Testing for Significance of Increased Correlation with Human Judgment

Yvette Graham **Timothy Baldwin**
Department of Computing and Information Systems
The University of Melbourne
graham.yvette@gmail.com, tb@ldwin.net

Abstract

Automatic metrics are widely used in machine translation as a substitute for human assessment. With the introduction of any new metric comes the question of just how well that metric mimics human assessment of translation quality. This is often measured by correlation with human judgment. Significance tests are generally not used to establish whether improvements over existing methods such as BLEU are statistically significant or have occurred simply by chance, however. In this paper, we introduce a significance test for comparing correlations of two metrics, along with an open-source implementation of the test. When applied to a range of metrics across seven language pairs, tests show that for a high proportion of metrics, there is insufficient evidence to conclude significant improvement over BLEU.

1 Introduction

Within machine translation (MT), efforts are ongoing to improve evaluation metrics and find better ways to automatically assess translation quality. The process of validating a new metric involves demonstration that it correlates better with human judgment than a standard metric such as BLEU (Papineni et al., 2001). However, although it is standard practice in MT evaluation to measure increases in automatic metric scores with significance tests (Germann, 2003; Och, 2003; Kumar and Byrne, 2004; Koehn, 2004; Riezler and Maxwell, 2005; Graham et al., 2014), this has not been the case in papers proposing new metrics. Thus it is possible that some reported improvements in correlation with human judgment are attributable to chance rather than a systematic improvement.

In this paper, we motivate and introduce a novel significance test to assess the statistical significance of differences in correlation with human judgment for pairs of automatic metrics. We apply tests to the WMT-12 shared metrics task to compare each of the participating methods, and find that for a high proportion of metrics, there is not enough evidence to conclude that they significantly outperform BLEU.

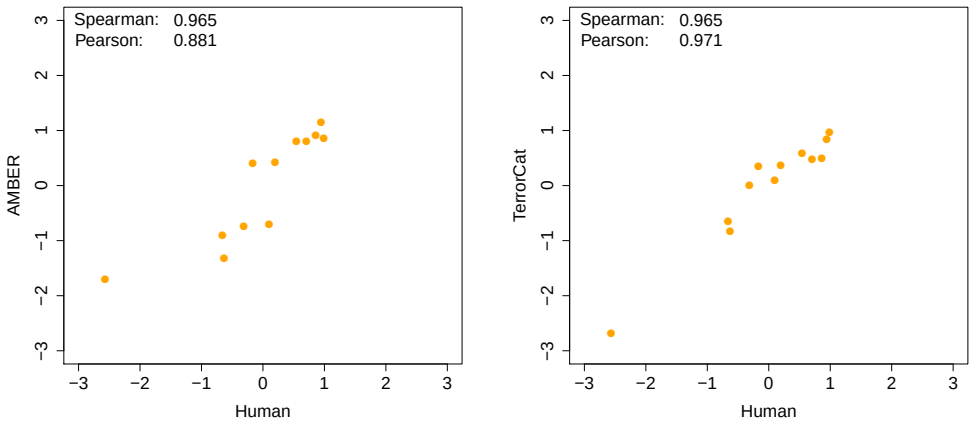
2 Correlation with Human Judgment

A common means of assessing automatic MT evaluation metrics is Spearman's rank correlation with human judgments (Melamed et al., 2003), which measures the relative degree of monotonicity between the metric and human scores in the range $[-1, 1]$. The standard justification for calculating correlations over ranks rather than raw scores is to: (a) reduce anomalies due to absolute score differences; and (b) focus evaluation on what is generally the primary area of interest, namely the ranking of systems/translations.

An alternative means of evaluation is Pearson's correlation, which measures the linear correlation between a metric and human scores (Leusch et al., 2003). Debate on the relative merits of Spearman's and Pearson's correlation for the evaluation of automatic metrics is ongoing, but there is an increasing trend towards Pearson's correlation, e.g. in the recent WMT-14 shared metrics task.

Figure 1 presents the system-level results for two evaluation metrics – AMBER (Chen et al., 2012) and TERRORCAT (Fishel et al., 2012) – over the WMT-12 Spanish-to-English metrics task. These two metrics achieved the joint-highest rank correlation ($\rho = 0.965$) for the task, but differ greatly in terms of Pearson's correlation ($r = 0.881$ vs. 0.971, resp.). The largest contributor to this artifact is the system with the lowest human score, represented by the leftmost point in both plots.

Graham, Yvette and Timothy Baldwin (to appear) Testing for Significance of Increased Correlation with Human Judgment,
In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), Doha, Qatar.



(a) AMBER (b) TERRORCAT

Figure 1: Scatter plot of human and automatic scores of WMT-12 Spanish-to-English systems for two MT evaluation metrics (AMBER and TERRORCAT)

Consistent with the WMT-14 metrics shared task, we argue that Pearson's correlation is more sensitive than Spearman's correlation. There is still the question, however, of whether an observed difference in Pearson's r is statistically significant, which we address in the next section.

3 Significance Testing

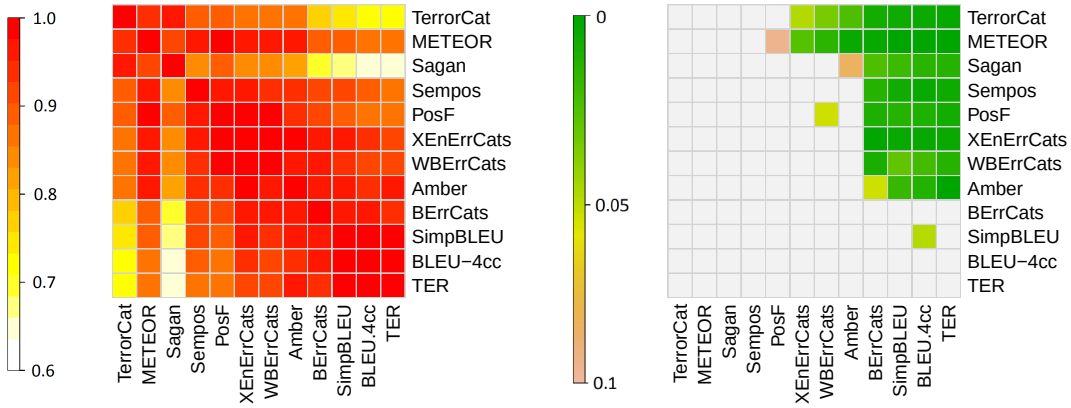
Evaluation of a new automatic metric, M_{new}, commonly takes the form of quantifying the correlation between the new metric and human judgment, $r(M_{new}, H)$, and contrasting it with the correlation for some baseline metric, $r(M_{base}, H)$. It is very rare in the MT literature for significance testing to be performed in such cases, however. We introduce a statistical test which can be used for this purpose, and apply the test to the evaluation of metrics participating in the WMT-12 metric evaluation task.

At first gloss, it might seem reasonable to perform significance testing in the following manner when an increase in correlation with human assessment is observed: apply a significance test separately to the correlation of each metric with human judgment, with the hope that the newly proposed metric will achieve a significant correlation where the baseline metric does not. However, besides the fact that the correlation between almost any document-level metric and human judgment will generally be significantly greater than zero, the logic here is flawed: the fact that one correlation is significantly higher than zero ($r(M_{new}, H)$) and that of another is not, does not necessarily mean that the *difference* between the two correlations is significant. Instead, a specific test should be applied to the difference in correlations on the data. For this same reason, confidence intervals for individual correlations with human judgment are also not particularly meaningful.

In psychological studies, it is often the case that samples that data are drawn from are independent, and differences in correlations are computed on independent data sets. In such cases, the Fisher r to z transformation is applied to test for significant differences in correlations. In the case of automatic metric evaluation, however, the data sets used are almost never independent. This means that if $r(M_{base}, H)$ and $r(M_{new}, H)$ are both > 0, the correlation between the metric scores themselves, $r(M_{base}, M_{new})$, must also be > 0. The strength of this correlation, directly between pairs of metrics, should be taken into account using a significance test of the difference in correlation between $r(M_{base}, H)$ and $r(M_{new}, H)$.

3.1 Correlated Correlations

Correlations computed for two separate automatic metrics on the same data set are not independent, and for this reason in order to test the difference in correlation between them, the degree to which the pair of metrics correlate with each other should be taken into account. The Williams test (Williams,



(a) Pearson's correlation (b) Statistical significance

Figure 2: (a) Pearson's correlation between pairs of automatic metrics; and (b) p-value of Williams significance tests, where a colored cell in row i (named on y-axis), col j indicates that metric i (named on x-axis) correlates significantly higher with human judgment than metric j; all results are based on the WMT-12 Spanish-to-English data set.

1959)[1] evaluates significance in a difference in dependent correlations (Steiger, 1980). It is formulated as follows, as a test of whether the population correlation between X_1 and X_3 equals the population correlation between X_2 and X_3:

$$t(n-3) = \frac{(r_{13} - r_{23})\sqrt{(n-1)(1+r_{12})}}{\sqrt{2K\frac{(n-1)}{(n-3)} + \frac{(r_{23}+r_{13})^2}{4}(1-r_{12})^3}},$$

where r_{ij} is the Pearson correlation between X_i and X_j, n is the size of the population, and:

$$K = 1 - {r_{12}}^2 - {r_{13}}^2 - {r_{23}}^2 + 2r_{12}r_{13}r_{23}$$

The Williams test is more powerful than the equivalent for independent samples (Fisher r to z), as it takes the correlations between X_1 and X_2 (metric scores) into account. All else being equal, the higher the correlation between the metric scores, the greater the statistical power of the test.

4 Evaluation and Discussion

Figure 2a is a heatmap of the degree to which automatic metrics correlate with one another when computed on the same data set, in the form of the Pearson's correlation between each pair of metrics that participated in the WMT-12 metrics task for Spanish-to-English evaluation. Metrics are ordered in all tables from highest to lowest *correlation with human assessment*. In addition, for the purposes of significance testing, we take the absolute value of all correlations, in order to compare error-based metrics with non-error based ones.

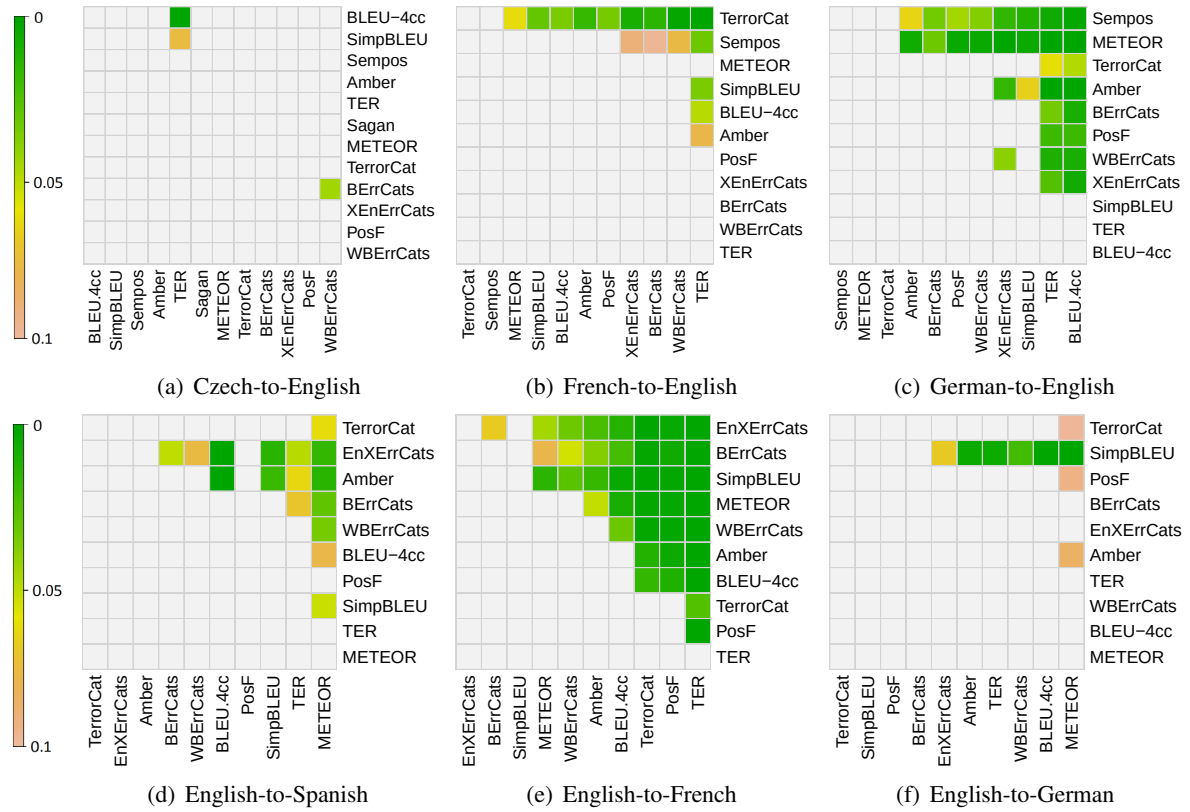
In general, the correlation is high amongst all pairs of metrics, with a high proportion of paired metrics achieving a correlation in excess of $r = 0.9$. Two exceptions to this are TERRORCAT (Fishel et al., 2012) and SAGAN (Castillo and Estrella, 2012), as seen in the regions of yellow and white.

Figure 2b shows the results of Williams significance tests for all pairs of metrics. Since we are interested in not only identifying significant differences in correlations, but ultimately ranking competing metrics, we use a one-sided test. Here again, the metrics are ordered from highest to lowest (absolute) correlation with human judgment.

For the Spanish-to-English systems, approximately 60% of WMT-12 metric pairs show a significant difference in correlation with human judgment at $p < 0.05$ (for one of the two metric directions).[2] As expected, the higher the correlation with human judgment, the more metrics a given method is superior to at a level of statistical significance. Although TERRORCAT (Fishel et al., 2012) achieves the highest absolute correlation with human judgment, it is not significantly better ($p \geq 0.05$) than the four next-best metrics (METEOR (Denkowski and Lavie, 2011), SAGAN (Castillo and Estrella, 2012), SEMPOS (Macháček and Bo-

[1] Also sometimes referred to as the Hotelling–Williams test.

[2] Correlation matrices (red) are maximally filled, in contrast to one-sided significance test matrices (green), where, at a maximum, fewer than half of the cells can be filled.



(a) Czech-to-English (b) French-to-English (c) German-to-English

(d) English-to-Spanish (e) English-to-French (f) English-to-German

Figure 3: Significance results for pairs of automatic metrics for each WMT-12 language pair.

jar, 2011) and POSF (Popovic, 2012)). There is not enough evidence to conclude, therefore, that this metric is any better at evaluating Spanish-to-English MT system quality than the next four metrics.

Figure 3 shows the results of significance tests for the six other language pairs used in the WMT-12 metrics shared task.[3] For no language pair is there an outright winner amongst the metrics, with proportions of significant differences between metrics for a given language pair ranging from 3% for Czech-to-English to 82% for English-to-French ($p < 0.05$). The number of metrics that significantly outperform BLEU for a given language pair is only 34% ($p < 0.05$), and no method significantly outperforms BLEU over all language pairs – indeed, even the best methods achieve statistical significance over BLEU for only a small minority of language pairs. This underlines the dangers of assessing metrics based solely on correlation numbers, and emphasizes the importance of statistical testing.

It is important to note that the number of competing metrics a metric significantly outperforms should not be used as the criterion for ranking competing metrics. This is due to the fact that the power of the Williams test to identify significant differences between correlations changes depending on the degree to which the pair of metrics correlate with each other. Therefore, a metric that happens to correlate strongly with many other metrics would be at an unfair advantage, were numbers of significant wins to be used to rank metrics. For this reason, it is best to interpret pairwise metric tests in isolation.

As part of this research, we have made available an open-source implementation of statistical tests tailored to the assessment of MT metrics available at `https://github.com/ygraham/significance-williams`.

5 Conclusions

We have provided an analysis of current methodologies for evaluating automatic metrics in machine translation, and identified an issue with respect to the lack of significance testing. We introduced the Williams test as a means of calculating the statistical significance of differences

[3] We omit English-to-Czech due to some metric scores being omitted from the WMT-12 data set.

in correlations for dependent samples. Analysis of statistical significance in the WMT-12 metrics shared task showed there is currently insufficient evidence for a high proportion of metrics to conclude that they outperform BLEU.

Acknowledgments

We wish to thank the anonymous reviewers for their valuable comments. This research was supported by funding from the Australian Research Council.

References

Julio Castillo and Paula Estrella. 2012. Semantic textual similarity for MT evaluation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 52–58, Montréal, Canada.

Boxing Chen, Roland Kuhn, and George Foster. 2012. Improving AMBER, an MT evaluation metric. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 59–63, Montréal, Canada.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 85–91, Edinburgh, UK.

Mark Fishel, Rico Sennrich, Maja Popović, and Ondřej Bojar. 2012. TerrorCat: a translation error categorization-based MT quality metric. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 64–70, Montréal, Canada.

Ulrich Germann. 2003. Greedy decoding for statistical machine translation in almost linear time. In *Proceedings of the 2003 Conference of the North American Chapter of the Assoc. Computational Linguistics on Human Language Technology-Volume 1*, pages 1–8, Edmonton, Canada.

Yvette Graham, Nitika Mathur, and Timothy Baldwin. 2014. Randomized significance tests in machine translation. In *Proceedings of the ACL 2014 Ninth Workshop on Statistical Machine Translation*, pages 266–274, Baltimore, USA.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of Empirical Methods in Natural Language Processing 2004 (EMNLP 2004)*, pages 388–395, Barcelona, Spain.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the 4th International Conference on Human Language Technology Research and 5th Annual Meeting of the NAACL (HLT-NAACL 2004)*, pages 169–176, Boston, USA.

Gregor Leusch, Nicola Ueffing, and Hermann Ney. 2003. A novel string-to-string distance measure with applications to machine translation evaluation. In *Proceedings 9th Machine Translation Summit (MT Summit IX)*, pages 240–247, New Orleans, USA.

Matouš Macháček and Ondřej Bojar. 2011. Approximating a deep-syntactic metric for MT evaluation and tuning. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 92–98, Edinburgh, UK.

Dan Melamed, Ryan Green, and Joseph Turian. 2003. Precision and recall of machine translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL 2003) — Short Papers*, pages 61–63, Edmonton, Canada.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: A method for automatic evaluation of machine translation. Technical Report RC22176 (W0109-022), IBM Research, Thomas J. Watson Research Center.

Maja Popovic. 2012. Class error rates for evaluation of machine translation output. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 71–75, Montréal, Canada.

Stefan Riezler and John T. Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for mt. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 57–64, Ann Arbor, USA.

James H. Steiger. 1980. Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, 87(2):245.

Evan J. Williams. 1959. *Regression Analysis*, volume 14. Wiley, New York, USA.