METEOR, M-BLEU and M-TER: Evaluation Metrics for High-Correlation with Human Rankings of Machine Translation Output

Abhaya Agarwal and **Alon Lavie**
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, 15213, USA
{abhayaa,alavie}@cs.cmu.edu

Abstract

This paper describes our submissions to the machine translation evaluation shared task in ACL WMT-08. Our primary submission is the METEOR metric tuned for optimizing correlation with human rankings of translation hypotheses. We show significant improvement in correlation as compared to the earlier version of metric which was tuned to optimized correlation with traditional adequacy and fluency judgments. We also describe M-BLEU and M-TER, enhanced versions of two other widely used metrics BLEU and TER respectively, which extend the exact word matching used in these metrics with the flexible matching based on stemming and Wordnet in METEOR .

1 Introduction

Automatic Metrics for MT evaluation have been receiving significant attention in recent years. Evaluating an MT system using such automatic metrics is much faster, easier and cheaper compared to human evaluations, which require trained bilingual evaluators. The most commonly used MT evaluation metric in recent years has been IBM's BLEU metric (Papineni et al., 2002). BLEU is fast and easy to run, and it can be used as a target function in parameter optimization training procedures that are commonly used in state-of-the-art statistical MT systems (Och, 2003). Various researchers have noted, however, various weaknesses in the metric. Most notably, BLEU does not produce very reliable sentence-level scores. METEOR , as well as several other proposed metrics such as GTM (Melamed et al., 2003), TER (Snover et al., 2006) and CDER (Leusch et al., 2006) aim to address some of these weaknesses.

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 describes our experiments in re-tuning the metric for improving correlation with the human rankings of translation hypotheses corresponding to a single source sentence. Our experiments show significant improvement in correlation as a result of re-tuning which shows the importance of having a metric tunable to different testing conditions. 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 between 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 us-

Proceedings of the Third Workshop on Statistical Machine Translation, pages 115–118,
Columbus, June 2008.

ing 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 harmonic 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:

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

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).

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.

To get meaningful BLEU scores at segment level, we compute smoothed BLEU as described in (Lin and Och, 2004).

4 Re-tuning METEOR for Rankings

(Callison-Burch et al., 2007) reported that the intercoder agreement on the task of assigning ranks to a given set of candidate hypotheses is much better than the intercoder agreement on the task of assigning a score to a hypothesis in isolation. Based on that finding, in WMT-08, only ranking judgments are being collected from the human judges.

The current version of METEOR uses parameters optimized towards maximizing the Pearson's correlation with human judgments of adequacy scores. It is not clear that the same parameters would be optimal for correlation with human rankings. So we would like to re-tune the parameters in the metric for maximizing the correlation with ranking judgments instead. This requires computing full rankings according to the metric and the humans and then computing a suitable correlation measure on those rankings.

4.1 Computing Full Rankings

METEOR assigns a score between 0 and 1 to every translation hypothesis. This score can be converted

Language	Judgments	
	Binary	Sentences
English	3978	365
German	2971	334
French	1903	208
Spanish	2588	284

Table 1: Corpus Statistics for Various Languages

to rankings trivially by assuming that a higher score indicates a better hypothesis.

In development data, human rankings are available as binary judgments indicating the preferred hypothesis between a given pair. There are also cases where both the hypotheses in the pair are judged to be equal. In order to convert these binary judgments into full rankings, we do the following:

1. Throw out all the equal judgments.
2. Construct a directed graph where nodes correspond to the translation hypotheses and every binary judgment is represented by a directed edge between the corresponding nodes.
3. Do a topological sort on the resulting graph and assign ranks in the sort order. The cycles in the graph are broken by assigning same rank to all the nodes in the cycle.

4.2 Measuring Correlation

Following (Ye et al., 2007), we first compute the Spearman correlation between the human rankings and METEOR rankings of the translation hypotheses corresponding to a single source sentence. Let N be the number of translation hypotheses and D be the difference in ranks assigned to a hypothesis by two rankings, then Spearman correlation is given by:

$$r = 1 - \frac{6\sum D^2}{N(N^2-1)}$$

The final score for the metric is the average of the Spearman correlations for individual sentences.

5 Experiments

5.1 Data

We use the human judgment data from WMT-07 which was released as development data for the evaluation shared task. Amount of data available for various languages is shown in Table 1. Development data contains the majority judgments (not every hypotheses pair was judged by same number of judges) which means that in the cases where multiple judges judged the same pair of hypotheses, the judgment given by majority of the judges was considered.

	English	German	French	Spanish
α	0.95	0.9	0.9	0.9
β	0.5	3	0.5	0.5
γ	0.45	0.15	0.55	0.55

Table 2: Optimal Values of Tuned Parameters for Various Languages

	Original	Re-tuned
English	0.3813	0.4020
German	0.2166	0.2838
French	0.2992	0.3640
Spanish	0.2021	0.2186

Table 3: Average Spearman Correlation with Human Rankings for METEOR on Development Data

5.2 Methodology

We do an exhaustive grid search in the feasible ranges of parameter values, looking for parameters that maximize the average Spearman correlation over the training data. To get a fair estimate of performance, we use 3-fold cross validation on the development data. Final parameter values are chosen as the best performing set on the data pooled from all the folds.

5.3 Results

5.3.1 Re-tuning METEOR for Rankings

The re-tuned parameter values are shown in Table 2 while the average Spearman correlations for various languages with original and re-tuned parameters are shown in Table 3. We get significant improvements for all the languages. Gains are specially pronounced for German and French.

Interestingly, weight for recall becomes even higher than earlier parameters where it was already high. So it seems that ranking judgments are almost entirely driven by the recall in all the languages. Also the re-tuned parameters for all the languages except German are quite similar.

5.3.2 M-BLEU and M-TER

Table 4 shows the average Spearman correlations of M-BLEU and M-TER with human rankings. For English, both M-BLEU and M-TER show considerable improvements. For other languages, improvements in M-TER are smaller but consistent. M-BLEU , however, doesn't shows any improvements in this case. A possible reason for this behavior can be the lack of a "WN synonymy" module for languages other than English which results in fewer extra matches over the exact matching baseline. Additionally, French, German and Spanish have a richer morphology as compared to English. The morphemes in these languages

	Exact Match	Flexible Match
English: Bleu	0.2486	**0.2747**
Ter	0.1598	**0.2033**
French: Bleu	**0.2906**	0.2889
Ter	0.2472	**0.2604**
German: Bleu	**0.1829**	0.1806
Ter	0.1509	**0.1668**
Spanish: Bleu	0.1804	**0.1847**
Ter	0.1787	**0.1839**

Table 4: Average Spearman Correlation with Human Rankings for M-BLEU and M-TER

carry much more information and different forms of the same word may not be as freely replaceable as in English. A more fine grained strategy for matching words in these languages remains an area of further investigation.

6 Conclusions

In this paper, we described the re-tuning of Meteor parameters to better correlate with human rankings of translation hypotheses. Results on the development data indicate that the re-tuned version is significantly better at predicting ranking than the earlier version. We also presented enhanced Bleu and Ter that use the flexible word matching module from Meteor and show that this results in better correlations as compared to the default exact matching versions. The new version of Meteor will be soon available on our website at: `http://www.cs.cmu.edu/~alavie/METEOR/` . This release will also include 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

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.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (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.

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.

Gregor Leusch, Nicola Ueffing, and Hermann Ney. 2006. CDER: Efficient MT Evaluation Using Block Movements. In *Proceedings of the Thirteenth Conference of the European Chapter of the Association for Computational Linguistics.*

Chin-Yew Lin and Franz Josef Och. 2004. Orange: a method for evaluating automatic evaluation metrics for machine translation. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, page 501, Morristown, NJ, USA. Association for Computational Linguistics.

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.

I. Dan Melamed, Ryan Green, and Joseph Turian. 2003. Precision and Recall of Machine Translation. In *Proceedings of the HLT-NAACL 2003 Conference: Short Papers*, pages 61–63, Edmonton, Alberta.

Franz Josef Och. 2003. Minimum Error Rate Training for Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics.*

Kishore Papineni, 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.

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.

Yang Ye, Ming Zhou, and Chin-Yew Lin. 2007. Sentence level machine translation evaluation as a ranking. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 240–247, Prague, Czech Republic, June. Association for Computational Linguistics.