IJCLA VOL. 4, NO. 2, JUL-DEC 2013, PP. 29–44
RECEIVED 08/12/12 ACCEPTED 04/03/13 FINAL 05/03/13

BLEU Deconstructed: Designing a Better MT Evaluation Metric

XINGYI SONG, TREVOR COHN, AND LUCIA SPECIA

University of Sheffield, UK

ABSTRACT

BLEU is the de facto standard automatic evaluation metric in machine translation. While BLEU is undeniably useful, it has a number of limitations. Although it works well for large documents and multiple references, it is unreliable at the sentence or sub-sentence levels, and with a single reference. In this paper, we propose new variants of BLEU which address these limitations, resulting in a more flexible metric which is not only more reliable, but also allows for more accurate discriminative training. Our best metric has better correlation with human judgements than standard BLEU, despite using a simpler formulation. Moreover, these improvements carry over to a system tuned for our new metric.

1 INTRODUCTION

Automatic machine translation evaluation metrics provide a cheaper and faster way to evaluate translation quality than using human judgements. The standard evaluation metric in machine translation (MT) is BLEU [1], which is a simple, language independent metric that has been shown to correlate reasonably well with human judges. It is not only used in evaluation, but is also commonly used as a loss function for discriminative training [2, 3].

BLEU was designed for evaluating MT output against multiple references, and over large documents. However, evaluating translations at sentence level with single a reference is very common in MT research. Popular evaluation campaigns such as those organised by the WMT workshop only provide one reference for test and development corpora. In addition, many state-of-the-art discriminative training algorithms require sentence level evaluation metrics [4–6]. Often this means using a sentence-based approximation of BLEU, which can unduly bias the system and affect overall performance. BLEU performs less well when applied at the sentence level or sub-sentence level, and when using only one reference [7–10]. One reason is that in this setting BLEU has many zero or low counts for higher (tri-gram or higher) n-grams, and this has a disproportional effect on the overall score. Other problems with BLEU include its brevity penalty which has been shown to be a poor substitute for recall [10, 7], and the clipping of n-gram counts such that they do not exceed the count of each n-gram in the references, which complicates sub-sentential application.

Previous research has sought to address these problems. [11] suggest using arithmetic average instead of geometric mean. [12] shows that uni-gram and bi-gram precision contribute over 95 percent of overall precision, and they also state that adding higher order n-gram precision introduces a bias towards fluency over precision. This led us to question the effect of removing or substituting some components especially for sentence level evaluation. In this paper, we provide experimental analysis of each component in BLEU aiming to design better evaluation metrics for sentence level MT evaluation and MT system tuning with a single reference. On the WMT 2012 evaluation workshop [13], our variant of BLEU had better correlation with human judgements than any other for out-of-English document level evaluation.

The remainder of this paper is structured as follows: We will give brief a review of BLEU and its limitations in Section 2. In Section 3 we present experiments testing different variants of BLEU against human evaluation data, and also optimise the MT system parameters using these variant metrics. We found that our simplified BLEU improves over standard BLEU in terms of human judgements in both cases.

2 BLEU Review

The rationale behind BLEU [1] is that high quality translations will share many n-grams with human translations. BLEU is defined as

$$BLEU = BP \times \left(\prod_{n=1}^{4} p_n \right)^{\frac{1}{4}} \tag{1}$$

where p_n measures the modified n-gram precision between a document with candidate translations and a set of human authored reference documents, and the brevity penalty (BP) down-scales the score for outputs shorter than the reference. These are defined as

$$p_n = \frac{\sum\limits_{C \in \{Candidates\}} \sum\limits_{n-gram \in C} Count_{clip}(\text{n-gram})}{\sum\limits_{C' \in \{Candidates\}} \sum\limits_{n-gram' \in C'} Count(\text{n-gram'})} \tag{2}$$

$$BP = \begin{cases} 1 & \text{if } c > r \\ \exp(1 - \frac{r}{c}) & \text{if } c \leq r \end{cases}$$

where $Candidates$ are the set of sentences to be evaluated, c are their aggregate length and r is the length of the reference. $Count(\text{n-gram})$ counts the number of times the n-gram appears in the candidate sentence, and $Count_{clip}(\text{n-gram})$ is the same albeit clipped such that it does not exceed the number of times it appears in one of the reference sentences (which may be zero).

We now look at each of BLEU's component in detail.

N-gram precision BLEU is a precision-oriented evaluation method. Each precision component measures the proportion of predicted n-grams of a given n that appear in the reference translation. If multiple-references are used, the count of n-gram matching is based on the maximum number of matches against any of the references. For example in Table 1, candidate 1 matches 'It is a guide to action' and 'ensure that the military' with reference 1, matches 'which', 'always' and 'the commands of the party.' with reference 2. Therefore, the uni-gram precision will be 18/19, as only the word 'obeys' is not found in any of the references.

Clipping Clipping aims at penalising over-generated reference words in the candidate translation, such that repetitions of a word will not be rewarded. For example, candidate 2 in Table 2 is not a good translation, but still has very high uni-gram score (8/8). Clipping limits the count of a candidate n-gram to the maximum count of the n-gram in references. In this case the clipped uni-gram precision for candidate 2 will be 4/8: only one 'there' and one 'is' are treated as correct, and the repeats are counted as errors.

Brevity Penalty BLEU does not consider recall explicitly. In order to ensue reasonable coverage of reference, an alternative to recall is used: the

Table 1. Example of candidate and reference translations, adapted from [1].

Candidate 1:	It is a guide to action which ensures that the military always obeys the commands of the party.
Reference 1:	It is a guide to action that ensures that the military will forever heed Party commands.
Reference 2:	It is the guiding principle which guarantees the military forces always being under the command of the Party.
Reference 3:	It is the practical guide for the army always to heed the directions of the party.

Table 2. Without clipping and brevity penalty, candidates 1–3 will have same uni-gram score. Example taken from [1].

Reference:	there is a cat on the blue mat
Candidate 1:	there is
Candidate 2:	there there there is is is a cat
Candidate 3:	the cat is on the blue mat

brevity penalty. For example, candidate 1 in Table 2 has a uni-gram precision of 1. [1] state that in the multiple reference case, different words may be used in each reference, which makes it difficult to measure recall (we can never expect a good translation to include all these words). Therefore the Brevity Penalty is used instead to penalise short sentences. The intuition is that the candidate should have a similar length to the reference(s), and shorter candidates will be missing information.

2.1 *BLEU Limitations*

BLEU has become the standard evaluation metric since it was introduced in 2002, but it has several limitations. Firstly, in a short document or sentence, there is a high probability of obtaining zero tri-gram or 4-gram precision, which makes the overall BLEU score equal zero due to the use of geometric mean. Similarly, very low but non-zero counts disproportionately affect the score. A common method to ameliorate this effect is smoothing the counts [14–17], e.g. adding α both to the numerator and denominator of Equation 2. This avoids zero precision scores and zero overall BLEU score. However, different α values will affect the accuracy of the approximation of BLEU, and it is unclear what is a reasonable value to use. [11] suggest that using arithmetic average rather than geometric average, which avoids the problems of zero BLEU scores without resort to smoothing.

BLEU supports multiple references, which makes it hard to obtain an estimate of recall. Therefore, recall is replaced by the BP, but [10] state that BP is a poor substitute for recall. [10, 18, 7] include recall in their metrics and achieve better correlation with human judgements compared with BLEU.

[14] analysed BLEU at the sentence level with Pearson's correlation with human judgements over 1 to 9 grams. In order to apply BLEU for sentence level, they add one to the count of each n-gram. Results shows that BLEU with only uni-gram precision has the highest adequacy correlation (0.87), while adding higher order n-gram precision factors decreases the adequacy correlation and increases fluency. Overall they recommend using up to 5-gram precision to achieve the best balance. [12]'s experiments show that uni-gram and bi-gram precisions contribute over 95% of the overall precision. They also found that adding higher n-gram precision leads to a bias towards fluency over precision. However, it is not clear which of fluency or adequacy is more important, with recent evaluation favouring ranking judgements that implicitly consider both fluency and adequacy [13, 19–21].

These limitations affect the possible applications of BLEU, particularly for MT tuning. In tuning, the references are given, and we want the decoder to produce translations with high BLEU score. Current solutions rank translations in n-best lists [4, 22] or explicitly search for the maximum BLEU translation and use this for discriminative updates [23, 4, 24, 5]. In order to efficiently search for the maximum BLEU translation we need to be able to evaluate BLEU over partial sentences. However, the clipping and high order n-grams make this hard to apply BLEU during decoding. Thus the process relies on coarse approximations.

3 EXPERIMENTS

To address the above mentioned limitations, we analyse each component of BLEU and seek to address these shortcomings. Our prime motivation is to allow for better sentence level evaluation. In what follows, we test the effect of replacing and adjusting each component in BLEU – swapping the precision terms for recall, moving to an arithmetic mean, considering only smaller n-grams, dropping clipping of counts etc. In each instance, we test how each component contributes to BLEU in terms of correlation with human judgement data from previous translation evaluations. Hereinafter we use the following notation to denote each component in our metric:

P n-gram precision
R n-gram recall used in place of precision in Equation 2
F n-gram F-measure used in place of precision, balanced to weight recall 9 times higher than precision
A P/R/F terms are combined using an arithmetic mean
G P/R/F terms are combined using a geometric mean, as in Equation 1
B the brevity penalty term is included
1–4 include P/R/F terms for n-grams up to the given size
C clipping of counts used in P/R/F computation.

Note that our short-hand for standard BLEU is `PGBC4`, while a metric for clipped recall over unigrams and bigrams with no brevity penalty is labelled `RGC2`.

Our experiments are divided in two parts. In the first part we modify BLEU into several variants and compare the evaluation results of variants with human judgements, at both the sentence and document levels. In the second part, BLEU variants are used for parameter tuning, and the system output of each variant is evaluated by human judges. Our baseline BLEU is David Chiang's implementation, and add-1 smoothing is used for sentence level evaluation.

3.1 *Sentence Level evaluation*

For sentence level evaluation we follow the procedure from WMT11 [19], which uses Kendall's tau correlation (equation 3) to measure metrics' quality,

$$\tau = \frac{\text{num concordant pairs - num discordant pairs}}{\text{total pairs}} \tag{3}$$

where two ranked lists of translations according to humans and metrics are compared by counting the number of concordant and discordant relative ordering of pairs of translations, ignoring ties in either human or metric rankings.

We use τ to compare the sentence rankings produced by BLEU and all of our variants against human rankings. The human rankings were collected from WMT 09–11 [21, 20, 19], pooling together the data from

Table 3. Sentence level evaluation results showing τ between the metric-derived rankings and the human rankings. The label in the three columns denotes precision (P), recall (R) or F-measure (F), as used to combine n-gram matches according to each row's metric specification.

	P	**R**	**F**
GBC4	0.2116	0.1942	0.1905
GB4	0.2102	0.1913	0.1868
GC4	0.1879	0.2387	0.2054
ABC4	0.2288	0.2126	0.2076
AB4	0.2267	0.2411	0.2036
AC4	0.2055	**0.2462**	0.2178

Table 4. Results for sentence level evaluation without smoothing counts. Show are Kendall's tau correlations against human rankings. The ^{u} superscript denotes unsmoothed counts and ^{b} denotes smoothed brevity penalty.

	P	**R**	**F**
ABC4^u	0.2351	0.2209	0.2157
GBC4^{u,b}	0.2128	0.1935	0.1900
AC4^u	0.2176	**0.2462**	0.2178

Table 5. Sentence level evaluation results for metrics with various sized n-grams. Results are τ values and bolding shows the best score in each column.

	PGBC	**PGB**	**PABC**^{u}	**RAC**
1-4 grams	0.2116	0.2102	0.2351	0.2462
1-3 grams	0.2252	0.2230	**0.2375**	0.2491
1-2 grams	**0.2295**	**0.2278**	0.2353	0.2501
unigram	0.2284	0.2181	0.2293	**0.2726**

English-Spanish, English-French and English-German, in both translation directions. We selected only sentence pairs that were judged by at least two human annotators and where at least 60% of annotators agreed on their judgements. Our final test set contains 10,278 sentence pairs and has a Kappa of 0.8576.

Tables 3–5 show the results of sentence level evaluation with precision, recall and F-measure. Table 3 shows the results for BLEU variants with add-one smoothing. It is clear that the recall based metrics generally outperform those using precision and F-measure. The best performing metric is the RAC4 variant which combines 1-4-gram recall scores in arithmetic mean with no brevity penalty. This configuration has 3%

higher τ compared to standard BLEU (PGBC4), 0.2462 versus 0.2116. Overall, variants using the arithmetic mean perform better than those using the geometric mean. When clipping is removed, the performance uniformly decreases, but only slightly. More notable is the effect of the brevity penalty. When it is omitted, the performance drops heavily for precision metrics, but increases for recall and F-measure metrics. This is unsurprising as these metrics already disprefer short output. The F-measure based metrics are worse than both precision and recall variants when BP is included, but slightly outperform precision when BP is omitted.

A natural question is how important smoothing of counts is to sentence-level evaluation. Table 4 presents the correlation results for a number of variants.[1] Compared to the smoothed versions in Table 3, the unsmoothed arithmetic mean variants have better performance. We also found that smoothing the brevity penalty, $BP = \exp(1 - \frac{r+\alpha}{c+\alpha})$, using the same value of $\alpha = 1$ gave better performance compared unsmoothed BP.

All the results thus far have used $n = 4$-grams and smaller, following in the footsteps of BLEU. Our next experimental question is revisit this choice and test different values of n. Table 5 shows the sentence-level correlation results for various n-gram sizes, applied to some of the more successful metrics identified above. The most striking result is that RAC1 far exceeds all other metrics, and is one of the simplest in that it only uses unigrams. The arithmetic mean uniformly outperforms the geometric mean (including standard BLEU, PGBC4, in the top left corner). Also interesting is the pattern in the other columns, where the performance is relatively insensitive to the choice of n, with the maximum at $n = 2$ or $n = 3$. Overall the story is clear: large n-grams are not appropriate in this setting, and harm performance.

3.2 *Document Metric Evaluation*

In this section, the performance of BLEU variants will be tested at document level. We follow the WMT08 [25] document level procedure: we compare rankings based on evaluation metrics against human rankings using Spearman's rho correlation, defined as

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{4}$$

[1] Un-smoothed PBCG4 is not reported as it has very low Kendal's tau correlation.

where d_i measures the difference between the rank value assigned to sentence i by the system versus the human, and the n is number of sentences in the document.

Our test corpora are taken from all systems that were submitted as part of WMT08 for the *test2008* dataset.[2] We selected Spanish, French and German into and out-of-English for our tests. The final score is the average of the BLEU variant Spearman's rho correlation with human ranking in three tasks of *ranking*, *constituent* and *yes/no*. Please see [25] for a full exposition. In brief for the *ranking* and *constituent* the human judges were asked to rank a small set of candidate sentences in order of quality, focusing on a specific syntactic constituent for the latter case, and for *yes/no* they made a binary judgement of acceptability of the translation. Documents level rankings were constructed by counting how often each system outperformed the others, or the ratio of yes to no judgements. For the purpose of our experiments, we present average ρ values over the three different tasks.

Table 6. Document level correlation, measured using ρ.

	PGBC4	RGBC4	PABC4	PGB4	RAC4	PGBC2
es-en	0.7995	0.8111	0.7995	0.7995	**0.8135**	0.7925
fr-en	**0.9501**	0.9267	0.9443	**0.9501**	0.9414	0.9428
de-en	**0.5939**	0.5818	**0.5939**	**0.5939**	**0.5939**	**0.5939**
en-es	0.7757	0.7545	**0.8060**	0.7757	0.7545	**0.8060**
en-fr	**0.9388**	**0.9388**	**0.9388**	**0.9388**	**0.9388**	**0.9388**
en-de	0.7151	0.7151	**0.7212**	0.7151	0.7151	**0.7212**
avg.	0.7955	0.7881	**0.8006**	0.7955	0.7928	0.7992

Table 6 shows the results for document level evaluation, where we have selected promising metrics from the sentence level experiments. All the results are very close together, making it hard to draw concrete conclusions. However we do notice some contrary findings compared to the sentence level results. Most notably, the recall based metric with arithmetic mean (RAC4) performs worse than BLEU (PGBC4). Our earlier finding regarding clipping still holds here, i.e., that it has a negligible

[2] The reason for using a different dataset than for the earlier sentence level evaluation experiments is that only the WMT08 data provides the official document level human ranking results.

difference (compare PGBC4 and PGB4).[3] The overall best performing variant is PABC4, the arithmetic mean using 4-gram precision, brevity penalty and clipping. This metric is very similar to BLEU, simply swapping the geometric mean for the arithmetic mean.

3.3 *Discriminative Training*

Until now we have applied our metrics to human evaluation data, testing whether our variant metrics result in better ranking of MT outputs. However, it remains to be seen whether the metrics might also work effectively as a loss function for tuning a translation system. This is a better test of the metric, as it will encounter a much wider variety of outputs than present in MT evaluation data. For instance, empty sentences, overly long output, output from models with a negatively weighted language model, etc.

In this experiment we investigate parameter tuning of a statistical machine translation system. The system we used for this evaluation is Moses, a phrase-based decoder [3], which we tune using cmert-0.5, David Chiang's implementation of MERT [22]. We use the following (default) features:

- Translation probabilities, including forward & backward lexical probabilities, word count and phrase count.
- Lexicalised distortion model.
- A tri-gram language model, trained on the target side of the parallel corpus.

The training data for this experiment is Europarl-v6 German to English corpus, which is tuned on *dev-newstest2010* from WMT10 [20]. For the test, we use the *de-en* test set from WMT11 [19]. We tuned five different systems, each minimising a different loss function, and then used them to decode the test set. We randomly picked 50 unique output sentences from five systems' outputs for human ranking, asking our judges to rank them best to worst.

The human ranking used in this paper was done on Amazon Mechanical Turk using MAISE [26]. For each ranking judgement, source and reference sentences are provided, and the five candidate sentences are given in random order. The user then decides how to rank the five outputs. We

[3] In further experiments, not reported her, clipping also had little effect on performance for lower orders of n-gram.

repeat each ranking five times with different annotators. Pairwise annotation agreement in this paper is measured by the kappa coefficient [27],

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} \tag{5}$$

where P(A) is percentage of annotators agree with each other, and P(E) is the probability of agreement by chance, here $P(E) = \frac{1}{3}$. We also measure the self-agreement of each annotator, and discard all data from annotators with low self-agreement. We used 42 annotators and produced a total of 250 rankings, leaving 143 rankings after the self-agreement filtering. The kappa value for the filtered data was $K = 0.40$, with $P(A) = 0.61$.

The results of the human evaluation are shown in Table 7. The key result is that the most consistently good metric from our earlier experiments, PABC4, also did very well here. It outperformed BLEU (PGBC4) in 31% of cases and underperformed in 27% of the cases, for an overall 4% improvement. This improvement is significant with $p < 0.07$, as measured using the paired bootstrap resampling test [28]. Another interesting result is that PGBC2 and PGBC4 have the same performance, i.e., there is no effect of using larger n-grams. Surprisingly BLEU with clipping is only slightly better than the version without clipping (0.29 vs 0.28). We would expect that the unclipped system might systematically over-predict function words, as these will be less heavily penalised, and therefore produce inchorent output (so-called 'gaming' of the metric). However it appears that the larger n-grams stop this degenerate behaviour.

To further analyse the outputs of the various systems, Table 8 shows the various BLEU components of each tuned system's output. The BLEU (PGBC4) tuned system has the highest tri-gram and 4-gram precision and overall BLEU score, but the PGBC2 tuned system output has the highest uni-gram and bi-gram precision, as expected. The recall variant (RGBC4) has the longest sentence length, while omitting clipping had very little effect on sentence length. Overall the differences in BLEU scores are very small, which is surprising given the significant differences in human evaluation results.

4 CONCLUSIONS

In this paper we set out to simplify BLEU, revisiting each of the decisions made when it was originally proposed and evaluating the effect on large

Table 7. Results of human evaluations of de→en output from different systems, each trained to optimise a different metric. The values in each cell show how often the system in the column was judged to be better than the system in the row. To see whether a was better than b, one much look at the difference between cells (a, b) and (b, a), i.e., its reflection. Bold values indicate that the system in the column outperformed the system in the row.

	PABC4	**PGBC4**	**PGBC2**	**PGB4**	**RGBC4**
PABC4	–	0.27	0.26	0.25	0.29
PGBC4	**0.31**	–	**0.29**	0.28	0.28
PGBC2	**0.33**	**0.29**	–	0.21	0.26
PGB4	**0.28**	**0.29**	**0.23**	–	0.24
RGBC4	**0.33**	**0.32**	**0.29**	**0.28**	–

Table 8. A comparison of the BLEU components for the de→en translations produced by MT systems optimising different evaluation metrics, shown as columns. The rows P1-4 denote 1 to 4-gram precision, and LR is the ratio of lengths between system output and the reference, as used in the brevity penalty.

	PABC4	**PGBC4**	**PGBC2**	**PGB4**	**RGBC4**
P1	0.4684	0.4761	**0.4763**	0.4711	0.4742
P2	0.1659	0.1691	**0.1705**	0.1676	0.1683
P3	0.0811	**0.0824**	0.0807	0.0816	0.0785
P4	0.0369	**0.0388**	0.0367	0.0380	0.0360
LR	1.0043	0.9985	0.9906	0.9985	**1.0072**
BLEU	0.1236	**0.1265**	0.1234	0.1250	0.1226

collections of human annotated MT evaluation data. Our objectives were to allow BLEU to be applied accurately at the sentence level, and pave the way for simpler sub-sentential usage in the future. The experiments turned up a number of interesting results: bi-grams are at least as effective as 4-grams, clipping makes little difference, and recall based metrics often outperform precision based metrics. The most consistent finding was that the arithmetic mean outperforms the geometric mean. Together the findings about clipping and the arithmetic mean augur well for discriminative training, as these together greatly simplify the decomposition of the metric to partial sentences, as required during decoding to find the best scoring hypothesis. Some of the improvements evaporated when moving from human evaluation data to the discriminative training setting, where the models were tuned to optimise each metric. This suggests that human evaluation data in WMT is biased towards similar models (those trained for BLEU), and that it is inherently dangerous to design a metric solely from WMT evaluation data without also evaluating on additional, more varied, data.

Our overall results show an improvement of sentence level correlation to $\tau = 0.2726$ from $\tau = 0.2116$ for sentence-level BLEU, and for a much simpler metric. We therefore recommend that MT researchers consider using one of our simplified metrics in their experiments where single-reference per-sentence application is required. Our intension is to develop a discriminative algorithm to optimise the simplified metric, which will allow for more accurate optimisation while also resulting in higher quality translations.

REFERENCES

1. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. (2002) 311–318
2. Li, Z., Callison-Burch, C., Dyer, C., Ganitkevitch, J., Khudanpur, S., Schwartz, L., Thornton, W.N.G., Weese, J., Zaidan, O.F.: Joshua: an open source toolkit for parsing-based machine translation. In: Proceedings of the Fourth Workshop on Statistical Machine Translation. StatMT '09, Stroudsburg, PA, USA, Association for Computational Linguistics (2009) 135–139
3. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open source toolkit for statistical machine translation. In: Annual Meeting of the Association for Computational Linguistics. (2007)
4. Liang, P., Bouchard-Côté, A., Klein, D., Taskar, B.: An end-to-end discriminative approach to machine translation. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. ACL-44, Stroudsburg, PA, USA, Association for Computational Linguistics (2006) 761–768
5. Chiang, D., Marton, Y., Resnik, P.: Online large-margin training of syntactic and structural translation features. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. EMNLP '08, Stroudsburg, PA, USA, Association for Computational Linguistics (2008) 224–233
6. Hopkins, M., May, J.: Tuning as ranking. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Edinburgh, Scotland, UK., Association for Computational Linguistics (July 2011) 1352–1362
7. Song, X., Cohn, T.: Regression and ranking based optimisation for sentence level mt evaluation. In: Proceedings of the Sixth Workshop on Statistical Machine Translation. (2011) 123–129

8. Chiang, D., DeNeefe, S., Chan, Y.S., Ng, H.T.: Decomposability of translation metrics for improved evaluation and efficient algorithms. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. EMNLP '08 (2008) 610–619
9. Callison-Burch, C., Osborne, M., Koehn, P.: Re-evaluating the role of bleu in machine translation research. In: In EACL. (2006) 249–256
10. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. Proceedings of the ACL-05 Workshop (2005)
11. Doddington, G.: Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: Proceedings of the second international conference on Human Language Technology Research. HLT '02 (2002) 138–145
12. Zhang, Y., Vogel, S., Waibel, A.: Interpreting bleu/nist scores: How much improvement do we need to have a better system. In: In Proceedings of Proceedings of Language Resources and Evaluation (LREC-2004. (2004) 2051–2054
13. Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., Specia, L.: Findings of the 2012 workshop on statistical machine translation. In: Proceedings of the Seventh Workshop on Statistical Machine Translation. (2012) 10–51
14. Lin, C.Y., Och, F.J.: Orange: a method for evaluating automatic evaluation metrics for machine translation. In: Proceedings of the 20th international conference on Computational Linguistics. COLING '04 (2004)
15. Owczarzak, K., Groves, D., Van Genabith, J., Way, A.: Contextual bitext-derived paraphrases in automatic mt evaluation. In: Proceedings of the Workshop on Statistical Machine Translation. StatMT 06 (2006) 86–93
16. Koehn, P., Arun, A., Hoang, H.: Towards better machine translation quality for the german–english language pairs. In: Proceedings of the Third Workshop on Statistical Machine Translation. StatMT '08 (2008) 139–142
17. Hanneman, G., Huber, E., Agarwal, A., Ambati, V., Parlikar, A., Peterson, E., Lavie, A.: Statistical transfer systems for french–english and german–english machine translation. In: Proceedings of the Third Workshop on Statistical Machine Translation. StatMT '08 (2008) 163–166
18. Liu, C., Dahlmeier, D., Ng, H.T.: Tesla: Translation evaluation of sentences with linear-programming-based analysis. In: Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR. (2010) 354–359
19. Callison-Burch, C., Koehn, P., Monz, C., Zaidan, O.: Findings of the 2011 workshop on statistical machine translation. In: Proceedings of the Sixth Workshop on Statistical Machine Translation. (2011) 22–64
20. Callison-Burch, C., Koehn, P., Monz, C., Peterson, K., Przybocki, M., Zaidan, O.: Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In: Proceedings of the Joint Fifth

Workshop on Statistical Machine Translation and MetricsMATR. (2010) 17–53 Revised August 2010.

21. Callison-Burch, C., Koehn, P., Monz, C., Schroeder, J.: Findings of the 2009 Workshop on Statistical Machine Translation. In: Proceedings of the Fourth Workshop on Statistical Machine Translation. (2009) 1–28
22. Och, F.J.: Minimum error rate training in statistical machine translation. In: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1. ACL '03, Stroudsburg, PA, USA, Association for Computational Linguistics (2003) 160–167
23. Arun, A., Koehn, P.: Online learning methods for discriminative training of phrase based statistical machine translation. In: Proc MT Summit XI. (2007)
24. Tillmann, C., Zhang, T.: A discriminative global training algorithm for statistical mt. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. ACL-44, Stroudsburg, PA, USA, Association for Computational Linguistics (2006) 721–728
25. Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., Schroeder, J.: Further meta-evaluation of machine translation. In: Proceedings of the Third Workshop on Statistical Machine Translation. (2008) 70–106
26. Zaidan, O.: Maise: A flexible, configurable, extensible open source package for mass ai system evaluation. In: Proceedings of the Sixth Workshop on Statistical Machine Translation. (2011) 130–134
27. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement **20:37** (1960)
28. Koehn, P.: Statistical significance tests for machine translation evaluation. In: Proceedings of 2004 EMNLP. (2004)

XINGYI SONG
DEPARTMENT OF COMPUTER SCIENCE,
UNIVERSITY OF SHEFFIELD,
SHEFFIELD, S1 4DP, UK
E-MAIL: <XSONG2@SHEFFIELD.AC.UK>

TREVOR COHN
DEPARTMENT OF COMPUTER SCIENCE,
UNIVERSITY OF SHEFFIELD,
SHEFFIELD, S1 4DP, UK
E-MAIL: <T.COHN@SHEFFIELD.AC.UK>

LUCIA SPECIA
DEPARTMENT OF COMPUTER SCIENCE,
UNIVERSITY OF SHEFFIELD,
SHEFFIELD, S1 4DP, UK
E-MAIL: <L.SPECIA@SHEFFIELD.AC.UK>