Decomposability of Translation Metrics for Improved Evaluation and Efficient Algorithms

David Chiang and **Steve DeNeefe**
Information Sciences Institute
University of Southern California
4676 Admiralty Way, Suite 1001
Marina del Rey, CA 90292 USA
{chiang,sdeneefe}@isi.edu

Yee Seng Chan and **Hwee Tou Ng**
Department of Computer Science
National University of Singapore
Law Link
Singapore 117590
{chanys,nght}@comp.nus.edu.sg

Abstract

Bleu is the *de facto* standard for evaluation and development of statistical machine translation systems. We describe three real-world situations involving comparisons between different versions of the same systems where one can obtain improvements in Bleu scores that are questionable or even absurd. These situations arise because Bleu lacks the property of *decomposability*, a property which is also computationally convenient for various applications. We propose a very conservative modification to Bleu and a cross between Bleu and word error rate that address these issues while improving correlation with human judgments.

1 Introduction

Bleu (Papineni et al., 2002) was one of the first automatic evaluation metrics for machine translation (MT), and despite being challenged by a number of alternative metrics (Melamed et al., 2003; Banerjee and Lavie, 2005; Snover et al., 2006; Chan and Ng, 2008), it remains the standard in the statistical MT literature. Callison-Burch et al. (2006) have subjected Bleu to a searching criticism, with two real-world case studies of significant failures of correlation between Bleu and human adequacy/fluency judgments. Both cases involve comparisons between statistical MT systems and other translation methods (human post-editing and a rule-based MT system), and they recommend that the use of Bleu be restricted to comparisons between related systems or different versions of the same systems. In Bleu's defense, comparisons between different versions of the same system were exactly what Bleu was designed for.

However, we show that even in such situations, difficulties with Bleu can arise. We illustrate three ways that properties of Bleu can be exploited to yield improvements that are questionable or even absurd. All of these scenarios arose in actual practice and involve comparisons between different versions of the same statistical MT systems. They can be traced to the fact that Bleu is not *decomposable* at the sentence level: that is, it lacks the property that improving a sentence in a test set leads to an increase in overall score, and degrading a sentence leads to a decrease in the overall score. This property is not only intuitive, but also computationally convenient for various applications such as translation reranking and discriminative training. We propose a minimal modification to Bleu that reduces its nondecomposability, as well as a cross between Bleu and word error rate (WER) that is decomposable down to the subsentential level (in a sense to be made more precise below). Both metrics correct the observed problems and correlate with human judgments better than Bleu.

2 The Bleu metric

Let $g_k(w)$ be the multiset of all k-grams of a sentence w. We are given a sequence of candidate translations $\mathbf{c}$ to be scored against a set of sequences of reference translations, $\{\mathbf{r}^j\} = \mathbf{r}^1, \ldots, \mathbf{r}^R$:

$$\begin{aligned}\mathbf{c} &= c_1, c_2, c_3, \ldots, c_N \\ \mathbf{r}^1 &= r^1_1, r^1_2, r^1_3, \ldots, r^1_N \\ &\vdots \\ \mathbf{r}^R &= r^R_1, r^R_2, r^R_3, \ldots, r^R_N\end{aligned}$$

Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 610–619,
Honolulu, October 2008.

Then the BLEU score of $\mathbf{c}$ is defined to be

$$\text{BLEU}(\mathbf{c}, \{\mathbf{r}^j\}) = \prod_{k=1}^{4} pr_k(\mathbf{c}, \{\mathbf{r}^j\})^{\frac{1}{4}} \times bp(\mathbf{c}, \{\mathbf{r}^j\}) \quad (1)$$

where[1]

$$pr_k(\mathbf{c}, \{\mathbf{r}^j\}) = \frac{\sum_i \left| g_k(c_i) \cap \bigcup_j g_k(r_i^j) \right|}{\sum_i |g_k(c_i)|} \quad (2)$$

is the k-gram precision of $\mathbf{c}$ with respect to $\{\mathbf{r}^j\}$, and $bp(\mathbf{c}, \mathbf{r})$, known as the *brevity penalty*, is defined as follows. Let $\phi(x) = \exp(1 - 1/x)$. In the case of a single reference $\mathbf{r}$,

$$bp(\mathbf{c}, \mathbf{r}) = \phi\left(\min\left\{1, \frac{\sum_i |c_i|}{\sum_i |r_i|}\right\}\right) \quad (3)$$

In the multiple-reference case, the length $|r_i|$ is replaced with an *effective reference length*, which can be calculated in several ways.

- In the original definition (Papineni et al., 2002), it is the length of the reference sentence whose length is *closest* to the test sentence.
- In the NIST definition, it is the length of the *shortest* reference sentence.
- A third possibility would be to take the *average* length of the reference sentences.

The purpose of the brevity penalty is to prevent a system from generating very short but precise translations, and the definition of effective reference length impacts how strong the penalty is. The NIST definition is the most tolerant of short translations and becomes more tolerant with more reference sentences. The original definition is less tolerant but has the counterintuitive property that decreasing the length of a test sentence can eliminate the brevity penalty. Using the average reference length seems attractive but has the counterintuitive property that an exact match with one of the references may not get a 100% score. Throughout this paper we use the NIST definition, as it is currently the definition most used in the literature and in evaluations.

The brevity penalty can also be seen as a stand-in for recall. The fraction $\frac{\sum_i |c_i|}{\sum_i |r_i|}$ in the definition of the brevity penalty (3) indeed resembles a weak recall score in which every guessed item counts as a match. However, with recall, the per-sentence score $\frac{|c_i|}{|r_i|}$ would never exceed unity, but with the brevity penalty, it can. This means that if a system generates a long translation for one sentence, it can generate a short translation for another sentence without facing a penalty. This is a serious weakness in the BLEU metric, as we demonstrate below using three scenarios, encountered in actual practice.

3 Exploiting the BLEU metric

3.1 The sign test

We are aware of two methods that have been proposed for significance testing with BLEU: bootstrap resampling (Koehn, 2004b; Zhang et al., 2004) and the sign test (Collins et al., 2005). In bootstrap resampling, we sample with replacement from the test set to synthesize a large number of test sets, and then we compare the performance of two systems on those synthetic test sets to see whether one is better 95% (or 99%) of the time. But Collins et al. (2005) note that it is not clear whether the conditions required by bootstrap resampling are met in the case of BLEU, and recommend the sign test instead. Suppose we want to determine whether a set of outputs $\mathbf{c}$ from a test system is better or worse than a set of baseline outputs $\mathbf{b}$. The sign test requires a function $f(b_i, c_i)$ that indicates whether c_i is a better, worse, or same-quality translation relative to b_i. However, because BLEU is not defined on single sentences, Collins et al. use an approximation: for each i, form a composite set of outputs $\mathbf{b}' = \{b_1, \ldots, b_{i-1}, c_i, b_{i+1}, \ldots, b_N\}$, and compare the BLEU scores of $\mathbf{b}$ and $\mathbf{b}'$.

The goodness of this approximation depends on to what extent the comparison between $\mathbf{b}$ and $\mathbf{b}'$ is dependent only on b_i and c_i, and independent of the other sentences. However, BLEU scores are highly context-dependent: for example, if the sentences in $\mathbf{b}$ are on average ϵ words longer than the reference sentences, then c_i can be as short as $(N-1)\epsilon$ words

[1] We use the following definitions about multisets: if X is a multiset, let $\#_X(a)$ be the number of times a occurs in X. Then:

$$|X| \equiv \sum_a \#_X(a)$$
$$\#_{X \cap Y}(a) \equiv \min\{\#_X(a), \#_Y(a)\}$$
$$\#_{X \cup Y}(a) \equiv \max\{\#_X(a), \#_Y(a)\}$$

shorter than r_i without incurring the brevity penalty. Moreover, since the c_i are substituted in one at a time, we can do this for all of the c_i. Hence, **c** could have a disastrously low BLEU score (because of the brevity penalty) yet be found by the sign test to be significantly better than the baseline.

We have encountered this situation in practice: two versions of the same system with BLEU scores of 29.6 (length ratio 1.02) and 29.3 (length ratio 0.97), where the sign test finds the second system to be significantly better than the first (and the first system significantly better than the second). Clearly, in order for a significance test to be sensible, it should not contradict the observed scores, and should certainly not contradict itself. In the rest of this paper, except where indicated, all significance tests are performed using bootstrap resampling.

3.2 Genre-specific training

For several years, much statistical MT research has focused on translating newswire documents. One likely reason is that the DARPA TIDES program used newswire documents for evaluation for several years. But more recent evaluations have included other genres such as weblogs and conversation. The conventional wisdom has been that if one uses a single statistical translation system to translate text from several different genres, it may perform poorly, and it is better to use several systems optimized separately for each genre.

However, if our task is to translate documents from multiple known genres, but they are evaluated together, the BLEU metric allows us to use that fact to our advantage. To understand how, notice that our system has an optimal number of words that it should generate for the entire corpus: too few and it will be penalized by BLEU's brevity penalty, and too many increases the risk of additional non-matching k-grams. But these words can be distributed among the sentences (and genres) in any way we like. Instead of translating sentences from each genre with the best genre-specific systems possible, we can generate longer outputs for the genre we have more confidence in, while generating shorter outputs for the harder genre. This strategy will have mediocre performance on each individual genre (according to both intuition and BLEU), yet will receive a higher BLEU score on the combined test set than the combined systems optimized for each genre.

In fact, knowing which sentence is in which genre is not even always necessary. In one recent task, we translated documents from two different genres, without knowing the genre of any given sentence. The easier genre, newswire, also tended to have shorter reference sentences (relative to the source sentences) than the harder genre, weblogs. For example, in one dataset, the newswire reference sets had between 1.3 and 1.37 English words per Arabic word, but the weblog reference set had 1.52 English words per Arabic word. Thus, a system that is uniformly verbose across both genres will apportion more of its output to newswire than to weblogs, serendipitously leading to a higher score. This phenomenon has subsequently been observed by Och (2008) as well.

We trained three Arabic-English syntax-based statistical MT systems (Galley et al., 2004; Galley et al., 2006) using max-BLEU training (Och, 2003): one on a newswire development set, one on a weblog development set, and one on a combined development set containing documents from both genres. We then translated a new mixed-genre test set in two ways: (1) each document with its appropriate genre-specific system, and (2) all documents with the system trained on the combined (mixed-genre) development set. In Table 3, we report the results of both approaches on the entire test dataset as well as the portion of the test dataset in each genre, for both the genre-specific and mixed-genre trainings.

The genre-specific systems each outperform the mixed system on their own genre as expected, but when the same results are combined, the mixed system's output is a full BLEU point higher than the combination of the genre-specific systems. This is because the mixed system produces outputs that have about 1.35 English words per Arabic word on average: longer than the shortest newswire references, but shorter than the weblog references. The mixed system does worse on each genre but better on the combined test set, whereas, according to intuition, a system that does worse on the two subsets should also do worse on the combined test set.

3.3 Word deletion

A third way to take advantage of the BLEU metric is to permit an MT system to delete arbitrary words

in the input sentence. We can do this by introducing new phrases or rules into the system that match words in the input sentence but generate no output; to these rules we attach a feature whose weight is tuned during max-BLEU training. Such rules have been in use for some time but were only recently discussed by Li et al. (2008).

When we add word-deletion rules to our MT system, we find that the BLEU increases significantly (Table 6, line 2). Figure 1 shows some examples of deletion in Chinese-English translation. The first sentence has a proper name, 麦格赛赛/*maigesaisai* 'Magsaysay', which has been mistokenized into four tokens. The baseline system attempts to translate the first two phonetic characters as "wheat Georgia," whereas the other system simply deletes them. On the other hand, the second sentence shows how word deletion can sacrifice adequacy for the sake of fluency, and the third sentence shows that sometimes word deletion removes words that could have been translated well (as seen in the baseline translation).

Does BLEU reward word deletion fairly? We note two reasons why word deletion might be desirable. First, some function words should truly be deleted: for example, the Chinese particle 的/*de* and Chinese measure words often have no counterpart in English (Li et al., 2008). Second, even content word deletion might be helpful if it allows a more fluent translation to be assembled from the remnants. We observe that in the above experiment, word deletion caused the absolute number of k-gram matches, and not just k-gram precision, to increase for all $1 \leq k \leq 4$.

Human evaluation is needed to conclusively determine whether BLEU rewards deletion fairly. But to control for these potentially positive effects of deletion, we tested a *sentence-deletion* system, which is the same as the word-deletion system but constrained to delete *all* of the words in a sentence or none of them. This system (Table 6, line 3) deleted 8–10% of its input and yielded a BLEU score with no significant decrease ($p \geq 0.05$) from the baseline system's. Given that our model treats sentences independently, so that it cannot move information from one sentence to another, we claim that deletion of nearly 10% of the input is a grave translation deficiency, yet BLEU is insensitive to it.

What does this tell us about word deletion? While acknowledging that some word deletions can improve translation quality, we suggest in addition that because word deletion provides a way for the system to translate the test set selectively, a behavior which we have shown that BLEU is insensitive to, part of the score increase due to word deletion is likely an artifact of BLEU.

4 Other metrics

Are other metrics susceptible to the same problems as the BLEU metric? In this section we examine several other popular metrics for these problems, propose two of our own, and discuss some desirable characteristics for any new MT evaluation metric.

4.1 Previous metrics

We ran a suite of other metrics on the above problem cases to see whether they were affected. In none of these cases did we repeat minimum-error-rate training; all these systems were trained using max-BLEU. The metrics we tested were:

- METEOR (Banerjee and Lavie, 2005), version 0.6, using the exact, Porter-stemmer, and WordNet synonmy stages, and the optimized parameters $\alpha = 0.81$, $\beta = 0.83$, $\gamma = 0.28$ as reported in (Lavie and Agarwal, 2007).
- GTM (Melamed et al., 2003), version 1.4, with default settings, except $e = 1.2$, following the WMT 2007 shared task (Callison-Burch et al., 2007).
- MAXSIM (Chan and Ng, 2008), more specifically MAXSIM_n, which skips the dependency relations.

On the sign test (Table 2), all metrics found significant differences consistent with the difference in score between the two systems. The problem related to genre-specific training does not seem to affect the other metrics (see Table 4), but they still manifest the unintuitive result that genre-specific training is sometimes worse than mixed-genre training. Finally, all metrics but GTM disfavored both word deletion and sentence deletion (Table 7).

4.2 Strict brevity penalty

A very conservative way of modifying the BLEU metric to combat the effects described above is to im-

(a)
source 费 孝通 被 授予 麦 格 赛 赛 奖
reference fei xiaotong awarded magsaysay prize
baseline fei xiaotong was awarded the wheat georgia xaixai prize
delete fei xiaotong was awarded xaixai award

(b)
source 雨 花 石 正中 是 一 幅 十分 清晰 的 中华人民共和国 版图 的 图象 。
reference the center of the yuhua stone bears an image which very much resembles the territory of the people 's republic of china .
baseline rain huashi center is a big clear images of chinese territory .
delete rain is a clear picture of the people 's republic of china .

(c)
source 城建 成为 外商 投资 青海 新 热点
reference urban construction becomes new hotspot for foreign investment in qinghai
baseline urban construction become new hotspot for foreign investment qinghai
delete become new foreign investment hotspot

Figure 1: Examples of word deletion. Underlined Chinese words were deleted in the word-deletion system; underlined English words correspond to deleted Chinese words.

pose a stricter brevity penalty. In Section 2, we presented the brevity penalty as a stand-in for recall, but noted that unlike recall, the per-sentence score $\frac{|c_i|}{|r_i|}$ can exceed unity. This suggests the simple fix of clipping the per-sentence recall scores in a similar fashion to the clipping of precision scores:

$$bp(\mathbf{c}, \mathbf{r}) = \phi\left(\frac{\sum_i \min\{|c_i|, |r_i|\}}{\sum_i |r_i|}\right) \quad (4)$$

Then if a translation system produces overlong translations for some sentences, it cannot use those translations to license short translations for other sentences. Call this revised metric BLEU-SBP (for BLEU *with strict brevity penalty*).

We can test this revised definition on the problem cases described above. Table 2 shows that BLEU-SBP resolves the inconsistency observed between BLEU and the sign test, using the example test sets from Section 3.1 (no max-BLEU-SBP training was performed). Table 5 shows the new scores of the mixed-genre example from Section 3.2 after max-BLEU-SBP training. These results fall in line with intuition—tuning separately for each genre leads to slightly better scores in all cases. Finally, Table 8 shows the BLEU-SBP scores for the word-deletion example from Section 3.3, using both max-BLEU training and max-BLEU-SBP training. We see that BLEU-SBP reduces the benefit of word deletion to an insignificant level on the test set, and severely punishes sentence deletion. When we retrain using max-BLEU-SBP, the rate of word deletion is reduced and sentence deletion is all but eliminated, and there are no significant differences on the test set.

4.3 4-gram recognition rate

All of the problems we have examined—except for word deletion—are traceable to the fact that BLEU is not a sentence-level metric. Any metric which is defined as a weighted average of sentence-level scores, where the weights are system-independent, will be immune to these problems. Note that any metric involving micro-averaged precision (in which the sentence-level counts of matches and guesses are summed separately before forming their ratio) cannot have this property. Of the metrics surveyed in the WMT 2007 evaluation-evaluation (Callison-Burch et al., 2007), at least the following metrics have this property: WER (Nießen et al., 2000), TER (Snover et al., 2006), and ParaEval-Recall (Zhou et al., 2006).

Moreover, this evaluation concern dovetails with a frequent engineering concern, that sentence-level scores are useful at various points in the MT pipeline: for example, minimum Bayes risk decoding (Kumar and Byrne, 2004), selecting oracle translations for discriminative reranking (Liang

et al., 2006; Watanabe et al., 2007), and sentence-by-sentence comparisons of outputs during error analysis. A variation on BLEU is often used for these purposes, in which the k-gram precisions are "smoothed" by adding one to the numerator and denominator (Lin and Och, 2004); this addresses the problem of a zero k-gram match canceling out the entire score, but it does not address the problems illustrated above.

The remaining issue, word deletion, is more difficult to assess. It could be argued that part of the gain due to word deletion is caused by BLEU allowing a system to selectively translate those *parts* of a sentence on which higher precision can be obtained. It would be difficult indeed to argue that an evaluation metric, in order to be fair, must be decomposable into subsentential scores, and we make no such claim. However, there is again a dovetailing engineering concern which is quite legitimate. If one wants to select the minimum-Bayes-risk translation from a *lattice* (or shared forest) instead of an n-best list (Tromble et al., 2008), or to select an oracle translation from a lattice (Tillmann and Zhang, 2006; Dreyer et al., 2007; Leusch et al., 2008), or to perform discriminative training on all the examples contained in a lattice (Taskar et al., 2004), one would need a metric that can be calculated on the edges of the lattice.

Of the metrics surveyed in the WMT 2007 evaluation-evaluation, only one metric, to our knowledge, has this property: word error rate (Nießen et al., 2000). Here, we deal with the related *word recognition rate* (McCowan et al., 2005),

$$\begin{aligned} \text{WRR} &= 1 - \text{WER} \\ &= 1 - \min \frac{I + D + S}{|r|} \\ &= \max \frac{M - I}{|r|} \end{aligned} \tag{5}$$

where I is the number of insertions, D of deletions, S of substitutions, and $M = |r| - D - S$ the number of matches. The dynamic program for WRR can be formulated as a Viterbi search through a finite-state automaton: given a candidate sentence c and a reference sentence r, find the highest-scoring path matching c through the automaton with states $0, \ldots, |r|$, initial state 0, final state $|r|$, and the following transitions (a $\star$ matches any symbol):

For $0 \leq i < |r|$:

$i \xrightarrow{r_{i+1}:1} i+1$	match
$i \xrightarrow{\epsilon:0} i+1$	deletion
$i \xrightarrow{\star:0} i+1$	substitution

For $0 \leq i \leq |r|$:

$i \xrightarrow{\star:-1} i$	insertion

This automaton can be intersected with a typical stack-based phrase-based decoder lattice (Koehn, 2004a) or CKY-style shared forest (Chiang, 2007) in much the same way that a language model can, yielding a polynomial-time algorithm for extracting the best-scoring translation from a lattice or forest (Wagner, 1974). Intuitively, the reason for this is that WRR, like most metrics, implicitly constructs a word alignment between c and r and only counts matches between aligned words; but unlike other metrics, this alignment is constrained to be monotone.

We can combine WRR with the idea of k-gram matching in BLEU to yield a new metric, the *4-gram recognition rate*:

$$\text{4-GRR} = \max \frac{\sum_{k=1}^{4} M_k - \alpha I - \beta D}{\sum_{k=1}^{4} |g_k(r)|} \tag{6}$$

where M_k is the number of k-gram matches, α and β control the penalty for insertions and deletions, and g_k is as defined in Section 2. We presently set $\alpha = 1, \beta = 0$ by analogy with WRR, but explore other settings below. To calculate 4-GRR on a whole test set, we sum the numerators and denominators as in micro-averaged recall.

The 4-GRR can also be formulated as a finite-state automaton, with states $\{(i, m) \mid 0 \leq i \leq |r|, 0 \leq m \leq 3\}$, initial state $(0, 0)$, final states $(|r|, m)$, and the following transitions:

For $0 \leq i < |r|$, $0 \leq m \leq 3$:

$(i,m) \xrightarrow{r_{i+1}:m+1} (i+1, \min\{m+1, 3\})$	match
$(i,m) \xrightarrow{\epsilon:-\beta} (i+1, 0)$	deletion
$(i,m) \xrightarrow{\star:0} (i+1, 0)$	substitution

Metric	Adq	Flu	Rank	Con	Avg
Sem. role overlap	77.4	83.9	80.3	74.1	78.9
ParaEval recall	71.2	74.2	76.8	79.8	75.5
METEOR	70.1	71.9	74.5	66.9	70.9
BLEU	68.9	72.1	67.2	60.2	67.1
WER	51.0	54.2	34.5	52.4	48.0
BLEU-SBP	73.9	76.7	73.5	63.4	71.9
4-GRR	72.3	75.5	74.3	64.2	71.6

Table 1: Our new metrics correlate with human judgments better than BLEU (case-sensitive). Adq = Adequacy, Flu = Fluency, Con = Constituent, Avg = Average.

For $0 \leq i \leq |r|$, $0 \leq m \leq 3$:

$$(i, m) \xrightarrow{\star:-\alpha} (i, 0) \qquad \text{insertion}$$

Therefore 4-GRR can also be calculated efficiently on lattices or shared forests.

We did not attempt max-4-GRR training, but we evaluated the word-deletion test sets obtained by max-BLEU and max-BLEU-SBP training using 4-GRR. The results are shown in Table 7. In general, the results are very similar to BLEU-SBP except that 4-GRR sometimes scores word deletion slightly lower than baseline.

5 Correlation with human judgments

The shared task of the 2007 Workshop on Statistical Machine Translation (Callison-Burch et al., 2007) was conducted with several aims, one of which was to measure the correlation of several automatic MT evaluation metrics (including BLEU) against human judgments. The task included two datasets (one drawn from the Europarl corpus and the other from the News Commentary corpus) and across three language pairs (from German, Spanish, and French to English, and back). In our experiments, we focus on the tasks where the target language is English.

For human evaluations of the MT submissions, four different criteria were used:

- Adequacy: how much of the meaning expressed in the reference translation is also expressed in the hypothesis translation.
- Fluency: how well the translation reads in the target language.
- Rank: each translation is ranked from best to worst, relative to the other translations of the same sentence.
- Constituent: constituents are selected from source-side parse-trees, and human judges are asked to rank their translations.

We scored the workshop shared task submissions with BLEU-SBP and 4-GRR, then converted the raw scores to rankings and calculated the Spearman correlations with the human judgments. Table 1 shows the results along with BLEU and the three metrics that achieved higher correlations than BLEU: semantic role overlap (Giménez and Márquez, 2007), ParaEval recall (Zhou et al., 2006), and METEOR (Banerjee and Lavie, 2005). We find that both our proposed metrics correlate with human judgments better than BLEU does.

However, recall the parameters α and β in the definition of 4-GRR that control the penalty for inserted and deleted words. Experimenting with this parameter reveals that $\alpha = -0.9, \beta = 1$ yields a correlation of 78.9%. In other words, a metric that unboundedly rewards spuriously inserted words correlates better with human judgments than a metric that punishes them. We assume this is because there are not enough data points (systems) in the sample and ask that all these figures be taken with a grain of salt. As a general remark, it may be beneficial for human-correlation datasets to include a few straw-man systems that have very short or very long translations.

6 Conclusion

We have described three real-world scenarios involving comparisons between different versions of the same statistical MT systems where BLEU gives counterintuitive results. All these issues center around the issue of decomposability: the sign test fails because substituting translations one sentence at a time can improve the overall score yet substituting them all at once can decrease it; genre-specific training fails because improving the score of two halves of a test set can decrease the overall score; and sentence deletion is not harmful because generating empty translations for selected sentences does not necessarily decrease the overall score.

We proposed a minimal modification to BLEU, called BLEU-SBP, and showed that it ameliorates these

problems. We also proposed a metric, 4-GRR, that is decomposable at the sentence level and is therefore guaranteed to solve the sign test, genre-specific tuning, and sentence deletion problems; moreoever, it is decomposable at the subsentential level, which has potential implications for evaluating word deletion and promising applications to translation reranking and discriminative training.

Acknowledgments

Our thanks go to Daniel Marcu for suggesting modifying the BLEU brevity penalty, and to Jonathan May and Kevin Knight for their insightful comments. This research was supported in part by DARPA grant HR0011-06-C-0022 under BBN Technologies subcontract 9500008412.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proc. Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pages 65–72.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *Proc. EACL 2006*, pages 249–256.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta-) evaluation of machine translation. In *Proc. Second Workshop on Statistical Machine Translation*, pages 136–158.

Yee Seng Chan and Hwee Tou Ng. 2008. MAXSIM: A maximum similarity metric for machine translation evaluation. In *Proc. ACL-08: HLT*, pages 55–62.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *Proc. ACL 2005*, pages 531–540.

Markus Dreyer, Keith Hall, and Sanjeev Khudanpur. 2007. Comparing reordering constraints for SMT using efficient BLEU oracle computation. In *Proc. 2007 Workshop on Syntax and Structure in Statistical Translation*, pages 103–110.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In *Proc. HLT-NAACL 2004*, pages 273–280.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. 2006. Scalable inference and training of context-rich syntactic translation models. In *Proc. COLING-ACL 2006*, pages 961–968.

Jesús Giménez and Lluís Márquez. 2007. Linguistic features for automatic evaluation of heterogeneous MT systems. In *Proc. Second Workshop on Statistical Machine Translation*, pages 256–264.

Philipp Koehn. 2004a. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *Proc. AMTA 2004*, pages 115–124.

Philipp Koehn. 2004b. Statistical significance tests for machine translation evaluation. In *Proc. EMNLP 2004*, pages 388–395.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proc. HLT-NAACL 2004*, pages 169–176.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proc. Second Workshop on Statistical Machine Translation*, pages 228–231.

Gregor Leusch, Evgeny Matusov, and Hermann Ney. 2008. Complexity of finding the BLEU-optimal hypothesis in a confusion network. In *Proc. EMNLP 2008*. This volume.

Chi-Ho Li, Dongdong Zhang, Ming Zhou, and Hailei Zhang. 2008. An empirical study in source word deletion for phrase-based statistical machine translation. In *Proc. Third Workshop on Statistical Machine Translation*, pages 1–8.

Percy Liang, Alexandre Bouchard-Côté, Dan Klein, and Ben Taskar. 2006. An end-to-end discriminative approach to machine translation. In *Proc. COLING-ACL 2006*, pages 761–768.

Chin-Yew Lin and Franz Josef Och. 2004. ORANGE: a method for evaluating automatic evaluation metrics for machine translation. In *Proc. COLING 2004*, pages 501–507.

Iaian McCowan, Darren Moore, John Dines, Daniel Gatica-Perez, Mike Flynn, Pierre Wellner, and Hervé Bourlard. 2005. On the use of information retrieval measures for speech recognition evaluation. Research Report 04-73, IDIAP Research Institute.

I. Dan Melamed, Ryan Green, and Joseph P. Turian. 2003. Precision and recall of machine translation. In *Proc. HLT-NAACL 2003*, pages 61–63. Companion volume.

Sonia Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney. 2000. An evaluation tool for machine translation: Fast evaluation for MT research. In *Proc. LREC 2000*, pages 39–45.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. ACL 2003*, pages 160–167.

sys	BLEU	BLEU-SBP	METEOR	GTM	MAXSIM
1	29.6^{++}	28.0	53.1^{++}	45.5^{++}	40.7^{++}
2	29.3^{++}	27.8	52.2^{--}	44.8^{--}	39.6^{--}

Table 2: The sign test yields inconsistent results with BLEU but not with other metrics. Significances are relative to other system.

	mixed-genre		genre-specific		
test set	BLEU	length	BLEU	length	ΔBLEU
nw	47.9	1.14	51.1	0.98	+3.2
web	16.3	0.87	16.8	0.95	+0.5
nw+web	31.5	0.97	30.4	0.96	−1.1

Table 3: When performing two genre-specific max-BLEU trainings instead of a single mixed-genre training, we expect that improvements in the newswire (nw) and web subsets should result in a similar improvement in the combined test set (nw+web), but this is not the case. Key: length = length ratio relative to effective reference length.

test set	ΔMETEOR	ΔGTM	ΔMAXSIM
nw	−2.2	−1.3	−2.8
web	+0.8	+0.7	+1.3
nw+web	−0.7	−0.6	−0.2

Table 4: Contradictory effects of genre-specific training were not observed with other metrics.

	mixed-genre	genre-specific	
test set	BLEU-SBP	BLEU-SBP	ΔBLEU-SBP
nw	49.6	49.9	+0.3
web	15.3	15.7	+0.4
nw+web	29.3	29.5	+0.2

Table 5: When performing two genre-specific max-BLEU-SBP trainings instead of a single mixed-genre training, we find as expected that improvements in the newswire (nw) and web subsets correlate with a similar improvement in the combined test set (nw+web).

	dev		test	
deletion	del%	BLEU	del%	BLEU
none	0	37.7	0	39.3
word	8.4	38.6^{++}	7.7	40.1^{++}
sentence	10.2	37.7	8.6	39.1

Table 6: Use of word-deletion rules can improve the BLEU score, and use of sentence-deletion rules shows no significant degradation, even though they are used heavily. Significances are relative to baseline (no deletion); all other differences are not statistically significant.

	test			
deletion	METEOR	GTM	MAXSIM	4-GRR
none	59.2	41.0	45.6	18.7
word	57.9	41.9	45.0	18.6
sentence	57.2	41.3	44.0	17.1

Table 7: Word and sentence deletion are punished by most of the other metrics. All systems used max-BLEU training. Significance testing was not performed.

max-BLEU training

deletion	dev BLEU-SBP	test BLEU-SBP
none	35.3	36.9
word	35.8^{+}	37.1
sentence	33.0^{--}	34.5^{--}

max-BLEU-SBP training

	dev		test	
deletion	del%	BLEU-SBP	del%	BLEU-SBP
none	0	35.8	0	37.1
word	5.3	36.3^{+}	5.0	37.3
sentence	0.02	35.9	0	37.5

Table 8: BLEU-SBP severely punishes the max-BLEU-trained sentence-deletion system; when we perform max-BLEU-SBP training, word deletion occurs less frequently and sentence deletion is nearly unused. Significances are relative to baseline (no deletion); other differences are not statistically significant.

Key:	+, ++	significant improvement ($p < 0.05$ or $p < 0.01$, respectively)
	−, −−	significant degradation ($p < 0.05$ or $p < 0.01$, respectively)
	Δ*metric*	change in *metric* due to genre-specific training
	del%	percentage of words deleted

Franz Josef Och. 2008. The Google statistical machine translation system for the 2008 NIST MT Evaluation. Presentation at the NIST Open Machine Translation 2008 Evaluation Workshop.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. ACL 2002*, pages 311–318.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proc. AMTA 2006*, pages 223–231.

Ben Taskar, Carlos Guestrin, and Daphne Koller. 2004. Max-margin markov networks. In *Proc. NIPS 2003*.

Christoph Tillmann and Tong Zhang. 2006. A discriminative global training algorithm for statistical MT. In *Proc. COLING-ACL 2006*, pages 721–728.

Roy W. Tromble, Shankar Kumar, Franz Och, and Wolfgang Macherey. 2008. Lattice minimum Bayes-risk decoding for statistical machine translation. In *Proc. EMNLP 2008*. This volume.

Robert A. Wagner. 1974. Order-*n* correction for regular languages. *Communications of the ACM*, 17(5):265.

Taro Watanabe, Jun Suzuki, Hajime Tsukuda, and Hideki Isozaki. 2007. Online large-margin training for statistical machine translation. In *Proc. EMNLP 2007*, pages 764–773.

Ying Zhang, Stephan Vogel, and Alex Waibel. 2004. Interpreting BLEU/NIST scores: How much improvement do we need to have a better system? In *Proc. LREC 2004*, pages 2051–2054.

Liang Zhou, Chin-Yew Lin, and Eduard Hovy. 2006. Re-evaluating machine translation results with paraphase support. In *Proc. EMNLP 2006*, pages 77–84.