Word-level Translation Quality Estimation Based on Optimal Transport

Yuto Kuroda* kuroda@ai.cs.ehime-u.ac.jp
Graduate School of Science and Engineering, Ehime University,
3 Bunkyo-cho, Matsuyama, Ehime, 790-8577, Japan

Atsushi Fujita atsushi.fujita@nict.go.jp
National Institute of Information and Communications Technology,
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan

Tomoyuki Kajiwara kajiwara@cs.ehime-u.ac.jp
Graduate School of Science and Engineering, Ehime University,
3 Bunkyo-cho, Matsuyama, Ehime, 790-8577, Japan

Abstract

Word-level translation quality estimation (TQE) is the task of identifying erroneous words in a translation with respect to the source. State-of-the-art methods for TQE exploit large quantities of synthetic training data generated from bilingual parallel corpora, where pseudo-quality labels are determined by comparing two independent translations for the same source text, i.e., an output from a machine translation (MT) system and a reference translation in the parallel corpora. However, this process is sorely reliant on the surface forms of words, with acceptable synonyms and interchangeable word orderings regarded as erroneous. This can potentially mislead the pre-training of models. In this paper, we describe a method that integrates a degree of uncertainty in labeling the words in synthetic training data for TQE. To estimate the extent to which each word in the MT output is likely to be correct or erroneous with respect to the reference translation, we propose to use the concept of optimal transport (OT), which exploits contextual word embeddings. Empirical experiments using a public benchmarking dataset for word-level TQE demonstrate that pre-training TQE models with the pseudo-quality labels determined by OT produces better predictions of the word-level quality labels determined by manual post-editing than doing so with surface-based pseudo-quality labels.

1 Introduction

Translation quality estimation (TQE) (Blatz et al., 2004; Specia et al., 2018) is the task of predicting quality labels or scores for a given translation, typically generated by machine translation (MT) systems, with respect to the source text, without referring to a reference translations. Predictions can be made at different levels of granularity, such as sentence and word levels. Sentence-level quality labels help users determine whether to use an MT output as it is or after post-editing (PE). Word-level quality labels better guide post-editors in the translation production process (ISO/TC37, 2017), i.e., identifying words that require revision.

In this paper, we focus on word-level TQE. The data for training and evaluating word-level TQE models consist of tuples of a source text, an MT output, and quality labels for each word in the MT output. In the TQE shared tasks at the Workshop on Machine Translation (WMT) (Specia et al., 2020, 2021; Zerva et al., 2022), binary labels, i.e., {"OK," "BAD"}, are used as the quality labels. As illustrated in the top part of Figure 1, TQE data are pro-

*This work was done during an internship of the first author at National Institute of Information and Communications Technology.

Proceedings of the 16th Conference of the Association for Machine Translation in the Americas,
Chicago, USA, September 30 - October 2, 2024. Volume 1: Research Papers

duced through manual PE of MT outputs, where revisions indicate that the words in the MT output are erroneous. It is therefore straightforward to determine gold-standard labels using the Translation Edit Rate (TER) toolkit (Snover et al., 2006)[1] by identifying the minimum edit distance between two sequences of words relying on surface-level matching.

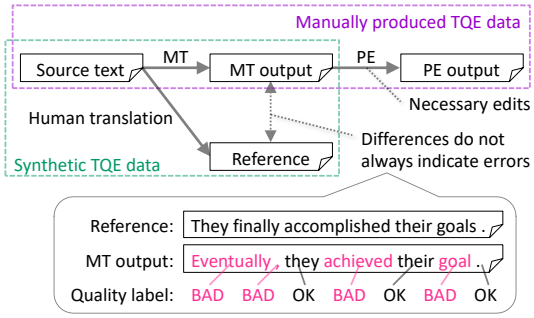


Figure 1: Framework for obtaining TQE data: (a) manual PE and (b) comparison of MT output with a reference translation independently produced by a human translator. An example for the latter exemplifies that the conventional TER-based method regards unessential differences as errors.

To improve the accuracy further, state-of-the-art methods for word-level TQE exploit large quantities of synthetic TQE data for pre-training (Liu et al., 2017; Lee, 2020; Tuan et al., 2021; Rubino et al., 2021; Yang et al., 2023). Figure 1 also shows the typical process of generating synthetic TQE data. First, the source side of a given bilingual parallel corpus is translated with an MT system, and then a pseudo-quality label for each word in the MT output is determined by comparing two independent translations for the same source text: the MT output and the target side of the parallel corpus, i.e., reference translation. Previous work (Liu et al., 2017; Lee, 2020; Tuan et al., 2021; Rubino et al., 2021) has used the TER toolkit for the comparison. However, surface-level differences between independent translations do not necessarily indicate errors. For instance, as shown in Figure 1, they can differ in the use of synonyms, interchangeable word orderings, and so forth, even if the MT output is error-free. Application of the TER toolkit to such pairs inevitably produces incorrect quality labels and consequently misleads the pre-training of TQE models.

In this paper, we describe a method that considers the degree of uncertainty in labeling words in synthetic training data for TQE. To estimate the extent in which each word in the MT output is likely to be correct or erroneous with respect to a reference translation, we propose to use the concept of optimal transport (OT). Given a pair of an MT output and a reference translation, our method first obtains contextual word embeddings. It then determines the optimal alignments between words in the MT output and the reference translation with their likelihood. Following Arase et al. (2023), we expect this approach to identify negligible semantic differences between synonymous expressions and corresponding position-free grammatical elements, and properly label them as "OK." Empirical experiments using a public benchmarking dataset for word-level TQE, i.e., MLQE-PE (Fomicheva et al., 2022), demonstrate that pre-training TQE models with the OT-based quality labels produces better predictions of the word-level quality labels determined by manual PE than models pre-trained on surface-based quality labels determined by TER.

2 Standard Framework

A word-level TQE model is trained on $D_{\mathrm{QE}} = (S_k, T'_k, Y_k)_{k=1}^{N}$, i.e., a set of N triplets of a source text S_k, its machine-translated text T'_k, and a sequence of quality labels Y_k corresponding to the words in T'_k.

2.1 Data for Word-level TQE

Data for word-level TQE can be obtained through post-editing the machine-translated text T'_k into R_k, or annotating errors in T'_k (Freitag et al., 2021). For the former, we can automatically identify the words that have been dropped or revised by comparing T'_k and R_k, typically using the TER toolkit (Snover et al., 2006), and regard them as errors. The post-editing process requires workers who are highly competent in both the source and target languages, and is a laborious task. Therefore, only limited quantities of data are available for a limited number of translation directions and content domains. For instance, the MLQE-PE dataset (Fomicheva et al., 2022) covers only 11 translation directions (see Section 4.1 for details).

[1] http://www.cs.umd.edu/~snover/tercom/

Label type	Problem/Arch.	W_o	$\sigma(\cdot)$	Loss function
Hard	Classification	$\mathbb{R}^{d \times c}$	$\arg\max(softmax(\cdot))$	e.g., Cross-entropy
Soft	Regression	$\mathbb{R}^{d \times 1}$	$sigmoid(\cdot)$	e.g., Mean squared error

Table 1: Architectures and components of word-level TQE models: d indicates the dimension of the contextual word embeddings and c represents the number of possible hard labels ($c = 2$ for {"OK," "BAD"}).

To improve the accuracy, overcoming the data sparseness issue, researchers have exploited synthetic TQE data, which are readily available at a large scale (Liu et al., 2017; Lee, 2020; Tuan et al., 2021; Rubino et al., 2021; Yang et al., 2023). Synthetic TQE data can be generated from a bilingual parallel corpus, $D_{\text{para}} = (S_k, T_k)_{k=1}^{N}$. The parallel corpus can be filtered with some metrics as exemplified in Section 5.1. Typically, T'_k is first generated by translating each source text S_k with an MT model. Alternatively, T'_k can be obtained by rewriting each target text T_k with a masked language model (Tuan et al., 2021) or translating T_k into another language and translating it back into the target language, i.e., round-trip translation (Ding et al., 2021), which can also be applied to monolingual data of the target language. Then, the pseudo-quality label for each word in T'_k is determined by comparing T'_k with the corresponding human translation in the bilingual parallel corpus, i.e., T_k. Most previous work has employed the TER toolkit for this purpose; however, as exemplified in Figure 1, this results in inaccurate pseudo-quality labels, which would mislead the pre-training of TQE models.

2.2 Training Word-level TQE Models

To train a word-level TQE model, large quantities of synthetic data, such as those obtained by the procedure explained in Section 2.1, are used for pre-training (Liu et al., 2017; Lee, 2020; Tuan et al., 2021; Rubino et al., 2021; Yang et al., 2023). In contrast, small quantities of manually produced data are used for fine-tuning the model.

State-of-the-art approaches for word-level TQE rely on a pre-trained multilingual encoder, such as XLM-RoBERTa (Conneau et al., 2020) and INFOXLM (Chi et al., 2021), to obtain contextual embeddings for the words in the source text S and its machine-translated text T'. To exploit cross-lingual relationships between S and T', previous work (Zerva et al., 2021; Rei et al., 2022) jointly encodes the sequences of words in S and T' with a pre-trained multilingual encoder, and obtains $[h_1, \ldots, h_n]$, i.e., d-dimensional contextual embeddings, for the n words in T'. Then, the label for each word t'_i in T' is predicted as follows:

$$\hat{y}_i = \sigma(W_o L(h_i)), \tag{1}$$

where $L(\cdot)$ denotes additional task-specific transformation layers, W_o is a projection matrix, and $\sigma(\cdot)$ is a normalization function. There are two major options for the labels: (a) a hard label, such as {"OK," "BAD"}, or (b) the degree of badness (or goodness). W_o and $\sigma(\cdot)$ are implemented depending on this choice, as summarized in Table 1. Appropriate loss function is also set according to the label type.

3 Determining Pseudo-Quality Labels with Optimal Transport

This paper describes how better pseudo-quality labels can be assigned to the synthetic TQE data. We assume that the triples $D_{\text{syn}} = (S_k, T'_k, T_k)_{k=1}^{N}$ are generated from a bilingual parallel corpus, $D_{\text{para}} = (S_k, T_k)_{k=1}^{N}$, and determine the pseudo-quality label for each word in T'_k by comparing T'_k with the corresponding T_k, as in previous work (Section 2.1).

In the proposed approach, we apply optimal transport (OT), which identifies the optimal way of converting one distribution into another. The application of OT is inspired by its application to monolingual word alignment (Arase et al., 2023). Let $[t'_1, \ldots, t'_n]$ be a sequence of n words in a given machine-translated text T' and $[t_1, \ldots, t_m]$ be a sequence of m words in the corresponding reference translation T. The goal of OT is to identify a matrix $P \in \mathbb{R}_{+}^{n \times m}$ that best aligns the words in T' and T, where $P_{i,j}$ represents the likelihood of the alignment between t'_i and t_j. To solve our problem with OT, we define the following two concepts:

Mass of each word: this is a probability simplex, i.e., $\sum_l = \{v \in \mathbb{R}_{+}^{l} \mid \sum_{i=1}^{l} v_i = 1\}$. We

denote the mass of n words in T' as $a \in \sum_n$ and that of m words in T as $b \in \sum_m$.

Cost for transportation: a cost function for each pair of words, $c(t'_i, t_j) \in \mathbb{R}_+$, can be defined as their dissimilarity. A matrix $C \in \mathbb{R}_+^{n \times m}$, where $C_{i,j} = c(t'_i, t_j)$, represents a summary of the cost for all pairs of words. The cost is typically computed on the basis of contextual word embeddings. In the process of obtaining the embeddings, such as by using a pre-trained multilingual encoder, we can also refer to the source text (see Appendix A), which is an advantage of this method over the TER toolkit.

A matrix $P \in \mathbb{R}_+^{n \times m}$ that minimizes the total cost for transportation is then identified as follows:

$$P = \underset{P' \in U(a,b)}{\arg\min} \sum_{i,j} C_{i,j} P'_{i,j}, \tag{2}$$

where $U(a, b)$ is a set of matrices ($\in \mathbb{R}_+^{n \times m}$) that satisfy a certain constraint. For instance, the following constraint preserves the mass of the source in the target:

$$U(a, b) = \{P \in \mathbb{R}_+^{n \times m} \mid P\mathbb{1}_n = a, P^\top \mathbb{1}_m = b\}, \tag{3}$$

where $\mathbb{1}_l$ is an l-dimensional vector in which all elements are 1. Equation (3) assumes that T' and T can be completely aligned, which conflicts with the motivation of word-level TQE, i.e., the necessity of spotting errors in T'. Therefore, we introduce a constraint that bounds the mass to be transported up to λ_m following the formulation of Partial OT (Figalli, 2010; Caffarelli and McCann, 2010):

$$U(a, b) = \{P \in \mathbb{R}_+^{n \times m} \mid P\mathbb{1}_n \leq a, P^\top \mathbb{1}_m \leq b, \mathbb{1}_n^\top P^\top \mathbb{1}_m = \lambda_m\}. \tag{4}$$

Having obtained the optimal transportation, P, which represents the most plausible alignments between T' and T, we determine the pseudo-quality label for each word t'_i in T'. We consider two variants: soft label ($y_i^{\text{soft}} \in [0, 1]$) and hard label ($y_i^{\text{hard}} \in$ {"OK," "BAD"}).

Soft label is a real number between 0.0 and 1.0, where 0.0 indicates that nothing is transported from the word, strongly suggesting that the word is erroneous, while 1.0 indicates that the word perfectly aligns with a word in T.

$$y_i^{\text{soft}} = \max(P_{i,0}, \ldots, P_{i,m}), \tag{5}$$

$$Y^{\text{soft}} = [y_1^{\text{soft}}, \ldots, y_n^{\text{soft}}]. \tag{6}$$

Hard label is a binary label, {"OK," "BAD"}, which is determined by thresholding the soft label. We introduce this merely for a comparison with the conventional binary labels determined by the TER toolkit.

$$y_i^{\text{hard}} = \begin{cases} \text{"OK"} & y_i^{\text{soft}} > \lambda \\ \text{"BAD"} & \text{otherwise} \end{cases} \tag{7}$$

$$Y^{\text{hard}} = [y_1^{\text{hard}}, \ldots, y_n^{\text{hard}}]. \tag{8}$$

Finally, we obtain two sets of synthetic data for word-level TQE: $D_{\text{QE}}^{\text{soft}} = (S_k, T'_k, Y_k^{\text{soft}})_{k=1}^N$ with the soft labels and $D_{\text{QE}}^{\text{hard}} = (S_k, T'_k, Y_k^{\text{hard}})_{k=1}^N$ with the hard labels.

4 Experiments

To confirm the effectiveness of the proposed method, we conducted experiments using a public dataset for word-level TQE, MLQE-PE (Fomicheva et al., 2022).[2] Following recent shared tasks on word-level TQE (Specia et al., 2020, 2021; Zerva et al., 2022) and Fomicheva et al. (2022), we evaluated TQE models using the Matthews correlation coefficient (MCC) (Matthews, 1975).

4.1 Word-level TQE Dataset

MLQE-PE (Fomicheva et al., 2022) contains test sets for 11 translation directions, each consisting of 1k triplets of source text, an MT output for it, and binary quality labels, i.e., {"OK", "BAD"}, determined by manual PE for the MT output and comparing the result with the raw MT output using the TER toolkit. We used Test20 (`data/post-editing/test`) and Test21 (`data/test21*`) in this repository. For seven[3] translation directions, the MT outputs have been generated by a unidirectional Transformer model (Vaswani et al., 2017) trained

[2] https://github.com/sheffieldnlp/mlqe-pe

[3] English-to-German (En→De), English-to-Chinese (En→Zh), Romanian-to-English (Ro→En), Estonian-to-English (Et→En), Nepali-to-English (Ne→En), Sinhalese-to-English (Si→En), and Russian-to-English (Ru→En).

with the fairseq toolkit (Ott et al., 2019);[4,5] training and development data consisting of 7k and 1k triplets, respectively, are also available (`data/post-editing/{train,dev}`). We used the training data for fine-tuning the TQE models and the development data for selecting the hyper-parameters and models, except for Ru→En. We regarded the remaining four translation directions,[6] for which the MT outputs have been generated by mBART50 (Tang et al., 2021), and Ru→En as zero-shot, since we used neither bilingual parallel data nor TQE data for them.[7]

4.2 Synthetic TQE Data

To generate synthetic TQE data, we used the bilingual parallel corpora[8] officially provided by the organizers of WMT21 TQE Task 2 and M2M-100 (Fan et al., 2021).[9] Table 2 summarizes their sizes and our groupings.

Group	Language pair	Bilingual	Synthetic
High	En–De	23,360,441	22,701,552
	En–Zh	20,305,268	16,201,271
Medium	Ro–En	3,901,501	3,027,243
	Et–En	877,769	855,680
Low	Ne–En	498,271	166,893
	Si–En	646,766	570,770

Table 2: Numbers of sentence pairs in the bilingual parallel corpora and the synthetic TQE data.

Before generating machine-translated texts, we fine-tuned M2M-100 for each translation direction on a sample from the bilingual parallel corpora: 1M, 200k, and 50k sentence pairs for the high-, medium-, and low-resource language pairs, respectively; in each pair, both source and target sides were composed of up to 128 sub-word tokens. Fine-tuning of M2M-100 on the sample was carried out with HuggingFace Transformers (Wolf et al., 2020), the AdamW optimizer (Loshchilov and Hutter, 2019) ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$), batches consisting of 16 sentence pairs, and a learning rate of 3×10^{-5}. This process was terminated after one epoch for each of the high- and medium-resource language pairs and after three epochs for each of the low-resource language pairs. After deduplication, we then decoded the source side of the entire parallel corpora, using beam search with a beam size of 5 and length penalty of 1.0. After decoding, we discarded MT outputs containing more than 128 sub-word tokens together with their corresponding parallel sentences. The numbers of retained sentence pairs are listed in the "Synthetic" column in Table 2.

To determine the pseudo-quality labels, we first obtained the word embeddings using INFOXLM$_{\text{Base}}$ (Chi et al., 2021),[10] inputting a concatenation of MT output T', source text S, and reference translation T in this order with an [SEP] token as the delimiter, and determining the embedding for each word by average pooling of its sub-word embeddings.[11] Then, we determined word alignment by solving OT using OTAlign (Arase et al., 2023);[12] more specifically, we used the entropy-regularized OT (Cuturi, 2013) formulated by Equation (9), which is superior to Equation (2) (Arase et al., 2023).

$$P = \underset{P' \in U(a,b)}{\arg\min} \sum_{i,j} C_{i,j} P'_{i,j} - \xi H(P'), \quad (9)$$

where $H(\cdot)$ is the entropy of a candidate matrix, and ξ is a weight for the regularizer, which we set to 0.1. We used a uniform distribution as the mass for each word, i.e., a and b, and took the cosine distance between contextual word embeddings[13] as the cost function, i.e., $C_{i,j}$. In contrast, we optimized the two hyper-parameters of OT for each translation direction through a grid search for λ_m in the range $[0.02, 1.00]$ with a step size of 0.02 and λ in the range $[0.01, 0.99]$ with a step size of 0.01, using the MLQE-PE development data and computing the MCC between the OT-based hard labels and the

[4] https://github.com/pytorch/fairseq
[5] https://github.com/facebookresearch/mlqe/tree/main/nmt_models
[6] English-to-Czech (En→Cs), English-to-Japanese (En→Ja), Khmer-to-English (Km→En), and Pashto-to-English (Ps→En).
[7] Bilingual parallel data for these language pairs could have been used for pre-training the MT models and multilingual encoders.
[8] https://www.statmt.org/wmt21/quality-estimation-task.html
[9] https://huggingface.co/facebook/m2m100_418M
[10] https://huggingface.co/microsoft/infoxlm-base
[11] Some decisions were made through a preliminary experiment. See Appendix A for details.
[12] https://github.com/yukiar/OTAlign
[13] $1 - \cos(h'_i, h_j)$, which has the range $[0.0, 2.0]$, where h'_i and h_j are word embeddings of t'_i and t_j, respectively.

gold-standard labels. Table 3 presents the values for λ_m and λ that achieved the highest MCC.

Translation direction	λ_m	λ	MCC	
			Dev	Syn
En→De	0.02	0.37	0.870	0.805
En→Zh	0.24	0.51	0.833	0.698
Ro→En	0.14	0.33	0.876	0.819
Et→En	0.02	0.35	0.804	0.776
Ne→En	0.14	0.37	0.680	0.777
Si→En	0.02	0.36	0.699	0.849

Table 3: Hyper-parameters that maximize MCC for the MLQE-PE development data (Dev), and MCC between OT-based and TER-based hard labels for the synthetic TQE data (Syn).

Finally, we determined the pseudo-quality labels with the optimal λ_m and λ as explained in Section 3. Table 3 also lists the MCCs between OT-based hard labels in our synthetic TQE data, derived with the optimized hyper-parameters, and TER-based pseudo-quality labels (Section 4.4).

4.3 TQE Model Training

We trained TQE models using OpenKiWi (Kepler et al., 2019) with the necessary modifications for training regression models and using multiple GPUs. As the backbone pre-trained multilingual encoder, we used INFOXLM$_{\text{Large}}$.[14] For each configuration, we trained a single multi-directional model to deal with all of the test sets, using the training data for six translation directions together: the synthetic TQE data for pre-training (Section 4.2) and manually labeled MLQE-PE training data (Section 4.1) for fine-tuning.

Pre-training was carried out for one epoch with the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$), batches consisting of $2,048$ sentence pairs, and a learning rate of 1×10^{-5}. To accommodate the imbalanced distribution of labels, we weighted the "BAD" labels as 3.0 times the "OK" labels when computing the cross-entropy loss for the classification models. When evaluating the regression models, we computed the MCC by thresholding the predicted value at 0.5.

We then fine-tuned the models on the MLQE-PE training data. When fine-tuning a regression model on the manually produced data with the TER-based hard labels, i.e., the MLQE-PE training data, the "BAD" and "OK" labels were casted as 0.0 and 1.0, respectively. We used Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$), batches consisting of 64 sentence pairs, and a learning rate of 1×10^{-5}. During ten epochs, the model was saved every after 0.5 epochs, and the model that maximized the MCC for the MLQE-PE development data was selected from the 20 checkpoints. For the regression models, we also performed a grid search for the threshold in the range $[0.1, 0.9]$ with a step size of 0.1, using the MLQE-PE development data, and used the results to convert the predictions into binary labels.

4.4 Baseline Methods

We compared our method against the models trained on the synthetic data with pseudo-quality labels determined by the TER toolkit as in MLQE-PE.[15] To re-confirm the impact of pre-training on synthetic TQE data, we also trained classification and regression models only on the MLQE-PE training data.

4.5 Main Results

For each model, we report on the average MCC over three training runs with different random seeds. To confirm the statistical significance of the difference between two sets of predictions, we used paired bootstrap resampling (Koehn, 2004) with $30,000$ sub-samples ($10,000$ for each random seed) and a significance level of 0.05.

Tables 4 and 5 summarize the MCCs for the non-zero-shot translation directions in Test20 and Test21, respectively, where models #1 and #6 based on TER-based pseudo-quality labels and model #4 based only on manually created training data are the baselines. The upper block presents the results in the pseudo-supervised setting, i.e., models trained only on the synthetic TQE data. The model trained on OT-based soft labels (#3) outperformed those trained on either TER-based (#1) or OT-based hard labels (#2). The lower block shows the results of fine-tuned models, i.e., those directly trained or fine-tuned on the MLQE-PE training data. In this setting, the model pre-trained on OT-based soft la-

[14] https://huggingface.co/microsoft/infoxlm-large
[15] https://github.com/deep-spin/qe-corpus-builder

ID	Arch.	PT	FT	Test20					
				En→De	En→Zh	Ro→En	Et→En	Ne→En	Si→En
#1	Class.	TER-Hard	—	0.196	0.163	0.240	0.307	0.320	0.372
#2	Class.	OT-Hard	—	0.208	0.154	0.252	0.312	0.344	0.374
#3	Reg.	OT-Soft	—	**0.258**$^{(1)}$	**0.196**$^{(1)}$	**0.301**$^{(1)}$	**0.358**$^{(1)}$	**0.356**	**0.413**$^{(1)}$
#4	Class.	—	TER-Hard	0.449	0.380	0.623	0.552	0.511	0.552
#5	Reg.	—	TER-Hard	0.439	0.373	**0.650**$^{(4)}$	0.537	0.510	0.550
#6	Class.	TER-Hard	TER-Hard	0.485$^{(4)}$	0.398	0.620	**0.577**$^{(4)}$	0.521	**0.571**
#7	Class.	OT-Hard	TER-Hard	0.486$^{(4)}$	0.397	0.615	0.571$^{(4)}$	0.518	0.564
#8	Reg.	OT-Soft	TER-Hard	**0.491**$^{(4)}$	**0.409**$^{(4)}$	0.634	0.569	**0.530**$^{(4)}$	**0.571**

Table 4: MCCs for the non-zero-shot translation directions in Test20: "Arch." indicates the model architecture while "PT" and "FT" denote the type of labels used for pre-training and fine-tuning, respectively. **Bold** signifies the highest value in each block and translation direction. Values with superscripts (~~deleted~~) are statistically significantly higher (lower) than that for the system with the indicated IDs.

ID	Arch.	PT	FT	Test21					
				En→De	En→Zh	Ro→En	Et→En	Ne→En	Si→En
#1	Class.	TER-Hard	—	0.217	0.129	0.248	0.302	0.322	0.348
#2	Class.	OT-Hard	—	0.236$^{(1)}$	0.117	0.258	0.316	0.346	0.355
#3	Reg.	OT-Soft	—	**0.282**$^{(1)}$	**0.151**$^{(1)}$	**0.304**$^{(1)}$	**0.362**$^{(1)}$	**0.366**$^{(1)}$	**0.406**$^{(1)}$
#4	Class.	—	TER-Hard	0.434	0.320	0.636	0.580	0.540	0.558
#5	Reg.	—	TER-Hard	0.406	0.316	**0.657**$^{(4)}$	0.570	0.537	0.554
#6	Class.	TER-Hard	TER-Hard	**0.496**$^{(4)}$	0.329	0.626	**0.605**$^{(4)}$	0.551	**0.586**$^{(4)}$
#7	Class.	OT-Hard	TER-Hard	0.486$^{(4)}$	0.322	0.629	0.591$^{(6)}$	0.544	0.571
#8	Reg.	OT-Soft	TER-Hard	0.485$^{(4)}$	**0.332**	0.643$^{(6)}$	0.596	**0.555**	0.582$^{(4)}$

Table 5: MCCs for the non-zero-shot translation directions in Test21.

bels (#8) achieved a higher MCC than the TER-based baseline (#6) for seven out of the 12 test sets. As in previous work (Liu et al., 2017; Lee, 2020; Tuan et al., 2021; Yang et al., 2023), pre-training on the synthetic TQE data brought a consistent improvement over the baseline (#4). However, only for Ro→En, the regression model with the same supervised signals (#5) significantly outperformed the classification-based baseline (#4) and even surpassed all models with pre-training. This suggests some peculiar characteristics of the MLQE-PE training data for this translation direction.

The MCCs for the zero-shot translation directions in Test20 and Test21 are presented in Table 6. There were similar trends as for the non-zero-shot translation directions. The synthetic TQE data with OT-based soft labels (#3) gave the best results in the pseudo-supervised setting. For the settings with fine-tuning, the MCCs for all translation directions benefited from supervised signals for other translation directions. They were further improved by pre-training, especially with OT-based soft labels (#8)

5 Analyses

We investigated the quality of the synthetic TQE data and the potential utility of OT-based labels for manually post-edited data. We used the non-zero-shot translation directions of Test20 because the post-edited texts for the MT outputs are available, enabling contrastive experiments.

5.1 Impact of Quality of Synthetic TQE Data

As mentioned in Section 2.1, bilingual parallel corpora used as the source of synthetic TQE data may include sentence pairs that are less likely to be translations. Pseudo-quality labels derived from seman-

ID	Arch.	PT	FT	Test20	Test21				
				Ru→En	En→Cs	En→Ja	Km→En	Ps→En	Ru→En
#1	Class.	TER-Hard	—	0.132	0.224	0.086	0.177	0.234	0.171
#2	Class.	OT-Hard	—	0.147	0.238	0.101$^{(1)}$	0.201$^{(1)}$	0.233	0.172
#3	Reg.	OT-Soft	—	**0.156**$^{(1)}$	**0.265**$^{(1)}$	**0.131**$^{(1)}$	**0.311**$^{(1)}$	**0.263**$^{(1)}$	**0.173**
#4	Class.	—	TER-Hard	0.280	0.326	0.148	0.444	0.348	0.313
#5	Reg.	—	TER-Hard	0.286	0.301	0.154	0.451	0.362	0.308
#6	Class.	TER-Hard	TER-Hard	0.282	0.379$^{(4)}$	0.170	0.469$^{(4)}$	0.368	**0.340**
#7	Class.	OT-Hard	TER-Hard	**0.289**	**0.381**$^{(4)}$	0.169	0.473$^{(4)}$	0.374$^{(4)}$	0.332
#8	Reg.	OT-Soft	TER-Hard	0.287	0.374$^{(4)}$	**0.190**$^{(4)}$	**0.480**$^{(4)}$	**0.381**$^{(4)}$	0.334

Table 6: MCCs for the zero-shot translation directions in Test20 and Test21.

tically isolated pairs of machine-translated text and reference translation could mislead the pre-training of models. To gauge the impact of the quality of parallel data, as well as the quality of synthetic TQE data, we conducted a corpus filtering experiment.

For each pair of sentences in the given bilingual parallel corpora, we computed the cosine similarity between their corresponding sentence embeddings determined by LaBSE (Feng et al., 2022),[16] and then filtered out pairs for which the similarity was lower than a pre-determined threshold. Figure 2 depicts the percentages of retained sentence pairs, depending on the threshold. We found that the Ro–En parallel corpus contained lots of noise, with approximately 40% of sentence pairs having a similarity lower than 0.5.

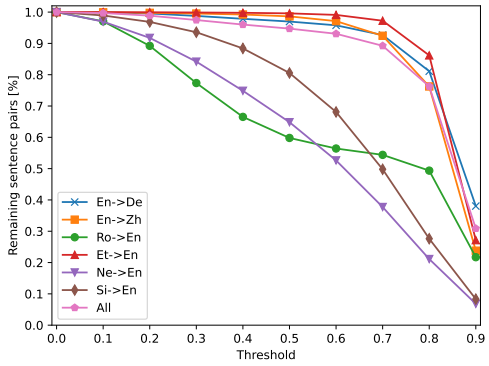


Figure 2: Percentages of remaining sentence pairs after the LaBSE-based filtering.

From the filtered bilingual parallel data, we generated synthetic TQE data and trained TQE models as described in Section 4.3. Finally, we evaluated the model accuracy in terms of MCC, using the non-zero-shot translation directions in Test20.

Table 7 presents the results. In the pseudo-supervised setting, a more aggressive filtering of the parallel corpus produced a higher MCC, suggesting that the quality of synthetic TQE data matters. Among the six translation directions, Ro→En benefited the most; this is to be expected from the statistics shown in Figure 2. In contrast, when fine-tuning was carried out after pre-training, the impact of pre-training, i.e., the gain over the directly supervised model (#5), was often diminished. This implies that the quantity of synthetic TQE data matters when the quality can be guaranteed by fine-tuning on manually produced training data. Besides a slight improvement with corpus filtering, pre-training still had a negative impact on Ro→En, i.e., models #8a and #8b underperformed model #5. In-depth analyses of the MLQE-PE training data of this translation direction is left for future work.

5.2 Fine-tuning on OT-based Labels

Figure 1 illustrated our motivation for obtaining pseudo-labels of better quality, especially for synthetic TQE data. In this section, we examine whether OT also brings some advantages for the authentic data derived through manual PE.

To this end, we first determined the quality labels for the MLQE-PE training data in the same manner as for the synthetic TQE data (Section 4.2). We then fine-tuned the pre-trained models (#2 and

[16] https://huggingface.co/sentence-transformers/LaBSE

ID	Arch.	Synthetic Data			FT	Test20					
		Label	Th	Size		En→De	En→Zh	Ro→En	Et→En	Ne→En	Si→En
#3	Reg.	OT-Soft	—	43.5M	—	0.258	0.196	0.301	0.358	0.356	0.413
#3a	Reg.	OT-Soft	0.5	41.2M	—	0.298$^{(3)}$	0.199	0.426$^{(3)}$	0.351	0.379$^{(3)}$	0.414
#3b	Reg.	OT-Soft	0.7	38.8M	—	0.305$^{(3)}$	0.205	0.448$^{(3)}$	0.363	**0.384**$^{(3)}$	**0.425**
#3c	Reg.	OT-Soft	0.9	13.4M	—	**0.316**$^{(3)}$	**0.241**$^{(3)}$	**0.489**$^{(3)}$	**0.392**$^{(3)}$	0.368	0.410
#8	Reg.	OT-Soft	—	43.5M	TER-Hard	**0.491**	0.409	0.634	0.569	**0.530**	**0.571**
#8a	Reg.	OT-Soft	0.5	41.2M	TER-Hard	0.488	0.407	**0.641**	0.571	0.527	**0.571**
#8b	Reg.	OT-Soft	0.7	38.8M	TER-Hard	0.489	**0.410**	**0.641**	**0.573**	0.527	0.568
#8c	Reg.	OT-Soft	0.9	13.4M	TER-Hard	0.484	0.401	0.637	0.565	0.519	0.555
#5	Reg.	—-	—	—	TER-Hard	0.439	0.373	0.650	0.537	0.510	0.550

Table 7: MCCs for the non-zero-shot translation directions in Test20 with several threshold values ("Th") for the similarity of parallel sentences: "Size" denotes the number of sentence pairs having a similarity higher than or equal to the threshold.

ID	Arch.	PT	FT	Test20					
				En→De	En→Zh	Ro→En	Et→En	Ne→En	Si→En
#4	Class.	—	TER-Hard	**0.449**	**0.380**	0.623	**0.552**	**0.511**	**0.552**
#5	Reg.	—	TER-Hard	0.439	0.373	**0.650**$^{(4)}$	0.537	0.510	0.550
#4'	Class.	—	OT-Hard	0.431	0.334$^{(4)}$	0.609	0.514$^{(4)}$	0.462$^{(4)}$	0.506$^{(4)}$
#5'	Reg.	—	OT-Soft	0.413$^{(4)}$	0.326$^{(4)}$	0.626	0.484$^{(4)}$	0.443$^{(4)}$	0.482$^{(4)}$
#6	Class.	TER-Hard	TER-Hard	0.485$^{(4)}$	0.398	0.620	**0.577**$^{(4)}$	0.521	**0.571**
#7	Class.	OT Hard	TER-Hard	0.486$^{(4)}$	0.397	0.615	0.571$^{(4)}$	0.518	0.564
#8	Reg.	OT-Soft	TER-Hard	**0.491**$^{(4)}$	**0.409**$^{(4)}$	**0.634**	0.569	**0.530**$^{(4)}$	**0.571**
#7'	Class.	OT-Hard	OT-Hard	0.464	0.350$^{(4,6)}$	0.589$^{(4,6)}$	0.523$^{(4,6)}$	0.467$^{(4,6)}$	0.508$^{(4,6)}$
#8'	Reg.	OT-Soft	OT-Soft	0.444$^{(6)}$	0.344$^{(4,6)}$	0.633	0.503$^{(4,6)}$	0.453$^{(4,6)}$	0.491$^{(4,6)}$

Table 8: MCCs with TER-based hard labels for the non-zero-shot translation directions in Test20.

#3) using these labels, as described in Section 4.3, and directly trained the models on them, as described in Section 4.4.

Table 8 presents the results for Test20. Irrespective of whether the pre-training was carried out, the models trained or fine-tuned on the OT-based pseudo-quality labels (#4' to #8') resulted in lower MCCs than the corresponding models trained on TER-based hard labels (#4 to #8). We consider this result to be natural because the gold-standard labels have been determined by the TER toolkit.

5.3 Predicting OT-based Labels

We also evaluated the predicted results with respect to the OT-based labels for Test20, with the labels determined by OT in the same manner as for the synthetic TQE data (Section 4.2).

The MCCs with OT-based hard labels are summarized in Table 9. Compared with those in Table 8, the MCCs of the TER-based models (#4 to #8) were lower, except for the pseudo-supervised models (#4 and #5) for Et→En, while the MCCs of the OT-based models (#4' to #8') were higher. For all translation directions, except for En→De, the models trained or fine-tuned on OT-based labels scored significantly higher MCCs than those based on TER-based labels. This also revealed that pre-training has little gain for all translation directions, implying that the distributions of OT-based labels for the synthetic TQE data and PE-derived data (see Figure 1) are similar.

We also evaluated the accuracy of the regression models against OT-based soft labels with Pearson's product-moment correlation coefficient (Pear-

ID	Arch.	PT	FT	Test20					
				En→De	En→Zh	Ro→En	Et→En	Ne→En	Si→En
#4	Class.	—	TER-Hard	0.437	0.344	0.589	0.555	0.497	0.521
#5	Reg.	—	TER-Hard	0.431	0.334	0.622$^{(4)}$	0.545	0.503	0.519
#4'	Class.	—	OT-Hard	**0.446**	**0.384**$^{(4)}$	0.647$^{(4)}$	**0.600**$^{(4)}$	**0.599**$^{(4)}$	**0.637**$^{(4)}$
#5'	Reg.	—	OT-Soft	0.430	0.376$^{(4)}$	**0.672**$^{(4)}$	0.577$^{(4)}$	0.587$^{(4)}$	0.623$^{(4)}$
#6	Class.	TER-Hard	TER-Hard	0.454	0.356	0.575	0.562	0.493	0.518
#7	Class.	OT-Hard	TER-Hard	0.472$^{(4)}$	0.367	0.585	0.564	0.501	0.526
#8	Reg.	OT-Soft	TER-Hard	**0.488**$^{(4,6)}$	0.381$^{(4,6)}$	0.603$^{(6)}$	0.565	0.516	0.533
#7'	Class.	OT-Hard	OT-Hard	0.483$^{(4,6)}$	**0.406**$^{(4,6)}$	0.629$^{(4,6)}$	**0.610**$^{(4,6)}$	**0.608**$^{(4,6)}$	**0.643**$^{(4,6)}$
#8'	Reg.	OT-Soft	OT-Soft	0.468	0.401$^{(4,6)}$	**0.679**$^{(4,6)}$	0.594$^{(4,6)}$	0.601$^{(4,6)}$	0.631$^{(4,6)}$

Table 9: MCCs with OT-based hard labels for the non-zero-shot translation directions in Test20.

ID	Arch.	PT	FT	Test20					
				En→De	En→Zh	Ro→En	Et→En	Ne→En	Si→En
#5	Reg.	—	TER-Hard	0.505	0.389	0.697	0.622	0.625	0.651
#5'	Reg.	—	OT-Soft	**0.581**$^{(5)}$	**0.486**$^{(5)}$	**0.773**$^{(5)}$	**0.703**$^{(5)}$	**0.714**$^{(5)}$	**0.751**$^{(5)}$
#8	Reg.	OT-Soft	TER-Hard	0.558	0.444	0.675	0.653	0.647	0.666
#8'	Reg.	OT-Soft	OT-Soft	**0.637**$^{(8)}$	**0.540**$^{(8)}$	**0.779**$^{(8)}$	**0.734**$^{(8)}$	**0.740**$^{(8)}$	**0.766**$^{(8)}$

Table 10: Pearson's r with OT-based soft labels for the non-zero-shot translation directions in Test20.

son's r), performing a statistical significance testing in the same manner as for the MCCs. Table 10 demonstrates that training or fine-tuning on OT-based labels leads to higher correlation. Unlike the results for predicting hard labels, pre-training consistently improved the correlation, irrespective of the types of labels used for fine-tuning, with the exception of "TER-Hard" for Ro→En.

These results confirm that the labels for fine-tuning should be consistent with those to be predicted, as discussed by Yang et al. (2023).

6 Conclusion

This paper has described the application of optimal transport (OT) to determine pseudo-quality labels in synthetic data for word-level TQE. Through experiments, we confirmed that OT-based labels better guide pre-training on large quantities of synthetic TQE data and result in higher accuracy in word-level TQE tasks, as measured by MCC. Our method achieved consistently better results for pseudo-supervised settings and in zero-shot translation directions, encouraging future applications to less-studied translation directions.

In future work, we plan to investigate better and finer-grained specifications of the hyper-parameters for OT. While we determined a single value of λ_m, the upper bound of the mass to be transported, for each translation direction, we consider it should be possible to approximate this value for each sentence pair. We have only evaluated our method for predicting target labels; doing so for source labels is another avenue for extension (Appendix B).

Acknowledgments

We would like to thank the anonymous reviewers, including those for past submissions, for their insightful comments and suggestions on earlier versions of this paper. This work was partly supported by JSPS KAKENHI Grant-in-Aid for Scientific Research (S) 19H05660 and a commissioned research (No. 22501) by National Institute of Information and Communications Technology (NICT), Japan.

References

Arase, Y., Bao, H., and Yokoi, S. (2023). Unbalanced Optimal Transport for Unbalanced Word Alignment. In

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3966–3986.

Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S., Goutte, C., Kulesza, A., Sanchis, A., and Ueffing, N. (2004). Confidence Estimation for Machine Translation. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 315–321.

Caffarelli, L. A. and McCann, R. J. (2010). Free Boundaries in Optimal Transport and Monge-Ampère Obstacle Problems. *Annals of Mathematics*, 171(2):673–730.

Chi, Z., Dong, L., Wei, F., Yang, N., Singhal, S., Wang, W., Song, X., Mao, X.-L., Huang, H., and Zhou, M. (2021). InfoXLM: An Information-Theoretic Framework for Cross-Lingual Language Model Pre-Training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588.

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.

Cuturi, M. (2013). Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *Proceedings of the 26th Conference on Neural Information Processing Systems*, pages 2292–2300.

Ding, S., Junczys-Dowmunt, M., Post, M., and Koehn, P. (2021). Levenshtein training for word-level quality estimation. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6724–6733.

Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., Chaudhary, V., Goyal, N., Birch, T., Liptchinsky, V., Edunov, S., Auli, M., and Joulin, A. (2021). Beyond English-Centric Multilingual Machine Translation. *Journal of Machine Learning Research*, 22(107):1–48.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., and Wang, W. (2022). Language-agnostic BERT Sentence Embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 878–891.

Figalli, A. (2010). The Optimal Partial Transport Problem. *Archive for Rational Mechanics and Analysis*, 195:533–560.

Fomicheva, M., Sun, S., Fonseca, E., Zerva, C., Blain, F., Chaudhary, V., Guzmán, F., Lopatina, N., Specia, L., and Martins, A. F. T. (2022). MLQE-PE: A Multilingual Quality Estimation and Post-Editing Dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4963–4974.

Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., and Macherey, W. (2021). Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Goyal, N., Du, J., Ott, M., Anantharaman, G., and Conneau, A. (2021). Larger-Scale Transformers for Multilingual Masked Language Modeling. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 29–33.

ISO/TC37 (2017). ISO 18587:2017 Translation Services: Post-editing of Machine Translation Output: Requirements.

ISO/TC37 (2024). ISO 5060:2024 Translation Services: Evaluation of Translation Output: General Guidance.

Kepler, F., Trénous, J., Treviso, M., Vera, M., and Martins, A. F. T. (2019). 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.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain.

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.

Liu, L., Fujita, A., Utiyama, M., Finch, A., and Sumita, E. (2017). Translation Quality Estimation Using Only Bilingual Corpora. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 25(9):1762–1772.

Loshchilov, I. and Hutter, F. (2019). Decoupled Weight Decay Regularization. In *Proceedings of the 7th International Conference on Learning Representations*.

Matthews, B. (1975). Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442–451.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

Rei, R., Treviso, M., Guerreiro, N. M., Zerva, C., Farinha, A. C., Maroti, C., C. de Souza, J. G., Glushkova, T., Alves, D., Coheur, L., Lavie, A., and Martins, A. F. T. (2022). CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 634–645.

Rubino, R., Fujita, A., and Marie, B. (2021). Error Identification for Machine Translation with Metric Embedding and Attention. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 146–156.

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.

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.

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.

Specia, L., Scarton, C., and Paetzold, G. H. (2018). Quality Estimation for Machine Translation. *Synthesis Lectures on Human Language Technologies*, 11(1):1–162.

Tang, Y., Tran, C., Li, X., Chen, P.-J., Goyal, N., Chaudhary, V., Gu, J., and Fan, A. (2021). Multilingual Translation from Denoising Pre-Training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466.

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.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, u., and Polosukhin, I. (2017). Attention is All You Need. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, pages 5998–6008.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Yang, Z., Meng, F., Yan, Y., and Zhou, J. (2023). Rethinking the Word-level Quality Estimation for Machine Translation from Human Judgement. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2012–2025.

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*, pages 69–99.

Zerva, C., van Stigt, D., Rei, R., Farinha, A. C., Ramos, P., C. de Souza, J. G., Glushkova, T., Vera, M., Kepler, F., and Martins, A. F. T. (2021). IST-Unbabel 2021 Submission for the Quality Estimation Shared Task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 961–972.

A Preliminary Investigation

In our preliminary experiment, we first selected a pre-trained multilingual encoder to obtain contextual word embeddings for OT, using the MLQE-PE development data and varying λ_m and λ in the same manner as described in Section 4.2. We also compared two encoding patterns. Table 11 summarizes the MCCs between OT-based hard labels and TER-based labels for the MLQE-PE development data. Among the four candidate encoders, INFOXLM$_{\text{Base}}$ achieved consistently high MCCs across all translation directions. Interestingly, "Large" models consistently underperformed their "Base" counterpart. We also confirmed that referring to the source text leads to higher MCCs in general.

Then, we investigated the ordering of the source text S, its MT output T', and its post-edited version R as the input for INFOXLM$_{\text{Base}}$, even though R must be replaced with an independently produced human reference when generating synthetic TQE data. Table 12 presents the results. Among the six permutations of these three elements, (S, T', R) resulted in the highest MCC in average, but other permutations also achieved comparable MCCs. Assuming that focusing on T' would be effective for synthetic data, we used (T', S, R) in our experiment. This experiment reconfirmed the usefulness of the source text S and revealed that determining pseudo-labels using only the source text and MT output, i.e., (S, T'), is infeasible.

Figure 3 visualizes the sensitivity of the two hyper-parameters λ_m and λ with INFOXLM$_{\text{Base}}$ and the (T', S, R) layout for its input.

Figure 4 depicts that the soft labels determined by OT are highly correlated with the TER-based binary labels. Nevertheless, we consider the continuity of the labels and some discrepancies to improve the prediction; discrepancies include high values with "BAD" label, such as those illustrated in Figure 1, and potentially low values with "OK" label for identical but unrelated word correspondences, such as articles for different nominal elements.

B Label Types to Predict

In the MLQE-PE dataset, word-level quality labels are assigned to both the words and gaps between each pair of adjacent words. The former, the so-called target label, indicates the quality of each word in the MT output, where "BAD" indicates that the word needs to be deleted or substituted with another one. On the other hand, the latter, the so-called gap label, represents whether some words must be inserted in the gap between the adjacent words ("BAD") or not ("OK").

We consider the task of predicting gap labels itself is arguable, because the correct position of a missing word is not necessarily unique: while the positions of missing articles are deterministic, there are multiple possible solutions for inserting untranslated words and phrases. Please refer to ISO/TC37 (2024) and the MQM-based TQE task tackled at WMT since 2022 (Zerva et al., 2022) for further discussion of the inutility of gap labels for translations in the translation production workflow.

C Computation Time

Table 13 summarizes the computation time in GPU hours for each process.

D Limitations

Our experiment covered only 11 translation directions, and our results do not guarantee the same conclusions on other translation directions. As demonstrated by our experiments, the accuracy can be substantially different even for the same translation direction (see Tables 4 and 5). This implies that the difficulty of the task depends on the characteristics of the test data, the MT systems used for generating MT outputs, and human annotators recruited for manual PE.

All experiments were carried out with up to eight NVIDIA Tesla V100 GPUs. If we had a more powerful environment, higher accuracy could be achieved, for instance, by employing larger pre-trained multilingual encoders, such as XLM-RoBERTa$_{\text{XL}}$ and XLM-RoBERTa$_{\text{XXL}}$ (Goyal et al., 2021), larger batch sizes, longer training, and ensembling multiple models.

E Ethics Statement

As shown in our experiments, the predicted labels do not perfectly correlate with the gold-standard labels obtained through manual PE. Therefore, such predicted labels could mislead potential users. This is not specific to our work, but common in the TQE task.

Backbone encoder	Input	En→De	En→Zh	Ro→En	Et→En	Ne→En	Si→En
XLM-RoBERTa$_{\text{Base}}$	(T', R)	0.855	0.802	0.864	0.796	0.671	0.685
XLM-RoBERTa$_{\text{Base}}$	(T', S, R)	0.854	0.771	**0.881**	0.801	**0.687**	0.693
XLM-RoBERTa$_{\text{Large}}$	(T', R)	0.650	0.682	0.655	0.639	0.562	0.579
XLM-RoBERTa$_{\text{Large}}$	(T', S, R)	0.669	0.700	0.702	0.665	0.606	0.610
INFOXLM$_{\text{Base}}$	(T', R)	0.865	0.829	0.869	0.796	0.677	0.689
INFOXLM$_{\text{Base}}$	(T', S, R)	**0.870**	**0.833**	0.876	**0.804**	0.680	**0.699**
INFOXLM$_{\text{Large}}$	(T', R)	0.713	0.710	0.743	0.704	0.628	0.640
INFOXLM$_{\text{Large}}$	(T', S, R)	0.752	0.760	0.772	0.714	0.645	0.654

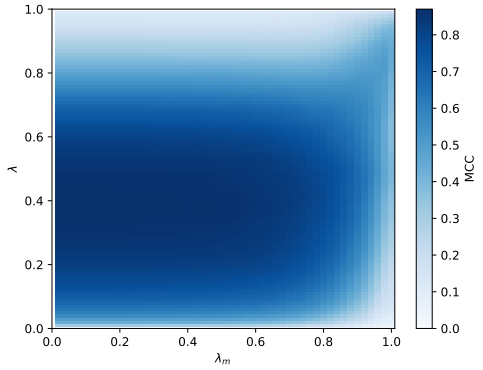
Table 11: MCCs between OT-based hard labels and TER-based labels for the MLQE-PE development data with different pre-trained multilingual encoders: S, T', and R denote the source text, its MT output, and its post-edited version, respectively. **Bold** and underline indicate the highest and second-highest values, respectively.

Backbone encoder	Input	En→De	En→Zh	Ro→En	Et→En	Ne→En	Si→En
INFOXLM$_{\text{Base}}$	(S, T')	0.052	0.048	0.167	0.042	0.032	0.077
	(T', R)	0.865	0.829	0.869	0.796	0.677	0.689
	(R, T')	0.862	0.826	0.868	0.796	0.680	0.688
	(S, T', R)	0.866	0.838	0.873	0.808	**0.686**	**0.705**
	(S, R, T')	0.866	0.838	0.875	0.805	0.684	0.702
	(T', S, R)	**0.870**	0.833	**0.876**	0.804	0.680	0.699
	(R, S, T')	0.867	0.835	0.874	0.802	0.680	0.697
	(T', R, S)	0.865	**0.839**	0.875	**0.811**	0.679	0.699
	(R, T', S)	0.869	0.835	0.871	0.805	0.682	0.697

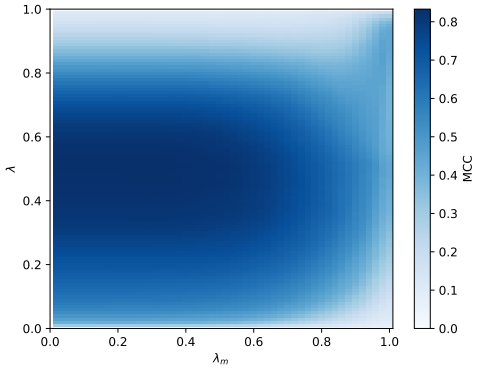
Table 12: MCCs for the MLQE-PE development data with different orderings of S, T', and R.

Step	En→De	En→Zh	Ro→En	Et→En	Ne→En	Si→En
Generating synthetic TQE data						
Fine-tuning M2M-100	9	10	2	2	2	2
Translation with M2M-100	963	856	173	31	6	7
OT-based labeling	103	80	15	4	1	2
TQE model training						
Pre-training with TER-based hard labels			372			
Pre-training with OT-based hard labels			372			
Pre-training with OT-based soft labels			366			
Fine-tuning a classification model			5			
Fine-tuning a regression model			5			
Direct training a classification model			5			
Direct training a regression model			5			

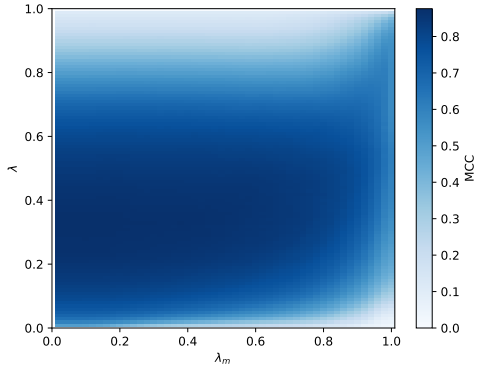
Table 13: GPU hours spent for each phase of TQE model training.



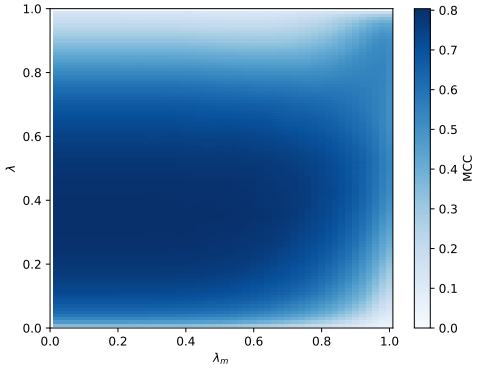
(a) En→De



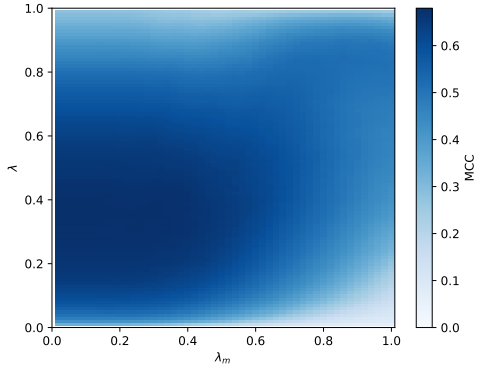
(b) En→Zh



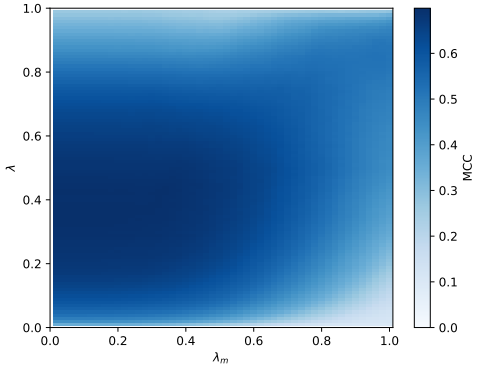
(c) Ro→En



(d) Et→En

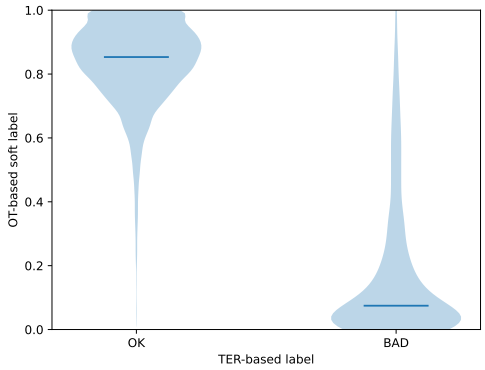


(e) Ne→En

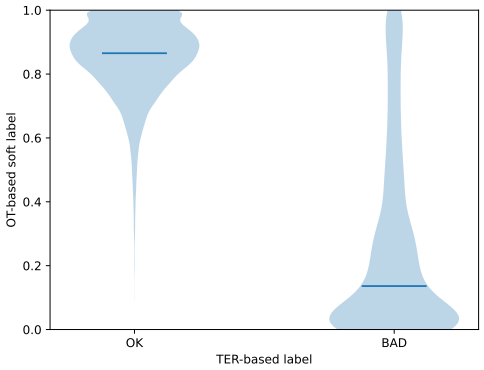


(f) Si→En

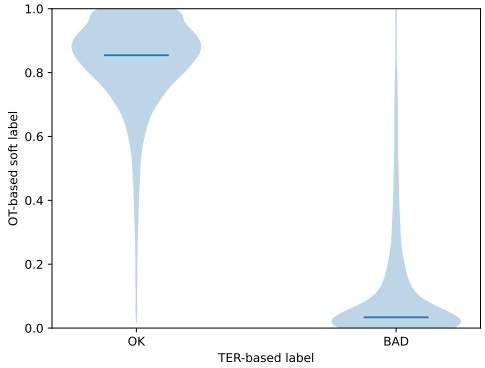
Figure 3: MCC for the MLQE-PE development data with different values for λ_m and λ.



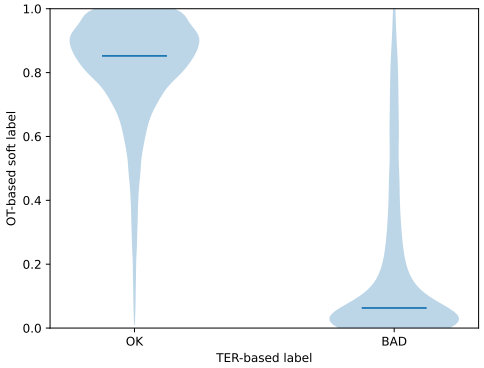
(a) En→De



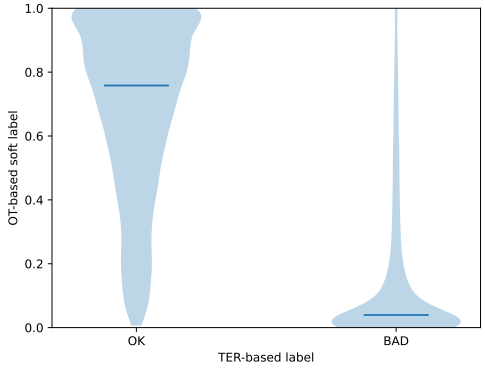
(b) En→Zh



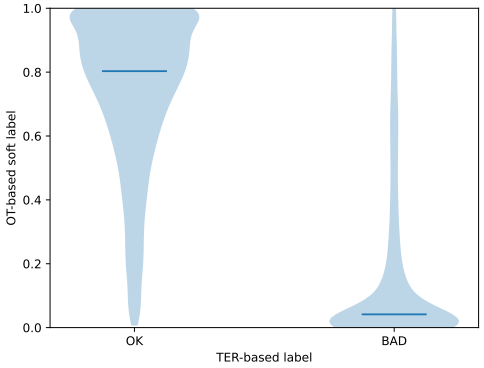
(c) Ro→En



(d) Et→En



(e) Ne→En



(f) Si→En

Figure 4: Distribution of OT-based soft labels for each of the {"OK," "BAD"} labels in the MLQE-PE development data, determined by the optimal λ_m in Table 3: the dark bar indicates the median.