

# SEMANTIC TRANSLATION ERROR RATE FOR EVALUATING TRANSLATION SYSTEMS

*Krishna Subramanian, Dave Stallard, Rohit Prasad, Shirin Saleem, Prem Natarajan*

BBN Technologies, 10 Moulton Street, Cambridge MA 02138  
{ksubrama, stallard, rprasad, ssaleem, pnataraj}@bbn.com

## ABSTRACT

In this paper, we introduce a new metric which we call the Semantic Translation Error Rate, or STER, for evaluating the performance of machine translation systems. STER is based on the previously published Translation Error Rate (TER)[1] and METEOR[2] metrics. Specifically, STER extends TER in two ways: first, by incorporating word equivalence measures (WordNet and Porter stemming) standardly used by METEOR, and second, by disallowing alignments of concept words to non-concept words (aka stop words). We show how these features make STER alignments better suited for human-driven analysis than standard TER. We also present experimental results that show that STER is better correlated to human judgments than TER. Finally, we compare STER to METEOR, and illustrate that METEOR scores computed using the STER alignments have similar statistical properties to METEOR scores computed using METEOR alignments.

**Index Terms**—*Automated Metric, Statistical Machine Translation*

## 1. INTRODUCTION

With the rapid spread of information technology and, in particular, the World Wide Web to most of the world, a significant amount of online electronic content is authored in languages other than English. At the same time, the globalization juggernaut has resulted in a need for efficient and effective conversations between people who speak different languages.

Statistical machine translation (SMT) systems have the potential to satisfy this demand for communicating and accessing information across language barriers and have, therefore, become very popular, both in actual use and as an area of ongoing research. Recently, researchers have shown that given an adequate parallel corpus for training the translation models, an SMT system can generate useful translations, from one language to another, of written text[3] and of spoken language[4,5,6]. Since there is no single *correct* translation of any text (different people will render the same information content in different, equally adequate,

ways), automatically evaluating the performance and utility of a SMT system is a very challenging task.

Evaluation metrics for machine translation (MT) can be categorized into two broad categories: (a) automated, and (b) manual. Compared with manual metrics, automated metrics are typically inexpensive because they require minimal investments of time and human resources. Examples of automated metrics are BLEU[7], METEOR[2], and the Translation Error Rate (TER)[1]. Manual metrics such as Likert scores[8] and Human Translation Error Rate (HTER)[1] require human involvement at various stages of the evaluation process and, while reflecting human judgment, are slow and resource intensive.

The automated metrics that have been proposed thus far, exhibit different levels of fidelity to human assessment of translation performance. The differences between the metrics are a result of differences in the attributes upon which the metrics are based. For examples, BLEU measures  $n$ -gram precision, TER measures word sequence similarity between hypothesis and reference, and METEOR is based on unigram matches with an emphasis on semantic equivalence.

Ideally, an automated metric should be correlated to the human judgments as well as provide useful information for human driven analysis. The alignments produced by TER are somewhat useful for both automated and human driven error analysis. However, the indifference of these alignments to the semantic equivalence and relationships between words makes it deficient in the sense that it underestimates translation performance. In addition, highlighting correct translations as errors may lower the productivity of a human performing error analysis using the TER alignments.

The primary theme of this paper is to incorporate a notion of semantic equivalence in TER, so that the measurements correlate more closely with human assessments and also improve the usability of the resulting alignments for both manual and automated error analysis. The new metric which we call the Semantic Translation Error Rate (STER) is described in detail along with experimental analysis to study the correlation of the metric with human judgments of the translation performance. The rest of the paper is as follows. In Section 2, we review TER

and METEOR evaluation metrics. Section 3 describes the STER alignment algorithm and other metrics derived from STER alignments. In Section 4, we present experimental results which indicate that STER is better correlated to human judgments than TER. In Section 5, we describe an error analysis tool, which uses STER alignments. Section 6 offers conclusions and directions for future work.

## 2. OVERVIEW OF EXISTING METRICS

In general, automated metrics are designed to improve the similarity of a translated sentence to one or more human generated reference translations. In this section, we describe the salient features of two automated metrics, TER and METEOR, which form the starting points for the proposed STER metric.

### 2.1. Translation Error Rate

The TER metric is based on an optimal alignment (in terms of edit distance) of words in the hypothesis sentence with words in the reference sentence. Every alignment is seen to consist of a set of edits which transform the hypothesis into the reference. Each such edit is associated with a cost. Consequently, the edit distance for an alignment is defined as the sum of edit costs over the set of edits in the alignment.

There are five basic types of edits that characterize a TER alignment: a) *match* - when a word in the hypothesis is exactly the same as a word in the reference, b) *insertion* - when a word in the hypothesis isn't aligned to any word in the reference, c) *deletion* - when a word in the reference isn't aligned to any word in the hypothesis, d) *substitution* - when a word in the hypothesis is aligned to another word in the reference, and e) *shift* - when a substring of consecutive words in the hypothesis is shifted from one to another position in the hypothesis.

Generating the TER alignment can be viewed as a two stage process. In the first stage, all possible substrings of the hypothesis which are also part of the reference are enumerated. These substrings are potential shifts that are considered for alignment. Given a substring, each position in the hypothesis which is less than a predefined maximum shift distance is a candidate position for the shift. In the next stage, we compute an alignment between the shifted hypothesis and the reference. The alignment is based on minimizing the edit cost using a dynamic programming algorithm. The total edit cost is the sum of the cost of aligning the shifted reference with the reference and the cost of shifting the substring in the hypothesis.

The hypothesis candidate which has the lowest total edit cost is used for computing the TER metric. The total edit distance for this candidate divided by the number of words in the reference is used as the TER score. Furthermore, the overall TER score over a corpus is the ratio of the accumulated total edit distance to the accumulated average

reference length. An example of a TER alignment and its associated TER edit distance is shown below. This example shows all the five basic types of edits used for alignment. The cost associated with each of the four edits (insertions, deletions, substitutions and shifts) is unity. The cost associated with a match is zero.

```
Best Ref: a b c d e f
Orig Hyp: d e b c g h
REF:   A  b c  d e  * F
HYP:   * @ b c [ d e ] G H
EVAL:  D      I S
SHFT:   1      1      1
TER Score: 66.67 ( 4.0/ 6.0)
```

In the example above, we see a match ('b c'), shift ('d e') from position in the hypothesis to position five, an insertion ('g'), a deletion ('a') and a substitution ('h' for 'f'). The total number of edits is four. Since the number of words in the reference is 6, the TER metric for this pair of hypothesis and reference is  $4/6=0.667$ .

### 2.2. METEOR

METEOR, like TER is also based on an alignment of the hypothesis sentence with the reference sentence. Unlike TER, however, it is not on based edit distance. Instead, the alignment is performed by first finding all possible unigram matches of words in the hypothesis to words in the reference. METEOR then finds subsets of these matches such that each word is aligned to at most one other word, and picks the largest such subset – that is, the one aligning the most words. If there is more than one such subset with the cardinality, METEOR chooses the one with the least crossing between word matches. METEOR then computes the precision, recall and fragmentation from this alignment.

Precision is defined as the ratio of the number of word matches to the number of words in the hypothesis. Recall is defined as the ratio of word matches to the total number of words in the reference. Fragmentation is defined as the number of chunks in the hypothesis to the total number of word matches. A chunk is defined as the longest substring in the hypothesis that has each word matched to another word in the reference. The METEOR uses a more liberal notion of a match between two words than TER. In particular, two words are said to match if:

1. They are identical words.
2. They are identical after stemming both of them using the Porter stemmer[9].
3. They are synonyms as defined in the WordNet database[10].

The matching of words between the hypothesis and the reference is done in multiple stages. In each stage, a match

```
REF: a big truck comes in the nighttime and delivers it
HYP: a big truck comes at night and delivered it
```

FIGURE 1: Example of METEOR alignment.

between words in the hypothesis and reference which haven't already been matched by in the stages is performed. Each stage uses a different notion of a match as enumerated above. Typically, a predefined set of stop words ("the", "a", etc) is removed before the alignment is performed. An example of a METEOR alignment is shown in Figure 1.

In the example above, 'delivered' is matched to 'delivers' since both words have the same Porter stem. 'Nighttime' is matched with 'night' since they are synonyms according to the WordNet. The total number of matches is 8. The precision is 8/9. The recall is 8/10. 'a big truck comes' and 'night and delivered it' are the only two chunks in the hypothesis. Hence, the fragmentation is 2/8.

Once the precision, recall and fragmentation are computed for a given METEOR alignment, the METEOR score is computed as:

$$10/Fmean = 1/Precision + 9/Recall \quad (1)$$

$$Penalty = 0.5 * Fragmentation^3 \quad (2)$$

$$METEOR\ Score = Fmean * (1 - Penalty)(3)$$

### 3. SEMANTIC TRANSLATION ERROR RATE

In addition to providing a score for evaluating translation performance, TER provides a detailed alignment of the hypothesis and the reference words. The alignment information which includes substitutions, insertions, deletions, and shifts, can be used for automated as well as human driven analysis of the MT performance. For instance, we could find all the *source:target* phrase pairs in the translation table which caused the most errors. Or we could find phrases that are most frequently inserted or deleted.

Although very detailed, the TER alignments are not ideal for human driven analysis in their standard form. This is due to the fact that TER alignment process treats all words as having the same value, and has no notion of word equivalence except simple identity.

A first level of control over the alignment process is introduced in standard TER by keeping the edit cost of a match equal to zero while the other edit costs equal to one. This encourages the alignment of a word occurring in the hypothesis and the reference to align with one another.

In the following we first describe the different constraints we use to improve the word alignment quality between reference and hypothesis sentences. These alignments are central for computing the proposed metrics like STER and other metrics derived from STER alignments.

#### 5.1. STER Alignment

In STER, we constrain the standard TER alignment to enforce multiple notions of word similarity. Specifically, we exercise control over the TER alignment process in two ways:

1. Expand word equivalence in TER by considering two words to be matches if:
  - a. They match after Porter stemming.
  - b. They are considered synonyms according to WordNet.
2. Prevent stop words from aligning to concept words.

The first level of control over the alignment process is motivated from the METEOR metric. The second constraint, a novel one, prevents non-concept words to be aligned to concept words. We enforce these controls on the alignments by making edit costs be context dependent. For instance, the substitution cost of aligning a concept and a stop word is infinity, whereas the substitution cost for two stop words or two concept words is . This assignment of costs has the effect of enforcing point 2 in the list above.

We use two sets of edit costs, one set for alignments involving only stop words and another set for alignments involving only concept words. The framework for aligning a stop word in the hypothesis with a stop word in the reference is as follows:

1. Ins+del is preferred over shift+sub.
2. Shift+match is preferred over ins+del.

Error	Cost	
Ins	$C$	
Del	$C$	
Sub	$2(C - \varepsilon)$	
Match	Exact	0
	Porter	$2C - s - \varepsilon$
	WordNet	$2C - s - \varepsilon$
Shift	$s$	

TABLE 2: Edit costs for alignments involving concept words.

The edit costs used to enforce above policies are summarized in Table 1.

The constraints for aligning concept words are the following:

1. For matches involving one concept words in the hypothesis with one concept words in the reference:
  - a. We make the distinction between exact matches and matches due to either Porter stemming or WordNet.
  - b. Shift+match (of any type) is preferred over ins+del.

Error	Cost
Ins	$c$
Del	$c$
Sub	$2c - s + \epsilon$
Match	0
Shift	$s$

TABLE 1: Edit costs for alignment of stop words.

2. For alignment involving two concept words in the hypothesis with two concept words in the reference, we prefer ins+del+shift+match (of any type) to sub+sub.

The edit costs used to enforce these policies are summarized in Table 2.

In the example in Figure 2, we show the TER and STER alignments for a hypothesis sentence and a reference sentence. The TER alignment, being indifferent to different words, considers shifts only if it sees a possibility of an exact match. In this example, it aligns ‘house’ with ‘smoke’

```
Best Ref: the house is smoking
Orig Hyp: smoke is came from the home
REF : the HOUSE is **** SMOKING
TER : [the] SMOKE is CAME FROM @ HOME
REF : **** the house is smoking
STER:@@ CAME FROM the home [is] [smoke ]
```

**FIGURE 2. Comparison of STER and TER alignments**

and ‘smoking’ with ‘home’). The STER alignment, by contrast, pairs ‘house’ with ‘home’ and ‘smoking’ with ‘smoke’. It realizes that to pair ‘smoke’ and ‘smoking’ requires ‘smoke’ to be shifted in the hypothesis and does so accordingly. The resulting alignment is much more useful for further analysis.

### 3.2 STER Metric

The STER alignments described earlier are used to compute the STER score. The STER score, like the TER score, is computed as the ratio of total edit cost over the average length of the references. However, unlike TER, the edit costs used to compute the STER score are different from those used to generate the STER alignments. In Table 3, we show the edit costs used for computing the STER scores. The shift cost is set to 1. As shown in Table 3, the errors associated with stop words are penalized less than errors associated with concept words.

### 3.3 SMET Metric

Performing human driven error analysis with METEOR is difficult because METEOR only computes unigram matches and lacks any notion of alignment errors between reference and hypothesis. In the following we extend METEOR to use STER alignments instead of using the alignment based on unigram matches. We refer to the new metric as the SMET.

In computing the SMET score, first, the correctly aligned words in the STER alignment are considered as word matches in the METEOR sense. These matched words are used to compute the precision, recall and fragmentation, which in turn are used to compute the SMET score in exactly the same way as one would compute the METEOR score. For utterances with more than one reference, the STER alignment for the reference which gives the least SMET score is used to compute the SMET score for the utterance.

Error	Stop	Concept
Ins	0.5	1
Del	0.5	1
Sub	0.5	1
Match	Exact	0
	Porter	0
	WordNet	0

**TABLE 3. Edit costs for computing the STER score.**

## 4. EXPERIMENTAL ANALYSIS

In this section, we evaluate the new metrics proposed in Section 3, in terms of their correlation with human judgments. The corpus used in all the experiments is the offline evaluation set used in the March 2006 TransTac evaluations (Stallard et al., 2006). This set consists of 1440 spoken Iraqi Arabic utterances spanning four different domains: general survey, intelligence, medical, and municipal services. Each utterance was transcribed in Arabic, and given four reference translations into English.

All SMT experiments provide two sets of scores. One set of scores evaluates translation performance on the reference transcriptions of the utterances (T2T). Another set of scores evaluates translation performance on speech recognition output as the source (S2T). The parameters used to set the edit costs are  $s=1$ ,  $c=1$ ,  $C=1$ , and  $\varepsilon=0.05$ .

In the first experiment, we compare the correlation of the various metrics with human judgment. A judge who was a native speaker of Arabic and fluent in English assigned 1-5 Likert scores to each translation output as a rating of their quality. In **Error! Reference source not found.** 4, we show the Pearson correlation coefficient,  $R$ , of Likert scores for every utterance against TER, METEOR, and STER scores respectively.

From Table 4, we see that the STER metric is better correlated to human judgment than TER. Since TER and STER metrics have different edit costs for edits involving stop words, we performed another experiment to ensure the improved correlation results from the quality of the alignment and not due to edit costs.

In Table 5, we provide correlation scores for STER and TER when the stop words have been removed from the hypothesis and reference. Since the edit costs for both the metrics are identical, the improvement in correlation reflects the improvement in the quality of the word alignments.

The TER scores in Table 4 and Table 5 show that the removal of stop words results in a slight improvement the correlation coefficient. However, removing stop words

Metric	R(T2T)	R(S2T)
TER	0.4450	0.5536
STER	0.4827	0.6077
METEOR	0.5342	0.6295

**TABLE 4: Comparison of Pearson correlation coefficient computed w.r.t Likert scores across different**

Metric	R(T2T)	R(S2T)
TER	0.4550	0.5661
STER	<b>0.4662</b>	<b>0.5875</b>

TABLE 5. Comparison of TER and STER after removing stop words.

reduces correlation with human judgment for STER. Given STER aligns stop words independently of concept words, the reduction in correlation shows that human judgment is sensitive to non-concept words too. We believe this is due to the fact that stop words positively correlate with human judgment when it can be aligned with other stop words (as in STER) but negatively correlates with human judgment when it can align with both stop words and concept words (as in TER).

Based on the results in Table 4, we can conclude that METEOR correlates best with human judgment. In Table 6, we compare METEOR to SMET, a metric derived from STER alignments as described in Section 3.3. The SMET instead of using the METOR alignments uses the STER alignment for computing the METEOR-equivalent score. As shown in Table 6, SMET has similar score as METEOR and is equally well correlated with human judgment. These results highlight the utility of SMET as a metric which is well correlated to human judgments and at the same time provides useful alignments for human driven analysis of the system output.

An interesting possible consequence of the results in Table 6 is that scores equivalent to METEOR can be computed with a simpler algorithm than METEOR itself uses. We performed a detailed comparison of the individual unigram alignments for concept words in STER and METEOR, and found that only 0.5% of them were different. Notably, SMET does not require METEOR’s multiple stages of unigram matching for different word equivalency measures (Wordnet, Porter stemming, etc). The STER alignment can also be viewed as enforcing the constraints that METEOR enforces on unigram alignments, namely making them one-to-one, and minimizing alignment crossings. Further study is needed to investigate this in more

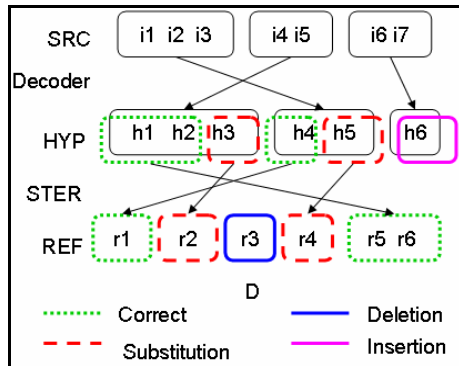


FIGURE 3. Sample STER alignments generated from the system output.

Metric	T2T		S2T	
	R	Score	R	Score
METEOR	<b>0.5342</b>	0.6540	<b>0.6295</b>	0.5430
SMET	0.5331	<b>0.6556</b>	0.6270	<b>0.5462</b>

Table 6: Comparison of METEOR and SMET metrics. detail.

## 5. ERROR ANALYSIS USING STER ALIGNMENTS

The STER alignments, as seen from the example in Section 3, show the potential benefit of using STER alignments over TER alignments for human-driven analysis of a SMT system. In this section, we describe a tool designed to aid in such a human-driven analysis. For each individual insertions and substitution error, the tool automatically finds the phrase pair instance which caused the error. It displays all instances of each phrase pair, linking them to the errors they cause in an easily navigable web-based interface.


### 5.1. Finding Phrase Pairs Causing Errors

To find which phrase pair caused an insertion or substitution error, we combine the STER alignment with the phrasal alignment produced by the MT decoder output (Most phrase based MT decoders produce such an alignment). As shown in Figure 3, the MT decoder associates each phrase (sub-string) in the source with a phrase in the hypothesis. For instance, the source phrase ‘i1 i2 i3’ is associated with the hypothesis phrase ‘h4 h5’. This source phrase and hypothesis (target) phrase constitute a phrase pair. STER alignments are then computed between the decoder-generated hypothesis and the reference translation. From these alignments, we can determine which target phrases have errors, and thus generate statistics leading to which phrase pairs “caused” the error.

Phrase pair errors could be due to: a) the phrase pair mapping being totally incorrect, b) the phrase pair mapping being incorrect in this context, i.e. being context-dependent, c) the phrase pair mapping’s target phrase being actually synonymous with the error alignment the system found, and finally d) the reference being incorrect. We have found that each phrase pair error can, in general, be categorized into one of these four broad categories of error classes. Once the phrase pair errors have been categorized, it is easier to apply corrective measures to improve overall system performance.

Since concept words are rich in information content, concept errors provide valuable insights into the manner in which information is being transmitted by the SMT system. Phrasal substitution errors analyze longer runs of STER alignments in which a phrase in the reference is substituted by another phrase in the hypothesis. For instance, a frequently occurring phrasal substitution is ‘i mean’ => ‘you know’. On analysis of utterances in which this kind phrasal substitution occurs, it is seen that these phrases are





Source Phrase	Target Phrase	TotErr	InstW	InstC	ProbW
يعنى	i mean	512	276	140	0.6635
ما	i do not	78	57	20	0.7403
اى	yes	55	55	361	0.1322
لاين	okay	52	52	98	0.3467

FIGURE 2. Display of phrase pair errors in the tool

frequently used in between sentences and functionally equivalent in this context. The tool thus aids in obtaining similar insights. In FIGURE 4, a screenshot of the tool displaying phrase pair errors is shown.

## 5.2. Description of the Analysis Tool

The analysis tool is designed to perform three primary functions:

1. Accumulate statistics for the various types of errors.
2. For each set of errors, show all instances in which these errors occurred.
3. Provide a simple user interface for navigation between these errors.

The analysis tool provides a mechanism for viewing all instances of an error in any of the error categories that are of interest to us for our analysis. For each instance, the tool displays, the reference (OREF), the source sentence (SRC), the hypothesis (OHYP), the STER alignment between the hypothesis and reference, the TER, STER and CTER scores. A screen shot of a section of the page displaying phrase pair errors for the phrase pair, “يعنى” => “I mean” are shown in Figure 4.

On placing the mouse over a word in the source or hypothesis, the corresponding phrase pair of which that word is a member is highlighted. As shown in Figure 5, the phrase pair “احياناً موجود” => “sometimes there is” is highlighted when the mouse is brought over the word ‘sometimes’ in OHYP. Substitution errors are highlighted in red and insertion errors are highlighted in blue (not shown in the figure).

Source:	يعنى
Target:	i mean
Errors:	Correct
APIQAR0609_126_f710i126_393-007787	
OREF:	well sometime fuel is available fuel
SRC:	يعنى احياناً موجود الوقود وغيره
OHYP:	i mean sometimes there is fuel
REF:	* WELL sometime FUEL ***** is AVAILABLE fuel
HYP:	I MEAN sometimes ***** there is ***** fuel
EVAL:	I S D D
SHFT:	
SCORE:	TER: 100.00 ( 6.0/ 6.0) STER: 75.00 ( 4.5/ 6.0)

FIGURE 5. Screenshot for a phrase pair error web page.

## 6. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed an improved version of the TER alignment algorithm. We call this improved version the STER alignment algorithm. On a test corpus of 1440 utterances, the TER score generated using STER alignment correlates better with human judgment than when the TER score is computed using TER alignment. We also proposed a variant of the METEOR metric, which instead of using the METEOR alignments uses STER alignments to compute the METEOR-equivalent score. This metric, named the SMET, preserves the strong correlation with human judgment property of the METEOR, with the additional benefit of generating alignments which are useful for human driven analysis.

Our use of METEOR’s word equivalence rules with the TER algorithm is a first step towards making TER more sensitive to semantic similarity in the hypothesis and the reference. A natural next step would be to incorporate phrasal equivalence in STER. We also propose to extend the studies in this paper with more human judges.

## 7. REFERENCES

- [1] Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J., “A Study of Translation Edit Rate with Targeted Human Annotation,” Proceedings of Association for Machine Translation in the Americas, 2006.
- [2] Banerjee, S. and Lavie, A., “METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments,” Proceedings of ACL, 2005.
- [3] Koehn, P., Och, F., and Marcu, D., “Statistical Phrase-Based Translation,” Proc. of the HLT and NAACL Conference, 2003.
- [4] Zhou, B., Dechelotte, D., and Gao, Y., “Two-way Speech-to-Speech Translation on Handheld Devices,” Int. Conf. of Spoken Language Processing (ICSLP), Korea, Oct. 2004.
- [5] Kathol, A., Precoda, K., Vergyri, D., Wang, W., and Riehemann, S., “Speech translation for low-resource languages: the case of Pashto,” Proc. INTERSPEECH-2005, 2273-2276, Lisbon, Portugal, 2006.
- [6] Stallard, D., Choi, F., Krstovski, K., Natarajan, P., Prasad, R., Saleem, S., “A Hybrid Phrase-based/Statistical Speech Translation System,” Proc. of Interspeech, Pittsburgh, PA, 2006.
- [7] Papineni, K., Roukos, S., Ward, T., and Zhu, W., “BLEU: A method for automatic evaluation of machine translation,” Proceedings of ACL, 311-318, 2002.
- [8] Barnett, V., “Sample Survey principles and methods,” Hodder Publisher, 1991.
- [9] Jones, K., and Willet, P., “Readings in Information Retrieval,” San Francisco: Morgan Kaufmann, 1997.
- [10] Fellbaum, C., “WordNet: An Electronic Lexical Database,” The MIT Press, Cambridge, MA, 1998.