Learning from Past Mistakes: Quality Estimation from Monolingual Corpora and Machine Translation Learning Stages

Thierry Etchegoyhen[1] tetchegoyhen@vicomtech.org
David Ponce[1,2] adponce@vicomtech.org

[1] Vicomtech Foundation, Basque Research and Technology Alliance (BRTA)
[2] University of the Basque Country - UPV/EHU

Abstract

Quality Estimation (QE) of Machine Translation output suffers from the lack of annotated data to train supervised models across domains and language pairs. In this work, we describe a method to generate synthetic QE data based on Neural Machine Translation (NMT) models at different learning stages. Our approach consists in training QE models on the errors produced by different NMT model checkpoints, obtained during the course of model training, under the assumption that gradual learning will induce errors that more closely resemble those produced by NMT models in adverse conditions. We test this approach on English-German and Romanian-English WMT QE test sets, demonstrating that pairing translations from earlier checkpoints with translations of converged models outperforms the use of reference human translations and can achieve competitive results against human-labelled data. We also show that combining post-edited data with our synthetic data yields to significant improvements across the board. Our approach thus opens new possibilities for an efficient use of monolingual corpora to generate quality synthetic QE data, thereby mitigating the data bottleneck.

1 Introduction

Significant improvements have been achieved in Machine Translation (MT) in recent years, in particular with the advent of Neural Machine Translation (NMT) (Sutskever et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017). However, the quality of automated translations can vary significantly depending on training data volumes, domain of application, language pairs or the complexity of specific source segments. Machine translation errors can significantly increase the risks and costs of using MT and the automatic estimation of MT quality becomes increasingly necessary to pinpoint or discard erroneous automatic translations.

Traditionally, the quality of MT output has been assessed against human references, via automated metrics such as BLEU (Papineni et al., 2002) or TER (Snover et al., 2006). However, such references are not always available and are costly to produce, which has led to the development of Quality Estimation (QE) approaches based on the sole properties of the source and machine-translated sentences (Blatz et al., 2004; Specia et al., 2010). Most approaches to QE are based on supervised learning, traditionally via feature engineering (Specia et al., 2013), and, in recent years via neural models (Kim and Lee, 2016; Kim et al., 2017; Fonseca et al., 2019;

Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track, pages 84–98
September 4–8, 2023, Macau SAR, China.

Specia et al., 2021). Although they provide the most accurate estimates to date, supervised methods depend on human annotations or post-edited translations to perform the task. The cost of producing quality QE training datasets hinders the development of QE models for the large number of possible domains and language pairs.

Two main alternatives address the lack of training QE data. On the one hand, unsupervised and self-supervised approaches (Moreau and Vogel, 2012; Popović, 2012; Etchegoyhen et al., 2018; Fomicheva et al., 2020; Zheng et al., 2021) discard the need for QE training data altogether, but typically fail to consistently meet the accuracy of supervised alternatives or may require access to additional information such as internal states of the MT model. On the other hand, methods that exploit synthetic training data have also been proposed in recent years, leveraging parallel dataset references. Under this approach, parallel training data can be exploited, for instance, by taking a target reference translation as the approximated post-edited version of a machine-translated source segment and generating artificial QE labels (Lee, 2020). The two may differ significantly however, thereby introducing noise in the QE training data. Alternatively, synthetic data can be generated by devising QE error generation pipelines from the parallel data (Baek et al., 2020; Tuan et al., 2021), although this requires approximating errors that may not correspond to actual MT ones.

In this work, we describe and evaluate a novel approach to synthetic QE data generation by exploiting the actual errors committed by NMT models at different learning stages. The hypothesis underlying this approach is that this type of errors might resemble more closely the errors produced by MT systems in scenarios where they typically fail, such as language pairs for which parallel training data are insufficient, or domains that deviate from those represented in the training sets. To test this hypothesis, we train NMT models on generic parallel data and select model checkpoints of varying quality to contrast their translations with either human reference translations or translations from the best converged NMT models. The generated synthetic data are then used to train neural QE estimators, either in isolation or in combination with human-generated data. We demonstrate the potential of this novel approach on WMT 2021 datasets in English-German and Romanian-English. We notably show that it outperforms the use of human reference translations, directly or via self-supervised learning, is competitive with the use of human post-edited data, and can complement the latter to achieve further gains in QE accuracy. Additionally, contrasting checkpoint translations with those of converged NMT models allows for a direct exploitation of monolingual data, thus opening new possibilities for the effective generation of synthetic QE data across languages and domains.

2 Related Work

Machine translation quality estimation has been standardly tackled via supervised approaches, with annotated or post-edited machine-translated segments being used to train machine learning classifiers (Blatz et al., 2004; Quirk, 2004) or regressors (Specia et al., 2009). Several approaches have been explored using different feature sets or underlying learning models such as Support Vector Machines or Gaussian Processes (Callison-Burch et al., 2012; Bojar et al., 2014; Specia et al., 2013; Felice and Specia, 2012; Forcada et al., 2017).

In recent years, approaches based on artificial neural networks have been successfully applied to the task as well, either as additional features (Shah et al., 2015, 2016) or as end-to-end quality estimation systems (Kim and Lee, 2016; Martins et al., 2017; Ive et al., 2018; Fan et al., 2019). The Predictor-Estimator framework proposed by Kim et al. (2017) can be considered the current standard, since it outperformed alternatives in recent WMT QE tasks (Bojar et al., 2017; Specia et al., 2018) and now serves as baseline in the latest editions of the task (Specia et al., 2020, 2021; Zerva et al., 2022). In this framework, a contextual word Predictor component acts as a feature extractor and an Estimator exploits the extracted features

to predict QE labels. A neural word prediction model can be trained on the parallel data (Kim et al., 2017; Zhou et al., 2019), though in recent years, pretrained large language models, such as BERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020), have also been successfully employed for this task (Kim et al., 2019; Kepler et al., 2019a; Specia et al., 2020).

As previously noted, supervised approaches depend on the availability of annotated datasets, typically HTER scores obtained from post-edited machine translation output, quality values on a predefined quality scale, or OK/BAD annotations at the word-level. Creating quality annotated datasets is a costly process, hindering the development of quality supervised QE models. To date, most publicly available QE datasets are those prepared for the WMT shared tasks, which are only available for a limited number of language pairs and domains.

To address this data bottleneck, alternatives to supervised modelling have been explored for the QE task. Thus, Moreau and Vogel (2012) tackled weakly supervised and single-feature unsupervised methods, as a means to minimise the dependency on annotated data. Popović (2012) describes an unsupervised method based on combining IBM1 models with language models over morphemes and part-of-speech tags, with a dependence on external tagging tools. In Etchegoyhen et al. (2018), unsupervised quality estimation is performed via lexical translation overlaps and n-gram language model scores, outperforming some feature-based supervised models but falling short against more sophisticated neural QE models. An unsupervised glass-box approach, based on the confidence of NMT models, was proposed by Fomicheva et al. (2020), achieving promising results, though it requires access to the NMT models that generate the evaluated translations. Recently, Zheng et al. (2021) proposed a self-supervised approach based on target token masking in parallel data, outperforming other methods based on unsupervised modelling or synthetic data generation.

Another approach to the lack of human-annotated training QE data is to leverage existing parallel corpora, similarly to what was suggested for automated post-editing (Negri et al., 2018). Thus, Lee (2020) and Tuan et al. (2021) explored the use of target reference translations as post-edited versions of machine-translated source sentences, showing that it can provide a basis for supervised QE models. In this type of approach, however, target references may differ significantly from MT output and therefore introduce noisy training tuples in the QE data. Alternatively, synthetic data can be generated by devising QE error generation pipelines from the parallel data (Baek et al., 2020; Tuan et al., 2021), although this requires approximating errors that may not correspond to the actual ones produced by MT models.

The study most related to ours is that of Ding et al. (2021), who evaluated their Levenshtein Transformer approach to word-level quality estimation using synthetic data, part of which was generated by using the output from a weaker MT model and contrasting it with the output of another MT model of higher quality, taken as reference translator. Although the idea of contrasting weaker and stronger MT models is similar, this differs from our approach in important respects: their synthetic data results are only established for their proposed QE framework based on the Levenshtein Transformer, only for word-level QE, and, most importantly, they use the output of unrelated converged translation models, instead of the related learning stages of the same model which we explore in this work.

3 Methodology

As previously indicated, our approach is based on the assumption that NMT models at different training stages might produce errors that resemble those committed by fully trained MT systems, in scenarios where they fail to properly translate such as domain shifts or insufficient training data. The methodology can be summarised as follows:

1. Train an NMT model on parallel data from language L_1 to language L_2.

		English-German		Romanian-English	
Corpus	Type	Sentences	Tokens	Sentences	Tokens
WMT21-QE train	Post-edited	7,000	114,980	7,000	120,247
WMT21-QE dev	Post-edited	1,000	16,519	1,000	17,279
WMT21-QE test	Post-edited	1,000	16,371	1,000	17,359
WikiMatrix	Comparable	696,880	15,386,735	102,106	2,120,383
WMT21-MT	Parallel	22,782,867	490,297,937	3,080,304	72,004,236
WikiDump	Monolingual	1,923,782	38,456,268	1,392,034	25,320,444

Table 1: Corpora statistics (number of tokens computed over source sentences)

2. Select model checkpoints at different stages of training. We used three different checkpoints in our experiments, though more could be defined as needed:
   - *b50*: the checkpoint whose development set BLEU score is the closest to 50% of the score of the converged NMT model.
   - *b75*: the checkpoint whose development set BLEU score is the closest to 75% of the score of the converged NMT model.
   - *best*: the checkpoint corresponding to the converged NMT model.
3. Translate a source corpus in L_1 using the selected checkpoints.
4. Extract tuples <*src*, *mt*, *ref*>, where *src* is the source sentence, *mt* is the translation generated by a given checkpoint, and *ref* is either a human reference translation (*hrt*), or the output of the *best* model when *b50* and/or *b75* are used to generate the translations.
5. Train the estimator of a Predictor-Estimator QE model (Kim et al., 2017) on the generated tuples.

Under this approach, synthetic QE data can be generated from monolingual or parallel source data, in any domain or language pair for which an NMT model was trained. Several aspects need to be examined to determine an optimal setup for this method, mainly the impact of: (i) using *best* model translations as opposed to an existing reference in parallel data; (ii) using different volumes of synthetic data; (iii) creating synthetic data from different domains; (iv) combining synthetic data from different model checkpoints; and (v) combing synthetic data and human post-edited translations. In the next sections, we describe the experimental protocols to tackle these aspects and evaluate the potential of our approach.

4 Experimental Setup

Our experiments centred on two language pairs, English-German (EN-DE) and Romanian-English (RO-EN), and the datasets of the WMT 2021 shared QE task (Specia et al., 2021) . The selected datasets and models for our experiments are described in turn below.

4.1 Data

We selected the WMT 2021 datasets from the quality estimation task[1] (hereafter, WMT21-QE) as development and test data for our QE models, on the translation pairs English-German and Romanian-English. For the experiments described in Section 7, we also merged our synthetic

[1]https://www.statmt.org/wmt21/quality-estimation-task.html

data with the human post-edited train dataset from the task. Our choice of datasets was mainly motivated by the balanced datasets introduced for the 2020 shared task, following work by Sun et al. (2020). English-German was selected as representative of a language pair with significant volumes of parallel data to train NMT models; Romanian-English features lower volumes of such data and was also selected to represent translation from a different source language.

To train the NMT models from which we extract the different checkpoints, we used the parallel training and development data provided in the 2021 QE shared task (WMT21-MT) for the two selected language pairs. To generate synthetic QE data, we used the following datasets:[2]

- *WikiMatrix*: since the domain for the selected language pairs in the WMT 2021 QE shared task was Wikipedia, we used the WikiMatrix dataset (Schwenk et al., 2021), selecting the top pairs with a LASER score (Artetxe and Schwenk, 2019) above a 1.06 threshold, following Tuan et al. (2021). With this dataset, either the aligned comparable target sentences or the *best* model translations were used as references, depending on the method at hand.

- *WMT21-MT*: to assess the impact of synthetic QE data generated from a different domain, we used a subset of the WMT21-MT data, selecting 2M sentence pairs via uniform sampling. As with the previous dataset, we evaluated the use of either the parallel translation or the *best* model translation as reference.

- WikiDump: this dataset is strictly monolingual and was only used for the experiments reported in Section 7, as there are no reference translations to perform the full set of experiments. We used Wikipedia dumps in both English and Romanian[3], as an additional monolingual test case, translating the source with model checkpoints and using *best* model translations as references.

The data were tokenised and truecased, using scripts from the Moses tookit (Koehn et al., 2007). Truecasing models were trained on the WMT21-MT datasets, and only applied on the QE datasets; for the NMT models, we used inline casing (Berard et al., 2019; Etchegoyhen and Gete, 2020), where all words are lowercased and casing information, if any, is prepended as symbols. The output of the NMT models was then recased and subsequently truecased for QE training and inference. For NMT training, subwords were generated via Byte Pair Encoding (Sennrich et al., 2016), training BPE models on WMT21-MT data with 32K operations.

4.2 Models

To compare different approaches to QE without human-labeled data, we selected the models described below.

Baseline. As a QE baseline, we followed the setup in the WMT 2021 QE shared task and trained Predictor-Estimator models on WMT21-QE data with OpenKiwi v2.1.0 (Kepler et al., 2019b), using XLM-R (Conneau et al., 2020) as Predictor. The baselines were trained separately for each language pair on the selected data.

Checkpoint-based QE. For our approach, we used MarianNMT (Junczys-Dowmunt et al., 2018) to train Transformer-base NMT models (Vaswani et al., 2017), with 6 encoder layers, 6 decoder layers, and 8 attention heads. We saved checkpoints every 5000 steps and translated the selected source datasets with a beam search of 6. The converged models obtained BLEU scores of 39.4 and 41.0 on the EN-DE and RO-EN WMT21-MT devsets, respectively. The QE models trained on data translated with checkpoint NMT models followed the setup of the baseline.

[2]In all cases, we filtered sentences containing more than 100 tokens, empty lines and duplicates.
[3]https://dumps.wikimedia.org/. Accessed 2022/12.

NMT QE. In this approach, the output of the NMT models is contrasted with the target references in the comparable or parallel dataset. This is similar to the approach denoted as NMT in Tuan et al. (2021), which obtained better results overall than their synthetic error generation method, with further gains obtained when both were used in combination. QE models based on this approach also followed the same setup as the baseline. Note that this approach is also similar, in a sense, to the use of unrelated contrastive NMT models as in Ding et al. (2021): in our case, the weaker model would be the NMT model, and the stronger model would be represented by the human translator, who can be assumed to provide the highest possible translation quality. Differences may arise from contrasting the output of the weaker model with human translations instead of the output of a strong MT model, although the results in Ding et al. (2021) indicate only minor differences in this respect.

Self-supervised QE. We selected the approach of Zheng et al. (2021), which is based on retrieving masked target words considering the source and target context, as it outperformed alternatives such as synthetic error generation (Tuan et al., 2021) in their experiments. We trained self-supervised models on the selected datasets where reference translations were available, i.e. WMT21-MT and WikiMatrix, using the publicly available code with default parameters.[4]

All models were trained until convergence. To evaluate their performance, we used the setup of the WMT 2021 QE shared task for Task 2, which measures word and sentence level post-editing effort. At the word level, targets are word level OK/BAD tags to signify the correctness of words and gaps in the source and translated sentences. The primary metric in this case is the Matthews correlation coefficient (MCC) (Matthews, 1975). For comparison purposes, we only report MCC results over the translated tags, as these are the only word-level predictions generated by the self-supervised approach. At the sentence-level, the targets are the HTER scores contrasting the machine translated output against the human reference, and the primary metric is the Pearson *r* correlation score. We used the evaluation scripts provided for the shared task to compute the results.

4.3 Checkpoint-based Variants

Under our approach, synthetic data may be generated via different configurations, in terms of data combination, type of data and volumes of data used to train the QE models. We describe our experimental setup for each one of these aspects below.

Checkpoint combination. Since our method allows for any model checkpoint to be used for synthetic data generation, different combinations may be exploited. We trained QE models that merged datasets generated by the following combinations of the selected checkpoints described in Section 3, using as reference either the parallel or comparable human reference (*hrt*) or the translation from the converged model (*best*): *<b50, hrt>*, *<b50, best>*, *<b75, hrt>*, *<b75, best>*, *<b50+b75, hrt>*, *<b50+b75, best>*, *<b50+b75+best, hrt>*.[5] We also indicate the results obtained with *<best, hrt>*, which corresponds to the NMT QE model described above.

Data type. Synthetic data may be generated from source data close to, or differing from, the domain of interest in a given QE task. As domain proximity may impact the usefulness of the synthetic data, we applied our method to the WikiMatrix data, closer in nature to the Wikipedia data used in the QE task, and the parallel data from the WMT 2021 MT task, which merges data from different domains.

Data size. The amount of potential synthetic data for a given language pair, under our approach, is only limited by the availability of monolingual source data, which may be available

[4] https://github.com/THUNLP-MT/SelfSupervisedQE.

[5] The notation + indicates concatenation of the data translated with each indicated checkpoint model.

in large quantities. However, synthetic data might differ significantly from human-labelled data and may feature noisy data. Therefore, adding large quantities of synthetic data might be detrimental to the quality of QE models. To determine the impact of synthetic data volumes, we trained different QE models based on: 7K synthetic data (*small* dataset), matching the amount of human post-edited data used in the WMT 2021 QE task; 70K (*medium*) to increase the initial size by an order of magnitude; and, finally, the maximum amount of data (*large*) available in the WikiMatrix dataset, using the same amount for the WMT 2021 MT training data.

5 Checkpoint-based QE Results

		English-German			Romanian-English		
Model	Dataset	Small	Medium	Large	Small	Medium	Large
<best, hrt>	WMT21-MT	0.213	0.277	0.207	0.576	0.598	0.609
<b50, hrt>	WMT21-MT	0.304	0.394	0.366	0.622	0.660	0.608
<b75, hrt>	WMT21-MT	0.355	0.397	0.385	0.557	0.611	0.604
<b50+b75+best, hrt>	WMT21-MT	0.369	0.435	0.427	0.541	0.628	0.610
<b50, best>	WMT21-MT	0.383	0.425	0.419	0.724	0.746	0.765
<b75, best>	WMT21-MT	0.366	**0.464**	0.424	0.731	0.787	0.786
<b50+b75, best>	WMT21-MT	0.431	0.421	0.462	0.767	0.783	**0.798**
<best, hrt>	WikiMatrix	0.259	0.306	0.089	0.774	0.803	0.791
<b50, hrt>	WikiMatrix	0.341	0.343	0.158	0.752	0.736	0.745
<b75, hrt>	WikiMatrix	0.352	0.370	0.159	0.747	0.777	0.784
<b50+b75+best, hrt>	WikiMatrix	0.370	0.400	0.143	0.786	0.781	0.788
<b50, best>	WikiMatrix	0.403	0.374	0.345	0.781	0.774	0.776
<b75, best>	WikiMatrix	0.411	0.436	0.390	0.801	**0.829**	0.828
<b50+b75, best>	WikiMatrix	**0.448**	0.425	0.413	0.808	0.814	0.809

Table 2: Pearson correlation results on WMT21-QE test sets for Task2 Sentence-level HTER prediction, using *small*, *medium* and *large* synthetic datasets. Best results across dataset splits are indicated in bold; best results per dataset split are underlined.

We first evaluated the impact of using different combinations of synthetic data, and either the human reference translation or the *best* model translation as references. The results at the sentence-level, for the two domains where comparable or parallel references were available, are shown in Table 2. The most notable result is that contrasting checkpoint translations with the output of the converged model markedly outperformed the alternatives in both language pairs and across datasets. In particular, these models obtained significantly better results than the NMT QE approach based on <best, hrt> coupling. These results at the sentence level thus indicate that directly exploiting monolingual source data via checkpoint and converged model translations can provide a better basis for QE than unrelated parallel or comparable references. Among models that used human reference translations, the checkpoint-based variants performed better than <best, hrt> in all cases and datasets for EN-DE. For RO-EN, the results featured less differences in scores, although <best, hrt> performed slightly better overall.

In terms of data size, in three out of four cases, the checkpoint-based models that relied on *best* translations as references obtained the best results with small (7K) or medium samples (70K). The larger datasets led to the best performance only in RO-EN on WMT21-MT and was competitive overall, but smaller data volumes seemed sufficient for the most part to reach the highest Pearson correlations on the test sets.

		English-German		Romanian-English	
Model	Dataset	Pearson	MCC	Pearson	MCC
Baseline	WMT21-QE	**0.541**	**0.37**4	**0.829**	**0.575**
NMT QE	WMT21-MT	0.277	0.213	0.609	0.180
Self-supervised QE	WMT21-MT	0.238	0.253	0.565	0.386
<b50, best>	WMT21-MT	0.425	0.320	0.765	0.489
<b75, best>	WMT21-MT	0.464	0.336	0.787	0.450
<b50+b75, best>	WMT21-MT	0.462	0.335	0.798	0.423
NMT QE	WikiMatrix	0.306	0.272	0.803	0.445
Self-supervised QE	WikiMatrix	0.286	0.283	0.731	0.500
<b50, best>	WikiMatrix	0.403	0.314	0.781	0.469
<b75, best>	WikiMatrix	0.436	0.343	**0.829**	0.543
<b50+b75, best>	WikiMatrix	0.448	0.325	0.814	0.520

Table 3: Comparative results on the WMT 2021 Task2 test sets for the Pearson (sentence-level) and MCC (word-level on MT tags) primary metrics. Baselines trained on human post-edited (PE) data. Best results overall are indicated in bold; best results among methods that do not rely on PE data are underlined.

Among the top-performing methods, *<b75, best>* and *<b50+b75, best>* outperformed *<b50, best>* overall, and the best results were distributed among the two depending on the dataset and language pair: *<b75, best>* was optimal in EN-DE with WMT21-MT and RO-EN with WikiMatrix using medium sized datasets, whereas *<b50+b75, best>* was optimal on WikiMatrix with the small dataset for EN-DE and on WMT21-MT with the large dataset for RO-EN. Either method might thus be a reasonable choice to generate synthetic QE data, and future experiments would be needed to further distinguish between the two options.

Finally, although the QE test sets were based on data from Wikipedia for these language pairs, using synthetic data generated from a different domain like WMT21-MT did not seem significantly detrimental, as it even led to better scores than WikiMatrix-based synthetic data in EN-DE on the medium and large datasets. The best scores in most cases for the two top-performing variants were nonetheless still achieved with synthetic data generated from the WikiMatrix datasets, which is closer in nature to the QE test data.

6 Comparative Results

In this Section, we compare our results with the selected alternative approaches, namely: baselines trained on the 7K post-edited data of the WMT-QE-Train datasets; Self-supervised models trained on the available parallel and comparable corpora, as these models require aligned data; the NMT QE model based on contrasting the NMT translation and the parallel or comparable target human reference (*<best, hrt>*); and the best variants of our approach as determined in the previous Section, all based on checkpoint translations of the source data and translations of the converged NMT models as references. In Table 3, we present the comparative results at the sentence and word levels, according to the primary metric in each case.[6]

The baselines obtained the best results overall, at both the sentence and word levels, which is not unexpected as they were trained on the post-edited data from the task. However, our best

[6]For each method, we indicate the best score obtained at the sentence and word level independently, irrespective of QE training data partition size.

		English-German		Romanian-English	
Model	Dataset	Pearson	MCC	Pearson	MCC
Baseline	WMT21-QE	0.541	0.374	0.829	0.575
WMT21-QE 7K + Synthetic 7K	WikiDump	0.583	0.407	0.815	0.571
WMT21-QE 7K + Synthetic 70K	WikiDump	0.567	0.399	0.836	0.551
WMT21-QE 70K + Synthetic 70K	WikiDump	**0.594**	**0.429**	0.827	0.579
WMT21-QE 7K + Synthetic 7K	WikiMatrix	0.563	0.398	0.842	0.570
WMT21-QE 7K + Synthetic 70K	WikiMatrix	0.552	0.390	0.838	0.555
WMT21-QE 70K + Synthetic 70K	WikiMatrix	0.588	0.414	**0.844**	**0.578**
WMT21-QE 7K + Synthetic 7K	WMT21-MT	0.558	0.387	0.838	0.556
WMT21-QE 7K + Synthetic 70K	WMT21-MT	0.573	0.403	0.825	0.522
WMT21-QE 70K + Synthetic 70K	WMT21-MT	0.591	0.409	0.831	0.560

Table 4: Sentence and word level results on the WMT 2021 Task2 test sets for QE models trained on combined human post-edited data and synthetic data generated from different datasets. Best results overall are indicated in bold; best results per dataset are underlined.

variant matched the best sentence-level score in RO-EN and obtained competitive results in all other cases at both sentence and word level. Considering that the training data were randomly sampled monolingual source sentences from datasets differing from the shared task post-edited training data, these results confirm the potential of the checkpoint-based approach to create synthetic QE data that can match or approximate the usefulness of human post-edited data.

Across metrics, both the NMT QE and the Self-supervised QE approaches were markedly outperformed by all variants of our approach, except for RO-EN on the WikiMatrix dataset, where NMT QE obtained better results than the least accurate <b50, best> variant at the sentence level. Self-supervised QE performed better than NMT QE on word-level accuracy in all cases, with opposite results at the sentence level. Note that the use of *unrelated* contrastive translations, at least in the form of NMT QE with high quality human translations contrasted with translations from a baseline NMT model, was outperformed by the use of translations from related NMT stages overall.

7 Natural and Synthetic Data Combination

Synthetic data can be used to fully train QE models when no human-labelled data are available, thus alleviating the training data bottleneck for supervised models. When human post-edited data are available however, it remains to be determined if checkpoint-based synthetic data can be used in a complementary manner to further improve the accuracy of QE models.

To study this question, we trained QE models on datasets that merged the QE training data of the WMT shared task with synthetic data generated from a separate dataset. For both English and Romanian, we thus randomly sampled sentences from the selected source monolingual datasets and generated synthetic data with the *<b75, best>* variant, which provided robust results across the board.[7] Since the shared task training datasets consist of 7K data points, we considered three different merged data partitions: (i) merging the 7K WMT QE training data with 7K tuples from the synthetic data; (ii) merging the WMT QE 7K with 70K synthetic tuples, corresponding to our medium datasets in the previous experiments; and (iii) upsampling the QE

[7]For the WikiMatrix and WMT21-MT datasets, the selected source sentences were the same as in the previous experiments.

training data to 70K and merging them with 70K synthetic tuples. There were thus two balanced datasets, and one unbalanced with an order of magnitude more synthetic data points.

The results of these experiments are shown in Table 4. At the sentence level, combinations of synthetic and human data outperformed the baseline in all cases for EN-DE and in 6 out of 9 combinations in RO-EN. At the word level, in RO-EN the baseline was outperformed by the balanced 70K models trained on WikiDump and WikiMatrix data, but obtained better results in the other configurations. In EN-DE, all combinations outperformed the baseline at the word level as well. Regarding data combination volumes, balancing the amount of human and synthetic data proved optimal on all three datasets. Slight improvements were obtained with the larger datasets, although the impact of upsampling the human QE data should be further analysed to measure eventual overfitting side-effects with this data augmentation approach. Finally, the top-performing variants were obtained by mixing the post-edited Wikipedia data with the synthetic data from WikiMatrix and WikiDump, but, as was the case in the previous experiments, the results obtained with the WMT21-MT corpus were competitive overall.

The synthetic data generated via checkpoint translation can thus provide additional accuracy to QE models based on human post-edited data, at both word and sentence levels. We left further experimentation for future research, notably the combination of natural data with mixed synthetic data sampled from different domains.

8 Conclusions

In this work, we described a novel approach to synthetic data generation for translation quality estimation, based on translation models at different learning stages. We exploited NMT model checkpoints, derived from standard training processes, to generate faulty translations that can be contrasted with either human references in parallel datasets, or the translations produced by the converged NMT model. We showed that the latter approach outperformed the use of human references by a significant margin, demonstrating the effectiveness of our method to directly exploit monolingual corpora for synthetic QE data generation. We also showed that checkpoint-based QE performed markedly better than both self-supervised QE and contrasting MT output with human references on parallel data.

The synthetic data generated under our approach was shown to match, or be competitive with, human post-edited data, with a relatively minor impact of domain relatedness between the synthetic training data and the test data in our experiments. We also demonstrated that combining human-generated and synthetic data led to significant improvements on the QE tasks, showing the potential of our approach as both a standalone solution when no human-labelled data are available, and as a complementary option when such data are available.

The main drawback of the checkpoint-based approach is the need to train a separate NMT model for synthetic data generation. However, since the goal of these models is to generate pairs of translations of differing relative quality, there is no requirement for them to be trained on large volumes of data to achieve high translation quality. As shown by our results in Romanian-English, using a relatively small MT training corpus can lead to quality QE synthetic datasets.

Our approach could be further explored along different lines. In this work, we only selected two arbitrary checkpoint models for our experiments, based on their distance to the converged model in terms of BLEU. Additional checkpoints could be used to enrich the synthetic datasets, exploiting earlier or later training stages. The relative distances between checkpoints, or alternative selection metrics beyond BLEU, could also be used to determine optimal checkpoints for QE data generation. Further experimentation will also be relevant to assess optimal data sampling and combination strategies, for specific domains in particular. Finally, determining if the errors learned from checkpoints may bias the QE system towards model-specific error types would require a dedicated analysis as well. We leave these research questions for future work.

References

Artetxe, M. and Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Baek, Y., Kim, Z. M., Moon, J., Kim, H., and Park, E. (2020). PATQUEST: Papago translation quality estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 991–998, Online. Association for Computational Linguistics.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Berard, A., Calapodescu, I., and Roux, C. (2019). Naver Labs Europe's Systems for the WMT19 Machine Translation Robustness Task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 526–532, Florence, Italy. Association for Computational Linguistics.

Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S., Goutte, C., Kulesza, A., Sanchis, A., and Ueffing, N. (2004). Confidence estimation for machine translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 315–321, Geneva, Switzerland. COLING.

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., and Tamchyna, A. (2014). Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., Huck, M., Koehn, P., Liu, Q., Logacheva, V., Monz, C., Negri, M., Post, M., Rubino, R., Specia, L., and Turchi, M. (2017). Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (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.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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.

Ding, S., Junczys-Dowmunt, M., Post, M., and Koehn, P. (2021). Levenshtein training for word-level quality estimation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6724–6733, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Etchegoyhen, T. and Gete, H. (2020). To case or not to case: Evaluating casing methods for neural machine translation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3752–3760, Marseille, France. European Language Resources Association.

Etchegoyhen, T., Martínez Garcia, E., and Azpeitia, A. (2018). Supervised and unsupervised minimalist quality estimators: Vicomtech's participation in the WMT 2018 quality estimation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 782–787, Belgium, Brussels. Association for Computational Linguistics.

Fan, K., Wang, J., Li, B., Zhou, F., Chen, B., and Si, L. (2019). "bilingual expert" can find translation errors. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Felice, M. and Specia, L. (2012). Linguistic features for quality estimation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 96–103, Montréal, Canada. Association for Computational Linguistics.

Fomicheva, M., Sun, S., Yankovskaya, L., Blain, F., Guzmán, F., Fishel, M., Aletras, N., Chaudhary, V., and Specia, L. (2020). Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

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.

Forcada, M. L., Esplà-Gomis, M., Sánchez-Martínez, F., and Specia, L. (2017). One-parameter models for sentence-level post-editing effort estimation. In *Proceedings of Machine Translation Summit XVI: Research Track*, pages 132–143, Nagoya Japan.

Ive, J., Blain, F., and Specia, L. (2018). deepQuest: A framework for neural-based quality estimation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3146–3157, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A. F., Bogoychev, N., Martins, A. F. T., and Birch, A. (2018). Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Kepler, F., Trénous, J., Treviso, M., Vera, M., Góis, A., Farajian, M. A., Lopes, A. V., and Martins, A. F. T. (2019a). Unbabel's participation in the WMT19 translation quality estimation shared task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 78–84, Florence, Italy. Association for Computational Linguistics.

Kepler, F., Trénous, J., Treviso, M., Vera, M., and Martins, A. F. T. (2019b). OpenKiwi: An open source framework for quality estimation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 117–122, Florence, Italy. Association for Computational Linguistics.

Kim, H. and Lee, J.-H. (2016). Recurrent neural network based translation quality estimation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 787–792, Berlin, Germany. Association for Computational Linguistics.

Kim, H., Lee, J.-H., and Na, S.-H. (2017). Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. In *Proceedings of the Second Conference on Machine Translation*, pages 562–568, Copenhagen, Denmark. Association for Computational Linguistics.

Kim, H., Lim, J.-H., Kim, H.-K., and Na, S.-H. (2019). QE BERT: Bilingual BERT using multi-task learning for neural quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 85–89, Florence, Italy. Association for Computational Linguistics.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Lee, D. (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.

Martins, A. F. T., Kepler, F., and Monteiro, J. (2017). Unbabel's participation in the WMT17 translation quality estimation shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 569–574, Copenhagen, Denmark. Association for Computational Linguistics.

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

Moreau, E. and Vogel, C. (2012). Quality estimation: an experimental study using unsupervised similarity measures. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 120–126, Montréal, Canada. Association for Computational Linguistics.

Negri, M., Turchi, M., Chatterjee, R., and Bertoldi, N. (2018). ESCAPE: a large-scale synthetic corpus for automatic post-editing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Popović, M. (2012). Morpheme- and POS-based IBM1 and language model scores for translation quality estimation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 133–137, Montréal, Canada. Association for Computational Linguistics.

Quirk, C. B. (2004). Training a sentence-level machine translation confidence measure. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Schwenk, H., Chaudhary, V., Sun, S., Gong, H., and Guzmán, F. (2021). WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Shah, K., Bougares, F., Barrault, L., and Specia, L. (2016). SHEF-LIUM-NN: Sentence level quality estimation with neural network features. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 838–842, Berlin, Germany. Association for Computational Linguistics.

Shah, K., Logacheva, V., Paetzold, G., Blain, F., Beck, D., Bougares, F., and Specia, L. (2015). SHEF-NN: Translation quality estimation with neural networks. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 342–347, Lisbon, Portugal. Association for Computational Linguistics.

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 the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

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., Blain, F., Fomicheva, M., Zerva, C., Li, Z., Chaudhary, V., and Martins, A. F. T. (2021). Findings of the WMT 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 684–725, Online. Association for Computational Linguistics.

Specia, L., Blain, F., Logacheva, V., F. Astudillo, R., and Martins, A. F. T. (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.

Specia, L., Raj, D., and Turchi, M. (2010). Machine translation evaluation versus quality estimation. *Machine translation*, 24:39–50.

Specia, L., Shah, K., de Souza, J. G., and Cohn, T. (2013). QuEst - a translation quality estimation framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 79–84, Sofia, Bulgaria. 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.

Sun, S., Guzmán, F., and Specia, L. (2020). Are we estimating or guesstimating translation quality? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6262–6267, Online. Association for Computational Linguistics.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Tuan, Y.-L., El-Kishky, A., Renduchintala, A., Chaudhary, V., Guzmán, F., and Specia, L. (2021). Quality estimation without human-labeled data. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 619–625, Online. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Zerva, C., Blain, F., Rei, R., Lertvittayakumjorn, P., C. De Souza, J. G., Eger, S., Kanojia, D., Alves, D., Orăsan, C., Fomicheva, M., Martins, A. F. T., and Specia, L. (2022). Findings of the WMT 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Zheng, Y., Tan, Z., Zhang, M., Maimaiti, M., Luan, H., Sun, M., Liu, Q., and Liu, Y. (2021). Self-supervised quality estimation for machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3322–3334, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhou, J., Zhang, Z., and Hu, Z. (2019). SOURCE: SOURce-conditional elmo-style model for machine translation quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 106–111, Florence, Italy. Association for Computational Linguistics.