Computational Linguistics in the Netherlands Journal 12 (2022) 99-123 Submitted 12/2022; Published 12/2022

Integrating Fuzzy Matches into Sentence-level Quality Estimation for Neural Machine Translation

Arda Tezcan ARDA.TEZCAN@UGENT.COM

LT³, Language and Translation Technology Team
Ghent University
Belgium

Abstract

Previous studies show that neural machine translation (NMT) systems produce translations with higher quality when highly similar sentences (i.e. fuzzy matches; FMs) to a given input sentence can be found in the NMT training data. This study explores the usefulness of FMs for the task of sentence-level quality estimation (QE) for NMT. To this end, fuzzy matches are integrated into the QE architecture that utilizes a pre-trained XLM-RoBERTa model, through a data augmentation methodology. The results show that FMs improve QE performance in domain-specific scenarios when using translation edit rate (TER) as quality labels. However, similar improvements are not observed when the same methodology is applied to a general-domain setting when quality labels were generated through direct (manual) assessment of translation quality or by measuring the technical post-editing effort required for transforming the MT output to its post-edited version.

1. Introduction

Quality estimation (QE) is defined as the task of predicting the document-, sentence- or word-level quality of machine-translated texts without any access to gold-standard human (i.e. reference) translations (Blatz et al. 2004, Specia et al. 2009, Wong and Kit 2012). QE has been an actively explored area in the field of machine translation (MT) given its many applications, which include error analysis (Ueffing and Ney 2007), comparing the translation quality of different MT systems (Rosti et al. 2007), filtering out low-quality translations for human post-editing (Specia et al. 2009) and selecting high-quality translations to be published as they are (Soricut and Echihabi 2010).

In sentence-level QE, state-of-the-art predictive models commonly rely on the information provided by source-MT pairs to assess the quality of a given MT output, since the quality of the MT output is estimated in terms of its correctness compared to the given source text. In two recent studies, it has been shown that transformer-based domain-specific neural machine translation (NMT) systems, which are trained on source-target sentence pairs, produce translations with higher quality when highly similar sentences to the input sentence (i.e. high fuzzy matches, FMs) were present in the NMT training data (Bulté and Tezcan 2019, Tezcan and Bulté 2022). These findings suggest that, in the context of domain-specific NMT, the degree of similarity between the input sentence and the NMT training data plays an important role in the quality of translations produced by the same system.

With the hypothesis that, for a given source sentence, the FMs retrieved from the NMT training data are informative for estimating the quality of the NMT output, the current study explores their usefulness for the sentence-level QE task. More concretely, this study aims to answer the following research questions:

- RQ1: How does the performance of the FM-augmented QE models compare to that of a baseline QE model?

- RQ2: What is the optimal data size for training FM-augmented QE models in the context of transfer learning?
- RQ3: How does QE performance differ in domain-specific and general-domain scenarios? Does predicting different quality labels result in different QE performances?
- RQ4: How does FM similarity score influence the performance of FM-augmented QE models?

To answer these research questions, this study adopts a data augmentation approach, which is originally proposed in the context of NMT (Bulté and Tezcan 2019). Using this approach, for a given source sentence, the best FM retrieved from the NMT training data is simply concatenated to a typical QE input sequence, which consists of the source-MT pair. Using the FM-augmented data sets, sentence-level QE models are trained in the context of transfer learning, by utilizing the XLM-RoBERTa (XLM-R) model (Conneau et al. 2020) within the TransQuest framework (Ranasinghe et al. 2020).

The first set of experiments is conducted on two domain-specific data sets and language pairs, namely the TM of the European Commission's translation service (DGT-TM) for the English–Dutch (EN–NL) language pair and the United Nations corpus (UN) for the English–French (EN–FR) language pair. The results obtained in domain-specific settings show that this simple data-augmentation method improves sentence-level QE performance compared to the baseline systems and better QE performance is achieved with increasing FM scores. The results also demonstrate that, by relying only on the information obtained from the source sentences and the NMT training data (i.e. in the absence of the MT output), decent QE performances can be achieved. To obtain a better picture of the impact of FM-augmentation on sentence-level QE, the experiments are repeated in a general-domain setting using the WMT 2020 Wikipedia data set for the English–German (EN–DE) language pair, while using a different set of quality labels. The experiments conducted in the general-domain scenario, however, show that the FM-augmentation approach does not lead to any statistically significant differences in QE performances. This study additionally identifies two potential reasons for the discrepancy between the QE performances observed in the two scenarios. Firstly, the different quality labels used in both scenarios potentially measure different aspects of translation quality, and the FMs, which are retrieved using cosine similarity between sentence embeddings, might only be informative for predicting only certain types of quality labels. Secondly, the similarity scores for the FMs retrieved for the test sentences used in the general-domain scenario are observed to be much lower on average compared to the domain-specific scenario. Given that the domain-specific QE systems perform better with higher FM scores, the FMs retrieved in the general-domain scenario are potentially less informative than the ones retrieved in the domain-specific scenario.

The remainder of this paper is structured as follows: in the next part, relevant previous research is discussed (Section 2). The methodology is described in Section 3, followed by the results (Section 4), and their discussion (Section 5). In the final section (Section 6), conclusions are drawn up.

2. Related Research

This section briefly describes the different types of QE tasks (Section 2.1) and QE systems (Section 2.2), before discussing previous attempts that utilized FMs for the task of QE (Section 2.3). We then summarize FM utilization approaches within the NMT framework and outline the neural fuzzy repair methodology, which is adopted in this study for the sentence-level QE task (Section 2.4).

2.1 QE tasks

QE is typically addressed as a supervised machine learning task where the goal is to predict the quality of MT output on word, sentence, or document levels, in the absence of reference (i.e. gold-standard) translations (Specia et al. 2020, Specia et al. 2021).

Word-level QE focuses on detecting word-level translation errors in a machine-translated text, where the aim is to predict binary quality labels per word as correct or incorrect. In literature, word-level quality labels have been generated in different ways, such as by utilizing coarse- and fine-grained human error annotations (Popović and Ney 2011, Tezcan 2018, Popović 2018, Specia et al. 2018) or by automatically marking words that require post-editing (i.e. technical post-editing effort) using automatic MT evaluation metrics, such as translation edit rate (TER) (Snover et al. 2006, Specia et al. 2020, Specia et al. 2021)

In sentence-level QE the goal is to often predict a (continuous) quality score, which reflects how close a machine-translated text is to a gold-standard translation. In the past, quality scores have been obtained through direct assessment (human evaluation) (Specia et al. 2020), by comparing MT output to its post-edited version, using automatic MT evaluation metrics (Specia and Farzindar 2010), or by measuring the time it takes to post-edited a given MT output (Bojar et al. 2013). More recently a new variant of the sentence-level QE task has been introduced, which aims to predict a sentence-level binary score indicating whether a translation contains (at least one) critical error, which may carry health, safety, or legal implications (Specia et al. 2021). Word- and sentence-level QE tasks, which utilize different quality labels or scores, can also be extended to a document level, which is referred to as document-level QE (Specia et al. 2020, Specia et al. 2021).

2.2 QE systems

Earlier work on QE mostly focused on machine learning methods that rely on linguistic processing and feature engineering (Hardmeier et al. 2012, Specia et al. 2013, Specia et al. 2015). In feature-based approaches to QE, linguistic features were mostly extracted from MT systems (glass-box features) or obtained from source-machine-translated sentence pairs, and external resources (black-box features).

In the last decades, (deep) neural networks (NNs) have shown outstanding performance in the field of natural language processing (NLP) and have led to improvements in various NLP tasks, including QE for MT. Firstly, recurrent NNs (RNNs) were successfully adopted for the QE task (Kim and Lee 2016, Patel and Mukundan 2016). Using RNNs, Kim et al. (2017) proposed a two-stage QE architecture referred to as predictor-estimator which consists of an encoder-decoder RNN (predictor) trained on parallel data for a word prediction task and a unidirectional RNN (estimator) that estimates the quality of a given MT output by using the representations generated by the predictor model. With the advances in the context of transfer learning, recent work on QE focused on fine-tuning large-scale, pre-trained models for the different QE tasks. For example, Kepler et al. (2019) replaced the predictor component in the predictor/estimator architecture with pre-trained BERT (Devlin et al. 2019) and XLM (Conneau and Lample 2019) models. More recently, the cross-lingual XLM-R model has been successfully utilized in the QE tasks, either as the predictor component in the predictor/estimator architecture (Chen et al. 2021), as a stand-alone model (Ranasinghe et al. 2020), or as an ensemble of different XLM-R checkpoints (Zerva et al. 2021). Such XLM-R-based QE models achieved state-of-the-art results in the WMT shared tasks on word- and sentence-level QE in recent years (Specia et al. 2020).[1][2]. It should also be noted that, in multiple studies, cross-lingual models, such as XLM-R, have been reported to perform better than multilingual models, such as mBERT (Pires et al. 2019) and mBART (Liu et al. 2020) on different QE tasks (Ranasinghe et al. 2020, Eo et al. 2021, Zerva et al. 2021).

2.3 Integration of FMs into QE for MT

To our knowledge, this is the first study that directly integrates FMs retrieved for a given source sentence from the MT training data into a sentence-level QE architecture. However, in the field of

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

[2] https://www.statmt.org/wmt20/quality-estimation-task_results.html

QE, the idea of utilizing the level of similarity between a given source sentence to the MT training data is not new. For example, the QuEst toolkit utilized a wide range of features extracted from source texts, MT output, external language resources, as well as from the MT training data (Specia et al. 2013). Regarding the similarity of a given source sentence to the MT training data, the QuEst toolkit utilized features that indicated how frequent source n-grams of different sizes also appeared in the MT training data set. In a more recent study, Wang et al. (2021) integrated a set of features into a neural QE architecture based on the highest n similarity scores measured between a given input sentence and the NMT training data. Similar to the QE architecture used in this study, Wang et al. (2021) utilized the XLM-R model in the context of transfer learning. The results of the experiments conducted on the WMT 2020 sentence-level QE data set showed that, when combined with other linguistic features, such similarity-based features led to a decrease in QE performance.

2.4 Integration of FMs into NMT

While the usefulness of integrating FMs into neural QE architectures is yet to be demonstrated, FMs have been successfully utilized to improve NMT models in the past. Within the NMT paradigm, various modifications to the NMT architecture and search algorithms were proposed to leverage information from FMs. For example, Cao and Xiong (2018) added an encoder to the NMT architecture, which specifically utilized FMs retrieved from the NMT training data. Alternatively, NMT models have also been adapted to incorporate lexical constraints obtained from FMs during the decoding stage (Gu et al. 2018), to take into account rewards attached to text fragments that are found in FMs (Zhang et al. 2018), or to generate tokens in the target language by employing a nearest neighbor classifier, which utilizes similar translations retrieved from the NMT training data (Khandelwal et al. 2020, Meng et al. 2022). Whereas most of the approaches that utilize FMs for NMT require modifications to the NMT architectures or decoding algorithms, to this end FMs have also been integrated into NMT through augmenting source sentences with translations of FMs retrieved from the NMT training data at training and inference times (Bulté and Tezcan 2019, Xu et al. 2020, Tezcan et al. 2021). This data augmentation approach, referred to as neural fuzzy repair (NFR), forms the basis of the QE methodology in this study.

In NFR, for a given TM, or a bilingual data set used for training the NMT model, consisting of source/target sentence pairs S, T, each source sentence $s_i \in S$ is augmented with the translations $\{t_1, \ldots, t_n\} \in T$ of n FMs $\{s_1, \ldots, s_n\} \in S$, where $s_i \notin \{s_1, \ldots, s_n\}$, given that the FM score is sufficiently high (i.e., above a given threshold λ). Sentence similarity is either measured using token-based edit distance (Bulté and Tezcan 2019) or cosine similarity between sentence embeddings (Tezcan et al. 2021).

The NMT model is then trained using the combination of the original TM, which consists of the source/target sentence pairs S, T, and the augmented TM, consisting of augmented-source/target sentence pairs S', T. At inference, each source sentence is augmented using the same method. If no FMs are found with a match score above λ, the non-augmented (i.e., original) source sentence is used as input to obtain translations. Figure 1 illustrates the NFR method for the EN–NL language pair, which utilizes a single FM for data augmentation.

3. Methodology

This section first describes the data sets (Section 3.1) and the NMT systems (Section 3.2) used for the domain-specific QE experiments conducted in the study. We then provide more details on the two core methodologies used for integrating FMs into the sentence-level QE task: FM retrieval (Section 3.3), and the FM-augmented QE architecture, which utilizes FM-augmented data sets in the context of transfer learning (Section 3.4).

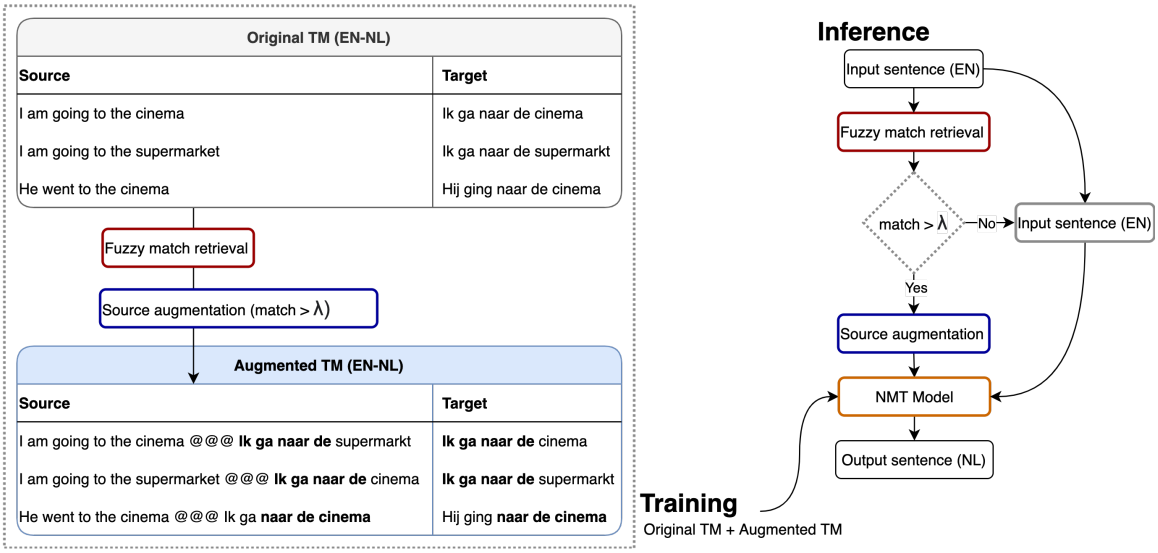


Figure 1: Neural fuzzy repair: training and inference (Tezcan et al. 2021).

3.1 Data

3.1.1 Domain-specific Setting

For the domain-specific QE experiments, data sets from two domains and in two language directions were used: the TM of the European Commission's translation service (DGT-TM), which consists of texts written for mostly legal purposes, such as contracts, reports, regulations, directives, policies and plans within the Commission (EN–NL) (Steinberger et al. 2012), and the UN parallel corpus, which consists of the official records and parliamentary documents of the United Nations (EN–FR) (Ziemski et al. 2016).

Each data set was partitioned into a training set, two validation sets, and a test set. To this end, random sentences were selected from the two data sets consisting of a minimum of 3 and a maximum of 85 tokens. In the first step, the DGT and UN data sets were used to train domain-specific NMT systems (by using the training, validation 1, and test sets). The data sets were also used to build FM-augmented sentence-level QE systems, in the second step (by using the training, validation 2, and test sets).

To train QE systems, sentence-level quality labels were automatically extracted using Translation Edit Rate (TER) (Snover et al. 2006)[3]. On the test set, the TER scores were calculated by comparing the reference translations with the translations obtained from the NMT model, which was trained with the full training set (as outlined in Table 1). As this study initially aims to utilize the whole NMT training data also for the QE task and additionally to measure the impact of different training sizes on QE performance, 4-fold jackknifing was used to obtain unbiased translations on the NMT training set. To this end, the NMT training set was further partitioned into four equal splits and four additional NMT models were trained, each using a different combination of the three splits as training data and the original validation set. These four NMT models were then used to obtain translations on the remaining split in each case. The number of sentences used for each domain, language pair, and partition, is provided in Table 1. In both data sets, the different partitions did not contain any overlapping sentences.

[3] Version 0.7.25: `https://github.com/snover/terp`

Table 1: NMT data set: number of sentences in training, validation (Val. 1 and Val. 2), and test sets, per domain and language pair.

Domain (Language Pair)	**Train**	**Val. 1**	**Val. 2**	**Test**	**Quality Labels**
DGT (EN–NL)	708695	2500	2500	2500	TER
UN (EN–FR)	887611	2500	2500	2500	TER

3.1.2 General-domain Setting

For the general-domain QE experiments, the EN–DE data set of the WMT 2020 shared task on sentence-level QE estimation was used (Fomicheva et al. 2020b, Specia et al. 2020). The WMT QE data set[4] consists of source sentences in the general domain, which were extracted from Wikipedia articles, as well as NMT output and quality labels per source sentence. The NMT output in this data set was produced by transformer-based NMT models trained on approximately 23 million sentence pairs extracted from Wikipedia articles, using the fairseq toolkit (Ott et al. 2019)[5]

In the WMT 2020 data set, sentence-level quality labels were produced both manually and automatically. For each sentence, the manual quality (i.e. direct assessment, DA) scores were obtained independently from at least three different professional translators from a single language service provider (sentence-level direct assessment task). To this end, each sentence was scored between 0 and 100, following the FLORES guidelines (Guzmán et al. 2019), according to the perceived translation quality. DA scores were normalized by taking the mean z-scores per sentence.

The automatic quality scores were obtained using HTER (Human-targeted Translation Edit Rate) as the minimum edit distance between a given MT output and its manually post-edited version (sentence-level post-editing task). In the context of general-domain QE, this study investigates the performance of FM-augmented QE models for both sub-tasks. The number of sentences in the WMT 2020 shared tasks on sentence-level QE is provided in Table 2.

Table 2: WMT 2020 data set used for the sentence-level QE shared tasks: number of sentences in the training, validation, and test sets.

Domain (Language Pair)	**Train**	**Validation**	**Test**	**Quality labels**
Wikipedia (EN–DE)	7000	1000	1000	DA/HTER

3.2 NMT systems

All the in-domain NMT systems were trained using the Transformer architecture (Vaswani et al. 2017) and the OpenNMT toolkit (Klein et al. 2017). Prior to training the NMT models, the data sets were segmented into sub-words using SentencePiece (Kudo and Richardson 2018), using the XLM-R (base) tokenizer[6]. The resulting vocabulary sizes were approximately 36K and 44K for the DGT (EN–NL) and the UN (EN–FR) data sets, respectively. For each training, a total of 1 million steps were used with validation at every 5000 steps. All of the models were trained with early stopping: the training ended when the system has not improved for 10 validation rounds in terms of both accuracy and perplexity. All training runs were initialized using the same seed to avoid differences between systems due to the effect of randomness. Other details regarding the hyper-parameters used for training the NMT models are provided in Appendix A.1

[4] https://www.statmt.org/wmt20/quality-estimation-task.html

[5] The official page of the WMT 2020 shared task points to a data set that consists of approximately 23 million sentence pairs as the parallel data used to train the NMT models. However, the exact number of sentences used in training, validation, or test sets, or the implementation of the details of the NMT modes are not further provided.

[6] https://huggingface.co/docs/transformers/v4.22.2/en/model_doc/xlm-roberta\#overview

To provide an indication of the average quality of each NMT system, Table 1 presents automated evaluation results obtained on each test set, in terms of the BLEU[7] (Papineni et al. 2002) and TER scores. The evaluations were performed after the MT output was detokenized, in each case. It should be noted that, while the evaluation for the domain-specific data sets was performed against reference translations, in the general-domain case, post-edited MT output was used for this purpose. This information is also provided in Table 1.

Table 3: Performance of the NMT models on the test set of each data set.

Domain (Language Pair)	**BLEU**	**TER**	**Evaluation vs.**
DGT (EN–NL)	44.10	0.489	reference translations
UN (EN–FR)	47.42	0.416	reference translations
Wikipedia (EN–DE)	72.37	0.172	post-edited MT output

3.3 FM retrieval

Fuzzy matching is a key functionality in NFR, as the quality of the generated translations is determined by the similarity level of the retrieved FMs (Bulté and Tezcan 2019, Tezcan et al. 2021). In the original NFR approach, when augmenting source sentences with FMs, a minimum FM similarity threshold is used ($\lambda = 0.5$), as the FMs with low similarity to the input did not lead to improved translation quality. As a result, to allow the NMT model to translate both the original and the augmented source sentences at inference time, the original training data is combined with its augmented version during training (see Figure 1).

With the hypothesis that FMs with low similarity can also be informative for predicting the quality of NMT output, in the QE task, we do not set a minimum FM similarity threshold. To this end, for each source sentence in the QE data set (training, validation, and test sets), we seek FMs against the NMT training data and retrieve the FM with the highest similarity score (i.e. best FM). Similar to (Tezcan et al. 2021), the sentence similarity score $SE(s_i, s_j)$ between two source sentences s_i and s_j is measured as the cosine similarity of their sentence embeddings e_i and e_j, that is,

$$SE(s_i, s_j) = \frac{e_i \cdot e_j}{\|e_i\| \times \|e_j\|} \tag{1}$$

where $\|e\|$ is the magnitude of vector e. To generate sentence embeddings, we use sent2vec (Pagliardini et al. 2018), and for efficient retrieval of FMs, we build a FAISS index (Johnson et al. 2021). FAISS is a library specifically designed for efficient similarity search and vector clustering and is compatible with the large data sets used in this study. The hyper-parameters used for generating sentence embeddings and for building the FAISS index are provided in Appendices A.2 and A.3, respectively. Prior to retrieving FMs, all sentences were segmented into sub-words using SentencePiece, with the same methodology described in Section 3.3.

Table 4 illustrates the FM retrieval process for the source sentence (SRC) *"This Common Position shall take effect on the day of its publication."*. In this table, the FM similarity score, the source (English) and target side (Dutch) of the best FM found for the source sentence are indicated as *FM Score*, *FMsrc* and *FMtgt*, respectively. The table also includes the machine-translated version of the source text (MT), and the reference (i.e. gold-standard) translation (REF) in the target language.

[7] `https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl`

Table 4: An example of FM retrieval for the English-Dutch language pair.

SRC	This Common Position shall take effect on the day of its publication.
FM Score	0.946
FMsrc	This Common Position shall take effect on the day of its adoption.
FMtgt	Dit gemeenschappelijk standpunt wordt van kracht op de dag van zijn aanneming.
MT	Dit gemeenschappelijk standpunt wordt van kracht op de dag van de bekendmaking ervan.
REF	Dit gemeenschappelijk standpunt wordt van kracht op de dag van zijn bekendmaking.

3.4 QE architecture

To build sentence-level QE models, we adopt a transfer-learning approach and use the MonoTransQuest architecture (MTQ) (Ranasinghe et al. 2020), which was the winner of the WMT 2020 sentence-level direct assessment task for all language pairs (Specia et al. 2020). The MTQ architecture uses a single cross-lingual XLM-R model, which is fine-tuned for the sentence-level QE task during training. The input of this model is a concatenation of the original source sentence and its translation (MT output), separated by a separator ([SEP]) token. The output of the classification token ([CLS]) is used as the input of a softmax layer that predicts the quality score of the translation, using mean-squared error loss as the objective function.

To integrate FMs into the MTQ architecture, the original input representation (SRC–MT pair) was extended by concatenating the source (FMsrc) or target (FMtgt) side of the best FM retrieved for each source sentence using an additional [SEP] token. The original MTQ architecture and the modification we made to this architecture in this study are illustrated in Figure 2.

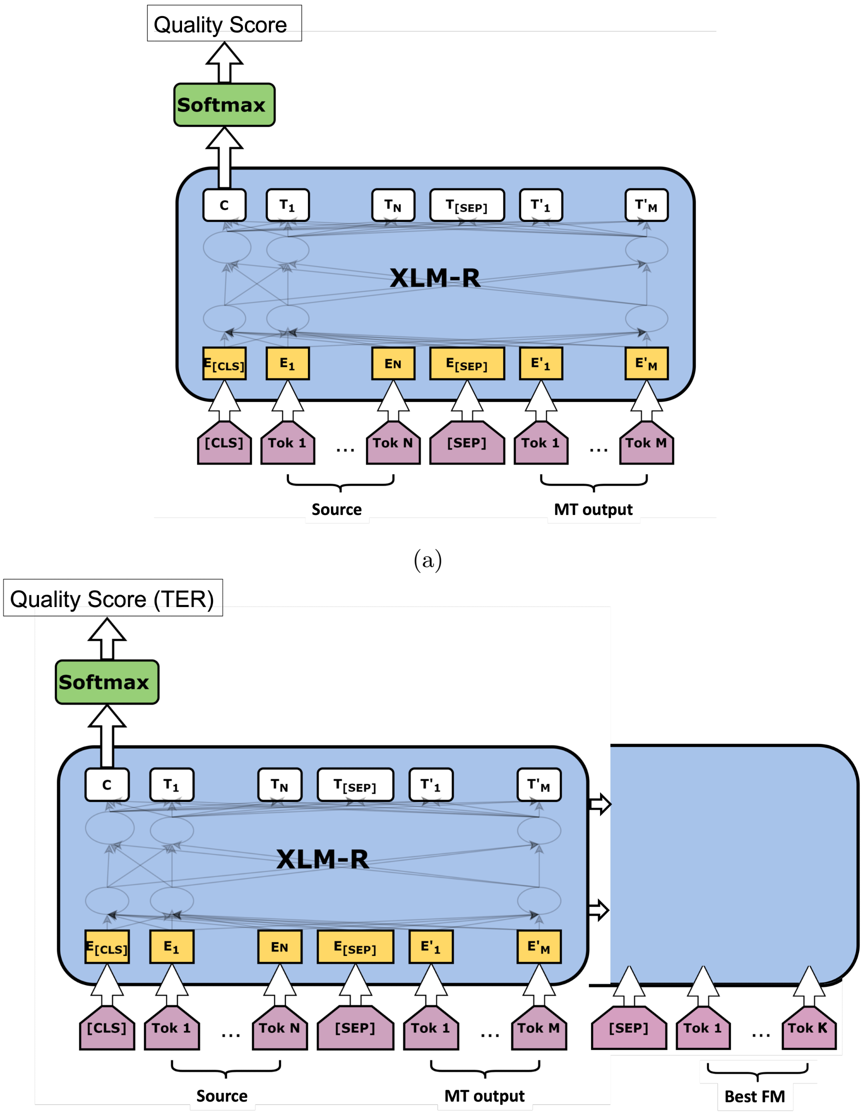
Besides extending the original SRC–MT pairs with the source or target sides of the best FMs (SRC–MT–FMsrc vs. SRC–MT–FMtgt), we also experimented with other input configurations. To further investigate the usefulness of FMs in the absence of the MT output, additional experiments were conducted by concatenating a given input source sentence with the source, target, or both source and target sides of the best FM obtained for each source sentence (SRC–FMsrc, SRC–FMtgt, and SRC–FMsrc–FMtgt, respectively). For all different configurations, the data augmentation methodology was applied to all sentences in the training data, as well as the sentences in the validation and the test sets.

For all language pairs we tested, two baseline QE models were trained: an MTQ model using the standard source–MT pairs as input, and a simple linear regression model[8], which uses the FM score for each given input as a single feature. While the purpose of training the baseline MTQ model was to compare the performance of the QE approaches proposed in this study to a state-of-the-art QE approach, the linear regression model was trained to observe the informativeness of the FM score as a single feature on the sentence-level QE task.

All the MTQ models are trained using the TransQuest toolkit[9] with close to standard settings. To train the baseline QE model with source-MT pairs as input the maximum input sequence was set to 170 tokens (instead of the default value of 80). To avoid memory issues, the training batch size was adapted from 8 input sequences to 4, while also changing the gradient accumulation steps from 1 to 2. Using these settings, the QE performance of the baseline model in the WMT 2020 shared task (EN–DE) could be reproduced by further modifying the learning rate to 1×10^{-6} (from 2×10^{-5}) and the number of validation steps to 100 (from 300). For training QE models with augmented input, which utilized three input sentences (for example the SRC-MT-FMsrc triplet), the maximum input sequence length was increased to 256 tokens. All QE models were trained with early stopping: the training ended when the system has not improved for 10 validation rounds in terms of mean

[8] Linear regression models were built using the SciPy library (Virtanen et al. 2020): `https://github.com/scipy/scipy`.

[9] TransQuest toolkit: `https://github.com/TharinduDR/TransQuest`



(a)

(b)

Figure 2: The original MonoTransQuest architecture (Ranasinghe et al. 2020) (a), and the modified MonoTransQuest architecture, in which the input representation is extended by the source or target side of the best FM retrieved for each given source sentence (b).

squared error (MSE). The same hyper-parameter values were also used for training domain-specific QE models with the exception that the number of validation steps was proportionally increased with increasing training data set sizes (up to 6400) (see Section 4.1.2 for the experiments that use different sizes of training data). All models were trained on a single Tesla V100 SXM2 GPU with 16GB of memory. An overview of the hyper-parameters used for training the MTQ models is provided in Appendix A.4.

4. Results

In this section, first, an analysis of the results for the domain-specific settings is provided (EN-NL and EN–FR) (Section 4.1). Subsequently, Section 4.2 shows the results for the general-domain setting (EN–DE). Similar to the WMT 2020 shared task on sentence-level quality estimation, Pearson's correlation coefficient (Pearson r) is used as the primary evaluation metric to evaluate the QE performance. Statistical significance on Pearson r was computed using William's test.[10]

4.1 Domain-specific setting

Table 5 shows the performances of the QE models, which were trained using the full training data sets for the EN–NL (DGT) and the EN–FR (UN) language pairs. The table consists of three sections. From top to bottom, it shows the results for (a) the two baseline systems, (b) the QE systems, which additionally utilize the source or target sentences of the FMs retrieved for each source sentence, and (c) the QE systems that incorporate source, target, or both the source and target sides of the retrieved FMs, without the MT output. The table reports Pearson r, as well as Spearman's correlation coefficient (r_s), root mean squared error (RMSE), and mean absolute error (MAE) values for all QE systems.[11].

Table 5: Results of the automated evaluations for DGT (English–Dutch) and UN (English–French) data sets. The best scores are highlighted in bold.

	DGT (EN–NL)				**UN (EN–FR)**			
	$r \uparrow$	$r_s \uparrow$	**RMSE** $\downarrow$	**MAE** $\downarrow$	r	r_s	**RMSE**	**MAE**
Linear Regression	0.489	0.517	0.239	0.295	0.408	0.409	0.172	0.217
SRC–MT	0.556	0.576	0.206	0.277	0.638	0.619	0.139	0.182
SRC–MT–FMsrc	0.572	0.593	0.209	0.271	0.661	0.646	0.136	0.177
SRC–MT–FMtgt	**0.621**	**0.648**	**0.197**	**0.260**	**0.672**	**0.653**	**0.134**	**0.175**
SRC–FMsrc	0.555	0.575	0.214	0.275	0.596	0.590	0.145	0.191
SRC–FMtgt	0.563	0.585	0.207	0.274	0.605	0.592	0.144	0.189
SRC–FMsrc–FMtgt	0.562	0.584	0.210	0.274	0.585	0.575	0.145	0.192

Looking at the baseline QE systems (upper section), we see that for both translation directions, the baseline MTQ model (SRC–MT), outperforms the linear regression model, which utilizes the FM score as a single feature, by 0.067 and 0.230 in terms of Pearson r. Similar improvements can be observed for the remaining evaluation metrics. While observing improvement in QE performance by using state-of-the-art QE models over simple linear regression models is not surprising, it should be noted that, by using a single feature, the linear regression models remarkably achieve a moderate correlation with the gold standard quality labels in both data sets.

In the middle section, we see that the QE systems, which additionally utilize the source or target side of the FMs retrieved for a given source sentence (SRC–MT–FMsrc and SRC–MT–FMtgt, respectively) outperform both baseline systems for both language pairs and evaluation metrics, with

[10] https://github.com/ygraham/nlp-williams

[11] All calculations are performed using the SciPy library (Virtanen et al. 2020): https://github.com/scipy/scipy.

the exception of RMSE for the SRC–MT–FMsrc configuration of the DGT (EN–NL) data set (0.209 vs. 0.206). The results also show that the target side of a retrieved FM is more informative on the sentence-level QE task than the source side when such information is combined with source–MT pairs. Combining source–MT pairs with FMtgt leads to improvements in r values, for the DGT (EN–NL) (+0.049) and EN-FR (+0.011) data sets, respectively.

The same trend can be seen in the lower section of Table 5, which shows the QE performances of the systems that utilize source sentences with the source, target, or both source and target side of the retrieved FMs, in the absence of the MT output. The QE system, which utilizes the target side of the retrieved FMs (SRC–FMtgt) does not only outperform its counterpart, which utilizes FMsrc instead (SRC–FMsrc), but also the system that utilizes both FMsrc and FMtgt (SRC–FMsrc–FMtgt), for all evaluation metrics. This section also shows a clear decrease in QE performance, when the MT output is removed from the SRC–MT–FMtgt triplets. Another interesting observation is that by using only the SRC–FMtgt pair as input (without using the MT output), a better QE performance than the baseline MTQ model (SRC–MT) was achieved for the DGT data set (0.563 r vs. 0.556 r). This observation, however, cannot be made for the UN data set (0.605 r vs. 0.638 r).

When the results are analyzed together, we see that the system which utilizes the SRC–MT–FMtgt triplet outperforms all other systems tested for both data sets and all evaluation metrics, including the baseline MTQ system (SRC–MT) by 0.065 (+12%) and 0.034 (+5%) points in terms of r. For both data sets, the r improvements achieved by the best-performing systems compared to the baseline MTQ systems are statistically significant ($p < 0.001$).

4.1.1 Impact of FM similarity score

In this section, the impact of the FM similarity score on QE performance is analyzed. To this end, the QE performances of the best performing MTQ models (SRC–MT–FMtgt) and the baseline MTQ models (SRC–MT) are provided in Table 6, per FM similarity range (i.e. 0.50-0.59 ... 0.90-0.99) and data set. For both the DGT and the UN data sets, the minimum FM range is observed as 0.50-0.59, when the similarity between source sentence embeddings is measured in terms of their cosine similarity. Additionally, for both language pairs, Table 6 provides the total number of sentences that are found in the corresponding test sets, per FM similarity range.

Table 6: QE performance per FM similarity range in terms of Pearson r. The number of sentences in each FM similarity range is indicated with #Sent.

	DGT (EN–NL)			**UN (EN–FR)**		
	SRC–MT	**SRC–MT–FMtgt**	**#Sent**	**SRC–MT**	**SRC–MT–FMtgt**	**#Sent**
0.50-0.59	0.178	0.355	15	0.542	0.533	36
0.60-0.69	0.361	0.360	387	0.486	0.485	832
0.70-0.79	0.429	0.412	840	0.578	0.583	830
0.80-0.89	0.366	0.436	500	0.595	0.625	375
0.90-0.99	0.539	0.607	758	0.575	0.699	427

Table 6 shows that the QE performance of the best-performing MTQ model, which combines the source-MT pair with FMtgt (SRC–MT–FMtgt) increases with increasing FM similarity score, except for the lowest FM range (0.50-0.59) for the UN data set. It should also be noted that this FM range corresponds to a very small portion of the full test sets (15 and 36 sentences out of 2500 sentences in the DGT and the UN test sets, respectively). A similar positive correlation between the FM similarity ranges and QE performance, however, cannot be observed for the baseline MTQ systems.

When we compare the QE performances of the two systems per FM range, for both data sets, we see a clear improvement in r values, when the FM-augmented model utilizes FMs in high similarity ranges (above 0.80). The total number of sentences that are observed above the 0.80 similarity range corresponds to 50% (1258 sentences) and 32% (802 sentences) of the whole test set for the DGT and UN data sets, respectively. While such clear improvements cannot be observed for the FM-augmented MTQ models in the lower FM similarity ranges (below 0.80), the results show that, for both data sets, these models still achieve similar QE performance to the baseline MTQ models (i.e. without any evident decrease in QE performance), when the lowest FM range for the DGT data set is excluded from this analysis (due to the low number of sentences found in this range).

4.1.2 Impact of training data size

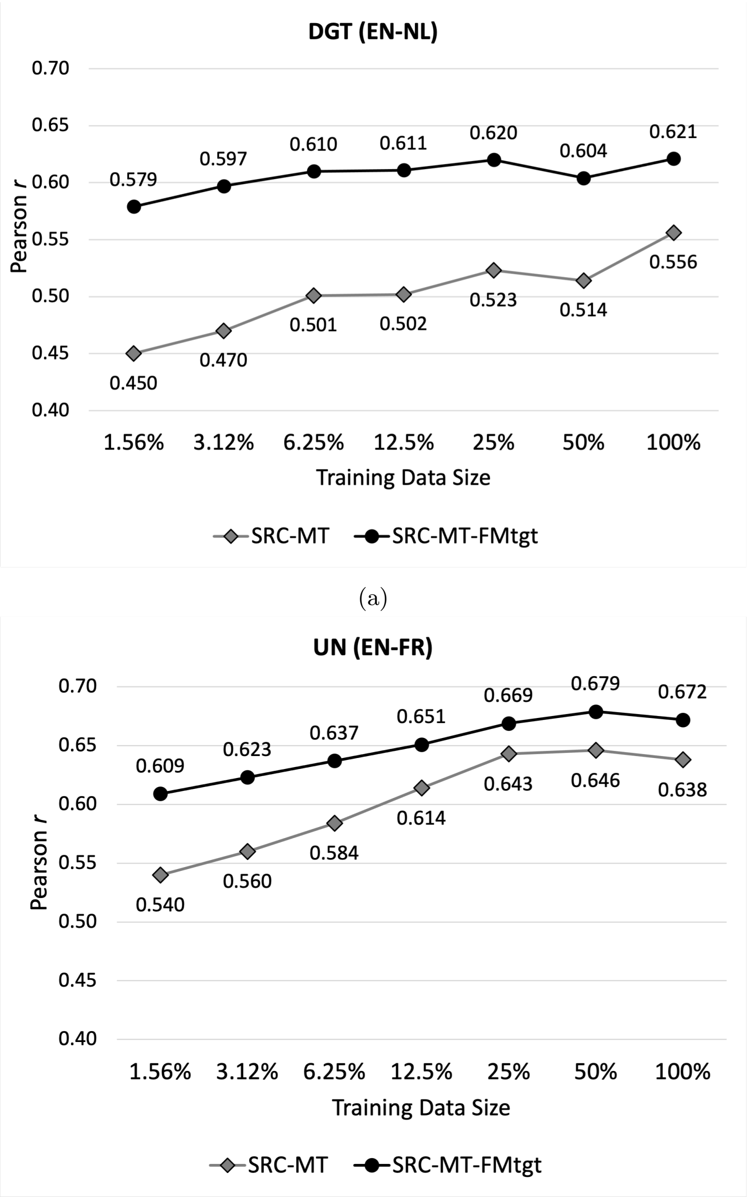
In the domain-specific setting, to analyze the impact of the training data size on QE performance, the performance of the best-performing (SRC–MT–FMtgt) and the baseline (SRC–MT) MTQ models are evaluated for increasingly smaller subsets of the DGT and the UN QE training data. To this end, after training the MTQ models with the full training data (100%), first, a random subset of half the size of the full data is collected. This operation is repeated by taking a random subset of the training set from the previous step until 1.56% of the full training data is reached. The smallest QE training data sets consist of approximately 11K (DGT) and 13K (UN) sentence pairs (SRC–MT) or triplets (SRC–MT–FMtgt). The motivation for training MTQ models in the domain-specific setting with such small data sets is twofold. Firstly, it allows us to measure the potential improvements over the baseline MTQ model in low-resource QE scenarios. Secondly, despite the differences in the quality labels used in the two settings, the QE performances in the domain-specific and general-domain scenarios become comparable to a certain extent, as in the general-domain scenario the full training data set consists of 7000 sentence pairs. Figure 3 shows the QE performance of the MTQ models, which are trained with different data sizes. It should also be noted that, when the QE models are trained with smaller training sets, the NMT models and the NMT output remain the same in each case.

The results in Figure 3 show a clear overall picture for both the DGT and the UN data sets, with an advantage for integrating FMs into the QE architecture, as the FM-augmented MTQ models outperform the baseline models for each QE training set size. For both data sets and each different training set size, the Pearson r improvements achieved by the FM-augmented system compared to the baseline MTQ system are statistically significant ($p < 0.001$). Moreover, the improvements observed over the baseline MTQ models increase with decreasing training set sizes. When trained with the smallest data sets (1.56% of the full training set), the FM-augmented models yield 0.13 (+29%) and 0.17 (+31%) r improvements, for the DGT and the UN data sets, respectively.

These results also reveal the training set sizes that yield near-optimal performance for the MTQ models. For both the DGT and the UN data sets, increasing the QE training set size of the FM-augmented MTQ models above 25% of the full training size does not yield notable improvements in QE performances. In the case of the baseline MTQ models, however, we see that a larger training set size (100%) leads best QE performance for the DGT data set (0.56 r). For the UN data set, similar to the FM-augmented QE model, the baseline MTQ model yields near-optimal performance when trained with 25% of the full training set (0.64 r).

4.2 General-domain setting

The general-domain QE models were trained using the WMT (EN-DE) data set, for the two sentence-level QE tasks in WMT 2020, namely the sentence-level direct assessment task, which uses the manually generated DA labels, and the sentence-level post-editing effort task, which uses HTER scores, as quality labels. FM-augmented MTQ models were built using the best-performing QE configuration in the domain-specific setting (SRC–MT–FMtgt). Similar to the general-domain setting, the performance of the FM-augmented MTQ models was compared to two baseline systems: an MTQ model using the standard source-MT pairs as input (SRC–MT), and a linear regression model, which uses the FM score for each given input as the single feature. Table 7 shows the QE performances



(a)

(b)

Figure 3: The impact of training data sizes on QE performance (Pearson r) for the best performing (SRC–MT–FMtgt) and the baseline (SRC–MT) MTQ models, for (a) the DGT (EN–NL) and (b) the UN (EN–FR) data sets.

of the three QE systems per quality label. The table reports Pearson's correlation coefficient (r), as well as Spearman's correlation coefficient (r_s) RMSE, and MAE values for all QE systems.

Table 7: Results of the automated evaluations for the WMT 2020 data set, for the sentence-level direct assessment and post-editing effort tasks. The best scores are highlighted in bold.

	Direct assessment (DA)				**Post-editing effort (HTER)**			
	$r \uparrow$	$r_s \uparrow$	**RMSE** $\downarrow$	**MAE** $\downarrow$	r	r_s	**RMSE**	**MAE**
Linear Regression	0.081	0.103	0.546	0.696	0.069	0.088	0.160	0.195
SRC–MT	**0.443**	**0.450**	**0.439**	**0.621**	0.485	0.473	0.134	0.175
SRC–MT–FMtgt	0.418	0.422	0.469	0.639	**0.505**	**0.475**	**0.128**	**0.169**

From Table 7, it can be seen that, unlike in the domain-specific setting, the predictions obtained from the linear regression models yield almost zero correlation with the gold-standard DA and HTER scores (0.081 r and 0.069 r). When we look at the Pearson r results for the DA task, we see that the baseline MTQ model (SRC–MT) outperforms the MTQ model that additionally utilizes the target side of the retrieved FMs in the input (SRC–MT–FMtgt) (0.443 vs. 0.418). We see a different picture, however, for the task of predicting post-editing effort. When HTER scores are used as quality labels, the additional information presented by the target side of FMs to the QE model leads to an increase in QE performance (0.485 vs. 0.505). While all evaluation metrics agree on the better-performing system in both the direct assessment task and the task of predicting post-editing effort, for the MTQ models, the differences in Pearson r are measured to be statistically insignificant ($p = 0.08$ and $p = 0.06$, respectively).

4.2.1 Impact of FM similarity score

To obtain more insight into the impact of FM similarity score for both QE tasks in the general domain, we analyze the QE performance of the baseline (SRC–MT) and the FM-augmented MTQ models (SRC–MT–FMtgt) per FM similarity score range. QE performance is measured in terms of Pearson r and the results are provided in Table 8. In the WMT QE test set, the minimum and maximum similarity range for the FMs retrieved from the NMT training data, which consists of approximately 23 million sentence pairs, are observed as 0.3-0.39 and 0.80-0.89, respectively. However, given that only a few sentences are found in these two FM ranges (4 and 5, respectively), they are excluded from Table 8.

Table 8: QE performance per FM similarity range in terms of Pearson r. The number of sentences in each FM similarity range is indicated with #Sent.

	Direct assessment (DA)		**Post-editing effort (HTER)**		
	SRC–MT	**SRC–MT–FMtgt**	**SRC–MT**	**SRC–MT–FMtgt**	**#Sent**
0.40-0.49	0.501	0.452	0.461	0.496	258
0.50-0.59	0.425	0.410	0.494	0.496	484
0.60-0.69	0.477	0.383	0.553	0.538	209
0.70-0.79	0.505	0.447	0.335	0.516	40

For the direct assessment task, we see that the baseline MTQ model (SRC–MT) outperforms the FM-augmented model (SRC–MT–FMtgt) in all FM similarity ranges. A similar trend can be seen for the task of predicting post-editing effort, where the overall best-performing MTQ model (SRC–MT–FMtgt) achieves higher Pearson r in all FM ranges, except the range of 0.60-0.68 (0.553 vs. 0.538). Unlike in the case of the domain-specific setting, we do not observe a clear increase in the QE performance of the FM-augmented MTQ models, with increasing FM similarity scores.

5. Discussion

5.1 Domain-specific setting

Our detailed analyses for the domain-specific setting demonstrate the usefulness of integrating FMs into the sentence-level QE task when TER scores are used as quality labels. The informativeness of FMs in this context is, firstly, shown by the baseline linear regression models. By utilizing the similarity score for the best FM retrieved for a given source sentence as a single feature, simple linear regression models were able to achieve a moderate positive correlation (r) with gold-standard TER scores, for both the DGT and UN data sets.

The usefulness of FMs in the domain-specific QE setting is confirmed by the neural QE models, which were built in the context of transfer learning. For both data sets, the FM-augmented MTQ models (SRC-MT-FMtgt and SRC-MT-FMsrc) outperformed the baseline MTQ models, which rely on SRC-MT input sequences. When we compared the impact of integrating the source (FMsrc) or the target (FMtgt) side of the best FM retrieved for a given source into MTQ models, we observed a clear trend: FMtgt is not only more informative on the sentence-level QE task than FMsrc (when concatenated to SRC or SRC–MT pairs as input) but also than the FMsrc–FMtgt pair (when concatenated to SRC as input). For both data sets, the MTQ systems which utilized the SRC–MT–FMtgt triplets outperformed all other configurations according to all evaluation metrics, achieving the overall best QE performance in the domain-specific setting. This configuration also yielded significant improvements in QE performances compared to the baseline MTQ systems, with +12% and +5% relative improvement in Pearson r for the UN and the WMT data sets respectively.

Even though the best QE performances were achieved by using the SRC–MT–FMtgt input sequence, our analyses also revealed that high QE performance could be achieved by only using SRC–FMtgt input sequence, without utilizing the MT output for the QE task. In fact, for the DGT data set, the MTQ model that utilized SRC–FMtgt input pairs outperformed the baseline MTQ model, which utilized SRC–MT input pairs instead. Considering that state-of-the-art QE models typically rely on SRC–MT pairs, these findings can lead to new research directions in the field of QE, which focus on building QE systems that only rely on the information collected from source sentences and NMT training data. Provided that high FMs can be retrieved from the MT training data, such QE approaches could also be beneficial for real-time applications of QE systems, as they do not require the generation of MT output prior to estimating translation quality.

The impact of integrating FMtgt into the QE architecture became clearer when the QE performances were measured for different FM similarity ranges. In the domain-specific setting, we noticed a positive correlation between QE performance and the similarity score of the best FM that is utilized in the QE task. For both data sets, the FMs above 0.80 (cosine) similarity score brought clear improvements to QE performance, and the most informative FMs on the QE task were observed in the highest (i.e. 0.90-0.99) similarity range. While the FMs in low similarity ranges did not affect the QE performance negatively, they also did not present any additional value to the MTQ models. Based on these results, the hypothesis that FMs in all similarity ranges are informative on the sentence-level QE task is rejected. Interestingly, these results confirm previous findings in the context of NMT, where it has been demonstrated that integrating FMs with higher similarity scores into the NMT architecture led to higher improvements in estimated translation quality (Bulté and Tezcan 2019, He et al. 2021, Tezcan and Bulté 2022). Furthermore, these results also explain to a certain extent why integrating FMtgt into the MTQ architecture yields larger improvements in average QE performance on the DGT data set compared to the UN data set: the ratio of source sentences in the DGT QE test set, for which highly informative FMs (i.e. FMs above the 0.8 similarity score) could be retrieved, was marked to be higher than in the WMT test set (50% vs. 32%).

One of the advantages of utilizing cross-lingual pre-trained models for the QE task in the context of transfer learning is their ability to achieve competitive results in low-resource scenarios (Ranasinghe et al. 2020). The full QE training data sets in the domain-specific setting can be considered large data sets that lead to computationally-expensive QE training processes. In order to see whether similar QE performances could be achieved with smaller training sets, we re-trained the best performing MTQ models (using the SRC–MT-FMtgt triplets as input), with increasingly smaller training sets. The results of these experiments revealed two important facts: for both data sets, (i) near-optimal QE performance was achieved by using 25% of the original QE training data sets, and (ii) the improvements observed in QE performance grew relatively larger, with decreasing training set sizes. In the domain-specific setting, the quality labels on the training data could only be obtained through the jackknifing method, which required training four additional NMT models using different subsets of the training data (see Section 3.1 for the details of the jackknifing methodology used to generate the QE training data set). Achieving near-optimal QE performance by using only 25% of the full training data suggests that the necessary quality labels for training a well-performing QE model can be obtained by training a single NMT model (using the remaining 75% of the NMT training data). As a result, the FM-augmented QE models proposed in this study can be built in a computationally more efficient manner. The larger relative improvements achieved in QE performance when using increasingly smaller training sets become important for comparing the results in domain-specific and general-domain settings, which is further discussed in Section 5.3.

5.2 General-domain setting

The analyses made in the general-domain setting showed that the similarity score of the best FM retrieved for a given source sentence is not informative for predicting the quality of the corresponding NMT output when this information is used as a single feature in a linear regression model. Such linear regression models yielded almost zero correlation for both QE tasks: predicting (i) manually obtained direct assessment (DA) scores, and (ii) HTER scores, which were obtained by comparing the NMT output to its post-edited version. These results were confirmed when the QE performance of the baseline MTQ configuration (SRC–MT) was compared with the best-performing FM-augmented model in the domain-specific setting (SRC–MT–FMtgt) for both QE tasks. While the FM-augmented model outperformed the baseline MTQ model for the task of predicting HTER scores, it achieved a lower QE performance than the baseline MTQ model for the DA task. However, the differences between the QE performances of the two configurations were not measured to be significant. These results also confirm the findings of Wang et al. (2021), who demonstrated that integrating a set of features into a neural QE architecture based on the highest similarity scores measured between source sentences and the NMT training data led to a slight decrease in QE performance when combined with other features. Similarly, when we analyzed the QE performance of both MTQ models for source sentences that are grouped in different FM similarity ranges, no clear patterns emerged.

5.3 Domain-specific vs. general-domain settings

There are two main potential explanations for the apparent discrepancy between the informativeness of FMs in sentence-level QE tasks in the domain-specific and general-domain settings. First, the quality labels in both scenarios were obtained through different methods, which capture different aspects of translation quality and can be considered also one of the main limitations of this study. While in the domain-specific setting the TER scores were obtained by comparing a given NMT output to a gold-standard, reference translation, in the general-domain setting, quality labels were either obtained manually by taking the average direct assessment (DA) scores or automatically by comparing a given NMT output to its post-edited version (HTER). In this context, while the DA scores can capture the severity of errors, this often cannot be measured by TER or HTER scores (Fomicheva et al. 2020a). On the other hand, while HTER scores measure the post-editing effort based on a given MT output, TER scores are measured again one possible correct translation, which can differ greatly from the post-edited MT output, especially in such general-domain scenarios, where the post-editing process is not driven by specialized guidelines or terminology requirements

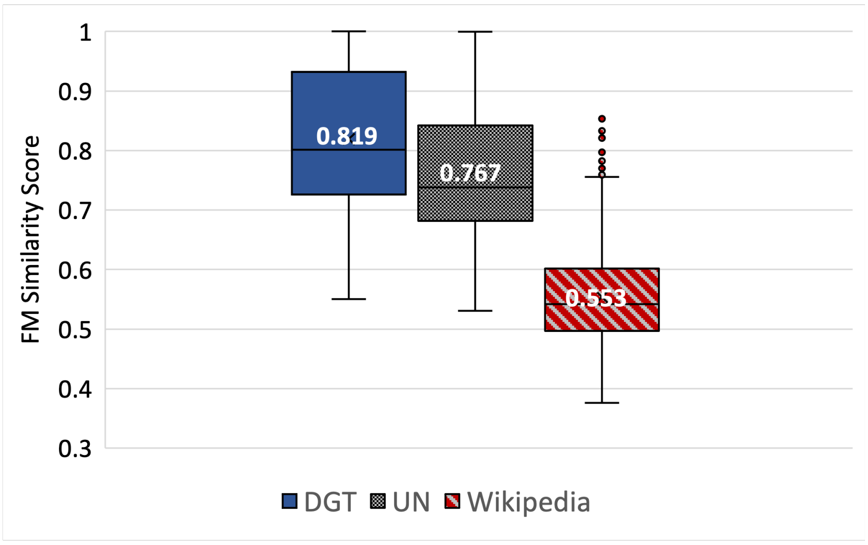


Figure 4: FM similarity score distributions in the DGT, WMT, and Wikipedia QE test sets. The mean FM score in each data set is displayed in white.

(Fomicheva et al. 2020a), unlike the domain-specific scenarios investigated in this study (Steinberger et al. 2012, Ziemski et al. 2016).

When the FM-augmented systems were trained with similar QE training set sizes in both scenarios (full training set in the general-domain setting and 1.25% of the full training set in the domain-specific setting), we could observe a clear difference in QE performances. While in the domain-specific setting (using TER), the FM-augmented QE model resulted in up to 31% relative improvement in Pearson r compared to the baseline MTQ model, in the general-domain scenario (using DA and HTER scores), we did not observe any significant differences in QE performances compared to the baseline MTQ models. Although this observation needs to be confirmed in subsequent studies, the differences between the QE performances in both scenarios can potentially be attributed to the different aspects of translation quality measured by these three different assessment techniques. It should also be noted that, in domain-specific scenarios, it is often not possible to manually assess the quality (DA) or obtain post-edited MT output (to calculate HTER scores) for a large number of sentences (e.g. a full TM). As this study demonstrates, while decent QE performances could be achieved with relatively small QE training sets (e.g. fewer than 14K input sequences), the QE performances were maximized when larger QE training sets were used in the context of transfer learning. As a result, automatic extraction of quality labels through automatic evaluation metrics (such as TER) might be the only practical solution to build FM-augmented QE models utilizing large training data sets.

A second potential reason for the difference in the QE performances of the FM-augmented MTQ models is related to the difference in the FM-similarity levels observed in the domain-specific and in-domain scenarios. In the domain-specific setting, we notice that FM augmentation led to higher QE performance when the best FM retrieved for a given source sentence yielded a minimum cosine similarity score of 0.80. In the general-domain scenario, for each source sentence in the QE test set, even though FMs were retrieved from a very large NMT training data that consists of approximately 23 million sentence pairs, we observed only five sentences for which FMs above the 0.80 similarity score could be retrieved. For an easier comparison between the DGT, WMT, and Wikipedia data sets, Figure 4 presents the box plot distribution of the scores for the retrieved FMs from the corresponding NMT data sets for each source sentence in each test set.

Given that the same quality label was used in the domain-specific setting for both the DGT and UN data sets, we already argued that the ratio of informative FMs in the QE test data sets could explain the differences in the QE performances of the FM-augmented MTQ models (see 5.1). Considering the differences in FM similarity score distributions in the domain-specific and general-domain data sets, this argument can be extended to the general-domain setting. In Figure 4, we do not only see that the Wikipedia QE test set contains only a few sentences with highly informative FMs (above 0.80 similarity level), but the mean (0.553), minimum (0.376), and maximum (0.853) FM similarity scores in this data set are much lower than the DGT and UN data sets. These results suggest that the FM-augmentation approach for sentence-level QE tasks is potentially beneficial and most effective in domain-specific scenarios when the NMT models are trained with data sets that have a certain amount of repetition and consistent writing style.

6. Conclusion

This study proposes a simple data augmentation method for integrating FMs, which are retrieved for a given source sentence from the NMT training data, to predict the sentence-level quality of the corresponding NMT output. The study was carried out for two data sets (DGT-TM and UN corpus) and language pairs (EN–NL and EN–FR, respectively) in a domain-specific setting, and for one data set (WMT 2020 Wikipedia data set) and language pair (EN–DE) in a general-domain setting. The experiments conducted in this study demonstrate that, in certain settings, FMs retrieved from NMT training data can be highly informative for predicting the sentence-level quality of the translations obtained from the corresponding NMT model.

In the domain-specific setting, augmenting the commonly used QE input representations (SRC–MT pairs) with the target side of the best FM retrieved for each source sentence (SRC–MT–FMtgt) led to the best QE performance for both data sets when sentence-level quality labels are automatically extracted by comparing NMT output with reference translations, using TER. Domain-specific experiments also revealed a positive correlation between the similarity scores of the utilized FMs and the QE performance, showing that the FM-augmented models achieved the best QE performances when the retrieved FMs were highly similar to the source sentences (above 0.90 cosine similarity). Moreover, by relying only on the information obtained from the source sentences and the NMT training data (without utilizing the MT output), high QE performance could be achieved.

The results of the QE experiments conducted in the general-domain setting, however, drew a different picture. In this setting, the FMs did not lead to any statistically significant difference in QE performances, for two different sentence-level QE tasks: predicting direct (manual) assessment scores, and post-editing effort, which is measured by comparing NMT output to its post-edited version using HTER. We argue that there are two main reasons for the difference in domain-specific and general-domain settings: the difference in quality labels used in both settings and the lower FM similarity scores observed in the general-domain setting.

There are several lines of research arising from this work that can be pursued. Firstly, in the context of domain-specific settings, we would like to investigate imposing a minimum similarity threshold for FM augmentation and study the usefulness of FM-augmented models for the word-level QE task. In the context of general-domain scenarios, a different line of research is to investigate the usefulness of utilizing sub-segment-level similarities between sentences for the word- and sentence-level QE tasks. Finally, in the future, we would like to repeat these experiments by integrating FMs into different QE architectures.

6.1 Acknowledgments

This research was funded by Research Foundation—Flanders (FWO).

References

Blatz, John, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing (2004), Confidence estimation for machine translation, *Proceedings of the 20th international conference on computational linguistics*, Geneva, Switzerland, pp. 315–321.

Bojar, Ondřej, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia (2013), Findings of the 2013 Workshop on Statistical Machine Translation, *Proceedings of the Eighth Workshop on Statistical Machine Translation*, Association for Computational Linguistics, Sofia, Bulgaria, pp. 1–44. https://aclanthology.org/W13-2201.

Bulté, Bram and Arda Tezcan (2019), Neural fuzzy repair: Integrating fuzzy matches into neural machine translation, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, pp. 1800–1809. https://aclanthology.org/P19-1175.

Cao, Qian and Deyi Xiong (2018), Encoding gated translation memory into neural machine translation, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Brussels, Belgium, pp. 3042–3047. https://www.aclweb.org/anthology/D18-1340.

Chen, Yimeng, Chang Su, Yingtao Zhang, Yuxia Wang, Xiang Geng, Hao Yang, Shimin Tao, Guo Jiaxin, Wang Minghan, Min Zhang, Yujia Liu, and Shujian Huang (2021), HW-TSC's participation at WMT 2021 quality estimation shared task, *Proceedings of the Sixth Conference on Machine Translation*, Association for Computational Linguistics, Online, pp. 890–896. https://aclanthology.org/2021.wmt-1.92.

Conneau, Alexis and Guillaume Lample (2019), *Cross-Lingual Language Model Pretraining*, Curran Associates Inc., Red Hook, NY, USA.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov (2020), Unsupervised cross-lingual representation learning at scale, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, pp. 8440–8451. https://aclanthology.org/2020.acl-main.747.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019), BERT: Pre-training of deep bidirectional transformers for language understanding, *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)*, Association for Computational Linguistics, Minneapolis, Minnesota, USA, pp. 4171–4186. https://aclanthology.org/N19-1423.

Eo, Sugyeong, Chanjun Park, Hyeonseok Moon, Jaehyung Seo, and Heuiseok Lim (2021), Comparative analysis of current approaches to quality estimation for neural machine translation, *Applied Sciences*. https://www.mdpi.com/2076-3417/11/14/6584.

Fomicheva, Marina, Shuo Sun, Erick Fonseca, Chrysoula Zerva, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André FT Martins (2020a), Mlqe-pe: A multilingual quality estimation and post-editing dataset, *arXiv preprint arXiv:2010.04480.*

Fomicheva, Marina, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia (2020b), Unsupervised Quality Estimation for Neural Machine Translation, *Transactions of the Association for Computational Linguistics* **8**, pp. 539–555. https://doi.org/10.1162/tacl_a_00330.

Gu, Jiatao, Yong Wang, Kyunghyun Cho, and Victor O. K. Li (2018), Search engine guided neural machine translation, *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, Association for the Advancement of Artificial Intelligence, New Orleans, Louisiana, USA, pp. 5133–5140.

Guzmán, Francisco, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato (2019), The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, Hong Kong, China, pp. 6098–6111. https://aclanthology.org/D19-1632.

Hardmeier, Christian, Joakim Nivre, and Jörg Tiedemann (2012), Tree kernels for machine translation quality estimation, *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Association for Computational Linguistics, Montréal, Canada, pp. 109–113. https://aclanthology.org/W12-3112.

He, Qiuxiang, Guoping Huang, Qu Cui, Li Li, and Lemao Liu (2021), Fast and accurate neural machine translation with translation memory, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Association for Computational Linguistics, Online, pp. 3170–3180. https://aclanthology.org/2021.acl-long.246.

Johnson, J., M. Douze, and H. Jegou (2021), Billion-scale similarity search with gpus, *IEEE Transactions on Big Data* **7** (03), pp. 535–547, IEEE Computer Society, Los Alamitos, CA, USA.

Kepler, Fabio, Jonay Trénous, Marcos Treviso, Miguel Vera, António Góis, M. Amin Farajian, António V. Lopes, and André F. T. Martins (2019), Unbabel's participation in the WMT19 translation quality estimation shared task, *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, Association for Computational Linguistics, Florence, Italy, pp. 78–84. https://aclanthology.org/W19-5406.

Khandelwal, Urvashi, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis (2020), Nearest neighbor machine translation, *arXiv preprint arXiv:2010.00710.*

Kim, Hyun and Jong-Hyeok Lee (2016), Recurrent neural network based translation quality estimation, *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, Association for Computational Linguistics, Berlin, Germany, pp. 787–792. https://aclanthology.org/W16-2384.

Kim, Hyun, Jong-Hyeok Lee, and Seung-Hoon Na (2017), Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation, *Proceedings of the Second Conference on Machine Translation*, Association for Computational Linguistics, Copenhagen, Denmark, pp. 562–568. https://aclanthology.org/W17-4763.

Klein, Guillaume, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush (2017), OpenNMT: Open-source toolkit for neural machine translation, *Computing Research Repository.* https://arxiv.org/abs/1701.02810.

Kudo, Taku and John Richardson (2018), SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Association for Computational Linguistics, Brussels, Belgium, pp. 66–71. https://aclanthology.org/D18-2012.

Liu, Yinhan, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer (2020), Multilingual denoising pre-training for neural machine translation, *Transactions of the Association for Computational Linguistics* **8**, pp. 726–742, MIT Press, Cambridge, MA. https://aclanthology.org/2020.tacl-1.47.

Meng, Yuxian, Xiaoya Li, Xiayu Zheng, Fei Wu, Xiaofei Sun, Tianwei Zhang, and Jiwei Li (2022), Fast nearest neighbor machine translation, *Findings of the Association for Computational Linguistics: ACL 2022*, Association for Computational Linguistics, Dublin, Ireland, pp. 555–565. https://aclanthology.org/2022.findings-acl.47.

Ott, Myle, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli (2019), fairseq: A fast, extensible toolkit for sequence modeling, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, Association for Computational Linguistics, Minneapolis, Minnesota, USA, pp. 48–53. https://aclanthology.org/N19-4009.

Pagliardini, Matteo, Prakhar Gupta, and Martin Jaggi (2018), Unsupervised learning of sentence embeddings using compositional n-gram features, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Association for Computational Linguistics, New Orleans, Louisiana, USA, pp. 528–540. https://www.aclweb.org/anthology/N18-1049.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002), Bleu: a method for automatic evaluation of machine translation, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pp. 311–318. https://www.aclweb.org/anthology/P02-1040.

Patel, Raj Nath and Sasikumar Mukundan (2016), Translation quality estimation using recurrent neural network, *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, Association for Computational Linguistics, Berlin, Germany, pp. 819–824. https://aclanthology.org/W16-2389.

Pires, Telmo, Eva Schlinger, and Dan Garrette (2019), How multilingual is multilingual BERT?, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, pp. 4996–5001. https://aclanthology.org/P19-1493.

Popović, Maja (2018), *Error Classification and Analysis for Machine Translation Quality Assessment*, Springer International Publishing, Cham, pp. 129–158.

Popović, Maja and Hermann Ney (2011), Towards automatic error analysis of machine translation output, *Computational Linguistics* **37** (4), pp. 657–688, MIT Press, Cambridge, MA. https://aclanthology.org/J11-4002.

Ranasinghe, Tharindu, Constantin Orasan, and Ruslan Mitkov (2020), TransQuest: Translation quality estimation with cross-lingual transformers, *Proceedings of the 28th International Conference on Computational Linguistics*, International Committee on Computational Linguistics, Barcelona, Spain (Online), pp. 5070–5081. https://aclanthology.org/2020.coling-main.445.

Rosti, Antti-Veikko, Necip Fazil Ayan, Bing Xiang, Spyros Matsoukas, Richard Schwartz, and Bonnie Dorr (2007), Combining outputs from multiple machine translation systems, *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, Association for Computational Linguistics, Rochester, New York, USA, pp. 228–235. https://aclanthology.org/N07-1029.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul (2006), A study of translation edit rate with targeted human annotation, *Proceedings of the 2006 Conference of the Association for Machine Translation in the Americas*, AMTA, Cambridge, Massachusetts, USA, pp. 223–231.

Soricut, Radu and Abdessamad Echihabi (2010), TrustRank: Inducing trust in automatic translations via ranking, *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Uppsala, Sweden, pp. 612–621. https://aclanthology.org/P10-1063.

Specia, Lucia and Atefeh Farzindar (2010), Estimating machine translation post-editing effort with HTER, *Proceedings of the Second Joint EM+/CNGL Workshop: Bringing MT to the User: Research on Integrating MT in the Translation Industry*, Association for Machine Translation in the Americas, Denver, Colorado, USA, pp. 33–43. https://aclanthology.org/2010.jec-1.5.

Specia, Lucia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André F. T. Martins (2021), Findings of the WMT 2021 shared task on quality estimation, *Proceedings of the Sixth Conference on Machine Translation*, Association for Computational Linguistics, Online, pp. 684–725. https://aclanthology.org/2021.wmt-1.71.

Specia, Lucia, 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, *Proceedings of the Fifth Conference on Machine Translation*, Association for Computational Linguistics, Online, pp. 743–764. https://aclanthology.org/2020.wmt-1.79.

Specia, Lucia, Frédéric Blain, Varvara Logacheva, Ramón F. Astudillo, and André F. T. Martins (2018), Findings of the WMT 2018 shared task on quality estimation, *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, Association for Computational Linguistics, Belgium, Brussels, pp. 689–709. https://aclanthology.org/W18-6451.

Specia, Lucia, Gustavo Paetzold, and Carolina Scarton (2015), Multi-level translation quality prediction with quest++, *Proceedings of ACL-IJCNLP 2015 system demonstrations*, pp. 115–120.

Specia, Lucia, Kashif Shah, José GC De Souza, and Trevor Cohn (2013), Quest – a translation quality estimation framework, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Sofia, Bulgaria, pp. 79–84.

Specia, Lucia, Marco Turchi, Nicola Cancedda, Nello Cristianini, and Marc Dymetman (2009), Estimating the sentence-level quality of machine translation systems, *Proceedings of the 13th Annual conference of the European Association for Machine Translation*, European Association for Machine Translation, Barcelona, Spain. https://aclanthology.org/2009.eamt-1.5.

Steinberger, Ralf, Andreas Eisele, Szymon Klocek, Spyridon Pilos, and Patrick Schlüter (2012), DGT-TM: A freely available translation memory in 22 languages, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, European Language Resources Association (ELRA), Istanbul, Turkey, pp. 454–459.

Tezcan, Arda (2018), *Informative quality estimation of machine translation output*, PhD thesis, Ghent University.

Tezcan, Arda and Bram Bulté (2022), Evaluating the impact of integrating similar translations into neural machine translation, *Information.* https://www.mdpi.com/2078-2489/13/1/19.

Tezcan, Arda, Bram Bulté, and Bram Vanroy (2021), Towards a better integration of fuzzy matches in neural machine translation through data augmentation, *Informatics.* https://www.mdpi.com/2227-9709/8/1/7.

Ueffing, Nicola and Hermann Ney (2007), Word-level confidence estimation for machine translation, *Computational Linguistics* **33** (1), pp. 9–40. https://aclanthology.org/J07-1003.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017), Attention is all you need, *Advances in neural information processing systems*, pp. 5998–6008.

Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors (2020), SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python, *Nature Methods* **17**, pp. 261–272.

Wang, Jiayi, Ke Wang, Boxing Chen, Yu Zhao, Weihua Luo, and Yuqi Zhang (2021), QEMind: Alibaba's submission to the WMT21 quality estimation shared task, *Proceedings of the Sixth Conference on Machine Translation*, Association for Computational Linguistics, Online, pp. 948–954. https://aclanthology.org/2021.wmt-1.100.

Wong, Billy T. M. and Chunyu Kit (2012), Extending machine translation evaluation metrics with lexical cohesion to document level, *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Association for Computational Linguistics, Jeju Island, Korea, pp. 1060–1068. https://aclanthology.org/D12-1097.

Xu, Jitao, Josep Crego, and Jean Senellart (2020), Boosting neural machine translation with similar translations, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, pp. 1580–1590. https://www.aclweb.org/anthology/2020.acl-main.144.

Zerva, Chrysoula, Daan van Stigt, Ricardo Rei, Ana C Farinha, Pedro Ramos, José G. C. de Souza, Taisiya Glushkova, Miguel Vera, Fabio Kepler, and André F. T. Martins (2021), IST-unbabel 2021 submission for the quality estimation shared task, *Proceedings of the Sixth Conference on Machine Translation*, Association for Computational Linguistics, Online, pp. 961–972. https://aclanthology.org/2021.wmt-1.102.

Zhang, Jingyi, Masao Utiyama, Eiichro Sumita, Graham Neubig, and Satoshi Nakamura (2018), Guiding neural machine translation with retrieved translation pieces, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Association for Computational Linguistics, New Orleans, Louisiana, USA, pp. 1325–1335. https://www.aclweb.org/anthology/N18-1120.

Ziemski, Michał, Marcin Junczys-Dowmunt, and Bruno Pouliquen (2016), The United Nations parallel corpus v1.0, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, European Language Resources Association (ELRA), Portorož, Slovenia, pp. 3530–3534. https://www.aclweb.org/anthology/L16-1561.

Appendix A. Hyper-parameters

A.1 NMT models

Table 9: Hyper-parameters for training NMT models.

Hyper-Parameter	**Value**
source/target embedding dimension	512
size of hidden layers	512
feed-forward layers	2048
number of heads	8
number of layers	6
batch size	64
gradient accumulation	1
dropout	0.1
warm-up steps	8000
optimizer	Adam

A.2 Sent2vec

To train our sent2vec models, we use the same hyper-parameters that are suggested in the description paper (Pagliardini et al. 2018) for a sent2vec model trained on Wikipedia data containing both unigrams and bigrams. In our experiments, we distributed training of a sent2vec model over 40 threads. The hyper-parameters are provided in Table 10.

Table 10: Hyper-parameters for training sent2vec models.

Hyper-Parameter	**Value**
embedding dimension	700
minimum word count	8
minimum target word count	20
initial learning rate	0.2
epochs	9
sub-sampling hyper-parameter	5×10^{-6}
bigrams dropped per sentence	4
number of negatives sampled	10

A.3 FAISS

Because our goal is to find matches over all available sentences in the FAISS index, we create a Flat index with an inner product metric to do a brute-force search. By adding the L2-normalised vectors of the sentence representation to the index, and using an L2-normalised sentence vector as an input query, we are effectively using cosine similarity as match metric. More information can be found here: `https://github.com/facebookresearch/faiss/wiki`.

A.4 TransQuest

Table 11: Hyper-parameters for training TransQuest models.

Hyper-Parameter	**Value**
max. sequence length	200 (300)
training batch size	4
gradient accumulation step	2
evaluation batch size	8
epochs	30
weight decay	0
learning rate	1×10^{-6}
adam epsilon	1×10^{-8}
validation steps	100