Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems

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 describes Meteor 1.3, our submission to the 2011 EMNLP Workshop on Statistical Machine Translation automatic evaluation metric tasks. New metric features include improved text normalization, higher-precision paraphrase matching, and discrimination between content and function words. We include Ranking and Adequacy versions of the metric shown to have high correlation with human judgments of translation quality as well as a more balanced Tuning version shown to outperform BLEU in minimum error rate training for a phrase-based Urdu-English system.

1 Introduction

The Meteor[1] metric (Banerjee and Lavie, 2005; Denkowski and Lavie, 2010b) has been shown to have high correlation with human judgments in evaluations such as the 2010 ACL Workshop on Statistical Machine Translation and NIST Metrics MATR (Callison-Burch et al., 2010). However, previous versions of the metric are still limited by lack of punctuation handling, noise in paraphrase matching, and lack of discrimination between word types. We introduce new resources for all WMT languages including text normalizers, filtered paraphrase tables, and function word lists. We show that the addition of these resources to Meteor allows tuning versions of the metric that show higher correlation with human translation rankings and adequacy scores on unseen test data. The evaluation resources are modular, usable with any other evaluation metric or MT software.

We also conduct a MT system tuning experiment on Urdu-English data to compare the effectiveness of using multiple versions of Meteor in minimum error rate training. While versions tuned to various types of human judgments do not perform as well as the widely used BLEU metric (Papineni et al., 2002), a balanced Tuning version of Meteor consistently outperforms BLEU over multiple end-to-end tune-test runs on this data set.

The versions of Meteor corresponding to the translation evaluation task submissions, (Ranking and Adequacy), are described in Sections 3 through 5 while the submission to the tunable metrics task, (Tuning), is described in Section 6.

2 New Metric Resources

2.1 Meteor Normalizer

Whereas previous versions of Meteor simply strip punctuation characters prior to scoring, version 1.3 includes a new text normalizer intended specifically for translation evaluation. The normalizer first replicates the behavior of the tokenizer distributed with the Moses toolkit (Hoang et al., 2007), including handling of non-breaking prefixes. After tokenization, we add several rules for *normalization*, intended to reduce meaning-equivalent punctuation styles to common forms. The following two rules are particularly helpful:

- Remove dashes between hyphenated words. (Example: `far-off` → `far off`)

[1]The metric name has previously been stylized as "METEOR" or "METEOR". As of version 1.3, the official stylization is simply "Meteor".

- Remove full stops in acronyms/initials. (Example: `U.N.` → `UN`)

Consider the behavior of the Moses tokenizer and Meteor normalizers given a reference translation containing the phrase "`U.S.-based organization`":

Moses:	`U.S.-based organization`
Meteor ≤1.2:	`U S based organization`
Meteor 1.3:	`US based organization`

Of these, only the Meteor 1.3 normalization allows metrics to match all of the following stylizations:

```
U.S.-based organization
US-based organization
U.S. based organization
US based organization
```

While intended for Meteor evaluation, use of this normalizer is a suitable preprocessing step for other metrics to improve accuracy when reference sentences are stylistically different from hypotheses.

2.2 Filtered Paraphrase Tables

The original Meteor paraphrase tables (Denkowski and Lavie, 2010b) are constructed using the phrase table "pivoting" technique described by Bannard and Callison-Burch (2005). Many paraphrases suffer from word accumulation, the appending of unaligned words to one or both sides of a phrase rather than finding a true rewording from elsewhere in parallel data. To improve the precision of the paraphrase tables, we filter out all cases of word accumulation by removing paraphrases where one phrase is a substring of the other. Table 1 lists the number of phrase pairs found in each paraphrase table before and after filtering. In addition to improving accuracy, the reduction of phrase table sizes also reduces the load time and memory usage of the Meteor paraphrase matcher. The tables are a modular resource suitable for other MT or NLP software.

2.3 Function Word Lists

Commonly used metrics such as BLEU and earlier versions of Meteor make no distinction between content and function words. This can be problematic for ranking-based evaluations where two system outputs can differ by a single word, such as mistranslating either a main verb or a determiner. To improve Meteor's discriminative power in such cases, we introduce a function word list for each WMT language and a new δ parameter to adjust the relative weight given to content words (any word not on the list) versus function words (see Section 3). Function word lists are estimated according to relative frequency in large monolingual corpora. For each language, we pool freely available WMT 2011 data consisting of Europarl (Koehn, 2005), news (sentence-uniqued), and news commentary data. Any word with relative frequency of 10^{-3} or greater is added to the function word list. Table 2 lists corpus size and number of function words learned for each language. In addition to common words, punctuation symbols consistently rise to the tops of function word lists.

Language	Phrase Pairs	After Filtering
English	6.24M	5.27M
Czech	756K	684K
German	3.52M	3.00M
Spanish	6.35M	5.30M
French	3.38M	2.84M

Table 1: Sizes of paraphrase tables before and after filtering

Language	Corpus Size (sents)	FW Learned
English	836M	93
Czech	230M	68
French	374M	85
German	309M	92
Spanish	168M	66

Table 2: Monolingual corpus size (words) and number of function words learned for each language

3 Meteor Scoring

Meteor evaluates translation hypotheses by aligning them to reference translations and calculating sentence-level similarity scores. This section describes our extended version of the metric.

For a hypothesis-reference pair, the search space of possible alignments is constructed by identifying all possible matches between the two sentences according to the following matchers:

Exact: Match words if their surface forms are iden-

tical.

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

Synonym: Match words if they share membership in any synonym set according to the WordNet (Miller and Fellbaum, 2007) database.

Paraphrase: Match phrases if they are listed as paraphrases in the paraphrase tables described in Section 2.2.

All matches are generalized to phrase matches with a start position and phrase length in each sentence. Any word occurring less than *length* positions after a match start is considered covered by the match. The exact and paraphrase matchers support all five WMT languages while the stem matcher is limited to English, French, German, and Spanish and the synonym matcher is limited to English.

Once matches are identified, the final alignment is resolved as the largest subset of all matches meeting the following criteria in order of importance:

1. Require each word in each sentence to be covered by zero or one matches.
2. Maximize the number of covered words across both sentences.
3. Minimize the number of *chunks*, where a *chunk* is defined as a series of matches that is contiguous and identically ordered in both sentences.
4. Minimize the sum of absolute distances between match start positions in the two sentences. (Break ties by preferring to align words and phrases that occur at similar positions in both sentences.)

Given an alignment, the metric score is calculated as follows. Content and function words are identified in the hypothesis (h_c, h_f) and reference (r_c, r_f) according to the function word lists described in Section 2.3. For each of the matchers (m_i), count the number of content and function words covered by matches of this type in the hypothesis ($m_i(h_c)$, $m_i(h_f)$) and reference ($m_i(r_c)$, $m_i(r_f)$). Calculate weighted precision and recall using matcher weights ($w_i ... w_n$) and content-function word weight (δ):

$$P = \frac{\sum_i w_i \cdot (\delta \cdot m_i(h_c) + (1-\delta) \cdot m_i(h_f))}{\delta \cdot |h_c| + (1-\delta) \cdot |h_f|}$$

Target	WMT09	WMT10	Combined
English	20,357	24,915	45,272
Czech	11,242	9,613	20,855
French	2,967	5,904	7,062
German	6,563	10,892	17,455
Spanish	3,249	3,813	7,062

Table 3: Human ranking judgment data from 2009 and 2010 WMT evaluations

$$R = \frac{\sum_i w_i \cdot (\delta \cdot m_i(r_c) + (1-\delta) \cdot m_i(r_f))}{\delta \cdot |r_c| + (1-\delta) \cdot |r_f|}$$

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 gaps and differences in word order, a fragmentation penalty is calculated using the total number of matched words (m, average over hypothesis and reference) and number of chunks (ch):

$$Pen = \gamma \cdot \left(\frac{ch}{m}\right)^{\beta}$$

The Meteor score is then calculated:

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

The parameters α, β, γ, δ, and $w_i ... w_n$ are tuned to maximize correlation with human judgments.

4 Parameter Optimization

4.1 Development Data

The 2009 and 2010 WMT shared evaluation data sets are made available as development data for WMT 2011. Data sets include MT system outputs, reference translations, and human rankings of translation quality. Table 3 lists the number of judgments for each evaluation and combined totals.

4.2 Tuning Procedure

To evaluate a metric's performance on a data set, we count the number of pairwise translation rankings preserved when translations are re-ranked by metric score. We then compute Kendall's τ correlation coefficient as follows:

$$\tau = \frac{\text{concordant pairs} - \text{discordant pairs}}{\text{total pairs}}$$

	Tune τ (WMT09)		Test τ (WMT10)	
Lang	Met1.2	Met1.3	Met1.2	Met1.3
English	0.258	**0.276**	0.320	**0.343**
Czech	0.148	**0.162**	**0.220**	0.215
French	0.414	**0.437**	0.370	**0.384**
German	0.152	**0.180**	**0.170**	0.155
Spanish	0.216	**0.240**	0.310	**0.326**

Table 5: Meteor 1.2 and 1.3 correlation with ranking judgments on tune and test data

	Meteor-1.2 r		Meteor-1.3 r	
Tune / Test	MT08	MT09	MT08	MT09
MT08	0.620	0.625	0.650	0.636
MT09	0.612	0.630	0.642	0.648
Tune / Test	P2	P3	P2	P3
P2	-0.640	-0.596	-0.642	-0.594
P3	-0.638	-0.600	-0.625	-0.612

Table 6: Meteor 1.2 and 1.3 correlation with adequacy and H-TER scores on tune and test data

For each WMT language, we learn Meteor parameters that maximize τ over the combined 2009 and 2010 data sets using an exhaustive parametric sweep. The resulting parameters, listed in Table 4, are used in the default Ranking version of Meteor 1.3.

For each language, the δ parameter is above 0.5, indicating a preference for content words over function words. In addition, the fragmentation penalties are generally less severe across languages. The additional features in Meteor 1.3 allow for more balanced parameters that distribute responsibility for penalizing various types of erroneous translations.

5 Evaluation Experiments

To compare Meteor 1.3 against previous versions of the metric on the task of evaluating MT system outputs, we tune a version for each language on 2009 WMT data and evaluate on 2010 data. This replicates the 2010 WMT shared evaluation task, allowing comparison to Meteor 1.2. Table 5 lists correlation of each metric version with ranking judgments on tune and test data. Meteor 1.3 shows significantly higher correlation on both tune and test data for English, French, and Spanish while Czech and German demonstrate overfitting with higher correlation on tune data but lower on test data. This overfitting effect is likely due to the limited number of systems providing translations into these languages and the difficulty of these target languages leading to significantly noisier translations skewing the space of metric scores. We believe that tuning to combined 2009 and 2010 data will counter these issues for the official Ranking version.

5.1 Generalization to Other Tasks

To evaluate the impact of new features on other evaluation tasks, we follow Denkowski and Lavie (2010a), tuning versions of Meteor to maximize length-weighted sentence-level Pearson's r correlation coefficient with adequacy and H-TER (Snover et al., 2006) scores of translations. Data sets include 2008 and 2009 NIST Open Machine Translation Evaluation adequacy data (Przybocki, 2009) and GALE P2 and P3 H-TER data (Olive, 2005). For each type of judgment, metric versions are tuned and tested on each year and scores are compared. We compare Meteor 1.3 results with those from version 1.2 with results shown in Table 6. For both adequacy data sets, Meteor 1.3 significantly outperforms version 1.2 on both tune and test data. The version tuned on MT09 data is selected as the official Adequacy version of Meteor 1.3. H-TER versions either show no improvement or degradation due to overfitting. Examination of the optimal H-TER parameter sets reveals a mismatch between evaluation metric and human judgment type. As H-TER evaluation is ultimately limited by the TER aligner, there is no distinction between content and function words, and words sharing stems are considered non-matches. As such, these features do not help Meteor improve correlation, but rather act as a source of additional possibility for overfitting.

6 MT System Tuning Experiments

The 2011 WMT Tunable Metrics task consists of using Z-MERT (Zaidan, 2009) to tune a pre-built Urdu-English Joshua (Li et al., 2009) system to a new evaluation metric on a tuning set with 4 reference translations and decoding a test set using the resulting parameter set. As this task does not provide a

Language	α	β	γ	δ	w_{exact}	w_{stem}	w_{syn}	w_{par}
English	0.85	0.20	0.60	0.75	1.00	0.60	0.80	0.60
Czech	0.95	0.20	0.60	0.80	1.00	–	–	0.40
French	0.90	1.40	0.60	0.65	1.00	0.20	–	0.40
German	0.95	1.00	0.55	0.55	1.00	0.80	–	0.20
Spanish	0.65	1.30	0.50	0.80	1.00	0.80	–	0.60

Table 4: Optimal Meteor parameters for WMT target languages on 2009 and 2010 data (Meteor 1.3 Ranking)

devtest set, we select a version of Meteor by exploring the effectiveness of using multiple versions of the metric to tune phrase-based translation systems for the same language pair.

We use the 2009 NIST Open Machine Translation Evaluation Urdu-English parallel data (Przybocki, 2009) plus 900M words of monolingual data from the English Gigaword corpus (Parker et al., 2009) to build a standard Moses system (Hoang et al., 2007) as follows. Parallel data is word aligned using the MGIZA++ toolkit (Gao and Vogel, 2008) and alignments are symmetrized using the "grow-diag-final-and" heuristic. Phrases are extracted using standard phrase-based heuristics (Koehn et al., 2003) and used to build a translation table and lexicalized reordering model. A standard SRI 5-gram language model (Stolke, 2002) is estimated from monolingual data. Using Z-MERT, we tune this system to baseline metrics as well as the versions of Meteor discussed in previous sections. We also tune to a balanced Tuning version of Meteor designed to minimize bias. This data set provides a single set of reference translations for MERT. To account for the variance of MERT, we run end-to-end tuning 3 times for each metric and report the average results on two unseen test sets: newswire and weblog. Test set translations are evaluated using BLEU, TER, and Meteor 1.2. The parameters for each Meteor version are listed in Table 7 while the results are listed in Table 8.

The results are fairly consistent across both test sets: the Tuning version of Meteor outperforms BLEU across all metrics while versions of Meteor that perform well on other tasks perform poorly in tuning. This illustrates the differences between evaluation and tuning tasks. In evaluation tasks, metrics are engineered to score 1-best translations from systems most often tuned to BLEU. As listed in Table 7, these parameters are often skewed to emphasize the differences between system outputs. In the tuning scenario, MERT optimizes translation quality with respect to the tuning metric. If a metric is biased (for example, assigning more weight to recall than precision), it will guide the MERT search toward pathological translations that receive lower scores across other metrics. Balanced between precision and recall, content and function words, and word choice versus fragmentation, the Tuning version of Meteor is significantly less susceptible to gaming. Chosen as the official submission for WMT 2011, we believe that this Tuning version of Meteor will further generalize to other tuning scenarios.

Newswire			
Tuning Metric	BLEU	TER	Met1.2
BLEU	23.67	72.48	50.45
TER	25.35	59.72	48.60
TER-BLEU/2	26.25	61.66	49.69
Meteor-tune	24.89	69.54	51.29
Meteor-rank	19.28	94.64	49.78
Meteor-adq	22.86	77.27	51.40
Meteor-hter	25.23	66.71	50.90
Weblog			
Tuning Metric	BLEU	TER	Met1.2
BLEU	17.10	76.28	41.86
TER	17.07	64.32	39.75
TER-BLEU/2	18.14	65.77	40.68
Meteor-tune	18.07	73.83	42.78
Meteor-rank	14.34	98.86	42.75
Meteor-adq	16.76	81.63	43.43
Meteor-hter	18.12	70.47	42.28

Table 8: Average metric scores for Urdu-English systems tuned to baseline metrics and versions of Meteor

Task	α	β	γ	δ	w_{exact}	w_{stem}	w_{syn}	w_{par}
Ranking	0.85	0.20	0.60	0.75	1.00	0.60	0.80	0.60
Adequacy	0.75	1.40	0.45	0.70	1.00	1.00	0.60	0.80
H-TER	0.40	1.50	0.35	0.55	1.00	0.20	0.60	0.80
Tuning	0.50	1.00	0.50	0.50	1.00	0.50	0.50	0.50

Table 7: Parameters for Meteor 1.3 tasks

7 Conclusions

We have presented Ranking, Adequacy, and Tuning versions of Meteor 1.3. The Ranking and Adequacy versions are shown to have high correlation with human judgments except in cases of overfitting due to skewed tuning data. We believe that these overfitting issues are lessened when tuning to combined 2009 and 2010 data due to increased variety in translation characteristics. The Tuning version of Meteor is shown to outperform BLEU in minimum error rate training of a phrase-based system on small Urdu-English data and we believe that it will generalize well to other tuning scenarios. The source code and all resources for Meteor 1.3 and the version of Z-MERT with Meteor integration will be available for download from the Meteor website.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proc. of ACL WIEEMMTS 2005*.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proc. of ACL2005*.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. 2010. Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. In *Proc. of ACL WMT/MetricsMATR 2010*.

Michael Denkowski and Alon Lavie. 2010a. Choosing the Right Evaluation for Machine Translation: an Examination of Annotator and Automatic Metric Performance on Human Judgment Tasks. In *Proc. of AMTA 2010*.

Michael Denkowski and Alon Lavie. 2010b. METEOR-NEXT and the METEOR Paraphrase Tables: Improve Evaluation Support for Five Target Languages. In *Proc. of ACL WMT/MetricsMATR 2010*.

Qin Gao and Stephan Vogel. 2008. Parallel Implementations of Word Alignment Tool. In *Proc. of ACL WSETQANLP 2008*.

Hieu Hoang, Alexandra Birch, Chris Callison-burch, Richard Zens, Rwth Aachen, Alexandra Constantin, Marcello Federico, Nicola Bertoldi, Chris Dyer, Brooke Cowan, Wade Shen, Christine Moran, and Ondej Bojar. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proc. of ACL 2007*.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proc. of NAACL/HLT 2003*.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proc. of MT Summit 2005*.

Zhifei Li, Chris Callison-Burch, Chris Dyer, Juri Ganitkevitch, Sanjeev Khudanpur, Lane Schwartz, Wren Thornton, Jonathan Weese, and Omar Zaidan. 2009. Joshua: An Open Source Toolkit for Parsing-based Machine Translation. In *Proc. of WMT 2009*.

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 ACL 2002*.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2009. English Gigaword Fourth Edition. Linguistic Data Consortium, LDC2009T13.

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

Mark Przybocki. 2009. NIST Open Machine Translation 2009 Evaluation. http://www.itl.nist.gov/iad/mig/tests/mt/2009/.

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*.

Andreas Stolke. 2002. SRILM - an Extensible Language Modeling Toolkit. In *Proc. of ICSLP 2002*.

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

Omar F. Zaidan. 2009. Z-MERT: A Fully Configurable Open Source Tool for Minimum Error Rate Training of Machine Translation Systems. *The Prague Bulletin of Mathematical Linguistics*.