics for MT Evaluation. In *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (AMTA-2004)*, pages 134–143, Washington, DC, September.
- Ding Liu and Daniel Gildea. 2006. Stochastic iterative alignment for machine translation evaluation. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 539–546, Morristown, NJ, USA. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Weijing 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.
- Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA-2006)*, pages 223–231, Cambridge, MA, August.
- C. van Rijsbergen, 1979. *Information Retrieval*. Butterworths, London, UK, 2nd edition.