Extending the METEOR Machine Translation Evaluation Metric to the Phrase Level

Michael Denkowski and **Alon Lavie**
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15232, USA
{mdenkows,alavie}@cs.cmu.edu

Abstract

This paper presents METEOR-NEXT, an extended version of the METEOR metric designed to have high correlation with post-editing measures of machine translation quality. We describe changes made to the metric's sentence aligner and scoring scheme as well as a method for tuning the metric's parameters to optimize correlation with human-targeted Translation Edit Rate (HTER). We then show that METEOR-NEXT improves correlation with HTER over baseline metrics, including earlier versions of METEOR, and approaches the correlation level of a state-of-the-art metric, TER-plus (TERp).

1 Introduction

Recent focus on the need for accurate automatic metrics for evaluating the quality of machine translation output has spurred much development in the field of MT. Workshops such as WMT09 (Callison-Burch et al., 2009) and the MetricsMATR08 challenge (Przybocki et al., 2008) encourage the development of new MT metrics and reliable human judgment tasks.

This paper describes our work extending the METEOR metric to improve correlation with human-targeted Translation Edit Rate (HTER) (Snover et al., 2006), a semi-automatic post-editing based metric which measures the distance between MT output and a targeted reference. We identify several limitations of the original METEOR metric and describe our modifications to improve performance on this task. Our extended metric, METEOR-NEXT, is then tuned to maximize segment-level correlation with HTER scores and tested against several baseline metrics. We show that METEOR-NEXT outperforms earlier versions of METEOR when tuned to the same HTER data and approaches the performance of a state-of-the-art TER-based metric, TER-plus.

2 The METEOR-NEXT Metric

2.1 Traditional METEOR Scoring

Given a machine translation hypothesis and a reference translation, the traditional METEOR metric calculates a lexical similarity score based on a word-to-word alignment between the two strings (Banerjee and Lavie, 2005). When multiple references are available, the hypothesis is scored against each and the reference producing the highest score is used. Alignments are built incrementally in a series of stages using the following METEOR matchers:

Exact: Words are matched if and only if their surface forms are identical.

Stem: Words are stemmed using a language-appropriate Snowball Stemmer (Porter, 2001) and matched if the stems are identical.

Synonym: Words are matched if they are both members of a synonym set according to the WordNet (Miller and Fellbaum, 2007) database. This matcher is limited to translations into English.

At each stage, one of the above matchers identifies all possible word matches between the two translations using words not aligned in previous stages. An alignment is then identified as the largest subset of these matches in which every word in each sentence aligns to zero or one words in the other sen-

Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 250–253,
Los Angeles, California, June 2010.

tence. If multiple such alignments exist, the alignment is chosen that best preserves word order by having the fewest crossing alignment links. At the end of each stage, matched words are fixed so that they are not considered in future stages. The final alignment is defined as the union of all stage alignments.

Once an alignment has been constructed, the total number of unigram matches (m), the number of words in the hypothesis (t), and the number of words in the reference (r) are used to calculate precision ($P = m/t$) and recall ($R = m/r$). The parameterized harmonic mean of P and R (van Rijsbergen, 1979) is then calculated:

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

To account for differences in word order, the minimum number of "chunks" (ch) is calculated where a chunk is defined as a series of matched unigrams that is contiguous and identically ordered in both sentences. The fragmentation ($frag = ch/m$) is then used to calculate a fragmentation penalty:

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

The final METEOR score is then calculated:

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

The free parameters α, β, and γ can be tuned to maximize correlation with various types of human judgments (Lavie and Agarwal, 2007).

2.2 Extending the METEOR Aligner

Traditional METEOR is limited to unigram matches, making it strictly a word-level metric. By focusing on only one match type per stage, the aligner misses a significant part of the possible alignment space. Further, selecting partial alignments based only on the fewest number of per-stage crossing alignment links can in practice lead to missing full alignments with the same number of matches in fewer chunks. Our extended aligner addresses these limitations by introducing support for multiple-word phrase matches and considering all possible matches in a single alignment stage.

We introduce an additional **paraphrase** matcher which matches *phrases* (one or more successive words) if one phrase is considered a paraphrase of the other by a paraphrase database. For English, we use the paraphrase database developed by Snover et al. (2009), using techniques presented by Bannard and Callison-Burch (2005).

The extended aligner first constructs a search space by applying all matchers in sequence to identify all possible matches between the hypothesis and reference. To reduce redundant matches, stem and synonym matches between pairs of words which have already been identified as exact matches are not considered. Matches have start positions and *lengths* in both sentences; a word occurring less than *length* positions after a match start is said to be *covered* by the match. As exact, stem, and synonym matches will always have length one in both sentences, they can be considered phrase matches of length one. Since other matches can cover phrases of different lengths in the two sentences, matches are now said to be one-to-one at the *phrase* level rather than the *word* level.

Once all possible matches have been identified, the aligner identifies the final alignment as the largest subset of these matches meeting the following criteria in order of importance:

1. Each word in each sentence is covered by zero or one matches
2. Largest number of covered words across both sentences
3. Smallest number of chunks, where a chunk is now defined as a series of matched phrases that is contiguous and identically ordered in both sentences
4. Smallest sum of absolute distances between match start positions in the two sentences (prefer to align words and phrases that occur at similar positions in both sentences)

The resulting alignment is selected from the full space of possible alignments and directly optimizes the statistics on which the the final score will be calculated.

2.3 Extended METEOR Scoring

Once an alignment has been chosen, the METEOR-NEXT score is calculated using extended versions of

the traditional METEOR statistics. We also introduce a tunable weight vector used to dictate the relative contribution of each match type. The extended METEOR score is calculated as follows.

The number of words in the hypothesis (t) and reference (r) are counted. For each of the matchers (m_i), count the number of words covered by matches of this type in the hypothesis ($m_i(t)$) and reference ($m_i(r)$) and apply the appropriate module weight (w_i). The weighted Precision and Recall are then calculated:

$$P = \frac{\sum_i w_i \cdot m_i(t)}{t} \qquad R = \frac{\sum_i w_i \cdot m_i(r)}{r}$$

The minimum number of chunks (ch) is then calculated using the new chunk definition. Once P, R, and ch are calculated, the remaining statistics and final score can be calculated as in Section 2.1.

3 Tuning for Post-Editing Measures of Quality

Human-targeted Translation Edit Rate (HTER) (Snover et al., 2006), is a semi-automatic assessment of machine translation quality based on the number of edits required to correct translation hypotheses. A human annotator edits each MT hypothesis so that it is meaning-equivalent with a reference translation, with an emphasis on making the minimum possible number of edits. The Translation Edit Rate (TER) is then calculated using the human-edited translation as a targeted reference for the MT hypothesis. The resulting scores are shown to correlate well with other types of human judgments (Snover et al., 2006).

3.1 Tuning Toward HTER

The GALE (Olive, 2005) Phase 2 unsequestered data includes HTER scores for multiple Arabic-to-English and Chinese-to-English MT systems. We used HTER scores for 10838 segments from 1045 documents from this data set to tune both the original METEOR and METEOR-NEXT. Both were exhaustively tuned to maximize the length-weighted segment-level Pearson's correlation with the HTER scores. This produced globally optimal α, β, and γ values for METEOR and optimal α, β, γ values plus stem, synonym, and paraphrase match weights for METEOR-NEXT (with the weight of exact matches fixed at 1). Table 1 compares the new HTER parameters to those tuned for other tasks including adequacy and fluency (Lavie and Agarwal, 2007) and ranking (Agarwal and Lavie, 2008).

Task	α	β	γ
Adequacy & Fluency	0.81	0.83	0.28
Ranking	0.95	0.50	0.50
HTER	0.70	1.95	0.50
HTER (extended)	0.65	1.95	0.45
	Stem	Syn	Par
	0	0.4	0.9

Table 1: Parameter values for various METEOR tasks for translations into English.

As observed by Snover et al. (2009), HTER prefers metrics which are more balanced between precision and recall: this results in the lowest values of α for any task. Additionally, non-exact matches receive lower weights, with stem matches receiving zero weight. This reflects a weakness in HTER scoring where words with matching stems are treated as completely dissimilar, requiring full word substitutions (Snover et al., 2006).

4 Experiments

The GALE (Olive, 2005) Phase 3 unsequestered data includes HTER scores for Arabic-to-English MT output. We created a test set from HTER scores of 2245 segments from 195 documents in this data set. Our evaluation metric (METEOR-NEXT-hter) was tested against the following established metrics: BLEU (Papineni et al., 2002) with a maximum N-gram length of 4, TER (Snover et al., 2006), versions of METEOR based on release 0.7 tuned for adequacy and fluency (METEOR-0.7-af) (Lavie and Agarwal, 2007), ranking (METEOR-0.7-rank) (Agarwal and Lavie, 2008), and HTER (METEOR-0.7-hter). Also included is the HTER-tuned version of TER-plus (TERp-hter), a metric with state-of-the-art performance in recent evaluations (Snover et al., 2009). Length-weighted Pearson's and Spearman's correlation are shown for all metrics at both the segment (Table 2) and document level (Table 3). System level correlations are not shown as the Phase 3 data only contained the output of 2 systems.

Metric	Pearson's r	Spearman's ρ
BLEU-4	-0.496	-0.510
TER	0.539	0.510
METEOR-0.7-af	-0.573	-0.561
METEOR-0.7-rank	-0.561	-0.556
METEOR-0.7-hter	-0.574	-0.562
METEOR-NEXT-hter	-0.600	-0.581
TERp-hter	0.627	0.610

Table 2: Segment level correlation with HTER.

Metric	Pearson's r	Spearman's ρ
BLEU-4	-0.689	-0.686
TER	0.675	0.679
METEOR-0.7-af	-0.696	-0.699
METEOR-0.7-rank	-0.691	-0.693
METEOR-0.7-hter	-0.704	-0.705
METEOR-NEXT-hter	-0.719	-0.713
TERp-hter	0.738	0.747

Table 3: Document level correlation with HTER.

METEOR-NEXT-hter outperforms all baseline metrics at both the segment and document level. Bootstrap sampling indicates that the segment-level correlation improvements of 0.026 in Pearson's r and 0.019 in Spearman's ρ over METEOR-0.7-hter are statistically significant at the 95% level. TERp's correlation with HTER is still significantly higher across all categories. Our metric does run significantly faster than TERp, scoring approximately 120 segments per second to TERp's 3.8.

5 Conclusions

We have presented an extended METEOR metric which shows higher correlation with HTER than baseline metrics, including traditional METEOR tuned on the same data. Our extensions are not specific to HTER tasks; improved alignments and additional features should improve performance on any task having sufficient tuning data. Although our metric does not outperform TERp, it should be noted that HTER incorporates TER alignments, providing TER-based metrics a natural advantage. Our metric also scores segments relatively quickly, making it a viable choice for tuning MT systems.

Acknowledgements

This work was funded in part by NSF grants IIS-0534932 and IIS-0915327.

References

Abhaya Agarwal and Alon Lavie. 2008. Meteor, m-bleu and m-ter: Evaluation Metrics for High-Correlation with Human Rankings of Machine Translation Output. In *Proc. of WMT08*, pages 115–118.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proc. of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proc. of ACL05*, pages 597–604.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proc. of WMT09*, pages 1–28.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Proc. of WMT07*, pages 228–231.

George Miller and Christiane Fellbaum. 2007. WordNet. http://wordnet.princeton.edu/.

Joseph Olive. 2005. *Global Autonomous Language Exploitation (GALE)*. DARPA/IPTO Proposer Information Pamphlet.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proc. of ACL02*, pages 311–318.

Martin Porter. 2001. Snowball: A language for stemming algorithms. http://snowball.tartarus.org/texts/.

M. Przybocki, K. Peterson, and S Bronsart. 2008. Official results of the NIST 2008 "Metrics for MAchine TRanslation" Challenge (MetricsMATR08). http://nist.gov/speech/tests/metricsmatr/2008/results/.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proc. of AMTA-2006*, pages 223–231.

Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. Fluency, Adequacy, or HTER? Exploring Different Human Judgments with a Tunable MT Metric. In *Proc. of WMT09*, pages 259–268.

C. van Rijsbergen, 1979. *Information Retrieval*, chapter 7. 2nd edition.