METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments

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

Abstract

METEOR is an automatic metric for Machine Translation evaluation which has been demonstrated to have high levels of correlation with human judgments of translation quality, significantly outperforming the more commonly used BLEU metric. It is one of several automatic metrics used in this year's shared task within the ACL WMT-07 workshop. This paper recaps the technical details underlying the metric and describes recent improvements in the metric. The latest release includes improved metric parameters and extends the metric to support evaluation of MT output in Spanish, French and German, in addition to English.

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. Automatic metrics are useful for comparing the performance of different systems on a common translation task, and can be applied on a frequent and ongoing basis during MT system development. 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) have described the details underlying the metric and have extensively compared its performance with BLEU and several other MT evaluation metrics. This paper recaps the technical details underlying METEOR and describes recent improvements in the metric. The latest release extends METEOR to support evaluation of MT output in Spanish, French and German, in addition to English. Furthermore, several parameters within the metric have been optimized on language-specific training data. We present experimental results that demonstrate the improvements in correlations with human judgments that result from these parameter tunings.

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 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

Proceedings of the Second Workshop on Statistical Machine Translation, pages 228–231,
Prague, June 2007. ©2007 Association for Computational Linguistics

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 parameterized 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:

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

In all previous versions of METEOR , the values of the three parameters mentioned above were set to be: $\alpha = 0.9$, $\beta = 3.0$ and $\gamma = 0.5$, based on experimentation performed in early 2004. In the latest release, we tuned these parameters to optimize correlation with human judgments based on more extensive experimentation, as reported in section 4.

3 METEOR Implementations for Spanish, French and German

We have recently expanded the implementation of METEOR to support evaluation of translations in Spanish, French and German, in addition to English. Two main language-specific issues required adaptation within the metric: (1) language-specific word-matching modules; and (2) language-specific parameter tuning. The word-matching component within the English version of METEOR uses stemming and synonymy modules in constructing a word-to-word alignment between translation and reference. The resources used for stemming and synonymy detection for English are the Porter Stemmer (Porter, 2001) and English WordNet (Miller and Fellbaum, 2007). In order to construct instances of METEOR for Spanish, French and German, we created new language-specific "stemming" modules. We use the freely available *Perl* implementation packages for Porter stemmers for the three languages (Humphrey, 2007). Unfortunately, we have so far been unable to obtain freely available WordNet resources for these three languages. METEOR versions for Spanish, French and German therefore currently include only "exact" and "stemming" matching modules. We are investigating the possibility of developing new synonymy modules for the various languages based on alternative methods, which could then be used in place of WordNet. The second main language-specific issue which required adaptation is the tuning of the three parameters within METEOR , described in section 4.

4 Optimizing Metric Parameters

The original version of METEOR (Banerjee and Lavie, 2005) has instantiated values for three parameters in the metric: one for controlling the relative weight of precision and recall in computing the Fmean score (α); one governing the shape of the penalty as a function of fragmentation (β) and one for the relative weight assigned to the fragmentation penalty (γ). In all versions of METEOR to date, these parameters were instantiated with the values $\alpha = 0.9$, $\beta = 3.0$ and $\gamma = 0.5$, based on early data experimentation. We recently conducted a more thorough investigation aimed at tuning these parameters based on several available data sets, with the goal of finding parameter settings that maximize correlation with human judgments. Human judgments come in the form of "adequacy" and "fluency" quantitative scores. In our experiments, we looked at optimizing parameters for each of these human judgment types separately, as well as optimizing parameters for the sum of adequacy and fluency. Parameter adapta-

Corpus	Judgments	Systems
NIST 2003 Ara-to-Eng	3978	6
NIST 2004 Ara-to-Eng	347	5
WMT-06 Eng-to-Fre	729	4
WMT-06 Eng-to-Ger	756	5
WMT-06 Eng-to-Spa	1201	7

Table 1: Corpus Statistics for Various Languages

tion is also an issue in the newly created METEOR instances for other languages. We suspected that parameters that were optimized to maximize correlation with human judgments for English would not necessarily be optimal for other languages.

4.1 Data

For English, we used the NIST 2003 Arabic-to-English MT evaluation data for training and the NIST 2004 Arabic-to-English evaluation data for testing. For Spanish, German and French we used the evaluation data provided by the shared task at last year's WMT workshop. Sizes of various corpora are shown in Table 1. Some, but not all, of these data sets have multiple human judgments per translation hypothesis. To partially address human bias issues, we *normalize* the human judgments, which transforms the raw judgment scores so that they have similar distributions. We use the normalization method described in (Blatz et al., 2003). Multiple judgments are combined into a single number by taking their average.

4.2 Methodology

We performed a "hill climbing" search to find the parameters that achieve maximum correlation with human judgments on the training set. We use Pearson's correlation coefficient as our measure of correlation. We followed a "leave one out" training procedure in order to avoid over-fitting. When n systems were available for a particular language, we train the parameters n times, leaving one system out in each training, and pooling the segments from all other systems. The final parameter values are calculated as the mean of the n sets of trained parameters that were obtained. When evaluating a set of parameters on test data, we compute segment-level correlation with human judgments for each of the systems in the test set and then report the mean over all systems.

4.3 Results

4.3.1 Optimizing for Adequacy and Fluency

We trained parameters to obtain maximum correlation with normalized adequacy and fluency judgments separately and also trained for maximal correlation with the sum of the two. The resulting optimal parameter values on the training corpus are shown in Table 2. Pearson correlations with human judgments on the test set are shown in Table 3.

	Adequacy	Fluency	Sum
α	0.82	0.78	0.81
β	1.0	0.75	0.83
γ	0.21	0.38	0.28

Table 2: Optimal Values of Tuned Parameters for Different Criteria for English

	Adequacy	Fluency	Sum
Original	0.6123	0.4355	0.5704
Adequacy	0.6171	0.4354	0.5729
Fluency	0.6191	0.4502	0.5818
Sum	0.6191	0.4425	0.5778

Table 3: Pearson Correlation with Human Judgments on Test Data for English

The optimal parameter values found are somewhat different than our previous metric parameters (lower values for all three parameters). The new parameters result in moderate but noticeable improvements in correlation with human judgments on both training and testing data. Tests for statistical significance using bootstrap sampling indicate that the differences in correlation levels are all significant at the 95% level. Another interesting observation is that precision receives slightly more "weight" when optimizing correlation with fluency judgments (versus when optimizing correlation with adequacy). Recall, however, is still given more weight than precision. Another interesting observation is that the value of γ is higher for fluency optimization. Since the fragmentation penalty reflects word-ordering, which is closely related to fluency, these results are consistent with our expectations. When optimizing correlation with the sum of adequacy and fluency, optimal values fall in between the values found for adequacy and fluency.

4.3.2 Parameters for Other Languages

Similar to English, we trained parameters for Spanish, French and German on the available WMT-06 training data. We optimized for maximum correlation with human judgments of adequacy, fluency and for the sum of the two. Resulting parameters are shown in Table 4.3.2. For all three languages, the parameters that were found to be optimal were quite different than those that were found for English, and using the language-specific optimal parameters re-

	Adequacy	Fluency	Sum
French:α	0.86	0.74	0.76
β	0.5	0.5	0.5
γ	1.0	1.0	1.0
German:α	0.95	0.95	0.95
β	0.5	0.5	0.5
γ	0.6	0.8	0.75
Spanish:α	0.95	0.62	0.95
β	1.0	1.0	1.0
γ	0.9	1.0	0.98

Table 4: Tuned Parameters for Different Languages

sults in significant gains in Pearson correlation levels with human judgments on the training data (compared with those obtained using the English optimal parameters)[1]. Note that the training sets used for these optimizations are comparatively very small, and that we currently do not have unseen test data to evaluate the parameters for these three languages. Further validation will need to be performed once additional data becomes available.

5 Conclusions

In this paper we described newly developed language-specific instances of the METEOR metric and the process of optimizing metric parameters for different human measures of translation quality and for different languages. Our evaluations demonstrate that parameter tuning improves correlation with human judgments. The stability of the optimized parameters on different data sets remains to be investigated for languages other than English. We are currently exploring broadening the set of features used in METEOR to include syntax-based features and alternative notions of synonymy. The latest release of METEOR is freely available on our website at: `http://www.cs.cmu.edu/~alavie/METEOR/`

Acknowledgements

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.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2003. Confidence Estimation for Machine Translation. Technical Report Natural Language Engineering Workshop Final Report, Johns Hopkins University.

Marvin Humphrey. 2007. Perl Interface to Snowball Stemmers. http://search.cpan.org/ creamyg/Lingua-Stem-Snowball-0.941/lib/Lingua/Stem/Snowball.pm.

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.*

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.

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

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.

Martin Porter. 2001. The Porter Stemming Algorithm. http://www.tartarus.org/ martin/PorterStemmer/index.html.

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.

[1]Detailed tables are not included for lack of space.