A Structured Review of the Validity of BLEU

Ehud Reiter
University of Aberdeen
Department of Computing Science
e.reiter@abdn.ac.uk

The BLEU metric has been widely used in NLP for over 15 years to evaluate NLP systems, especially in machine translation and natural language generation. I present a structured review of the evidence on whether BLEU is a valid evaluation technique—in other words, whether BLEU scores correlate with real-world utility and user-satisfaction of NLP systems; this review covers 284 correlations reported in 34 papers. Overall, the evidence supports using BLEU for diagnostic evaluation of MT systems (which is what it was originally proposed for), but does not support using BLEU outside of MT, for evaluation of individual texts, or for scientific hypothesis testing.

1. Introduction

BLEU (Papineni et al. 2002) is a metric that is widely used to evaluate Natural Language Processing (NLP) systems which produce language, especially machine translation (MT) and Natural Language Generation (NLG) systems. Because BLEU itself just computes word-based overlap with a gold-standard reference text, its use as an evaluation metric depends on an assumption that it correlates with and predicts the real-world utility of these systems, measured either extrinsically (e.g., by task performance) or by user satisfaction. From this perspective, it is similar to *surrogate endpoints* in clinical medicine, such as evaluating an AIDS medication by its impact on viral load rather than by explicitly assessing whether it leads to longer or higher-quality life.

Hence the usage of BLEU to evaluate NLP systems is only sensible in the presence of *validation studies* which show that BLEU scores correlate with direct evaluations of the utility of NLP systems. In rough terms, a validation study involves evaluating a number of NLP systems (or individual output texts) using both a metric such as BLEU and a gold standard human evaluation, and then calculating how well the metric correlates with the gold-standard human evaluation.

Many such studies have been published, and in this paper I present a **structured review** of these studies. Structured reviews are literature reviews that are designed to be objective, repeatable, and comprehensive. In other words, whereas the author of a normal literature review uses their knowledge of the field to identify key papers and qualitatively summarize their findings, the author of a structured review identifies relevant papers via objective search criteria and extracts from each paper key information

Submission received: 30 August 2017; revised version received: 23 January 2018; accepted for publication: 26 March 2018.

doi:10.1162/COLI_a_00322

accurate evaluation	automatic human	automatic measures
automatic metrics	automatic validity	bleu
correlation human	empirical metrics	”evaluating evaluation”
improving evaluation	intrinsic extrinsic	meta evaluation
metric human	metric validity	pitfalls evaluation

Figure 1
Title search terms for identifying papers in the ACL Anthology. I also included all papers that presented results of the WMT metric evaluation shared tasks.

in a structured fashion. Structured reviews are widely used in clinical medicine to integrate evidence from many separate studies. In this paper, I use this methodology to address the question of whether BLEU is a valid evaluation technique.

2. Protocol

2.1 Identifying Papers

The first step in the Prisma process for structured reviews (Moher et al. 2009) is to identify candidate papers by doing a well-defined search on relevant archive(s).

I used the ACL Anthology[1] as the archive of candidate papers. The ACL Anthology is far from ideal, because it does not include many relevant studies. However, despite this problem I believe it is the best archive for a structured review in NLP. I also looked at Arxiv,[2] but it is not suitable because it contains many papers that are drafts (not final versions) and also many papers that have not gone through a peer review process.

I searched the ACL Anthology using a title search on the search words listed in Figure 1; this was done in late June 2017. Title search is not ideal, but unfortunately the ACL Anthology does not support abstract or keyword search. I created an initial list of search terms myself, and expanded this by analyzing the titles of relevant papers suggested by colleagues. I quoted one search term, ”evaluating evaluation,” as otherwise it matched every paper that mentioned ”evaluation” in its title. Normal search was used on all other terms. I also automatically included all papers that presented results of the WMT conferences’ metric evaluation shared tasks, such as Bojar et al. (2016a).

Colleagues have subsequently pointed out to me several relevant papers that were published after June 2017, such as Novikova et al. (2017), or otherwise were missed by my survey. I have not added these papers to my survey (it is not appropriate to add individual papers to a structured review; the only way to include these papers would be to redo the entire survey with new criteria and end date). However, I have read all of these papers, and they are consistent with the core findings of my survey.

My survey is limited to BLEU, and does not look at other popular metrics such as METEOR (Banerjee and Lavie 2005). I focus on BLEU because I believe it is the most popular metric; indeed, Papineni et al. (2002) is one of the most cited NLP papers, according to Google Scholar, and has been given a NAACL Test of Time award.

[1] http://aclanthology.info/.
[2] https://arxiv.org/.

2.2 Screening Papers

I screened the candidate papers by reading through them and selecting papers that met the criteria presented in Figure 2. If a paper presented several correlations between BLEU and human evaluations, I looked at the correlations individually and in some cases accepted some but rejected others.

Language: Written in English, because this is the only language I am fluent in.

BLEU: Looked at a standard version of BLEU. If a paper looked at different variants of standard BLEU (e.g., with different tokenization), I just reported on the author's preferred variant; if no preference was expressed, I used the variant that correlated best with human studies. I did not look at the NIST metric (Doddington 2002), that is, I did not consider NIST to be a standard version of BLEU.

For example, Bouamor et al. (2014) presents correlations for both standard BLEU and a modified version called AL-BLEU. I included the correlation with standard BLEU, but not AL-BLEU.

Correlation: Presented a correlation between BLEU and a human evaluation of NLP systems. I insisted on a correlation (Pearson, Spearman, or Kendall); I excluded studies that assessed the agreement between BLEU and human evaluations using other techniques. I also excluded studies that assessed how well BLEU correlated with human evaluations of human-written texts (since this is not my research question). I accepted papers that included human-written texts as one of the ″systems″ being evaluated, provided that the study also looked at at least one computer NLP system.

For example, Kilickaya et al. (2017) presents BLEU–human correlations on several data sets. I only included the correlation with the COMPOSITE data set, which contains a mixture of human-written and computer-generated texts; I excluded correlations computed on the other data sets (e.g., PASCAL-50S) because they did not contain any computer-generated texts.

Size: Included at least 5 NLP systems if BLEU scores are computed at the system level, or at least 5 texts if scores are computed at the text level. Five data points are the minimum needed to be able to have a statistically significant Spearman correlation.

For example, Callison-Burch, Osborne, and Koehn (2006) present an initial set of correlations based on seven systems (their Figures 2 and 3) and a second set of correlations based on three systems (their Figure 4); I included the former but not the latter.

Originality: Did not re-present results that had been presented in another paper; this prevents double-counting. I always preferred the primary source describing a study over a secondary source. However, if a paper presented an improved version of a study piloted in an earlier paper, I only included the later study.

For example, I excluded Belz and Reiter (2006) because Reiter and Belz (2009) presents an improved version of that study.

Figure 2
Inclusion criteria for studies.

- *NLP Systems in the study*
  - Type (e.g., MT) and subtype (e.g., Chinese-to-English) of NLP system
  - Output language (e.g., English)
  - Domain (e.g., news)
- *BLEU scoring details*
  - Granularity: Whether BLEU scores were calculated for NLP *systems* or for individual *texts* produced by these systems. Note that some papers use the term *segment* for what I refer to as *text* granularity.
  - Number of reference texts (e.g., 1)
  - Source of reference texts (e.g., professional translators)
- *Gold-standard human evaluation*
  - Type (e.g., ranking)
  - Participants (e.g., Mechanical Turk)
  - Aspect ranked or rated (e.g., fluency)
  - Inter-annotator agreement between participants
- *Result*
  - Type of correlation (Pearson, Spearman, Kendall)
  - Actual correlation
  - Any potential bias—for example, if the paper was presenting an alternative metric that was supposed to be better than BLEU.

Figure 3
Information extracted from studies.

2.3 Extracting Information from Papers

I tried to extract the information described in Figure 3 from each paper that made it through the screening process. However, in many cases I could not find all of this information in the paper.

Some of the papers surveyed (as well as many of the papers I excluded) gave interesting qualitative analyses of cases when BLEU provides misleading results. For example, Bouamor et al. (2014) explain BLEU's weaknesses in evaluating texts in morphologically rich languages such as Arabic, and Espinosa et al. (2010) point out that BLEU inappropriately penalizes texts that have different adverbial placement compared with reference texts. These comments are interesting and valuable research contributions, but in this structured review my focus is on quantitative correlations between BLEU and human evaluations.

3. Results

The full results of the survey are presented in the data file associated with this article.[3] I summarize key findings here.

[3] `http://dx.doi.org/10.20392/766c9dd8-75a7-4761-915d-856c0f7cc3c4`.

One important question is what level of correlation is sufficient for BLEU to be regarded as a valid proxy for human evaluation. I will use the following classification.

- *High:* Correlation is 0.85 or higher
- *Medium:* Correlation is between 0.70 and 0.85
- *Low:* Correlation is between 0 and 0.70
- *Negative:* Correlation is below 0

The High, Medium, and Low classification is based on the classification of surrogate endpoints in Prasad et al. (2015), which in turn is based on criteria from the Institute for Quality and Efficiency in Health Care (IQWiG) in Germany, which assesses the benefits of medical interventions. IQWiG essentially only considers surrogate endpoints to be meaningful if reliable validation studies show Medium or High correlation with actual clinical outcomes. I have added the category Negative for correlations below 0.

Many papers present more than one correlation between BLEU and human evaluations. The results presented here are based on treating each correlation as a separate data point. I also computed results on a paper-weighted basis, where if a paper presents n correlations, each of these correlations is given a weight of $1/n$. Paper-weighted results are similar to the unweighted results, so I do not present them separately here.

One clear finding is that BLEU–human correlations are poor for NLG (Figure 4). For MT, they are poor for text-level correlations, but reasonable for system-level correlations (Figure 5). Hence the only kind of BLEU–human correlation that is mostly Medium or High in the surveyed papers is system-level BLEU–human correlations for MT.

My survey also included six BLEU-correlations for other types of NLP systems. I have not shown box plots for these because the data set is so small. Five of these correlations are Low; the sixth (reported in Graham [2015]) is High.

Callison-Burch, Osborne, and Koehn (2006) and others (e.g., Figure 3 in Bojar et al. [2016b]) have suggested that BLEU is biased against certain technologies, and hence it correlates better with human judgments when it is used to evaluate systems built with similar technologies. Unfortunately, this survey cannot shed light on this question, because many papers do not say which technologies are used in the systems evaluated.

One striking result from the survey is the wide range of BLEU–human correlations reported, even for similar tasks. For example, this survey includes 10 papers that report

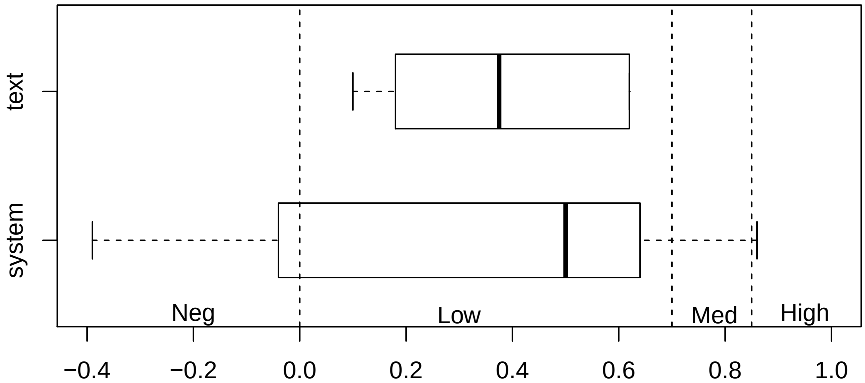

Figure 4
Box plot of BLEU–human correlations for NLG, at system and text granularities.

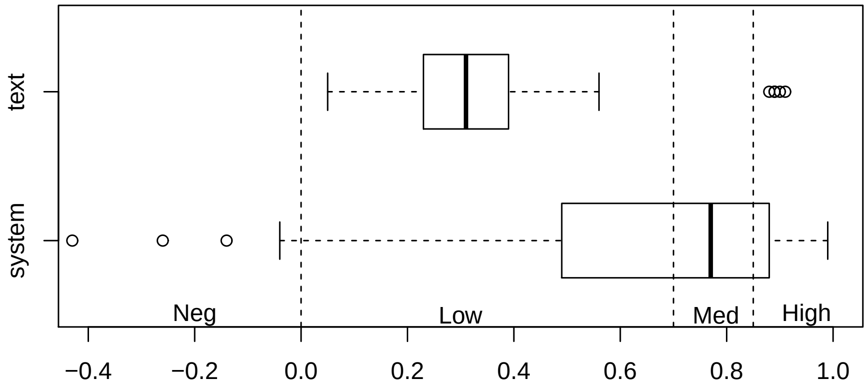

Figure 5
Box plot of BLEU–human correlations for MT, at system and text granularities.

results from WMT events (WMT07 to WMT16). Every one of these papers reports the correlation of BLEU with ranking-based human assessments of English–German and German–English MT systems in a news domain. One would hope that BLEU–human correlations would be similar in such a constrained context, but in fact correlations vary widely, as shown in Table 1. This suggests that whether BLEU correlates with human evaluations is very dependent on the details of the systems being evaluated, the exact corpus texts used, and the exact protocol used for human evaluations. If this is the case, then it is difficult to predict whether BLEU will correlate well with human evaluations in a new context. This is a concern, because surrogate endpoints are only useful if they can reliably predict outcomes (e.g., human evaluations) in new contexts.

4. Discussion: What Is a Good Gold-Standard Human Evaluation?

Surrogate endpoints such as BLEU are useful if they can reliably predict an outcome that is of real-world importance or is the core of a scientific hypothesis we wish to test. In a medical context, validation studies are expected to correlate the surrogate endpoint against direct measurements of the outcome of interest.

In NLP, human evaluations can be based on human ratings or rankings (intrinsic) or on measurement of an outcome such as task performance (extrinsic); they can also be carried out in laboratory or real-world contexts. The strongest and most meaningful evaluation is a real-world outcome-based evaluation, where a system is operationally deployed and we measure its impact on real-world outcomes.[4]

From this perspective, it is striking that few of the surveyed papers looked at task/outcome measures. Indeed, only one paper correlated system-level BLEU scores with task outcomes (Belz and Gatt 2008), and all such correlations in that paper were Low or Negative. *None* of the surveyed papers used real-world human evaluations; that is, they all used human evaluations performed in an artificial context (usually by paid individuals, crowdsourced workers, or the researchers themselves), rather than looking at the impact of systems on real-world users.

[4] Examples are given in `https://ehudreiter.com/2017/01/19/types-of-nlg-evaluation/`.

Table 1
Correlation of BLEU with ranking-based human evaluation reported in WMT events, for German–English and English–German MT in a news domain.

Event	correlation type	correlation for German–English MT	correlation for English–German MT
WMT07	Spearman	0.40	0.26
WMT08	Spearman	0.12	0.58
WMT09	Spearman	0.41	−0.43
WMT10	Spearman	0.52	0.39
WMT11	Spearman	0.48	0.44
WMT12	Spearman	0.67	0.22
WMT13	Spearman	0.90	0.83
WMT14	Pearson	0.83	0.22
WMT15	Pearson	0.86	0.57
WMT16	Pearson	0.88	0.78

The most common way to measure real-world effectiveness in computing is with A/B testing, where different real-world users of a service are given access to different systems. A/B testing is most often used to measure user satisfaction, but it could also be used to measure extrinsic outcomes such as post-edit time in an MT context. I suspect that many commercial providers of online NLP services have carried out a considerable amount of A/B testing. I realize that the results of such tests are commercially confidential, but if it were possible for such providers to publish correlations between their A/B tests and BLEU, that would be very helpful in assessing the validity of BLEU.

I am *not* suggesting that academic researchers evaluate systems using task/outcome-based real-world A/B testing—this is clearly not feasible. What I am saying is that the results of real-world A/B testing could be used to determine contexts in which BLEU reliably had good correlation with real-world effectiveness. Researchers could then confidently use BLEU as a surrogate endpoint in these contexts.

5. Conclusion: Is BLEU Valid?

BLEU was originally proposed for *diagnostic evaluations* of MT systems, that is, as a technique for allowing researchers and developers to quickly "weed out bad ideas from good ideas" (Papineni et al. 2002, page 311). I think the surveyed papers support this use of BLEU, since most of the system-level BLEU–human correlations for MT reported in the survey are Medium or High (Figure 5).

However, the evidence does *not* support using BLEU to evaluate other types of NLP systems (outside of MT), and it does *not* support using BLEU to evaluate individual texts rather than NLP systems.

Also, BLEU should not be the primary evaluation technique in NLP papers. Researchers can use it for diagnostic evaluation when developing their ideas. However, when they present evidence to the community (e.g., via an ACL paper) that their approach is effective (*scientific hypothesis testing*), this evidence should not be based primarily on BLEU. This is because of the following concerns about the validity and reliability of BLEU:

- There are a wide range of correlations between BLEU and human evaluations, even in very similar tasks (e.g., Table 1). This suggests that the correlation is dependent on contextual factors.

- The human evaluations in the validation studies surveyed do not directly measure real-world outcomes. I suspect these studies would not be regarded as acceptable by the standards of medical research (Section 4).
- BLEU has technological biases that we do not understand. This is especially worrying as new technologies such as neural networks become more prominent; we do not know if BLEU is ”fair” to such technologies.

These recommendations would change if solid evidence was presented in high-quality validation studies that clearly indicated the contexts in which BLEU reliably had good correlation with real-world extrinsic human evaluations of NLP systems. Such validation studies are not cheap or easy, but they are necessary if the NLP community wishes to confidently use BLEU for testing scientific hypotheses.

Acknowledgments

Many thanks to the anonymous reviewers and my colleagues at Aberdeen for their very helpful comments. I’d also like to thank the many people who responded to my talks and blog posts about metric evaluation for their suggestions and comments.

References

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

Belz, Anja and Albert Gatt. 2008. Intrinsic vs. extrinsic evaluation measures for referring expression generation. In *Proceedings of ACL-2008*, pages 197–200.

Belz, Anja and Ehud Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *Proceedings of EACL 2006*, pages 313–320.

Bojar, Ondřej, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016a. Results of the WMT16 metrics shared task. In *Proceedings of WMT-2016*, pages 199–231.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016b. Findings of the 2016 conference on machine translation. In *Proceedings of WMT-2016*, pages 131–198.

Bouamor, Houda, Hanan Alshikhabobakr, Behrang Mohit, and Kemal Oflazer. 2014. A human judgment corpus and a metric for Arabic MT evaluation. In *Proceedings of EMNLP-2014*, pages 207–213.

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

Doddington, George. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of HLT 2002*, pages 138–145.

Espinosa, Dominic, Rajakrishnan Rajkumar, Michael White, and Shoshana Berleant. 2010. Further meta-evaluation of broad-coverage surface realization. In *Proceedings of EMNLP-2010*, pages 564–574.

Graham, Yvette. 2015. Re-evaluating automatic summarization with BLEU and 192 shades of ROUGE. In *Proceedings of EMNLP-2015*, pages 128–137.

Kilickaya, Mert, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. 2017. Re-evaluating automatic metrics for image captioning. In *Proceedings of EACL-2017*, pages 199–209.

Moher, David, Alessandro Liberati, Jennifer Tetzlaff, and Douglas G. Altman. 2009. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *BMJ*, 339:b2535.

Novikova, Jekaterina, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of EMNLP-2017*, pages 2241–2252.

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

Prasad, Vinay, Chul Kim, Mauricio Burotto, and Andrae Vandross. 2015. The strength of association between surrogate end points and survival in oncology: A systematic review of trial-level meta-analyses. *JAMA Internal Medicine*, 175(8):1389–1398.

Reiter, Ehud and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.