Towards Efficient Machine Translation Evaluation by Modelling Annotators

Nitika Mathur **Timothy Baldwin** **Trevor Cohn**
School of Computing and Information Systems
The University of Melbourne
Victoria 3010, Australia

nmathur@student.unimelb.edu.au {tbaldwin,tcohn}@unimelb.edu.au

Abstract

Current machine translation evaluations use Direct Assessment, based on crowdsourced judgements from a large pool of workers, along with quality control checks, and a robust method for combining redundant judgements. In this paper we show that the quality control mechanism is overly conservative, increasing the time and expense of the evaluation. We propose a model that does not filter workers, and takes into account varying annotator reliabilities. Our model effectively weights each worker's scores based on the inferred precision of the worker, and is much more reliable than the mean of either the raw or standardised scores.

1 Introduction

Accurate evaluation is critical for measuring progress in machine translation (MT). Despite progress over the years, automatic metrics are still biased, and human evaluation is still a fundamental requirement for reliable evaluation. The process of collecting human annotations is time-consuming and expensive, and the data is always noisy. The question of how to efficiently collect this data has evolved over the years, but there is still scope for improvement. Furthermore, once the data has been collected, there is no consensus on the best way to reason about translation quality.

Direct Assessment ("DA": Graham et al. (2017)) is currently accepted as the best practice for human evaluation, and is the official method at the Conference for Machine Translation (Bojar et al., 2017a). Every annotator scores a set of translation-pairs, which includes quality control items designed to filter out unreliable workers.

However, the quality control process has low recall for good workers: as demonstrated in Section 3, about one third of good data is discarded, increasing expense. Once good workers are identified, their outputs are simply averaged to produce the final 'true' score, despite their varying accuracy.

In this paper, we provide a detailed analysis of these shortcomings of DA and propose a Bayesian model to address these issues. Instead of standardising individual worker scores, our model can automatically infer worker offsets using the raw scores of all workers as input. In addition, by learning a worker-specific precision, each worker effectively has a differing magnitude of vote in the ensemble. When evaluated on the WMT 2016 Tr-En dataset which has a high proportion of unskilled annotators, these models are more efficient than the mean of the standardised scores.

2 Background

The Conference on Machine Translation (WMT) annually collects human judgements to evaluate the MT systems and metrics submitted to the shared tasks. The evaluation methodology has evolved over the years, from 5 point adequacy and fluency rating, to relative rankings ("RR"), to DA. With RR, annotators are asked to rank translations of 5 different MT systems. In earlier years, the final score of a system was the expected number of times its translations score better than translations by other systems (expected wins). Bayesian models like Hopkins and May (Hopkins and May, 2013) and Trueskill (Sakaguchi et al., 2014) were then proposed to learn the relative ability of the MT systems. Trueskill was adopted by WMT in 2015 as it is more stable and efficient than the expected wins heuristic.

DA was trialled at WMT 2016 (Bojar et al.,

Nitika Mathur, Timothy Baldwin and Trevor Cohn. 2018. Towards Efficient Machine Translation Evaluation by Modelling Annotators. In *Proceedings of Australasian Language Technology Association Workshop*, pages 77–82.

2016a), and has replaced RR since 2017 (Bojar et al., 2017a). It is more scalable than RR as the number of systems increases (we need to obtain one annotation per system, instead of one annotation per system pair). Each translation is rated independently, minimising the risk of being influenced by the relative quality of other translations. Ideally, it is possible that evaluations can be compared across multiple datasets. For example, we can track the progress of MT systems for a given language pair over the years.

Another probabilistic model, EASL (Sakaguchi and Van Durme, 2018), has been proposed that combines some advantages of DA with Trueskill. Annotators score translations from 5 systems at the same time on a sliding scale, allowing users to explicitly specify the magnitude of difference between system translations. Active learning to select the systems in each comparison to increase efficiency. But it does not model worker reliability, and is, very likely, not compatible with longitudinal evaluation, as the systems are effectively scored relative to each other.

In NLP, most other research on learning annotator bias and reliability has been on categorical data (Snow et al., 2008; Carpenter, 2008; Hovy et al., 2013; Passonneau and Carpenter, 2014).

3 Direct Assessment

To measure adequacy, in DA, annotators are asked to rate how adequately an MT output expresses the meaning of a reference translation using a continuous slider, which maps to an underlying scale of 0–100. These annotations are crowdsourced using Amazon Mechanical Turk, where "workers" complete "Human Intelligence Tasks" (HITs) in the form of one or more micro-tasks.

Each HIT consists of 70 MT system translations, along with an additional 30 control items:

1. degraded versions of 10 of these translations;
2. 10 reference translations by a human expert, corresponding to 10 system translations; and
3. repeats of another 10 translations.

The scores on the quality control items are used to filter out workers who either click randomly or on the same score continuously. A conscientious worker would give a near perfect score to reference translations, give a lower score to degraded translations when compared to the corresponding MT system translation, and be consistent with scores for repeat translations.

The paired Wilcoxon rank-sum test is used to test whether the worker scored degraded translations worse than the corresponding system translation. The (arbitrary but customary) cutoff of $p < 0.05$ is used to determine good workers. The paired Wilcoxon rank-sum test ($p < 0.05$) is used to test whether the worker scored degraded translations worse than the corresponding system translation. The remaining workers are further tested to check that there is no significant difference between their scores for repeat-pairs.

Worker scores are manually examined to filter out workers who obviously gave the same score to all translations, or scored translations at random. Only these workers are rejected payment. Thus, other workers who do not pass the quality control check are paid for their efforts, but their scores are unused, increasing the overall cost.

Some workers might have high standards and give consistently low scores for all translations, while others are more lenient. And some workers may only use the central part of the scale. Standardising individual workers' scores makes them more comparable, and reduces noise before calculating the mean.

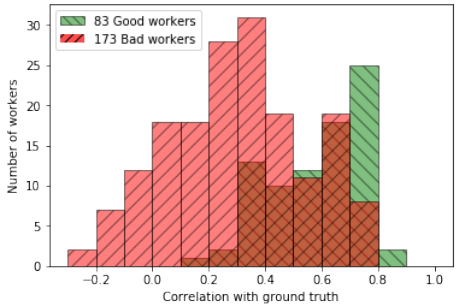
The final score of an MT system is the mean standardised score of its translations after discarding scores that do not meet quality control criteria. The noise in worker scores is cancelled out when a large number of translations are averaged.

To obtain accurate scores of individual translations, multiple judgments are collected and averaged. As we increase the number of annotators per translation, there is greater consistency and reliability in the mean score. This was empirically tested by showing that there is high correlation between the mean of two independent sets of judgments, when the sample size is greater than 15 (Graham et al., 2015).

However, both these tests are based on a sample-size of 10 items, and, as such, the first test has low power; we show that it filters out a large proportion of the total workers. One solution would be to increase the sample size of the degraded-reference-pairs, but this would be at the expense of the number of useful worker annotations. It is better to come up with a model that would use the scores of all workers, and is more robust to low quality scores.

Automatic metrics such as BLEU (Papineni et al., 2002) are generally evaluated using the Pear-

(a) all language pairs



(b) Tr-En language pair

Figure 1: Accuracy of "good" vs "bad" workers in the WMT 2016 dataset.

son correlation with the mean standardised score of the good workers. We similarly evaluate a worker's accuracy using the Pearson correlation of the worker's scores with this ground truth. Over all the data collected for WMT16, the group of good workers are, on average, more accurate than the group of workers who failed the significance test. However, as seen in Figure 1a, there is substantial overlap in the accuracies of the two groups. We can see that very few inaccurate workers were included. However, about a third of the total workers whose scores have a correlation greater than 0.6 were not approved. In particular, over the Tr-En Dataset, the significance test was not very effective, as seen in Figure 1b.

Workers whose scores pass the quality control check are given equal weight, despite the variation in their reliability. Given that quality control is not always reliable (as with the Tr-En dataset, e.g.), this could include worker with scores as low as $r = 0.2$ correlation with the ground truth.

While worker standardisation succeeds in increasing inter-annotator consistency, this process discards information about the absolute quality of the translations in the evaluation set. When using the mean of standardised scores, we cannot compare MT systems across independent evaluations. In the evaluation of the WMT 17 Neural MT Training Task, the baseline system trained on 4GB GPU memory was evaluated separately from the baseline trained on 8 GB GPU memory and the other submissions. In this setup of manual evaluation, Baseline-4GB scores slightly higher than Baseline-8GB when using raw scores, which is possibly due to chance. However, it scores significantly higher when using standardised scores, which goes against our expectations (Bojar et al., 2017b).

4 Models

We use a simple model, assuming that a worker score is normally distributed around the true quality of the translation. Each worker has a precision parameter τ that models their accuracy: workers with high τ are more accurate. In addition, we include a worker-specific offset β, which models their deviation from the true score.

For each translation $i \in T$, we draw the true quality μ from the standard normal distribution.[1] Then for each worker $j \in W$, we draw their accuracy τ_j from a gamma distribution with shape parameter k and rate parameter θ.[2] The offset β_j is again drawn from the standard normal distribution. The worker's score r_{ij} is drawn from a normal distribution, with mean $\mu_i + \beta_j$, and precision τ_j.

$$r_{ij} = \mathcal{N}\left(\mu_i + \beta_j, \tau_j^{-1}\right) \quad (1)$$

To help the model, we add constraints on the quality control items: the true quality of the degraded translation is lower than the quality of the corresponding system translation. In addition, the true quality of the repeat items should be approximately equal.

We expect that the model will learn a high τ for good quality workers, and give their scores higher weight when estimating the mean. We believe that the additional constraints will help the model to infer the worker precision.

DA can be viewed as the Maximum Likelihood Estimate of this model, with the following substitutions in Equation (1): s_{ij} is the standardised score of worker j, β_j is 0 for all workers, and τ is

[1] We first standardise scores (across all workers together) in the dataset

[2] We use $k = 2$ and $\theta = 1$ based on manual inspection of the distribution of worker precisions on a development dataset (WMT18 Cs-En)

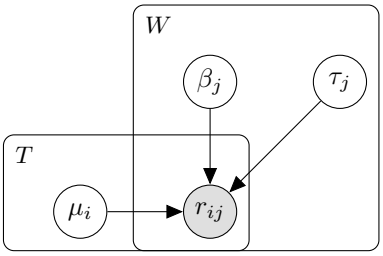


Figure 2: The proposed model, where worker $j \in W$ has offset β_j and precision τ_j, translation $i \in T$ has quality μ_i, and worker j scores translation i with r_{ij}

.

constant for all workers.

$$s_{ij} = \mathcal{N}\left(\mu_i, \tau^{-1}\right) \tag{2}$$

The choice of a Gaussian distribution to model worker scores is technically deficient as a Gaussian is unbounded, but it is still a reasonable approximation. This could be remedied, for example, by using a truncated Gaussian distribution, which we leave to future work.

We want to maximise the likelihood of the observed judgments:

$$\begin{aligned} P(r) = &\iint_{j=1}^{W} P(\beta_j)P(\tau_j) \int_{i=1}^{T} P(\mu_i) \\ &P(r_{i,j}|\mu_i, \beta, \tau)\, d\beta\, d\tau\, d\mu \\ = &\iint_{j=1}^{W} \mathcal{N}\left(\beta_j|0,1\right) \Gamma\left(\tau_j|k,\theta\right) \int_{i=1}^{T} \mathcal{N}\left(\mu_i|0,1\right) \\ &\mathcal{N}\left(r_{ij}|\mu_i, \tau^{-1}\right)\, d\beta\, d\tau\, d\mu \end{aligned} \tag{3}$$

We use the Expectation Propagation algorithm (Minka, 2001) to infer posteriors over μ and worker parameters β and τ.[3] Expectation Propagation is a technique for approximating distributions which can be written as a product of factors. It iteratively refines each factor by minimising the KL divergence from the approximate to the true distribution.

5 Experiments

We evaluate our models on data from the segment-level WMT 16 dataset (Bojar et al., 2016b). We choose the Turkish to English (Tr-En) dataset, which consists of 256 workers, of which about two thirds (67.58%) fail the quality control measures. It consists of 560 translations, with at least 15 "good" annotations for each of these translations (see Figure 1b).

[3]We use the Infer.NET (Minka et al., 2018) framework to implement our models.

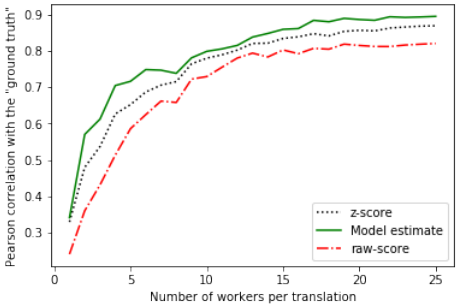


Figure 3: Pearson's r of the estimated true score with the "ground truth" as we increase the number of workers per translation.

We use the mean of 15 good standardised annotations as a proxy for the gold standard when evaluating efficiency, and starting from one worker, increase the number of workers to the maximum available. Figure 3 shows that our models are consistently more accurate than the mean of the standardised scores.

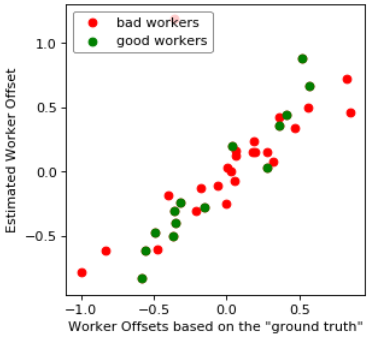
Figure 4 shows the learned precision and offset for 5 annotators per translation, against the precision and offset of worker scores calculated with respect to the "ground truth". This shows that the model is learning worker parameters even when the number of workers is very small, and is using this information to get a better estimate of the mean (the model obtains $r = 0.72$, compared to $r = 0.65$ for the mean z-score).

On further examination of the outlier in Figure 4a, we find that this worker is pathologically bad. They give a 0 score for all the translations in one HIT, and mostly 100s to the other half. This behaviour is not captured by our model.

6 Discussion and Future Work

We showed that significance tests over a small set of quality control items are ineffective at identifying good and bad workers, and propose a model that does not depend on this step. Instead, it uses constraints on the quality control items to learn worker precision, and returns a more reliable estimate of the mean using fewer worker scores per translation. This model does not tell us when to stop collecting judgments. It would be useful to know to have a method to determine when to stop

(a) Worker Precision



(b) Worker Offset

Figure 4: Scatter plot of worker precision/offset inferred by the model with only 5 workers per translation, against the precision/offset of the deltas of the worker score and the "ground truth".

collecting annotations based on scores received, instead of relying on a number obtained from one-time experiments.

More importantly, we need to have ways to calibrate worker scores to ensure consistent evaluations across years, so we can measure progress in MT over time. Even if a better model is found to calibrate workers, this does not ensure consistency in judgments, and we believe the HIT structure needs to be changed. We propose to replace the 30 quality control items with items of reliably known quality from the previous year. The correlation between the worker scores and the known scores can be used to assess the reliability of the worker. Moreover, we can scale the worker scores based on these known items, to ensure consistent scores over years.

Acknowledgements

We thank the anonymous reviewers for their valuable feedback and suggestions. This work was supported in part by the Australian Research Council.

References

Ondřej 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. 2017a. Findings of the 2017 Conference on Machine Translation (WMT17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*. Copenhagen, Denmark, pages 169–214. http://www.aclweb.org/anthology/W17-4717.

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. 2016a. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation*. Berlin, Germany, pages 131–198.

Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016b. Results of the WMT16 Metrics Shared Task. In *Proceedings of the First Conference on Machine Translation*. Berlin, Germany, pages 199–231.

Ondřej Bojar, Jindřich Helcl, Tom Kocmi, Jindřich Libovický, and Tomáš Musil. 2017b. Results of the WMT17 Neural MT Training task. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*. Copenhagen, Denmark, pages 525–533. http://www.aclweb.org/anthology/W17-4757.

Bob Carpenter. 2008. Multilevel Bayesian models of categorical data annotation. Technical report, Alias-i.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering* 23(1):330.

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, USA, pages 1183–1191.

Mark Hopkins and Jonathan May. 2013. Models of translation competitions. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*. Sofia, Bulgaria, pages 1416–1424.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard H Hovy. 2013. Learning whom to trust

with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2013)*. Atlanta, USA, pages 1120–1130.

T. Minka, J.M. Winn, J.P. Guiver, Y. Zaykov, D. Fabian, and J. Bronskill. 2018. Infer.NET 0.3. Microsoft Research Cambridge. `http://dotnet.github.io/infer`.

Thomas P. Minka. 2001. Expectation propagation for approximate Bayesian inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. Seattle, USA, pages 362–369.

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 (ACL 2002)*. Philadelphia, USA, pages 311–318.

J. Rebecca Passonneau and Bob Carpenter. 2014. The benefits of a model of annotation. *Transactions of the Association of Computational Linguistics* 2(1):311–326.

Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. 2014. Efficient elicitation of annotations for human evaluation of machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Baltimore, USA, pages 1–11.

Keisuke Sakaguchi and Benjamin Van Durme. 2018. Efficient online scalar annotation with bounded support. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pages 208–218. http://aclweb.org/anthology/P18-1020.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast — but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Honolulu, USA, pages 254–263.