Improving Evaluation of Machine Translation Quality Estimation

Yvette Graham ADAPT Centre School of Computer Science and Statistics Trinity College Dublin yvette.graham@scss.tcd.ie

Abstract

Quality estimation evaluation commonly takes the form of measurement of the error that exists between predictions and gold standard labels for a particular test set of translations. Issues can arise during comparison of quality estimation prediction score distributions and gold label distributions, however. In this paper, we provide an analysis of methods of comparison and identify areas of concern with respect to widely used measures, such as the ability to gain by prediction of aggregate statistics specific to gold label distributions or by optimally conservative variance in prediction score distributions. As an alternative, we propose the use of the unit-free Pearson correlation, in addition to providing an appropriate method of significance testing improvements over a baseline. Components of WMT-13 and WMT-14 quality estimation shared tasks are replicated to reveal substantially increased conclusivity in system rankings, including identification of outright winners of tasks.

1 Introduction

Machine Translation (MT) Quality Estimation (QE) is the automatic prediction of machine translation quality without the use of reference translations (Blatz et al., 2004; Specia et al., 2009). Human assessment of translation quality in theory provides the most meaningful evaluation of systems, but human assessors are known to be inconsistent and this causes challenges for quality estimation evaluation. For instance, there is a general lack of consensus both with respect to what provides the most meaningful gold standard representation, as well as best method of comparison of gold labels and system predictions. For example, in the 2014 Workshop on Statistical Machine Translation (WMT), which since 2012 has provided a main venue for evaluation of systems, sentence-level systems were evaluated with respect to three distinct gold standard representations and each of those compared to predictions using four different measures, resulting in a total of 12 different system rankings, 6 identified as official rankings (Bojar et al., 2014).

Although the aim of several methods of evaluation is to provide more insight into performance of systems, this also produces conflicting results and raises the question which method of evaluation really identifies the system(s) or method(s) that best predicts translation quality. For example, an extreme case in WMT-14 occurred for sentence-level quality estimation for English-to-Spanish. In each of the 12 system rankings, many systems were tied and this resulted in a total of 22 official winning systems for this language pair. Besides leaving potential users of quality estimation systems at a loss as to what the best system may be, a large number of inconclusive evaluation methodologies is also likely to lead to confusion about which evaluation methods should be applied in general in QE research, or worse still, researchers simply choosing the methodology that favors their system from among the many different methodologies.

In this paper, we provide an analysis of each of the methodologies used in WMT and widely applied to evaluation of quality estimation systems in general. Our analysis reveals potential flaws in existing methods and we subsequently provide detail of a single method that overcomes previous challenges. To demonstrate, we replicate com-

1804

ponents of evaluations previously carried out at WMT-13 and WMT-14 sentence-level quality estimation shared tasks. Results reveal substantially more conclusive system rankings, revealing outright winners that had not previously been identified.

2 Relevant Work

The Workshop on Statistical Machine Translation (WMT) provides a main venue for evaluation of quality estimation systems, in addition to the rare and highly-valued effort of provision of publicly available data sets to facilitate further research. We provide an analysis of current evaluation methodologies applied not only in the most recent WMT shared task but also widely within quality estimation research.

2.1 WMT-style Evaluation

WMT-14 quality estimation evaluation at the sentence-level, Task 1, is comprised of three subtasks. In Task 1.1, human gold labels comprise three levels of translation quality or "perceived post-edit effort" (1 = perfect translation; 2 = near miss translation; 3 = very low quality translation). A possible downside of the evaluation methodology applied in Task 1.1 is firstly that the gold standard representation may be overly coarse-grained. Considering the vast range of possible errors occurring in translations, limiting the levels of translation quality to only three may impact negatively on systems' ability to discriminate between translations of various quality.

More importantly, however, the combination of such coarse-grained gold labels (1, 2 or 3) and comparison of gold labels and system predictions by mean absolute error (MAE) has a counterintuitive effect on system rankings, as systems that produce continuous predictions are at an advantage over those that produce discrete predictions even though gold labels are also discrete. Figure 1(a) shows discrete gold label distributions for the scoring variant of Task 1.1 in WMT-14 and Figure 1(b) prediction distributions for an example system that was at a disadvantage because it restricted its predictions to discrete ratings like those of gold labels, and Figure 1(c) a system that achieves apparent better performance (lower MAE) despite prediction representations mismatching the discrete nature of gold labels.

Evaluation of the ranking variant of Task 1.1

	r
Post-edit Time	0.36
Post-edit Rate	0.69***

Table 1: Pearson correlation with HTER scores of post-edit times (PETs) and post-edit rates (PERs) for WMT-14 Task 1.2 and Task 1.3 gold labels, correlation marked with *** is significantly greater at p < 0.001.

again includes a significant mismatch between representations used as gold labels, which again were limited to the ratings 1, 2 or 3, while systems were required to provide a total-order ranking of test set translations, for example ranks 1-600 or 1-450, depending on language pair. Evaluation methodologies applied to ranking tasks may be better facilitated by application of more finegrained gold standard labels that more closely represent total-order rankings of system predictions.

Evaluation methodologies applied in Task 1.3 employ the more fine-grained post-edit times (PETs) as translation quality gold labels. PETs potentially provide a good indication of the underlying quality of translations, as a translation that takes longer to manually correct is thought to have lower quality. However, we propose what may correspond more directly to translation quality is an alteration of this, a post-edit rate (PER), where PETs are normalized by the number of words in translations. This takes into account the fact that, all else being equal, longer translations simply take a greater amount of time to post-edit than shorter ones. To investigate to what degree PERs may correspond better to translation quality than PETs, we compute correlations of each with HTER gold labels of translations from Task 1.2. Table 6 reveals a significantly higher correlation that exists between PER and HTER compared to PET and HTER (p < 0.001), and we conclude therefore that the PER of a translation provides a more faithful representation of translation quality than PET, and convert PETs for both predictions and gold labels to PERs (in seconds per word) in our later replication of Task 1.3.

In Task 1.2 of WMT-14, gold standard labels used to evaluate systems were in the form of human translation error rates (HTERs) (Snover et al., 2009). HTER scores provide an effective representation for evaluation of quality estima-

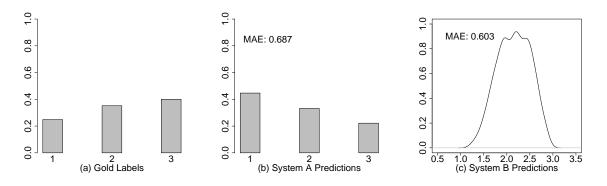


Figure 1: WMT-14 English-to-German quality estimation Task 1.1 where mismatched prediction/gold labels achieves apparent better performance, where (a) gold label distribution; (b) example system disadvantaged by its discrete predictions; (c) example system gaining advantage by its continuous predictions.

tion systems, as scores are individually computed per translation using custom post-edited reference translations, avoiding the bias that can occur with metrics that employ generic reference translations. In our later evaluation, we therefore use HTER scores in addition to PERs as suitable gold standard labels.

2.2 Mean Absolute Error

Mean absolute error is likely the most widely applied comparison measure of quality estimation system predictions and gold labels, in addition to being the official measure applied to scoring variants of tasks in WMT (Bojar et al., 2014). MAE is the average absolute difference that exists between a system's predictions and gold standard labels for translations, and a system achieving a lower MAE is considered a better system. Significant issues arise for evaluation of quality estimation systems with MAE when comparing distributions for predictions and gold labels, however. Firstly, a system's MAE can be lowered not only by individual predictions closer to corresponding gold labels, but also by prediction of aggregate statistics specific to the distribution of gold labels in the particular test set used for evaluation. MAE is most susceptible in this respect when gold labels have a unimodal distribution with relatively low standard deviation. For example, Figure 2(a) shows test set gold label HTER distribution for Task 1.2 in WMT-14 where the bulk of HTERs are located around one main peak with relatively low variance in the distribution. Unfortunately with MAE, a system that correctly predicts the location of the mode of the test set gold distribution and centers predictions around it with an optimally conservative variance can achieve lower MAE and apparent better performance. Figure 2(b) shows a lower MAE can be achieved by rescaling the original prediction distribution for an example system to a distribution with lower variance.

A disadvantage of an ability to gain in performance by prediction of such features of a given test set is that prediction of aggregates is, in general, far easier than individual predictions. In addition, inclusion of confounding test set aggregates such as these in evaluations will likely lead to both an overestimate of the ability of some systems to predict the quality of unseen translations and an underestimate of the accuracy of systems that courageously attempt to predict the quality of translations in the tails of gold distributions, and it follows that systems optimized for MAE can be expected to perform badly when predicting the quality of translations in the tails of gold label distributions (Moreau and Vogel, 2014).

Table 2 shows how MAEs of original predicted score distributions for *all systems* participating in Task 1.2 WMT-14 can be reduced by shifting and rescaling the prediction score distribution according to gold label aggregates. Table 3 shows that for similar reasons other measures commonly applied to evaluation of quality estimation systems, such as root mean squared error (RMSE), that are also not unit-free, encounter the same problem.

2.3 Significance Testing

In quality estimation, it is common to apply bootstrap resampling to assess the likelihood that a decrease in MAE (an improvement) has occurred by chance. In contrast to other areas of MT, where the accuracy of randomized methods of signifi-

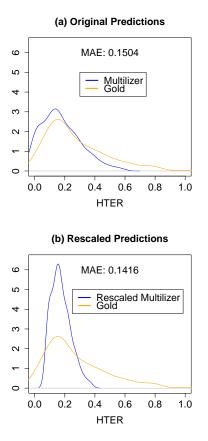


Figure 2: Comparison of example system from WMT-14 English-to-Spanish Task 1.2 (a) original prediction distribution and gold labels and (b) the same when the prediction distribution is rescaled to half its original standard deviation, showing a lower MAE can be achieved by reducing the variance in prediction distributions.

	Original	Rescaled
	MAE	MAE
FBK-UPV-UEDIN-wp	0.129	0.125
DCU-rtm-svr	0.134	0.127
USHEFF	0.136	0.133
DCU-rtm-tree	0.140	0.129
DFKI-svr	0.143	0.132
FBK-UPV-UEDIN-nowp	0.144	0.137
SHEFF-lite-sparse	0.150	0.141
Multilizer	0.150	0.135
baseline	0.152	0.149
DFKI-svr-xdata	0.161	0.146
SHEFF-lite	0.182	0.168

Table 2: MAE of WMT-14 Task 1.2 systems for original HTER prediction distributions and when distributions are shifted and rescaled to the mean and half the standard deviation of the gold label distribution.

	Original RMSE	Rescaled RMSE
FBK-UPV-UEDIN-wp	0.167	0.166
DCU-rtm-svr	0.167	0.165
DCU-rtm-tree	0.175	0.169
DFKI-svr	0.177	0.171
USHEFF	0.178	0.178
FBK-UPV-UEDIN-nowp	0.181	0.180
SHEFF-lite-sparse	0.184	0.179
baseline	0.195	0.194
DFKI-svr-xdata	0.195	0.187
Multilizer	0.209	0.181
SHEFF-lite	0.234	0.216

Table 3: RMSE of WMT-14 Task 1.2 systems for original HTER prediction distributions and when distributions are shifted and rescaled to the mean and half the standard deviation of the gold label distribution.

cance testing such as bootstrap resampling in combination with BLEU and other metrics have been empirically evaluated (Koehn, 2004; Graham et al., 2014), to the best of our knowledge no research has been carried out to assess the accuracy of similar methods specifically for quality estimation evaluation. In addition, since data used for evaluation of quality estimation systems are not independent, methods of significance testing differences in performance will be inaccurate unless the dependent nature of the data is taken into account.

3 Quality Estimation Evaluation by Pearson Correlation

The Pearson correlation is a measure of the linear correlation between two variables, and in the case of quality estimation evaluation this amounts to the linear correlation between system predictions and gold labels. Pearson's r overcomes the outlined challenges of previous approaches, such as mean absolute error, for several reasons. Firstly, Pearson's r is a unit-free measure with a key property being that the correlation coefficient is invariant to separate changes in location and scale in either of the two variables. This has the obvious advantage over MAE that the coefficient cannot be altered by shifting or rescaling prediction score distributions according to aggregates specific to the test set.

To illustrate, Figure 3 depicts a pair of systems for which the baseline system appears to outperform the other when evaluated with MAE, but this is only due to the conservative variance in its pre-

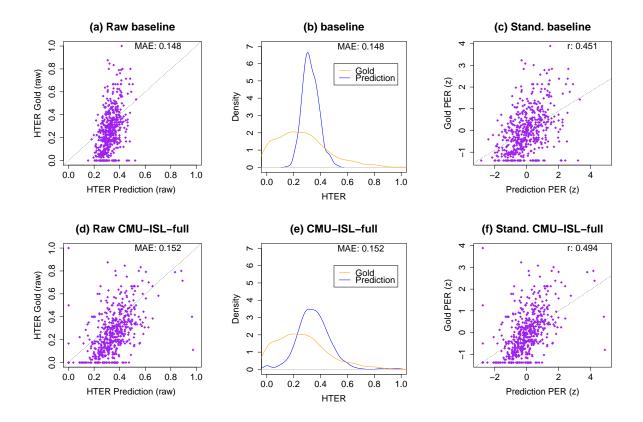


Figure 3: WMT-13 Task 1.1 systems showing baseline with better MAE than CMU-ISL-FULL only due to conservative variance in prediction distribution and despite its weaker correlation with gold labels.

diction score distribution, as can be seen by the narrow blue spike in Figure 3(b). Figure 3(e) shows how the prediction distribution of CMU-ISL-FULL, on the other hand, has higher variance, and subsequently higher MAE. Figures 3(c) and 3(f) depict what occurs in computation of the Pearson correlation where raw prediction and gold label scores are replaced by standardized scores, i.e. numbers of standard deviations from the mean of each distribution, where CMU-ISL-FULL in fact achieves a significantly higher correlation than the baseline system at p < 0.001.

An additional advantage of the Pearson correlation is that coefficients do not change depending on the representation used in the gold standard in the way they do with MAE, making possible a comparison of performance across evaluations that employ different gold label representations. Additionally, there is no longer a need for training and test representations to directly correspond to one another. To demonstrate, in our later evaluation we include the evaluation of systems trained on both HTER and PETs for prediction of both HTER and PERs. Finally, when evaluated with the Pearson correlation significance tests can be applied without resorting to randomized methods, in addition to taking into account the dependent nature of data used in evaluations.

The fact that the Pearson correlation is invariant to separate shifts in location and scale of either of the two variables is nonproblematic for evaluation of quality estimation systems. Take, for instance, the possible counter-argument: a pair of systems, one of which predicts the precise gold distribution, and another system predicting the gold distribution + 1, would unfairly receive the same Pearson correlation coefficient. Firstly, it is just as difficult to predict the gold distribution + 1, as it is to predict the gold distribution itself. More importantly, however, the scenario is extremely unlikely to occur in practice, it is highly unlikely that a system would ever accurately predict the gold distribution + 1, as opposed to the actual gold distribution unless training labels were adjusted in the same manner, or indeed predict the gold distribution shifted or rescaled by any other constant value. It is important to understand that invariance of the Pearson correlation to a shift in location or scale means that the measure is only invariant to a shift in location or scale applied to the entire distribution (of either of the two variables), such as the shift in location and scale that can be used to boost apparent performance of systems when measures like MAE and RMSE, that are not unit-free, are employed. Increasing the distance between system predictions and gold labels for anything less than the entire distribution, a more realistic scenario, or by something other than a constant across the entire distribution, will result in an appropriately weaker Pearson correlation.

4 Quality Estimation Significance Testing

Previous work has shown the suitability of Williams significance test (Williams, 1959) for evaluation of automatic MT metrics (Graham and Baldwin, 2014; Graham et al., 2015), and, for similar reasons, Williams test is appropriate for significance testing differences in performance of competing quality estimation systems which we detail further below.

Evaluation of a given quality estimation system, P_{new} , by Pearson correlation takes the form of quantifying the correlation, $r(P_{new}, G)$, that exists between system prediction scores and corresponding gold standard labels, and contrasting this correlation with the correlation for some baseline system, $r(P_{base}, G)$.

At first it might seem reasonable to perform significance testing in the following manner when an increase in correlation with gold labels is observed: apply a significance test separately to the correlation of each quality estimation system with gold labels, with the hope that the new system will achieve a significant correlation where the baseline system does not. The reasoning here is flawed however: the fact that one correlation is significantly higher than zero $(r(P_{new}, G))$ 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 gold labels are also not useful.

In psychology, 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. Data used for evaluation of quality estimation systems are not independent, however, and this means that if $r(P_{base}, G)$ and $r(P_{new}, G)$ are both > 0, the correlation between both sets of predictions themselves, $r(P_{base}, P_{new})$, must also be > 0. The strength of this correlation, directly between predictions of pairs of quality estimation systems, should be taken into account using a significance test of the difference in correlation between $r(P_{base}, G)$ and $r(P_{new}, G)$.

Williams test ¹ (Williams, 1959) evaluates the significance of 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 correlation between X_i and X_j , n is the size of the population, and:

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

As part of this research, we have made available an open-source implementation of statistical tests tailored to the assessment of quality estimation systems, at https://github.com/ ygraham/mt-qe-eval.

5 Evaluation and Discussion

To demonstrate the use of the Pearson correlation as an effective mechanism for evaluation of quality estimation systems, we rerun components of previous evaluations originally carried out at WMT-13 and WMT-14.

Table 4 shows Pearson correlations for systems participating in WMT-13 Task 1.1 where gold labels were in the form of HTER scores. System rankings diverge considerably from original rankings, notably the top system according to the Pearson correlation is tied in fifth place when evaluated with MAE.

Table 5 shows Pearson correlations of systems that took part in Task 1.2 of WMT-14, where gold labels were again in the form of HTER scores,

¹Also known as Hotelling-Williams.

System	r	MAE
DCU-SYMC-rc	0.595	0.135
SHEFMIN-FS	0.575	0.124
DCU-SYMC-ra	0.572	0.135
CNGL-SVRPLS	0.560	0.133
CMU-ISL-noB	0.516	0.138
CNGL-SVR	0.508	0.138
CMU-ISL-full	0.494	0.152
fbk-uedin-extra	0.483	0.144
LIMSI-ELASTIC	0.475	0.133
SHEFMIN-FS-AL	0.474	0.130
LORIA-INCTRA-CONT	0.474	0.148
fbk-uedin-rsvr	0.464	0.145
LORIA-INCTRA	0.461	0.148
baseline	0.451	0.148
TCD-CNGL-OPEN	0.329	0.148
TCD-CNGL-RESTR	0.291	0.152
UMAC-EBLEU	0.113	0.170

Table 4: Pearson correlation and MAE of system HTER predictions and gold labels for English-to-Spanish WMT-13 Task 1.1.

and to demonstrate the ability of evaluation of systems trained on a representation distinct from that of gold labels made possible by the unit-free Pearson correlation, we also include evaluation of systems originally trained on PET labels to predict HTER scores. Since PET systems also produce predictions in the form of PET, we convert predictions for all systems to PERs prior to computation of correlations, as PERs more closely correspond to translation quality. Results reveal that systems originally trained on PETs in general perform worse than HTER trained systems, and this is not all that surprising considering the training representation did not correspond well to translation quality. Again system rankings diverge from MAE rankings with the second best system according to MAE moved to the initial position.

Table 6 shows Pearson correlations for predictions of PER for systems trained on either PETs or HTER, and predictions for systems trained on PETs are converted to PER for evaluation. System rankings diverge most for this data set from the original rankings by MAE, as the system holding initial position according to MAE moves to position 13 according to the Pearson correlation.

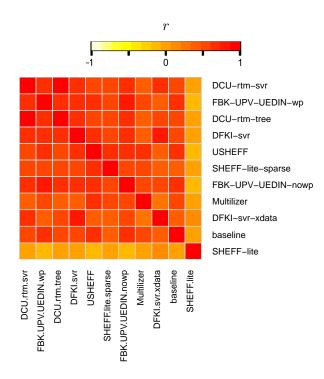
Many of the differences in correlation between systems in Tables 4, 5 and 6 are small and instead of assuming that an increase in correlation of one system over another corresponds to an improvement in performance, we first apply significance testing to differences in correlation with gold labels that exist between correlations for each pair

Training	QE		
Labels	System	r	MAE
HTER	DCU-rtm-svr	0.550	0.134
HTER	FBK-UPV-UEDIN-wp	0.540	0.129
HTER	DCU-rtm-tree	0.518	0.140
HTER	DFKI-svr	0.501	0.143
HTER	USHEFF	0.432	0.136
HTER	SHEFF-lite-sparse	0.428	0.150
HTER	FBK-UPV-UEDIN-nowp	0.414	0.144
HTER	Multilizer	0.409	0.150
PET	DCU-rtm-rr	0.350	_
HTER	DFKI-svr-xdata	0.349	0.161
PET	FBK-UPV-UEDIN-wp	0.346	_
PET	Multilizer-2	0.331	_
PET	Multilizer-1	0.328	_
PET	DCU-rtm-svr	0.315	_
HTER	baseline	0.283	0.152
PET	FBK-UPV-UEDIN-nowp	0.279	_
PET	USHEFF	0.246	_
PET	baseline	0.246	_
PET	SHEFF-lite-sparse	0.229	_
PET	SHEFF-lite	0.194	_
HTER	SHEFF-lite	0.052	0.182

Table 5: Pearson correlation and MAE of system HTER predictions and gold labels for English-to-Spanish WMT-14 Task 1.2 and 1.3 systems trained on either HTER or PET labelled data.

Training	QE		
Labels	System	r	MAE
HTER	FBK-UPV-UEDIN-wp	0.529	_
PET	FBK-UPV-UEDIN-wp	0.472	0.972
HTER	FBK-UPV-UEDIN-nowp	0.452	_
HTER	USHEFF	0.444	_
HTER	DCU-rtm-svr	0.444	_
HTER	DCU-rtm-tree	0.442	_
HTER	SHEFF-lite-sparse	0.441	_
PET	DCU-rtm-rr	0.430	0.932
PET	FBK-UPV-UEDIN-nowp	0.423	1.012
HTER	DFKI-svr	0.412	_
PET	USHEFF	0.394	1.358
PET	baseline	0.394	1.359
PET	DCU-rtm-svr	0.365	0.915
HTER	Multilizer	0.361	_
PET	SHEFF-lite-sparse	0.337	0.951
PET	SHEFF-lite	0.323	0.940
PET	Multilizer-1	0.288	0.993
HTER	baseline	0.286	_
HTER	DFKI-svr-xdata	0.277	_
PET	Multilizer-2	0.271	0.972
HTER	SHEFF-lite	0.011	_

Table 6: Pearson correlation of system PER predictions and gold labels for English-to-Spanish WMT-14 Task 1.2 and 1.3 systems trained on either HTER or PET labelled data, mean absolute error (MAE) provided are in seconds per word.



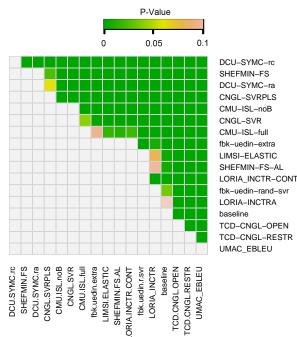


Figure 4: Pearson correlation between prediction scores for all pairs of systems participating in WMT-14 Task 1.2

Figure 5: HTER prediction significance test outcomes for all pairs of systems from English-to-Spanish WMT-13 Task 1.1, colored cells denote a significant increase in correlation with gold labels for row i system over column j system.

of systems.

5.1 Significance Tests

When an increase in correlation with gold labels is present for a pair of systems, significance tests provide insight into the likelihood that such an increase has occurred by chance. As described in detail in Section 4, the Williams test (Williams, 1959), a test also appropriate for MT metrics evaluated by the Pearson correlation (Graham and Baldwin, 2014), is appropriate for testing the significance of a difference in dependent correlations and therefore provides a suitable method of significance testing for quality estimation systems. Figure 4 provides an example of the strength of correlations that commonly exist between predictions of quality estimation systems.

Figure 5 shows significance test outcomes of the Williams test for systems originally taking part in WMT-13 Task 1.1, with systems ordered by strongest to least Pearson correlation with gold labels, where a green cell in (row i, column j) signifies a significant win for row i system over column j system, where darker shades of green signify conclusions made with more certainty. Test outcomes allow identification of significant increases in correlation with gold labels of one system over another, and subsequently the systems shown to outperform others. Test outcomes in Figure 5 reveal substantially increased conclusivity in system rankings made possible with the application of the Pearson correlation and Williams test, with almost an unambiguous total-order ranking of systems and an outright winner of the task.

Figure 6 shows outcomes of Williams significance tests for prediction of HTER and Figure 7 shows outcomes of tests for PER prediction for WMT-14 English-to-Spanish, again showing substantially increased conclusivity in system rankings for tasks.

It is important to note that the number of competing systems a system significantly outperforms should not be used as the criterion for ranking competing quality estimation systems, since the power of the Williams test changes depending on the degree to which predictions of a pair of systems correlate with each other. A system with predictions that happen to correlate strongly with predictions of many other systems would be at an unfair advantage, were numbers of significant wins to be used to rank systems. For this reason, it is

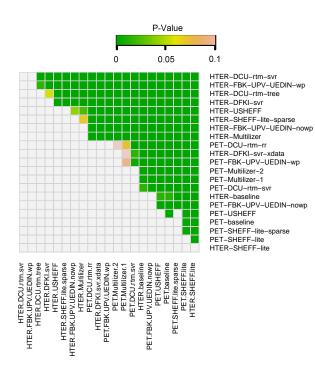


Figure 6: HTER prediction significance test outcomes for all pairs of systems from English-to-Spanish WMT-14 Task 1.2, colored cells denote a significant increase in correlation with gold labels for row i system over column j system.

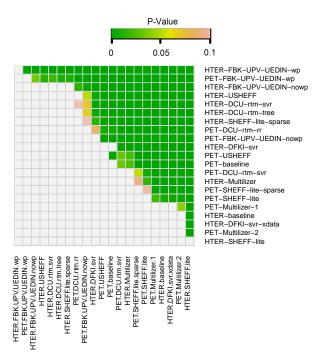


Figure 7: PER prediction significance test outcomes for all pairs of systems from English-to-Spanish WMT-14 Task 1.3, colored cells denote a significant increase in correlation with gold labels for row i system over column j system.

best to interpret pairwise system tests in isolation.

6 Conclusion

We have provided a critique of current widely used methods of evaluation of quality estimation systems and highlighted potential flaws in existing methods, with respect to the ability to boost scores by prediction of aggregate statistics specific to the particular test set in use or conservative variance in prediction distributions. We provide an alternate mechanism, and since the Pearson correlation is a unit-free measure, it can be applied to evaluation of quality estimation systems avoiding the previous vulnerabilities of measures such as MAE and RMSE. Advantages also outlined are that training and test representations no longer need to directly correspond in evaluations as long as labels comprise a representation that closely reflects translation quality. We demonstrated the suitability of the proposed measures through replication of components of WMT-13 and WMT-14 quality estimation shared tasks, revealing substantially increased conclusivity of system rankings.

Acknowledgements

We wish to thank the anonymous reviewers for their valuable comments and WMT organizers for the provision of data sets. This research is supported by Science Foundation Ireland through the CNGL Programme (Grant 12/CE/I2267) in the ADAPT Centre (www.adaptcentre.ie) at Trinity College Dublin.

References

- J. Blatz, E. Fitzgerald, G. Foster, S. Gandrabur, C. Goutte, A. Kulesza, A. Sanchis, and N. Ueffing. 2004. Confidence estimation for machine translation. In *Proceedings of the 20th international conference on Computational Linguistics*, pages 315– 321. Association for Computational Linguistics.
- O. Bojar, C. Buck, C. Federmann, B. Haddow, P. Koehn, J. Leveling, C. Monz, P. Pecina, M. Post, H. Saint-Amand, R. Soricut, L. Specia, and A. Tamchyna. 2014. Findings of the 2014 Workshop on Statistical Machine Translation. In *Proc. 9th Wkshp. Statistical Machine Translation*, Baltimore, MA. Association for Computational Linguistics.
- Y. Graham and T. Baldwin. 2014. Testing for significance of increased correlation with human judgment. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 172–176, Doha, Qatar. Association for Computational Linguistics.
- Y. Graham, N. Mathur, and T. Baldwin. 2014. Randomized significance tests in machine translation.

In *Proceedings of the ACL 2014 Ninth Workshop on Statistical Machine Translation*, pages 266–274. Association for Computational Linguistics.

- Yvette Graham, Nitika Mathur, and Timothy Baldwin. 2015. Accurate evaluation of segment-level machine translation metrics. In *Proceedings of the* 2015 Conference of the North American Chapter of the Association for Computational Linguistics Human Language Technologies, Denver, Colorado.
- P. Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proc. of Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.
- E. Moreau and C. Vogel. 2014. Limitations of mt quality estimation supervised systems: The tails prediction problem. In *Proceedings of COLING 2014*, *the 25th International Conference on Computational Linguistics*, pages 2205–2216.
- M. Snover, N. Madnani, B.J. Dorr, and R. Schwartz. 2009. Fluency, adequacy, or hter?: exploring different human judgments with a tunable mt metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 259–268. Association for Computational Linguistics.
- L. Specia, M. Turchi, N. Cancedda, M. Dymetman, and N. Cristianini. 2009. Estimating the sentence-level quality of machine translation systems. In *13th Conference of the European Association for Machine Translation*, pages 28–37.
- J.H. Steiger. 1980. Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, 87(2):245.
- E.J. Williams. 1959. *Regression analysis*, volume 14. Wiley New York.