Investigating the Helpfulness of Word-Level Quality Estimation for Post-Editing Machine Translation Output

Raksha Shenoy[1,2], **Nico Herbig**[1], **Antonio Krüger**[1], **Josef van Genabith**[1,2]
[1]German Research Center for Artificial Intelligence (DFKI)
Saarland Informatics Campus, Germany
[2]Department of Language Science and Technology
Saarland University, Germany
rakshamanel08shenoy@gmail.com
{nico.herbig, krueger, josef.van_genabith}@dfki.de

Abstract

Compared to fully manual translation, post-editing (PE) machine translation (MT) output can save time and reduce errors. Automatic word-level quality estimation (QE) aims to predict the correctness of words in MT output and holds great promise to aid PE by flagging problematic output. Quality of QE is crucial, as incorrect QE might lead to translators missing errors or wasting time on already correct MT output. Achieving accurate automatic word-level QE is very hard, and it is currently not known (i) at what quality threshold QE is actually beginning to be useful for human PE, and (ii), how to best present word-level QE information to translators. In particular, should word-level QE visualization indicate uncertainty of the QE model or not? In this paper, we address both research questions with real and simulated word-level QE, visualizations, and user studies, where time, subjective ratings, and quality of the final translations are assessed. Results show that current word-level QE models are not yet good enough to support PE. Instead, quality levels of $\geq$ 80% F1 are required. For helpful quality levels, a visualization reflecting the uncertainty of the QE model is preferred. Our analysis further shows that speed gains achieved through QE are not merely a result of blindly trusting the QE system, but that the quality of the final translations also improves. The threshold results from the paper establish a quality goal for future word-level QE research.

1 Introduction

Advances in Machine Translation (MT) have made MT a key component in many professional translation workflows, where human post-editors identify and correct mistakes in raw MT output. Overall, Post-Editing (PE) saves time and reduces errors (Moorkens and Brien, 2017; Lagoudaki, 2009; Yamada, 2014; Green et al., 2013), however, for sentences of low MT quality, PE can take longer than manually translating from scratch (Specia et al., 2018). In order to better support and guide post-editors, it would therefore be helpful to indicate MT quality and flag potential errors.

Quality Estimation (QE) focuses on this goal by predicting how good MT output is based on factors such as fluency, comprehensibility, and adequacy (Moorkens et al., 2018). In contrast to standard MT evaluation methods such as BLEU (Papineni et al., 2002) and TER (Snover et al., 2006), QE does not rely on reference translations for comparison. Instead, QE can be framed as a supervised machine learning task, predicting the quality of an MT output for a source sentence (Turchi et al., 2013). QE models can be created at the word-level, sentence-level, and document-level (Turchi et al., 2014). Sentence-level scores can be leveraged for the basic decision whether to post-edit a sentence or to translate from scratch. Word-level quality estimates provide more fine-grained output that allows post-editors to quickly identify incorrect words, thereby aiming for faster translations of higher quality. However, if a QE prediction is incorrect, translators might waste time on already correct MT output or overlook errors. Apart from the quality of the QE system, the way its output is presented to the user can also have an impact on its usefulness. In particular, should the uncertainty of the model be communicated to the user or is a classification into correct/incorrect more suitable?

This paper focuses on word-level QE, which holds great potential in supporting PE but is also a very difficult machine learning task: even the best models currently only achieve F1 scores in the range 60% to 63% (Lee, 2020; Specia et al., 2020) depending on the text domain and the underlying MT system used to generate the data. While publicly available datasets and shared tasks are rapidly advancing QE research, we are not aware of any research that investigated the level of quality required to make word-level QE useful in prac-

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 10173–10185
November 7–11, 2021. ©2021 Association for Computational Linguistics

tical applications. In this paper, we address this research question in terms of a well-controlled experiment with professional translators who are presented with word-level QE output of varying quality, based on state-of-the-art QE systems and simulated QE output, with the goal of determining minimum quality thresholds required to support the PE process, thereby setting a target for future QE model developers. Since visualization of QE output might also have an impact on the helpfulness in PE, we compare two visualization techniques in our study. The results indicate that current state-of-the-art word-level QE models are not yet good enough to support PE, but that quality levels of at least 80% F1 are required. For the desired QE quality levels, a visualization showing the model's uncertainty is preferred.

2 Related Work

Word-level QE attempts to automatically mark words in an MT proposal, such that words requiring post-editing are labelled as 'BAD', while all others are tagged as 'OK' (Esplà-Gomis et al., 2018). This is usually cast as a binary classification task for each word (and gap) in an MT output. QE research is strongly advanced through shared tasks organized by the Workshop/Conference for Machine Translation (WMT) every year since 2012 (Callison-Burch et al., 2012). Until recently, WMT used the target F1 score as performance metric for word-level QE. In 2019 (Fonseca et al., 2019), the Matthews Correlation Coefficient (MCC, Matthews (1975)) was introduced in addition to the primary metric (F1), and became the new main metric in 2020 (Specia et al., 2020) due to its usefulness on unbalanced classes, since in word-based QE, OK tags are much more common than BAD tags. However, as works prior to 2019 did not report MCC, we will focus on F1 on the remainder of the paper.

The word-level model of Wang et al. (2018) won the WMT 2018 QE shared task (Specia et al., 2018) by achieving an F1 score of 62.46%. Their "QE-Brain" uses a pre-trained neural bilingual expert model (Fan et al., 2019), extracting semantic features from both the source and translation output for estimating translation quality with a bidirectional LSTM (Graves and Schmidhuber, 2005). In particular, three important strategies were utilized, namely incorporating human-crafted features, artificial QE data augmentation (Junczys-Dowmunt and Grundkiewicz, 2016) for more diversified training data, and a model ensemble with a greedy algorithm (Partalas et al., 2008). Kepler et al. (2019a) won the WMT 2019 word-level QE task (Fonseca et al., 2019) by combining linear, neural, and predictor-estimator systems with new transfer learning approaches using BERT (Devlin et al., 2019) and XLM (Lample and Conneau, 2019) pre-trained models, achieving an F1 score of 47.5%. In the most recent word-level QE shared task of 2020 (Specia et al., 2020), Lee (2020) proposed an XLM-R-based neural network architecture, which was trained in two phases. First, the model was trained with a very large artificially generated dataset based on a parallel corpus with OpenNMT-py (Klein et al., 2017) and the TER tool (Snover et al., 2006). Then, the model was fine-tuned with a human-labeled dataset. This approach achieved the highest F1 score of 61.89% among the submissions.

Note that the scores across the years are not directly comparable because the datasets used are different: as MT gets better, QE models need to adapt to work well on this changed output. Even though there has been considerable progress in QE and detailed analyses on datasets are done, to date there is very little research on the usability and impact of word-level QE in the PE process. In particular, no research has investigated whether current word-level QE models are already good enough to reduce translators' PE effort and increase translation quality and throughput, and if not, which level of QE quality is required to achieve this.

As discussed in Section 1, the way QE output is visualized to the human translator is also important for the PE process. Kepler et al. (2019b) propose to simply visualize the OK and BAD labels predicted by their state-of-the-art word-level QE system as either red or green words. Additionally, a gap indicating missing words is depicted by a red-colored underscore. Similarly for the sentence level, Turchi et al. (2015) use binary color-coded labels for visualizing MT quality information and assess whether this can lead to noticeable gains in translators' productivity: indeed, the authors show a statistically significant PE speedup, however, only for source sentences containing 5-20 words and a MT quality of more than 0.1 HTER. Inspired by the work of Turchi et al. (2015), Parra Escartín et al. (2017) modified the Post-Editing Tool (PET, Aziz et al. (2012)) to present translators with a traffic light system of sentence-level QE visualizations that indicates whether they need to translate the source text from scratch or post-edit the MT. Their results indicate that good and accurate MT QE, is vital to the efficiency of the translation workflow, and can cut translation time and effort significantly. Teixeira and O'Brien (2017) explore the impact of expressing the scores from different sentence-level QE models as percentages between 20% and 99%. Their study with 20 professional translators that post-edited four pieces of text each indicates that just displaying sentence-level percentage scores is not enough. Instead, one should also visualize word-level QE predictions.

To summarize, while word-level QE research has considerably advanced, with novel word-level QE systems emerging on a year by year basis, existing studies usually focus on detailed technical analyses on the test data, without user studies to investigate if the quality levels achieved by the QE systems are already useful for enhancing translation throughput or reducing errors in practical PE. Today, our best word-level QE models achieve F1 scores in and around 60% to 63%. In this paper we test whether this is sufficient for practical use, and if not, what quality levels are required. We do this through a user study with professional translators, comparing state-of-the-art QE model output with simulated QE output of higher quality levels created by manipulating ground truth data. Furthermore, we compare the binary red/green visualization proposed by Kepler et al. (2019b) to a color gradient-based visualization showing the uncertainty of the model.

3 Concept and Implementation

State-of-the-art word-level QE models roughly attain a quality level of 60% to 63% F1. This quality level may not yet be good enough to guide post-editors well: if the QE fails to detect errors, a post-editor relying on the QE might miss them more easily than without QE. Furthermore, correct output marked as incorrect could lead to wasted time trying to figure out what the mistake might be. Hence, incorrect predictions can have a severe impact on productivity and translation quality.

To ascertain which quality level would be sufficient to start helping, we artificially generate data at quality levels higher than what can be achieved by current state-of-the-art word-level QE models. The process of simulating QE output, as well as integrating their visualization into a translation environment, is explained in the following sections.

3.1 Artificial Generation of QE Output

We are not able to predict what exactly the output of a word-level QE model achieving 95% F1 would look like, i.e., which kinds of MT errors the model could detect well and which might be classified wrongly. The best we can do is to assume that a higher quality QE model would be similar to current QE models, but gradually improving on all parts of the MT output where current models fail. We therefore first conduct a pre-analysis to understand the kinds of errors of a current QE model, which we then leverage to generate artificial QE output of higher quality. For this, we flip labels of the ground truth annotations, while taking into consideration the parts of speech that current QE models are more likely to classify incorrectly, instead of flipping labels fully randomly.

3.1.1 Pre-Analysis

We use PoS (Part of Speech) information as a way of capturing error types of current word-level QE systems. First, the MT output from the training set of the WMT 2019 QE shared task is PoS tagged using the TextBlob library[1], which provides a simple API for common natural language processing tasks[2]. The PoS-tagged MT is then fed into the "QEBrain" model (Wang et al., 2018) to generate quality predictions. This QE model was chosen because it was the best performing system according to the assessment carried out by Shterionov et al. (2019). Next, we check how often each PoS is incorrectly labelled by comparing the QE output to the reference annotations. The probability of each PoS and the corresponding conditional error probability (given as $(P(PoS),\ P(error|PoS))$) are: nouns are most often wrong (32%, 48%), followed by prepositions (10.9%, 15.6%), pronouns (8.69%, 14.8%), determiners (13.04%, 14.3%), conjunctions (7%, 13.9%), interjections (4.5%, 12.9%), verbs (28%, 9.6%), adjectives (4.34%, 7.9%) and adverbs (5%, 2.9%).

3.1.2 Data Generation Process

We simulate the error behaviour of QE models achieving a certain quality level by flipping the ground truth QE label (OK or BAD) of the words depending on the conditional probability of the corresponding PoS. As an example, con-

[1] https://pypi.org/project/textblob/

[2] Since PoS tagging accuracy is fairly high for high-resource languages and general text domains, we expect limited impact of PoS errors on our results.

sider a PoS-tagged MT output with 5 words (Ich/PRON bin/AUX ein/DET Berliner/NOUN ./PUNCT), where, PRON stands for pronoun, AUX stands for auxiliary verb, DET stands for determiner, NOUN stands for noun and PUNCT stands for punctuation. A QE model with 100% F1 score would label all the words in the MT output correctly according to the ground truth data. For lower desired F1 scores, errors need to be introduced by flipping ground truth labels. Since, the error likelihood of nouns is highest, it should be more likely to flip a noun's label than that of an adverb. We use the following equation to determine the flip probability per PoS:

$$P(flip|PoS) = P(error|PoS) * \frac{F1_{base}}{F1_{target}}$$

where $P(error|PoS)$ is the conditional probability computed in the pre-analysis, $F1_{base}$ is the F1 score of the real QE model used in the pre-analysis, and $F1_{target}$ is the F1 quality score that we artificially generate. To get confidence scores for simulated QE models, we simply randomly sample values above 0.5 for 'OK' predictions, and values below 0.5 for 'BAD' predictions.

The limitations of our approach are that it assumes a constant error distribution, in the sense that higher quality QE models would just make proportionally fewer errors in each category, and that confidence scores are simply randomized. Of course, this is debatable, but, given that we cannot know exactly what a higher quality QE model would look like, we believe that this simple approach is a reasonable starting point for estimating the threshold when word-level QE stops confusing and starts helping the PE process.

3.2 Visualization of QE output

Apart from QE quality, the visualization of QE output might also impact whether QE helps or hinders PE. We designed two alternatives, called *Binary*- and *Gradient*-based visualization schemes, as shown in Figure 1. In the binary visualization, quality is represented by simply coloring words in green and red depending on the QE output based on a level threshold of 0.5[3]. While this seems intuitive and easy to understand, uncertainties of the QE model cannot be depicted in the binary visualization. To tackle this, the gradient-based visualization directly shows the floating point number output of the QE model in the interval [0....1] by mapping the output to a color gradient ranging from red to green. Thus, the darker the shade of green, the more correct the model estimates the word to be. At the same time, this additional information about model uncertainties may well be confusing or overwhelming for the human post-editors.

3.3 QE Integration into CAT Environment

A Computer-Aided Translation (CAT) tool or PE environment allows the capture and correction of mistakes, as well as the selection, manipulation, adaptation and recombination of good segments (Herbig et al., 2020c). Our implementation was done within the MMPE CAT tool[4] (Herbig et al., 2019, 2020a,b,c; Jamara et al., 2021), as it is open source and easily extendable.

The frontend is the main component of MMPE and is developed using Angular. The backend is implemented using node.js and is used for saving and loading of projects from JSON files. The interface has a horizontal source (left) - target (right) layout with the current segment enlarged. MMPE offers undo, redo and segment confirmation features either using hotkeys or through buttons between source and target. We extend MMPE's project file structure for QE: the (real and simulated) QE models' quality predictions per word of the MT output, and a value indicating which visualization mode to use for the segment (binary or gradient-based), are stored in and loaded from a JSON file. Figure 2 illustrates the MMPE interface with the QE extension. The logging functionality was also extended to capture the QE condition and visualization.

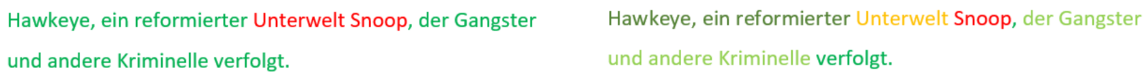
When a user manually changes a word flagged by the QE system, its color is changed to black because we assume post-edits done by a user to be correct. Ideally, we could re-run the QE to obtain new scores for all the unchanged parts of the segment after each edit. However, as our simulated QE models rely on ground truth data, which we only have for the original MT output without modification, this is not a possibility.

4 Evaluation

Using our implementation, we conducted a user study to assess the quality threshold when word-level QE starts facilitating and stops hindering the

[3] The chosen threshold can trade off how sensitive the shown QE annotations are, so it can potentially trade off editing time for correctness. Our approach of using 0.5 is a simple and straight-forward starting point, also often used in logistic regression and similar classification by just showing the tendency of the model; nevertheless future research should investigate different thresholds.

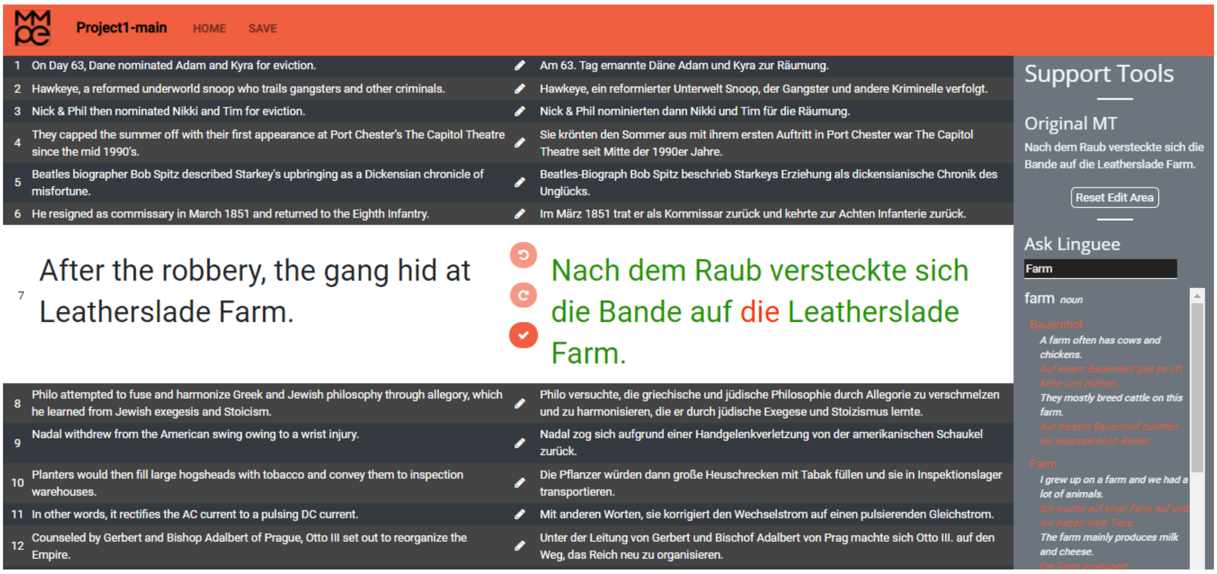
[4] https://github.com/NicoHerbig/MMPE



Hawkeye, ein reformierter Unterwelt Snoop, der Gangster und andere Kriminelle verfolgt.

Hawkeye, ein reformierter Unterwelt Snoop, der Gangster und andere Kriminelle verfolgt.

Figure 1: Binary (left) and gradient-based (right) visualization schemes.



1 On Day 63, Dane nominated Adam and Kyra for eviction. — Am 63. Tag ernannte Däne Adam und Kyra zur Räumung.

2 Hawkeye, a reformed underworld snoop who trails gangsters and other criminals. — Hawkeye, ein reformierter Unterwelt Snoop, der Gangster und andere Kriminelle verfolgt.

3 Nick & Phil then nominated Nikki and Tim for eviction. — Nick & Phil nominierten dann Nikki und Tim für die Räumung.

4 They capped the summer off with their first appearance at Port Chester's The Capitol Theatre since the mid 1990's. — Sie krönten den Sommer aus mit ihrem ersten Auftritt in Port Chester war The Capitol Theatre seit Mitte der 1990er Jahre.

5 Beatles biographer Bob Spitz described Starkey's upbringing as a Dickensian chronicle of misfortune. — Beatles-Biograph Bob Spitz beschrieb Starkeys Erziehung als dickensianische Chronik des Unglücks.

6 He resigned as commissary in March 1851 and returned to the Eighth Infantry. — Im März 1851 trat er als Kommissar zurück und kehrte zur Achten Infanterie zurück.

7 After the robbery, the gang hid at Leatherslade Farm. — Nach dem Raub versteckte sich die Bande auf die Leatherslade Farm.

8 Philo attempted to fuse and harmonize Greek and Jewish philosophy through allegory, which he learned from Jewish exegesis and Stoicism. — Philo versuchte, die griechische und jüdische Philosophie durch Allegorie zu verschmelzen und zu harmonisieren, die er durch jüdische Exegese und Stoizismus lernte.

9 Nadal withdrew from the American swing owing to a wrist injury. — Nadal zog sich aufgrund einer Handgelenkverletzung von der amerikanischen Schaukel zurück.

10 Planters would then fill large hogsheads with tobacco and convey them to inspection warehouses. — Die Pflanzer würden dann große Heuschrecken mit Tabak füllen und sie in Inspektionslager transportieren.

11 In other words, it rectifies the AC current to a pulsing DC current. — Mit anderen Worten, sie korrigiert den Wechselstrom auf einen pulsierenden Gleichstrom.

12 Counseled by Gerbert and Bishop Adalbert of Prague, Otto III set out to reorganize the Empire. — Unter der Leitung von Gerbert und Bischof Adalbert von Prag machte sich Otto III. auf den Weg, das Reich neu zu organisieren.

Support Tools

Original MT

Nach dem Raub versteckte sich die Bande auf die Leatherslade Farm.

Ask Linguee

Farm

farm *noun*

Bauernhof

A farm often has cows and chickens.

Auf einem Bauernhof gibt es oft Kühe und Hühner.

They mostly breed cattle on this farm.

Auf diesem Bauernhof züchten sie hauptsächlich Rinder.

Farm

I grew up on a farm and we had a lot of animals.

Ich wuchs auf einer Farm auf und wir hatten viele Tiere.

The farm mainly produces milk and cheese.

Die Farm produziert

Figure 2: Screenshot of the interface after integrating QE information.

PE process. We further want to find out which word-level QE visualization is preferred.

4.1 Method

Due to the ongoing COVID-19 pandemic the evaluation was conducted online[5] by 17 participants. The study took approximately one hour per participant and involved three separate stages. First, participants filled in a questionnaire capturing demographics as well as information on their translation skills, post-editing skills, and CAT usage. Stage two was the main experiment, where participants had to post-edit segments with different quality QE annotations in the two visualizations, as described in detail below. In order to successfully perform stage two of the experiment, the participants received an explanation of all of the prototype's features and pointers regarding the execution of the main experiment in a four minute introductory video[6]. After the video, translators were asked to try out MMPE with word-level QE in a trial project to get accustomed to the environment (for which we did not record data) before delving into the main experiment. In the final stage, the participants filled out another questionnaire capturing their experience during the experiment, pain points, and other feedback.

The main experiment was the central part of the study and entailed post-editing of 32 text segments (8 text blocks of 4 segments each) with QE support, chosen according to the text selection step described in Section 4.2 below. The segments were labeled by either the real QE model or simulated QE output with 75% F1, 85% F1 and 95% F1 quality levels created as detailed in Section 3.1.2. The real QE model proposed by Wang et al. (2018) was pre-trained on the training set of WMT 2018's QE shared task and fine-tuned on the Wikipedia domain, achieving a quality level of 63.5% F1 on the training set[7] of WMT 2020's QE shared task. QE information for each sentence was visualized either with the binary or gradient-based visualization. Since there are four quality levels and two visualization schemes, the experiment follows an 8*8 Balanced Latin Square: the text order is kept identical for each participant, but the QE quality and visualization are counter-balanced accordingly on the 8 text blocks of 4 sentences each. Thus, participants might be more exhausted for the same sentences towards the end of the experiment and better concentrated on the initial sentences; how-

[5] The study has been approved by the university's ethical review board.

[6] https://youtu.be/6LgUzia_3pM

[7] Note that the model was not trained on this data.

ever, the effects of text and tiredness should cancel out for quality level and visualization due to the counter-balancing. This methodology allows us to analyze if visualization-x with QE-y is better than visualization-x' with QE-y' across text blocks. Moreover, the impact of translation skill or technical skill of first-time users of MMPE factors into all the conditions equally due to the chosen within-subject design.

After post-editing each segment, participants had to press confirm, which we used to record the required time. When confirming, we also showed a pop-up asking "Was the word-level quality estimation helpful?" which the participant had to rate on a 9-point Likert scale. Apart from *duration* and *subjective ratings*, we measure per condition *edits* done and the *final quality* of the translations in terms of TER (Snover et al., 2006) by comparing the post-edited version of the translation to the reference. TER was used because of its popularity as an automatic translation quality metric, and because it counts edit operations and therefore reflects PE effort. Naturally, we also store the QE model and visualization used per segment to compare the different conditions.

4.2 Text Selection

In line with previous recent QE research, we used data from the training set of the WMT 2020 QE shared task (Specia et al., 2020), which relies on an up-to-date NMT model based on an attentional encoder-decoder architecture built using the fairseq toolkit (Ott et al., 2019). The training set consists of source, MT, reference, MT quality scores, tags, and source-MT alignment information. We used a common language pair (namely EN-DE) and selected text from a general domain (namely Wikipedia) to ease participant recruitment as it does not require specific domain expertise. In order to understand whether QE helps more or less with different MT quality, we selected sentences with low, medium and high MT quality for different sentence lengths.

The selection of the text segments follows the following steps:

- We analyzed the length distribution of all sentences in the dataset, leading to a categorization into three groups: short with 7 to 12 words, medium with 13 to 19 words and long with 20 to 30 words. Given that the experiment requires 32 sentences, roughly reflecting the frequency distributions this amounts to 9 short sentences, 15 medium length sentences, and 8 long sentences.
- We randomly sample sentences from these length categories. To ensure that their MT quality levels roughly match the overall MT quality distribution in the dataset, we plot and compare their MT quality distribution against the MT quality distribution of the whole dataset (for short, medium, and long sentences). This avoids accidental random fluctuations that might bias the data.

This approach provides us with a suitable dataset for our study which relies on high quality MT data, represents a realistic distribution of sentence lengths, and realistic MT qualities.

4.3 Participants

The prototype was evaluated by non color-blind professional translators and translation students who were working as freelancers on a platform called Upwork Global Inc[8]. We used EN–DE text, and therefore recruited participants that were well-versed in English and German, having either C1 or C2 level of proficiency in both languages.

Overall 16 (f=9, m=7) professional translators and one translation student (f=1) participated in the experiment. Their ages ranged from 20 to 65 (avg=33.26, sigma=9.11), with 8 months to 38 years of professional experience (avg=17.65, sigma=9.66), and offering a total of 7 language pairs (avg=2). For most participants the self-assessed CAT knowledge was good (8 times) or very good (2). However, participants were less confident about their PE skills (4 bad, 6 neutral, 2 good, 5 very good). Their years of CAT tool experience ranged from 0 to 18 years (avg=9, sigma=5.23), where participants had used between 1 and 9 distinct CAT tools (avg=4.39, sigma=2.18), most often Trados Studio, Across, Transit, and MemoQ.

4.4 Results

We present our results in 4 categories: (1) subjectively assessed helpfulness per QE quality, (2) preferred visualization, (3) editing duration per QE quality, and (4) translation quality per QE quality.

4.4.1 Subjective Helpfulness of QE Quality

We analyze subjective ratings across all segments with the same word-level QE quality level. To

[8] https://www.upwork.com/

ensure independence of samples, we average the ratings for the same QE quality level per participant. The box plot in Figure 3 shows that lower QE quality levels of 63.5% F1 and 75% F1 are consistently rated as less helpful, receiving mean values of 1.5 and 3.25 respectively on our 9-point Likert scale for the question "Was the word-level quality estimation helpful?". The scale ranged from 1 representing "very strongly disagree" to 9 representing "very strongly agree". In contrast, the higher quality levels of 85% F1 and 95% F1 are rated more helpful with mean subjective ratings of 7 and 8, respectively. This indicates that in comparison to high quality QE, bad QE is not considered helpful. Whether it is helpful in terms of productivity in comparison to "no-QE" is something we cannot show with our data, as there was no "no-QE" condition. We did consider adding such a condition to the experiment, but with 4 quality levels and 2 visualizations we had a rather complex study setup that we did not want to make even more complex.

We test the results for each group with a two-tailed t-test for significance against 5, which is the middle value along the subjective rating scale depicting neutrality. The results indicate positive significant differences, depicting helpfulness, for the quality levels 85% F1 (with $t(16) = 1.746, p < 0.05$, Cohen's $d = 0.90$) and 95% F1 (with $t(16) = 1.746, p < 0.05$, Cohen's $d = 0.93$). Furthermore, we find negative significant differences, meaning that the quality levels are not helpful, for the levels 63.5% F1 (with $t(16) = 1.746, p < 0.05$, Cohen's $d = 0.82$) and 75% F1 (with $t(16) = 1.746, p < 0.05$, Cohen's $d = 0.85$).

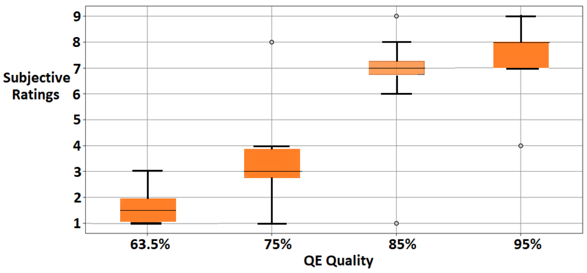


Figure 3: A box plot with QE quality levels along the X-axis and the subjective ratings for helpfulness on a 9-point Likert scale along the Y-axis.

4.4.2 Preferred Visualization per QE Quality

To analyze preference of visualization, we use the same ratings, however, we multiply the ratings corresponding to sentences shown in binary visualization by -1 (and multiply the ratings for sentences visualized using a color gradient by $+1$). As before, independence of samples is achieved by averaging ratings per participant across segments with the same QE quality and visualization. The obtained scores are then normalized to the range [0...1]. Therefore, values close to 0 indicate a preference for the binary visualization scheme, while values close to 1 indicate a preference for the gradient-based visualization.

The corresponding box-plot in Figure 4 shows that for word-level QE quality levels 63.5% F1 and 75% F1 binary is the preferred visualization scheme (preference value below 0.5). In contrast, for higher quality levels of 85% F1 and 95% F1, the gradient-based visualization is preferred (preference value above 0.5). We test the results for significance using a two-tailed t-tests against 0.5, which is the value corresponding to an equal preference towards both visualizations. The results indicate positive significant differences, meaning that there is a significant preference towards the gradient visualization scheme, for quality levels 85% F1 (with $t(16) = 1.746, p < 0.05$, Cohen's $d = 0.90$) and 95% F1 (with $t(16) = 1.746, p < 0.05$, Cohen's $d = 0.96$). Furthermore, we find negative significant differences, meaning that participants would rather prefer the binary visualization scheme, for quality levels 63.5% F1 (with $t(16) = 1.746, p < 0.05$, Cohen's $d = 0.83$) and 75% F1 (with $t(16) = 1.746, p < 0.05$, Cohen's $d = 0.88$).

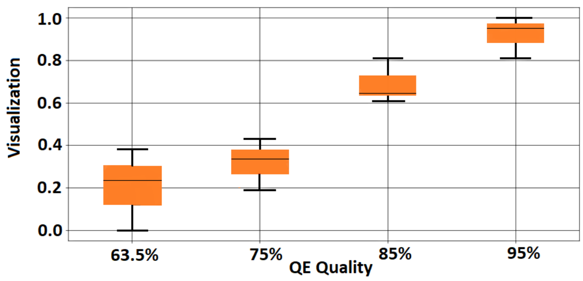


Figure 4: A box plot with QE quality levels along the X-axis and visualization along the Y-axis.

4.4.3 Editing Duration per QE Quality

Apart from subjective ratings (used for QE helpfulness and visualization preference), we also capture the time taken to post-edit the segments per QE quality. We average the duration across the segments having the same QE quality per participant to make the observations independent within and among the groups. The box plot in Figure 5 depicts that when the QE quality is low the duration taken to post-edit the segments is high, whereas translators are fast when the QE quality is high. In order to find out whether the differences in duration are significant, we first run a one-way ANOVA. The results indicate that there is a significant difference between at least two of these groups (with $F(3,31) = 2.9223, p < 0.05$, $\eta^2 = 0.35$). To examine where the group differences lie, we perform the Tukey HSD post-hoc test, showing that all pairs except for 85% F1 vs. 95% F1 are significantly different.

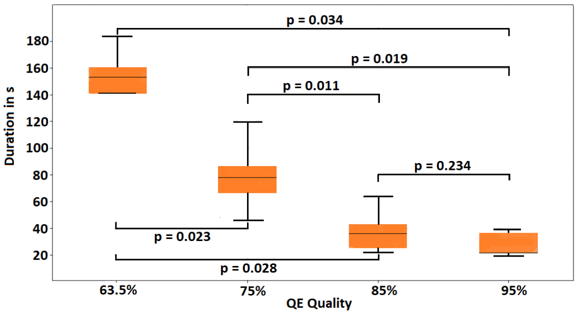


Figure 5: A box plot with QE quality levels along X-axis and duration in seconds along Y-axis, with p-values of Tukey HSD for pairwise comparisons.

4.4.4 Translation Quality per QE Quality

We have seen that the translators subjectively find high quality QE helpful and post-edit fast with it. In order to analyze whether they are fast just because they blindly trust and follow the QE system (even when they should not) or because the system actually helps, we evaluate the quality of the resulting translations. As before, the scores were averaged across segments having the same QE model per participant to make the observations independent within and among the groups. The box plot in Figure 6 shows that the TER of the post-edited version against the reference is low when the QE quality is high, and by contrast, it is high when the QE quality is low. Since a low TER score implies a high quality translation, translations get better with increasing QE quality.

We speculate that the reason for the high TER score for the 63.5% F1 QE model is that the translators produced a different translation than the reference. This different translation may or may not be accurate; we cannot know for sure without manual evaluation. Nonetheless, from our automatic quality evaluation we are certain that with better QE the final translations get closer to the reference. In order to find out whether the differences are significant, we first run a one-way ANOVA, showing that there is a significant difference between at least two of these groups (with $F(3,31) = 2.9223, p < 0.05$, $\eta^2 = 0.41$). The follow-up Tukey HSD test shows that indeed all pairs are significantly different.

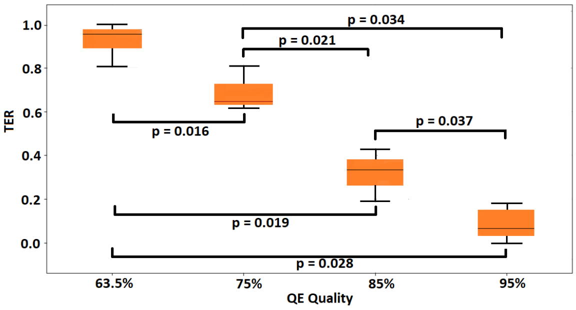


Figure 6: A box plot with QE quality levels along X-axis and TER scores along Y-axis, with p-values of Tukey HSD for pairwise comparisons.

5 Discussion

Our results show that existing state-of-the-art word-level QE systems are not yet good enough to be helpful during PE. Instead, all our analyses agree that QE systems need an F1 score of at least 80% to support PE in terms of subjective helpfulness, editing duration, and quality of the final translations. This establishes a target for future QE research. In terms of visualization, the word-level quality scores should be visualized using gradient-based visualization which also shows uncertainties of the model, since the binary approach was considered superior only in cases where QE was not helpful. This preference is interesting as the exact color chosen for the gradient was randomly sampled in the red/green range for BAD/OK ratings for the artificial QE output. A reason for the preference could be a stronger involvement in the decision process and hence a stronger feeling of control.

As expected, with increasing QE quality, PE becomes more efficient, where in particular the higher quality levels of 85% and 95% require less editing time than the lower quality levels. Lastly, we found that translation speed gains obtained by QE are not merely a result of blindly trusting the QE system, but indeed help producing higher quality translations. To sum up, a QE quality level of at least 80% F1 sets the approximate boundary where word-level QE starts helping translators, and for these QE quality levels, a gradient-based visualization is preferred.

6 Conclusion and Future Work

The goal of this paper was to estimate how accurate a word-level QE system would need to be in order to support the PE process. Furthermore, we also test how to best visualize the output of such a system. Since we hypothesized that state-of-the-art QE systems may not yet be good enough to aid the PE process, we developed an approach to generate higher quality levels artificially. We performed a user study where the output from a state-of-the-art QE model and three artificial QE models were presented to users in a CAT tool using either a binary or gradient-based red/green visualization, where the latter also shows uncertainties of the QE model. The results of the evaluation show that current word-level QE models are not yet good enough to guide post-editors; instead, quality levels of at least 80% F1 serves as a reasonable first approximate boundary required to aid the PE process. For these quality levels, gradient is the preferred visualization scheme.

In the future, we plan to explore the impact of QE models' performance on different language pairs and text domains. Furthermore, we will investigate how sentence-level and word-level QE can best be combined for efficient PE. As our current quality analysis only verified that a QE hinting towards a certain reference indeed leads to translations close to that reference, we would like to extend our analysis with a manual human translation quality analysis that is not bound to a comparison to a single reference. Besides, we want to explore further non-color-based visualization schemes. Finally, in addition to having a word quality prediction task, we want to also explore gap prediction and how this can be best visualized.

7 Acknowledgments

This research was funded in part by the German Research Foundation (DFG) under grant number GE 2819/2-1 (project MMPE). We thank all participants of the user study for their valuable feedback.

References

W. Aziz, Sheila Castilho, and Lucia Specia. 2012. Pet: a tool for post-editing and assessing machine translation. In *Proceedings of the 16th Annual conference of the European Association for Machine Translation*, page 99, Trento, Italy. European Association for Machine Translation.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Miquel Esplà-Gomis, Felipe Sánchez-Martínez, and Mikel Forcada. 2018. Predicting insertion positions in word-level machine translation quality estimation. *Applied Soft Computing*, 76:174–192.

Kai Fan, Bo Li, Fengming Zhou, and Jiayi Wang. 2019. "bilingual expert" can find translation errors. *Proceedings of the AAAI Conference on Artificial Intelligence*, abs/1807.09433:6367–6374.

Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the WMT 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602–610. International Joint Conference on Neural Networks.

Spence Green, Jeffrey Heer, and Christopher D. Manning. 2013. The efficacy of human post-editing for language translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, page 439–448. Association for Computing Machinery.

Nico Herbig, Tim Düwel, Santanu Pal, Kalliopi Meladaki, Mahsa Monshizadeh, Antonio Krüger, and Josef van Genabith. 2020a. MMPE: A multi-modal interface for post-editing machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1691–1702, Online. Association for Computational Linguistics.

Nico Herbig, Santanu Pal, Tim Düwel, Kalliopi Meladaki, Mahsa Monshizadeh, Vladislav Hnatovskiy, Antonio Krüger, and Josef van Genabith. 2020b. MMPE: A multi-modal interface using handwriting, touch reordering, and speech commands for post-editing machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 327–334. Association for Computational Linguistics.

Nico Herbig, Santanu Pal, Tim Düwel, Raksha Shenoy, Antonio Krüger, and Josef van Genabith. 2020c. Improving the multi-modal post-editing (MMPE) CAT environment based on professional translators' feedback. In *Proceedings of 1st Workshop on Post-Editing in Modern-Day Translation*, pages 93–108, Virtual. Association for Machine Translation in the Americas.

Nico Herbig, Santanu Pal, Josef van Genabith, and Antonio Krüger. 2019. Multi-modal approaches for post-editing machine translation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 231. Association for Computing Machinery.

Rashad Albo Jamara, Nico Herbig, Antonio Krüger, and Josef van Genabith. 2021. Mid-air hand gestures for post-editing of machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 6763–6773. Association for Computational Linguistics.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2016. Log-linear combinations of monolingual and bilingual neural machine translation models for automatic post-editing. In *Proceedings of the First Conference on Machine Translation*, pages 751–758, Berlin, Germany. Association for Computational Linguistics.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, António Góis, M. Amin Farajian, António V. Lopes, and André F. T. Martins. 2019a. Unbabel's participation in the WMT19 translation quality estimation shared task. In *Proceedings of the Fourth Conference on Machine Translation*, pages 78–84, Florence, Italy. Association for Computational Linguistics.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. 2019b. OpenKiwi: An open source framework for quality estimation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 117–122, Florence, Italy. Association for Computational Linguistics.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (2017)*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Elina Lagoudaki. 2009. Translation editing environments. In *In MT Summit XII: Workshop on Beyond Translation Memories (2009)*, pages 42–63, Ottawa, Canada.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Computing Research Repository*, abs/1901.07291.

Dongjun Lee. 2020. Two-phase cross-lingual language model fine-tuning for machine translation quality estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1024–1028, Online. Association for Computational Linguistics.

B.W. Matthews. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442–451.

Joss Moorkens and Sharon Brien. 2017. *Assessing User Interface Needs of Post-Editors of Machine Translation*, pages 109–130. Routledge.

Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty. 2018. *Translation Quality Assessment From Principles to Practice*. Springer.

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*, pages 48–53. Association for Computational Linguistics.

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 on Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Carla Parra Escartín, Hanna Béchara, and Constantin Orasan. 2017. Questing for quality estimation a user study. *The Prague Bulletin of Mathematical Linguistics*, pages 343–354.

Ioannis Partalas, Grigorios Tsoumakas, and I. Vlahavas. 2008. Focused ensemble selection: A diversity-based method for greedy ensemble selection. In *Proceedings of ECAI 2008, 18th European Conference on Artificial Intelligence*, pages 117–121. IOS Press.

Dimitar Shterionov, Félix Do Carmo, Joss Moorkens, Eric Paquin, Dag Schmidtke, Declan Groves, and Andy Way. 2019. When less is more in neural quality estimation of machine translation. an industry case study. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, pages 228–235, Dublin, Ireland. European Association for Machine Translation.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231. Association for Machine Translation in the Americas.

L. Specia, C. Scarton, G. H. Paetzold, and G. Hirst. 2018. *Quality Estimation for Machine Translation*. 162. Morgan and Claypool Publishers.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. 2020. Findings of the WMT 2020 shared task on quality estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 743–764. Association for Computational Linguistics.

Lucia Specia, Frédéric Blain, Varvara Logacheva, Ramón F. Astudillo, and André F. T. Martins. 2018. Findings of the WMT 2018 shared task on quality estimation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 689–709, Belgium, Brussels. Association for Computational Linguistics.

Carlos Teixeira and Sharon O'Brien. 2017. The impact of mt quality estimation on post-editing effort. In *Proceedings of MT Summit XVI*, Nagoya, Japan. academia.edu.

Marco Turchi, Antonios Anastasopoulos, José G. C. de Souza, and Matteo Negri. 2014. Adaptive quality estimation for machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (2014)*, pages 710–720. Association for Computational Linguistics.

Marco Turchi, Matteo Negri, and Marcello Federico. 2013. Coping with the subjectivity of human judgements in MT quality estimation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 240–251, Sofia, Bulgaria. Association for Computational Linguistics.

Marco Turchi, Matteo Negri, and Marcello Federico. 2015. MT quality estimation for computer-assisted translation: Does it really help? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 530–535, Beijing, China. Association for Computational Linguistics.

Jiayi Wang, Kai Fan, Bo Li, Fengming Zhou, Boxing Chen, Yangbin Shi, and Luo Si. 2018. Alibaba submission for WMT18 quality estimation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 809–815, Belgium, Brussels. Association for Computational Linguistics.

Masaru Yamada. 2014. Can college students be post-editors? an investigation into employing language learners in machine translation plus post-editing settings. *Machine Translation*, 29:49–67.

A Appendix

For completeness, this appendix provides further analyses regarding preferred visualization as well as the impact of the MT quality on editing duration and translation quality.

A.1 Editing Duration per Visualization

In the main paper, we have focused on the preference for the two visualization schemes across different QE quality levels. Here, we extend our results by also investigating the duration when editing within a visualization. As we found that word-level QE only facilitates PE with a quality level of at least 80% F1, we plot the duration taken to post-edit the segments per visualization for QE quality levels of 85% and 95% F1. For this, we average the duration across the segments having the same visualization per participant to make the observations independent within and among the groups. The box plot in Figure 7 suggests that in terms of duration, results are roughly comparable with a slightly smaller editing time for gradient than for binary. In order to find out if the difference in duration between the two groups is significant, we run Welch's t test. The results indicate that there is not sufficient evidence to say that the means of the two groups are significantly different ($p > 0.05$).

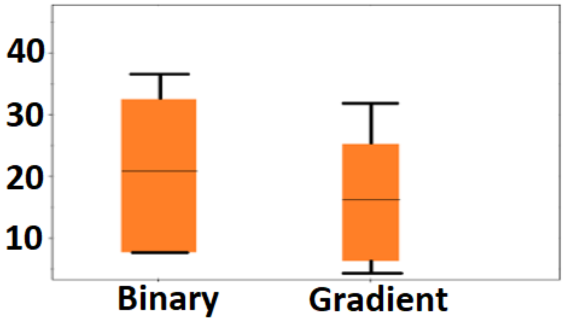


Figure 7: A box plot having visualization along X-axis and duration in seconds along Y-axis.

A.2 Translation Quality per Visualization

Similarly, we plot the TER of the post-edited version against the reference per visualization for QE quality levels of 85% and 95% F1 in Figure 8. We see that in terms of TER, results are also roughly comparable with a slight tendency for better results for binary compared to gradient. Note that TER is relative to the given reference translation (not just any correct translation), and the QE also targets this reference, thus, a lower TER only implies closeness to the reference but not necessarily an overall better translation. In order to find out whether the difference in TER scores between the two groups is significant, we run Welch's t test. The results indicate that there is not sufficient evidence to say that the means of the two groups are significantly different (p > 0.05).

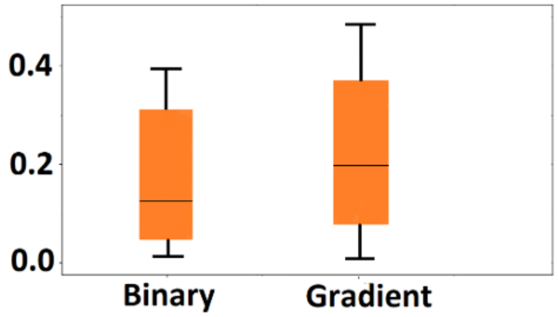


Figure 8: A box plot having Visualization along X-axis and TER scores along Y-axis.

A.3 Preferred Visualization per MT and QE Quality

In order to investigate whether the preference for a visualization scheme depends on the MT quality, we analyzed our data not just by QE quality, but also by MT quality: Figure 9 shows the preference score per QE quality and MT quality, where MT quality is categorized based on TER into three levels: High $[0, 0.192]$, Medium $[0.193, 0.33]$, and Low $[0.34, 0.55]$. The 3D plot shows that participants prefer the visualizations mainly based on QE quality, whereas the MT quality has less influence on visualization preference.

A.4 Editing Duration per MT and QE Quality

Similarly, we investigate if the editing duration depends not only on the QE quality but also on the MT quality: Figure 10 shows the MT quality distribution of the chosen sentences categorized into the same three levels of MT quality. Again, the 3D plot shows that the time taken to post-edit depends mainly on the QE quality, and only to a small extent on the MT quality.

A.5 Translation Quality per MT and QE Quality

Finally, we investigate the impact of MT quality (besides QE quality) on the final translation quality in terms of TER: Figure 11 shows the TER per QE and MT quality level (using the same quality ranges as above). Again, the 3D plot shows that the effects are mainly driven by the QE quality and rather independent of the MT quality.

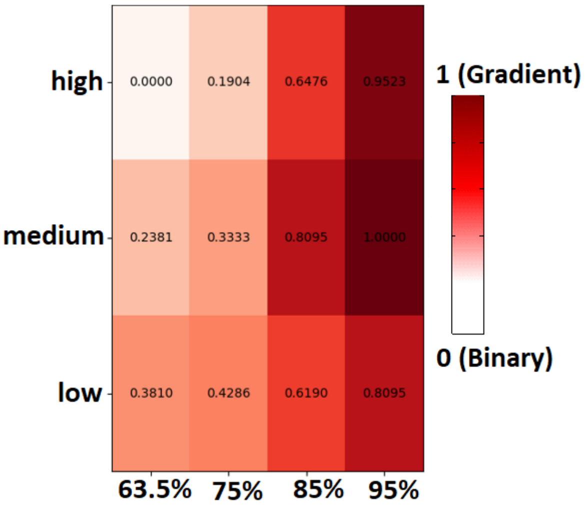


Figure 9: A 3D plot having QE quality levels along X-axis and MT quality along Y-axis. The color ranges from white depicting preference towards binary visualization to red depicting preference towards gradient visualization.

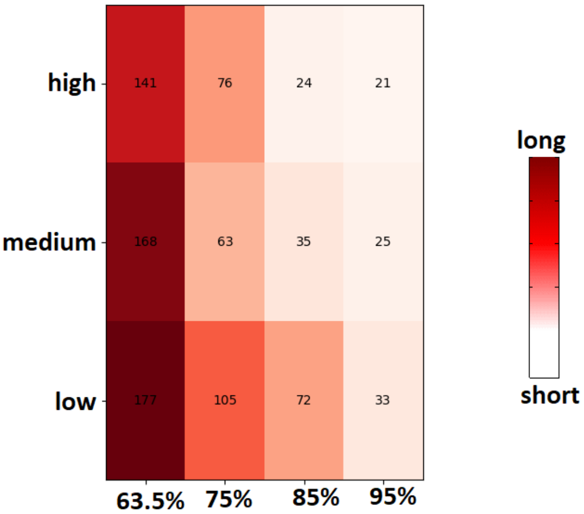


Figure 10: A 3D plot having QE quality levels along X-axis and MT quality along Y-axis. The color ranges from white depicting short durations to red depicting long durations.

A.6 Summary and Discussion

To summarize, our additional analyses on visualization suggest that while gradient is preferred for the relevant QE levels of more than 80% F1 score, there are no significant differences between the visualization schemes in terms of editing duration or

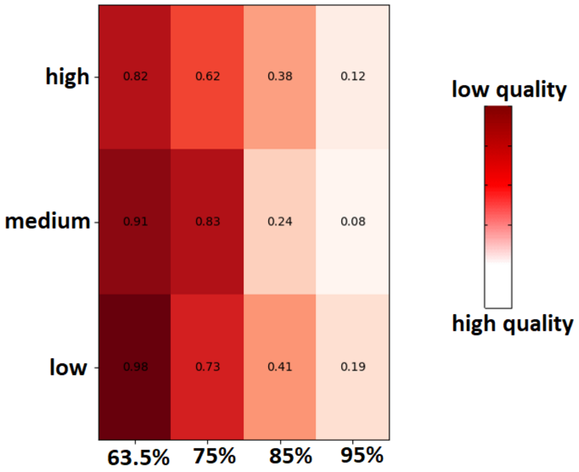


Figure 11: A 3D plot having QE quality levels along X-axis and MT quality along Y-axis. The color ranges from white depicting high quality translations to red depicting low quality translations.

quality of the final translation. Furthermore, the 3D plots investigating the effect of MT quality show that the effects shown in the main paper are a result mainly of the QE quality, whereas the MT quality plays only a minor role.