The METEOR Metric for Automatic Evaluation of Machine Translation

Alon Lavie & Michael Denkowski
{alavie,mdenkows}@cs.cmu.edu
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, 15213, USA

Abstract.

The METEOR Automatic Metric for Machine Translation evaluation, originally developed and released in 2004, was designed with the explicit goal of producing sentence-level scores which correlate well with sentence-level human judgments of translation quality. Several key design decisions were incorporated into METEOR in support of this goal. In contrast with IBM's BLEU , which uses only precision-based features, METEOR uses and emphasizes recall in addition to precision, a property that has been confirmed by several metrics as being critical for high correlation with human judgments. METEOR also addresses the problem of reference translation variability by utilizing flexible word matching, allowing for morphological variants and synonyms to be taken into account as legitimate correspondences. Furthermore, the feature ingredients within METEOR are parameterized, allowing for the tuning of the metric's free parameters in search of values that result in optimal correlation with human judgments. Optimal parameters can be separately tuned for different types of human judgments and for different languages. We discuss the initial design of the METEOR metric, subsequent improvements, and performance in several independent evaluations in recent years.

1. Introduction

The development of automatic metrics for MT evaluation has been an active research area in recent years. Evaluation of MT systems can be made faster, simpler, and less expensive by using automatic metrics in place of trained human evaluators. IBM's BLEU metric (Papineni et al., 2002) has been the most widely used automatic metric in recent years. BLEU is fast, easy to run, and can be used as a target function in parameter optimization training methods commonly used in state-of-the-art statistical MT systems (Och, 2003). However, while popular, weaknesses have been noted in BLEU in recent years, most notably the lack of reliable sentence-level scores. METEOR, along with other metrics such as GTM (Melamed et al., 2003), TER (Snover et al., 2006) and CDER (Leusch et al., 2006), were developed with the aim of specifically addressing this and other weaknesses identified in BLEU .

First developed and released in 2004, METEOR was explicitly designed with the goal of possessing high-levels of correlation with human judgments of MT output quality at the sentence level. To a large extent, METEOR is based on measures of lexical similarity between an MT translation that is being evaluated (the *hypothesis*) and reference translations for the same source sentence. To measure this similarity, METEOR first establishes an explicit word-to-word matching between each MT hypothesis and one or more reference translations. One key innovation of METEOR has been in the way it addresses translation variability. Since the same meaning can be reflected using different lexical choices, the word-to-word matcher used by METEOR can match not only exact words, but also morphological variants and synonyms. Similar approaches for flexible matching were later adopted by other automatic metrics. These *unigram matches*, based on surface forms, word stems, and word meanings (Banerjee and Lavie, 2005), form an alignment between the hypothesis and the reference. All possible alignments are scored based on a combination of features including unigram-precision, unigram-recall, and fragmentation with respect to the reference. The best scoring alignment among all possible alignments over all reference translations is selected to derive the segment-level score. The component statistics for this score are then also used in the calculation of the aggregate system-level score for the entire test set.

One of the early observations that motivated the design of METEOR was the importance of *recall* as a metric component (Lavie et al., 2004). Many other metrics have since confirmed this critical importance of recall and incorporated it as a metric component. Another key innovation in METEOR is the ability to tune several free parameters within the metric in order to optimize correlation with various forms of human judgments and for various languages (Lavie and Agarwal, 2007).

This paper describes the motivation and development of the METEOR metric. We include results from several independent evaluations from recent years that compare the performance of METEOR against other automatic metrics. We end the paper with a description of the main current and future work planned for the metric.

2. Weaknesses of the BLEU Metric Addressed by METEOR

The main ideas and principles that underline the development of METEOR arose out of a number of observations of potential weaknesses in the BLEU metric (Papineni et al., 2002). BLEU is based on the concept of n-gram precision over multiple reference translations. n-grams (consecutive sub-strings) from each MT hypothesis are checked against a

set of reference translations, and precision is calculated as the fraction of n-grams which can be matched in the reference translations out of the total number of n-grams in the hypothesis. This is performed for n-grams ranging in length from one to n. Precision is calculated independently for each n-gram order and combined into a single score through geometric averaging. BLEU does not directly measure recall, the fraction of matched n-grams in the hypothesis out of the total number of n-grams in *the reference translation*. The notion of recall in BLEU is not well defined, since BLEU was designed to match against multiple reference translations simultaneously. Rather, BLEU compensates for lack of recall with a *Brevity Penalty* which lowers the scores of MT hypotheses that are significantly shorter than the reference translations (thus artificially inflating precision scores).

Even though the BLEU metric is widely used and has greatly driven progress in statistical MT, it suffers from several weaknesses which we specifically aimed to address in the design of our METEOR metric:

- **Lack of Recall**: Our early experiments (Lavie et al., 2004) led us to believe that the lack of recall within BLEU was a significant weakness, and that the "Brevity Penalty" in the BLEU metric does not adequately compensate for the lack of recall. It has since been demonstrated by several evaluations of metrics that recall strongly correlates with human judgments of translation quality, and that recall is thus an extremely important feature component in automatic metrics (Lavie et al., 2004).

- **Use of Higher Order N-grams for Fluency and Grammaticality**: BLEU uses higher order n-grams to encapsulate and indirectly measure fluency and grammaticality in a translation hypothesis. We conjectured that flexible matching of unigrams was sufficient for assessing lexical similarity, and that a direct measure of reordering between hypothesis and reference can better capture the notions of fluency and grammaticality and can be incorporated as a feature in automatic metrics.

- **Use of Geometric Averaging of N-grams**: Geometric averaging of n-gram scores produces a result of zero whenever any of the individual n-gram scores are zero. As a result, sentence-level BLEU scores are highly unreliable. Although the BLEU metric was designed to be used on entire test sets, sentence-level scores are extremely important and useful for making fine-grained distinctions between systems. METEOR was thus designed to be a robust, sentence-level metric.

3. Design of the METEOR Metric

3.1. The METEOR Matcher

METEOR evaluates a translation hypothesis by computing a score based on an explicit word-to-word matching between the hypothesis and a given reference translation. If multiple references are provided, the hypothesis is scored against each independently and the best scoring pair is used (Banerjee and Lavie, 2005).

For each pair of translations to be compared, the METEOR Matcher is used to create a word alignment between the hypothesis and the reference. The alignment is a one-to-one mapping between words in both strings 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 in the Matcher:

- **Exact**: Words are matched based only on surface forms; a match is made if and only if the two words are identical.
- **Stem**: Words are stemmed using the Snowball Stemmer Collection (Porter, 2001). Two words match if they have identical stems.
- **Synonymy**: Words are matched if they are synonyms of one another. Words are considered synonymous if they share any synonym sets according to an external database. For English, we use the WordNet database (Miller and Fellbaum, 2007).

By default the Exact, Stem, and Synonymy modules are called in order. After each module is applied, the aligned words are fixed so that the next module only considers previously unaligned words. The largest subset of these mappings is then identified as our maximum cardinality alignment. If more than one such alignment is found, the Matcher selects the alignment which best preserves word order (fewest "crossing" unigram mappings).

3.2. The METEOR Scorer

Once a final alignment exists between a hypothesis translation and a reference translation, the METEOR score is produced as follows. Based on the number of mapped unigrams found between the two strings (m), the total number of unigrams in the hypothesis (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):

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

Our precision, recall, and Fmean are all based on single-word matches. To account for the extent to which the unigrams in both translations are in the same order, a fragmentation penalty is computed as follows. First, the sequence of matched unigrams between the two translations 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 translations is calculated as:

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

METEOR assigns a score between 0 and 1 to each individual segment. In addition, the aggregate precision, recall, and fragmentation are tracked over all segments, and a final system level score is calculated by applying the above formulas to these aggregate statistics.

3.3. Free Parameters

All versions of METEOR to date use three parameters in calculating the final score: one for controlling the relative weights of precision and recall in the Fmean score (α), one for controlling the shape of the penalty as function of fragmentation (β), and one for the relative weight assigned to the fragmentation penalty (γ).

In the earliest versions of METEOR, the values of the above parameters were set to $\alpha = 0.9$, $\beta = 3.0$ and $\gamma = 0.5$ (Banerjee and Lavie, 2005). The following section describes the adjustment of these parameters to improve correlation with human judgment.

Table I. Corpus Statistics for Various Languages

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

4. Tuning and Extending METEOR

4.1. Optimizing for Adequacy and Fluency Judgments

In 2007, we investigated tuning the three free parameters in the METEOR metric based on several available data sets, with the goal of finding an optimal set of parameters which maximized correlation with human judgments. We first explored tuning to "adequacy" and "fluency" quantitative scores, both separately and in conjunction (Lavie and Agarwal, 2007).

For our first optimization experiments in English, we used the NIST 2003 Arabic-to-English MT evaluation data for tuning and the 2004 Arabic-to-English evaluation data for testing. For optimization in Spanish, French, and German, described in the following section, we used the WMT 2006 evaluation data. Sizes of these corpora are shown in Table I. Scores from data sets with multiple human judgments per translation hypothesis were combined by taking their average. All human judgments were also normalized using the method described in (Blatz et al., 2003), so that judgment scores would have similar distributions, thus minimizing human bias.

We conducted a "hill climbing" search to find parameter values which achieve maximum correlation with human judgments on the training data, using Pearson's correlation coefficient as our measure of correlation. To avoid over-fitting, we used a "leave one out" approach. For our n available systems, we train the parameters n times, leaving one system out of each training run and pooling the segments from the other systems. We then calculate the final parameters as the mean of the n sets of trained parameters. To evaluate our trained parameters, we compute segment-level correlation with the human judgments for each system in the test set and report the mean over all systems.

We tuned parameters to maximize correlation with adequacy and fluency separately, as well as tuning to a sum of the two. The optimal

Table II. Optimal Values of Tuned Parameters for Different Criteria in English

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

parameter values for English, shown in Table II, are all lower than the original metric parameters. The result is a measurable improvement in correlation with human judgment on both training and test data. Furthermore, bootstrap sampling indicates that the differences in correlation are all statistically significant at the 95% level. An interesting observation is that precision receives noticeably more weight when tuning to fluency judgments than when tuning to adequacy judgments, though recall is always weighted more than precision. The value of *gamma* is higher for fluency optimization, which has the effect of increasing the fragmentation penalty. This reflects the fact that correct word ordering is more important for fluency.

4.2. Meteor for Different Languages

To fully support a language, Meteor requires a stemmer, a synonymy source such as WordNet, and sufficient human judgment data to optimize parameter weights. As the Snowball stemmer collection used by Meteor includes support for other European languages and MT evaluations such as NIST and WMT provide human judgment data in these languages, we were able to train Meteor systems for additional languages with both the surface form and stemming modules.

Using the WMT 2006 data, we conducted similar tuning experiments on Spanish, French, and German. Once again, we optimized parameters to adequacy, fluency and a sum of the two, producing the values listed in Table III. In each case, the optimal parameters that were found were quite different from those obtained for English, and using these language-tuned new parameters to score translations in their respective languages resulted in better Pearson correlation levels compared to the original English parameters (Lavie and Agarwal, 2007).

Table III. Optimal Values of Tuned Parameters for Different Criteria Across Languages

	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

4.3. Optimizing for Ranking Judgments

(Callison-Burch et al., 2007) reported that inter-coder agreement on the task of assigning ranks to translation hypotheses was found to be much higher than inter-coder agreement on the task of simply assigning a numeric score to a single hypothesis. This led to the adoption of ranking judgments in WMT 2008 and the increased availability of these judgments for metric tuning. Since Meteor parameters had previously been tuned to optimize correlation with adequacy and fluency, we decided to retrain Meteor to optimize correlation with ranking judgments. This required computing full rankings according to the metric and the human judges and computing a suitable correlation measure of these rankings. As Meteor assigns a score between zero and one to each hypothesis, we can easily obtain a ranking by simply ordering a list of hypotheses by their respective Meteor scores. Human rankings are available as binary judgments which create independent rankings for hypothesis pairs. In some cases, both hypotheses are judged to be equal. To obtain full rankings, we process the data in the following way:

1. Remove all equal judgments.

2. Construct a directed graph with nodes corresponding to translation hypotheses and edges corresponding to binary judgments between hypotheses.

Table IV. Corpus Statistics for Various Languages

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

3. Execute a topological sort on the directed graph, assigning ranks in the sort order. Cycles are broken by assigning the same rank to all nodes in the cycle.

To measure correlation, we compute the Spearman correlation between the human rankings and the METEOR rankings corresponding to each single source sentence (Ye et al., 2007). For N translation hypotheses where D is the difference in ranks assigned to a hypothesis by two rankings, the Spearman correlation is given by:

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

A final score is obtained by averaging the Spearman correlations for the individual sentences.

We used the human judgment data from the WMT 2007 shared evaluation task to tune our metric. The available data is summarized in Table IV. Since not every hypothesis pair was judged by the same number of judges, we considered the judgment given by the majority of judges in cases where multiple judgments were available. We performed an exhaustive grid search of the feasible parameter space to maximize correlation over the training data (Agarwal and Lavie, 2008). Using 3-fold cross-validation, we chose the best performing set of parameters on the pooled data from all folds.

The optimal parameter values are shown in Table V while the average Spearman correlations using the original and re-tuned parameters are compared in Table VI. There is significant improvement for all languages tested, with particularly significant increases in correlation for German and French. While recall was already weighted significantly, it seems that ranking judgments are driven almost entirely by recall across all the languages. Further, the re-tuned parameters are quite similar across the languages, with the exception of German.

Table V. Optimal Values of Tuned Parameters for Ranking

	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 VI. Average Spearman Correlation with Human Rankings for METEOR on Development Data

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

5. Performance in Open Evaluations

Multiple versions of the METEOR metric have been submitted to recent MT evaluations for independent analysis of correlation with various types of human judgments. All versions of METEOR are as described in Section 3, while versions "meteor-0.6" and "meteor-0.7" are tuned to adequacy and fluency judgments as described in Section 4.1 and "meteor-rank" is tuned to ranking judgments as described in Section 4.3.

5.1. WMT 2008 Evaluation Task

Raw human judgment scores for the WMT 2008 Translation Task systems were converted into three forms of ranks: the percent of time that sentences produced were judged better than or equal to those of any other system, the percent of time that constituent translations were judged better than or equal to those of any other system, and the percent of time that constituent translations were judged acceptable (Callison-Burch et al., 2008). Table VII reports the correlation performance of several evaluated metrics with these rank judgments using Spearman's rank correlation coefficient ρ.

Table VII. WMT 2008 Evaluation Task: System-level Correlation of Metrics with Human Judgments for Translations into English (Top 6 of 13 Entries)

	Rank	Constituent	Yes/No	Overall
meteor-rank	**.81**	.72	.77	**.76**
ULCh	.68	.79	.82	**.76**
meteor-0.7	.77	.75	.74	.75
posbleu	.77	**.8**	.66	.74
pos4gramFmeasure	.75	.62	.82	.73
ULC	.66	.67	**.84**	.72

Table VIII. MATR 2008 System-Level Scores for Top Performing Metrics Across Categories (10 of 39 Entries)

	Single Reference Track			Multiple Reference Track		
	Adequacy	Yes/No	Pairwise	Adequacy	Yes/No	Pairwise
TERp	-0.8685	**-0.8729**	**-0.7481**	-0.8659	-0.8761	-0.7685
SEPIA1	0.8689	0.8459	0.7015	0.9031	**0.8826**	**0.7718**
EDPM	0.8797	0.8509	0.6816	0.9218	0.8747	0.7522
meteor-0.7	0.8968	0.8554	0.6933	0.8830	0.8763	0.7312
BLEU-v12	0.8567	0.8483	0.6935	0.9008	0.8711	0.7571
CDer	**-0.9037**	-0.8390	-0.6797	-0.9167	-0.8575	-0.7258
SEPIA2	0.8620	0.8308	0.6812	0.9307	0.8646	0.7375
NIST-v11b	0.8775	0.8384	0.6866	**0.9314**	0.8569	0.7127
Bleu-sbp	0.8679	0.8407	0.6835	0.9113	0.8686	0.7314
meteor-rank	0.8906	0.8572	0.6982	0.8444	0.8616	0.7182

5.2. NIST Metrics MATR 2008

Introduced in 2008, the NIST MetricsMATR Challenge presents a series of challenge tracks aimed at promoting the development of more accurate MT evaluation metrics. For each submitted metric, scores were computed using single and multiple reference sets separately, and correlation with several types of human judgments was calculated. Table VIII reports the Spearman's rank correlation coefficient ρ for three types of human judgments: a 7-point adequacy scale judgment, a yes-no sufficient quality judgment, and a pair-wise ranking judgment.

Table IX. WMT 2009 Evaluation Task: System-level Correlation of Metrics with Human Judgments for Translations into English (Top 8 of 19 Entries)

	de-en	fr-en	es-en	cz-en	hu-en	Average
ulc	**.78**	.92	.86	**1**	.6	**.83**
maxsim	.76	.91	**.98**	.7	.66	.8
rte (absolute)	.64	.91	.96	.6	**.83**	.79
meteor-rank	.64	**.93**	.96	.7	.54	.75
rte (pairwise)	.76	.59	.78	.8	**.83**	.75
terp	-.72	-.89	-.94	-.7	-.37	-.72
meteor-0.6	.56	**.93**	.87	.7	.54	.72
meteor-0.7	.55	**.93**	.86	.7	.26	.66

5.3. WMT 2009 Evaluation Task

Similarly to WMT 2008, the raw human judgment scores for the WMT 2009 Translation Task systems were converted into ranking judgments of adequacy. Table IX reports the correlation performance of several metrics with these judgments, using Spearman's rank correlation coefficient ρ.

6. Discussion and Ongoing Work

6.1. Flexible Matching

As mentioned in previous sections, the Meteor matcher creates a word-level alignment between two sentences, matching surface forms, shared stems, or synonyms. The Meteor matcher can also be used as a "stand-alone" component, and cane be incorporated into other metrics and systems.

M-Bleu **and** M-Ter : Many widely used metrics such as Bleu (Papineni et al., 2002) and Ter (Snover et al., 2006) are based on measuring the similarity of sentences by matching words or groups of words between two strings. These metrics, however, are limited to finding exact, surface form matches. We conducted experiments using both Bleu and Ter in which we first aligned a translation hypothesis to a reference using our matcher, and then replaced any non-surface-form matches with the corresponding words in the reference translation. We could then run Bleu and Ter on the adjusted hypothesis, thereby simulating flexible matching. We also used the smoothed Bleu described in

(Lin and Och, 2004) to ensure meaningful segment-level BLEU scores. The resulting metrics, which we call M-BLEU and M-TER , achieve higher average Spearman correlation with human ranking judgments in almost all cases[1] (Agarwal and Lavie, 2008).

MT System Combination: MT system combination is an active area of research that aims to combine the output generated by multiple MT systems operating on the same input, with the goal of producing translations that are superior to all of the original MT systems. The system combination approach described by (Heafield et al., 2009) does this by creating alignments between translation hypotheses from various systems and selecting phrases based on the alignments. Using the METEOR flexible matcher, this system can better align hypotheses from systems which are prone to different vocabulary selection, and can use features based on these alignments when constructing synthetic combined hypotheses.

6.2. CURRENT WORK

In May 2009, we released a new version of METEOR that is much faster and specifically tailored to support Minimum Error Rate Training (MERT) for MT systems. A code reimplementation of the metric resulted in a several-fold speed increase and provides library functionality in both a traditional environment and a distributed "Map-Reduce" environment. The other improvements over previous versions discussed in earlier sections of this paper include:

Length Penalty: METEOR now employs a length cost which prevents exceedingly long hypotheses with high recall but low precision from receiving excessively high scores. The acceptable length envelope is a function of the length of the reference translation, and if multiple references are available, is applied on a per-reference basis.

Generic Synonymy: The synonymy module has been redesigned to support a generic synonymy source consisting of a list of synonymy-sets and a stemmer which produces word forms as they appear in the synonym-sets. Though we currently use data extracted from the WordNet database (Miller and Fellbaum, 2007), the module can now use synonymy data from any source, and can support languages other than English.

[1] Detailed results omitted for lack of space, see cited publication for detailed analysis

References

Agarwal, A. and A. Lavie: 2008, 'Meteor, m-bleu and m-ter: Evaluation Metrics for High-Correlation with Human Rankings of Machine Translation Output'. In: *Proceedings of the Third ACL Workshop on Statistical Machine Translation*. Columbus, Ohio.

Banerjee, S. and A. 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*. Ann Arbor, Michigan, pp. 65–72.

Blatz, J., E. Fitzgerald, G. Foster, S. Gandrabur, C. Goutte, A. Kulesza, A. Sanchis, and N. Ueffing: 2003, 'Confidence Estimation for Machine Translation'. Technical Report Natural Language Engineering Workshop Final Report, Johns Hopkins University.

Callison-Burch, C., C. Fordyce, P. Koehn, C. Monz, and J. Schroeder: 2007, '(Meta-) Evaluation of Machine Translation'. In: *Proceedings of the Second Workshop on Statistical Machine Translation*. Prague, Czech Republic, pp. 136–158.

Callison-Burch, C., C. Fordyce, P. Koehn, C. Monz, and J. Schroeder: 2008, 'Further Meta-Evaluation of Machine Translation'. In: *Proceedings of the Third Workshop on Statistical Machine Translation*. Columbus, Ohio, pp. 70–106.

Heafield, K., G. Hanneman, and A. Lavie: 2009, 'Machine Translation System Combination with Flexible Word Ordering'. In: *Proceedings of the Fourth Workshop on Statistical Machine Translation*. Athens, Greece, pp. 56–60.

Lavie, A. and A. 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*. Prague, Czech Republic, pp. 228–231.

Lavie, A., K. Sagae, and S. 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)*. Washington, DC, pp. 134–143.

Leusch, G., N. Ueffing, and H. 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*.

Lin, C.-Y. and F. J. 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*. Morristown, NJ, USA, p. 501.

Melamed, I. D., R. Green, and J. Turian: 2003, 'Precision and Recall of Machine Translation'. In: *Proceedings of the HLT-NAACL 2003 Conference: Short Papers*. Edmonton, Alberta, pp. 61–63.

Miller, G. and C. Fellbaum: 2007, 'WordNet'. http://wordnet.princeton.edu/.

Och, F. J.: 2003, 'Minimum Error Rate Training for Statistical Machine Translation'. In: *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu: 2002, 'BLEU: a Method for Automatic Evaluation of Machine Translation'. In: *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics (ACL)*. Philadelphia, PA, pp. 311–318.

Porter, M.: 2001, 'Snowball: A language for stemming algorithms'. http://snowball.tartarus.org/texts/introduction.html.

Snover, M., B. Dorr, R. Schwartz, L. Micciulla, and J. 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)*. Cambridge, MA, pp. 223–231.

van Rijsbergen, C.: 1979, *Information Retrieval*. London, UK: Butterworths, 2nd edition.

Ye, Y., M. Zhou, and C.-Y. Lin: 2007, 'Sentence Level Machine Translation Evaluation as a Ranking'. In: *Proceedings of the Second Workshop on Statistical Machine Translation*. Prague, Czech Republic, pp. 240–247.