On conducting better validation studies of automatic metrics in natural language generation evaluation

Anonymous NAACL submission

Abstract

Natural language generation (NLG) has received increasing attention, which has highlighted evaluation as a central methodological concern. Since human evaluations for these systems are costly, automatic metrics have broad appeal in NLG. Research in language generation often finds situations where it is appropriate to apply existing metrics or propose new ones. The application of these metrics are entirely dependent on validation studies - studies that determine a metric's correlation to human judgment. However, there are many details and considerations in conducting strong validation studies. This document is intended for those validating existing metrics or proposing new ones in the broad context of NLG: we 1) begin with a write-up of best practices in validation studies, 2) outline how to adopt these practices, 3) conduct analyses in the WMT'17 metrics shared task[1], and 4) highlight promising approaches to NLG metrics 5) conclude with our opinions on the future of this area.

1 Introduction

Increasing interest in tasks that require generating natural language such as image captioning (Lin et al., 2014), dialogue (Vinyals and Le, 2015), and style transfer (Fu et al., 2018) highlight evaluation as a central methodological concern. With human involvement, a system can be evaluated extrinsically (how well does the system fulfill its intended purpose? e.g. Reiter et al., 2003) or intrinsically (what is the quality of the output?) In domains such as machine translation (MT), a system's extrinsic value is both hard to define and measure, and intrinsic human judgments of a system's output quality have been the main indicator of progress in the field. (Bojar et al., 2016b)

---

[1]Our jupyter notebook containing the analyses is available at https://github.com

Source	Reference	Output	DA	BEER
双方很难，甚至不可能重新建立真正的信任。	rebuilding real trust will be hard , perhaps impossible .	it is difficult , if not impossible , to re-establish real trust .	0.55	0.39
如何在持枪攻击中使用马伽术保护自己	how to defend yourself from gun attacks using krav ...	how to use marcella to protect himself in a gun attack	-0.85	0.42

⋮

Figure 1: Examples from the WMT'17 metrics task for `zh-en` translation evaluation. Outputs produced from Sennrich et al. (2017). The DA score is a mean human judgment of translation quality. More on DA (Direct Assessment) in §2.2. Scores of a participating metric (BEER; Stanojevic and Sima'an, 2014) are shown. Metrics aim to achieve high correlation with DA scores. More on task details in §2.

This paper focuses on those domains best evaluated intrinsically. Since acquiring intrinsic human judgments is costly, automatic metrics, which both are computed automatically and correlate highly with human judgment, are ideal. If sufficiently correlated, a metric can be used as a surrogate evaluation, which may be useful in developmental cycles. Therefore, the application of such metrics are dependent on studies of their validity (Reiter, 2018, how well does the metric correlate with human judgment?). In MT, BLEU (Papineni et al., 2002) has seen widespread use, and, consequently, its validity has been extensively studied. (Callison-Burch et al., 2006, inter alia)

These automatic metrics are generally appealing to natural language generation (NLG) domains. During 2005-2014, Gkatzia and Mahamood (2015) found that a significant portion of NLG research in ACL reported results from automatic metrics. Research in under-explored NLG domains have begun with proposals of both models and novel evaluation metrics. (Fu et al., 2018; Yao et al., 2018) For the application of any existing metric (e.g. BLEU) or newly proposed metric to bear validity in various domains, researchers have

Figure 2: The interface for collection of *direct assessment* (DA) scores from Amazon Mechanical Turk workers. Figure taken from Bojar et al. (2017b). Each worker rates system outputs on a continuous 0-100 scale, which are converted to deviations from the worker's mean to normalize over scoring strategies.

attempted to conduct validation studies, for tasks such as surface realization (Novikova et al., 2017), open-domain dialogue (Liu et al., 2016), and image captioning (Kilickaya et al., 2017), at times reporting negative results.

However, there are many considerations when designing a validation study. There are at least two aspects that experimenters should be cognizant about in conducting a robust validation study: first are the assumptions made; the second is statistical methodology. Not unlike testing of our models, validation of our metrics should also be approached with rigor. At the time of writing, the metrics shared task in the Conference of Machine Translation (WMT; Bojar et al., 2018b) annually conducts strong validation studies and we recommend adopting best practices from this domain.

This document is intended for those validating existing metrics or proposing new ones in NLG. To this end, we present:

- an overview of the WMT metrics shared task validation procedures (§2) and an outline on how to adopt these best practices, generally, to other NLG domains (§3).

- proposals of basic analyses of metrics scoring, complemented by analysis on existing data from the WMT'17 metrics task (§4).

- a literature review in both metrics (§5) and their analysis, (§6) concluding with the authors' opinions on directions of the area (§7).

2 WMT metrics shared task

The metrics shared task (Ma et al., 2018) utilizes the WMT evaluation results of the popular translation shared task. The data provided from the
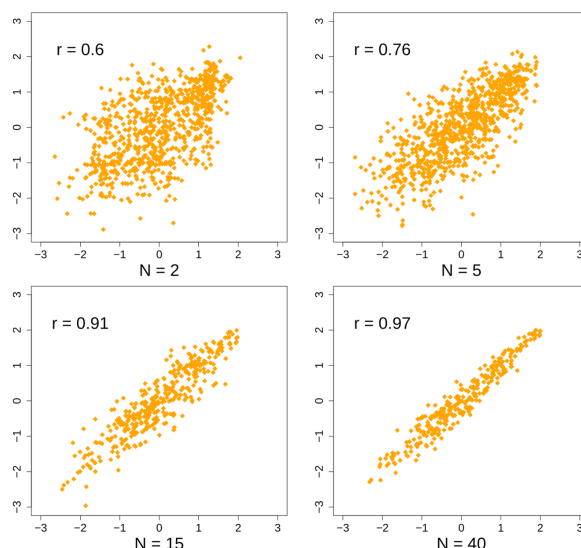

Figure 3: The effect of the number of assessors on the consistency of DA scores. Figure taken from Graham et al. (2015). The x-axis is a sample mean of translation quality calculated from n judgments, and the y-axis is a true mean estimated from a much larger population.

translation shared task is diverse, including output from a range of state-of-the-art systems. Over the years, WMT also has developed robust statistical methodology for collecting human judgments and significance testing of metric performance. For these reasons, the results from the metrics shared task are particularly strong. We highlight relevant methodological aspects in the following sections.

2.1 Direct Assessment

Beginning in WMT'17, the WMT human evaluation campaign has adopted *direct assessment* (DA; Graham et al., 2015) as the primary human evaluation. (Bojar et al., 2017b) DA scores are formulated on the law of large numbers - a sample mean of many human judgments is close to the population mean, which represents an intrinsic property of the translation quality.

However, we do not know the variance of human scores for a given translation, so an appropriate sample size n cannot be calculated. In addition, the variance in judgments can vary considerably across translations. To overcome this, DA uses a two step process: 1. Empirically determine the number of assessors n needed per translation for the desired consistency. 2. Collect n human judgments for remaining system outputs.

Refer to Figure 2. DA uses a continuous sliding bar from 1-100, which are averaged over many
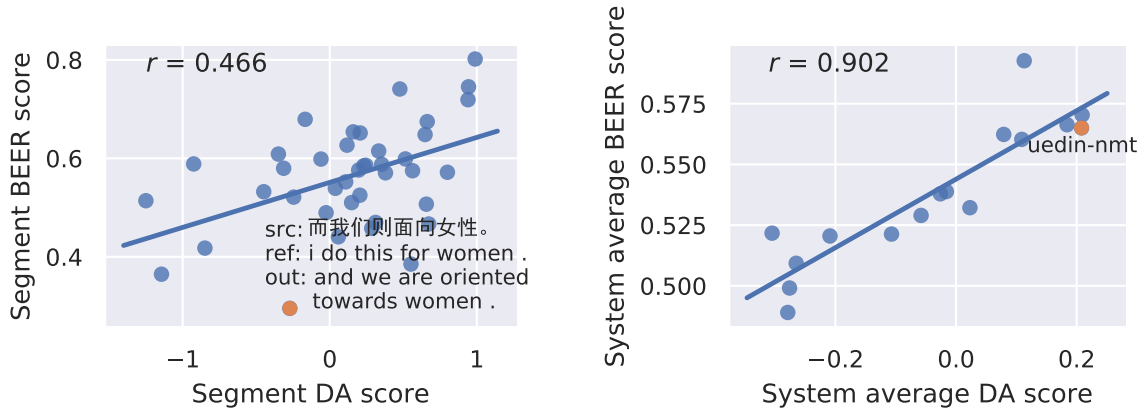
Figure 4: Sample Pearson r of segment/system-level BEER and DA scores from the WMT'17 metrics task for `zh-en` translation evaluation. **Left.** Each point denotes a test example from the metrics task. Examples are a subset from the `uedin-nmt` system (Sennrich et al., 2017). **Right.** Each point denotes a system participating in the `zh-en` translation task. DA and BEER scores have been averaged over all outputs for a given system.

workers for a sample mean. (Graham et al., 2013) Refer to Figure 3. In the first phase of DA, a large ($n = 100$) number of assessments are collected for a small number of output translations. Compute a simulated DA score for each segment with the first i judgments, and estimate a true mean with the remaining $n - i$ judgments. For the i where the correlation between DA scores and "true mean" is sufficient, gather i judgments for each segment in the full data collection.

Several aspects of DA have made it appealing for the WMT shared tasks. These points are stated in Bojar et al. (2017b): 1) DA score are more consistent than relative rankings of system output, which are often contradictory 2) The samples means from DA scores are absolute, facilitating comparisons across translations 3) Sliding bar judgments can be collected from crowdsourcing websites, and there are effective measures for quality control.

2.2 System and sentence-level correlation

Direct Assessment gives human judgment scores on a sentence (or segment) level. System level quality is then defined as the average of DA scores of a system's output over an entire test set, and is how the rankings for the annual WMT translation task are produced. (Bojar et al., 2018b; 2017b) Refer to Figure 4. Therefore, a metric can be evaluated on two correlations: at the segment-level - between segments' metric scores and DA scores, or system-level - between systems' average DA scores and aggregate metric scores (most

commonly a mean).

There are several notable distinctions between the two. For a metric to correlate highly on the system level does not imply it correlates on the segment level. In fact, in the MT domain, baselines such as BLEU have high correlations on the system-level for $\star$-en translation evaluation ($r > 0.9$), but have low segment-level correlations ($r < 0.5$, Ma et al., 2018). This is also seen for the SPICE (Anderson et al., 2016) metric in image captioning. Intuitively, a metric may only be competent in penalizing bad output, but cannot differentiate between average and good output (Refer to §4.1 for an analysis of BLEU). This causes the discrepancy in low correlation at the segment-level but high at the system-level. (Novikova et al., 2017; Chaganty et al., 2018)

In practice, system-level correlation is relevant. Research will often report results of system-level BLEU scores to show the effectiveness of a new system over baselines or existing literature. (Koehn, 2004) When making decisions about hyperparameters or model architectures, we rely on metric-produced system rankings. (Britz et al., 2017) However, existence of metrics with high segment-level correlations opens up research questions e.g. can we use such a metric as an alternative training objective? (Ranzato et al., 2015)
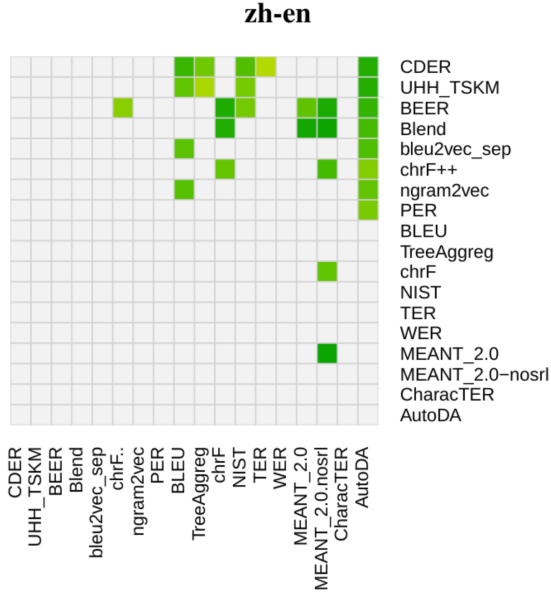
3

zh-en

Figure 5: p-values from the William's test on all pairs of metrics participating in system-level, `zh-en` translation evaluation in WMT'17. A green cell denotes ($p < 0.05$). Figure taken from (Bojar et al., 2017c). Significance in cell (i, j) denotes metric in row i significantly outperformed metric in column j.

2.3 Pearson correlation

The emerging consensus in WMT is the use of Pearson correlation[2] in segment and system-level evaluation of metrics. Given n paired data points $\{(x_1, y_1), ..., (x_n, y_n)\}$, the sample Pearson correlation is defined as:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

where $\overline{x}$ and $\overline{y}$ are the samples means for x_i and y_i, respectively. This correlation measures a linear association between two ordinal variables and ranges from $[-1, 1]$. In segment-level metric validation, (x_i, y_i) may represent the DA and metric scores of a system translation. On the system level, (x_i, y_i) may represent the average DA and aggregated metric scores of an MT system.

There are at least three reasons for motivation of the use of Pearson coefficient over Kendall's and Spearman. (These reasons are stated across Graham et al. 2015; 2014) 1) The linear measurement of Pearson correlation measure is more sensitive than rank-based correlations (i.e. Spearman), which measure any monotonic relationship. 2) Pearson correlations are absolute, facilitating

---

[2]https://libguides.library.kent.edu/SPSS/PearsonCorr

comparisons across testing settings. 3) Significance testing of differences in Pearson values are possible (see §2.4).

2.4 Significance tests for correlations

For the sample Pearson correlation r calculated between metric and DA scores, there are two significance tests that are often applied. The first test (which is *not* used in the WMT metrics task) calculates a p-value to reject the null hypothesis $H_0 : r = 0$. An unlikely null hypothesis means your metric has some true non-zero correlation. This is a common test, and p-values will be provided by most standard statistical packages.[3]

The WMT metrics task uses the William's test (Williams, 1959), which tests for the significance of differences in Pearson correlation. Concretely, a p-value is calculated to reject $H_0 :$ $Corr(X_1, X_3) = Corr(X_2, X_3)$, where X_1 and X_2 are correlated. In our case, X_1 and X_2 are generalizations of two different metric scores, and X_3 of DA scores. (Graham et al., 2015) An unlikely null hypothesis means that it is likely one metric is better than the other.

Refer to Figure 5. In the WMT metrics task, the winners are declared as those metrics which were not significantly outperformed by any other metric. In WMT'17 for `zh-en`, 9 different metrics were co-winners. (Bojar et al., 2017c, count the empty columns in Figure 5)

3 Conducting a validation study

We assume that a suitable NLG task has been chosen for evaluation. The section will outline considerations and statistical methodology step-by-step in your validation study. We synthesize from existing literature we read, and borrow heavily from the design of the WMT metrics shared task.

3.1 Collecting diverse system output

The results of your validation study will vary greatly based on testing conditions of the metrics. Conclusions generalize best to similar conditions in practice, and so it is important to cover as much ground as possible.

- If you are using Pearson correlation (recommended, refer to §2.2) to measure a metric's system-level correlation, choose at least five (5) systems. Five points is the least amount

---

[3]`cor.test` in R or `stats.pearsonr` in `scipy`.

of data points to make a statistically significant conclusion ($p < 0.05$) that a correlation is non-zero.[4] (Reiter, 2018)

- Produce outputs from a mix of baselines and state-of-the art systems using a variety of approaches. In MT, a few popular approaches may include transfer-based, statistical, and neural systems. It is known, for instance, that BLEU correlates poorly with rule-based systems, (Callison-Burch et al., 2006) and the exclusion of these systems have caused correlations between BLEU and human judgments to increase. (Reiter, 2018)

- If possible, include a few synthetic variations of a system, ideally variations seen in practice. (Callison-Burch et al., 2006) One such variation could be several identical models with ablated subsets of the training data.

- Report characteristics of the test set system output is elicited from. Metrics using linguistic resources (e.g. WordNet, parsers, taggers) will be sensitive to the language they are applied to. (Kilickaya et al., 2017)

3.2 Collecting consistent human judgments

Consistent human judgments are necessary to accurately evaluate metrics. Collect data in a manner that is replicable and publicly release the data.

- Choose a question to elicit intrinsic quality judgment from humans. Unfortunately, the choice of question is nearly art and considerations may be philosophical. (Gatt and Krahmer, 2018). However, a question that elicits consistent judgment saves effort in annotation, and may mean the question reflects a true intrinsic property. Note that BLEU was originally validated against human judgment of "general translation quality." (Papineni et al., 2002)

- Choose one intrinsic quality question. In WMT, aspects of adequacy and fluency were previously judged separately but later abandoned. (Bojar et al., 2016a) This is not ideal because two forms of results are confusing, and if you expect a metric to correlate with both aspects, only one question is needed.

- Direct assessment (§2.1) collects consistent human judgments, and can be crowdsourced

relatively hassle-free through Amazon Mechanical Turk. (Graham et al., 2015; Bojar et al., 2017b). This collection methodology has also been successfully applied NLG domains outside of MT. Refer to Graham et al. (2018) for an example in video captioning.

3.3 Producing automatic metric scores

- Use consistent tokenization across all metrics. Most metrics, especially n-gram based metrics, will be affected by tokenization. (Post, 2018)

- You will likely be using BLEU and/or sent-BLEU as a baseline. If so, note factors affecting scores (e.g. preprocessing, n-gram weighting, length penalty), and that SACREBLEU is an existing tool to manage these parameters. (Post, 2018)

3.4 Conducting significance tests

Applying robust statistical methodology will allow sound conclusion to be drawn about the performance of our metrics.

- Don't leave anything to chance - use the William's test (§2.4) to test for significance in increase of Pearson correlation. Graham and Baldwin (2014) provides an open source implementation at `https://github.com/ygraham/significance-williams`.

- Report both system and segment-level results with significant tests. (§2.2) You may also want to use Kendall's τ or Spearman's ρ as secondary metric evaluations. (§4)

4 Analyses of metrics in WMT'17

This section demonstrates several basic analysis methods of metric scoring on both the segment and system level. We will exclusively perform these analyses on BLEU and its sentence-level variant sentence-BLEU[5], (Koehn et al., 2007) for demonstration and for insight in a widely applied baseline metric.

4.1 Segment-level analysis: metric score distributions conditional on DA

To understand how your metric is scoring with respect to translations of different quality, you may visualize the distributions of your metric score on

---

[4]`https://ehudreiter.com/2018/07/10/how-to-validate-metrics/`

[5]`https://github.com/moses-smt/mosesdecoder/blob/master/mert/sentence-bleu.cpp`

Bin	Source	Reference	Output	DA	sentence-BLEU
good	如果两家工厂关闭，则电力市场的需求量会大大减少。	if both of those plants go from the market that 's a significant reduction in demand in the [electricity] market .	if the two factories closed , the power market demand will be greatly reduced .	0.355	0.066
	上次产油国召开会议已经是4月份的事情，OPEC成员国未能就任何措施达成协议。	it was april when the oil-producing countries had meeting . no agreement was reached among the opec member ...	the last time oil producers are meeting in april , opec member states that have failed to agree on any measures to reach an ...	0.247	0.079
	小鹏和吴言在一个大群里相识，因为两人都是积极发言者，很快就熟络起来了。	xiaopeng and wuyan meet each other in the group chatting . they two get familiar with each other because they are active ...	xiao peng and wu met in a large group , because both were active speakers and soon became familiar .	0.537	0.087
bad	朗兹曼写道。	lanzmann wrote .	朗兹曼 wrote .	-1.245	0.707
	云南省首届青运会13日开赛开幕式打民族牌展示青春活力	the opening ceremony of the first youth games of yunnan province on the 13th day of the month showed youth vitality	the opening ceremony of the opening ceremony of the first olympic games of yunnan province on the opening ceremony of yunnan province	-1.081	0.481
	远在千里之外，巨嘴鸟格雷西亚的故事感动了著名纪录片导演葆拉·埃雷迪亚和探索新闻频道制片人约翰·霍夫曼。	thousands of miles away , the story of the toucan gracia touched the famous documentary director paula el reidia and the news channel producer john hoffman .	far more than thousands of thousands , the story of greecia is moved by the famous documentary director paula mareda and the exploration of the producer of news ...	-0.776	0.282

Table 1: Selections of lowest scoring BLEU examples in the "good" bin and highest scoring BLEU examples in the "bad" bin for `zh-en` translation evaluation in WMT'17. Examples are binned by DA scores, and chosen from the top/bottom 10. Some words have been elided. Readers are encouraged to view other examples on our notebook.
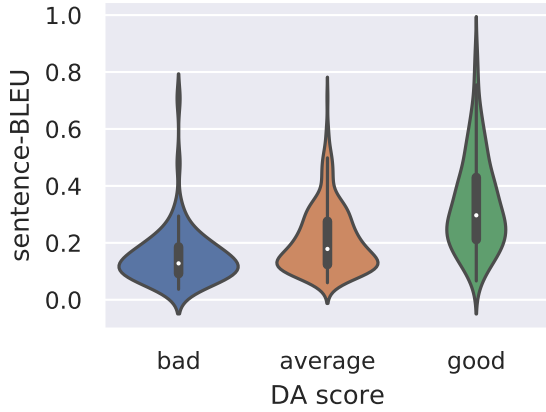

Figure 6: Distribution of sentence-BLEU scores conditional on bin of DA score for `zh-en` translation evaluation in WMT'17. The three intervals for binning are $(-1.73, -0.749] < (-0.749, 0.228] < (0.228, 1.205]$. Every bin has a near equal number of points.

"bins" of DA scores. Start by partitioning your examples by DA scores into a lower one-third ($\frac{1}{3}$) DA scores bin (bad translations), a medium one-third bin (average), and higher one-third bin (good). Each bin should have a near equal number of points. Produce violin plots of the conditional distributions as in Chaganty et al. (2018). Examining conditional correlations within bins *does not* provide the same information!

Refer to Figure 6. We see that sentence-BLEU has significant overlap in scoring bad and average translations, but not the good translation. Our interpretation is that sentence-BLEU mainly dif-

ferentiates good translations well. Points that lay high on the x-axis (high n-gram overlap) will almost always be good translations, as the reference and system output will be nearly exact matches. In `*-en` translation, this strategy is effective to achieve high system-level correlation.

4.2 Segment-level analysis: qualitative analysis on metric failure cases

One of the best ways to understand your metric is to examine failure cases - cases in which there is large disagreement between human DA scores and your metric score. We have found this analysis to be insightful even looking at 5-10 examples. After binning examples into three bins by DA score (as in §4.1), sort the "good" bin in ascending order, and the "bad" bin in descending order. You may also want to produce some features of your metric for comparison. Qualitative analysis has been used in Tao et al. (2018).

Refer to Table 1. One of the biggest shortcomings of BLEU is the lack of respect to the semantic content. In the good system outputs that received low BLEU scores, we see several meaning preserving paraphrases of the reference. However, we note that these instances are quite rare (observe Figure 6). In the bad output examples, there are at least two failure cases of BLEU: matching some words in really short sentences and repeating long phrases artificially inflates the BLEU score.
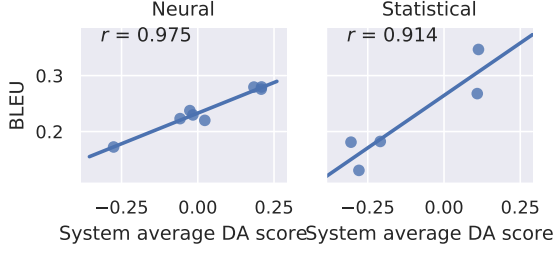
Figure 7: Sample Pearson correlation of BLEU with DA conditional on system type for `zh-en` translation evaluation in WMT'17. System types were classified by hand. Four out of five statistical systems were anonymized online translation systems.
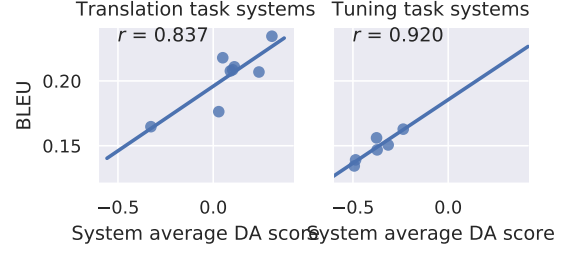

Figure 8: Sample Pearson correlation of BLEU with DA conditional on submission track for `en-cs` translation evaluation in WMT'17. All tuning task systems are instances of Neural Monkey. (Helcl and Libovický, 2017) The translation task includes different systems.

	n	$|r|$	$|\tau|$
`cs-en`	4	0.971	1.000
`de-en`	11	0.923	0.564
`fi-en`	6	0.903	0.867
`lv-en`	9	0.979	0.833
`ru-en`	9	0.912	0.778
`tr-en`	10	0.976	0.911
`zh-en`	16	0.864	0.767

Table 2: Sample Pearson's r and Kendall's τ correlations between BLEU and DA at the system level.

4.3 System-level analysis: metric correlation with DA conditional on system type

The correlation of a metric can vary greatly depending on the diversity of systems it is validated on. As shown in Callison-Burch et al. (2006), your correlation may decrease when validated against a diverse range of systems. In practice, if we know that a metric has weak correlation for comparing systems with different approaches, we may want to constrain metric use to comparing systems using the same approach (e.g. neural).

Refer to Figure 7. The correlation for neural systems is higher than that of the statistical systems. These recent switch from statistical to neural MT systems is a likely factor in Reiter (2018) observing human-BLEU correlation increasing over time. When comparing BLEU scores, it is more effective to compare neural systems to other neural systems. In validating our metrics, we must choose a mix of possible approaches to better understand our correlation.

Refer to Figure 8. The WMT'17 tuning task (Bojar et al., 2017d) calls for participants to tune the same neural MT model (Neural Monkey; Helcl and Libovický, 2017) for `en-cs` translation with varying training settings e.g. curriculum learning etc. Intuitively, models of the same architecture are likely to make similar mistakes, on average, over the same test set, so are penalized by humans equally. This causes the spread (or residuals) of the best-fit-line to be much tighter.

4.4 System-level analysis: Kendall's τ coefficient and its interpretation

In this section we propose the use of Kendall's τ as a secondary evaluation metric on the system level. While this coefficient does not have an appropriate significance test in our setting, it has an intuitive interpretation. For paired data points $\{(x_1, y_1), ..., (x_n, y_n)\}$ Kendall's τ is a rank-based correlation coefficient defined as

$$\frac{\text{(Concordant pairs)-(Discordant pairs)}}{\text{(Concordant pairs)+(Discordant pairs)}}$$

where the number of concordant pairs are the number of pairs such that $x_i < x_j \wedge y_i < y_j$ or $x_i > x_j \wedge y_i > y_j$ for all $1 \le i \le j \le n$. Pairs are discordant when there is disagreement in rank between the two variables. Kendall's τ is then a difference in the percentage (normalized over total pairs) of concordant and discordant pairs.

Intuitively, x_i and y_i represent the BLEU and DA scores. Refer to Table 2. Kendall's τ then tells us the percentage difference that BLEU would agree over disagree with DA pairwise judgment on system quality. For instance in `zh-en`, BLEU agrees more than disagrees 86 percentage points with DA for pairwise judgments. We leave the interpretation of correlation metrics more relevant in practice as future work.

7

5 Relevant work: metrics

This section attempts to make a selection of influential work that overviews the metrics development literature. Besides the well-known BLEU metric, other metrics primarily using n-gram features include the NIST (Doddington, 2002) metric for translation, which is similar to BLEU but weighs n-grams based on their rarity, and ROUGE (Lin, 2004) for summarization. Character-level n-gram features have also been proposed to capture subword information. The CHRF metric (Popovic, 2015) calculate F-scores based on character-level n-grams. CDER and BEER (Stanojevic and Sima'an, 2014) also use character features.

Several translation metrics calculate scores based on alignments of extracted surface linguistic features between hypotheses and references. METEOR (Denkowski and Lavie, 2014) uses alignments based on exact, stem, synonym, and paraphrase matches between words and phrases.[6] MEANT (Lo, 2017) evaluates translations by first aligning semantic frames and then aligning role fillers. Both MEANT and METEOR incorporate linguistic resources such as WordNet and semantic role parsers, respectively, and their incorporation, including other character-level and shallow linguistic approaches, have shown benefits in the WMT metrics task. (Bojar et al., 2016b)

More recently, both formal and distributed semantic representations have found success in metrics, even in NLG domains outside of translation. SPICE (Anderson et al., 2016) is an image captioning metric based on semantic parsing, where scores are based on overlap of semantic propositions in the graph meaning representations of both the hypothesis and reference captions. The RU-BER metric (Tao et al., 2018) for dialogue uses cosine similarities of word embeddings to predict response appropriateness, and a neural model trained with negative sampling to predict relevancy. The first metric in WMT'18 to use sentence embeddings, RUSE (Shimanaka et al., 2018) scores with a trained neural regression model on sentence embeddings and is the highest performing metric for $\star$-en translation.

Finally, we will highlight MT quality estimation systems (QE, evaluating how good a translation is without the reference) which includes both neural and grammar-based approaches. Recurrent

---

[6] http://www.cs.cmu.edu/~alavie/METEOR/

---

neural networks encoding both the source and MT output have been used for regression on human quality scores. (Dusek et al., 2017) The WMT quality estimation shared task (Specia et al., 2018) includes examples of many such systems, with a winning submission using neural models. (Wang et al., 2018) Approaches based on grammatical error correction systems are reference-less and have also been proposed for QE. (Napoles et al., 2016)

6 Conclusion

The authors believe a metric that applies across multiple NLG domains may be widely adopted. This metric might focus only on semantic similarity (with a reference), and be complemented with another domain-specific measure e.g. style preservation for style transfer, (Fu et al., 2018) summary length for summarization, or relevancy for dialogue. (Tao et al., 2018) To compute semantic similarity, a model may use meaning representations and parsing methods (Konstas et al., 2017) or continuous sentence/word embeddings, (Peters et al., 2018; Conneau et al., 2017) as these technologies are becoming more widespread and powerful in NLP.

For a few NLG tasks e.g. MT, image captioning, summarization, achieving usable system-level correlation is not impossible. In some cases, it can be argued that achieving usable system-level correlation is not difficult, as merely correctly differentiating good output is sufficient. (§4.1) However, segment-level correlation leaves much to be desired. The existence of metrics with high segment-level correlation opens up exciting research directions. Such a metric can also be useful in practice, as it can identify model failure modes, or detect low quality output and fallback on rule-based models as needed.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: semantic propositional image caption evaluation. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*, volume 9909 of *Lecture Notes in Computer Science*, pages 382–398. Springer.

Ondrej Bojar, Christian Buck, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno-Yepes, Philipp Koehn, and Julia Kreutzer, editors. 2017a. *Proceedings of the Second Conference on Machine Transla-*

tion, WMT 2017, Copenhagen, Denmark, September 7-8, 2017. Association for Computational Linguistics.

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno-Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana L. Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors. 2018a. *Proceedings of the Third Conference on Machine Translation: Shared Task Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*. Association for Computational Linguistics.

Ondrej 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. 2017b. Findings of the 2017 conference on machine translation (WMT17). In (Bojar et al., 2017a), pages 169–214.

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno-Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana L. Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin M. Verspoor, and Marcos Zampieri. 2016a. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation, WMT 2016, colocated with ACL 2016, August 11-12, Berlin, Germany*, pages 131–198. The Association for Computer Linguistics.

Ondrej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018b. Findings of the 2018 conference on machine translation (WMT18). In (Bojar et al., 2018a), pages 272–303.

Ondrej Bojar, Yvette Graham, and Amir Kamran. 2017c. Results of the WMT17 metrics shared task. In (Bojar et al., 2017a), pages 489–513.

Ondrej Bojar, Jindrich Helcl, Tom Kocmi, Jindrich Libovický, and Tomás Musil. 2017d. Results of the WMT17 neural MT training task. In (Bojar et al., 2017a), pages 525–533.

Ondřej Bojar, Christian Federmann, Barry Haddow, Philipp Koehn, and Lucia Specia Matt Post and. 2016b. Ten years of wmt evaluation campaigns: Lessons learnt. In *Workshop on Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem*, pages 27–36, Portoroz, Slovenia.

Denny Britz, Anna Goldie, Minh-Thang Luong, and Quoc Le. 2017. Massive exploration of neural machine translation architectures. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1442–1451. Association for Computational Linguistics.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluation the role of bleu in machine translation research. In *EACL 2006, 11st Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, April 3-7, 2006, Trento, Italy*. The Association for Computer Linguistics.

Arun Tejasvi Chaganty, Stephen Mussmann, and Percy Liang. 2018. The price of debiasing automatic metrics in natural language evalaution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 643–653. Association for Computational Linguistics.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In (Palmer et al., 2017), pages 670–680.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

George Doddington. 2002. 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, pages 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Ondrej Dusek, Jekaterina Novikova, and Verena Rieser. 2017. Referenceless quality estimation for natural language generation. *CoRR*, abs/1708.01759.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In (McIlraith and Weinberger, 2018), pages 663–670.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *J. Artif. Intell. Res.*, 61:65–170.

Dimitra Gkatzia and Saad Mahamood. 2015. A snapshot of NLG evaluation practices 2005 - 2014. In *ENLG 2015 - Proceedings of the 15th European Workshop on Natural Language Generation, 10-11 September 2015, University of Brighton, Brighton, UK*, pages 57–60. The Association for Computer Linguistics.

Yvette Graham, George Awad, and Alan Smeaton. 2018. Evaluation of automatic video captioning using direct assessment. *PLOS ONE*, 13(9):1–20.

Yvette Graham and Timothy Baldwin. 2014. Testing for significance of increased correlation with human judgment. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 172–176. ACL.

Yvette Graham, Timothy Baldwin, and Nitika Mathur. 2015. Accurate evaluation of segment-level machine translation metrics. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 1183–1191. The Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, LAW-ID@ACL 2013, August 8-9, 2013, Sofia, Bulgaria*, pages 33–41. The Association for Computer Linguistics.

Jindřich Helcl and Jindřich Libovický. 2017. Neural Monkey: An Open-source Tool for Sequence Learning. *The Prague Bulletin of Mathematical Linguistics*, (107):5–17.

Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. 2017. Re-evaluating automatic metrics for image captioning. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*, pages 199–209. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing , EMNLP 2004, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004, 25-26 July 2004, Barcelona, Spain*, pages 388–395. ACL.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ioannis Konstas, Srinivasan Iyer, Mark Yatskar, Yejin Choi, and Luke Zettlemoyer. 2017. Neural AMR: sequence-to-sequence models for parsing and generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 146–157. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: a package for automatic evaluation of summaries.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael 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 (Su et al., 2016), pages 2122–2132.

Chi-kiu Lo. 2017. MEANT 2.0: Accurate semantic MT evaluation for any output language. In (Bojar et al., 2017a), pages 589–597.

Qingsong Ma, Ondrej Bojar, and Yvette Graham. 2018. Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In (Bojar et al., 2018a), pages 671–688.

Sheila A. McIlraith and Kilian Q. Weinberger, editors. 2018. *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*. AAAI Press.

Courtney Napoles, Keisuke Sakaguchi, and Joel R. Tetreault. 2016. There's no comparison: Reference-less evaluation metrics in grammatical error correction. In (Su et al., 2016), pages 2109–2115.

Jekaterina Novikova, Ondrej Dusek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In (Palmer et al., 2017), pages 2241–2252.

Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors. 2017. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA.*, pages 311–318. ACL.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Maja Popovic. 2015. chrf: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation, WMT@EMNLP 2015, 17-18 September 2015, Lisbon, Portugal*, pages 392–395. The Association for Computer Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 186–191. Association for Computational Linguistics.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. *CoRR*, abs/1511.06732.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3).

Ehud Reiter, Roma Robertson, and Liesl Osman. 2003. Lessons from a failure: Generating tailored smoking cessation letters. *Artif. Intell.*, 144(1-2):41–58.

Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. 2017. The university of edinburgh's neural MT systems for WMT17. In (Bojar et al., 2017a), pages 389–399.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. RUSE: regressor using sentence embeddings for automatic machine translation evaluation. In (Bojar et al., 2018a), pages 751–758.

Lucia Specia, Frédéric Blain, Varvara Logacheva, Ramón Fernández Astudillo, and André F. T. Martins. 2018. Findings of the WMT 2018 shared task on quality estimation. In (Bojar et al., 2018a), pages 689–709.

Milos Stanojevic and Khalil Sima'an. 2014. BEER: better evaluation as ranking. In *Proceedings of the Ninth Workshop on Statistical Machine Translation, WMT@ACL 2014, June 26-27, 2014, Baltimore, Maryland, USA*, pages 414–419. The Association for Computer Linguistics.

Jian Su, Xavier Carreras, and Kevin Duh, editors. 2016. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*. The Association for Computational Linguistics.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. RUBER: an unsupervised method for automatic evaluation of open-domain dialog systems. In (McIlraith and Weinberger, 2018), pages 722–729.

Oriol Vinyals and Quoc V. Le. 2015. A neural conversational model. *CoRR*, abs/1506.05869.

Jiayi Wang, Kai Fan, Bo Li, Fengming Zhou, Boxing Chen, Yangbin Shi, and Luo Si. 2018. Alibaba submission for WMT18 quality estimation task. In (Bojar et al., 2018a), pages 809–815.

E.J. Williams. 1959. *Regression analysis*. Wiley series in probability and mathematical statistics. Probability and mathematical statistics. Wiley.

Lili Yao, Nanyun Peng, Ralph M. Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2018. Plan-and-write: Towards better automatic storytelling. *CoRR*, abs/1811.05701.