LAMP-TR-126 July 2005
CS-TR-4755
UMIACS-TR-2005-58

A STUDY OF TRANSLATION ERROR RATE WITH TARGETED HUMAN ANNOTATION

Matthew Snover†, Bonnie Dorr†, Richard Schwartz‡, John Makhoul‡, Linnea Micciulla‡, Ralph Weischedel‡

†Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742

‡BBN Technologies
10 Moulton Street
Cambridge, MA 02138

Abstract

We define a new, intuitive measure for evaluating machine translation output that avoids the knowledge intensiveness of more meaning-based approaches, and the labor-intensiveness of human judgments. Translation Error Rate (TER) measures the amount of editing that a human would have to perform to change a system output so it exactly matches a reference translation. We also compute a human-targeted TER (or HTER), where the minimum TER of the translation is computed against a human 'targeted reference' that preserves the meaning (provided by the reference translations) and is fluent, but is chosen to minimize the TER score for a particular system output. We show that: (1) The single-reference variant of TER correlates as well with human judgments of MT quality as the four-reference variant of BLEU; (2) The human-targeted HTER yields a 33% error-rate reduction and is shown to be very well correlated with human judgments; (3) The four-reference variant of TER and the single-reference variant of HTER yield higher correlations with human judgments than BLEU; (4) HTER yields higher correlations with human judgments than METEOR or its human-targeted variant (HMETEOR); and (5) The four-reference variant of TER correlates as well with a single human judgment as a second human judgment does, while HTER, HBLEU, and HMETEOR correlate significantly better with a human judgment than a second human judgment does.

This work has been supported, in part, by BBNT contract number 9500006806.

1 Introduction

Due to the large space of possible correct translations, automatic machine translation (MT) has proved a difficult task to evaluate. Human judgments of evaluation are expensive and noisy. Many automatic measures have been proposed to facilitate fast and cheap evaluation of MT systems, the most widely used of which is BLEU [7], an evaluation metric that matches n-grams from multiple references. A similar version of this metric, typically referred to as the "NIST" metric, was proposed by Doddington [2]. Other proposed methods for MT evaluation include METEOR [1], which uses unigram matches on the words and their stems, and a linear combination of automatic MT evaluation methods along with meaning-based features for identifying paraphrases [8].[1]

We seek to define a new, more intuitive measure of "goodness" of MT output—specifically, the number of edits needed to fix the output so that it semantically matches a correct translation. We attempt to avoid the knowledge intensiveness of more meaning-based approaches, and the labor-intensiveness of human judgments. We also seek to achieve higher correlations with human judgments by assigning lower costs to phrasal shifts than those assigned by n-gram-based approaches such as BLEU.

This paper presents a new measure called Translation Error Rate (TER) that determines the minimum number of edits (allowing phrasal shifts) needed to change a hypothesis so that it exactly matches one of the references. We also present a procedure to create targeted references where a fluent speaker of English creates a new reference translation targeted for this system output by editing the hypothesis until it is both fluent and has the same meaning as the reference(s). We then compute a human-targeted TER (or HTER), where the minimum TER of the translation is computed against this new reference.

We present results that indicate that HTER yields a 33% lower error rate than TER. We also compare both TER and HTER to BLEU and METEOR (and their human-targeted variants, HBLEU and HMETEOR). We show that the single-reference variant of TER correlates as well with human judgments of MT quality as the four-reference variant of BLEU. We also show that the four-reference variant of TER and the single-reference variant of HTER yield higher correlations with human judgments than BLEU—even when BLEU is given human-targeted references. Our results show that METEOR correlates with human judgments better than TER, when given the same number of references, but that HTER correlates with human judgments better than HMETEOR.

2 Related Work

The first attempts at machine translation evaluation relied on purely subjective human judgments [4]. Later work measured machine translation error by post editing machine translation output and counting the number of edits, typically measured in the number of keystrokes to convert the system output into a "canonical" human translation [3]. Attempts have been made to improve machine translation performance by automatic post-editing techniques [5]. Post editing measures have also been shown effective for text summarization evaluation [6] and natural language generation [9].

[1] One variant of the meaning-based approach incorporates the translation error rate described in this paper. We adopt a simpler evaluation paradigm that requires no meaning-based features, but still achieves correlations that are better than the existing standard, BLEU.

When developing machine translation systems, a purely automatic measure of accuracy is preferred for rapid feedback and reliability. Purely human based evaluation metrics fail in this regard and have largely been replaced by purely automatic machine translation evaluations. Automatic machine translation evaluation has traditionally relied upon string comparisons between a set of reference translations and a hypothesis translation. The quality of such automatic measures can only be determined by comparisons to human judgments. One difficulty in using these automatic measures is that their output is not meaningful except to compare one system against another.

BLEU [7] calculates the score of a translation by measuring the number of n-grams, or varying length, of the system output that occur within the set of references. This measure has contributed to the recent improvement in machine translation systems by giving developers a reliable, cheap evaluation measure on which to compare their systems. In addition to being a relatively unintuitive measure, BLEU relies upon a large number of references and a large number of sentences in order to correlate with human judgments. BLEU's highest correlations with human judgments occur when entire system data sets (often 1000 sentences) are correlated; correlations at the sentence level are much poorer.

METEOR [1] is an evaluation measure that counts the number of exact word matches between the system output and reference. Unmatched words are then stemmed and matched. Additional penalities are assessed for reordering the words between the hypothesis and reference. This method has been shown to correlate very well with human judgments.

An MT scoring measure that uses the notion of maximum matching string (MMS) has been demonstrated to yield significant correlations with human judges [10]. The MMS method is similar to our own approach, in that it only allows a string to be matched once, and also permits string reordering. The MMS approach explicitly favors long contiguous matches, whereas TER attempts to minimize the number of edits between the reference and the hypothesis. Their work also reported a poor correlation between human judges for machine translation quality.

Our work is largely inspired by the previous post-editing methodology of MT evaluation, with an eye towards a more automated solution. Our method provides both a purely automatic evaluation metric (TER), and a human-in-the-loop method (HTER), by providing a method to automatically measure post-editing distance.

3 Definition of Translation Error Rate

We define TER as the minimum number of edits needed to change a hypothesis so that it exactly matches one of the references, normalized by the average length of the references. Since we are concerned with the minimum number of edits needed to modify the hypothesis, we only measure the number of edits to the closest reference (as measured by the TER score). Specifically:

$$\text{TER} = \frac{\#\text{ of edits}}{\text{average }\#\text{ of reference words}}$$

Possible edits include the insertion, deletion, and substitution of single words as well as shifts of word sequences. A shift moves a contiguous sequence of words within the hypothesis to another location within the hypothesis. All edits, including shifts of any number of words, any distance, have equal cost. In addition, punctuation tokens are treated as normal words. In order to allow a shift, the words must be identical to the words in the reference, including capitalization.

Consider the reference/hypothesis pair below, where differences between the reference and hypothesis are indicated by upper case:

```
REF: SAUDI ARABIA denied THIS WEEK information published in the AMERICAN new york times
HYP: THIS WEEK THE SAUDIS denied information published in the new york times
```

Here, the hypothesis (HYP) is fluent and means the same thing (except for missing "American") as the reference (REF). However, TER does not consider this an exact match. First, we note that the phrase "this week" in the hypothesis is in a "shifted" position (at the beginning of the sentence rather than after the word "denied") with respect to the hypothesis. Second, we note that the phrase "Saudi Arabia" in the reference appears as "the Saudis" in the hypothesis (this counts as two separate substitutions). Finally, the word "American" appears only in the reference.

If we apply TER to this hypothesis and reference, the number of edits is 4 (1 Shift, 2 Substitutions, and 1 Insertion), giving a TER score of $\frac{4}{13} = 31\%$. BLEU also yields a poor score of 32.3% (or 67.7% when viewed as the error-rate analog to the TER score) on the hypothesis because it doesn't account for phrasal shifts adequately.

Clearly these scores do not reflect the acceptability of the hypothesis, but it would take human knowledge to determine that the hypothesis semantically matches the reference. A solution to this, using human annotators is discussed in section 4.

The number of edits for TER is calculated in two phases. The number of insertions, deletions, and substitutions is calculated using dynamic programming. A greedy search is used to find the set of shifts, by repeatedly selecting the shift that most reduces the number of insertions, deletions and substitutions, until no more beneficial shifts remain. Then dynamic programming is used to optimally calculate the remaining edit distance using a minimum-edit-distance (where insertions, deletions and substitutions all have cost 1). The number of edits is calculated for all of the references, and the best (lowest) score is used. The pseudo-code for calculating the number of edits is shown in Algorithm 1.

The greedy search is necessary to select the set of shifts because an optimal sequence of edits (with shifts) is very expensive to find. In order to further reduce the space of possible shifts, to allow for efficient computation, several other constraints are used:

1. The shifted words must match the reference words in the destination position exactly.

2. The word sequence of the hypothesis in the original position and the corresponding reference words must not exactly match.

3. The word sequence of the reference that corresponds to the destination position must be misaligned before the shift.

The min-edit-distance algorithm is $O(n^2)$ in the number of words. Therefore we use a beam search so that the evaluation code works efficiently on long sentences.

As an example, consider the following reference/hypothesis pair:

```
REF: a b c d e f   c
HYP: a     d e   b c f
```

The words "b c" in the hypothesis can be shifted to the left to correspond to the words "b c" in the reference, because there is a mismatch in the current location of "b c" in the hypothesis, and there is a mismatch of "b c" in the reference. After the shift the hypothesis is changed to:

Algorithm 1 Calculate Number of Edits

```
input: HYPOTHESIS $h$
input: REFERENCES $R$
$E \leftarrow \infty$
for all $r \in R$ do
    $h' \leftarrow h$
    $e \leftarrow 0$
    repeat
        Find shift, $s$, that most reduces min-edit-distance($h'$, $r$)
        if $s$ reduces edit distance then
            $h' \leftarrow$ apply $s$ to $h'$
            $e \leftarrow e + 1$
        end if
    until No shifts that reduce edit distance remain
    $e \leftarrow e+$ min-edit-distance($h'$, $r$)
    if $e < E$ then
        $E \leftarrow e$
    end if
end for
return $E$
```

```
REF: a b c d e f c
HYP: a b c d e f
```

TER as defined above, only calculates the number of edits between the best reference and the hypothesis. It most accurately measures the error rate of a hypothesis when that reference is the closest possible reference to the hypothesis. While predetermined references can be used to measure the error rate, the most accurate results require custom references generated with the assistance of a human annotator.

4 Human-in-the-loop Evaluation

As stated earlier, the acceptability of a hypothesis is not entirely indicated by the TER score, which ignores notions of semantic equivalence. This section describes an approach that employs human annotation for additional gains in accuracy of the edit rate described above.

Our human-in-the-loop evaluation, or HTER (for human targeted translation edit rate), involves a procedure for creating targeted references. In order to accurately measure the number of edits necessary to transform the hypothesis into a fluent English sentence with the same meaning as the references, one must do more than measure the distance between the hypothesis and the current references. Specifically, a more successful approach is one that finds the closest possible reference to the hypothesis from the space of all possible fluent references that have the same meaning as the original references. To approximate this, we use human annotators to generate a new targeted reference. We start with an automatic system output (hypothesis) and one or more human references. A fluent speaker of English creates a new reference translation targeted for this system output by editing the hypothesis until it is fluent and has the same meaning as the reference(s). We then compute the minimum TER (using the technique described above in Section 3) using this single targeted reference as a new human reference.[2]

Within this approach it is possible to reduce the cost for development within a system. Specifically, it is not necessary to create new references for each run; for many systems, most translations do not change from run to run. Moreover, we only need to create new references on sentences with a significantly increased edit rate (since the last run).

The human annotation tool that we used for our experiments displays all references and the system hypothesis. In the main window, the annotation tool shows where the hypothesis differs from the best reference, as determined by applying TER. The tool also shows the current number of edits for the "reference in progress". In addition, the surrounding reference sentences in the the document are also shown to give the annotator additional context. We found that annotators took an average of 3.5 minutes per sentence to provide a good targeted reference. The time was relatively consistent over the 4 annotators, but we believe this time could be reduced by a better annotation tool.

An example set of references, hypothesis and resulting human-targeted reference are shown below:

Ref 1: `The expert, who asked not to be identified, added, "This depends on the conditions of the bodies."`

Ref 2: `The experts who asked to remain unnamed said, "the matter is related to the state of the bodies."`

Hyp: `The expert who requested anonymity said that "the situation of the matter is linked to the dead bodies".`

Targ: `The expert who requested anonymity said that "the matter is linked to the condition of the dead bodies".`

Note, that the ambiguity regarding the term "dead bodies," which is a correct translation from the arabic, is absent in the NIST references, but present in the hypothesis and targeted reference. The annotator can reach such conclusions by using the surrounding context and world knowledge.

5 Experimental Design

In our experiments, we used the results of two MT systems, which we call S_1 and S_2. According to MTEval 2004 metrics, S_1 is one of the lower performing systems and S_2 is one of the best systems. We used 100 sentences from the MTEval 2004 Arabic evaluation data set. The sentences were chosen randomly from the set of sentences that had also been annotated with human judgments of fluency and adequacy. Four monolingual native English annotators corrected the system output. Two annotators were assigned to each sentence from each system. Annotators were coached on how to minimize the edit rate, while preserving the meaning of the reference translation.

After the initial generation of the targeted references, another pass of annotation was performed to ensure that the new targeted references were sufficiently accurate and fluent. During this second

[2] The targeted reference is the only human reference used for the purpose of measuring HTER. However, this reference is not used for computing the average reference length to avoid issues relating to crafting very long references to reduce the TER score.

pass, other annotators checked (and corrected) all targeted references for fluency and meaning without exposure to the system output. On average, this second pass changed 0.63 words per sentence. This correction pass raised the average TER score as the annotators had typically erred in favor of the system output.

6 Results

Table 1 shows that the HTER (i.e., TER with one human-targeted reference) reduces the edit rate by 33% relative to TER with 4 untargeted NIST references. Substitutions were reduced by the largest factor. In both TER and HTER, the majority of the edits were substitutions and deletions. Because TER is an edit-distance metric, lower numbers indicate better performance. In previous pilot studies with more experienced annotators, HTER yielded an even higher reduction of 50%.[3]

Condition	**Ins**	**Del**	**Sub**	**Shift**	**Total**
TER: 4 NIST Refs	4.6	12.0	25.8	7.2	49.6
HTER: 1 Targ Ref	3.5	10.5	14.6	4.9	33.5

Table 1: Untargeted (TER) and Human-targeted (HTER) Results: Average of S_1 and S_2

In an analysis of shift size and distance, we found that most shifts are short in length (1 word) and most shifts are by less than 7 words. We also did a side-by-side comparison of HTER with BLEU and METEOR (see Table 2),[4] and found that human-targeted references lower edit distance overall but the TER measure is aided more than BLEU by targeted references: HTER yields a reduction of 33% whereas HBLEU yields a reduction of 28%.[5]

We also did a study of the correlations among TER, HTER, BLEU, HBLEU, METEOR, HMETEOR and Human Judgments, as shown in Table 3. The table shows the Pearson Coefficients of correlation between each of the evaluation metrics that we measured. TER, BLEU and METEOR have been abbreviated as T, B and M, respectively. T(1) refers to the application of TER with only one NIST reference. (The reported correlation refers to an average of the four correlations, one for each reference.) B(1) and M(1) are analogously computed for BLEU and METEOR, respectively. T(4), B(4), and M(4) refer to the score computed using all four of the NIST references for TER, BLEU, and METEOR, respectively. HT, HB and HM refer to the application of TER, BLEU, and METEOR, respectively, with only one human-targeted reference.[6] (The reported correlation refers to an average over the two correlations, one for each human-targeted reference.)

[3] The annotators in this study were recent additions to our annotation group, as opposed to the veteran annotators we used in our pilot study. In addition the annotators in this study were placed under more strict time constraints, encouraging them not to spend more than five minutes on a single sentence. This tradeoff of annotation speed and annotation quality is important to consider as HTER is considered for use in machine translation evaluations.

[4] In the case of BLEU and METEOR (and the human-targeted variants), we must subtract the score from 1 to get numbers that are comparable to TER (and its human-targeted variant). That is, lower numbers indicate better scores for all three measures.

[5] It is possible that performance of HBLEU is biased, as the targeted references were designed to minimize TER scores rather than BLEU scores.

[6] The 4 NIST references were not used in calculating the HTER, HBLEU, HMETEOR metrics.

Condition	S_1	S_2	**Average**
BLEU: 4 NIST Refs	73.5	62.1	67.8
HBLEU: 1 Targ Ref	62.2	45.0	53.6
METEOR: 4 NIST Refs	46.0	39.2	42.1
HMETEOR: 1 Targ Ref	33.9	22.1	28.0
TER: 4 NIST Refs	53.2	46.0	49.6
HTER: 1 Targ Ref	39.7	27.2	33.5

Table 2: Untargeted and Human-targeted Scores: BLEU, HBLEU, METEOR, HMETEOR, TER, HTER

The Human Judgment scores are the average of fluency and adequacy judgments from both human judges. Sentences from both systems were used for a total of 200 data points—there were no significant differences from these results when only one of the systems was used. The values for the evaluation measures decrease for better values, whereas the values for human judgments increase for better values; however, for clarity, we report only the magnitude, not the sign of the correlations.

We found that correlation between HTER and Human Judgments was very high with a Pearson coefficient of 0.630, exceeding the correlation of all other metrics. T(1) and T(4) are both well correlated with HTER (r=0.606 and r=0.789, respectively), and are ideal to be used for system development where the cost of HTER is prohibitive. In addition, T(1) is shown to correlate as well with human judgments as B(4)[7] (r=0.390 and r=0.390, respectively), indicating that equally valuable results can be gained with TER using a fourth of the number of references—even without human targeting. While M(1) and M(4) correlate better with Human Judgments than T(1) and T(4), respectively, neither M(1) nor M(4) (nor even the human-targeted HM) correlate as well as with human judgments as HTER (0.493/0.550/0.602 vs. 0.630).

We also examined correlations between the two human judges and the evaluation metrics, as shown in Table 4. HJ-1 and HJ-2 refer to the two sets of human judgments, each of which is the average of fluency and adequacy. Correlating the two sets of human judgments against each other shows a correlation of only 0.478, less than the correlation of the average of the judges with HTER, or even the correlation of HTER with either individual human judgment set (r=0.506 and r=0.575). In fact, even the TER(4) measure (with untargeted references) correlates about as well with each of the human judgments (0.461 and 0.466) as each of the humans against each other (0.478).

In an analysis of standard deviation due to annotator differences (Table 5), we observed that HTER is much less sensitive to the number of references than BLEU and also that the standard deviation decreased somewhat with targeted references. To determine this, we compute variance across combinations with each sentence. We then took the average (weighted by length) across sentences and took the square root. Table 6 shows the means for each annotator. The standard deviation of these numbers is 2.8%. All annotators were given the same instructions, shown in Appendix A.

[7] Correlations on system level data points (sets of 1000 or more sentences) would have much higher correlation

Measure	T(1)									
T(1)	0.737	T(4)								
T(4)	0.792	1.000	HT							
HT	**0.606**	**0.789**	0.929	B(1)						
B(1)	0.473	0.521	0.457	0.709	B(4)					
B(4)	0.518	0.606	0.565	0.737	1.000	HB				
HB	0.502	0.624	0.794	0.535	0.687	0.919	M(1)			
M(1)	0.555	0.652	0.623	0.479	0.553	0.607	0.845	M(4)		
M(4)	0.586	0.727	0.675	0.488	0.596	0.643	0.888	1.000	HM	
HM	0.482	0.618	0.802	0.433	0.545	0.761	0.744	0.806	0.945	HJ
HJ	**0.390**	0.539	**0.630**	0.325	**0.391**	0.579	0.493	0.550	0.602	1.000

Table 3: Correlations among TER (T), BLEU (B), METEOR (M), human variants (HT, HB, HM), and Human Judgments (HJ)

Measure	T(1)	T(4)	HT	B(1)	B(4)	HB	M(1)	M(4)	HM	HJ-1	
HJ-1	0.332	0.461	**0.506**	0.270	0.341	0.461	0.446	0.502	0.511	1.000	HJ-2
HJ-2	0.339	0.466	**0.575**	0.288	0.331	0.532	0.403	0.446	0.525	**0.478**	1.000

Table 4: Correlations among TER (T), BLEU (B), METEOR (M), human variants (HT, HB, HM), and Individual Human Judgments

7 Conclusions and Future Work

We have shown that TER is adequate for research purposes as it correlates reasonably well with human judgments and also with HTER. However it gives an overestimate of the actual translation edit rate. Targeted references mitigates this issue. With 4 NIST references, the edit rate was reduced 33%.

In addition, HTER makes fine distinctions among correct, near correct, and bad translations: correct translations have HTER = 0 and bad translations have high HTER (< 1).

In our correlation experiments, we showed that BLEU and TER are highly correlated, and that HTER is more highly correlated to human judgments than BLEU or HBLEU. Although METEOR using NIST references is more highly correlated than TER using NIST references, human-targeted HTER correlates with human judgments better than METEOR, or its human-targeted variant (HMETEOR).

The correlations shown were only on single sentences; correlations on document length segments should also be explored. In addition, the HTER numbers do vary, depending on the training, skill, and effort of the annotators. In a previous pilot study, the reduction from TER to HTER was 50% instead of 33%.

TER is easy to explain to people outside of the machine translation community (i.e., the amount of work needed to correct the translations). Both TER and HTER appear to be good predictors of

values for both BLEU and TER.

Condition		Mean	Std Dev
B:	**4 NIST Refs**	67.8	-
B:	**3 of 4 NIST Refs**	71.0	5.7
B:	**1 of 4 NIST Refs**	83.2	8.9
HB:	**1 of 2 Targ Refs**	53.6	10.1
M:	**4 NIST Refs**	42.1	-
M:	**3 of 4 NIST Refs**	43.3	3.4
M:	**1 of 4 NIST Refs**	49.6	7.9
HM:	**1 of 2 Targ Refs**	28.0	6.7
T:	**4 NIST Refs**	49.6	-
T:	**3 of 4 NIST Refs**	51.0	3.4
T:	**1 of 4 NIST Refs**	57.0	8.1
HT:	**1 of 2 Targ Refs**	33.5	6.8

Table 5: Standard Deviation Due to Annotator Differences for TER (T), BLEU (B), METEOR (M), and human variants (HT, HB, HM)

Annotator	Mean HTER
1	31.5
2	36.4
3	31.1
4	29.9

Table 6: Variance Among Annotators

human judgments of translation quality. In addition, HTER may represent a method of capturing human judgments about translation quality without the need for noisy subjective judgments. The automatic TER score with 4 references correlates as well with a single human judgment as another human judgment does, while the scores with a human in the loop, such as HTER, correlate significantly better with a human judgment than a second human judgment does. This confirms that if human are to be used to judge the quality of MT output, it should be done by creating a new reference and counting errors, rather than by making subjective judgments.

Acknowledgments

This work has been supported, in part, by BBNT contract number 9500006806. We would like to thank Philip Resnik, Grazia Russo-Lassner, and Jimmy Lin for several fruitful discussions on the TER error metric and its variants, as well as Nathanial Waisbrot for discussions on the algorithmic aspects of TER. We are also thankful to Alon Lavie and his students for freely distributing the METEOR software, and to Audrey Le for providing access to the system output and human judgments upon which these experiments were conducted.

References

[1] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaulation Measures for MT and/or Summarization*, 2005.

[2] George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurence statistics. In *Human Language Technology: Notebook Proceedings*, pages 128–132, 2002.

[3] Robert Frederking and Sergei Nirenburg. Three heads are better than one. In *Proceedings of the Fourth Conference on Applied Natural Language Processing, ANLP-94*, 1994.

[4] Margaret King. Evaluating natural language processing systems. *Communication of the ACM*, pages 73–79, 1996.

[5] Kevin Knight and Ishwar Chander. Automated postediting of documents. In *Proceedings of National Conference on Artificial Intelligence (AAAI)*, 1994.

[6] Inderjeet Mani, Gary Klein, David House, and Lynette Hirschman. Summac: A text summarization evaluation. *Natural Language Engineering*, pages 43–68, 2002.

[7] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine traslation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002.

[8] Grazia Russo-Lassner, Jimmy Lin, and Philip Resnik. A paraphrase-based approach to machine translation evaluation. Technical Report LAMP-TR-125/CS-TR-4754/UMIACS-TR-2005-57, University of Maryland, College Park, 2005.

[9] Somayajulu Sripada, Ehud Reiter, and Lezan Hawizy. Evaluating an nlg system using post-editing. Technical Report AUCS/TR0402, Department of Computing Science, University of Aberdeen, 2004.

[10] Joseph P. Turian, Luke Shen, and I. Dan Melamed. Evaluation of machine translation and its evaluation. In *Proceedings of MT Summit IX*, 2003.

A TER Annotation Guidelines

This appendix presents the guidelines used by our annotators for the application of TER.

A.1 Purpose

The purpose of this task is to develop a method for improving evaluation of machine translation (MT) output. Current automatic MT evaluation requires a very large number of human translations, in order to accommodate as many legitimate machine translation variations as possible. This task aims to improve the evaluation mechanism by cutting back on the number of human translations required, through producing an acceptable translation that matches the MT output as closely as possible.

A.2 Overview

Annotators are presented with four human-generated translations (REFs) and one machine-generated (HYP) translation of the same original sentence. One of the REFs is identified as the closest match.

For each sentence, annotators create a "good enough" (TARG) translation by making a minimum number of corrections to the HYP, while preserving the meaning presented in the REFs.
(1) Example:

(a) REF1: The diplomat confirmed that "North Korea's withdrawal from the treaty starts as of today." (AFA20030111.3500-5)

(b) REF2: A diplomat confirmed that "the withdrawal (of North Korea from the treaty) will begin with effect from today". (AFA20030111.3500-5)

(c) REF3: The diplomat confirmed, "The withdrawal (of North Korea from the treaty) is effective starting today." (AFA20030111.3500-5)

(d) REF4: The diplomat confirmed that "the withdrawal [of North Korea from the Treaty] will begin today". (AFA20030111.3500-5)

(e) HYP: The diplomatic withdrawal " (North Korea of the treaty) will start today . " (AFA20030111.3500-5)

Potential TARG: The diplomat confirmed " the withdrawal (of North Korea from the
treaty) will start today . " (AFA20030111.3500-5)

A.3 Procedure

A.3.1 Recommended steps

1. Run cter.pl. Open xxx_targ.txt.

2. Read the CLOSEST REFERENCE.

3. Read the HYPOTHESIS.

4. If the HYPOTHESIS does not capture the meaning of the CLOSEST REFERENCE, you can read the other 3 REFERENCES.

5. Remove the # from the CLOSEST REFERENCE.

6. Edit the CLOSEST REFERENCE so that it is closer to the HYPOTHESIS, but still preserves the original meaning; ie. a) is grammatically correct and b) matches the meaning of at least one of the REFs.

7. Run cter.pl again.

8. Repeat steps 6-7 if you believe you can further reduce the number of edits required.

A.3.2 Using REFs

All REFs are considered valid translations.

TARG will often be a combination of 1 or more REFs. In example (1), the REF options for the final VP are:

REF1: starts as of today
REF2: will begin with effect from today
REF3: is effective starting today
REF4: will begin today

and the HYP reads:

HYP: will start today

Since at least one of the REFs does not mention "effect" or "effective" it is acceptable to leave this out of TARG, and no change should be made to HYP.

Any "errors" found in REF will be considered acceptable, unless these interfere with comprehension of the sentence.

A.3.3 How do I know when I'm done?

TARG creation should strive to reduce the number of edits as far as possible without spending a great amount of time on a single edit. Ideally, a TARG of average length should take less than 5 minutes.

A.4 Rules

There are three requirements for TARG:

1. The meaning expressed in the REFs must be preserved
2. The form must be such that it is easily understood by a native speaker of English
3. TARG must be as close to HYP as possible, without violating 1 and 2

A.4.1 Defining change

The following operations represent a change:

Deletion of a word
Insertion of a word
Substitution of a word
Shifting of a word or group of words

A *word* in the operations described above is any contiguous string of characters. Each punctuation mark is considered a single word. Any changes to a string, including seemingly minor ones such as capitalization, punctuation or inflection, will create distinct words. The following pairs constitute distinct words:

Ahmed — Ahmed's
The — the
speak — speaks

A.4.2 Preserving meaning

The compositional meaning of TARG must be equivalent to at least one REF. In the example below, the type of office referred to in REF is qualitatively different from the type provided in HYP; therefore, this HYP would not be a valid TARG.

(2) Example:
REF: an office for telecommunications
HYP: the office of special communications

Logical vs. connotative equivalence

We will consider HYP strings that are the logical equivalents of a REF string to be valid candidates for TARG. It is not required that the exact string be present in a REF.

(3) Example:
REF1: discuss the Committee's meeting to be held
REF2: to discuss a meeting of the committee which will be held
REF3: discusion of "the committee meeting due to be held
REF4: "discussed the meeting of the council to be held
HYP: "an examination of the meeting will be held

Although *examination* may connotatively suggest more thoroughness and detail than *discussion*, they can be used interchangeably to represent the same event. So in this case *examination* can be considered to be the logical equivalent of *discussion*. The following is therefore a candidate TARG:

TARG: "an examination of the meeting to be held"

Speech act verbs

Speech acts must be preserved as part of the original meaning. One type of speech act may be substituted for another, but it may not be omitted. The use of quotation marks is not equivalent to an overt reference to a speech act.

If all of the REFs are of the form, *He said, "X, Y, Z"* then it is permissible for the TARG to begin, *He said/stated/reported/claimed*, etc. It is not permissible for TARG to consist of simply *"X, Y, Z"*.

A.4.3 Maintaining grammaticality

Strict prescriptive rules do not need to be observed. However, basic agreement rules should be followed.

ACCEPTABLE: The two are leaving this evening.
NOT ACCEPTABLE: The two is leaving this evening.

So although the string, "He left today from Baghdad" may seem less natural than, "He left Baghdad today" both strings will be considered acceptable.

Proper nouns

Proper nouns referring to GPEs or organizations should be spelled according to one of the options provided in REF.

Arabic person names may have a very large number of possible spelling in English. In particular, single and double consonants are often used interchangeably in the same name, and vowels, especially {e, ee, i} and {o, u, ou} may be considered equally valid in transliteration of an Arabic name. Annotators may use the Internet to check for possible alternative spellings of Arabic names.

Proper nouns must be capitalized.

Capitalization and punctuation

The first word of the sentence and all proper names should be capitalized. Other stylistic capitalization norms are not required. The following strings are considered equivalent:

(4) Example:
REF: Turkish Prime Minister Rajab Tayyib Ardogan
TARG: Turkish prime minister Rajab Tayib Ardogan

All sentences should end in a period. Commas, semi-colons and quotation marks may be omitted.

Alternate spellings

Alternate spellings found in standard dictionaries are acceptable, including British spellings. http://dictionary.reference.com/ may be used as a reference.

Any legal contraction is permissible.

Determiner usage

Flexibility should be allowed for alternate determiners. The following substitutions are allowed:

(5) REF: An Israeli ministerial committee had decided...
TARG: The Israeli ministerial committee decided...

(6) REF: ...allow the inspectors to study the military technology of both North Korea and Iran.
TARG: ... allow the inspectors to study each military technology from North Korea and Iran.

A.5 Types/examples of acceptable/unacceptable equivalence

Types	**REFs**	**HYP/TARG**	**Acceptable?**
Abbreviations vs. full names	UN	United Nations	yes
	Associated French Press	APF	yes
Headline cap style	Tony Blair Puts Forward New Ideas to President Mubarak	Tony Blair offered to President Mubarak ideas of fresh impetus	yes
	these officers, like thousands of other Iraqis	these officers have the Iraqi and other	no
	these officers, like thousands of other Iraqis	these officers, like other Iraqi thousands	yes
Quotations	He said, “I believe that...”	“I believe that...”	no
	He said, “I believe that...”	He stated, “I believe that...”	yes