Fluency, Adequacy, or HTER?
Exploring Different Human Judgments with a Tunable MT Metric

Matthew Snover[§], **Nitin Madnani**[§], **Bonnie J. Dorr**[§ †] **& Richard Schwartz**[† ‡]
[§]Laboratory for Computational Linguistics and Information Processing
[§]Institute for Advanced Computer Studies
[§]University of Maryland, College Park
[†]Human Language Technology Center of Excellence
[‡]BBN Technologies
{snover,nmadnani,bonnie}@umiacs.umd.edu schwartz@bbn.com

Abstract

Automatic Machine Translation (MT) evaluation metrics have traditionally been evaluated by the correlation of the scores they assign to MT output with human judgments of translation performance. Different types of human judgments, such as Fluency, Adequacy, and HTER, measure varying aspects of MT performance that can be captured by automatic MT metrics. We explore these differences through the use of a new tunable MT metric: TER-Plus, which extends the Translation Edit Rate evaluation metric with tunable parameters and the incorporation of morphology, synonymy and paraphrases. TER-Plus was shown to be one of the top metrics in NIST's Metrics MATR 2008 Challenge, having the highest average rank in terms of Pearson and Spearman correlation. Optimizing TER-Plus to different types of human judgments yields significantly improved correlations and meaningful changes in the weight of different types of edits, demonstrating significant differences between the types of human judgments.

1 Introduction

Since the introduction of the BLEU metric (Papineni et al., 2002), statistical MT systems have moved away from human evaluation of their performance and towards rapid evaluation using automatic metrics. These automatic metrics are themselves evaluated by their ability to generate scores for MT output that correlate well with human judgments of translation quality. Numerous methods of judging MT output by humans have been used, including *Fluency*, *Adequacy*, and, more recently, Human-mediated Translation Edit Rate (*HTER*) (Snover et al., 2006). Fluency measures whether a translation is fluent, regardless of the correct meaning, while Adequacy measures whether the translation conveys the correct meaning, even if the translation is not fully fluent. Fluency and Adequacy are frequently measured together on a discrete 5 or 7 point scale, with their average being used as a single score of translation quality. HTER is a more complex and semi-automatic measure in which humans do not score translations directly, but rather generate a new reference translation that is closer to the MT output but retains the fluency and meaning of the original reference. This new *targeted* reference is then used as the reference translation when scoring the MT output using Translation Edit Rate (TER) (Snover et al., 2006) or when used with other automatic metrics such as BLEU or METEOR (Banerjee and Lavie, 2005). One of the difficulties in the creation of targeted references is a further requirement that the annotator attempt to minimize the number of edits, as measured by TER, between the MT output and the targeted reference, creating the reference that is as close as possible to the MT output while still being adequate and fluent. In this way, only true errors in the MT output are counted. While HTER has been shown to be more consistent and finer grained than individual human annotators of Fluency and Adequacy, it is much more time consuming and taxing on human annotators than other types of human judgments, making it difficult and expensive to use. In addition, because HTER treats all edits equally, no distinction is made between serious errors (errors in names or missing subjects) and minor edits (such as a difference in verb agreement or a missing determinator).

Different types of translation errors vary in importance depending on the type of human judgment being used to evaluate the translation. For example, errors in tense might barely affect the adequacy of a translation but might cause the translation be scored as less fluent. On the other hand, deletion of content words might not lower the fluency of a translation but the adequacy would suffer. In this paper, we examine these differences by taking an automatic evaluation metric and tuning it to these these human judgments and examining the resulting differences in the parameterization of the metric. To study this we introduce a new evaluation metric, TER-Plus (TERp)[1] that improves over the existing Translation Edit Rate (TER) metric (Snover et al., 2006), incorporating morphology, synonymy and paraphrases, as well as tunable costs for different types of errors that allow for easy interpretation of the differences between human judgments.

Section 2 summarizes the TER metric and discusses how TERp improves on it. Correlation results with human judgments, including independent results from the 2008 NIST Metrics MATR evaluation, where TERp was consistently one of the top metrics, are presented in Section 3 to show the utility of TERp as an evaluation metric. The generation of paraphrases, as well as the effect of varying the source of paraphrases, is discussed in Section 4. Section 5 discusses the results of tuning TERp to Fluency, Adequacy and HTER, and how this affects the weights of various edit types.

2 TER and TERp

Both TER and TERp are automatic evaluation metrics for machine translation that score a translation, the *hypothesis*, of a foreign language text, the *source*, against a translation of the source text that was created by a human translator, called a *reference* translation. The set of possible correct translations is very large—possibly infinite—and any single reference translation is just a single point in that space. Usually multiple reference translations, typically 4, are provided to give broader sampling of the space of correct translations. Automatic MT evaluation metrics compare the hypothesis against this set of reference translations and assign a score to the similarity; higher scores are given to hypotheses that are more similar to the references.

In addition to assigning a score to a hypothesis, the TER metric also provides an alignment between the hypothesis and the reference, enabling it to be useful beyond general translation evaluation. While TER has been shown to correlate well with human judgments of translation quality, it has several flaws, including the use of only a single reference translation and the measuring of similarity only by exact word matches between the hypothesis and the reference. The handicap of using a single reference can be addressed by the construction of a lattice of reference translations. Such a technique has been used with TER to combine the output of multiple translation systems (Rosti et al., 2007). TERp does not utilize this methodology[2] and instead focuses on addressing the exact matching flaw of TER. A brief description of TER is presented in Section 2.1, followed by a discussion of how TERp differs from TER in Section 2.2.

2.1 TER

One of the first automatic metrics used to evaluate automatic machine translation (MT) systems was Word Error Rate (WER) (Niessen et al., 2000), which is the standard evaluation metric for Automatic Speech Recognition. WER is computed as the Levenshtein (Levenshtein, 1966) distance between the words of the system output and the words of the reference translation divided by the length of the reference translation. Unlike speech recognition, there are many correct translations for any given foreign sentence. These correct translations differ not only in their word choice but also in the order in which the words occur. WER is generally seen as inadequate for evaluation for machine translation as it fails to combine knowledge from multiple reference translations and also fails to model the reordering of words and phrases in translation.

TER addresses the latter failing of WER by allowing block movement of words, called *shifts*. within the hypothesis. Shifting a phrase has the same edit cost as inserting, deleting or substituting a word, regardless of the number of words being shifted. While a general solution to WER with block movement is NP-Complete (Lopresti

[1] Named after the nickname–"terp"–of the University of Maryland, College Park, mascot: the diamondback terrapin.

[2] The technique of combining references in this fashion has not been evaluated in terms of its benefit when correlating with human judgments. The authors hope to examine and incorporate such a technique in future versions of TERp.

and Tomkins, 1997), TER addresses this by using a greedy search to select the words to be shifted, as well as further constraints on the words to be shifted. These constraints are intended to simulate the way in which a human editor might choose the words to shift. For exact details on these constraints, see Snover et al. (2006). There are other automatic metrics that follow the general formulation as TER but address the complexity of shifting in different ways, such as the CDER evaluation metric (Leusch et al., 2006).

When TER is used with multiple references, it does not combine the references. Instead, it scores the hypothesis against each reference individually. The reference against which the hypothesis has the fewest number of edits is deemed the closet reference, and that number of edits is used as the numerator for calculating the TER score. For the denominator, TER uses the average number of words across all the references.

2.2 TER-Plus

TER-Plus (TERp) is an extension of TER that aligns words in the hypothesis and reference not only when they are exact matches but also when the words share a stem or are synonyms. In addition, it uses probabilistic phrasal substitutions to align phrases in the hypothesis and reference. These phrases are generated by considering possible paraphrases of the reference words. Matching using stems and synonyms (Banerjee and Lavie, 2005) and using paraphrases (Zhou et al., 2006; Kauchak and Barzilay, 2006) have previously been shown to be beneficial for automatic MT evaluation. Paraphrases have also been shown to be useful in expanding the number of references used for parameter tuning (Madnani et al., 2007; Madnani et al., 2008) although they are not used directly in this fashion within TERp. While all edit costs in TER are constant, all edit costs in TERp are optimized to maximize correlation with human judgments. This is because while a set of constant weights might prove adequate for the purpose of measuring translation quality—as evidenced by correlation with human judgments both for TER and HTER—they may not be ideal for maximizing correlation.

TERp uses all the edit operations of TER—Matches, Insertions, Deletions, Substitutions and Shifts—as well as three new edit operations: Stem Matches, Synonym Matches and Phrase Substitutions. TERp identifies words in the hypothesis and reference that share the same stem using the Porter stemming algorithm (Porter, 1980). Two words are determined to be synonyms if they share the same synonym set according to WordNet (Fellbaum, 1998). Sequences of words in the reference are considered to be paraphrases of a sequence of words in the hypothesis if that phrase pair occurs in the TERp phrase table. The TERp phrase table is discussed in more detail in Section 4.

With the exception of the phrase substitutions, the cost for all other edit operations is the same regardless of what the words in question are. That is, once the edit cost of an operation is determined via optimization, that operation costs the same no matter what words are under consideration. The cost of a phrase substitution, on the other hand, is a function of the probability of the paraphrase and the number of edits needed to align the two phrases according to TERp. In effect, the probability of the paraphrase is used to determine how much to discount the alignment of the two phrases. Specifically, the cost of a phrase substitution between the reference phrase, p_1 and the hypothesis phrase p_2 is:

$$\begin{aligned}\text{cost}(p_1, p_2) = & w_1 + \\ & \text{edit}(p_1, p_2) \times \\ & (w_2 \log(\Pr(p_1, p_2)) \\ & + w_3 \Pr(p_1, p_2) + w_4)\end{aligned}$$

where w_1, w_2, w_3, and w_4 are the 4 free parameters of the edit cost, $\text{edit}(p_1, p_2)$ is the edit cost according to TERp of aligning p_1 to p_2 (excluding phrase substitutions) and $\Pr(p_1, p_2)$ is the probability of paraphrasing p_1 as p_2, obtained from the TERp phrase table. The w parameters of the phrase substitution cost may be negative while still resulting in a positive phrase substitution cost, as w_2 is multiplied by the log probability, which is always a negative number. In practice this term will dominate the phrase substitution edit cost.

This edit cost for phrasal substitutions is, therefore, specified by four parameters, w_1, w_2, w_3 and w_4. Only paraphrases specified in the TERp phrase table are considered for phrase substitutions. In addition, the cost for a phrasal substitution is limited to values greater than or equal to 0, i.e., the substitution cost cannot be negative. In addition, the shifting constraints of TERp are also relaxed to allow shifting of paraphrases, stems, and synonyms.

In total TERp uses 11 parameters out of which four represent the cost of phrasal substitutions. The match cost is held fixed at 0, so that only the 10 other parameters can vary during optimization. All edit costs, except for the phrasal substitution parameters, are also restricted to be positive. A simple hill-climbing search is used to optimize the edit costs by maximizing the correlation of human judgments with the TERp score. These correlations are measured at the sentence, or *segment*, level. Although it was done for the experiments described in this paper, optimization could also be performed to maximize document level correlation – such an optimization would give decreased weight to shorter segments as compared to the segment level optimization.

3 Correlation Results

The optimization of the TERp edit costs, and comparisons against several standard automatic evaluation metrics, using human judgments of Adequacy is first described in Section 3.1. We then summarize, in Section 3.2, results of the NIST Metrics MATR workshop where TERp was evaluated as one of 39 automatic metrics using many test conditions and types of human judgments.

3.1 Optimization of Edit Costs and Correlation Results

As part of the 2008 NIST Metrics MATR workshop (Przybocki et al., 2008), a development subset of translations from eight Arabic-to-English MT systems submitted to NIST's MTEval 2006 was released that had been annotated for Adequacy. We divided this development set into an optimization set and a test set, which we then used to optimize the edit costs of TERp and compare it against other evaluation metrics. TERp was optimized to maximize the segment level Pearson correlation with adequacy on the optimization set. The edit costs determined by this optimization are shown in Table 1.

We can compare TERp with other metrics by comparing their Pearson and Spearman correlations with Adequacy, at the segment, document and system level. Document level Adequacy scores are determined by taking the length weighted average of the segment level scores. System level scores are determined by taking the weighted average of the document level scores in the same manner.

We compare TERp with BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and TER (Snover et al., 2006). The IBM version of BLEU was used in case insensitive mode with an ngram-size of 4 to calculate the BLEU scores. Case insensitivity was used with BLEU as it was found to have much higher correlation with Adequacy. In addition, we also examined BLEU using an ngram-size of 2 (labeled as *BLEU-2*), instead of the default ngram-size of 4, as it often has a higher correlation with human judgments. When using METEOR, the exact matching, porter stemming matching, and WordNet synonym matching modules were used. TER was also used in case insensitive mode.

We show the Pearson and Spearman correlation numbers of TERp and the other automatic metrics on the optimization set and the test set in Tables 2 and 3. Correlation numbers that are statistically indistinguishable from the highest correlation, using a 95% confidence interval, are shown in bold and numbers that are actually not statistically significant correlations are marked with a †. TERp has the highest Pearson correlation in all conditions, although not all differences are statistically significant. When examining the Spearman correlation, TERp has the highest correlation on the segment and system levels, but performs worse than METEOR on the document level Spearman correlatons.

3.2 NIST Metrics MATR 2008 Results

TERp was one of 39 automatic metrics evaluated in the 2008 NIST Metrics MATR Challenge. In order to evaluate the state of automatic MT evaluation, NIST tested metrics across a number of conditions across 8 test sets. These conditions included segment, document and system level correlations with human judgments of preference, fluency, adequacy and HTER. The test sets included translations from Arabic-to-English, Chinese-to-English, Farsi-to-English, Arabic-to-French, and English-to-French MT systems involved in NIST's MTEval 2008, the GALE (Olive, 2005) Phase 2 and Phrase 2.5 program, Transtac January and July 2007, and CESTA run 1 and run 2, covering multiple genres. The version of TERp submitted to this workshop was optimized as described in Section 3.1. The development data upon which TERp was optimized was not part of the test sets evaluated in the Challenge.

							Phrase Substitution			
Match	**Insert**	**Deletion**	**Subst.**	**Stem**	**Syn.**	**Shift**	w_1	w_2	w_3	w_4
0.0	0.26	1.43	1.56	0.0	0.0	0.56	-0.23	-0.15	-0.08	0.18

Table 1: Optimized TERp Edit Costs

	Optimization Set			**Test Set**			**Optimization+Test**		
Metric	Seg	Doc	Sys	Seg	Doc	Sys	Seg	Doc	Sys
BLEU	0.623	0.867	0.952	0.563	0.852	**0.948**	0.603	0.861	0.954
BLEU-2	0.661	**0.888**	0.946	0.591	**0.876**	**0.953**	0.637	0.883	0.952
METEOR	0.731	**0.894**	0.952	0.751	**0.904**	**0.957**	0.739	**0.898**	0.958
TER	-0.609	-0.864	-0.957	-0.607	-0.860	**-0.959**	-0.609	-0.863	-0.961
TERp	**-0.782**	**-0.912**	**-0.996**	**-0.787**	**-0.918**	**-0.985**	**-0.784**	**-0.914**	**-0.994**

Table 2: Optimization & Test Set Pearson Correlation Results

Due to the wealth of testing conditions, a simple overall view of the official MATR08 results released by NIST is difficult. To facilitate this analysis, we examined the average rank of each metric across all conditions, where the rank was determined by their Pearson and Spearman correlation with human judgments. To incorporate statistical significance, we calculated the 95% confidence interval for each correlation coefficient and found the highest and lowest rank from which the correlation coefficient was statistically indistinguishable, resulting in lower and upper bounds of the rank for each metric in each condition. The average lower bound, actual, and upper bound ranks (where a rank of 1 indicates the highest correlation) of the top metrics, as well as BLEU and TER, are shown in Table 4, sorted by the average upper bound Pearson correlation. Full descriptions of the other metrics[3], the evaluation results, and the test set composition are available from NIST (Przybocki et al., 2008).

This analysis shows that TERp was consistently one of the top metrics across test conditions and had the highest average rank both in terms of Pearson and Spearman correlations. While this analysis is not comprehensive, it does give a general idea of the performance of all metrics by synthesizing the results into a single table. There are striking differences between the Spearman and Pearson correlations for other metrics, in particular the CDER metric (Leusch et al., 2006) had the second highest rank in Spearman correlations (after TERp), but was the sixth ranked metric according to the Pearson correlation. In several cases, TERp was not the best metric (if a metric was the best in all conditions, its average rank would be 1), although it performed well on average. In particular, TERp did significantly better than the TER metric, indicating the benefit of the enhancements made to TER.

4 Paraphrases

TERp uses probabilistic phrasal substitutions to align phrases in the hypothesis with phrases in the reference. It does so by looking up—in a precomputed phrase table—paraphrases of phrases in the reference and using its associated edit cost as the cost of performing a match against the hypothesis. The paraphrases used in TERp were extracted using the pivot-based method as described in (Bannard and Callison-Burch, 2005) with several additional filtering mechanisms to increase the precision. The pivot-based method utilizes the inherent monolingual semantic knowledge from bilingual corpora: we first identify English-to-F phrasal correspondences, then map from English to English by following translation units from English to F and back. For example, if the two English phrases `e1` and `e2` both correspond to the same foreign phrase `f`, then they may be considered to be paraphrases of each other with the following probability:

$$p(e1|e2) \approx p(e1|f) * p(f|e2)$$

If there are several pivot phrases that link the two English phrases, then they are all used in comput-

[3] System description of metrics are also distributed by AMTA: `http://www.amtaweb.org/AMTA2008.html`

	Optimization Set			**Test Set**			**Optimization+Test**		
Metric	Seg	Doc	Sys	Seg	Doc	Sys	Seg	Doc	Sys
BLEU	0.635	**0.816**	0.714†	0.550	**0.740**	**0.690†**	0.606	0.794	**0.738†**
BLEU-2	0.643	**0.823**	0.786†	0.558	**0.747**	**0.690†**	0.614	0.799	**0.738†**
METEOR	0.729	0.886	**0.881**	**0.727**	**0.853**	**0.738†**	0.730	**0.876**	**0.922**
TER	-0.630	**-0.794**	-0.810†	-0.630	**-0.797**	**-0.667†**	-0.631	-0.801	**-0.786†**
TERp	**-0.760**	**-0.834**	**-0.976**	**-0.737**	**-0.818**	**-0.881**	**-0.754**	-0.834	**-0.929**

Table 3: MT06 Dev. Optimization & Test Set Spearman Correlation Results

Metric	**Average Rank by Pearson**	**Average Rank by Spearman**
TERp	$1.49 \ll 6.07 \ll 17.31$	$1.60 \ll 6.44 \ll 17.76$
METEOR v0.7	$1.82 \ll 7.64 \ll 18.70$	$1.73 \ll 8.21 \ll 19.33$
METEOR ranking	$2.39 \ll 9.45 \ll 19.91$	$2.18 \ll 10.18 \ll 19.67$
METEOR v0.6	$2.42 \ll 10.67 \ll 19.11$	$2.47 \ll 11.27 \ll 19.60$
EDPM	$2.45 \ll 8.21 \ll 20.97$	$2.79 \ll 7.61 \ll 20.52$
CDER	$2.93 \ll 8.53 \ll 19.67$	$1.69 \ll 8.00 \ll 18.80$
BleuSP	$3.67 \ll 9.93 \ll 21.40$	$3.16 \ll 8.29 \ll 20.80$
NIST-v11b	$3.82 \ll 11.13 \ll 21.96$	$4.64 \ll 12.29 \ll 23.38$
BLEU-1 (IBM)	$4.42 \ll 12.47 \ll 22.18$	$4.98 \ll 14.87 \ll 24.00$
BLEU-4 (IBM)	$6.93 \ll 15.40 \ll 24.69$	$6.98 \ll 14.38 \ll 25.11$
TER v0.7.25	$8.87 \ll 16.27 \ll 25.29$	$6.93 \ll 17.33 \ll 24.80$
BLEU-4 v12 (NIST)	$10.16 \ll 18.02 \ll 27.64$	$10.96 \ll 17.82 \ll 28.16$

Table 4: Average Metric Rank in NIST Metrics MATR 2008 Official Results

ing the probability:

$$p(e1|e2) \approx \sum_{f'} p(e1|f') * p(f'|e2)$$

The corpus used for extraction was an Arabic-English newswire bitext containing a million sentences. A few examples of the extracted paraphrase pairs that were actually used in a run of TERp on the Metrics MATR 2008 development set are shown below:

(*brief* → *short*)
(*controversy over* → *polemic about*)
(*by using power* → *by force*)
(*response* → *reaction*)

A discussion of paraphrase quality is presented in Section 4.1, followed by a brief analysis of the effect of varying the pivot corpus used by the automatic paraphrase generation upon the correlation performance of the TERp metric in Section 4.2.

4.1 Analysis of Paraphrase Quality

We analyzed the utility of the paraphrase probability and found that it was not always a very reliable estimate of the degree to which the pair was semantically related. For example, we looked at all paraphrase pairs that had probabilities greater than 0.9, a set that should ideally contain pairs that are paraphrastic to a large degree. In our analysis, we found the following five kinds of paraphrases in this set:

(a) **Lexical Paraphrases.** These paraphrase pairs are not phrasal paraphrases but instead differ in at most one word and may be considered as lexical paraphrases for all practical purposes. While these pairs may not be very valuable for TERp due to the obvious overlap with WordNet, they may help in increasing the coverage of the paraphrastic phenomena that TERp can handle. Here are some examples:

(*2500 polish troops* → *2500 polish soldiers*)
(*accounting firms* → *auditing firms*)
(*armed source* → *military source*)

(b) **Morphological Variants.** These phrasal pairs only differ in the morphological form for one of the words. As the examples show, any knowledge that these pairs may provide is already available to TERp via stemming.

(*50 ton* → *50 tons*)
(*caused clouds* → *causing clouds*)
(*syria deny* → *syria denies*)

(c) **Approximate Phrasal Paraphrases.** This set included pairs that only shared partial semantic content. Most paraphrases extracted by the pivot method are expected to be of this nature. These pairs are not directly beneficial to TERp since they cannot be substituted for each other in all contexts. However, the fact that they share at least some semantic content does suggest that they may not be entirely useless either. Examples include:

(*mutual proposal* → *suggest*)
(*them were exiled* → *them abroad*)
(*my parents* → *my father*)

(d) **Phrasal Paraphrases.** We did indeed find a large number of pairs in this set that were truly paraphrastic and proved the most useful for TERp. For example:

(*agence presse* → *news agency*)
(*army roadblock* → *military barrier*)
(*staff walked out* → *team withdrew*)

(e) **Noisy Co-occurrences.** There are also pairs that are completely unrelated and happen to be extracted as paraphrases based on the noise inherent in the pivoting process. These pairs are much smaller in number than the four sets described above and are not significantly detrimental to TERp since they are rarely chosen for phrasal substitution. Examples:

(*counterpart salam* → *peace*)
(*regulation dealing* → *list*)
(*recall one* → *deported*)

Given this distribution of the pivot-based paraphrases, we experimented with a variant of TERp that did not use the paraphrase probability at all but instead only used the actual edit distance between the two phrases to determine the final cost of a phrase substitution. The results for this experiment are shown in the second row of Table 5. We can see that this variant works as well as the full version of TERp that utilizes paraphrase probabilities. This confirms our intuition that the probability computed via the pivot-method is not a very useful predictor of semantic equivalence for use in TERp.

4.2 Varying Paraphrase Pivot Corpora

To determine the effect that the pivot language might have on the quality and utility of the extracted paraphrases in TERp, we used paraphrase pairsmade available by Callison-Burch (2008). These paraphrase pairs were extracted from Europarl data using each of 10 European languages (German, Italian, French etc.) as a pivot language separately and then combining the extracted paraphrase pairs. Callison-Burch (2008) also extracted and made available syntactically constrained paraphrase pairs from the same data that are more likely to be semantically related.

We used both sets of paraphrases in TERp as alternatives to the paraphrase pairs that we extracted from the Arabic newswire bitext. The results are shown in the last four rows of Table 5 and show that using a pivot language other than the one that the MT system is actually translating yields results that are almost as good. It also shows that the syntactic constraints imposed by Callison-Burch (2008) on the pivot-based paraphrase extraction process are useful and yield improved results over the baseline pivot-method. The results further support our claim that the pivot paraphrase probability is not a very useful indicator of semantic relatedness.

5 Varying Human Judgments

To evaluate the differences between human judgment types we first align the hypothesis to the references using a fixed set of edit costs, identical to the weights in Table 1, and then optimize the edit costs to maximize the correlation, without realigning. The separation of the edit costs used for alignment from those used for scoring allows us to remove the confusion of edit costs selected for alignment purposes from those selected to increase correlation.

For Adequacy and Fluency judgments, the MTEval 2002 human judgement set[4] was used. This set consists of the output of ten MT systems, 3 Arabic-to-English systems and 7 Chinese-

[4] Distributed to the authors by request from NIST.

Paraphrase Setup	Pearson			Spearman		
	Seg	Doc	Sys	Seg	Doc	Sys
Arabic pivot	**-0.787**	**-0.918**	**-0.985**	**-0.737**	**-0.818**	**-0.881**
Arabic pivot and no prob	**-0.787**	**-0.933**	**-0.986**	**-0.737**	**-0.841**	**-0.881**
Europarl pivot	**-0.775**	**-0.940**	**-0.983**	**-0.738**	**-0.865**	**-0.905**
Europarl pivot and no prob	**-0.775**	**-0.940**	**-0.983**	**-0.737**	**-0.860**	**-0.905**
Europarl pivot and syntactic constraints	**-0.781**	**-0.941**	**-0.985**	**-0.739**	**-0.859**	**-0.881**
Europarl pivot, syntactic constraints and no prob	**-0.779**	**-0.946**	**-0.985**	**-0.737**	**-0.866**	**-0.976**

Table 5: Results on the NIST MATR 2008 test set for several variations of paraphrase usage.

Human Judgment	Match	Insert	Deletion	Subst.	Stem	Syn.	Shift	Phrase Substitution			
								w_1	w_2	w_3	w_4
Alignment	0.0	0.26	1.43	1.56	0.0	0.0	0.56	-0.23	-0.15	-0.08	0.18
Adequacy	0.0	0.18	1.42	1.71	0.0	0.0	0.19	-0.38	-0.03	0.22	0.47
Fluency	0.0	0.12	1.37	1.81	0.0	0.0	0.43	-0.63	-0.07	0.12	0.46
HTER	0.0	0.84	0.76	1.55	0.90	0.75	1.07	-0.03	-0.17	-0.08	-0.09

Table 6: Optimized Edit Costs

to-English systems, consisting of a total, across all systems and both language pairs, of 7,452 segments across 900 documents. To evaluate HTER, the GALE (Olive, 2005) 2007 (Phase 2.0) HTER scores were used. This set consists of the output of 6 MT systems, 3 Arabic-to-English systems and 3 Chinese-to-English systems, although each of the systems in question is the product of system combination. The HTER data consisted of a total, across all systems and language pairs, of 16,267 segments across a total of 1,568 documents. Because HTER annotation is especially expensive and difficult, it is rarely performed, and the only source, to the authors' knowledge, of available HTER annotations is on GALE evaluation data for which no Fluency and Adequacy judgments have been made publicly available.

The edit costs learned for each of these human judgments, along with the alignment edit costs are shown in Table 6. While all three types of human judgements differ from the alignment costs used in alignment, the HTER edit costs differ most significantly. Unlike Adequacy and Fluency which have a low edit cost for insertions and a very high cost for deletions, HTER has a balanced cost for the two edit types. Inserted words are strongly penalized against in HTER, as opposed to in Adequacy and Fluency, where such errors are largely forgiven. Stem and synonym edits are also penalized against while these are considered equivalent to a match for both Adequacy and Fluency. This penalty against stem matches can be attributed to Fluency requirements in HTER that specifically penalize against incorrect morphology. The cost of shifts is also increased in HTER, strongly penalizing the movement of phrases within the hypothesis, while Adequacy and Fluency give a much lower cost to such errors. Some of the differences between HTER and both fluency and adequacy can be attributed to the different systems used. The MT systems evaluated with HTER are all highly performing state of the art systems, while the systems used for adequacy and fluency are older MT systems.

The differences between Adequacy and Fluency are smaller, but there are still significant differences. In particular, the cost of shifts is over twice as high for the fluency optimized system than the adequacy optimized system, indicating that the movement of phrases, as expected, is only slightly penalized when judging meaning, but can be much more harmful to the fluency of a translation. Fluency however favors paraphrases more strongly than the edit costs optimized for adequacy. This might indicate that paraphrases are used to generate a more fluent translation although at the potential loss of meaning.

6 Discussion

We introduced a new evaluation metric, TER-Plus, and showed that it is competitive with state-of-the-art evaluation metrics when its predictions are correlated with human judgments. The inclusion of stem, synonym and paraphrase edits allows TERp to overcome some of the weaknesses of the TER metric and better align hypothesized translations with reference translations. These new edit costs can then be optimized to allow better correlation with human judgments. In addition, we have examined the use of other paraphrasing techniques, and shown that the paraphrase probabilities estimated by the pivot-method may not be fully adequate for judgments of whether a paraphrase in a translation indicates a correct translation. This line of research holds promise as an external evaluation method of various paraphrasing methods.

However promising correlation results for an evaluation metric may be, the evaluation of the final output of an MT system is only a portion of the utility of an automatic translation metric. Optimization of the parameters of an MT system is now done using automatic metrics, primarily BLEU. It is likely that some features that make an evaluation metric good for evaluating the final output of a system would make it a poor metric for use in system tuning. In particular, a metric may have difficulty distinguishing between outputs of an MT system that been optimized for that same metric. BLEU, the metric most frequently used to optimize systems, might therefore perform poorly in evaluation tasks compared to recall oriented metrics such as METEOR and TERp (whose tuning in Table 1 indicates a preference towards recall). Future research into the use of TERp and other metrics as optimization metrics is needed to better understand these metrics and the interaction with parameter optimization.

Finally, we explored the difference between three types of human judgments that are often used to evaluate both MT systems and automatic metrics, by optimizing TERp to these human judgments and examining the resulting edit costs. While this can make no judgement as to the preference of one type of human judgment over another, it indicates differences between these human judgment types, and in particular the difference between HTER and Adequacy and Fluency. This exploration is limited by the the lack of a large amount of diverse data annotated for all human judgment types, as well as the small number of edit types used by TERp. The inclusion of additional more specific edit types could lead to a more detailed understanding of which translation phenomenon and translation errors are most emphasized or ignored by which types of human judgments.

Acknowledgments

This work was supported, in part, by BBN Technologies under the GALE Program, DARPA/IPTO Contract No. HR0011-06-C-0022 and in part by the Human Language Technology Center of Excellence.. TERp is available on the web for download at: http://www.umiacs.umd.edu/~snover/terp/.

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.*

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with Bilingual Parallel Corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 597–604, Ann Arbor, Michigan, June.

Chris Callison-Burch. 2008. Syntactic constraints on paraphrases extracted from parallel corpora. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 196–205, Honolulu, Hawaii, October. Association for Computational Linguistics.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press. http://www.cogsci.princeton.edu/˜wn [2000, September 7].

David Kauchak and Regina Barzilay. 2006. Paraphrasing for Automatic Evaluation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, pages 455–462.

Gregor Leusch, Nicola Ueffing, and Hermann Ney. 2006. CDER: Efficient MT Evaluation Using Block Movements. In *Proceedings of the 11th Conferenceof the European Chapter of the Association for Computational Linguistics (EACL 2006).*

V. I. Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Physics Doklady*, 10:707–710.

Daniel Lopresti and Andrew Tomkins. 1997. Block edit models for approximate string matching. *Theoretical Computer Science*, 181(1):159–179, July.

Nitin Madnani, Necip Fazil Ayan, Philip Resnik, and Bonnie J. Dorr. 2007. Using paraphrases for parameter tuning in statistical machine translation. In *Proceedings of the Workshop on Statistical Machine Translation*, Prague, Czech Republic, June. Association for Computational Linguistics.

Nitin Madnani, Philip Resnik, Bonnie J. Dorr, and Richard Schwartz. 2008. Are Multiple Reference Translations Necessary? Investigating the Value of Paraphrased Reference Translations in Parameter Optimization. In *Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas*, October.

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*.

Martin F. Porter. 1980. An algorithm for suffic stripping. *Program*, 14(3):130–137.

Mark Przybocki, Kay Peterson, and Sebastian Bronsart. 2008. Official results of the NIST 2008 ”Metrics for MAchine TRanslation” Challenge (MetricsMATR08). http://nist.gov/speech/tests/metricsmatr/2008/results/, October.

Antti-Veikko Rosti, Spyros Matsoukas, and Richard Schwartz. 2007. Improved word-level system combination for machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 312–319, Prague, Czech Republic, June. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of Association for Machine Translation in the Americas*.

Liang Zhou, Chon-Yew Lin, and Eduard Hovy. 2006. Re-evaluating Machine Translation Results with Paraphrase Support. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, pages 77–84.