

USC/ISI Metric Description: BLEU-SBP and 4-GRR

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

1 Introduction

The USC/ISI submission to MetricsMATR included two metrics published elsewhere (Chiang et al., 2008), BLEU-SBP (BLEU with strict brevity penalty) and 4-GRR (4-gram recognition rate). These metrics were motivated by observations of improvements in BLEU scores that were questionable or even absurd, observations that center around 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. Our BLEU-SBP is a trivial modification to BLEU that reduces (though does not eliminate) its nondecomposability, and our 4-GRR is a cross between BLEU and word error rate (WER) that is decomposable down to the *subsential* level (in a sense to be made more precise below). Both metrics correlate with human judgments better than BLEU on the shared-task data from the 2007 Workshop on Machine Translation (Callison-Burch et al., 2007).

2 The BLEU metric

To provide a foundation for the notation used below, we provide a brief definition of BLEU. 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 trans-

lations, $\{\mathbf{r}^j\} = \mathbf{r}^1, \dots, \mathbf{r}^R$:

$$\begin{aligned}\mathbf{c} &= c_1, c_2, c_3, \dots, c_N \\ \mathbf{r}^1 &= r_1^1, r_2^1, r_3^1, \dots, r_N^1 \\ &\quad \vdots \\ \mathbf{r}^R &= r_1^R, r_2^R, r_3^R, \dots, r_N^R\end{aligned}$$

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

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

where¹

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

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 various ways; most commonly, the shortest reference is used. The purpose of the brevity penalty is to counterbalance the precision factors, preventing a system from generating very short but precise translations.

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

$$|X| \equiv \sum_a \#_X(a)$$

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

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

3 Strict brevity penalty

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 fact can be exploited in practice in various ways, described in another paper (Chiang et al., 2008).

A very conservative way of modifying the BLEU metric to combat this is to tighten the brevity penalty. We call this revised metric BLEU-SBP (for BLEU *with strict brevity penalty*). If the brevity penalty somewhat resembles recall, except that 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.

4 4-gram recognition rate

BLEU-SBP is designed to ameliorate undesirable effects of the nondecomposability of BLEU; any metric that is defined as a weighted average over sentence-level scores, where the weights are system-independent, would be guaranteed to be free from these problems. Translation error rate (Snover et al., 2006) is an example of such a metric.

But there are situations where it is desirable for a metric not only to be decomposable at the sentence level but at the subsentential level. For example, 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-competition, 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),

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

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 $0, \dots, |r|$, initial state 0, final state $|r|$, and the following transitions (a \star matches any symbol):

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

$$\begin{aligned} i &\xrightarrow{r_{i+1}:1} i + 1 && \text{match} \\ i &\xrightarrow{\epsilon:0} i + 1 && \text{deletion} \\ i &\xrightarrow{\star:0} i + 1 && \text{substitution} \end{aligned}$$

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

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

This automaton can be intersected with a typical stack-based phrase-based decoder lattice (Koehn, 2004) 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*:

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

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

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

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

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

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

5 Correlation with human judgments

We computed the correlations of BLEU-SBP and 4-GRR with the human judgements from the shared task of the 2007 Workshop on Statistical Machine Translation (Callison-Burch et al., 2007). 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.

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.

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.

Acknowledgments

Our thanks go to Daniel Marcu for suggesting modifying the BLEU brevity penalty and to Yee Seng Chan for help with calculating correlations with the WMT 2007 human judgements. 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, 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.
- David Chiang, Steve DeNeeffe, Yee Seng Chan, and Hwee Tou Ng. 2008. Decomposability of translation metrics for improved evaluation and efficient algorithms. In *Proc. EMNLP 2008*.

- David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.
- 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.
- 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. 2004. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *Proc. AMTA 2004*, pages 115–124.
- Gregor Leusch, Evgeny Matusov, and Hermann Ney. 2008. Complexity of finding the BLEU-optimal hypothesis in a confusion network. In *Proc. EMNLP 2008*.
- 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.
- 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.
- 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*.
- Robert A. Wagner. 1974. Order- n correction for regular languages. *Communications of the ACM*, 17(5):265.
- Liang Zhou, Chin-Yew Lin, and Eduard Hovy. 2006. Re-evaluating machine translation results with paraphrase support. In *Proc. EMNLP 2006*, pages 77–84.