Transformer-based unified neural network for quality estimation and Transformer-based re-decoding model for machine translation

Cong Chen, Qinqin Zong, Qi Luo, Bailian Qiu, and Maoxi Li*

Jiangxi Normal University, NanChang JiangXi, China
chencong.jxnu@gmail.com,
{zongqinqin,luoqi,mosesli}@jxnu.edu.cn,
qiubl@ecjtu.edu.cn

Abstract. In this paper, we describe our submitted system for CCMT 2020 sentence-level quality estimation subtasks and machine translation subtasks. We propose two models: (i) a Transformer-based unified neural network for the quality estimation of machine translation, which consists of a Transformer bottleneck layer and a bidirectional long short-term memory network that are jointly optimized and fine-tuned for quality estimation, and (ii) a Transformer-based re-decoding model for machine translation, which takes the generated machine translation outputs as the approximate contextual environment of the target language and synchronously re-decodes each token in the machine translation outputs. Experimental results on the development set show that the proposed approaches outperform the baseline systems.

Keywords: machine translation, quality estimation of machine translation, re-decoding, encoder-decoder model

1 Introduction

The 16th China Conference on Machine Translation (CCMT 2020) was organized around machine translation[10] evaluation tasks, which consist of four subtasks: bilingual translation, multilingual translation, speech translation, and the quality estimation of machine translation. The team of Jiangxi Normal University participated in two subtasks in the conference: the sentence-level quality estimation of machine translation and machine translation. The systems and related technologies we used for these two evaluation subtasks and the system performance for the development set are presented in this paper.

2 Model

2.1 Transformer-based Unified Neural Network for the Quality Estimation of Machine Translation

The quality estimation of machine translation output is performed without relying on reference translations. Quality estimation plays an important role in postediting[6]. Sentence-level translation quality estimation is generally regarded as a regression problem. Features are extracted from source sentences and their machine translation outputs[1], and then input into a regression model to obtain a sentence quality score for the machine translation[4].

A bottleneck layer is generally defined as a multilayer neural network that abstracts the raw instance into a high-dimensional embedding in the deep neural network. The bottleneck layer and its output embeddings play an important role in transfer learning[9]. To fully use the bilingual associative knowledge learned from the bilingual parallel corpus through the Transformer model, we propose a Transformer-based unified neural network for quality estimation (TUNQE) model, which is a combination of the bottleneck layer of the Transformer model with a bidirectional long short-term memory network (Bi-LSTM), as shown in Figure 1. The process by which the translation outputs to be estimated and the corresponding source sentences reach the top of the bottleneck layer through the trained unified neural network can be regarded as a feature extraction process for words in the machine translations. The Bi-LSTM layer converts word-level features into sentence-level features, which are input to a feed-forward neural network to predict the translation quality scores.

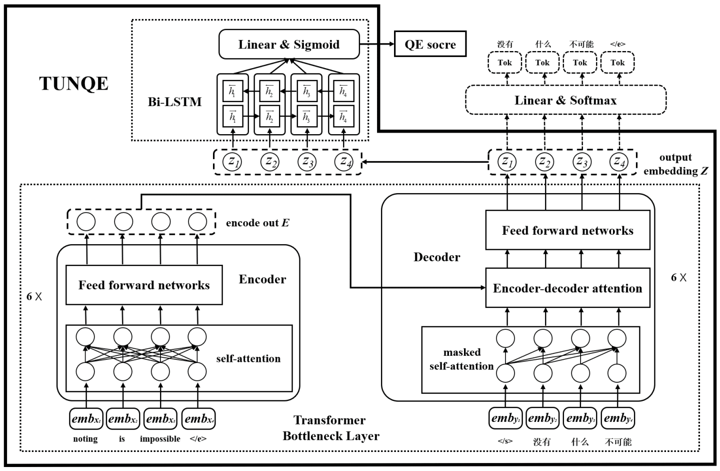

TUNQE
Bi-LSTM
Linear & Sigmoid
QE socre
$\overleftarrow{h_1}$ $\overleftarrow{h_2}$ $\overleftarrow{h_3}$ $\overleftarrow{h_4}$
$\overrightarrow{h_1}$ $\overrightarrow{h_2}$ $\overrightarrow{h_3}$ $\overrightarrow{h_4}$
z_1 z_2 z_3 z_4
没有 什么 不可能 </e>
Tok Tok Tok Tok
Linear & Softmax
z_1 z_2 z_3 z_4
output embedding Z
encode out E
Feed forward networks
Encoder
6 X
self-attention
emb_{x_1} emb_{x_2} emb_{x_3} emb_{x_4}
noting is impossible </e>
Transformer Bottleneck Layer
Decoder
Feed forward networks
Encoder-decoder attention
6 X
masked self-attention
emb_{y_1} emb_{y_2} emb_{y_3} emb_{y_4}
</s> 没有 什么 不可能

Figure 1. TUNQE model architecture.

Feature Extraction Module The feature extraction module is located in the bottleneck layer of the Transformer model (the bottom half of Figure 1). This module is used to extract the quality embedding of the words in the machine translations to be estimated, namely, the word embedding Z of the output of the Transformer bottleneck layer.

The first step in extracting the quality embedding Z of the words in the translations is to encode the input source sentences X to yield the encoder output E:

$$A_e = LN\left(W_{emb}X + Attention\left(W_{emb}X, W_{emb}X, W_{emb}X\right)\right) \tag{1}$$

$$E = LN\left(A_e + FFN\left(A_e\right)\right) \tag{2}$$

where $Attention()$ is the self-attention function in the Transformer, $LN()$ is the layer normalization function in the Transformer, $FFN()$ is the position feed-forward neural network function[8], the symbol A_e represents the output embedding of the encoder's self-attention, and W_{emb} is the word embedding matrix.

The source sentences are encoded to obtain the representation E, which is input into the decoder with the machine translation to be estimated Y, and the quality embedding Z of the machine translations is extracted:

$$A_d = LN\left(W_{emb}Y + Attention\left(W_{emb}Y, W_{emb}Y, W_{emb}Y\right)\right) \tag{3}$$

$$A_{ed} = LN\left(A_d + Attention\left(A_d, E, E\right)\right) \tag{4}$$

$$Z = LN\left(A_{ed} + FFN\left(A_{ed}\right)\right) \tag{5}$$

where $Z = (z_1, z_2, ..., z_{L_y})$ is the quality embedding of the words in the machine translation outputs. z_i is the quality embedding of the i_{th} word in the machine translations. A_d represents the self-attention of the encoder. A_{ed} is the attention of the encoder-decoder. L_y is the length of the machine translation Y.

Quality Estimation Module The quality estimation module is implemented by Bi-LSTM, and the quality embedding Z of the translation to be estimated is obtained by the feature extraction module and input to calculate the quality score QE_{sorre}:

$$\overrightarrow{h}_{1:L_y}; \overleftarrow{h}_{1:L_y} = BiLSTM\left(z_1, z_2, \ldots, z_{L_y}\right) \tag{6}$$

$$Z_{sen} = \frac{1}{L_y}\sum_{i=1}^{L_y}\left[\overrightarrow{h}_i; \overleftarrow{h}_i\right] \tag{7}$$

$$QE_{sorre} = sigmoid\left(W_{qe}Z_{sen}\right) \tag{8}$$

where the symbol $\overrightarrow{h}_i$ represents the hidden state of the i_{th} forward time-step of Bi-LSTM, and $\overleftarrow{h}_i$ represents the hidden state of the i_{th} backward time-step of Bi-LSTM. Z_{sen} is the sentence-level quality embedding of the machine translations, and W_{qe} is the weight parameters of the full connection layer in the quality estimation module.

2.2 Study of Re-decoding-based Neural Machine Translation

In recent years, the Transformer[8] , which exploited the self-attention mechanism in the encoder and in the decoder, significantly improved translation quality. However, the model usually generates a sequence token-by-token from left to right; hence, this autoregressive decoding procedure lacks the guidance of a future context, which is crucial to prevent undertranslation. To alleviate this issue, we propose a TransRedecoder model (Figure 2), which employs a Mask-CURRENT attention matrix (Figure 3(b)) to predict the re-decoding output sequence.

As shown in Figure 2, the same encoder structure is used in the TransRedecoder model as in the Transformer model. The decoder of the TransRedecoder model is an identical layer. Unlike the masked matrix used in the Transformer decoder (Figure 3(a)), the TransRedecoder model decoder employs the Mask-CURRENT attention matrix to fully use the machine translation generated by the Transformer as an approximate contextual environment of the target language. During the re-decoding, we enter the source language (src) and machine translation (mt) generated by the Transformer into the encoder and decoder, respectively. The former contents and the post contents of position j are combined in the machine translation of the target language generated by the Transformer to generate the re-decoding machine translation.

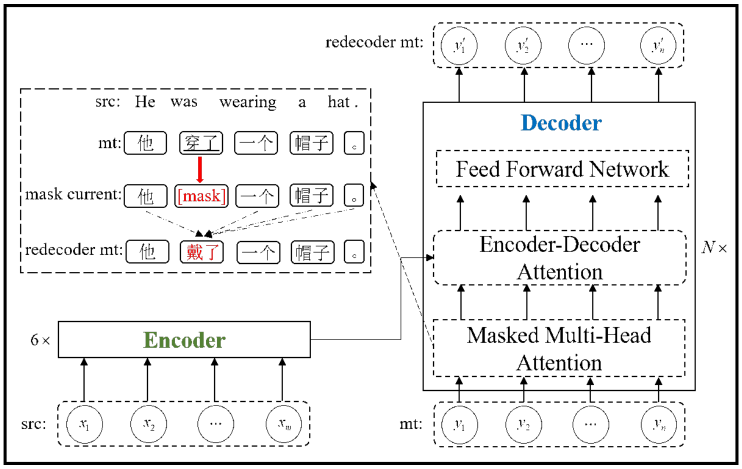


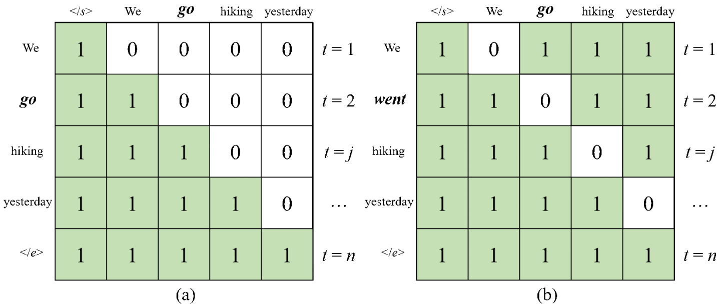
Figure 2. TransRedecoder model architecture.

Given an input sentence $x = (x_1, x_2, \ldots, x_m)$ and its translation outputs $y = (y_1, y_2, \ldots, y_n)$, the model successively re-decodes each token in the translation outputs and generates a new machine translation output y'. As a result, every token in the re-decoded sequence y' fully uses the contextual information:

$$P\left(y' \mid x, y; \theta'\right)=\prod_{i=1}^{n} P\left(y_i' \mid x, y; \theta'\right) \tag{9}$$

where θ' represents the parameters of the TransRedecoder model.

As shown in Figure 3(b), the attention matrix is utilized to combine the former contents and the post contents of position j in the machine translation of the target language generated by the Transformer to generate the re-decoding machine translation. When t = 2, the decoder applies the masked vector (1,1,0,1,1) for masking "go", utilizes the contextual information "$\langle/s\rangle$, we, hiking, yesterday" of "go", and modifies "go" into "went". This helps solve the under-translation problem caused by the absence of the future text.



	</s>	We	*go*	hiking	yesterday	
We	1	0	0	0	0	$t = 1$
go	1	1	0	0	0	$t = 2$
hiking	1	1	1	0	0	$t = j$
yesterday	1	1	1	1	0	…
</e>	1	1	1	1	1	$t = n$

(a)

	</s>	We	*go*	hiking	yesterday	
We	1	0	1	1	1	$t = 1$
went	1	1	0	1	1	$t = 2$
hiking	1	1	1	0	1	$t = j$
yesterday	1	1	1	1	0	…
</e>	1	1	1	1	1	$t = n$

(b)

Figure 3. Two attention matrices: (a) Transformer masked attention matrix and (b) Mask-CURRENT attention matrix.

3 Experiment

3.1 Setting

The configuration of the computer hardware and software environment is shown in Table 1. English sentences are normalized, lowercased, tokenized, and segmented using the BPE subword. Chinese sentences are segmented by the Stanford word segmenter.

Table 1. Computer operating system and hardware configuration.

Operating system	CPU	Memory	GPU
Ubuntu19.04 LTS	Intel i5-6500	32G	GeForce GTX 2080Ti

To evaluate quality estimation, we pre-train the Transformer model using the data provided by the CWMT2018 news translation task. The TUNQE model is jointly optimized and fine-tuned with the training set provided by the CCMT2020 quality estimation tasks. Statistics for the corpus size are shown in Table 2.

Table 2. Statistics for translation quality estimation evaluation corpus.

	Direction	Training set	Development set	Test set
CWMT2018	parallel corpus	6 M	3000	3000
CCMT2020	zh-en	10070	1143	4211
	en-zh	14789	1381	4355

The training set used to evaluate the machine translation subtask is entirely provided by the CCMT2020 machine translation subtask. Statistics for the corpus size are shown in Table 3.

Table 3. Statistics for translation quality estimation evaluation corpus.

	Pair of sentences			Number of tokens		
	Training set	Development set	Test set	Training set	Development set	Test set
en-zh	9 M	10K	1997	1375 M	2.5 M	0.2 M
zh-en	9 M	10K	2000	1375 M	2.5 M	0.2 M

The TUNQE model is developed based on the Facebook fairseq open source toolkit[5]. The Transformer bottleneck layer is pre-trained by using the Adam optimizer ($\beta_1 = 0.9, \beta_2 = 0.98, \varepsilon = 10^{-9}$), where the learning rate lr=0.0007 and the minimum learning rate min_lr=10^{-9}. The SGD optimizer is adopted when the Transformer bottleneck layer and the Bi-LSTM layer are jointly optimized and fine-tuned with a quality estimation training set. The learning rate is fixed at 0.05.

The parameters of the Transformer translation model are consistent with the Transformer-base proposed by Vaswani[8]. The decoder of the TransRedecoder model is an identical layer, and the remaining model parameters are consistent with those of the Transformer model. The feed forward neural network layer has a dimensionality of 2048. We employ 8 parallel attention layers or heads. The Adam optimizer is used to train the model at a learning rate lr=0.0003 and a minimum learning rate min_lr= 10^{-9}. To facilitate these residual connections, all the sublayers in the model, as well as the embedding layers, produce outputs of dimension 512.

3.2 Results

Sentence-Level Quality Estimation Task The performance of quality estimation is evaluated in terms of the Pearson correlation coefficient between the quality estimation and human judgments, and the Spearman correlation coefficient is used to measure the correlation between the rankings of the translation quality and human judgments. The higher the Pearson or Spearman correlation coefficient is, the higher the model performance is. The TUNQE model is tested on the CCMT2020 development set for sentence-level quality estimation. The experimental results are shown in Table 4.

Table 4. TUNQE results for CCMT2020 sentence-level QE dev set.

model	parallel corpus	en-zh		zh-en	
		Pearson	Spearman	Pearson	Spearman
$\text{TUNQE}_{\text{SEP}}$	CWMT 6 M	0.4476	0.3128	0.4877	0.4277
TUNQE		0.5055	0.3555	**0.5888**	0.4806
$\text{TUNQE}_{\text{BERT}}$		**0.5322**	0.3785	0.5735	0.4721

We assess the advantages of jointly training the unified neural network by comparing the performances of TUNQE and $\text{TUNQE}_{\text{SEP}}$. $\text{TUNQE}_{\text{SEP}}$ is a method of separately training the Transformer bottleneck and the Bi-LSTM layers using a bilingual parallel corpus and a sentence-level quality estimation training set, respectively. The experimental results in Table 4 show that the TUNQE method outperforms the $\text{TUNQE}_{\text{SEP}}$ method. The Pearson correlation coefficients of TUNQE are improved by 12.9% and 20.7% in the en-zh and zh-en directions, respectively, over those of $\text{TUNQE}_{\text{SEP}}$.

Li et al. verified that the integration of BERT contextual word embedding[2] can improve translation quality estimation by using the fluency features of the translation[11]. We apply this method to estimate the translation quality, where by BERT pre-trained word embedding in the translation is fused with the embedding extracted by TUNQE after the average pooling of the last 4 layers of representation, which is named $\text{TUNQE}_{\text{BERT}}$. The experimental results show that $\text{TUNQE}_{\text{BERT}}$ exhibits higher system performance than TUNQE.

Machine translation Task The final machine translation results of the CCMT 2020 en-zh and en-zh direction development sets are shown in Table 5. In the en-zh direction, the BLEU score of the re-decoding machine translation increases by 1.26, and the NIST score of the re-decoding machine translation increases by 0.15.

We verify the validity of the TransRedecoder model by using the same data post-processing and scoring approaches for the experimental results submitted by KSAI[3] and Baidu[7] for WMT2019 and utilize the TransRedecoder model to generate a re-decoding machine translation based on the original machine translation. KSAI used the Transformer[8] as a baseline system, trained the model with 24.22 M pairs of sentences, and then used data filtering, fine-tuning, back translation, model enhancement, model integration, and reordering techniques to improve the translation quality. Baidu used the big Transformer[8] as a baseline system. Baidu trained the model with 15.7 M pairs of sentences in the en-zh and zh-en directions, and back translation, joint training, knowledge distillation, fine-tuning, model integration and reordering technology were also used to improve the machine translation quality.

Table 5. Results of original and re-decoding machine translation for different machine translations of CCMT2020 dev sets. MT_O means the original machine translation and MT_R means the re-decoding machine translation.

	Transformer				KSAI				Baidu			
	en-zh		zh-en		en-zh		zh-en		en-zh		zh-en	
	BLEU	NIST	BLEU	NIST	BLEU	NIST	BLEU	NIST	BLEU	NIST	BLEU	NIST
MT_O	31.52	7.84	25.15	7.02	42.42	9.14	40.25	9.09	42.49	9.22	40.95	9.21
MT_R	32.78	7.99	26.51	7.18	42.61	9.15	40.79	9.13	42.65	9.24	41.45	9.25
Δ	1.26	0.15	1.36	0.16	0.19	0.01	0.54	0.04	0.16	0.02	0.5	0.04

The re-decoding experimental results are shown in Table 5. In the en-zh direction, the TransRedecoder model increases the BLEU score by 0.19 and 0.16, respectively, and the NIST score by 0.01 and 0.02, respectively. In the zh-en direction, the TransRedecoder model significantly improves the BLEU scores by 0.54 and 0.50, respectively, and both improve the NIST scores by 0.04. Although the Baidu/ KSAI submitted systems achieved the best translation performance for the WMT19 test sets, the results in Table 5 show there is room for improvement. The novel re-decoding-based neural machine translation model, TransRedecoder, improves upon the quality of the machine translation.

3.3 Analysis

The re-decoding-based neural machine translation method is validated by the results in Table 6, which demonstrate an example of the original machine translation and re-decoding machine translation generated by the TransRedecoder model for the en-zh and zh-en directional dev sets of the CCMT2020 machine translation evaluation subtask. A comparison with the human reference translation clearly shows that the TransRedecoder model can effectively correct inaccurate target words in the machine translation. In the en-zh direction, the future context of “禁止” is utilized to generate the re-decoding words “拒绝”, which is better with “ 提供 庇护” than “拒绝”; in the zh-en direction; “hope” is re-decoded by combining the future context of “meeting”, and the re-decoded output “intension” is a better translation of the source word “会谈”. The results demonstrate that the proposed TransRedecoder model can effectively utilize contextual information from the original machine translation to improve the quality of the re-decoding machine translation of the target language.

Table 6. Original machine translation and re-decoding machine translation.

Source	eighteen states and the district of columbia are supporting a legal challenge to a new u.s. policy that denies asylum to victims fleeing gang or domestic violence .
Reference	美国 18 个 州 和 哥伦比亚 特区 支持 对 一 项 新 政策 发起 法律 挑战 ， 这 项 政策 拒绝 向 逃离 帮派 或 家庭 暴力 的 受害者 提供 庇护 。
MT	十八 个 州 和 哥伦比亚 特区 正 支持 对 美国 的 一 项 新 政策 的 法律 挑战 ， 这 项 政策 禁止 为 逃避 帮派 或 家庭 暴力 的 受害者 提供 庇护 。
Re-decoding MT	18 个 州 和 哥伦比亚 特区 政府 支持 对 美国 的 一 项 新 政策 的 法律 挑战 ， 这 项 政策 拒绝 向 逃离 帮派 或 家庭 暴力 的 受害者 提供 庇护 。
Source	同时 他 也 表达 了 希望 与 玉城 会谈 的 意向 。
Reference	and he also expressed his intention to talk with tamaki .
MT	at the same time , he also expressed the hope of meeting with yucheng .
Re-decoding MT	at the same time , he also expressed his intention of talks with yucheng .

4 Conclusions

This paper is a description of a technical report from Jiangxi Normal University on CCMT2020 sentence-level translation quality estimation subtasks and machine translation evaluation subtasks. We propose a simple and effective unified neural network model based on the Transformer model to effectively improve the performance of sentence-level translation quality estimation, as well as propose a TransRedecoder model that employs the Mask-CURRENT attention matrix to use the context of the original machine translation to increase the quality of the machine translation.

Acknowledgements

This research has been funded by the Natural Science Foundation of China under Grant No.61662031 and 61462044. The authors would like to extend their sincere thanks to the anonymous reviewers who provided valuable comments.

References

1. Chen, Z., Tan, Y., Zhang, C., Xiang, Q., Zhang, L., Li, M., Wang, M.: Improving machine translation quality estimation with neural network features. In: Proceedings of the WMT. pp. 551–555 (2017)
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the NAACL. pp. 4171–4186 (2018)

3. Guo, X., Liu, C., Li, X., Wang, Y., Li, G., Wang, F., Xu, Z., Yang, L., Ma, L., Li, C.: Kingsoft's neural machine translation system for wmt19. In: Proceeding of the ACL. pp. 196–202 (2019)
4. Li, M., Xiang, Q., Chen, Z., Wang, M.: A unified neural network for quality estimation of machine translation. IEICE Transactions on Information and Systems **101**(9), 2417–2421 (2018)
5. Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., Auli, M.: fairseq: A fast, extensible toolkit for sequence modeling. In: Proceedings of the NAACL. pp. 48–53 (2019)
6. Specia, L., Shah, K., De Souza, J.G., Cohn, T.: Quest-a translation quality estimation framework. In: Proceedings of the ACL. pp. 79–84 (2013)
7. Sun, M., Jiang, B., Xiong, H., He, Z., Wu, H., Wang, H.: Baidu neural machine translation systems for wmt19. In: Proceeding of the ACL. pp. 374–381 (2019)
8. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Proceedings of the NIPS. pp. 5998–6008 (2017)
9. Yu, D., Seltzer, M.L.: Improved bottleneck features using pretrained deep neural networks. In: Proceedings of the INTERSPEECH. pp. 237–240 (2011)
10. 李亚超, 熊德意, 张民: 神经机器翻译综述. 计算机学报 **41**(12), 100–121 (2018)
11. 李培芸, 李茂西, 裘白莲, 王明文: 融合 bert 语境词向量的译文质量估计方法研究. 中文信息学报 **34**(3), 56–63