A Call for Clarity in Reporting BLEU Scores

Matt Post
Amazon Research
Berlin, Germany

Abstract

The field of machine translation faces an under-recognized problem because of inconsistency in the reporting of scores from its dominant metric. Although people refer to "the" BLEU score, BLEU is in fact a parameterized metric whose values can vary wildly with changes to these parameters. These parameters are often not reported or are hard to find, and consequently, BLEU scores between papers cannot be directly compared. I quantify this variation, finding differences as high as 1.8 between commonly used configurations. The main culprit is different tokenization and normalization schemes applied to the reference. Pointing to the success of the parsing community, I suggest machine translation researchers settle upon the BLEU scheme used by the annual Conference on Machine Translation (WMT), which does not allow for user-supplied reference processing, and provide a new tool, SACREBLEU,[1] to facilitate this.

1 Introduction

Science is the process of formulating hypotheses, making predictions, and measuring their outcomes. In machine translation research, the predictions are made by models whose development is the focus of the research, and the measurement, more often than not, is done via BLEU (Papineni et al., 2002). BLEU's relative language independence, its ease of computation, and its reasonable correlation with human judgments have led to its adoption as the dominant metric for machine translation research. On the whole, it has been a boon to the community, providing a fast and cheap way for researchers to gauge the performance of their models. Together with larger-scale controlled manual evaluations, BLEU has shepherded the field through a decade and a half of quality improvements (Graham et al., 2014).

This is of course not to claim there are no problems with BLEU. Its weaknesses abound, and much has been written about them (cf. Callison-Burch et al. (2006); Reiter (2018)). This paper is not, however, concerned with the shortcomings of BLEU as a proxy for human evaluation of quality; instead, our goal is to bring attention to the relatively narrower problem of the *reporting* of BLEU scores. This problem can be summarized as follows:

- BLEU is not a single metric, but requires a number of parameters (§2.1).
- Preprocessing schemes have a large effect on scores (§2.2). Importantly, BLEU scores computed against differently-processed references are not comparable.
- Papers vary in the hidden parameters and schemes they use, yet often do not report them (§2.3). Even when they do, it can be hard to discover the details.

Together, these issues make it difficult to evaluate and compare BLEU scores across papers, which impedes comparison and replication. I quantify these issues and show that they are serious, with variances bigger than many reported gains. After introducing the notion of *user-* versus *metric-supplied* tokenization, I identify user-supplied reference tokenization as the main cause of this incompatibility. In response, I suggest the community use only *metric-supplied* reference tokenization when sharing scores,[2] following the annual Conference on Machine Translation (Bojar et al., 2017, WMT). In support of this, I release a

[1] https://github.com/awslabs/sockeye/tree/master/contrib/sacrebleu

[2] Sometimes referred to as *detokenized BLEU*, since it requires that system output be detokenized prior to scoring.

Proceedings of the Third Conference on Machine Translation (WMT), Volume 1: Research Papers, pages 186–191
Belgium, Brussels, October 31 - Novermber 1, 2018.
https://doi.org/10.18653/v1/W18-64019

Python package, SACREBLEU,[3] which automatically downloads and stores references for common test sets, thus introducing a "protective layer" between them and the user. It also provides a number of other features, such as reporting a version string which records the parameters used and which can be included in published papers.

2 Problem Description

2.1 Problem: BLEU is underspecified

"BLEU" does not signify a single concrete method, but a constellation of parameterized methods. Among these parameters are:

- The number of references used;
- for multi-reference settings, the computation of the length penalty;
- the maximum n-gram length; and
- smoothing applied to 0-count n-grams.

Many of these are not common problems in practice. Most often, there is only one reference, and the length penalty calculation is therefore moot. The maximum n-gram length is virtually always set to four, and since BLEU is corpus level, it is rare that there are any zero counts.

But it is also true that people use BLEU scores as very rough guides to MT performance across test sets and languages (comparing, for example, translation performance into English from German and Chinese). Apart from the wide intra-language scores between test sets, the number of references included with a test set has a large effect that is often not given enough attention. For example, WMT 2017 includes two references for English–Finnish. Scoring the online-B system with one reference produces a BLEU score of 22.04, and with two, 25.25. As another example, the NIST OpenMT Arabic–English and Chinese–English test sets[4] provided four references and consequently yielded BLEU scores in the high 40s (and now, low 50s). Since these numbers are all gathered together under the label "BLEU", over time, they leave an impression in people's minds of very high BLEU scores for some language pairs or test sets relative to others, but without this critical distinguishing detail.

[3] `pip3 install sacrebleu`

[4] https://catalog.ldc.upenn.edu/LDC2010T21

2.2 Problem: Different reference preprocessings cannot be compared

The first problem dealt with parameters used in BLEU scores, and was more theoretical. A second problem, that of preprocessing, exists in practice.

Preprocessing includes input text modifications such as normalization (e.g., collapsing punctuation, removing special characters), tokenization (e.g., splitting off punctuation), compound-splitting, the removal of case, and so on. Its general goal is to deliver meaningful white-space delimited tokens to the MT system. Of these, tokenization is one of the most important and central. This is because BLEU is a precision metric, and changing the reference processing changes the set of n-grams against which system n-gram precision is computed. Rehbein and Genabith (2007) showed that the analogous use in the parsing community of F_1 scores as rough estimates of cross-lingual parsing difficulty were unreliable, for this exact reason. BLEU scores are often reported as being *tokenized* or *detokenized*. But for computing BLEU, both the system output and reference are always tokenized; what this distinction refers to is whether the reference preprocessing is *user-supplied* or *metric-internal* (i.e., handled by the code implementing the metric), respectively. And since BLEU scores can only be compared when the reference processing is the same, user-supplied preprocessing is error-prone and inadequate for comparing across papers.

Table 1 demonstrates the effect of computing BLEU scores with different reference tokenizations. This table presents BLEU scores where a single WMT 2017 system (online-B) and the reference translation were both processed in the following ways:

- *basic*. User-supplied preprocessing with the MOSES tokenizer (Koehn et al., 2007).[5]
- *split*. Splitting compounds, as in Luong et al. (2015a):[6] e.g., *rich-text* → *rich - text*.
- *unk*. All word types not appearing at least twice in the target side of the WMT training data (with "basic" tokenization) are mapped to UNK. This hypothetical scenario could

[5] Arguments `-q -no-escape -protected basic-protected-patterns -l LANG`.

[6] Their use of compound splitting is not mentioned in the paper, but only here: http://nlp.stanford.edu/projects/nmt.

	English→ ⋆						⋆ →English					
config	en-cs	en-de	en-fi	en-lv	en-ru	en-tr	cs-en	de-en	fi-en	lv-en	ru-en	tr-en
basic	20.7	25.8	22.2	16.9	33.3	18.5	26.8	31.2	26.6	21.1	36.4	24.4
split	20.7	26.1	22.6	17.0	33.3	18.7	26.9	31.7	26.9	21.3	36.7	24.7
unk	20.9	26.5	25.4	18.7	33.8	20.6	26.9	31.4	27.6	22.7	37.5	25.2
metric	20.1	26.6	22.0	17.9	32.0	19.9	27.4	33.0	27.6	22.0	36.9	25.6
range	0.6	0.8	0.6	1.0	1.3	1.4	0.6	1.8	1.0	0.9	0.5	1.2
basic_{lc}	21.2	26.3	22.5	17.4	33.3	18.9	27.7	32.5	27.5	22.0	37.3	25.2
split_{lc}	21.3	26.6	22.9	17.5	33.4	19.1	27.8	32.9	27.8	22.2	37.5	25.4
unk_{lc}	21.4	27.0	25.6	19.1	33.8	21.0	27.8	32.6	28.3	23.6	38.3	25.9
metric_{lc}	20.6	27.2	22.4	18.5	32.8	20.4	28.4	34.2	28.5	23.0	37.8	26.4
$\textit{range}_{lc}$	0.6	0.9	0.5	1.1	0.6	1.5	0.7	1.7	1.0	1.0	0.5	1.2

Table 1: BLEU score variation across WMT'17 language arcs for cased (top) and uncased (bottom) BLEU. Each column varies the processing of the "online-B" system output and its references. *basic* denotes basic user-supplied tokenization, *split* adds compound splitting, *unk* replaces words not appearing at least twice in the training data with UNK, and *metric* denotes the metric-supplied tokenization used by WMT. The *range* row lists the difference between the smallest and largest scores, excluding *unk*.

easily happen if this common user-supplied preprocessing were inadvertently applied to the reference.

- *metric*. Only the metric-internal tokenization of the official WMT scoring script, `mteval-v13a.pl`, is applied.[7]

The changes in each column show the effect these different schemes have, as high as 1.8 for one arc, and averaging around 1.0. The biggest is the treatment of case, which is well known, yet many papers are not clear about whether they report cased or case-insensitive BLEU.

Allowing the user to handle pre-processing of the reference has other traps. For example, many systems (particularly before sub-word splitting (Sennrich et al., 2016) was proposed) limited the vocabulary in their attempt to deal with unknown words. It's possible that these papers applied this same unknown-word masking to the references, too, which would artificially inflate BLEU scores. Such mistakes are easy to introduce in researcher pipelines.[8]

2.3 Problem: Details are hard to come by

User-supplied reference processing precludes direct comparison of published numbers, but if enough detail is specified in the paper, it is at least possible to reconstruct comparable numbers. Unfortunately, this is not the trend, and even for meticulous researchers, it is often unwieldy to include this level of technical detail. In any case, it creates uncertainty and work for the reader. One has to read the experiments section, scour the footnotes, and look for other clues which are sometimes scattered throughout the paper. Figuring out what another team did is not easy.

paper	configuration
Chiang (2005)	metric_{lc}
Bahdanau et al. (2014)	*(unclear)*
Luong et al. (2015)	user or metric *(unclear)*
Jean et al. (2015)	user
Wu et al. (2016)	user or user_{lc} *(unclear)*
Vaswani et al. (2017)	user or user_{lc} *(unclear)*
Gehring et al. (2017)	user, metric

Table 2: Benchmarks set by well-cited papers use different BLEU configurations (Table 1). Which one was used is often difficult to determine.

The variations in Table 1 are only some of the possible configurations, since there is no limit to the preprocessing that a group could apply. But assuming these represent common, concrete configurations, one might wonder how easy it is to determine which of them was used by a particular paper. Table 2 presents an attempt to recover this information from a handful of influential papers in the literature. Not only are systems not comparable due to different schemes, in many cases, no easy determination can be made.

[7] `https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/mteval-v13a.pl`

[8] This paper's observations stem in part from an early version of a research workflow I was using, which applied preprocessing to the reference, affecting scores by half a point.

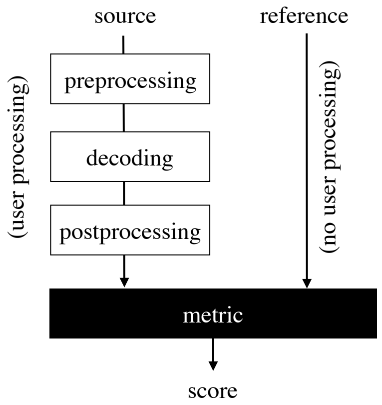

Figure 1: The proper pipeline for computing reported BLEU scores. White boxes denote user-supplied processing, and the black box, metric-supplied. The user should not touch the reference, while the metric applies its own processing to the system output and reference.

2.4 Problem: Dataset specification

Other tricky details exist in the management of datasets. It has been common over the past few years to report results on the English→German arc of the WMT'14 dataset. It is unfortunate, therefore, that for this track (and this track alone), there are actually *two* such datasets. One of them, released for the evaluation, has only 2,737 sentences, having removed about 10% of the original data after problems were discovered during the evaluation. The second, released after the evaluation, restores this missing data (after correcting the problem) and has 3,004 sentences. Many researchers are unaware of this fact, and do not specify which version they use when reporting, which itself contributes to variance.

2.5 Summary

Figure 1 depicts the ideal process for computing sharable scores. Reference tokenization must identical in order for scores to be comparable. The widespread use of user-supplied reference preprocessing prevents this, needlessly complicating comparisons. The lack of details about preprocessing pipelines exacerbates this problem. This situation should be fixed.

3 A way forward

3.1 The example of PARSEVAL

An instructive comparison comes from the evaluation of English parsing scores, where numbers have been safely compared across papers for decades using the PARSEVAL metric (Black et al., 1991). PARSEVAL works by taking labeled spans of the form (N, i, j) representing a nonterminal N spanning a constituent from word i to word j. These are extracted from the parser output and used to compute precision and recall against the gold-standard set taken from the correct parse tree. Precision and recall are then combined to compute the F_1 metric that is commonly reported and compared across parsing papers.

Computing parser F_1 comes with its own set of hidden parameters and edge cases. Should one count the `TOP` (`ROOT`) node? What about `-NONE-` nodes? Punctuation? Should any labels be considered equivalent? These boundary cases are resolved by that community's adoption of a standard codebase, `evalb`,[9] which included a parameters file that answers each of these questions.[10] This has facilitated almost thirty years of comparisons on treebanks such as the Wall Street Journal portion of the Penn Treebank (Marcus et al., 1993).

3.2 Existing scripts

MOSES[11] has a number of scoring scripts. Unfortunately, each of them has problems. Moses' `multi-bleu.perl` cannot be used because it requires user-supplied preprocessing. The same is true of another evaluation framework, MultEval (Clark et al., 2011), which explicitly advocates for user-supplied tokenization.[12] A good candidate is Moses' `mteval-v13a.pl`, which makes use of metric-internal preprocessing and is used in the annual WMT evaluations. However, this script inconveniently requires the data to be wrapped into XML. Nematus (Sennrich et al., 2017) contains a version (`multi-bleu-detok.perl`) that removes the XML requirement. This is a good idea, but it still requires the user to manually handle the reference translations. A better approach is to keep the reference away from the user entirely.

3.3 SACREBLEU

SACREBLEU is a Python script that aims to treat BLEU with a bit more reverence:

- It expects detokenized outputs, applying its own metric-internal preprocessing, and produces the same values as WMT;

[9] http://nlp.cs.nyu.edu/evalb/

[10] The configuration file, `COLLINS.PRM`, answers these questions as no, no, no, and ADVP=PRT.

[11] http://statmt.org/moses

[12] https://github.com/jhclark/multeval

- it automatically downloads and stores WMT (2008–2018) and IWSLT 2017 (Cettolo et al., 2017) test sets, obviating the need for the user to handle the references at all; and

- it produces a short version string that documents the settings used.

SACREBLEU can be installed via the Python package management system:

```
pip3 install sacrebleu
```

It can then be used to download the source side of test sets as decoder input—all WMT test sets are available, as well as recent IWSLT test sets, and others are being added. After decoding and detokenization, it can then used to produce BLEU scores.[13] The following command selects the WMT'14 EN-DE dataset used in the official evaluation:

```
cat output.detok \
  | sacrebleu -t wmt14 -l en-de
```

(The restored version that was released after the evaluation (§2.4) can be selected by using `-t wmt14/full`.) It prints out a version string recording all the parameters as '+' delimited KEY.VALUE pairs (here shortened with `--short`):

```
BLEU+c.mixed+l.en-de+#.1+s.exp
  +t.wmt14+tok.13a+v.1.2.10
```

recording:

- mixed case evaluation
- on EN-DE
- with one reference
- and exponential smoothing
- on the WMT14 dataset
- using the WMT standard '13a' tokenization
- with SACREBLEU 1.2.10.

SACREBLEU is open source software released under the Apache 2.0 license.

[13] The CHRF metric is also available via the `-m` flag.

4 Summary

Research in machine translation benefits from the regular introduction of test sets for many different language arcs, from academic, government, and industry sources. It is a shame, therefore, that we are in a situation where it is difficult to directly compare scores across these test sets. One might be tempted to shrug this off as an unimportant detail, but as was shown here, these differences are in fact quite important, resulting in large variances in the score that are often much higher than the gains reported by a new method.

Fixing the problem is relatively simple. Research groups should only report BLEU computed using a metric-internal tokenization and preprocessing scheme for the reference, and they should be explicit about the BLEU parameterization they use. With this, scores can be directly compared. For backwards compatibility with WMT results, I recommend the processing scheme used by WMT, and provide a new tool that makes it easy to do so.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *ArXiv e-prints*, abs/1409.0473.

E. Black, S. Abney, D. Flickenger, C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini, and T. Strzalkowski. 1991. A procedure for quantitatively comparing the syntactic coverage of english grammars. In *Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, February 19-22, 1991*.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214. Association for Computational Linguistics.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluation the role of bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stker, Katsuitho Sudoh, Koichiro Yoshino, and Christian Federmann. 2017.

Overview of the iwslt 2017 evaluation campaign. In *14th International Workshop on Spoken Language Translation*, pages 2–14, Tokyo, Japan.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 263–270. Association for Computational Linguistics.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 176–181. Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, and Yann Dauphin. 2017. A convolutional encoder model for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 123–135. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2014. Is machine translation getting better over time? In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 443–451. Association for Computational Linguistics.

Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. On using very large target vocabulary for neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015a. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421. Association for Computational Linguistics.

Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals, and Wojciech Zaremba. 2015b. Addressing the rare word problem in neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 11–19. Association for Computational Linguistics.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics, Volume 19, Number 2, June 1993, Special Issue on Using Large Corpora: II*.

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*.

Ines Rehbein and Josef van Genabith. 2007. Treebank annotation schemes and parser evaluation for german. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Ehud Reiter. 2018. A structured review of the validity of bleu. *Computational Linguistics*, 0(0):393–401.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde. 2017. Nematus: a toolkit for neural machine translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–68. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *ArXiv e-prints*, abs/1706.03762.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, ukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *ArXiv e-prints*, abs/1609.08144.