Improving Evaluation of Document-level Machine Translation Quality Estimation

Yvette Graham
Dublin City University
yvette.graham@dcu.ie

Qingsong Ma
Chinese Academy of Sciences
maqingsong@ict.ac.cn

Timothy Baldwin
University of Melbourne
tb@ldwin.net

Qun Liu
Dublin City University
qun.liu@dcu.ie

Carla Parra
Dublin City University
carla.parra@adaptcentre.ie

Carolina Scarton
University of Sheffield
c.scarton@sheffield.ac.uk

Abstract

Meaningful conclusions about the relative performance of NLP systems are only possible if the gold standard employed in a given evaluation is both valid and reliable. In this paper, we explore the validity of human annotations currently employed in the evaluation of document-level quality estimation for machine translation (MT). We demonstrate the degree to which MT system rankings are dependent on weights employed in the construction of the gold standard, before proposing direct human assessment as a valid alternative. Experiments show direct assessment (DA) scores for documents to be highly reliable, achieving a correlation of above 0.9 in a self-replication experiment, in addition to a substantial estimated cost reduction through quality controlled crowdsourcing. The original gold standard based on post-edits incurs a 10–20 times greater cost than DA.

1 Introduction

Evaluation of NLP systems commonly takes the form of comparison of system-generated outputs with a corresponding human-sourced gold standard. The suitability of the employed gold standard representation greatly impacts the *reliability* and *validity* of conclusions drawn in any such evaluation. With respect to reliability, measures such as inter-annotator agreement (IAA) enable the likelihood of replicability to be taken into account, were an evaluation to be repeated with a distinct set of human annotators. One approach to achieving high IAA is through the development of a strict set of annotation guidelines, while for machine translation (MT), human assessment is more subjective, making high IAA difficult to achieve. For example, in past large-scale human evaluations of MT, low IAA levels have been highlighted as a cause of concern (Callison-Burch et al., 2007; Bojar et al., 2016). Such problems cause challenges not only for evaluation of MT systems, but also for MT quality estimation (QE), where the ideal gold standard comprises human assessment.

Although concern surrounding the *reliability* of human annotations is by far the most common complaint with respect to human evaluation of MT, the *validity* of the particular gold standard representation used in a given evaluation is also highly important. When it comes to validity, conventionally speaking, the very fact that human annotators manually generate the gold standard provides reassurance of its validity, as results at least reflect the judgment of one or more members of the target audience, i.e. human users. In the case of there being some "interpretation" of the human annotations, tuned to the particulars of a given task, validity becomes a concern. In recent document-level QE shared tasks, for example, the gold standard is generated through a linear combination of two separate human evaluation components, with weights tuned to optimize mean absolute error (MAE) and variance with respect to gold label distributions. In this paper, we explore the validity of the gold standard, and investigate to what degree tuning the gold standard impacts the validity of the resultant system performance estimates. Our contribution shows the method used to generate the gold standard has a substantial impact on the resultant system ranking, and propose an alternate gold standard representation for document-level quality estimation that is both more reliable and more valid as a gold standard.

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 356–361,
Valencia, Spain, April 3-7, 2017.

2 Background

Document-level QE (Soricut and Echihabi, 2010) is a relatively new area, with only two shared tasks taking place to date (Bojar et al., 2015; Bojar et al., 2016).

In WMT-15, gold standard labels took the form of automatic metric scores for documents (specifically Meteor scores (Denkowski and Lavie, 2011)), and system predictions were compared to gold labels via MAE. A conclusion that emerged from the initial shared task was that automatic metric scores were not adequate, based on the following observation: if the average of the training set scores is used as a prediction value for all data points in the test set, this results in a system as good as the baseline system when evaluated with MAE. The fact that average scores are good predictors is more likely a consequence of the applied evaluation measure, MAE, however, as outlined in Graham (2015). When evaluated with the Pearson correlation, such a set of predictions would not be a reasonable entry to the shared task since the prediction distribution would effectively be a constant and its correlation with anything is therefore undefined. Regardless of the predictability of automatic metric scores when evaluated with MAE, they unfortunately do not provide a suitable gold standard, simply because they are known to provide an insufficient substitute for human assessment, often unfairly penalizing translations that happen to be superficially dissimilar to reference translations (Callison-Burch et al., 2006).

Consequently, for WMT-16, the gold standard was modified to take the form of a linear combination of two human-targeted translation edit rate (HTER) (Snover et al., 2006) scores assigned to a given document. Scores were produced via two human post-editing steps: firstly, sentences within a given MT-output document were post-edited independent of other sentences in that document, producing post-edition 1 (PE_1). Secondly, PE_1 sentences were concatenated to form a document-level translation, and post-edited a second time by the same annotator, with the aim of isolating errors only identifiable when more context is available, to produce post-edition 2 (PE_2). Next, two translation edit rate (TER) scores were computed by: (1) comparing the document-level MT output with PE_1, TER(PE_1, MT); and TER between PE_2 and PE_1, TER(PE_2, PE_1). Finally, these two scores were combined into a single gold standard label, G, as follows:

$$G = W_1\text{TER}(PE_1, MT) + W_2\text{TER}(PE_2, PE_1)$$

where weights, W_1 and W_2, are decided by the outcome of the following tuning process: W_1 is held static at 1; W_2 is increased by 1 from a starting value of 1 until either of the following stopping criteria is reached: (i) the ratio between the standard deviation and the mean is 0.5 for the official baseline QE system predictions, or (ii) a baseline prediction distribution is constructed by assigning to all prediction labels the expected value of the training set labels. This second case is designed to deal with the degenerate behaviour described above of assigning to each test item the average over the training data, with the stopping criteria being such that the difference between the MAE achieved by such a system and the official baseline MAE is at least 0.1. The final values used to produce official results were $W_1 = 1$ and $W_2 = 13$.

The way in which the gold standard is constructed deviates to quite a degree from conventional gold standards, therefore, which raises some important questions. Firstly, it appears that the optimization process is carried out with direct reference to the test set. If so, does such a process overly blur the lines with respect to what is considered true unseen test data?

Secondly, neither of the two TER scores corresponds to a straightforward human assessment, putting into doubt the conventional validity attributed to human-generated gold standards. For example, the component assigned most weight in the final evaluation is TER(PE_2, PE_1), and this unfortunately corresponds more closely to a measure of the dependence of the meaning of the sentences within a given document on other sentences in that document, as opposed to the overall quality of the MT output document.

Finally, and most importantly, assigning weights to components of the human evaluation through a somewhat arbitrary optimization process deviates from the expected interpretation of each reported correlation, i.e. the correlation between system predictions of translation quality and the actual quality of translated documents. Including such weights in the construction of a gold standard potentially invalidates the human evaluation, and is unfortunately very likely to exaggerate the apparent performance of some systems while under-rewarding others.

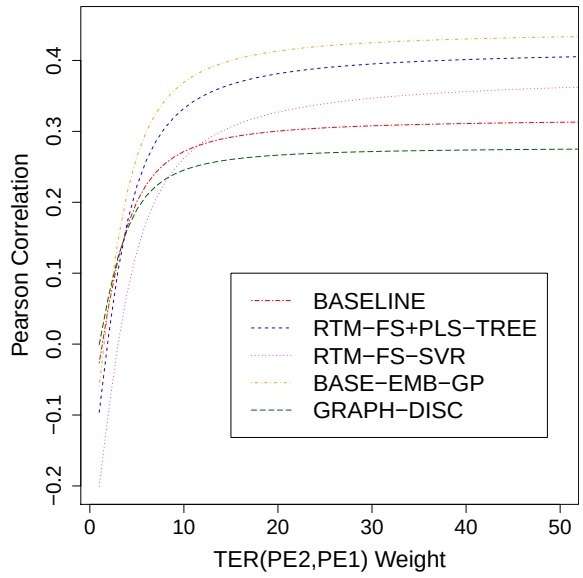

Figure 1: System performance as the weight of the TER(PE_2, PE_1) human evaluation component is increased to 13, as in official evaluation, and beyond (WMT-16 document-level QE English to Spanish shared task systems).

To demonstrate to what degree this could be the case, since post-editions employed in the creation of the actual gold standard used to produce results in the shared task are unavailable, we simulate a possible set of TER(PE_1, MT) and TER(PE_2, PE_1) labels for test documents in the following way: A possible set of TER(PE_1, MT) labels are simulated by relocation of the TER score distribution (of the MT output document with reference translations as opposed to post-edits) to more closely resemble scores of our later human evaluation, before rescaling that score distribution according to the mean and standard deviations (provided in the QE task findings paper) of TER(PE_1, MT). TER(PE_2, PE_1) scores were then reverse-engineered from the correspondence between TER(PE_1, MT) and gold labels.[1] Final gold labels arrived at through our simulation of TER(PE_1, MT) and TER(PE_2, PE_1) are identical to the original evaluation for $W_1 = 1$ and $W_2 = 13$.

Figure 1 shows correlations achieved by all systems participating in the shared task when the weight of our simulated TER(PE_2, PE_1) component is varied from 1 up towards the original weight of 13 and beyond. The correlation achieved by all systems varies dramatically with W_2, demonstrating how correlations achieved by QE systems are highly dependent on the chosen weights.

3 Alternate Human Gold Standard

A recent development in human evaluation of MT is direct assessment ("DA"), a human assessment shown to yield highly replicable segment-level scores, by combination of a minimum of 15 repeat human assessments per translation into mean scores (Graham et al., 2015).

Human adequacy assessments are collected via a 0–100 rating scale that facilitates reliable quality control of crowd-sourcing. Document-level DA scores are computed by repeat assessment of the individual segments within a given document, computation of the mean score for each segment (micro-average), and finally, combination of the mean segment scores into an overall mean document score (macro-average).[2]

DA assessments are carried out by comparison of a given MT output segment (rendered in black) with a human-generated reference translation (in gray), and human annotators rate the degree to which they agree with the statement: *The black text adequately expresses the meaning of the gray text in Spanish*.[3]

Reference translations employed in DA are manually translated by an expert with reference to the entire source document, thus ensuring individual reference segments retain any elements needed to stay faithful to the meaning of the source document as a whole. Since in creation of a test set in general in MT, the professional human translator will have access to and make use of the entire source document, reference translations found in standard MT test sets can directly be employed.

3.1 Self-replication Experiment

Although DA has been shown to produce highly reliable human scores for translations on the segment level, achieving a correlation of above 0.9 between scores for segments collected in separate data collection runs (Graham et al., 2015), the reliability of DA on the document level has yet to be tested. Similar to Graham et al.

[1] All data employed in this work is available at `http://github.com/ygraham/eacl2017`

[2] Micro-averaging before macro-averaging avoids weighting segments by the number of times they are assessed.

[3] Instructions are translated into the target language.

	Total	Post QC	Mean Assess. per Document
Run A	14,600	6,640	107
Run B	10,050	7,700	124

Table 1: Numbers of DA human assessments collected per data collection run on Mechanical Turk before ("Total") and after quality control filtering ("Post QC") for WMT-16 Document-level QE task (English to Spanish; 62 documents in total).

(2015), we therefore assess the reliability of DA for document-level human evaluation by quality-controlled crowd-sourcing in two separate data collection runs (Runs A and B) on Mechanical Turk, and compare scores for individual documents collected in each run.

Quality control is carried out by inclusion of pairs of genuine MT outputs and automatically degraded versions of them (bad references) within 100-translation HITs, before a difference of means significance test is applied to the ratings belonging to a given worker. The resulting p-value is employed as an estimate of the reliability of a given human assessor to accurately distinguish between the quality of translations (Graham et al., 2013; Graham et al., 2014). Table 1 shows numbers of judgments collected in total for each data collection run on Mechanical Turk, including numbers of assessments before and after quality control filtering, where only data belonging to workers with a p-value below 0.05 were retained.

Figure 2 shows the correlation between document-level DA scores collected in Run A with scores produced in Run B, where, for Run B, repeat assessments are down-sampled to show the increasing correspondence between scores as ever-increasing numbers of repeat assessments are collected for a given document. Correlation between scores collected in the two separate data collection runs reaches $r = 0.901$ by a minimum of 27 repeat assessments of the sentences of a given document, or by an average 107 sentence assessments per document.[4]

Since DA scores achieve a correlation of $r > 0.9$ in our self-replication experiment, we now know that DA provides reliable human evaluation

[4]Variance in numbers of repeat assessments per document is due to sentences of all documents being sampled without preference for documents made up of larger numbers of sentences.

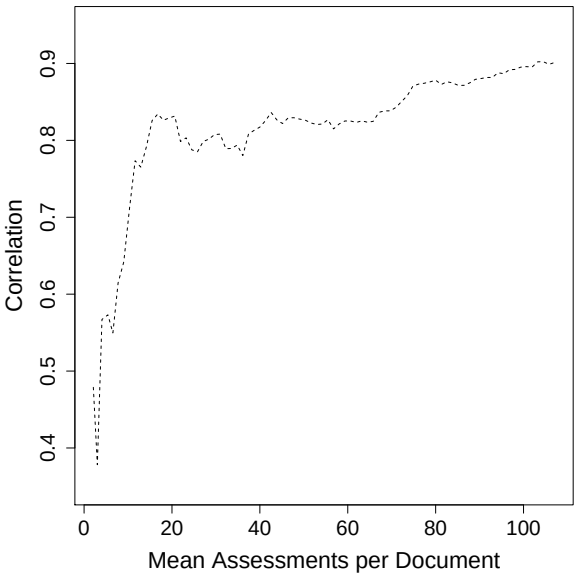

Figure 2: Correlation between scores for documents collected in initial data collection run and scores for the same documents as numbers of repeat assessments per document are increased.

scores for not only segments but also documents. The validity of DA is superior to the existing gold standard employed for document-level QE as it avoids arbitrary weighting or tuning of component scores to reach final gold standard labels. It is therefore highly unlikely to ever unfairly exaggerate (or under-reward) the performance of any QE system in a given evaluation.

With regard to resources required to construct each gold standard, a single DA data collection run cost USD$109 on average , while the cost estimate provided to us by a professional post-editor for the same test set came between USD$1,422 and USD$2,728. In other words, the cost of producing the gold standard is 10–20 times greater for post-editing than DA.[5]

3.2 Re-evaluating Doc-level QE WMT-16

In order to demonstrate DA's potential as a gold standard, Table 2 shows correlations for WMT-16 document-level QE shared task systems when evaluated with DA and the original gold standard. Results show system rankings that diverge from the original, as the original gold standard exaggerated the performance of three participating sys-

[5]Post-editing cost estimates are based on 0.06 and 0.12 Euro per source document word converted to USD$. Further details provided by the post-editor in relation to estimates can be found at `https://github.com/ygraham/eacl2017`

	DA	WMT-16
RTM-FS+PLS-TREE	0.38	0.36
GRAPH-DISC	0.32	0.26
BASE-EMB-GP	0.31	0.39
BASELINE	0.26	0.29
RTM-FS-SVR	0.23	0.29

Table 2: Correlation (r) of system predictions with direct assessment (DA) and original gold standard (WMT-16 QE English to Spanish)

tems, while under-rewarding two other systems. Notably, system GRAPH-DISC, which includes discourse features learned from document-level features, achieves a higher correlation when evaluated with DA compared to the original gold standard.

Differences in correlations are small, however, and can't be interpreted as differences in performance without significance testing. Differences in dependent correlations showed no significant difference for all pairs of competing systems according to Williams test (Williams, 1959; Graham and Baldwin, 2014).

3.3 Discussion of DA Fluency Omission

In development of the newly proposed variant of DA for document-level QE, the question arose if the assessment should also include an assessment of the fluency of documents (in addition to adequacy), as in Graham et al. (2016b). Besides the several other design criteria in DA aimed at avoiding possible sources of bias in general, the motivation for including a separate fluency assessment was originally to counter any bias resulting from comparison of the MT output with a reference translation in the adequacy assessment, similar to the reference bias encountered in automatic metrics scores. Although genuine human assessors of MT are unlikely to be biased by the reference by anything close to the degree to which automatic metrics will be, there still exists the possibility that reference bias could impact the accuracy of DA scores to *some* degree. Inclusion of fluency does of course have a trade-off, however, requiring additional resources, resources that could otherwise be employed to increase the number of translations in the test set, for example. It is important to investigate the degree to which reference bias may or may not be a problem for DA before including it in document-level QE evaluation therefore.

Graham et al. (2016a) provide an investigation into reference bias in monolingual evaluation of MT and despite the risk of reference bias that DA adequacy could potentially encounter, experiment results show no evidence of reference bias. Human assessors of MT appear to genuinely read and compare the meaning of the reference translation and the MT output, as requested with DA, applying their human intelligence to the task in a reliable way, and are not overly influenced by the generic reference.

Although DA fluency could still have its own applications, for the purpose of evaluating MT or MT QE, this additional insight into the lack of reference bias encountered by DA adequacy means that there is no longer any real motivation for including DA fluency when resources are constrained. Given the choice of inclusion of DA fluency in evaluation of document-level QE or expanding the test set (with respect to adequacy), there is no question that the latter is now the more sensible choice.

4 Conclusion

Methodological concerns were raised with respect to optimization of weights employed in construction of document-level QE gold standards in WMT-16. We demonstrated the degree to which MT system rankings are dependent on weights employed in the construction of the gold standard. Experiments showed with respect to the alternate gold standard we propose, direct assessment (DA), scores for documents are highly reliable, achieving a correlation of above 0.9 in a self-replication experiment. Finally, DA resulted in a substantial estimated cost reduction, with the original post-editing gold standard incurring a 10–20 times greater cost than that of DA.

Acknowledgments

This project has received funding from the European Union Horizon 2020 research and innovation programme under grant agreement 645452 (QT21) and Science Foundation Ireland in the ADAPT Centre for Digital Content Technology (`www.adaptcentre.ie`) at Dublin City University funded under the SFI Research Centres Programme (Grant 13/RC/2106) co-funded under the European Regional Development Fund.

References

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 Workshop on Statistical Machine Translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal, September. Association for Computational Linguistics.

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

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *Proc. 11th Conf. European Chapter of the ACL*, pages 249–256, Trento, Italy, April. Association for Computational Linguistics.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic, June. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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)*, pages 172–176, Doha, Qatar, October. 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 & Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2014. Is machine translation getting better over time? In *Proceedings of the European Chapter of the Association of Computational Linguistics*, pages 443–451, Gothenburg, Sweden. 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*, pages 1183–1191, Denver, Colorado. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Meghan Dowling, Maria Eskevich, Teresa Lynn, and Lamia Tounsi. 2016a. Is all that glitters in machine translation quality estimation really gold standard? In *Proceedings of the 26th International Conference on Computational Linguistics*, Osaka, Japan.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2016b. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, FirstView:1–28, 1.

Yvette Graham. 2015. Improving evaluation of machine translation quality estimation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 1804–1813, Beijing, China. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Scwartz, John Makhoul, and Linnea Micciula. 2006. A study of translation error rate with targeted human annotation. In *Proceedings of the 7th Biennial Conference of the Association for Machine Translaiton in the Americas*, pages 223–231, Boston, MA.

Radu Soricut and Abdessamad Echihabi. 2010. Trustrank: Inducing trust in automatic translations via ranking. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 612–621. Association for Computational Linguistics.

Evan James Williams. 1959. *Regression analysis*, volume 14. Wiley New York.