A Systematic Comparison of Smoothing Techniques for Sentence-Level BLEU

Boxing Chen and Colin Cherry
National Research Council Canada
first.last@nrc-cnrc.gc.ca

Abstract

BLEU is the *de facto* standard machine translation (MT) evaluation metric. However, because BLEU computes a geometric mean of n-gram precisions, it often correlates poorly with human judgment on the sentence-level. Therefore, several smoothing techniques have been proposed. This paper systematically compares 7 smoothing techniques for sentence-level BLEU. Three of them are first proposed in this paper, and they correlate better with human judgments on the sentence-level than other smoothing techniques. Moreover, we also compare the performance of using the 7 smoothing techniques in statistical machine translation tuning.

1 Introduction

Since its invention, BLEU (Papineni et al., 2002) has been the most widely used metric for both machine translation (MT) evaluation and tuning. Many other metrics correlate better with human judgments of translation quality than BLEU, as shown in recent WMT Evaluation Task reports (Callison-Burch et al., 2011; Callison-Burch et al., 2012). However, BLEU remains the *de facto* standard evaluation and tuning metric. This is probably due to the following facts:

1. BLEU is language independent (except for word segmentation decisions).
2. BLEU can be computed quickly. This is important when choosing a tuning metric.
3. BLEU seems to be the best tuning metric from a quality point of view - i.e., models trained using BLEU obtain the highest scores from humans and even from other metrics (Cer et al., 2010).

One of the main criticisms of BLEU is that it has a poor correlation with human judgments on the sentence-level. Because it computes a geometric mean of n-gram precisions, if a higher order n-gram precision (eg. $n = 4$) of a sentence is 0, then the BLEU score of the entire sentence is 0, no matter how many 1-grams or 2-grams are matched. Therefore, several smoothing techniques for sentence-level BLEU have been proposed (Lin and Och, 2004; Gao and He, 2013).

In this paper, we systematically compare 7 smoothing techniques for sentence-level BLEU. Three of them are first proposed in this paper, and they correlate better with human judgments on the sentence-level than other smoothing techniques on the WMT metrics task. Moreover, we compare the performance of using the 7 smoothing techniques in statistical machine translation tuning on NIST Chinese-to-English and Arabic-to-English tasks. We show that when tuning optimizes the expected sum of these sentence-level metrics (as advocated by Cherry and Foster (2012) and Gao and He (2013) among others), all of these metrics perform similarly in terms of their ability to produce strong BLEU scores on a held-out test set.

2 BLEU and smoothing

2.1 BLEU

Suppose we have a translation T and its reference R, BLEU is computed with precision $P(N, T, R)$ and brevity penalty BP(T,R):

$$\text{BLEU}(N, T, R) = P(N, T, R) \times \text{BP}(T, R) \quad (1)$$

where $P(N, T, R)$ is the geometric mean of n-gram precisions:

$$P(N, T, R) = \left(\prod_{n=1}^{N} p_n \right)^{\frac{1}{N}} \quad (2)$$

Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 362–367,
Baltimore, Maryland USA, June 26–27, 2014.

and where:

$$p_n = \frac{m_n}{l_n} \quad (3)$$

m_n is the number of matched n-grams between translation T and its reference R, and l_n is the total number of n-grams in the translation T. BLEU's brevity penalty punishes the score if the translation length len(T) is shorter than the reference length len(R), using this equation:

$$\text{BP}(T, R) = \min\left(1.0, \exp\left(1 - \frac{\text{len}(R)}{\text{len}(T)}\right)\right) \quad (4)$$

2.2 Smoothing techniques

The original BLEU was designed for the document-level; as such, it required no smoothing, as some sentence would have at least one 4-gram match. We now describe 7 smoothing techniques that work better for sentence-level evaluation. Suppose we consider matching n-grams for $n = 1 \ldots N$ (typically, $N = 4$). Let m_n be the original match count, and m'_n be the modified n-gram match count.

Smoothing 1: if the number of matched n-grams is 0, we use a small positive value ϵ to replace the 0 for n ranging from 1 to N. The number ϵ is set empirically.

$$m'_n = \epsilon, \text{ if } m_n = 0. \quad (5)$$

Smoothing 2: this smoothing technique was proposed in (Lin and Och, 2004). It adds 1 to the matched n-gram count and the total n-gram count for n ranging from 2 to N.

$$m'_n = m_n + 1, \text{ for } n \text{ in } 2 \ldots N, \quad (6)$$

$$l'_n = l_n + 1, \text{ for } n \text{ in } 2 \ldots N. \quad (7)$$

Smoothing 3: this smoothing technique is implemented in the NIST official BLEU toolkit *mteval-v13a.pl*.[1] The algorithm is given below. It assigns a geometric sequence starting from $1/2$ to the n-grams with 0 matches.

1. *invcnt* $= 1$
2. for n in 1 to N
3. if $m_n = 0$
4. *invcnt* $=$ *invcnt* $\times 2$
5. $m'_n = 1/$*invcnt*
6. endif
7. endfor

[1] available at http://www.itl.nist.gov/iad/mig/tests/mt/2009/

Smoothing 4: this smoothing technique is novel to this paper. We modify Smoothing 3 to address the concern that shorter translations may have inflated precision values due to having smaller denominators; therefore, we give them proportionally smaller smoothed counts. Instead of scaling *invcnt* with a fixed value of 2, we replace line 4 in Smoothing 3's algorithm with Equation 8 below.

$$\textit{invcnt} = \textit{invcnt} \times \frac{K}{\ln(\text{len}(T))} \quad (8)$$

It assigns larger values to *invcnt* for shorter sentences, resulting in a smaller smoothed count. K is set empirically.

Smoothing 5: this smoothing technique is also novel to this paper. It is inspired by the intuition that matched counts for similar values of n should be similar. To a calculate the n-gram matched count, it averages the $n-1$, n and $n+1$ –gram matched counts. We define $m'_0 = m_1 + 1$, and calculate m'_n for $n > 0$ as follows:

$$m'_n = \frac{m'_{n-1} + m_n + m_{n+1}}{3} \quad (9)$$

Smoothing 6: this smoothing technique was proposed in (Gao and He, 2013). It interpolates the maximum likelihood estimate of the precision p_n with a prior estimate p^0_n. The prior is estimated by assuming that the ratio between p_n and p_{n-1} will be the same as that between p_{n-1} and p_{n-2}. Formally, the precisions of lower order n-grams, i.e., p_1 and p_2, are not smoothed, while the precisions of higher order n-grams, i.e. $n > 2$, are smoothed as follows:

$$p_n = \frac{m_n + \alpha p^0_n}{l_n + \alpha} \quad (10)$$

where α is set empirically, and p^0_n is computed as

$$p^0_n = p_{n-1} \times \frac{p_{n-1}}{p_{n-2}} \quad (11)$$

Smoothing 7: this novel smoothing technique combines smoothing 4 and smoothing 5. That is, we first compute a smoothed count for those 0 matched n-gram counts using Smoothing 4, and then take the average of three counts to set the final matched n-gram count as in Equation 9.

3 Experiments

We carried out two series of experiments. The 7 smoothing techniques were first compared in

set	year	lang.	#system	#seg. pair
dev	2008	xx-en	43	7,804
test1	2012	xx-en	49	34,909
test2	2013	xx-en	94	281,666
test3	2012	en-xx	54	47,875
test4	2013	en-xx	95	220,808

Table 1: Statistics of the WMT dev and test sets.

the metric task as evaluation metrics, then they were compared as metrics for tuning SMT systems to maximize the sum of expected sentence-level BLEU scores.

3.1 Evaluation task

We first compare the correlations with human judgment for the 7 smoothing techniques on WMT data; the development set (dev) is the WMT 2008 all-to-English data; the test sets are the WMT 2012 and WMT 2013 all-to-English, and English-to-all submissions. The languages "all" ("xx" in Table 1) include French, Spanish, German, Czech and Russian. Table 1 summarizes the dev/test set statistics.

Following WMT 2013's metric task (Macháček and Bojar, 2013), for the segment level, we use Kendall's rank correlation coefficient τ to measure the correlation with human judgment:

$$\tau = \frac{\#\text{concordant-pairs} - \#\text{discordant-pairs}}{\#\text{concordant-pairs} + \#\text{discordant-pairs}} \quad (12)$$

We extract all pairwise comparisons where one system's translation of a particular segment was judged to be better than the other system's translation, i.e., we removed all tied human judgments for a particular segment. If two translations for a particular segment are assigned the same BLEU score, then the #concordant-pairs and #discordant-pairs both get a half count. In this way, we can keep the number of total pairs consistent for all different smoothing techniques.

For the system-level, we used Spearman's rank correlation coefficient ρ and Pearson's correlation coefficient γ to measure the correlation of the metric with human judgments of translation. If we compute document-level BLEU as usual, all 7 smoothing techniques actually get the same result, as document-level BLEU does not need smoothing. We therefore compute the document-level BLEU as the weighted average of sentence-level BLEU, with the weights being the reference lengths:

$$\text{BLEU}_d = \frac{\sum_{i=1}^{D} \text{len}(R_i)\text{BLEU}_i}{\sum_{i=1}^{D} \text{len}(R_i)} \quad (13)$$

where BLEU_i is the BLEU score of sentence i, and D is the size of the document in sentences.

We first set the free parameters of each smoothing method by grid search to optimize the sentence-level score on the dev set. We set ϵ to 0.1 for Smoothing 1; $K = 5$ for Smoothing 4; $\alpha = 5$ for Smoothing 6.

Tables 2 and 3 report our results on the metrics task. We compared the 7 smoothing techniques described in Section 2.2 to a baseline with no smoothing (Smoothing 0). All scores match n-grams $n = 1$ to 4. Smoothing 3 is implemented in the standard official NIST evaluation toolkit (*mteval-v13a.pl*). Results are averaged across the 4 test sets.

	Into-English		
smooth	seg τ	sys γ	sys ρ
crp	–	0.720	0.887
0	0.165	0.759	0.887
1	0.224	0.760	0.887
2	0.226	0.757	0.887
3	0.224	0.760	0.887
4	0.228	0.763	0.887
5	0.234	0.765	0.887
6	0.230	0.754	0.887
7	**0.236**	**0.766**	0.887

Table 2: Correlations with human judgment on WMT data for Into-English task. Results are averaged on 4 test sets. "crp" is the origianl IBM corpus-level BLEU.

All smoothing techniques improved sentence-level correlations (τ) over no smoothing. Smoothing method 7 got the best sentence-level results on both the Into-English and Out-of-English tasks.

On the system-level, our weighted average of sentence-level BLEU scores (see Equation 13) achieved a better correlation with human judgement than the original IBM corpus-level BLEU. However, the choice of which smoothing technique is used in the average did not make a very big difference; in particular, the system-level rank correlation ρ did not change for 13 out of 14 cases. These methods help when comparing one hypothesis to another, but taken as a part of a larger average, all seven methods assign relatively low scores

	Out-of-English		
smooth	seg τ	sys γ	sys ρ
crp	–	0.712	0.744
0	0.119	0.715	0.744
1	0.178	0.722	**0.748**
2	0.180	0.725	0.744
3	0.178	0.724	0.744
4	0.181	0.727	0.744
5	0.184	0.731	0.744
6	0.182	0.725	0.744
7	**0.187**	**0.734**	0.744

Table 3: Correlations with human judgment on WMT data for Out-of-English task. Results are averaged on 4 test sets. "crp" is the origianl IBM corpus-level BLEU.

corpus	# segs	# en tok
	Chinese-English	
train	10.1M	283M
tune	1,506	161K
MT06	1,664	189K
MT08	1,357	164K
	Arabic-English	
train	1,512K	47.8M
tune	1,664	202K
MT08	1,360	205K
MT09	1,313	187K

Table 4: Statistics of the NIST Chinese-English and Arabic-English data.

to the cases that require smoothing, resulting in similar system-level rankings.

3.2 Tuning task

In this section, we explore the various BLEU smoothing methods in the context of SMT parameter tuning, which is used to set the decoder's linear model weights w. In particular, we use a tuning method that maximizes the sum of expected sentence-level BLEU scores, which has been shown to be a simple and effective method for tuning with large feature sets by both Cherry and Foster (2012) and Gao and He (2013), but which requires a smoothed sentence-level BLEU approximation. For a source sentence f_i, the probability of the k^{th} translation hypothesis e_i^k is its exponentiated and normalized model score:

$$P_w(e_i^k|f_i) = \frac{\exp(\textit{score}_w(e_i^k, f_i))}{\sum_{k'} \exp(\textit{score}_w(e_i^{k'}, f_i))}$$

where k' ranges over all hypotheses in a K-best list.[2] We then use stochastic gradient descent (SGD) to minimize:

$$\lambda||w||^2 - \sum_i \left[\text{len}(R_i) \times E_{P_w}\left(\text{BLEU}(e_i^k, f_i)\right)\right]$$

Note that we scale the expectation by reference length to place more emphasis on longer sentences. We set the regularization parameter λ, which determines the trade-off between a high expected BLEU and a small norm, to $\lambda = 10$.

Following Cherry and Foster (2012), we tune with a MERT-like batch architecture: fixing a set of K-best lists, optimizing, and then re-decoding the entire dev set to K-best and aggregating with previous lists to create a better K-best approximation. We repeat this outer loop 15 times.

We carried out experiments in two different settings, both involving data from NIST Open MT 2012.[3] The first setting is based on data from the Chinese-to-English constrained track, comprising about 283 million English running words. The second setting uses NIST 2012 Arabic-to-English data, but excludes the UN data. There are about 47.8 million English running words in these training data. The dev set (*tune*) for the Chinese-to-English task was taken from the NIST 2005 evaluation set, augmented with some web-genre material reserved from other NIST corpora. We test on the evaluation sets from NIST 2006 and 2008. For the Arabic-to-English task, we use the evaluation sets from NIST 2006, 2008, and 2009 as our dev set and two test sets, respectively. Table 4 summarizes the training, dev and test sets.

Experiments were carried out with an in-house, state-of-the-art phrase-based system. Each corpus was word-aligned using IBM2, HMM, and IBM4 models, and the phrase table was the union of phrase pairs extracted from these separate alignments, with a length limit of 7. The translation model (TM) was smoothed in both directions with Kneser-Ney smoothing (Chen et al., 2011). We use the hierarchical lexicalized reordering model (RM) (Galley and Manning, 2008), with a distortion limit of 7. Other features include lexical weighting in both directions, word count, a distance-based RM, a 4-gram LM trained on the target side of the parallel data, and a 6-gram En-

[2] We use $K = 100$ in our experiments.

[3] http://www.nist.gov/itl/iad/mig/openmt12.cfm

	Tune	std	MT06	std	MT08	std
0	27.6	0.1	35.6	0.1	29.0	0.2
1	27.6	0.0	35.7	0.1	29.1	0.1
2	27.5	0.1	35.8	0.1	29.1	0.1
3	27.6	0.1	35.8	0.1	29.1	0.1
4	27.6	0.1	35.7	0.2	29.1	0.2
5	27.6	0.1	35.5	0.1	28.9	0.2
6	27.5	0.1	35.7	0.1	29.0	0.2
7	27.6	0.1	35.6	0.1	29.0	0.1

Table 5: Chinese-to-English Results for the small feature set tuning task. Results are averaged across 5 replications; *std* is the standard deviation.

	Tune	std	MT08	std	MT09	std
0	46.9	0.1	46.5	0.1	49.1	0.1
1	46.9	0.0	46.4	0.1	49.1	0.1
2	46.9	0.0	46.4	0.1	49.0	0.1
3	47.0	0.0	46.5	0.1	49.2	0.1
4	47.0	0.0	46.5	0.1	49.2	0.1
5	46.9	0.0	46.4	0.1	49.1	0.1
6	47.0	0.0	46.4	0.1	49.1	0.1
7	47.0	0.0	46.4	0.1	49.0	0.1

Table 6: Arabic-to-English Results for the small feature set tuning task. Results are averaged across 5 replications; *std* is the standard deviation.

glish *Gigaword* LM.

We also conducted a set of experiments with a much larger feature set. This system used only GIZA++ for word alignment, increased the distortion limit from 7 to 9, and is trained on a high-quality subset of the parallel corpora used earlier. Most importantly, it includes the full set of sparse phrase-pair features used by both Hopkins and May (2011) and Cherry and Foster (2012), which results in nearly 7,000 features.

Our evaluation metric is the original IBM BLEU, which performs case-insensitive matching of n-grams up to $n = 4$. We perform random replications of parameter tuning, as suggested by Clark et al. (2011). Each replication uses a different random seed to determine the order in which SGD visits tuning sentences. We test for significance using the MultEval tool,[4] which uses a stratified approximate randomization test to account for multiple replications. We report results averaged across replications as well as standard deviations, which indicate optimizer stability.

Results for the small feature set are shown in Tables 5 and 6. All 7 smoothing techniques, as well as the no smoothing baseline, all yield very similar results on both Chinese and Arabic tasks. We did not find any two results to be significantly different. This is somewhat surprising, as other groups have suggested that choosing an appropriate BLEU approximation is very important. Instead, our experiments indicate that the selected BLEU smoothing method is not very important.

The large-feature experiments were only conducted with the most promising methods according to correlation with human judgments:

[4]available at https://github.com/jhclark/multeval

	Tune	std	MT06	std	MT08	std
mira	29.9	0.1	38.0	0.1	31.0	0.1
0	29.5	0.1	37.9	0.1	31.4	0.3
2	29.6	0.3	38.0	0.2	31.1	0.2
4	29.9	0.2	38.1	0.1	31.2	0.2
6	29.7	0.1	37.9	0.2	31.0	0.2
7	29.7	0.2	**38.0**	0.2	31.2	0.1

Table 7: Chinese-to-English Results for the large feature set tuning task. Results are averaged across 5 replications; *std* is the standard deviation. Significant improvements over the no-smoothing baseline ($p \leq 0.05$) are marked in bold.

0: No smoothing (baseline)
2: Add 1 smoothing (Lin and Och, 2004)
4: Length-scaled pseudo-counts (this paper)
6: Interpolation with a precision prior (Gao and He, 2013)
7: Combining Smoothing 4 with the match interpolation of Smoothing 5 (this paper)

The results of the large feature set experiments are shown in Table 7 for Chinese-to-English and Table 8 for Arabic-to-English. For a sanity check, we compared these results to tuning with our very stable Batch k-best MIRA implementation (Cherry and Foster, 2012), listed as *mira*, which shows that all of our expected BLEU tuners are behaving reasonably, if not better than expected.

Comparing the various smoothing methods in the large feature scenario, we are able to see significant improvements over the no-smoothing baseline. Notably, Method 7 achieves a significant improvement over the no-smoothing baseline in 3 out of 4 scenarios, more than any other method. Unfortunately, in the Chinese-English MT08 scenario, the no-smoothing baseline significantly out-

	Tune	std	MT08	std	MT09	std
mira	47.9	0.1	47.3	0.0	49.3	0.1
0	48.1	0.1	47.2	0.1	49.5	0.1
2	48.0	0.1	**47.4**	0.1	**49.7**	0.1
4	48.1	0.2	**47.4**	0.1	49.6	0.1
6	48.2	0.0	47.3	0.1	**49.7**	0.1
7	48.1	0.1	**47.3**	0.1	**49.7**	0.1

Table 8: Arabic-to-English Results for the large feature set tuning task. Results are averaged across 5 replications; *std* is the standard deviation. Significant improvements over the no-smoothing baseline ($p \leq 0.05$) are marked in bold.

performs all smoothed BLEU methods, making it difficult to draw any conclusions at all from these experiments. We had hoped to see at least a clear improvement in the tuning set, and one does see a nice progression as smoothing improves in the Chinese-to-English scenario, but no corresponding pattern emerges for Arabic-to-English.

4 Conclusions

In this paper, we compared seven smoothing techniques for sentence-level BLEU. Three of them are newly proposed in this paper. The new smoothing techniques got better sentence-level correlations with human judgment than other smoothing techniques. On the other hand, when we compare the techniques in the context of tuning, using a method that requires sentence-level BLEU approximations, they all have similar performance.

References

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64, Edinburgh, Scotland, July. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada, June. Association for Computational Linguistics.

Daniel Cer, Christopher D. Manning, and Daniel Jurafsky. 2010. The best lexical metric for phrase-based statistical mt system optimization. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 555–563, Los Angeles, California, June. Association for Computational Linguistics.

Boxing Chen, Roland Kuhn, George Foster, and Howard Johnson. 2011. Unpacking and transforming feature functions: New ways to smooth phrase tables. In *MT Summit 2011*.

Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *NAACL 2012*.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *ACL 2011*.

Michel Galley and C. D. Manning. 2008. A simple and effective hierarchical phrase reordering model. In *EMNLP 2008*, pages 848–856, Hawaii, October.

Jianfeng Gao and Xiaodong He. 2013. Training mrf-based phrase translation models using gradient ascent. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 450–459, Atlanta, Georgia, June. Association for Computational Linguistics.

Mark Hopkins and Jonathan May. 2011. Tuning as ranking. In *EMNLP 2011*.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 605–612, Barcelona, Spain, July.

Matouš Macháček and Ondřej Bojar. 2013. Results of the WMT13 metrics shared task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 45–51, Sofia, Bulgaria, August. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, July. ACL.