How Good is Your Machine Translation?
Quality Estimation for Direct User Feedback

Andrea Brennen[1], **Ab Mosca**[2], **Remco Chang**[2], **Nina Lopatina**[3]
[1] IQT Labs, Waltham, MA & [3] IQT Labs, Menlo Park, CA
[2] Tufts University, Medford, MA
{abrennen, nlopatina}@iqt.org, {amosca01, remco}@cs.tufts.edu

Abstract

Mistranslation by AI has led to bad outcomes, including erroneous arrests. To address this problem, we designed a tool called VeriCAT, short for **Veri**fication of **C**omputer-**A**ssisted **T**ranslation. We used a Quality Estimation model to predict a sentence-level quality score for individual snippets of Russian text that had been translated into English and we designed a simple user interface to display these scores, along with the text snippets, in a way that helps users determine whether to trust a specific machine-translated sentence. We evaluated VeriCAT by conducting a quantitative user study to measure how the tool impacted users' ability to identify poor quality translations and we found that the tool significantly increased their accuracy on this task. We also demonstrated that users performed the task as accurately with VeriCAT's predicted quality scores as they did with human-generated quality scores.

1 Introduction

In 2017 Facebook's machine translation (MT) algorithm incorrectly translated a construction worker's Arabic-language post. The original post said "good morning" in Arabic, but was erroneously translated into Hebrew as "attack them," leading to the worker's arrest and several hours of questioning. No Arabic-speakers were asked to verify the machine translation of the post leading up to the arrest (Hern, 2017). For many people who use machine translation, it is easy to forget that translated output is susceptible to error and, as illustrated by this situation, that translation errors can lead to severe consequences. Our goal in this work was to develop and evaluate a tool to help users determine when, and whether, to trust machine translation.

Our prototype solution is called VeriCAT, short for **Veri**fication of **C**omputer-**A**ssisted **T**ranslation. The tool combines a Quality Estimation (QE) model with a simple user interface (UI). Our initial VeriCAT prototype assesses the translation quality of text snippets that have been translated from Russian into English by the FairSeq model (Ott et al., 2019). VeriCAT's QE model is a trained version of OpenKiwi's predictor-estimator model (Kim, et al., 2017). The UI displays output from this model – a predicted quality score, visualized as some value out of a possible 100 – for individual sentences of machine-translated text (Figure 1).

The objective of QE is to train a machine learning model to predict a quality score for translated text that is similar to what a human would assign to that translation (Maučec and Donaj, 2019). MT model developers typically use QE for validation and model improvement and previous work has demonstrated the value of QE in post-editing. VeriCAT is novel in its use of QE to provide feedback directly to MT end users.

Many MT accuracy metrics, such as BLEU score (Papineni et al., 2002), provide information about the accuracy of a MT model in general. For example, FairSeq has a BLEU score of 40.0 on Russian to English translation, calculated with the SacreBLEU standard (Post, 2018). These metrics, however, do not provide information about specific snippets of translated text. In contrast, a QE model outputs predictions about the translation quality of individual sentences. At the outset of this work, we hypothesized that predicted quality scores could benefit people using machine translated text, by helping them determine whether to trust a specific machine-translated sentence.

If a user of MT text is not a speaker of the original language and does not have additional contextual information, the only basis for judging translation quality is fluency, which refers to how well the text follows the target language's norms, taking into account grammar and clarity (Maučec and Donaj, 2019). In many cases, fluency is a reasonable proxy for translation quality. However, this is not always true; for example, one common source of MT error is incorrect or inconsistent translation of names and proper nouns.

To test our hypothesis that QE could help end users of MT, we designed and conducted a quantitative user study. In this study, we measured how well the VeriCAT prototype helped users perform a task that required them to assess the quality of Russian text snippets that were translated into English. The task incorporated examples of erroneous translations that appeared fluent, i.e. where fluency was not an adequate proxy for translation quality. The results of our user study indicated: (1) that study participants with access to VeriCAT's quality scores performed better on the task than those without; (2) that participants with access to sentence-level quality scores performed better than those who saw predictions of word-level errors; and (3) that participants who saw VeriCAT's predicted quality scores performed as well as those who saw human-generated quality scores (which we treated as ground-truth). These results illustrated the utility of VeriCAT and supported our hypothesis that QE can be a helpful source of feedback not only for MT model developers, but also to MT users.

2 Related Work

Recent advances have greatly improved the accuracy of machine translation, but human translators still outperform MT in overall accuracy and in preserving the original meaning of translated text (Maučec and Donaj, 2019). As MT becomes more widespread, translation inaccuracies are a greater concern. One way to address this issue is to use a second machine learning model to assess the quality of the MT model's output.

MT quality can be measured with a variety of metrics. Direct Assessment (DA) scores capture human judgements of translation fluency and adequacy (Snover et al., 2009); metrics such as BLEU, NIST, METEOR, and TER approximate human judgement through automated means (Maučec and Donaj, 2019); and HTER (human-mediated translation error rate) blends automated assessments and human judgments by capturing the number of post edits made to a MT by a human translator (Maučec and Donaj, 2019). Different metrics are suited for different applications. Some are appropriate for document-level quality assessment, whereas others can be used to train a QE model to predict the quality of specific sentences. HTER and DA have previously been used to train sentence-level QE models (Graham et al., 2017; Turchi et al., 2014).

There are some existing tools designed to make QE accessible to non-MT experts. Avramidis created a GUI for this purpose (2017), but it required users to be proficient in Python and the command line. Collins et al.'s lattice visualization illustrates uncertainty in MT text (2007), but we doubt that this tool (which was designed for an instant messaging use case) could scale to full passages of text. Albrecht et al.'s human-AI collaborative system uses visualization to help users gain intuition about a translation's source language to help them correct errors in MT (2009) and DeNeefe et al. developed an interactive tool called a DerivTool, which is intended to give users intuition about MT models (2005). However, VeriCAT is distinct in its use of QE to provide contextual information about particular snippets of text directly to MT users.

We found little prior work related to the usability of MT or QE systems, effective communication of MT error, or how QE might impact a user's ability to make decisions based on perceived translation quality. Martindale and Carpuat conducted a usability study to investigate whether revealing MT errors in fluency and adequacy might change users' trust in MT (2018). They found that poor fluency in translations significantly influenced users' trust of MT, but also, that trust was easily rebuilt. OpenKiwi (Kepler et al., 2019) performed a demo of a user interface for QE at ACL 2019; however, the team has not released their code, a demo, or a user study. There was a demonstration titled: "XAIT: An Interactive Website for Explainable AI for Text" at the IUI conference in 2020. However, when we reviewed the papers cited by this work we did not find any that related specifically to MT usability (Oduor et al., 2020).

3 VeriCAT System Overview

VeriCAT helps users assess the translation quality of Russian sentences that have been translated into English using FairSeq (Ott et al., 2019). Commonly used MT accuracy metrics such as BLEU score (Papineni et al., 2002) provide information about the accuracy of a MT model in general, but VeriCAT is unique in providing information about the translation quality of individual sentences through an accessible UI. The tool is particularly helpful when translation fluency is a poor proxy for translation quality.

3.1 Training Dataset

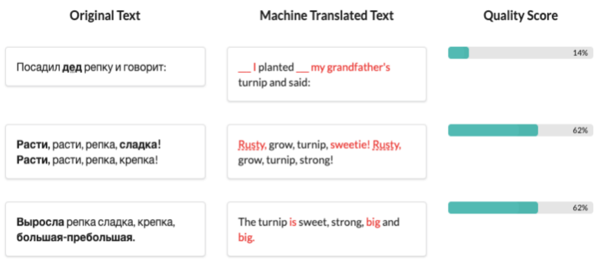
The VeriCAT QE model is finetuned on a dataset composed of 7,000 labeled sentence pairs. The sources of this text are passages from Reddit and Russian Proverbs from wikiquotes. The training dataset is curated from these sources because they represent types of text on which machine translation models are challenged. Each sentence is translated using the pretrained FairSeq model (Ott et al., 2019) that performed best at World Machine Translation 2019 (news task, the most recent results from the annual benchmark for MT). Each sentence has 3 Direct Assessment (DA) score quality judgments by human translators. These DA scores were labeled by ModelFront. Each DA score is rated on a scale from 1-100, with 100 representing a perfect translation. Across the dataset the average score was 68. These labeled data were contributed to the World Machine Translation Workshop (Nov 2020) as part of the Quality Estimation Shared Task (`https://statmt.org/wmt20`).

3.2 Quality Estimation (QE) Model

Quality Estimation benchmarks are set annually at the World Machine Translation QE Shared task. At the time of this study, the most accurate QE model available in the open source was the Predictor-Estimator model (Kim et al., 2017), open-sourced by OpenKiwi (Kepler et al., 2019) and the benchmark for WMT 2020. We pretrained the predictor model on the same parallel datasets that the FairSeq translation model (Ott et al., 2019) was trained on. We finetuned the estimator model on the novel Russian-English QE dataset described above, tuning the following hyperparameters from the baseline model: epochs, hidden LSTM layers, learning rate, batch size, and dropout. We obtained a Pearson correlation of 0.62 on the development set, which we used to test, since the shared task test set was not known to us. We ran inference with this model to generate the predictions used in our UI demo and confirmed the correlation between predicted and actual scores for this data subset was 0.67, which was in line with the model's expected performance.

3.3 User Interface (UI)

In the VeriCAT UI, a passage of text is broken down into individual sentences. For each sentence, users see the original (Russian) text, the FairSeq translation (English), and VeriCAT's quality score for that sentence (Figure 1). Quality scores are represented with a horizontal bar, where the percentage of the bar that is colored represents the score on a scale from 1 to 100. For clarity, the numerical value of the quality score is also displayed. These sentence-level quality scores are intended to help users quickly assess the translation quality for each sentence, to determine if the MT needs further inspection by a human. The demo UI also has the capacity to show predicted word-level errors in the translated text; these are highlighted in red.



Original Text	Machine Translated Text	Quality Score
Посадил **дед** репку и говорит:	__ I planted __ my grandfather's turnip and said:	14%
Расти, расти, репка, **сладка!** **Расти**, расти, репка, крепка!	Rusty, grow, turnip, sweetie! Rusty, grow, turnip, strong!	62%
Выросла репка сладка, крепка, **большая-пребольшая.**	The turnip is sweet, strong, big and big.	62%

Figure 1: The VeriCAT user interface.

4 Evaluation

To evaluate VeriCAT and test if QE could provide useful feedback directly MT users we conducted a between-subjects experiment with participants recruited via Amazon's Mechanical Turk.

4.1 Task & Study Procedure

In our user study experiment, we measured how well VeriCAT helped users perform a task that was analogous to that which motivated the design of the tool. We showed participants 3-sentence passages of text that had been translated from Russian into English with the FairSeq model (Ott et al., 2019). Participants were informed that a MT model had translated the passage and that they would be asked to answer comprehension questions based on the passage.

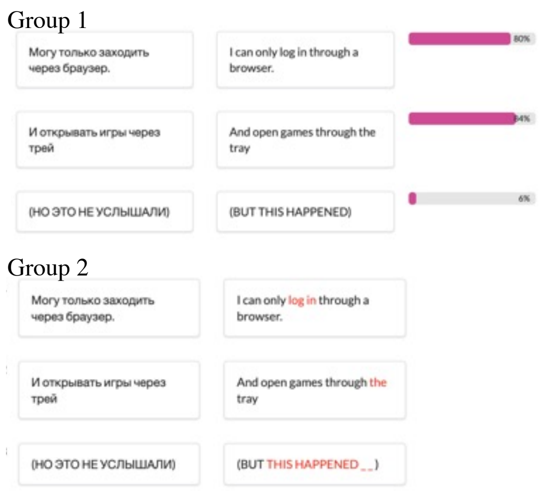
The quality of two of the translated sentences was good, but the third was poor. Prior to answering the comprehension questions, we gave participants (who did not speak Russian) the opportunity to select the sentence they believed was the poorest quality translation to be re-translated by a human. Participants also had the option not to select any sentence for re-translation. After providing the re-translation, we asked participants to answer two comprehension questions. We used their answers as an attention check, but for analysis, we scored participants' accuracy based on whether they choose the sentence with the lowest quality translation (measured by a human-generated Direct Assessment score) for re-translation.

4.2 Passage Type

Participants repeated this task for four passages. We designed these passages to test whether participants prioritized the quality scores or (their own intuition based on) fluency, as a primary indicator of poor translation quality. To do this we identified translated sentences in each of four categories: (1) good quality/good fluency, (2) good quality/ poor fluency, (3) poor quality/good fluency, (4) poor quality/poor quality. In our study, two passages include sentences of types (1), (2), and (3). The other two passages include sentences of types (1), (2) and (4). We expected categories (2) and (3) – where quality scores conflicted with fluency (and intuition) -- to be the most difficult for participants to assess correctly.

4.3 Experimental Conditions

We recruited 385 study participants via Amazon's Mechanical Turk and randomly assigned them to 5 groups. The **Baseline** group saw only the original (Russian) and translated (English) text snippets, with no additional information from the QE model. Participants in the other four groups were shown different versions of the VeriCAT UI each showing different information about translation quality. **Group 1** saw sentence-level quality scores, drawn from human-generated DA scores; **Group 2** saw word-level errors (highlighted in red) in each translated sentence; **Group 3** saw both sentence-level quality scores and word-level errors; and **Group 4** saw sentence-level quality scores predicted by VeriCAT's Quality Estimation model. These predicted scores were imperfect and reflected a smaller magnitude of difference between the scores of the “good” and “poor” quality sentences than the human-generated DA scores, but the poorly translated sentences always had the lowest predicted score. Two of these conditions are shown in Figure 2.



Group 1

Могу только заходить через браузер.

I can only log in through a browser.

И открывать игры через трей

And open games through the tray

(НО ЭТО НЕ УСЛЫШАЛИ)

(BUT THIS HAPPENED)

Group 2

Могу только заходить через браузер.

I can only log in through a browser.

И открывать игры через трей

And open games through the tray

(НО ЭТО НЕ УСЛЫШАЛИ)

(BUT THIS HAPPENED _ _)

Figure 2: Experimental conditions.

4.4 Hypotheses

[H1] We expected that participants in the Baseline condition (who received no information about translation quality from the QE model) would rely on fluency as a proxy for translation quality, following prior work by Martindale and Carpuat (2018). Therefore, we expected them to have lower accuracy on our experimental task.

[H2] We anticipated that different types of information about translation quality would not provide equivalent utility to users. Therefore, we expected to see differences in participants' performance across Groups 1, 2, and 3.

[H3, exploratory] We did not know how VeriCAT's predicted quality scores would compare to human-generated quality scores, in terms of their effect on users' performance, but this is what we hoped to investigate by comparing participants' performance in Groups 1 and 4.

4.5 Results Summary

To calculate participants' performance accuracy, we summed the number of correct answers across all four passages and divided by 4. Table 1 summarize results across all 4 passages and Figure 3 shows a more detailed breakdown of responses for Passages 1 and 2.

[H1] Participants with access to information about translation quality (displayed through the VeriCAT UI) performed the task better (i.e. they

	Correct selection	Incorrect selection	No selection
Baseline	6	44	50
Avg 1-4	39	34	27
Group 1	52	28	20
Group 2	24	36	40
Group 3	42	38	19
Group 4	37	33	30

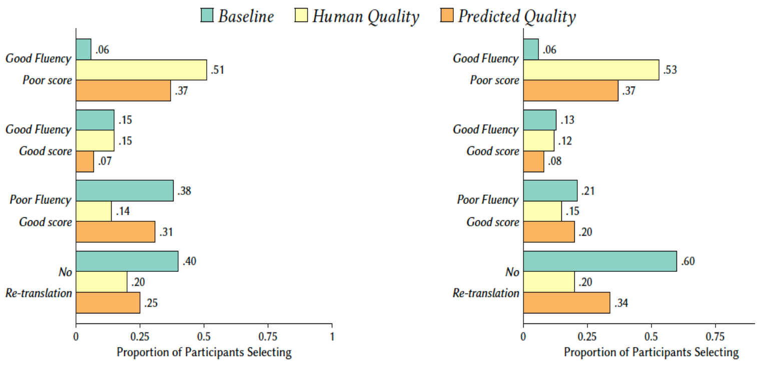
Table 1: Summary of Participant Responses (%)

more frequently selected the poorest quality sentence for re-translation by a human). Only 6% of participants in the baseline group selected correctly, compared to 39% of participants across Groups 1, 2, 3 and 4. Interestingly, we saw that participants in the baseline condition often opted for no-retranslation. Proportions of participants giving each answer for Passages 1 & 2 are shown in Figure 3. To test if the differences we observed were significant, we ran a Kruskal-Wallis test of *overall_score* ~ *condition* and found a significant difference across conditions ($H(2) = 29.7$, $p < 0.001$). A post-hoc Dunn's multiple comparisons test with a Bonferroni corrected alpha (0.02) showed significant pairwise differences between Human Quality and baseline ($Z = 5.2$, $p < 0.01$), and baseline and Predicted Quality ($Z = -4.3$, $p < 0.01$). **[H2]** Different types of information about MT quality did not provide equivalent utility to users. 52% of participants who saw sentence-level quality scores (Group 1) selected correctly, compared to 24% who saw word-level errors (Group 2). Access to additional information did not improve participants' performance. 52% of participants who saw sentence-level quality scores selected correctly (Group 1), compared to 42% who saw sentence-level quality scores *and* word-level errors (Group 3).

[H3] VeriCAT's predicted quality scores provided considerable utility. 37% of participants who saw predicted quality scores (Group 4) made the correct selection. While this was lower than the 52% in Group 1, it was notably higher than the 6% who selected correctly in the Baseline group. We interpreted this result as suggesting that the accuracy of VeriCAT's QE model, despite only .6 correlation with human judgement, still provided substantial value to end users. Figure 4 shows *overall_score* distributions by condition.

5 Lessons Learned

In this work, we sought to test whether QE could provide value directly to MT users, by helping them assess the translation quality of individual sentences. Our results indicated that VeriCAT, the tool we designed for this purpose, did substantially improve users' performance on a task that asked them to identify poor quality MT. While performance was slightly lower with predicted quality scores (from our QE model) than with human-generated quality scores, both led to significantly better performance than (users' intuitive assessment of) translation fluency. Additionally, our results suggested that some methods of displaying information about MT quality were better than others. Participants shown sentence-level quality scores significantly outperformed those who saw word-level errors. They also outperformed those who saw sentence-level scores *and* word-level errors, suggesting that the additional information provided by word-level errors was not helpful. This finding informed our thinking about VeriCAT's QE model and as a result, we decided not to develop word-level error prediction features. This is an example of how UI prototyping and evaluation can inform model design, by revealing which features provide the most value to end users.



(a) Passage 1

(b) Passage 2

Figure 3: Participant Responses for Passages 1 & 2.

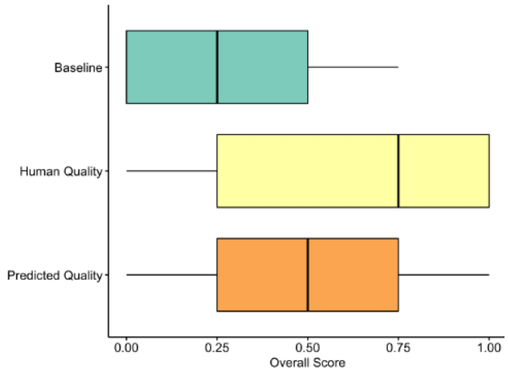


Figure 4: Distribution of Responses.

Links

Source code for VeriCAT is available at:
https://github.com/IQTLabs/VeriCAT
A demo is available here:
https://iqtlabs.github.io/VeriCAT-UI/

Acknowledgments

VeriCAT was designed, developed, and evaluated at IQT Labs, with support from Glenn Gaffney, Bob Gleichauf, and Zach Lyman. The authors would like to thank John Pope for his software development assistance, and Lucia Specia and Alex Fraser for their insightful comments and feedback.

References

Joshua Albrecht, Rebecca Hwa, and G Elisabeta Marai. 2009. The Chinese room: visualization and interaction to understand and correct ambiguous machine translation. In *Computer Graphics Forum*, Vol. 28. Wiley Online Library, 1047–1054.

Eleftherios Avramidis. 2017. QE:: GUI–A Graphical User Interface for Quality Estimation. *The Prague Bulletin of Mathematical Linguistics* 109 (10 2017). https://doi.org/10.1515/pralin-2017-0038

Christopher Collins, Sheelagh Carpendale, and Gerald Penn. 2007. Visualization of Uncertainty in Lattices to Support Decision-Making. In *Proceedings of Eurographics/IEEE VGTC Symposium on Visualization* (EuroVis 2007). 51–58.

Steve DeNeefe, Kevin Knight, and Hayward H Chan. 2005. Interactively exploring a machine translation model. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*. 97–100.

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 (2017), 3–30. https://doi.org/10.1017/S1351324915000339

Alex Hern. 2017. Facebook translates 'good morning' into 'attack them', leading to arrest. The Guardian (Oct 2017).

Fábio Kepler, Jonay Trénous, Marcos V. Treviso, Miguel Vera, and André F. T. Martins. 2019. OpenKiwi: An Open Source Framework for Quality Estimation. *CoRR* abs/1902.08646 (2019). arXiv:1902.08646 http://arxiv.org/abs/1902.08646

H. Kim, Jong-Hyeok Lee, and Seung-Hoon Na. 2017. Predictor-Estimator using Multilevel Task Learning with Stack Propagation for Neural Quality Estimation. In *WMT*.

Marianna J. Martindale and Marine Carpuat. 2018. Fluency Over Adequacy: A Pilot Study in Measuring User Trust in Imperfect MT. *CoRR* abs/1802.06041 (2018). arXiv:1802.06041 http://arxiv.org/abs/1802.06041

Mirjam Sepesy Maučec and Gregor Donaj. 2019. Machine Translation and the Evaluation of Its Quality. In *Recent Trends in Computational Intelligence*. IntechOpen.

Erick Oduor, Kun Qian, Yunyao Li, and Lucian Popa. 2020. XAIT: An Interactive Website for Explainable AI for Text. In *Proceedings of the 25th International Conference on Intelligent User Interfaces Companion* (Cagliari, Italy) (IUI '20). ACM, New York, NY, USA, 120–121. https://doi.org/10.1145/3379336.3381468

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics* (Demonstrations). ACL, Minneapolis, MN, 48–53. https://doi.org/10.18653/v1/N19-4009

Kishore Papineni, Salim Roukos, ToddWard, andWei-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, Philadelphia, PA, USA, 311–318. https://doi.org/10.3115/1073083.1073135

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers.* ACL, Brussels, Belgium, 186–191. https://doi.org/10.18653/v1/W18-6319

Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard 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.* ACL, Athens, Greece, 259–268. https://www.aclweb.org/anthology/W09-0441

Marco Turchi, Antonios Anastasopoulos, José GC de Souza, and Matteo Negri. 2014. Adaptive quality estimation for machine translation. In *Proceedings of the 52nd Annual Meeting of the ACL* (Volume 1: Long Papers). 710–720.