METEOR-Tuned Phrase-Based SMT: CMU French-English and Haitian-English Systems for WMT 2011

Michael Denkowski and Alon Lavie

CMU-LTI-11-011

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213
www.lti.cs.cmu.edu

Abstract

This report describes the machine translation system tuning experiments leading to CMU system submissions to the WMT 2011 French-English and Haitian-English translation tracks. For each language track, we tune a standard phrase-based SMT system to a variety of metrics including BLEU, TER, and several variations of METEOR. We select a balanced Tuning version of METEOR that performs well across tracks as the tuning metric for our official submissions.

1 Introduction

Shared evaluations such as the Workshop on Statistical Machine Translation and Metrics MATR (Callison-Burch et al., 2010) have spurred the development of many sophisticated automatic metrics for machine translation evaluation. While newer metrics have consistently outperformed the baseline BLEU in system evaluation, BLEU remains the de facto standard for system tuning, with attempts to tune systems to other metrics yielding mixed results. This is largely due to the mismatch between metrics designed to perform well on final 1-best outputs from BLEU-tuned systems and the thousands of erroneous and often pathological translations encountered in n-best MERT. For example, the version of the METEOR metric tuned to ranking judgments places a much larger weight on recall than precision, effectively guiding the MERT search toward pathologically verbose translations.

For WMT 2011, we tune standard phrase-based statistical machine translation systems to baseline metrics and several versions of METEOR designed specifically for the tuning task. Our official submissions to the French-English and Haitian-English tracks are tuned to a final Tuning version of METEOR (METEOR-T), selected for consistently strong performance and minimal bias.

2 Minimum Error Rate Training

Minimum error rate training (MERT) (Och, 2003) is an efficient technique for optimizing a log-linear machine translation system to fit an evaluation metric over a set of sentences. As the objective function is based only on the system's $\arg\max$ for each sentence in a tuning set, optimization is typically conducted using line searches over the non-smooth error surface. If metric scores are obtained for the critical points where a new $\arg\max$ is preferred by the system, it is possible to efficiently find the global optimum along any search direction due to the linear relationship between system parameters and system score. We use the Z-MERT implementation (Zaidan, 2009), which performs MERT over n-best lists with the convergence criteria that no previously unseen translation hypotheses are generated in an iteration. For each iteration, several random initial points are considered in addition to the best performing point from previous iterations to reduce the chances of converging on a poor local optimum.

3 Objective Functions for MERT

This section discusses the evaluation metrics we examine as objective functions for MERT. This includes the baseline BLEU and TER metrics, (often combined as TER-BLEU/2), and the METEOR metric, including the tuning versions introduced for this work.

3.1 BLEU

The BLEU metric (Papineni et al., 2002) scores translation hypotheses according to surface form n-gram precision. For every n-gram length up to N, ($N = 4$ in the widely used BLEU-4 variant), an individual precision score P_n is calculated as the percentage of n-grams in the hypothesis also found in a reference translation. Precision scores are combined using a geometric mean and scaled by a brevity penalty. As BLEU does not explicitly measure recall, the brevity penalty is used to prevent short, high precision translations from receiving inflated scores. The penalty (BP) is based on the length of the translation hypothesis (h) and reference (r):

$$\mathrm{BP} = \begin{cases} 1 & \text{if } h > r \\ e^{(1-r/h)} & \text{if } h \leq r \end{cases}$$

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{4} \frac{1}{4} \log P_n\right)$$

3.2 TER

Translation edit rate (TER) (Snover et al., 2006) measures the minimum number of edit operations required to transform a translation hypothesis into a reference translation, normalized by the average length of all available references. Edit operations include word-level insertions, deletions, and substitutions as well as span-level shifts. Shifts move contiguous spans of words from one location to another within the hypothesis, preventing misplaced but locally correct phrases from incurring excessive edit costs. All operations (including shifts) are considered single edits.

$$\text{TER} = \frac{\text{min edits}}{\text{avg reference words}}$$

TER scores range from 0 to infinity, (though in practice scores tend not to exceed 1 except for very short sentences), with lower scores indicating better translations.

3.3 METEOR

The METEOR metric (Banerjee and Lavie, 2005; Denkowski and Lavie, 2010), (current version METEOR-NEXT), scores machine translation hypotheses by aligning them to reference translations and computing sentence-level similarity scores based on the alignments. When multiple references are available, hypotheses are scored against each and the reference producing the highest score is selected.

For each hypothesis-reference pair, the space of possible alignments is identified using the following types of matches:

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

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

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

Paraphrase: Phrases are matched if they are listed as paraphrases in the METEOR paraphrase tables (Denkowski and Lavie, 2010).

All matches are generalized to the phrase level with a start position and length in each sentence. Words occurring within *length* of the start position in each sentence are considered to be covered by the match. A final alignment is selected as the largest subset of 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 matched phrases 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 aligning words and phrases that occur at similar positions in both sentences).

Once an alignment is selected, the metric score is calculated as follows. The number of words in the translation hypothesis (h) and reference (r) are counted. For each match type (m_i), count the number of words covered by these matches in the hypothesis ($m_i(h)$) and reference ($m_i(r)$) and apply matcher weight (w_i). The weighted Precision and Recall are then calculated:

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

The parameterized harmonic mean of P and R 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) and number of chunks (ch):

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

The final METEOR score is then calculated:

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

Parameters				
Variant	α	β	γ	
rank	0.75	0.60	0.35	
strict	0.50	0.10	0.90	
fair	0.50	1.00	0.90	
lenient	0.50	1.00	0.50	
Match Weights				
Variant	w_{ex}	w_{stem}	w_{syn}	w_{para}
rank	1.00	0.80	0.80	0.60
strict	1.00	0.10	0.10	0.10
fair	1.00	0.50	0.50	0.50
lenient	1.00	0.90	0.90	0.90

Table 1: Parameters and match weights for METEOR variants

The parameters α, β, γ, and $w_i...w_n$ are tuned to maximize correlation with human judgments. The current default version of METEOR, METEOR-NEXT-Rank (METEOR-R), is tuned to human rankings of translation hypotheses from WMT 2009 (Callison-Burch et al., 2009).

3.4 METEOR Variants

The number of free parameters in METEOR allows the metric to be adapted to emphasize various aspects of translation quality. Whereas previous work tunes the parameters to maximize correlation with human judgments, we set the parameters to encode various types of objective functions for MERT.

To develop tuning versions of METEOR, we first balance the relative weights of precision and recall (α=0.5) and impose harsher fragmentation penalties (β and γ) to prevent pathologically verbose translations from receiving inflated scores as reported by He and Way (2009) and Cer et al. (2010). We then explore the parameter spaces of (1) the fragmentation penalty and (2) the weight given to non-exact matches as applied to scoring n-best lists in MERT.

The fragmentation penalty jointly penalizes reordering and gaps between spans of matches in translation hypotheses. We test versions of METEOR with the following fragmentation penalties:

strict: Fragmentation is penalized harshly and a fragmented translation can lose up to 90% of its score. (Strongly emphasize word order and penalize extraneous words; precision and recall break ties.)

fair: More fragmentation is tolerated, but a very fragmented translation can still lose most of its score.

lenient: More fragmentation is tolerated and a very fragmented translation can lose up to half of its score. (Balance between precision and recall and fragmentation. This variant is still harsher than the fragmentation penalty in the Ranking version of METEOR.)

The weights given to non-exact matches determine the level of lexical variation tolerated in hypotheses. Versions of METEOR include the following weight configurations:

strict: Non-exact matches receive weight 0.1, effectively limiting their role to breaking ties among translations with the maximum number of exact matches.

fair: Non-exact matches receive half the weight of exact matches, encoding a preference for exact matches but allowing for flexibility in word choice and paraphrasing.

lenient: Non-exact matches receive only slightly less weight than exact matches so that two matches of any type will lead to a higher score than any one match.

Table 1 lists the parameters and match weights for each configuration as well as for the default Ranking version of METEOR. The cross product of these configurations plus the baseline version of METEOR provide a total of 10 versions of the metric. The variant corresponding to "lenient" fragmentation and "fair" match weighting, (the most balanced configurations), is ultimately chosen as the Tuning version of METEOR (METEOR-T) due to its consistent performance in experiments. This is the version used to tune our French-English and Haitian-English WMT systems.

4 Systems / Experiments

For each language track, we build a standard phrase-based Moses system (Hoang et al., 2007) using freely available WMT data. We then use the Z-MERT (Zaidan, 2009) MERT implementation to tune the system to BLEU, TER, TER-BLEU/2, METEOR-NEXT-Rank (METEOR-R), and the range of METEOR variants. To account for the natural instability of minimum error rate training, we run

French-English		
Parallel	Sentences	14M
Monolingual	Sentences	45M
	Words	1.2B
Tuning	Sentences	2,051
Test	Sentences	2,489
Haitian-English		
Parallel	Sentences	93K
Monolingual	Sentences	45M
(out of domain)	Words	1.2B
Monolingual	Sentences	17K
(in domain)	Words	366K
Clean Tuning	Sentences	900
Clean Test	Sentences	900
Raw Tuning	Sentences	900
Raw Test	Sentences	900

Table 2: Available data for selected language tracks

MERT to convergence twice for each metric in the French-English track and three times for each metric in the Haitian-English tracks, each time evaluating the resulting system's translations on the unseen test set. We then report more reliable average metric scores over all tune-test runs.

4.1 Data Sets

For each system, we build translation models using only freely available WMT data (Callison-Burch et al., 2011). For the French-English track, this consists of parallel French-English parliamentary proceedings, United Nations documents, and news commentary. We do not use the Giga-FrEn corpus. For the Haitian-English track, this includes out-of-domain parallel data such as medical and news documents, Wikipedia articles, and glossaries and dictionaries, plus in-domain parallel SMS data. For both tracks, large monolingual English news domain data is available in addition to the English side of each parallel data set.

The French-English track includes development and test data in the news domain while the Haitian-English track includes two distinct sets of development and test data: "clean" data, manually post-edited SMS messages, and "raw" data, unedited SMS messages. All SMS messages are anonymized, with named entity tags replacing proper names, email addresses, and phone numbers. Data sizes are listed in Table 2

4.2 Phrase-Based System

For each language track, we construct a phrase-based SMT system as follows. Parallel and monolingual data are preprocessed with language-appropriate Moses tokenizers[1] and word aligned with the MGIZA++ toolkit (Gao and Vogel, 2008), an efficient multi-threaded implementation of standard statistical word alignment models (Och and Ney, 2003). Alignments are symmetrized using the "grow-diag-final-and" heuristic. Phrases are extracted from the symmetrized alignments using standard phrase-based heuristics (Koehn et al., 2003) and used to build a translation table and lexicalized reordering model. A SRI 5-gram language model (Stolke, 2002) with modified Kneser-Ney smoothing is estimated from monolingual data. Due to the availability of small in-domain data for the Haitian-English track, domain-specific models are estimated separately and interpolated.[2] For the French-English track, monolingual data is pooled to estimate a single language model.

For each individual track, (the single French-English news track and the "clean" and "raw" Haitian English SMS tracks), the phrase-based system is tuned toward each metric listed in Section 3 using Z-MERT. French-English systems are tuned and evaluated twice per metric and Haitian-English systems three times per metric. Official submissions to WMT 2011 consist of the best performing tuning run for each track using the Tuning version of METEOR (METEOR-T). Results for these systems on unseen test sets are listed in Section 5.

5 Results

This section reports the performance of systems tuned to each metric when translating unseen test sets. METEOR variants are named for their behavior on fragmentation and lexical choice. For example, METEOR-strict-fair refers to a "strict" fragmentation penalty and "fair" non-exact match weights.

[1] We use the French language settings for Haitian tokenization as French is the closest supported language.

[2] It is notable that while the in-domain model is estimated on only 0.03% of the total data, it receives over 80% of the interpolation weight due to the specific nature of the target domain.

We evaluate translations with BLEU, TER, and METEOR-NEXT-Rank (METEOR-R), metrics placing varying degrees of importance on various aspects of translation quality. As such, we look for agreement across metrics to indicate significant improvement. The reported score for a system tuned to a particular metric for a particular track is the average of its scores over all tune-test runs. All results are shown in Table 3.

In the French-English track, BLEU, TER-BLEU/2, and the METEOR variants perform similarly while METEOR-R and TER produce better scores according to themselves at the expense of other metric scores.

The small amount of training data makes the Haitian-English tracks more problematic for metrics. BLEU, TER, and METEOR-R all overfit while METEOR variants are more balanced across metrics. In the "raw" track, tuning to METEOR-T produces a significantly improved TER score over tuning to BLEU while maintaining nearly identical BLEU and METEOR-R scores.

5.1 Runtime Considerations

Integrated with the Z-MERT framework, METEOR-T scores 500 hypotheses per second on a 2.33GHz processor, leading to no noticeable increase in per-iteration wall time compared to BLEU tuning. The number of iterations required for end-to-end METEOR-T MERT is comparable to BLEU subject to the natural variance of MERT.

6 Conclusion

A balanced Tuning version of the METEOR metric relying on flexible word and phrase matching and a balance between precision, recall, and fragmentation penalty is shown to perform well across language tracks. While some skewed METEOR configurations outperform METEOR-T in some tracks, we believe that its balanced nature will allow it to better combat overfitting in new data scenarios. Our official submissions to the WMT11 French-English and Haitian-English tracks consist of the phrase-based system we have described tuned to METEOR-T. Our version of Z-MERT with METEOR-T integration will be freely available for download under an open source license from the METEOR website.

French-English Track			
Tuning Metric	BLEU	TER	MET-R
BLEU	28.27	53.94	54.07
TER	26.16	52.51	52.22
TER-BLEU/2	28.26	52.95	53.77
MET-R	27.05	56.30	54.44
MET-strict-strict	27.99	54.24	54.05
MET-strict-fair	27.80	54.52	54.14
MET-strict-lenient	27.81	54.51	54.11
MET-fair-strict	28.07	54.17	54.03
MET-fair-fair	28.20	53.94	54.06
MET-fair-lenient	27.98	54.29	54.16
MET-lenient-strict	28.36	53.61	53.97
MET-T (lenient-fair)	28.14	54.14	54.11
MET-lenient-lenient	27.84	54.44	54.21
Haitian-English Clean Track			
Tuning Metric	BLEU	TER	MET-R
BLEU	30.25	54.84	50.89
TER	28.37	52.34	49.48
TER-BLEU/2	29.64	52.56	50.39
MET-R	30.07	55.67	51.22
MET-strict-strict	30.07	52.90	50.73
MET-strict-fair	29.97	52.84	50.88
MET-strict-lenient	29.85	52.98	50.70
MET-fair-strict	29.88	52.65	50.60
MET-fair-fair	29.74	52.75	50.58
MET-fair-lenient	29.61	53.07	50.57
MET-lenient-strict	29.61	53.09	50.44
MET-T (lenient-fair)	29.62	53.30	50.50
MET-lenient-lenient	29.74	53.24	50.72
Haitian-English Raw Track			
Tuning Metric	BLEU	TER	MET-R
BLEU	25.99	61.44	46.77
TER	24.06	58.38	44.94
TER-BLEU/2	25.14	58.28	45.78
MET-R	25.67	63.15	47.12
MET-strict-strict	25.84	59.87	46.64
MET-strict-fair	25.92	60.25	46.69
MET-strict-lenient	25.81	60.17	46.64
MET-fair-strict	25.61	59.71	46.41
MET-fair-fair	25.91	60.64	46.73
MET-fair-lenient	25.67	59.90	46.55
MET-lenient-strict	25.90	60.36	46.67
MET-T (lenient-fair)	25.95	59.85	46.80
MET-lenient-lenient	25.62	60.10	46.65

Table 3: Results on unseen test data (WMT11 devtest)

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*.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proc. of ACL WMT 2009*.

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*.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, and Omar Zaidan. 2011. Findings of the 2011 Joint Workshop on Statistical Machine Translation. In *Proc. of ACL WMT 2011*.

Daniel Cer, Christopher D. Manning, and Daniel Jurafsky. 2010. The Best Lexical Metric for Phrase-Based Statistical MT System Optimization. In *Proc. of NAACL/HLT 2010*.

Michael Denkowski and Alon Lavie. 2010. 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*.

Yifan He and Andy Way. 2009. Improving the Objective Function in Minimum Error Rate Training. In *Proc. of MT Summit 2009*.

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*.

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

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29.

Franz Josef Och. 2003. Minimum Error Rate Training for Statistical Machine Translation. In *Proc. of ACL 2003*.

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*.

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

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*.

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*.