# A Study of Translation Edit Rate with Targeted Human Annotation

**Matthew Snover and Bonnie Dorr**
Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742
{snover,bonnie}@umiacs.umd.edu

**Richard Schwartz, Linnea Micciulla, and John Makhoul**
BBN Technologies
10 Moulton Street
Cambridge, MA 02138
{schwartz,lmicciul,makhoul}@bbn.com

## Abstract

We examine 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 Edit 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 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 define a human-targeted TER (or HTER) and show that it yields higher correlations with human judgments than BLEU—even when BLEU is given human-targeted references. Our results indicate that HTER correlates with human judgments better than HMETEOR and that the four-reference variants of TER and HTER correlate with human judgments as well as—or better than—a second human judgment does.

## 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 (Papineni et al., 2002), an evaluation metric that matches n-grams from multiple references. A variant of this metric, typically referred to as the "NIST" metric, was proposed by Doddington (Doddington, 2002). Other proposed methods for MT evaluation include METEOR (Banerjee and Lavie, 2005), 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 (Russo-Lassner et al., 2005).[1]

We 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. In a less expensive variant, 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.

Recently the GALE (Olive, 2005) (Global Autonomous Language Exploitation) research program introduced a new error measure called Translation Edit Rate (TER)[2] that was originally designed to

[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.

[2] Within the GALE community, the TER error measure is referred to as Translation *Error* Rate , derived from the Word Error Rate (WER) metric in the automatic speech recognition community. The name is regrettable for its implication that it

*Proceedings of the 7th Conference of the Association for Machine Translation in the Americas,* pages 223-231,
Cambridge, August 2006. ©2006 The Association for Machine Translation in the Americas

count the number of edits (including phrasal shifts) performed by a human to change a hypothesis so that it is both fluent and has the correct meaning. This was then decomposed into two steps: defining a new reference and finding the minimum number of edits so that the hypothesis exactly matches one of the references. This measure was defined such that all edits, including shifts, would have a cost of one. Finding only the minimum number of edits, without generating a new reference is the measure defined as TER; finding the minimum of edits to a new *targeted references* is defined as human-targeted TER (or HTER). We investigate both measures and also present a procedure to create targeted references where a fluent speaker of the target language 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).

The next section describes work related to our MT evaluation approach. Following this we define TER and its human-targeted variant, HTER. We then present our experimental design and results of our experiments, comparing TER and HTER to BLEU and METEOR (and their human-targeted variants, HBLEU and HMETEOR). We compare these measures against human judgments of the fluency and adequateness of the system output. Finally, we conclude with a summary of results and future work.

## 2 Related Work

The first attempts at MT evaluation relied on purely subjective human judgments (King, 1996). Later work measured MT error by post editing MT output and counting the number of edits, typically measured in the number of keystrokes to convert the system output into a "canonical" human translation (Frederking and Nirenburg, 1994). Attempts have been made to improve MT performance by automatic post-editing techniques (Knight and Chander, 1994). Post editing measures have also been shown effective for text summarization evaluation (Mani et al., 2002) and natural language generation (Sripada et al., 2004).

When developing MT 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 MT evaluations. Automatic MT evaluation has traditionally relied upon string comparisons between a set of reference translations and a translation hypothesis. 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 (Papineni et al., 2002) calculates the score of a translation by measuring the number of n-grams, of varying length, of the system output that occur within the set of references. This measure has contributed to the recent improvement in MT systems by giving developers a reliable, cheap evaluation measure on which to compare their systems. However, BLEU is relatively unintuitive and relies upon a large number of references and a large number of sentences in order to correlate with human judgments.

METEOR (Banerjee and Lavie, 2005) 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 high correlations with human judges (Turian et al., 2003). The MMS method is similar to the approach used by TER, 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. TER assigns a lower cost to phrasal shifts than MMS, and does not explicitly favor longer matching strings.

is the *definitive* MT measure. The authors make no such claim, and have adopted the name Translation *Edit* Rate for use in this paper and the wider community.

## 3 Definition of Translation Edit Rate

TER is defined 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:

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

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, by any distance, have equal cost. In addition, punctuation tokens are treated as normal words and mis-capitalization is counted as an edit. [3]

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.

Optimal calculation of edit-distance with move operations has been shown to be NP-Complete(Shapira and Storer, 2002), causing us to use the following approximation to calculate TER scores. 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. Note that a shift that reduces the number of insertions, deletions, substitutions by just one has no net reduction in cost, due to the cost of 1 for the shift itself. However, in this case, we still adopt the shift, because we find that the alignment is more correct subjectively and often results in slightly lower edit distance later on. 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

[3] It is possible that this measure could be improved by a more careful setting of edit weights, but this paper is concerned with exploring the current TER measure in use. The choice for shifts to be cost one, regardless of length or distance, seems somewhat arbitrary and other cost measures for shifts might yield higher correlations with human judgments.

**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 $E$
```

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.

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:

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

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

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-targeted Translation Edit Rate

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 to make TER be a more accurate measure of translation quality. [4]

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 target language (English in our experiments) 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, who are fluent speakers of the target language, to generate a new targeted reference. We start with an automatic system output (hypothesis) and one or more pre-determined, or *untargeted*, reference translations. They could generate the targeted reference by editing the system hypothesis or the original reference translation. We find that most editors edit the hypothesis until it is fluent and has the same meaning as the untargeted 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. 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.[5]

[4] The human-in-the-loop variant of TER is the one that will be used in the GALE MT Evaluations this year.

[5] The targeted reference is not used to compute the average

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. New targeted references only need to be created for those sentences whose translations have changed. Moreover, we probably 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 minimum number of edits for the "reference in progress." In addition, the surrounding reference sentences in the document are also shown to give the annotator additional context. We found that annotators took an average of 3 to 7 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 the term "dead bodies," which is a correct translation from the Arabic, disagrees with both references, but is 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, from MTEval 2004, which we call $S_1$ and $S_2$. According to the distribution requirements set by NIST for the shared metrics data, these systems remain anonymous. According to the 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. Each sentence had four reference translations, which we will henceforth refer to as *untargeted references*, that were used to calculate BLEU scores in the MTEval 2004 evaluation. The sentences were chosen randomly from the set of sentences that had also been annotated with human judgments of fluency and adequacy for the MTEval 2004 evaluation.[6]

In order to create the targeted references for HTER, four monolingual native English annotators corrected the system output. Two annotators were assigned to each sentence from each system.[7] Annotators were coached on how to minimize the edit rate.[8] while preserving the meaning of the reference translations. Bilingual annotators could possibly give more accurate results, since they would be better able to understand the source sentence meaning, but are much more expensive to hire and would likely take more time to complete the task. Given that four references translations are available to convey the meaning, monolingual annotators are a cheaper, more readily available alternative.

If the annotators were told to simply edit the system output, without any attempt to minimize HTER, they would lack incentive to find a good targeted reference and would favor transforming the system output using the least effort possible (e.g., by copying one of the original references). In this case much higher HTER scores result, which do not then cor-

reference length, as this would change the denominator in the TER calculation, and crafty annotators could favor long targeted references in order to minimize HTER.

[6] We used the community standard human judgments from MTEval 2004. For time and cost reasons we did not generate our own independent human judgments.

[7] Two annotators were used in this study so that issues of annotator consistency could be examined. In practice one annotator per sentence would be adequate to create the targeted references.

[8] The coaching given to the annotators in order to minimize HTER, consisted mostly of teaching them which edits were considered by TER. The annotators also consulted with each other during training to compare various techniques they found to minimize HTER scores.

respond to calculating the minimum number of edits required to transform the system output into a fluent sentence with the same meaning as the references.

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 pass, other annotators, other HTER annotators who worked on different data, 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 HTER score as the annotators had typically erred in favor of the system output. The corrections in this pass consisted almost entirely of fixing small errors, that were a result of annotator oversight, such as verb or tense agreement.

## 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 references. Substitutions were reduced by the largest factor, presumably because words were often judged synonymous. In both TER and HTER, the majority of the edits were substitutions and deletions. Because TER and HTER are edit-distance metrics, lower numbers indicate better performance. In previous pilot studies with more experienced annotators, HTER yielded an even higher reduction of 50%.[9]

| Condition | Ins | Del | Sub | Shift | Total |
|---|---|---|---|---|---|
| **TER: 4 UnTarg 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 are by less than 7 words. We also did a side-by-side comparison of HTER with BLEU and METEOR (see Table 2)[10] 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%. It is possible that performance of HBLEU is biased, as the targeted references were designed to minimize TER scores rather than BLEU scores. 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.

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. Note that the correlations along the diagonal are not always 1, as several of these are the averages of multiple correlations.

TER, BLEU and METEOR are abbreviated as T, B and M, respectively. T(1) refers to the application of TER with only one untargeted reference. (The reported correlation refers to an average of the 4 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 4 untargeted 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. (The reported correlation refers to an average of the 2 correlations, one for each human-targeted reference.)

Human Judgments refer to the average of fluency and adequacy judgments from both human judges. Sentences from both systems were used for a total of 200 data points—no significant differences were found 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 corre-

[9] 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 community-wide MT evaluations.

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

| **Condition** | $S_1$ | $S_2$ | **Average** |
|---|---|---|---|
| **BLEU: 4 UnTarg Refs** | 73.5 | 62.1 | 67.8 |
| **HBLEU: 1 Targ Ref** | 62.2 | 45.0 | 53.6 |
| **METEOR: 4 UnTarg Refs** | 46.0 | 39.2 | 42.1 |
| **HMETEOR: 1 Targ Ref** | 33.9 | 22.1 | 28.0 |
| **TER: 4 UnTarg 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

| **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)

lations.

The correlation between HTER (HT) and Human Judgments (HJ) 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 HT (r=0.606 and r=0.789, respectively), and are ideal 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) (r=0.390 and r=0.391, respectively), indicating that equally valuable results can be gained with TER using one 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 the human-targeted HM) correlate as well as with human judgments as HTER (0.493/0.550/0.602 vs. 0.630).

In an analysis of standard deviation due to annotator differences (Table 4), we observed that TER is much less sensitive to the number of references than BLEU and also that the standard deviation decreased somewhat with targeted references. To compute variance among annotators, we first compute the variances from the mean for each sentence. Then we then took the average (weighted by length) across sentences and took the square root. Table 5 shows the means for each annotator. The standard deviation of these numbers is 2.8%.

We also examined correlations between the two human judges and the evaluation metrics (see Table 6). 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 almost as well with each of the human judgments (0.461 and 0.466) as each of the humans against each other (0.478).

That HTER correlates better with average human judgments than individual human judgments correlate with each other may seem paradoxical at first. This is due to the fact that the individual judgments actually have a high degree of disagreement. The average of many individual judgments actually forms the better benchmark of performance. Much of the dilemma arises in the difficulty of assigning a subjective numerical judgment to a translation, and the

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

Table 4: 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 5: Variance Among Annotators

| 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 6: Correlations among TER (T), BLEU (B), METEOR (M), human variants (HT, HB, HM), and Individual Human Judgments

high degree of variability it entails. Given its correlation to the average human judgment, HTER seems to be a possible substitute for such subjective human judgments.

## 7 Conclusions and Future Work

We have described a new measure for the 'goodness' of MT, HTER, which is less subjective than pure human judgments. This method is expensive, in that it requires approximately 3 to 7 minutes per sentence for a human to annotate. While the studies presented here use four reference translations to create a targeted reference, fewer references would likely be adequate, reducing the cost of the method, relative to methods that require many reference translations. HTER is not suitable for use in the development cycle of an MT system, although it could be employed on a periodic basis, but does appear to be a possible substitute for subjective human judgments of MT quality.

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 error rate. Targeted references mitigates this issue. When compared with TER with 4 untargeted references, the edit rate with HTER 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, around 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 untargeted references is more highly correlated than TER using untargeted references, human-targeted HTER correlates with human judgments better than METEOR, or its human-targeted variant (HMETEOR). Future studies might benefit from also examining correlations with the MMS evaluation metric(Turian et al., 2003).

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 MT community (i.e., the amount of work needed to correct the translations). Both TER and HTER appear to be good predictors of 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 humans are to be used to judge the quality of MT output, this should be done by creating a new reference and counting errors, rather than by making subjective judgments.

## Acknowledgments

This work has been supported under the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022. 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. We would also like to thank the reviewers for their valuable feedback.

## References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaulation Measures for MT and/or Summarization*.

George Doddington. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurence Statistics. In *Human Language Technology: Notebook Proceedings*, pages 128–132.

Robert Frederking and Sergei Nirenburg. 1994. Three Heads are Better than One. In *Proceedings of the Fourth Conference on Applied Natural Language Processing, ANLP-94*.

Margaret King. 1996. Evaluating Natural Language Processing Systems. *Communication of the ACM*, 29(1):73–79, January.

Kevin Knight and Ishwar Chander. 1994. Automated Postediting of Documents. In *Proceedings of National Conference on Artificial Intelligence (AAAI)*.

Inderjeet Mani, Gary Klein, David House, and Lynette Hirschman. 2002. SUMMAC: A Text Summarization Evaluation. *Natural Language Engineering*, 8(1):43–68.

S. Niessen, F.J. Och, G. Leusch, and H. Ney. 2000. An evaluation tool for machine translation: Fast evaluation for MT research. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC-2000)*, pages 39–45.

Joseph Olive. 2005. *Global Autonomous Language Exploitation (GALE)*. DARPA/IPTO Proposer Information Pamphlet.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Traslation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

Grazia Russo-Lassner, Jimmy Lin, and Philip Resnik. 2005. 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.

Dana Shapira and James A. Storer. 2002. Edit Distance with Move Operations. In *Proceedings of the 13th Annual Symposium on Computational Pattern Matching*, pages 85–98.

Somayajulu Sripada, Ehud Reiter, and Lezan Hawizy. 2004. Evaluating an NLG System Using Post-Editing. Technical Report AUCS/TR0402, Department of Computing Science, University of Aberdeen.

Joseph P. Turian, Luke Shen, and I. Dan Melamed. 2003. Evaluation of Machine Translation and its Evaluation. In *Proceedings of MT Summit IX*.