Mach Translat (2009) 23:105–115
DOI 10.1007/s10590-009-9059-4

The METEOR metric for automatic evaluation of machine translation

Alon Lavie · Michael J. Denkowski

Received: 16 May 2009 / Accepted: 13 October 2009 / Published online: 1 November 2009

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 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.

Keywords Machine translation · MT Evaluation · Automatic metrics

1 Introduction

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

A. Lavie (✉) · M. J. Denkowski
Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA
e-mail: alavie@cs.cmu.edu

M. J. Denkowski
e-mail: mdenkows@cs.cmu.edu

(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). 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 specifically to address these 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 establishes an explicit word-to-word matching between each MT hypothesis and one or more reference translations. One key innovation of Meteor has been its addressing of 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 used in the calculation of the aggregate system-level score for the full test set.

One early observation that motivated the design of Meteor was the importance of *recall* as a metric component (Lavie et al. 2004). Other metrics have since confirmed this critical importance and incorporated recall as a metric component. Another key innovation in Meteor is the ability to tune 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 an overview of the current and future work planned for the metric. All versions of Meteor are available for download at: http://www.cs.cmu.edu/~alavie/METEOR/.

2 Weaknesses of the Bleu metric addressed by Meteor

The main principles that underline the development of Meteor arose from 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. BLEU compensates for lack of recall with a *Brevity Penalty* which lowers the scores of hypotheses that are significantly shorter than the reference translations (thus artificially inflating precision scores).

Although 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 translation hypotheses. 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 zero result 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 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 a 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 translation pair, the Matcher creates a word alignment between the hypothesis string and reference string incrementally through a sequence of stages, each corresponding to one of METEOR's word-mapping modules:

- *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 a Snowball Stemmer (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).

Each stage begins with the identification of all possible unigram mappings between the two strings using the specified module. The largest subset of these mappings is then selected such that every word in each string maps to *at most one word* in the other string. If more than one such alignment is found, the Matcher selects the alignment which best preserves word order (fewest "crossing" unigram mappings). This process is implemented via greedy search with a limit on the maximum number of computations.

At the conclusion of each stage, the aligned words are fixed so that any subsequent module considers only words unaligned in previous stages. By default the Exact, Stem, and Synonymy modules are called in order.

3.2 The METEOR scorer

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

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

Our precision, recall, and F_{mean} are all based on single-word matches. To account for preservation of word order, a fragmentation penalty is computed as follows. First, the sequence of matched unigrams between the two strings is divided into the smallest number of "chunks" such that matched unigrams in each chunk are adjacent (in both strings) and in identical order. The counts of chunks (ch) and matches (m) are then used to calculate a fragmentation fraction: frag = ch/m. The penalty is then computed as:

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

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:

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

METEOR assigns a score between 0 and 1 to each individual segment. In addition, aggregate counts of matches (m), test unigrams (t), reference unigrams (r), and chunks (ch) are collected for the entire test set. The above formulas are then applied to these counts to calculate the system level F_{mean}, Pen, and METEOR score.

3.3 Free parameters

METEOR currently uses three free parameters when calculating final scores: one for controlling the relative weights of precision and recall in the F_{mean} 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).

The values of the above parameters were initially 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.

4 Tuning and extending METEOR

4.1 Optimizing for adequacy and fluency judgments

In 2007, we investigated tuning the free parameters in METEOR based on several available data sets to find 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 English, we used the NIST 2003 Arabic-to-English MT evaluation data for tuning and the 2004 Arabic-to-English data for testing. For optimization in Spanish, French, and German, described in the following section, we used the WMT 2006 evaluation data.[1] Scores from data sets with multiple human judgments per translation hypothesis were combined by taking their average. All judgments were 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. We used a "leave one out" training procedure in order to avoid over-fitting. When n systems were available for a particular language, we trained the parameters n times, leaving one system out in each training, and pooling the segments from all other systems. The final parameter values were 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.

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

[1] Corpus statistics omitted for lack of space, see Lavie and Agarwal (2007) for additional information.

Table 1 Optimal values of tuned parameters for 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

in Table 1, are all lower than the original metric parameters. The *alpha*, *beta*, and *gamma* tuned to adequacy-fluency sums are used in versions 0.6 and 0.7 of METEOR. The result is a measurable improvement in correlation with human judgment on both training and test data. Bootstrap sampling indicates that the differences in correlation are all statistically significant at the 95% level.[2] We observed 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 increases the fragmentation penalty. This reflects the fact that correct word ordering is more important for fluency.

4.2 METEOR for different languages

As the stemmers used by METEOR already include 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. Again, we optimized parameters to adequacy, fluency and a sum of the two, producing the values listed in Table 2. In each case, the final parameters were quite different from those obtained for English, and using these new language-tuned parameters to score translations in their respective languages resulted in better Pearson correlation levels compared to the original English parameters (Lavie and Agarwal 2007). The parameters tuned to adequacy-fluency sums are used in versions 0.6 and 0.7 of METEOR for French, German, and Spanish.

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 much higher than agreement on the task of 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. We decided to retrain METEOR to optimize correlation with these ranking judgments. This required computing full rankings according to the metric and the human judges and computing a suitable correlation measure. As METEOR assigns a score between zero and one to each hypothesis, we can obtain a ranking by ordering

[2] For details on correlation levels, see Lavie and Agarwal (2007).

Table 2 Optimal values of tuned parameters 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

a list of hypotheses by their 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.
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). 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.[3] In cases where multiple judgments were available, we considered the judgment given by the majority of judges. 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 3 while the average Spearman correlations using the original and re-tuned parameters are compared in Table 4. 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 lan-

[3] Judgment data statistics omitted for lack of space, see Agarwal and Lavie (2008) for additional information.

Table 3 Optimal values of tuned parameters for ranking

	English	German	French	Spanish
α	0.95	0.90	0.90	0.90
β	0.5	3.0	0.5	0.5
γ	0.45	0.15	0.55	0.55

Table 4 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

Table 5 WMT 2008 evaluation task: system-level correlation of metrics with human judgments for translations into English (top 5 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	**.80**	.66	.74
pos4gramFmeasure	.75	.62	**.82**	.73

guages. Further, the re-tuned parameters are quite similar across the languages, with the exception of German.

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 Sect. 3, while versions "meteor-0.6" and "meteor-0.7" are tuned to adequacy and fluency judgment sums as described in Sects. 4.1 and 4.2, and "meteor-rank" is tuned to ranking judgments as in Sect. 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 5 reports the correlation of several evaluated metrics with these rank judgments using Spearman's rank correlation coefficient ρ.

Table 6 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	**1.0**	.60	**.83**
maxsim	.76	.91	**.98**	.70	.66	.80
rte (absolute)	.64	.91	.96	.60	**.83**	.79
meteor-rank	.64	**.93**	.96	.70	.54	.75
rte (pairwise)	.76	.59	.78	.80	**.83**	.75
terp	−.72	−.89	−.94	−.70	−.37	−.72
meteor-0.6	.56	**.93**	.87	.70	.54	.72
meteor-0.7	.55	**.93**	.86	.70	.26	.66

Table 7 MetricsMATR 2008 Pearson's correlation for top performing metrics across categories in multiple reference track (10 of 39 entries)

	Segment level			System level		
	Adequacy	Yes/no	Pairwise	Adequacy	Yes/no	Pairwise
meteor-0.7	0.737	0.559	0.373	0.874	0.849	0.681
meteor-0.6	0.733	0.582	0.368	0.848	0.845	0.676
Terp	−0.722	−0.595	−0.371	−0.866	−0.861	−0.705
CDer	−0.720	−0.555	−0.345	−0.904	−0.834	−0.68
BleuSP	0.687	0.582	0.360	0.849	0.857	0.703
meteor-rank	0.710	0.580	0.357	0.849	0.851	0.683
SVM-rank	0.718	0.576	0.385	0.844	0.860	0.707
LET	0.678	0.495	0.381	0.920	0.792	0.684
ATEC3	0.647	0.493	0.358	0.923	0.782	0.660
SEPIA1	0.653	0.531	0.358	0.900	0.862	0.716

5.2 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 6 reports the correlation of several metrics with these judgments, using Spearman's rank correlation coefficient ρ (Callison-Burch et al. 2009).

5.3 NIST MetricsMATR 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 7 reports the Pearson's correlation coefficient for three types of human judgments on the multiple reference track.

In the adequacy task, evaluators judged how much meaning expressed in a reference translation was successfully captured by a hypothesis, assigning a score from 1 (none) to 7 (all). This was followed by a binary judgment of whether or not a hypothesis meant essentially the same as the reference. In a pair-wise ranking task, judges were asked which of two hypotheses they preferred given a reference, with an option for no preference. Detailed analyses and results of additional evaluation tasks can be found in the official results for MetricsMATR 2008 (Przybocki et al. 2008).

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. This matcher can also be used as a "stand-alone" component, and can be incorporated into other metrics, systems, and applications. One concrete example of such an application is MT system combination. MT system combination 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 reimplemented version of METEOR that is much faster and specifically tailored to support Minimum Error Rate Training (MERT) for MT systems in both traditional or distributed environments. Other improvements beyond the versions discussed in this paper include:

Length penalty METEOR now supports a length cost intended to prevent exceedingly long hypotheses with high recall but low precision from receiving excessively high scores. An acceptable length envelope is implemented as a parametric function of the length of the reference translation, and if multiple references are available, is applied on a per-reference basis. Current work includes fine tuning the function parameters to yield the best cost function to guide system tuning.

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.

Acknowledgements This work was supported in part by NFS grant IIS-0534932. We would like to thank the following individuals for contributing to the development and implementation of the METEOR metric: Satanjeev Banerjee, Abhaya Agarwal, Rachel Reynolds, Kenji Sagae, Jeremy Naman, and Shyamsundar Jayaraman.

References

Agarwal A, Lavie A (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, OH, pp 115–118

Banerjee S, Lavie A (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, MI, pp 65–72

Blatz J, Fitzgerald E, Foster G, Gandrabur S, Goutte C, Kulesza A, Sanchis A, Ueffing N (2003) Confidence estimation for machine translation. Technical report natural language engineering workshop final report, Johns Hopkins University

Callison-Burch C, Fordyce C, Koehn P, Monz C, Schroeder J (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, Fordyce C, Koehn P, Monz C, Schroeder J (2008) Further meta-evaluation of machine translation. In: Proceedings of the third workshop on statistical machine translation. Columbus, OH, pp 70–106

Callison-Burch C, Koehn P, Monz C, Schroeder J (2009) Findings of the 2009 workshop on statistical machine translation. In: Proceedings of the fourth workshop on statistical machine translation. Athens, Greece, pp 1–28

Heafield K, Hanneman G, Lavie A (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, Agarwal A (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, Sagae K, Jayaraman S (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, Ueffing N, Ney H (2006) CDER: efficient MT evaluation using block movements. In: Proceedings of the thirteenth conference of the European chapter of the association for computational linguistics. pp 241–248

Melamed ID, Green R, Turian J (2003) Precision and recall of machine translation. In: Proceedings of the HLT-NAACL 2003 conference: short papers. Edmonton, Alberta, pp 61–63

Miller G, Fellbaum C (2007) WordNet. http://wordnet.princeton.edu/

Och FJ (2003) Minimum error rate training for statistical machine translation. In: Proceedings of the 41st annual meeting of the association for computational linguistics. pp 160–167

Papineni K, Roukos S, Ward T, Zhu W-J (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

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

Snover M, Dorr B, Schwartz R, Micciulla L, Makhoul J (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, Chap. 7, 2nd edn. Butterworths, London, UK

Ye Y, Zhou M, Lin C-Y (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