Error Identification for Machine Translation with Metric Embedding and Attention

Raphael Rubino and **Atsushi Fujita** and **Benjamin Marie**
National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan
raphael.rubino,atsushi.fujita,bmarie@nict.go.jp

Abstract

Quality Estimation (QE) for Machine Translation has been shown to reach relatively high accuracy in predicting sentence-level scores, relying on pretrained contextual embeddings and human-produced quality scores. However, the lack of explanations along with decisions made by end-to-end neural models makes the results difficult to interpret. Furthermore, word-level annotated datasets are rare due to the prohibitive effort required to perform this task, while they could provide interpretable signals in addition to sentence-level QE outputs. In this paper, we propose a novel QE architecture which tackles both the word-level data scarcity and the interpretability limitations of recent approaches. Sentence-level and word-level components are jointly pretrained through an attention mechanism based on synthetic data and a set of MT metrics embedded in a common space. Our approach is evaluated on the *Eval4NLP* 2021 shared task and our submissions reach the first position in all language pairs. The extraction of metric-to-input attention weights show that different metrics focus on different parts of the source and target text, providing strong rationales in the decision-making process of the QE model.

1 Introduction

Quality Estimation (QE) for Machine Translation (MT) (Blatz et al., 2004; Quirk, 2004; Specia et al., 2009) aims at providing quality scores or labels to MT output when translation references are not available. Sentence-level QE is usually conducted using human produced direct assessments (DA) (Graham et al., 2013) or post-edits. The latter allows to derive token-level quality indicators such as *good* and *bad* tags (Fonseca et al., 2019; Specia et al., 2020). Token-level QE is particularly useful for applications such as source pre-editing or focused MT post-editing, but requires high-quality fine-grained annotated data for supervised learning. Furthermore, token-level quality indicators can be seen as explanations for sentence-level scores, whether given by humans or automatically produced. However, explainability of QE models decisions is obscured by contemporary approaches relying on large data-driven neural-based models, making use of pretrained contextual language models (LM) such as BERT (Devlin et al., 2019) and XLM (Conneau and Lample, 2019), albeit showing steady performance increase as reported in the QE shared tasks (Fonseca et al., 2019; Specia et al., 2020). Yet, the QE layers and architectures are rarely investigated, neither for performance nor for interpretability purposes, and the center of attention is mainly on large pretrained models and generating additional (synthetic) training corpora.

In this paper, we present a novel QE architecture which encompasses a metric-to-input attention mechanism allowing for several extensions of the habitual QE approach. First, since sentence-level QE scores are usually obtained with surface-level MT metrics computed between translation outputs and human produced references or post-edits such as HTER (Snover et al., 2006), we propose to make use of several metrics simultaneously in order to model translation errors at various granularities, i.e. at the character, token, and phrase levels. Second, we design a *metric embeddings model* which represents metrics in their own space through a dedicated set of learnable parameters, allowing for straightforward extensions of the number and type of metrics. Third, by employing an attention mechanism between metric embeddings and bilingual input representations, the metric-to-input attention weights indicate where each metric focuses given an input sequence, increasing the interpretability of the QE components. We conduct a set of experiments on the *Eval4NLP* 2021 shared task dataset (Fomicheva et al., 2021) using only the training data along with sentence-level scores officially released for the tasks (illustrated in Figure 1). In addition, we

Source	Religioon pakub vaimu puhastamiseks teatud vahendeid .
MT	Religion offers certain means of cleansing the spirit .
PE	Religion offers certain means of cleansing the spirit .
Sentence-level scores: DA 0.905 – chrF 1.0 – TER 0.0 – BLEU 1.0	

Source	Tänu	Uku	kalastamiskirele	pääseb Õnne	13 maja	põlengust .		
MT	Thanks to	the breath of fresh	fishing ,	13 houses are escaped	from	contempt	.	
PE	Thanks to Uku 's passion for fishing, the house at Õnne 13 is saved from fire.							
Sentence-level scores: DA 0.132 – chrF 0.366 – TER 0.667 – BLEU 0.0								

Figure 1: Samples of source sentences, automatic translations and human post-editions, along with direct assessment (DA) scores, taken from the *Eval4NLP* 2021 shared task Estonian–English validation set representing high and low quality translations. Additional metrics are presented, namely chrF, TER and BLEU, to illustrate variations related to metrics granularity. Green and red colors are tokens annotated with classes 0 and 1 respectively.

produce a large synthetic corpus for QE pretraining using publicly available resources.

The contributions of our work are the following: *(i)* a novel QE architecture using metric embeddings and attention-based interpretable neural components allowing for unsupervised token-level quality indicators, *(ii)* an extensible framework designed for unrestricted sentence-level QE scores or labels where new metrics can be added through fine-tuning, *(iii)* the reproducibility guaranteed by the use of publicly available datasets, tools, and models, and *(iv)* word and sentence-level QE results on par or outperforming top-ranked approaches based on the official *Eval4NLP* 2021 shared task results.

The remainder of this paper is organized as follows. In Section 2, we introduce some background in QE based on contextual language model, followed in Section 3 with the detailed implementation of the proposed model using metric embedding and attention. In Section 4, the experimental setup is presented, including the data and tools used, as well as the training procedure of our models. Section 5 contains the results obtained in our experiments along with their analysis and interpretation. A comparison of our method and results with previous work is made in Section 6. Finally, we conclude and suggest future research directions in Section 7.

2 Background

Current state-of-the-art QE approaches are commonly based on sentence encoders taking as input source–translation pairs (Ranasinghe et al., 2020a; Wang et al., 2020; Rubino, 2020). Encoders are usually contextual LMs pretrained on large amount of multilingual data. Existing QE implementations commonly rely on additional layers added on top of a pretrained LM, which enables multi-task learning for word and sentence-level QE.

Pretraining of contextual LMs is done by optimizing a prediction function given input sequences of tokens containing randomly masked tokens, or tokens randomly replaced by other tokens sampled from the vocabulary. Formally, given an input sequence Z of n tokens $z_{1:n}$, corresponding word (or subword) embeddings $x_{1:n}$ with dimension d ($x_{1:n} \in \mathbb{R}^{n \times d}$) are learned, and output contextual embeddings $h^l_{1:n} \in \mathbb{R}^{n \times d}$ are computed at each layer $l \in [1, L] \subset \mathbb{N}$ of a Transformer encoder (Vaswani et al., 2017). Usually based on the output of the last encoder layer, the model optimizes the following loss function: $\ell(s, t) = -\mathbb{E}_{r \sim [1,n]} \log P(z_r | \bar{z}_r)$, where z_r are randomly sampled tokens from z to be masked or replaced and $\bar{z}_r$ are the remaining tokens from z with $r \in [1, n] \subset \mathbb{N}$. To perform QE, QE-specific layers are commonly added on top of pretrained contextual LMs, being fed with contextual token embeddings from the topmost (i.e., L-th) layer of the LM.

For sentence-level QE the specific component is a regression head formalized by $y^s = \sigma(\phi(h^L_{1:n}) \cdot W^s + b^s)$, where $y^s \in [0, 1] \subset \mathbb{R}$. W^s and b^s are trainable parameters of the linear output layer, ϕ is a pooling function, and σ is the sigmoid function. The output y^s of this QE component is a score indicating the sentence-level translation quality.

For token-level QE a classification head is implemented as $y^t_{1:n} = softmax(h^L_{1:n} \cdot W^t + b^t)$, where $y^t_{1:n} \in \mathbb{R}^{n \times |C|}$ with C the set of word-level QE classes, W^t and b^t are trainable parameters of the linear output layer. The output $y^t_{1:n}$ of this QE component is a vector of labels indicating the

translation quality of corresponding input tokens. Source tokens are annotated according to the accuracy of their translation, while annotations of target tokens also take into account their position in the target sequence.

Multiple losses are then computed, one for the sentence-level and one for each word-level outputs (source and target tokens), based on gold labels to train (or finetune) the contextual LM and the QE model in an end-to-end fashion using back-propagation (Kim et al., 2017; Lee, 2020; Rubino and Sumita, 2020). Commonly used losses are cross-entropy and mean-squared error for classification and regression respectively.

However, this approach has limitations While the token-level QE implementation makes use of each input token representation in context thanks to the pretrained LM, the sentence-level QE components relies only on the pooled representation of the input sequence. This approach drastically limits the amount of information flowing through the sentence-level specific set of layers and may force the network to focus more on cues and data artifacts which correlate with QE scores, instead of encoding translation-related features from source and target inputs (Sun et al., 2020). These findings corroborate with the empirical observation made by Kepler et al. (2019), where the authors obtained the best word-level QE results using BERT and ignoring target language features when predicting source quality labels and vice-versa. Additionally, most recent QE approaches do not allow for the interpretability of sentence-level QE predictions at test time and leads to the current state of QE as a set of *black-box* components. Furthermore, token-level error annotations is costly to produce.

3 Metric Embedding and Attention

Motivated by the limitations to contextual LM based QE, we propose a novel architecture employing metric embeddings and attention, which is computed between the contextual embeddings and the embedded QE criteria of the supervised learning task, namely MT automatic metrics or direct assessment scores provided by human annotators.

The metric embedding matrix $E \in \mathbb{R}^{g \times d}$, randomly initialized at the beginning of training, is added on top of the pretrained LM to model metrics in their own space, with a predefined set of sentence-level metrics $M = \{m_1, \ldots, m_g\}$. Each metric is initially represented as a one-hot vector, noted $m_j \in \mathbb{R}^g$ with $j \in [1, g] \subset \mathbb{N}$. Its corresponding embedding is retrieved with $m_j \cdot E$, forming the query used in the attention mechanism (eqn. 1):

$$Q_{i,j} = (m_j \cdot E) \cdot W_i^Q \tag{1}$$

where $i \in [1, u] \subset \mathbb{N}$ is the head index from a predefined number of heads, $Q_{i,j} \in \mathbb{R}^d$ is the metric embedding corresponding to the one-hot vector m_j, $W_i^Q \in \mathbb{R}^{d \times (d/u)}$ is a matrix of learnable parameters projecting the metric embedding into the dimensionality of the attention head (d/u). Note that we present the query computation for a single metric but our implementation allows several metrics to be packed into a single query, sharing the parameter matrix W_i^Q (biases are omitted for the sake of simplicity).

Keys and values which are the two other components of the attention mechanism, noted K_i and V_i respectively, are computed based on the output of the topmost layer of the pretrained LM, which is first fed into a position-wise feed-forward layer following (eqn. 2):

$$\begin{aligned} ff_{1:n} &= ReLU(h^L_{1:n} \cdot W^{s,D_1}) \cdot W^{s,D_2} \\ K_i = ff_{1:n} \cdot W_i^K \ &, \ V_i = ff_{1:n} \cdot W_i^V \end{aligned} \tag{2}$$

where W_i^K and $W_i^V \in \mathbb{R}^{d \times (d/u)}$ are the parameter matrices for the keys and values respectively, $W^{s,D_1} \in \mathbb{R}^{d \times b}$ and $W^{s,D_2} \in \mathbb{R}^{b \times d}$ are parameter matrices of the linear layers with dimensionality b and a $ReLU$ activation function in between, leading to $ff_{1:n} \in \mathbb{R}^{n \times d}$.

Metrics to tokens attention weights aim to represent the focus made by a given metric on specific parts of the input sequences. These attention weights are computed between the embedding of a metric and the contextually encoded input tokens (eqn. 3):

$$\alpha_{i,j,1:n} = \sigma\left(\frac{Q_{i,j} K_i^T}{\sqrt{d/u}}\right) \tag{3}$$

where σ is the sigmoid function. Note that the common way to compute attention weights, as presented in Vaswani et al. (2017), relies on $softmax$ which is based on the exponential function, well-suited for tasks such as machine translation as it results in few *alignments* between tokens involved in the attention mechanism. However, in the case of

unsupervised sequence labeling such as token-level QE without annotated data, zero to many tokens may influence sentence-level scores given a metric. Thus, to allow more flexibility in the distribution of attention weights over input tokens and following the approach presented in (Rei and Søgaard, 2018), we replaced *softmax* by *sigmoid*.

Sentence-level scores are obtained for each metric with the weighted sum of value vectors for each attention head (eqn. 4):

$$attn_{i,j} = \sum \alpha_{i,j,1:n} V_i, \quad (4)$$

where $attn_{i,j} \in \mathbb{R}^{(d/u)}$, before concatenating the output of each head and projecting the result back in the dimensionality of the model (eqn. 5):

$$y_j^{s'} = (attn_{1,j} \oplus \ldots \oplus attn_{u,j}) \cdot W^O \quad (5)$$

with $W^O \in \mathbb{R}^{d \times d}$. Finally, we project $y_j^{s'}$ from the model dimensionality to a single score through a metric specific linear layer: $y_j^s = y_j^{s'} \cdot W_j^s$ with $W_j^s \in \mathbb{R}^{d \times 1}$ and $y_j^s \in [0, 1] \subset \mathbb{R}$.

Token-level QE scores are computed by using the attention weights (see eqn. 3) followed by three transformations: attention heads combination through a linear transformation, concatenation of token embeddings and combined attention heads, combination of metrics through a final linear transformation (eqn. 6):

$$\begin{aligned} y_{j,1:n}^{t'} &= \alpha_{j,1:n} \cdot W^{t,H} \\ y_{1:n}^{t} &= (y_{1:n}^{t'} \oplus h_{1:n}^{L}) \cdot W^{t,O} \end{aligned} \quad (6)$$

where $\alpha_{j,1:n} = (\alpha_{1,j,1:n} \oplus \ldots \oplus \alpha_{u,j,1:n})$, $y_{1:n}^{t'} = (y_{1,1:n}^{t'} \oplus \ldots \oplus y_{g,1:n}^{t'})$, $W^{t,H} \in \mathbb{R}^{u \times 1}$ and $W^{t,O} \in \mathbb{R}^{(d+g) \times 1}$ are parameter matrices of linear layers, leading to $y^t \in [0, 1] \subset \mathbb{R}$ for each token in the input sequence $z_{1:n}$.

The learning process allows for supervised or unsupervised token-level QE. Note that all learnable parameters for the sentence-level QE components except W_j^s are shared between metrics, including the metric embeddings matrix E. We believe that such an approach enables to capture translation errors at different granularities according to the specificity of each metric, e.g., characters, tokens and phrases, while keeping a reasonable total amount of learnable parameters. The loss functions for sentence-level QE are mean-squared error while the losses for token-level QE are cross-entropy. The final loss is obtained by linearly combining all losses computed for each output of the model. The general architecture of our QE model is illustrated in Figure 2.

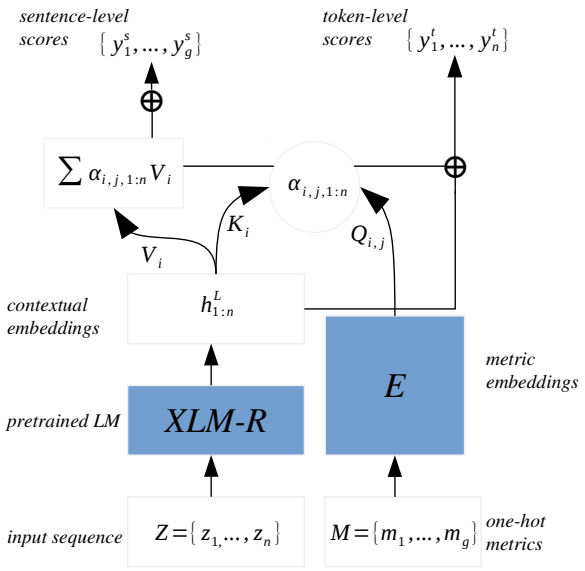

Figure 2: Architecture of the metric embeddings and attention mechanism. Shaded elements, curved arrows and $\oplus$ are parameters of the model, i and j are the attention head and the metric indexes respectively.

Supervised learning is conducted by computing losses according to each output of the model and their corresponding gold labels from the training data. Thus, for the sentence-level QE layers, we compute one loss per metric (Mean Squared Error), while for the token-level QE layers, if token-level annotations are available, two losses allow to optimize the model for source and target tokens separately (Cross-entropy).

Unsupervised learning is conducted when token-level annotations are not available, which is one of the objectives in the constrained task of *Eval4NLP* 2021. In this case, only sentence-level losses are used to optimize the parameter of the model through backpropagation. Following the guidelines of the shared task, we do not use the direct assessment annotations made by humans at the word-level.

4 Experimental Setup

This section presents our experimental setup, including the pretrained models, the datasets and the training procedure. All pretrained models and

scripts used in our experiments are based on PyTorch (Paszke et al., 2019) and all computations are conducted on NVIDIA V100 GPUs with CUDA v10.2.

4.1 Pretrained Models

Two types of pretrained models were necessary to conduct our experiments: contextual embedding LMs to encode bilingual input sequences and MT models to produce synthetic data required for QE pretraining.

Contextual embedding LMs used in our experiments are based on a pretrained XLM-R checkpoint, namely *xlm-roberta-large* from the HuggingFace Transformers library (Wolf et al., 2020). This model, initially introduced in (Conneau et al., 2020), was pretrained on 2.5TB of filtered CommonCrawl data, covering 100 languages with a vocabulary of $250k$ BPE tokens (Sennrich et al., 2016), $1,024$ embedding and hidden-state dimensions, $4,096$-dimensional feed-forward layers and 16 attention heads.

MT models used in our experiments are transformer-based neural MT (NMT) models. For two language pairs and translation directions of the *Eval4NLP* 2021 shared task, namely Estonian→English (ET–EN) and Romanian→English (RO–EN), we used pretrained NMT models made available by the WMT'20 QE shared task organizers (Specia et al., 2020).[1] For German→Chinese (DE–ZH) and Russian→German (RU–DE), the two zero-shot pairs of the shared task, we used the mBART50 model (Liu et al., 2020; Tang et al., 2020).[2] All NMT models are based on the fairseq library (Ott et al., 2019).

4.2 Datasets

Two datasets were used in our experiments: a synthetic dataset for QE pretraining, and the shared task dataset consisting of training, validation and test sets. Details of the latter dataset are presented in Table 1 while we give more information about the synthetic data in this section.

[1]Models available at https://github.com/facebookresearch/mlqe/blob/master/nmt_models/README-models.md

[2]Model available at https://github.com/pytorch/fairseq/tree/master/examples/multilingual

	Lang.	Sent.	Tokens	Types
Train	ET–EN	7.0k	98.1k / 136.6k	28.9k / 14.6k
	RO–EN	7.0k	120.2k / 123.3k	23.5k / 15.2k
Valid	ET–EN	1.0k	14.4k / 20.1k	6.9k / 4.7k
	RO–EN	1.0k	17.3k / 17.7k	6.4k / 4.8k
Test	ET–EN	1.0k	14.0k / 19.6k	6.9k / 4.7k
	RO–EN	1.0k	17.4k / 17.8k	6.3k / 4.8k
	DE–ZH	1.4k	24.9k / 52.8k	8.4k / 2.2k
	RU–DE	1.2k	25.4k / 28.8k	10.2k / 7.5k

Table 1: Official training, validation and test data released for the *Eval4NLP* 2021 shared task. DE–ZH and RU–DE are zero-shot language pairs thus have neither training nor validation corpora. Tokens and types columns contain source / MT counts, *k* stands for thousands, Chinese tokens and types are characters.

Lang.	Sent.	Tokens	Types
ET–EN	24.9M	322.5M / 411.0M	4.8M / 2.8M
RO–EN	42.1M	600.5M / 601.2M	4.0M / 3.6M
DE–ZH	19.8M	422.8M / 708.1M	4.5M / 3.3k
RU–DE	19.5M	256.9M / 262.7M	4.4M / 4.4M

Table 2: Synthetic data produced for QE pretraining. Tokens and types columns contain source / MT counts, *M* stands for millions and *k* for thousands, Chinese tokens and types are characters.

Synthetic data generation was based on gathered parallel corpora translated by the NMT systems presented in Section 4.1. The translated sentences were compared to the target side of the parallel corpora to produce sentence-level scores based on chrF (Popović, 2016), TER (Snover et al., 2006) and BLEU (Papineni et al., 2002) metrics. Additionally, only for the synthetic data, we produced token-level scores following the usual procedure to determine post-editing effort (Specia et al., 2020).[3] For this step, word alignments were required to obtain source-side token-level quality indicators. We used the same parallel corpora to produce synthetic data and to train word alignments based on the IBM 2 model (Brown et al., 1993) and trained using *fast_align* (Dyer et al., 2013). Details about the synthetic data are presented in Table 2.[4]

The special case of DE–ZH resulting from preliminary experiments, we noticed for this language pair that the translation quality of the synthetic data was low compared to the three other language pairs. We assumed that it was due to two issues:

[3]Scripts and procedure available at https://github.com/deep-spin/qe-corpus-builder

[4]Parallel corpora were collected from the WMT news translation task (Tiedemann, 2016) and OPUS (Tiedemann, 2016).

the quality of the DE–ZH parallel corpora and the performance of the NMT model. To tackle the first issue, we generated our own DE–ZH parallel corpora by pivot-based (back-) translation, starting from a monolingual Chinese corpus composed of CommonCrawl and NewsCrawl 2018 to 2020, translating it into English using an in-house NMT model trained with Marian (Junczys-Dowmunt et al., 2018) on the WMT'21 QE ZH–EN parallel corpus, then translating the English output into German using the EN–DE NMT model released by the WMT'20 QE shared task organizers, resulting in a synthetic DE–ZH parallel corpus. To tackle the second issue, we finetuned mBART50 using NewsCommentary and MultiUN DE–ZH retrieved from OPUS, as these two corpora appeared to be the cleanest among the available ones.

4.3 Training Procedure

We detail in this section the training procedures employed for QE pretraining on the synthetic data and finetuning on the officially released training data.

QE pretraining was conducted per language pair starting from the XLM-R checkpoint presented in Section 4.1 using two different random seeds and learning rates. Additionally, four QE pretraining were conducted on the concatenation of all synthetic data using four different random seeds and two learning rates. Training was ran for two epochs for the language specific models and for a single epoch for the remaining ones. We restricted the length of training samples to a minimum of 5 and a maximum of 128 subword tokens for the bilingual models and a maximum of 96 subword tokens for the multilingual ones. Training was conducted with batches of 128 source and target sequences with the AdamW optimizer (Loshchilov and Hutter, 2019) (parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1 \times 10^{-6}$). A linear learning rate warmup was employed during the first 50k updates to reach a maximum value of 5×10^{-6} or 2×10^{-6} depending on the model and random seed, which remained without decay until the end of the first epoch. The dropout rates were set to 0.1 for both the embeddings and the transformer blocks (feed-forward and attention layers), the model dimensionality and embedding size was $1,024$, feed-forward layers had a dimensionality of $4,096$ and the numbers of attention heads were set to 16 for the language model and 8 for the metric attention block.

Finetuning was conducted using the officially released data presented in Table 1 during 20 epochs, monitoring the performance of each model using the validation set. No length restriction was applied on these corpora. In addition to the three automatic metrics used during QE pretraining, namely chrF, TER and BLEU, the direct assessment scores provided by the shared task organizers were used by simply adding an entry in the metric embeddings matrix. A few hyperparameters, namely the batch size, learning rate, as well as embedding and Transformer dropout rates, were optimized in a grid-search manner. The best resulting models according to token-level source and target performances based on the official metrics (Area Under the Curve, AUC, and Average Precision, AP) were kept for ensembling and predicting scores on the validation and test sets.

5 Results and Analysis

We present in Table 3 the results obtained on the *Eval4NLP* 2021 shared task as reported by the organizers, including our baselines and final submissions along with the three baselines proposed by the organizers, namely random scores, TransQuest (Ranasinghe et al., 2020b) combined with LIME (Ribeiro et al., 2016) (noted *Official baseline 1*), and XMoverScore (Zhao et al., 2020) combined with SHAP (Lundberg and Lee, 2017) (noted *Official baseline 2*). Our baselines are composed of ensembles of two finetuned language-specific QE pretrained models while our final submissions are composed of ensembles of eight finetuned models for each of token-level tasks (source and target) and eight models for sentence-level tasks. The eight token-level models are, for each random seed, the best language-specific finetuned models according to source or target AUC and AP, and the best multilingual models according to source or target AUC. The eight sentence-level models are the best direct assessment Pearson's ρ for both bilingual and multilingual models, as well as the best direct assessment RMSE for the bilingual model.

Results show that our baselines and final submissions largely outperform the organizer's baselines on the four language pairs, while our final submissions reach higher source and target token-level performances compared to our baselines. Except for the DE–ZH pair, our final submissions are outperforming our baselines on the sentence-level Pearson's ρ evaluation. Because the main objec-

Lang.	Model	Source Token AUC	Source Token AP	Source Token top-K R	Target Token AUC	Target Token AP	Target Token top-K R	Sentence Pearson's ρ
ET–EN	Random baseline	0.488	0.338	0.193	0.496	0.358	0.247	-0.029
	Official baseline 1	0.545	0.440	0.309	0.624	0.536	0.426	0.772
	Official baseline 2	0.535	0.370	0.231	0.616	0.441	0.339	0.494
	Our baseline	0.926	0.848	0.761	0.887	0.808	0.714	0.793
	Our submission	**0.932**	**0.852**	**0.771**	**0.896**	**0.824**	**0.734**	**0.845**
RO–EN	Random baseline	0.501	0.281	0.149	0.515	0.312	0.188	0.017
	Official baseline 1	0.478	0.351	0.243	0.635	0.523	0.415	0.899
	Official baseline 2	0.535	0.293	0.148	0.667	0.536	0.437	0.695
	Our baseline	0.937	0.826	0.727	0.942	0.860	0.770	0.855
	Our submission	**0.947**	**0.851**	**0.752**	**0.946**	**0.869**	**0.778**	**0.918**
DE–ZH	Random baseline	0.499	0.300	0.172	0.495	0.293	0.172	0.000
	Official baseline 1	0.486	0.317	0.196	0.461	0.271	0.145	**0.335**
	Official baseline 2	0.474	0.288	0.158	0.545	0.333	0.220	0.176
	Our submission	**0.847**	**0.645**	**0.509**	**0.849**	**0.679**	**0.571**	0.286
RU–DE	Random baseline	0.506	0.340	0.238	0.494	0.309	0.217	-0.017
	Official baseline 1	0.535	0.427	0.320	0.403	0.263	0.165	0.498
	Official baseline 2	0.522	0.356	0.261	0.523	0.329	0.227	0.252
	Our submission	**0.922**	**0.804**	**0.709**	**0.927**	**0.829**	**0.736**	**0.679**

Table 3: Official test results of the *Eval4NLP* 2021 shared task, according to three metrics for source and target token-level QE (AUC, AP and Recall at top-K), and one metric for sentence-level QE (Pearson's ρ).

tives of the shared task was token-level evaluation, we did not focus on improving the sentence-level scores. We assume that further improvements are achievable on this aspect of QE. Additionally, due to the lack of validation sets for the zero-shot language pairs, we could not try to improve over a baseline and thus only provided a unique and final submission. Note that we did not use the official word-level training data at all, neither during pretraining nor during finetuning of our models. Only the provided validation set was used for monitoring purposes.

In order to evaluate the impact of QE pretraining and finetuning, as well as the difference in performance between ensemble and single models trained using language pair specific (bilingual) and multilingual datasets, we present an ablation study conducted on the word-level QE in Table 4 for the two non zero-shot language pairs. The results obtained without ensemble models are the average of results from individual models. The ablation study shows that individual models (- *Ensemble*) are outperformed by the ensemble (*Submission*), while removing language pair specific training data (- *Bilingual*) has limited impact on performances for ET–EN, which motivates the use of multilingual pretrained models and ensembling. Comparing removing finetuning and QE pretraining, the latter leads to the largest performance drop while the former has a relatively limited impact. This indicates that large amount of synthetic data combined with our approach performs well even without using any of the provided manually annotated data for the shared task.

Lang.	Model	Source Token AUC	Source Token AP	Target Token AUC	Target Token AP
ET–EN	Submission	0.876	0.781	0.904	0.843
	- Ensemble	0.869	0.773	0.898	0.831
	- Bilingual	0.869	0.771	0.898	0.831
	- Finetuning	0.856	0.752	0.875	0.791
	- Pretraining	0.633	0.515	0.631	0.513
RO–EN	Submission	0.928	0.860	0.950	0.891
	- Ensemble	0.918	0.843	0.941	0.873
	- Bilingual	0.916	0.839	0.937	0.866
	- Finetuning	0.903	0.815	0.925	0.835
	- Pretraining	0.582	0.350	0.573	0.417

Table 4: Results of the ablation study on non zero-shot pairs obtained on the validation set for token-level QE.

As an explanation of sentence-level scores predicted by our model, we propose to extract the attention weights computed between the metric embeddings and the contextually encoded input sequences. A few samples extracted from the validation set are presented in Figure 3. We can see on these examples that individual metrics do not correlate with human annotations. However, the multimetric approach, which relies on heads and metrics combination through linear layers, provides a potential error identification method.

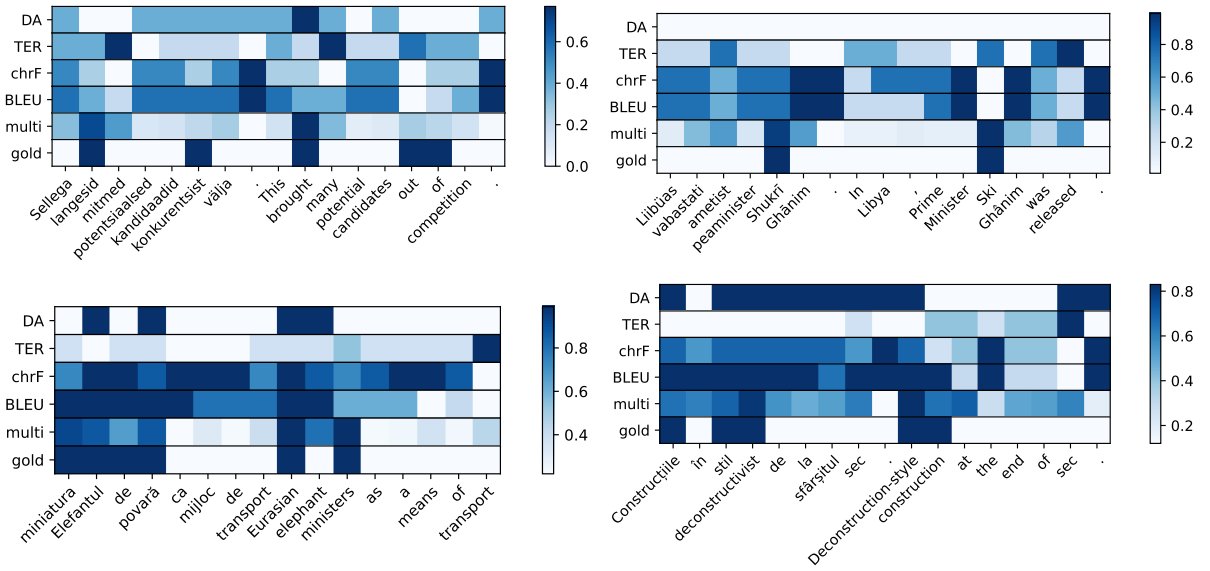

Figure 3: Attention weights computed between individual metric embeddings, namely DA, TER, chrF and BLEU, along with the multimetric approach (see eqn. 6) and the human annotations (noted *gold*). Samples extracted from the ET–EN and RO–EN validation sets (top and bottom respectively).

6 Previous Work

Since the shift of most NLP tasks towards using large pretrained contextual LMs as basis for task-specific finetuning, the research community working on QE for MT moved from a classic two-step process of feature engineering followed by machine learning (Blatz et al., 2004; Quirk, 2004; Specia et al., 2009) to an end-to-end training neural-based paradigm. First attempts in this direction were conducted by Kim et al. (2017) with the predictor-estimator, which inspired further work in using various types of encoders (Wang et al., 2020), enriching the model with features extracted from NMT models (Moura et al., 2020; Fomicheva et al., 2020a) or modifying the pretraining objective of contextual LMs for QE adaptation (Rubino and Sumita, 2020).

More recently, due to the costly nature of data acquisition for supervised learning of QE models, unsupervised approaches were proposed by relying mostly on signals given by NMT systems when translating source sentences (Fomicheva et al., 2020b), the so-called *glass-box* features. Alternatively, when the NMT systems which produced data to perform QE on are not available (i.e. *black-box* setting), relying on large amount of synthetic data for contextual LM continued training prior to finetuning appears to be an effective way to approximate human judgments of translation quality (Lee, 2020; Tuan et al., 2021).

7 Conclusion

This paper presented a novel QE architecture for unsupervised token-level quality prediction providing sentence-level explainable decisions from the model. We implemented a metric embeddings and attention mechanism on top of a widely used pretrained contextual LM, allowing to add metrics during finetuning and enabling high performance QE both at the levels of token and sentence. This extensible framework was shown to produce results on par or outperforming state-of-the-art QE approaches without relying on human-produced token-level annotations, which could be approximated with the use of relatively cost-effective synthetic data and automatic metrics. Our pivot-based translation approach also tackled a recurrent issue in MT when parallel data are scarce and final results for zero-shot language pairs validated this method.

Acknowledgements

We would like to thank the reviewers for their insightful comments and suggestions. A part of this work was conducted under the commissioned research program "Research and Development of Advanced Multilingual Translation Technology" in the "R&D Project for Information and Communications Technology (JPMI00316)" of the Ministry of Internal Affairs and Communications (MIC), Japan, and supported by JSPS KAKENHI grant numbers 20K19879 and 19H05660.

References

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence Estimation for Machine Translation. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 315–321. International Committee on Computational Linguistics.

Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, and Robert L Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational linguistics*, 19(2):263–311.

Alexis Conneau, 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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual Language Model Pretraining. In *Advances in Neural Information Processing Systems*, pages 7057–7067. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648. Association for Computational Linguistics.

Marina Fomicheva, Piyawat Lertvittayakumjorn, Wei Zhao, Steffen Eger, and Yang Gao. 2021. The eval4nlp shared task on explainable quality estimation: Overview and results. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Vishrav Chaudhary, Mark Fishel, Francisco Guzmán, and Lucia Specia. 2020a. BERGAMOT-LATTE Submissions for the WMT20 Quality Estimation Shared Task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1010–1017. Association for Computational Linguistics.

Marina Fomicheva, 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*, 8:539–555.

Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the WMT 2019 Shared Tasks on Quality Estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–12, Florence, Italy. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous Measurement Scales in Human Evaluation of Machine Translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, et al. 2018. Marian: Fast Neural Machine Translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121. Association for Computational Linguistics.

Fábio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, António Góis, M Amin Farajian, António V Lopes, and André FT Martins. 2019. Unbabel's Participation in the WMT19 Translation Quality Estimation Shared Task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 80–86. Association for Computational Linguistics.

Hyun Kim, Jong-Hyeok Lee, and Seung-Hoon Na. 2017. Predictor-Estimator using Multilevel Task Learning with Stack Propagation for Neural Quality Estimation. In *Proceedings of the Second Conference on Machine Translation*, pages 562–568. Association for Computational Linguistics.

Dongjun Lee. 2020. Two-Phase Cross-Lingual Language Model Fine-Tuning for Machine Translation Quality Estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1024–1028. Association for Computational Linguistics.

Yinhan Liu, 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*, 8:726–742.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *Proceedings of the Seventh International Conference on Learning Representations*.

Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

João Moura, miguel vera, Daan van Stigt, Fabio Kepler, and André F. T. Martins. 2020. IST-Unbabel Participation in the WMT20 Quality Estimation Shared Task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1029–1036. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL*, pages 311–318. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, pages 8026–8037. Curran Associates, Inc.

Maja Popović. 2016. chrF Deconstructed: β Parameters and n-gram Weights. In *Proceedings of the First Conference on Machine Translation*, pages 499–504. Association for Computational Linguistics.

Christopher Quirk. 2004. Training a Sentence-Level Machine Translation Confidence Measure. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, pages 825–828. European Language Resources Association (ELRA).

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020a. TransQuest at WMT2020: Sentence-Level Direct Assessment. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1049–1055. Association for Computational Linguistics.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020b. TransQuest: Translation Quality Estimation with Cross-lingual Transformers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5070–5081. International Committee on Computational Linguistics.

Marek Rei and Anders Søgaard. 2018. Zero-shot sequence labeling: Transferring knowledge from sentences to tokens. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 293–302. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. Association for Computing Machinery.

Raphael Rubino. 2020. NICT Kyoto Submission for the WMT'20 Quality Estimation Task: Intermediate Training for Domain and Task Adaptation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1042–1048. Association for Computational Linguistics.

Raphael Rubino and Eiichiro Sumita. 2020. Intermediate Self-supervised Learning for Machine Translation Quality Estimation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4355–4360. International Committee on Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 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. Association for Machine Translation in the Americas.

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

Lucia Specia, Nicola Cancedda, Marc Dymetman, Marco Turchi, and Nello Cristianini. 2009. Estimating the Sentence-Level Quality of Machine Translation Systems. In *Proceedings of the 13th Annual conference of the European Association for Machine Translation*, pages 28–35. European Association for Machine Translation.

Shuo Sun, Francisco Guzmán, and Lucia Specia. 2020. Are we Estimating or Guesstimating Translation Quality? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6262–6267. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual Translation with Extensible Multilingual Pretraining and Finetuning. *arXiv preprint arXiv:2008.00401*.

Jörg Tiedemann. 2016. OPUS – Parallel Corpora for Everyone. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation: Projects/Products*. Baltic Journal of Modern Computing.

Yi-Lin Tuan, Ahmed El-Kishky, Adithya Renduchintala, Vishrav Chaudhary, Francisco Guzmán, and Lucia Specia. 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. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Advances in Neural Information Processing Systems*, pages 5998–6008. Curran Associates, Inc.

Minghan Wang, Hao Yang, Hengchao Shang, Daimeng Wei, Jiaxin Guo, Lizhi Lei, Ying Qin, Shimin Tao, Shiliang Sun, Yimeng Chen, et al. 2020. HW-TSC's Participation at WMT 2020 Quality Estimation Shared Task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1056–1061. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 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. Association for Computational Linguistics.

Wei Zhao, Goran Glavaš, Maxime Peyrard, Yang Gao, Robert West, and Steffen Eger. 2020. On the Limitations of Cross-lingual Encoders as Exposed by Reference-Free Machine Translation Evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1656–1671. Association for Computational Linguistics.