Validating Quality Estimation in a Computer-Aided Translation Workflow: Speed, Cost and Quality Trade-off

Fernando Alva-Manchego f.alva@sheffield.ac.uk
Department of Computer Science, University of Sheffield, UK

Lucia Specia l.specia@imperial.ac.uk
Department of Computing, Imperial College London, UK

Sara Szoc sara.szoc@crosslang.com
Tom Vanallemeersch tom.vanallemeersch@crosslang.com
Heidi Depraetere heidi.depraetere@crosslang.com
CrossLang, Kerkstraat 106, 9050 Gentbrugge, Belgium

Abstract

In modern computer-aided translation workflows, Machine Translation (MT) systems are used to produce a draft that is then checked and edited where needed by human translators. In this scenario, a Quality Estimation (QE) tool can be used to score MT outputs, and a threshold on the QE scores can be applied to decide whether an MT output can be used as-is or requires human post-edition. While this could reduce cost and turnaround times, it could harm translation quality, as QE models are not 100% accurate. In the framework of the APE-QUEST project (Automated Post-Editing and Quality Estimation), we set up a case-study on the trade-off between speed, cost and quality, investigating the benefits of QE models in a real-world scenario, where we rely on end-user acceptability as quality metric. Using data in the public administration domain for English-Dutch and English-French, we experimented with two use cases: assimilation and dissemination. Results shed some light on how QE scores can be explored to establish thresholds that suit each use case and target language, and demonstrate the potential benefits of adding QE to a translation workflow.

1 Introduction

Quality Estimation (QE) for Machine Translation (MT) predicts how good or reliable automatic translations are without access to gold-standard references (Specia et al., 2009; Fonseca et al., 2019; Specia et al., 2020). This is especially useful in real-world settings, such as within a translation company, where it can improve post-editing efficiency by filtering out segments that require more effort to correct than to translate from scratch (Specia, 2011; Martins et al., 2017), or select high-quality segments to be published as they are (Soricut and Echihabi, 2010). However, while the utility of MT is widely accepted nowadays, thus far no research has looked into validating the utility of QE in practice, in a realistic setting. To address this gap, in this paper we ask ourselves the following questions: 1) Can QE make the translation process more efficient (i.e. faster and cheaper)? 2) What is the impact of a QE-based filter on the quality of the final translations? and 3) How does varying the threshold for this filter affect these two competing goals (efficiency and quality)?

In the APE-QUEST project for Automated Post-Editing and Quality Estimation (Van den Bogaert et al., 2019; Depraetere et al., 2020),[1] we set up a proof-of-concept environment combining MT with QE. This Quality Gate was integrated within the workflow of the two companies in the consortium (CrossLang and Unbabel), specialized in computer-aided translation: predicted QE scores are used to decide whether an MT output can be used as-is (predicted as *acceptable* quality) or should be post-edited (predicted as *unacceptable* quality). It is expected that this Quality Gate speeds up the translation workflow and reduces costs since not all MT outputs would require human post-edition, but having humans read translations to make this decision is time-consuming. However, without a good understanding of the effects of QE-based filtering, there is a risk that the workflow becomes biased towards maximising throughput, i.e. towards selecting more low-quality translations as acceptable, and thus compromising the quality of the final translations. We propose a simple approach to studying the trade-off between speed, cost and quality, and show how important it is in allowing the Quality Gate to provide sufficiently-good MT while employing humans to only post-edit "difficult" sentences. We also show that this varies depending on the intended use of the translations.

Our experiments with the Quality Gate use state-of-the-art neural MT (NMT) and QE models with texts in the public administration domain, and translation use cases with different quality requirements (Section 3). To elaborate a realistic trade-off model, stakeholder input is important. As such, we collected human post-edits (along with post-editing time) and end-user acceptability judgements (binary scores) for two use cases (assimilation and dissemination) and two language pairs (English-Dutch and English-French) to evaluate the Quality Gate in different scenarios (Section 4). This data served to analyse how varying thresholds of QE scores affect post-editing time, overall cost and end-user acceptability, where we compare the Quality Gate against a human-only translation workflow (all MTs are checked and post-edited) and an MT-only translation workflow (all MTs are used as-is). Results (Section 5) show that QE scores can be used to establish thresholds that reduce cost and time, while maintaining similar quality levels as the human-only workflow, for all use cases and target languages. The gains are even greater when using oracle scores instead of predicted scores, signalling the benefits of improving this type of technology. This trade-off methodology for establishing QE thresholds proved helpful to demonstrate the benefits of incorporating QE in real-world computer-aided translation workflows (Section 6).

2 Related Work

Previous studies on the benefits of QE in translation workflows compared translators' productivity when post-editing selected MT outputs (based on QE scores) versus translating from scratch. Turchi et al. (2015) found that significant gains depend on the length of the source sentences and the quality of the MT output. Similarly, Parra Escartín et al. (2017) showed that translators spent less time post-editing sentences with "good" QE scores, i.e. scores that accurately predicted low PE effort. Different from these studies, we do not investigate impact on post-editor productivity, but rather whether it is possible to rely on QE scores to selectively bypass human post-edition and still achieve similar levels of translation quality. In addition, we experiment with state-of-the-art neural QE systems instead of feature-based ones.

The applicability of neural QE was investigated by Shterionov et al. (2019) when translating software UI strings from Microsoft products. The authors compared three systems in terms of business impact (using Microsoft's business metrics, such as throughput), model performance (using standard metrics, such as Pearson correlation), and cost (in terms of training and inference times). Different from theirs, our work relies on end-user translation acceptability as primary evaluation metric.

[1] https://ape-quest.eu/

Finally, some work attempted to determine thresholds for metrics' scores to identify ranges where post-editing productivity gains can be obtained (Parra Escartín and Arcedillo, 2015), or improvement in the quality of the raw MT output can be expected (Guerrero, 2020). However, they were based on post-hoc computations of TER (translation edit rate) or edit distance, respectively, instead of predicted QE scores as in our case. In addition, we experiment with thresholds of QE scores that benefit the overall translation workflow for different use cases and language pairs.

3 Quality Gate

We describe the technical components of the Quality Gate, the translation workflows that it compares to, and the translation use cases we considered.

3.1 Core Technologies

Machine Translation Module: The Quality Gate uses eTranslation[2] as backend NMT service. This service provides state-of-the-art NMT systems for more than 24 languages, and is targeted mainly at European public administrations and small and medium-sized enterprises.

Quality Estimation Module: The Quality Gate incorporates QE models built using TransQuest (Ranasinghe et al., 2020), the winning toolkit in the WMT20 Quality Estimation Shared Task for sentence-level QE (Specia et al., 2020). In these models, the original sentence and its translation are concatenated using the [SEP] token, and then passed through a pre-trained Transformer-based language model to obtain a joint representation via the [CLS] token. This serves as input to a softmax layer that predicts translation quality.

We trained language-specific models by fine-tuning Multilingual BERT (Devlin et al., 2019) with the dataset of Ive et al. (2020), which contains (source, MT output, human post-edition, target) tuples of sentences in the legal domain. We chose this data since it is the closest to our application domain, and contains instances in the language pairs of our interest: 11,249 for English-Dutch (EN-NL) and 9,989 for English-French (EN-FR). In order to obtain gold QE scores, we used `tercom` (Snover et al., 2006) to compute a TER value for each sentence. We trained our models using the same data splits as Ive et al. (2020), obtaining better results than the ones originally reported with ensembles of 5 models per language (Table 1).

	EN-NL		EN-FR	
Model	r	MAE	r	MAE
Ive et al. (2020)	0.38	0.14	0.58	0.14
Ours	**0.51**	**0.10**	**0.69**	**0.10**

Table 1: Performance of QE models in terms of Pearson's r correlation coefficient and Mean Absolute Error (MAE) in the test set of Ive et al. (2020).

Whilst the performance of the models is moderate according to Pearson, the error is relatively low (0.1 in a 0-1 range), and thus we believe the predictions can be useful to analyse the utility of current state-of-the-art QE in a real-word setting.

3.2 Workflows

In the **Quality Gate** workflow, given an automatic translation, the QE module provides a score that needs to be thresholded such that: (1) acceptable-quality MT will be left unchanged and

[2] `https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation`

passed directly to the end-user; and (2) unacceptable-quality MT will be sent to a Human Post-Editing (HPE) pipeline. This workflow will be compared to a **Traditional** workflow, where all MT outputs are manually checked and edited as needed, as well as to an **MT-Only** workflow, where translations from MT are not checked/post-edited but used as-is.

3.3 Use Cases

In our experiments, we used source texts snippets composed by texts sampled from a European public administration handling consumer complaints.[3] We devised two use cases that correspond to two distinct well-established uses of MT:

Assimilation: Translations are to be used for internal communication purposes (e.g. emails) or for general text understanding. Translation quality is expected to be *good enough* to understand the main message of the text.

Dissemination: Translations are to be published in any form (online or in print), so they need to be of *very high* quality, only requiring final quality checks (i.e. proofreading).

The input to the workflows are individual sentences, but they are post-edited and assessed in the context of the surrounding sentences.

4 Evaluation Protocol

Our trade-off model should help to answer the following questions:

- When compared to the Traditional workflow, does the Quality Gate workflow help to improve speed (i.e. time to get to final translation) and reduce cost (how many translations need HPE), while maintaining translation quality?
- When compared to the MT-Only workflow, does the Quality Gate workflow help to improve translation quality?

In addition, we investigate how the answers to these questions vary for: (1) different thresholds on the predicted quality of translations; (2) each of the two use cases (assimilation and dissemination); (3) different target languages; and (4) different quality of the QE scores (predicted vs oracle).

4.1 Measurable Criteria

The measurable criteria we compute for each use case and target language are:

Quality: Percentage of sentences considered of acceptable quality by independent human raters.

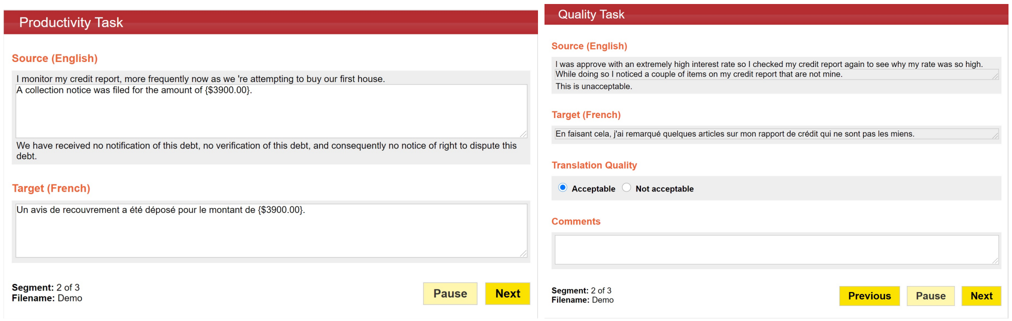
Cost: Percentage of sentences that require HPE, versus being fit for purpose.

Speed: Time required for HPE. The time to predict QE scores is negligible so it is not considered.

4.2 Datasets

For our evaluation, we used English text snippets from the public administration for each use case and target language. This type of text is challenging for the Quality Gate since it is out-

[3] For reasons of confidentiality, we cannot disclose the name of this administration. Therefore, the examples provided in this paper are taken from a publicly available dataset provided by the U.S. government: `https://catalog.data.gov/dataset/consumer-complaint-database`.



Productivity Task

Source (English)

I monitor my credit report, more frequently now as we 're attempting to buy our first house.
A collection notice was filed for the amount of {$3900.00}.
We have received no notification of this debt, no verification of this debt, and consequently no notice of right to dispute this debt.

Target (French)

Un avis de recouvrement a été déposé pour le montant de {$3900.00}.

Segment: 2 of 3
Filename: Demo

Quality Task

Source (English)

I was approve with an extremely high interest rate so I checked my credit report again to see why my rate was so high.
While doing so I noticed a couple of items on my credit report that are not mine.
This is unacceptable.

Target (French)

En faisant cela, j'ai remarqué quelques articles sur mon rapport de crédit qui ne sont pas les miens.

Translation Quality

● Acceptable ○ Not acceptable

Comments

Segment: 2 of 3
Filename: Demo

(a) Human post-edits

(b) Binary acceptability ratings

Figure 1: Screenshot of the MT Evaluation tool used to collect manual annotations.

of-domain compared to the texts used to train the NMT system (mainly general public administration) and the QE models (legal domain). The decision on the target languages – Dutch (NL) and French (FR) – is based on the availability of the human raters.

Assimilation Dataset: It consists of user complaints received by the public administration. This data is particularly interesting since it corresponds to conversational language. Sentence segmentation was applied before sending the texts to the MT system. After all preprocessing steps, we ended up with 25 complaints, totalling 966 English source sentences with an average length of 22.51 words per sentence.

Dissemination Dataset: Original texts were obtained from the website of the public administration. The data was segmented into sentences and then sent to the MT system. This resulted in 114 input sentences, with an average length of 18.32 words per sentence.

4.3 Human Annotations

We collected human annotations in two forms: **post-edits (HPE)** and **acceptability ratings**. While sentences that go through HPE are expected to have acceptable quality, we still collected human ratings for them to validate this assumption.

HPEs were obtained for all MT outputs available in each use case and target language. Post-editors were experienced professional translators in the domain of interest and for each use case. For each target language, three post-editors were hired, and each sentence was post-edited once.

Ratings were elicited for all MT outputs and their corresponding HPEs. Raters were professional translators that judged the quality of the sentences as Acceptable/Unacceptable for each use case. Raters were not informed of whether the sentences being judged were an MT output or HPE. For each target language and use case, two raters scored each translation (either MT or HPE) once.

HPEs and ratings were collected using the in-house MT Evaluation tool of one of the consortium's companies. Following recommended practice (Läubli et al., 2018; Toral et al., 2018), sentences were post-edited and rated within the document context of the source language, i.e. the preceding and the following sentences. For HPEs (Figure 1a) we also collected timestamps of when an editor started the editing job and of when the final job was delivered, at the sentence level. For collecting ratings (Figure 1b), the tool is flexible regarding the type of judgements that can be collected. In our case, we used binary ones for each use case.

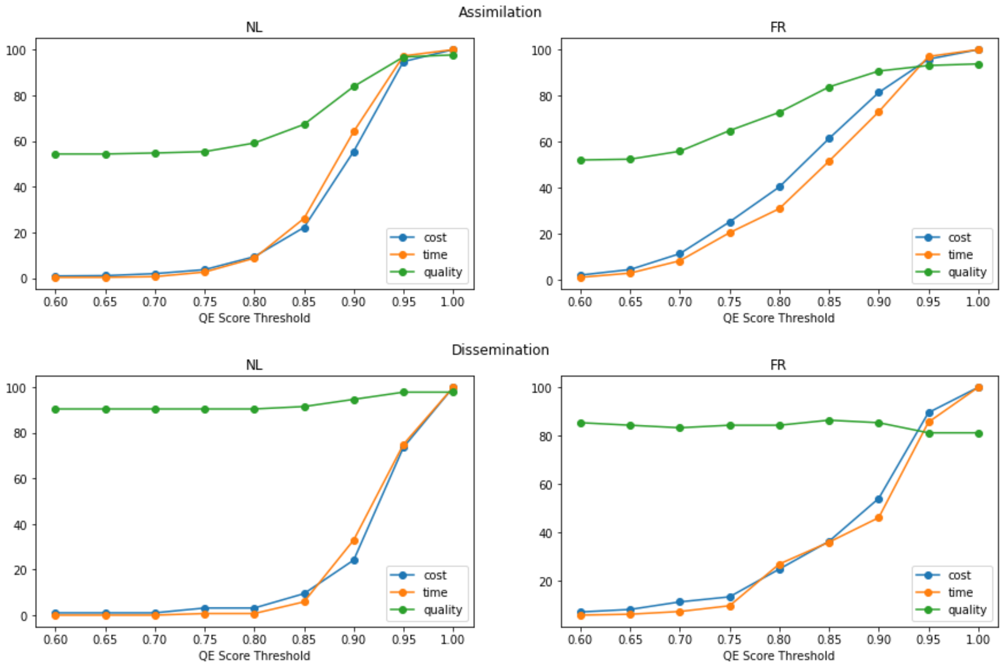


Figure 2: Variation of Cost, Time and Quality based on the QE score (**predicted**) threshold in the Quality Gate workflow for each use case and language (NL = English-Dutch; FR = English-French).

5 Experiments and Results

We evaluate the performance of the Quality Gate workflow depending on different values of QE score thresholds. We present the three evaluation metrics: Quality, Cost and Time. For better visualisation, we normalized Time as a percentage with respect to the Traditional workflow.

We set up thresholds from 0 to 1 in 0.05 increments, and computed the evaluation metrics under the assumption that sentences whose QE score was below the threshold required HPE. More specifically, for these sentences we took their post-editing time and quality judgement after HPE to calculate the metrics. For the rest (i.e. sentences not "requiring HPE") their time is 0 and their quality judgement is that of the MT output.

We first use the predicted QE scores to evaluate the current performance of the Quality Gate (Section 5.1). Then, we experiment with an oracle scenario where the QE scores are perfect, in order to measure the potential best-case-scenario performance of the Quality Gate workflow (Section 5.2).

5.1 Predicted QE Scores

Figure 2 shows how the three evaluation criteria vary depending on the threshold selected for the predicted QE score in the Quality Gate workflow. Table 2 details and compares the values to those from the Traditional (post-edit everything) and MT-only (do not post-edit anything) workflows.[4]

For all target languages and use cases, it is possible to set up a QE score threshold that allows the Quality Gate Workflow to obtain Quality with a value similar to the Traditional

[4] The QE < 1.0 threshold is excluded since no instance had a predicted QE score between 0.95 and 1.0.

Lang	Threshold	Assimilation			Dissemination		
		Cost	Time	Quality	Cost	Time	Quality
NL	Traditional	100.00	100.00	97.67	100.00	100.00	97.89
	QE < 0.95	94.77	97.24	96.80	73.68	74.98	97.89
	QE < 0.90	55.52	64.18	83.87	24.21	33.02	94.74
	QE < 0.85	22.24	26.17	67.30	9.47	5.87	91.58
	MT-Only	0.00	0.00	54.07	0.00	0.00	90.53
FR	Traditional	100.00	100.00	93.79	100.00	100.00	81.25
	QE < 0.95	95.86	97.02	93.10	89.58	85.64	81.25
	QE < 0.90	81.38	73.00	90.69	54.17	46.22	85.42
	QE < 0.85	61.38	51.54	83.79	36.46	36.11	86.40
	MT-Only	0.00	0.00	50.34	0.00	0.00	86.46

Table 2: Cost (% of sentences that need HPE), Time (% of HPE time with respect to Traditional) and Quality (% of acceptable translations) for varying thresholds of **predicted** QE scores in the Quality Gate compared to the Traditional and MT-only workflows.

Workflow, with reductions in Cost and Time. This QE score threshold is 0.95 for most cases. The gains in Time and Cost vary depending on the target language and use case.

For both use cases, the Quality Gate workflow achieves better results in NL than the MT-only one. The gains in Time and Cost vary according to the threshold selected. In the case of FR, the gains are evident for the Assimilation use case. However, MT-only obtains a better Quality score in the Dissemination use case, even superior to the Traditional workflow. This is because, for a few sentences, whilst one rater judged their MT outputs as acceptable, the other rater judged their HPE versions as unacceptable. We hypothesize that this is caused by disagreements in the human judgements rather than HPE being worse than MT. More analysis with multiple human ratings per translation would be needed to test this hypothesis.

5.2 Oracle QE Scores

Since we have HPEs for all MT outputs, we use them to compute oracle QE scores, that is, their "real" QE scores. This models an ideal scenario where the Quality Gate perfectly determines the QE score of an MT output. This could be seen as an upper bound of the potential benefits of the Quality Gate workflow. Figure 3 and Table 3 show our results in this setting.

In this ideal scenario, the gains are higher for all target languages in both use cases. This evidences the potential of the Quality Gate for reducing Cost and Time while preserving high Quality. We would expect the Quality Gate workflow to be able to move towards this ideal scenario as it is put in place and post-edits in the actual domain of interest are collected to better train the QE models.

6 Conclusions

In this paper, we provided evidence of the benefits of introducing QE into the computer-aided translation workflow of a company. In the framework of the APE-QUEST project, we implemented a Quality Gate that decides, based on predicted QE scores, whether MT outputs can be used as-is (acceptable quality) or if they require post-edition (unacceptable quality). We performed a trade-off study to establish thresholds on the QE scores that allow reducing time and cost, while keeping translation quality more or less stable. We collected human post-edits and acceptability ratings from real use case scenarios and real end-users, and demonstrated that the

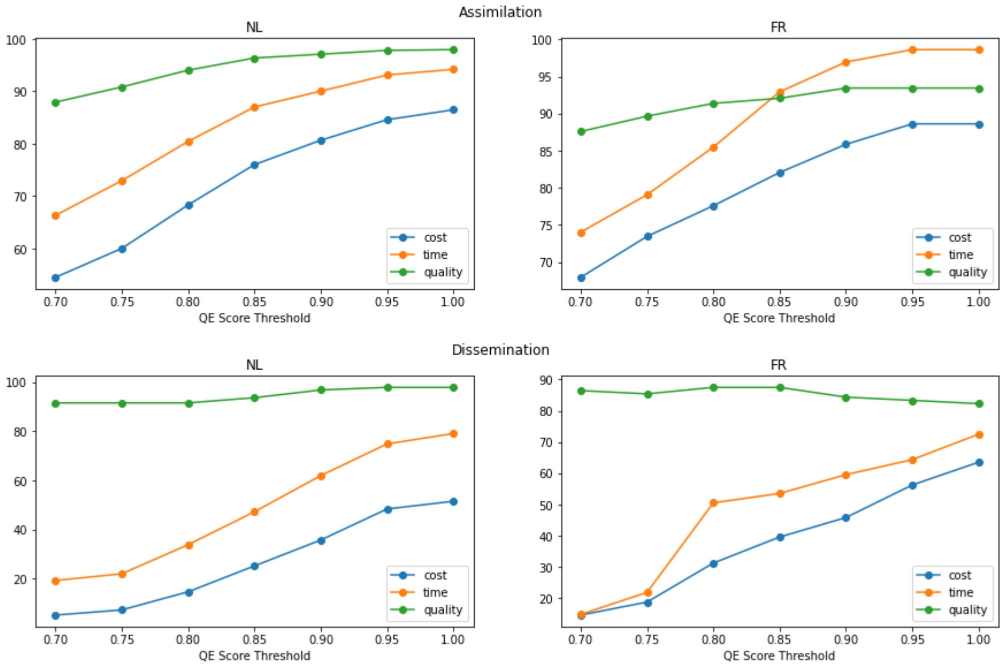


Figure 3: Variation of Cost, Time and Quality based on the QE score (**oracle**) threshold in the Quality Gate workflow for each use case and language (NL = English-Dutch; FR = English-French).

Lang	Threshold	Assimilation			Dissemination		
		Cost	Time	Quality	Cost	Time	Quality
NL	Traditional	100.00	100.00	97.67	100.00	100.00	97.89
	QE < 1.00	86.48	94.20	97.97	51.58	79.12	97.89
	QE < 0.95	84.59	93.12	97.82	48.42	74.89	97.89
	QE < 0.90	80.67	90.04	97.09	35.79	62.00	96.84
	MT-Only	0.00	0.00	54.07	0.00	0.00	90.53
FR	Traditional	100.00	100.00	93.79	100.00	100.00	81.25
	QE < 1.00	88.62	98.63	93.45	63.54	72.51	82.29
	QE < 0.95	88.62	98.63	93.45	56.25	64.37	83.33
	QE < 0.90	85.86	96.95	93.45	45.83	59.55	84.38
	MT-Only	0.00	0.00	50.34	0.00	0.00	86.46

Table 3: Cost (% of sentences that need HPE), Time (% of HPE time with respect to Traditional) and Quality (% of acceptable translations) for varying thresholds of **oracle** QE scores in the Quality Gate compared to the Traditional and MT-only workflows.

Quality Gate can obtain similar levels of quality to the current human-only workflow, for all use cases and target languages explored. In addition, when the predicted QE scores are changed to oracle ones, the gains are higher, illustrating the potential benefits of improving the predictive abilities of the QE models.

Acknowledgements

APE-QUEST was funded by the EC's CEF Telecom programme (project 2017-EU-IA-0151). The project ran from 2018 to 2020.

References

Depraetere, H., Van den Bogaert, J., Szoc, S., and Vanallemeersch, T. (2020). APE-QUEST: an MT quality gate. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 473–474, Lisboa, Portugal. European Association for Machine Translation.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Fonseca, E., Yankovskaya, L., Martins, A. F. T., Fishel, M., and Federmann, C. (2019). Findings of the WMT 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

Guerrero, L. (2020). In search of an acceptability/unacceptability threshold in machine translation post-editing automated metrics. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track)*, pages 32–47, Virtual. Association for Machine Translation in the Americas.

Ive, J., Specia, L., Szoc, S., Vanallemeersch, T., Van den Bogaert, J., Farah, E., Maroti, C., Ventura, A., and Khalilov, M. (2020). A post-editing dataset in the legal domain: Do we underestimate neural machine translation quality? In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3692–3697, Marseille, France. European Language Resources Association.

Läubli, S., Sennrich, R., and Volk, M. (2018). Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Martins, A. F. T., Junczys-Dowmunt, M., Kepler, F. N., Astudillo, R., Hokamp, C., and Grundkiewicz, R. (2017). Pushing the limits of translation quality estimation. *Transactions of the Association for Computational Linguistics*, 5:205–218.

Parra Escartín, C. and Arcedillo, M. (2015). Living on the edge: productivity gain thresholds in machine translation evaluation metrics. In *Proceedings of 4th Workshop on Post-Editing Technology and Practice (WPTP4)*, pages 46–56, Miami.

Parra Escartín, C., Béchara, H., and Orasan, C. (2017). Questing for quality estimation a user study. *The Prague Bulletin of Mathematical Linguistics*, 108:343–354.

Ranasinghe, T., Orasan, C., and Mitkov, R. (2020). TransQuest: Translation quality estimation with cross-lingual transformers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5070–5081, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Shterionov, D., Carmo, F. D., Moorkens, J., Paquin, E., Schmidtke, D., Groves, D., and Way, A. (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.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of 7th Biennial Conference of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge, Massachusetts, USA.

Soricut, R. and Echihabi, A. (2010). TrustRank: Inducing trust in automatic translations via ranking. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 612–621, Uppsala, Sweden. Association for Computational Linguistics.

Specia, L. (2011). Exploiting objective annotations for minimising translation post-editing effort. In *Proceedings of the 15th Annual conference of the European Association for Machine Translation*, pages 73–80, Leuven, Belgium. European Association for Machine Translation.

Specia, L., Blain, F., Fomicheva, M., Fonseca, E., Chaudhary, V., Guzmán, F., and Martins, A. F. T. (2020). Findings of the WMT 2020 shared task on quality estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 743–764, Online. Association for Computational Linguistics.

Specia, L., Turchi, M., Cancedda, N., Cristianini, N., and Dymetman, M. (2009). Estimating the sentence-level quality of machine translation systems. In *Proceedings of the 13th Annual conference of the European Association for Machine Translation*, Barcelona, Spain. European Association for Machine Translation.

Toral, A., Castilho, S., Hu, K., and Way, A. (2018). Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

Turchi, M., Negri, M., and Federico, M. (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 (Volume 2: Short Papers)*, pages 530–535, Beijing, China. Association for Computational Linguistics.

Van den Bogaert, J., Depraetere, H., Szoc, S., Vanallemeersch, T., Van Winckel, K., Everaert, F., Specia, L., Ive, J., Khalilov, M., Maroti, C., Farah, E., and Ventura, A. (2019). APE-QUEST. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, pages 110–111, Dublin, Ireland. European Association for Machine Translation.