

Explanation of data files for BLEU structured review

Data files

Two sets of data files are presented, each in both CSV and Excel format

- **bleu-validation-review-papers** . This table lists the 34 papers in the structured review. Each row represents a different paper.
- **bleu-validation-review-corr** . This table lists the 284 individual correlations reported in the structured review. Each row represents a different correlation.

The appendix of this document lists all of the papers in the review in standard bibliographic form.

Information included in the spreadsheets

Note: ?? means that the information was not specified in the paper

Paper Information: Information about the paper

ACL Anth ID: The ID of the paper in the ACL Anthology (<http://aclanthology.info/>)

Year: The year the paper was published

Num corr: Number of correlations reported in the paper

Corr weight: Weight of this correlation in paper-weighted averages. *Only in corr spreadsheet*

Title: Title of paper. *Only in papers spreadsheet*

URL: URL of paper. *Only in papers spreadsheet*

Systems Evaluated: Information about the NLP systems in the study. These are the systems which were evaluated using both BLEU and a human evaluation, providing data for BLEU-human correlation.

Type: One of

- MT
- NLG
- other NLP

Subtype: additional information about the type of system (eg, Chinese-English MT)

Domain: domain of systems, such as news or medical. Some papers did not specify this.

Language: Language produced by the systems (eg, English). “multiple” means that the paper reported multiple studies, and these studies cover more than one language.

Note: I originally also tried to get information about the technology used in the systems, eg rule-based, statistical, or neural. It was often difficult to get this information from papers, so I dropped this.

BLEU details: Information about how BLEU scores were calculated

Granularity: One of

- system: BLEU scores were calculated for NLP systems.
- text: BLEU scores were calculated for individual texts. Ie, if an NLP system was run on 100 texts, a BLEU score was calculated for each of these 100 texts. This is often referred to as “segment” in the literature.
- multiple: paper reports multiple studies, some of which use system granularity, and others of which use text granularity.

Num ref texts: Number of reference texts used to calculate BLEU score. Some papers did not specify this.

Source of ref texts: Who wrote the reference texts (eg professional translators). MTurk means Amazon Mechanical Turk. Some papers did not specify this.

Note: I originally tried to get information about the exact BLEU algorithm used, but this often was not specified, and also when it was specified, it was sometimes a dead URL. So I dropped this.

Human study details: Information about the human evaluation

Type: One of

- Rating: Subjects were asked to rate texts, often on a Likert scale
- Ranking: Subjects were shown multiple texts and asked to rank them
- Task: Subjects were asked to perform a task (eg, post-editing), and their performance was measured by some metric (eg, time)
- Multiple: paper reported multiple studies, and these studies used more than one of the above.

Subjects: Who the subjects were (eg students). MTurk means Amazon Mechanical Turk. Some papers did not specify this.

Measure: What was ranked, rated, or measured (eg, adequacy).

Interannotator agreement: Interannotator agreement between subjects. This information was often not specified, but it is useful, so I included this when it was specified. *Only in papers spreadsheet*

Note: I originally also wanted to assess whether subjects understood the texts they were reading, and also whether they understood the input data/text to the NLP system. But this is difficult to define precisely, and was rarely discussed in the papers, so I dropped this.

Result: Correlations, and how they were computed

Correlation type: one of

- Pearson
- Spearman
- Kendall
- Multiple: paper reported multiple studies, and these studies used more than one type of correlation
- ??: not specified in paper

Reversed polarity: For a few studies, better outcome was indicated by a lower rather than higher human evaluation score. For example, we want post-edit time to be low rather than high. Since higher BLEU score is supposed to indicate a better system, I reversed the polarity of the correlation in such cases. This is indicated by “reversed” in this column.

Corr: Correlation between BLEU and human evaluation. *Only in corr spreadsheet*

Category: One of the below. This is explained in Section 3 of the paper (*Only in corr spreadsheet*)

- High: correlation ≥ 0.85
- Medium: $0.85 > \text{correlation} \geq 0.70$
- Low: $0.70 > \text{correlation} \geq 0$
- Negative: $0 > \text{correlation}$

Corr minimum: Minimum correlation reported in the paper. If a paper only reports one correlation, it will be shown here. Note that there is a separate spreadsheet which lists each individual correlation (study) in the papers reviewed. *Only in papers spreadsheet*

Where min: Which correlation in the paper was lowest. This is blank if a paper only reports one correlation. *Only in papers spreadsheet*

Corr max: Maximum correlation reported in the paper. If a paper only reports one correlation, it will be shown here. Note that there is a separate spreadsheet which lists each individual correlation (study) in the papers reviewed. *Only in papers spreadsheet*

Where max: Which correlation in the paper was highest. This is blank if a paper only reports one correlation. *Only in papers spreadsheet*

Notes: Additional information

Bias?: Potential bias, if any (eg, paper proposes a competing metric to BLEU). *Only in papers spreadsheet*

Comment: Free text comment with additional relevant information about the paper. Comments in paper spreadsheet apply to every correlation in a paper. Comments in corr spreadsheet only apply to a specific correlation

Appendix: Papers included in the survey

- Babych, Bogdan and Anthony Hartley. 2008. Sensitivity of automated MT evaluation metrics on higher quality MT output: BLEU vs task-based evaluation methods. In *Proc of LREC-2008*.
- Banerjee, Satanjeev and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proc 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 *Proc of ACL-2008*, pages 197–200.
- Bojar, Ondřej et al. 2016. Findings of the 2016 conference on machine translation. In *Proc of WMT-2016*, pages 131–198.
- Bojar, Ondřej, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016. Results of the WMT16 metrics shared task. In *Proc of WMT-2016*, pages 199–231.
- Bouamor, Houda, Hanan Alshikhabobakr, Behrang Mohit, and Kemal Oflazer. 2014. A human judgement corpus and a metric for Arabic MT evaluation. In *Proc of EMNLP-2014*, pages 207–213.
- Cahill, Aoife. 2009. Correlating human and automatic evaluation of a german surface realiser. In *Proc of ACL-2009*, pages 97–100.
- Callison-Burch, Chris, Miles Osborne, and Philipp Koehn. 2006. Re-evaluation the role of BLEU in machine translation research. In *Proc of EACL-2006*, pages 249–256.
- Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta-) evaluation of machine translation. In *Proc of WMT-2007*, pages 136–158.
- Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proc of WMT-2008*, pages 70–106.
- Callison-Burch, Chris, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proc of the WMT-2009*, pages 1–28.
- Callison-Burch, Chris, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proc of WMT-2010*, pages 17–53.
- Callison-Burch, Chris, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 Workshop on Statistical Machine Translation. In *Proc of WMT-2011*, pages 22–64.
- Callison-Burch, Chris, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 Workshop on Statistical Machine Translation. In *Proc of WMT-2012*, pages 10–51.
- Echizen'ya, Hiroshi, Kenji Araki, and Eduard Hovy. 2013. Automatic evaluation metric for machine translation that is independent of sentence length. In *Proc of RANLP-2013*, pages 230–236.
- Elliott, Desmond and Frank Keller. 2014. Comparing automatic evaluation measures for image description. In *Proc of ACL-2014*, pages 452–457.
- Espinosa, Dominic, Rajakrishnan Rajkumar, Michael White, and Shoshana Berleant. 2010. Further meta-evaluation of broad-coverage surface realization. In *Proc of EMNLP-2010*, pages 564–574.

- Finch, Andrew, Yasuhiro Akiba, and Eiichiro Sumita. 2004. How does automatic machine translation evaluation correlate with human scoring as the number of reference translations increases? In *Proc of LREC-2004*.
- Graham, Yvette. 2015. Re-evaluating automatic summarization with BLEU and 192 shades of ROUGE. In *Proc of EMNLP-2015*, pages 128–137.
- Graham, Yvette, Timothy Baldwin, and Nitika Mathur. 2015. Accurate evaluation of segment-level machine translation metrics. In *Proc of NAACL-2015*, pages 1183–1191.
- Kilickaya, Mert, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. 2017. Re-evaluating automatic metrics for image captioning. In *Proc of EACL-2017*, pages 199–209.
- Kim, Su Nam, Timothy Baldwin, and Min-Yen Kan. 2010. Evaluating n-gram based evaluation metrics for automatic keyphrase extraction. In *Proc of Coling 2010*, pages 572–580.
- Lin, Chin-Yew and Franz Josef Och. 2004. ORANGE: a method for evaluating automatic evaluation metrics for machine translation. In *Proc of COLING 2004*.
- Liu, Chia-Wei, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proc of EMNLP-2016*, pages 2122–2132.
- Lo, Chi-kiu and Dekai Wu. 2011. MEANT: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on semantic roles. In *Proc of ACL-2011*, pages 220–229.
- Macháček, Matouš and Ondřej Bojar. 2013. Results of the WMT13 metrics shared task. In *Proc of WMT-2013*, pages 45–51.
- Machacek, Matous and Ondrej Bojar. 2014. Results of the WMT14 metrics shared task. In *Proc of WMT-2014*, pages 293–301.
- Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc of ACL 2002*, pages 311–318.
- 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).
- Stanojević, Miloš, Amir Kamran, Philipp Koehn, and Ondřej Bojar. 2015. Results of the WMT15 metrics shared task. In *Proc of WMT-2015*, pages 256–273.
- Sun, Yanli. 2010. Mining the correlation between human and automatic evaluation at sentence level. In *Proc of LREC-2010*.
- Yasuda, Keiji, Fumiaki Sugaya, Toshiyuki Takezawa, Seiichi Yamamoto, and Masuzo Yanagida. 2003. Applications of automatic evaluation methods to measuring a capability of speech translation system. In *Proc of EACL-2003*.
- Zhang, Ying, Stephan Vogel, and Alex Waibel. 2004. Interpreting BLEU/NIST scores: How much improvement do we need to have a better system? In *Proc of LREC-2004*.
- Zhu, Junguo, Muyun Yang, BoWang, Sheng Li, and Tiejun Zhao. 2010. All in strings: a powerful string-based automatic MT evaluation metric with multiple granularities. In *Proc of Coling 2010*, pages 1533–1540.