Research Article

Hongxia Li* and Xin Tuo

Research on an English translation method based on an improved transformer model

https://doi.org/10.1515/jisys-2022-0038
received January 11, 2022; accepted March 22, 2022

Abstract: With the expansion of people's needs, the translation performance of traditional models is increasingly unable to meet current demands. This article mainly studied the Transformer model. First, the structure and principle of the Transformer model were briefly introduced. Then, the model was improved by a generative adversarial network (GAN) to improve the translation effect of the model. Finally, experiments were carried out on the linguistic data consortium (LDC) dataset. It was found that the average Bilingual Evaluation Understudy (BLEU) value of the improved Transformer model improved by 0.49, and the average perplexity value reduced by 10.06 compared with the Transformer model, but the computation speed was not greatly affected. The translation results of the two example sentences showed that the translation of the improved Transformer model was closer to the results of human translation. The experimental results verify that the improved Transformer model can improve the translation quality and be further promoted and applied in practice to further improve the English translation and meet application needs in real life.

Keywords: English translation, Transformer model, generative adversarial network, perplexity value

1 Introduction

With the progress and development of society, the need for cross-language communication has increased [1]; therefore, translation between different languages has become particularly important. Traditionally, translation is done by humans [2], with high accuracy; however, with the development of globalization, the speed and cost of human translation cannot meet the current demand, so machine translation (MT) has emerged [3]. MT is a technology that enables the interconversion of different languages through computers, with high speed and low cost, and has been well used in many large-scale translation scenarios. To better serve the society, improving the quality of MT has become a very important issue nowadays. Lee et al. [4] introduced a neural machine translation (NMT) model that mapped the source character sequence to the target character sequence without any segmentation and used a character-level convolutional network with maximum pooling in the encoder part. By experimenting on a many-to-one translation task, the model was found to have high translation quality. Wu et al. [5] improved the NMT model using source and target dependency trees. The new encoder enriched each source state with a dependency relationship in the tree. During decoding, the tree structure was used as a context to facilitate word generation. The experiment found that the model was found to be effective in improving translation quality. Choi et al. [6] contextualized the word embedding vectors using a nonlinear bag-of-words representation of the source sentence and represented special tokens with typed symbols to facilitate translation of the words that are less

* **Corresponding author: Hongxia Li,** Xi'an Innovation College, Yan'an University, Yan'an, Shaanxi 716000, China, e-mail: a6h1x5@yeah.net
Xin Tuo: Xi'an Innovation College, Yan'an University, Yan'an, Shaanxi 716000, China

suitable for translation through continuous vectors. The experimental results of En-Fr and En-De demonstrated the effectiveness of the model in improving the translation quality. Hewavitharana and Vogel [7] proposed a phrase alignment method that aims to align parallel sections bypassing nonparallel sections of a sentence and verified the effectiveness of the method in translation systems for Arabic English and Urdu English, which resulted in improvements up to 1.2 Bilingual Evaluation Understudy (BLEU) over the baseline. The current NMT has some problems, such as over-translation and under-translation, which lead to poor and ineffective translation, and further improvement and research are still needed. Therefore, this work studied the current mainstream Transformer model and innovatively improved it by combining it with a generative adversarial network (GAN). In the experiments, the reliability of the improved model in Chinese-English translation was verified by comparing the traditional Transformer model with the improved Transformer model. It was found that the translation results of the improved Transformer model were closer to the semantics of the source language and had smaller differences with the reference translations. The improved Transformer model is conducive to further improving the translation quality and effect and the better application of the Transformer model in Chinese-English translation.

2 Transformer model

With the development of artificial intelligence technology, MT has gradually developed from the earliest rule-based MT [8] to the early statistical MT [9], and the more common one now is NMT [10], which is mainly based on an "encoder-decoder" framework. NMT uses an encoder to map the source language sentence to a computable semantic vector and uses a decoder to decode the semantic vector to generate the target language sentence. Improving the translation effect of NMT is a key and difficult content in the current research. This work mainly studied the Transformer model.

Compared with the traditional NMT model, the Transformer model [11] completely abandons the recurrent neural network (RNN) structure and uses only the Attention mechanism to implement MT [12], which is good for reducing the computation and improving the translation effect. In the Transformer model, for an input $(x_1, x_2,..., x_n)$, it is mapped to $\vec{z} = (z_1, z_2, \cdots z_n)$ through an encoder, and an output sequence $(y_1, y_2,..., y_n)$ is generated through a decoder. The overall structure of the model is shown in Figure 1.

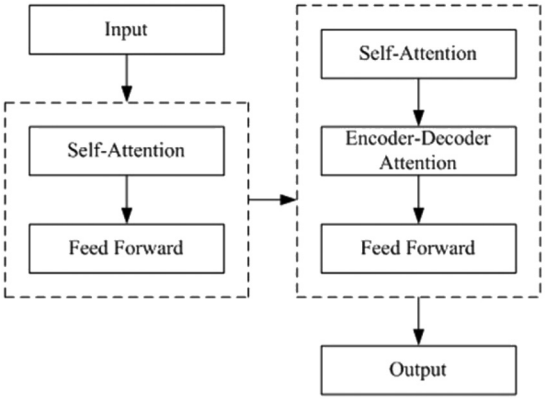


Figure 1: The structure of the Transformer model.

"Attention" in the Transformer model refers to Scale Dot-Product Attention. Let the dimension of the input query be q and key be d_k and the dimension of value be d_v. Query, keys, and value are processed into Q, K, and V. The output matrix can be written as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{Q^T k}{\sqrt{d_k}}\right), \tag{1}$$

$Q \in R^{m\times d_k}$, $K \in R^{m\times d_k}$, and $V \in R^{m\times d_v}$. The dimension of the output matrix is $R^{m\times d_v}$.

Multi-head attention processing is used in the Transformer model. First, a linear mapping is performed on Q, K, and $V \cdot Q$, K and V matrices whose input dimension is d_{model} is mapped to $Q \in R^{m\times d_k}$, $K \in R^{m\times d_k}$, and $V \in R^{m\times d_v}$. Then, the result is calculated through Scale Dot-Product Attention. The above steps are repeated. Attentions obtained through h times of operations are put together to obtain multi-head attention. The detailed calculation formula is:

$$\text{MultiHeadAttention}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, \cdots, \text{head}_h), \tag{2}$$

$$\text{Head}_i = \text{Attention}(QW_i^Q, KW_i^k, VW_i^V), \tag{3}$$

where W_i^Q, W_i^k, and W_i^V are parameter matrices.

Since the encoder and decoder of the model cannot capture the sequence order information, the position encoding is used in the Transformer model, which can be written as:

$$\text{PE}_{(\text{pos},2i)} = \sin(\text{pos}/10,000^{2i/d_{\text{model}}}), \tag{4}$$

$$\text{PE}_{(\text{pos},2i+1)} = \cos(\text{pos}/10,000^{2i/d_{\text{model}}}), \tag{5}$$

where pos refers to the position of a word in a sentence and i is the dimensional subscript of a word.

3 The improved Transformer model

In order to further improve the performance of the Transformer model, this paper improved it by combining GAN, which is based on the Nash equilibrium of game theory [13] and has extensive applications in image processing [14] and natural language processing [15]. GAN obtains the desired data through confrontation and gaming of a generator (G) and a discriminator (D) [16], using the back-propagation of D as the parameter update of G, thus enabling the model to learn what kind of utterance is considered a good translation.

The improved Transformer model uses the Transformer model as the generator (G) and the convolutional neural network (CNN) [17] as the discriminator (D). The goal of G is to generate a sequence from the initial state that maximizes the final desired reward, written as:

$$J(\theta) = \sum_{Y_{1:T}} G_\theta(Y_{1:T}|X) R_{D,Q}^{G_\theta}(Y_{1:T} \cdot X, y_T \cdot Y^\star), \tag{6}$$

where θ refers to the parameter in the generator G, $Y_{1:T}$ refers to the generated target sentence, X refers to the source sentence, $Y^\star$ refers to the given standard target sentence, and $R_{D,Q}^{G_\theta}$ refers to the action-value function from source sentence X to target sequence.

The BLEU value of a sentence is used as a generator, and the n-element syntactic precision of sentence y_g is computed. Based on the target standard sentence y_d, reward $Q(y_g, y_d)$ is calculated. The calculation formula of $R_{D,Q}^{G_\theta}$ is:

$$R_{D,Q}^{G_\theta}(Y_{1:T} \cdot X, y_T \cdot Y^\star) = \gamma(D(X, Y_{1:T}) - b(X, Y_{1:T})) + (1 - \gamma)Q(Y_{1:T}, Y^\star), \tag{7}$$

where $b(X, Y)$ stands for the base value, which is set as 0.5 for simple calculation, and γ stands for a hyperparameter. To obtain a stable reward, an N-time Monte Carlo search is used to obtain the action-value, and the formula is:

$$\{Y_{1:T_1}^1, \cdots, Y_{1:T_N}^N\} = \text{MC}^{G_\theta}((Y_{1:T}, X), N), \tag{8}$$

where T_i refers to the length of the sequence at the ith Monte Carlo search, $(Y_{1:T}, X)$ refers to the present state, $Y_{1:T_N}^N$ refers to the sentence generated according to the policy G_θ, and MC refers to the Monte Carlo search function.

The generator is optimized by updating the discriminator used as the reward function, and the relevant calculation formula is:

$$\min(E_{X,Y\in P_{\text{data}}}[\lg(D(X, Y))] - E_{X,Y\in G}[\lg(1 - D(X, Y))]), \tag{9}$$

where E is an anticipation function. The derivative of the target function $J(\theta)$ to the generator parameter θ is:

$$c_c \pi \nabla J(\theta) = \frac{1}{T}\sum_{t=1}^{T}\sum_{y_t} R_{D,Q}^{G_\theta}(Y_{1:T}, X, y_T, Y^\star)\cdot\nabla_\theta(G_\theta(y_t|Y_{1:T}, X)), \tag{10}$$

The generator parameter is updated as follows:

$$\theta \leftarrow \theta + a_h \nabla_\theta J(\theta), \tag{11}$$

where a_h refers to the learning rate at the h step.

4 Experimental analysis

The previous sections introduced two NMT models: the traditional Transformer model and the improved Transformer model combined with GAN. In this section, the experimental setup was described, the evaluation indicators of the model performance were introduced, the experiments were conducted on the linguistic data consortium (LDC) dataset, and the experimental results were analyzed in detail.

4.1 Experimental setup

The baseline system was the Transformer model in the open-source framework THUMT from Tsinghua University, whose parameters are shown in Table 1.

Table 1: Parameters of the Transformer model

Parameters	Value
Number of layers in the encoder	6
Number of layers in the decoder	6
Size of Chinese and English word lists	32k
Word vector dimension	512
The hidden layer state dimension of the feedforward neural network	2,048
The number of heads in multi-head attention	8
Dropout ratio	0.1
Number of words in each batch	6,250
The largest number of words in a sentence	50

The experimental task was a Chinese-English translation task on the LDC data set. NIST06 was used as the development set, and NIST02, NIST03, NIST04, NIST05, and NIST08 were used as the test sets to compare the performance of the Transformer model with the improved Transformer model.

4.2 Evaluation criteria

BLEU [18]: the more similar the results of model translation and human translation were, the better the performance of the model was. The calculation method of BLEU is shown below.

(1) The maximum number of possible occurrences of an n-gram word in the reference translation, i.e., maxrefcount(n-gram) was calculated. Then, the number of occurrences of the n-gram word in the translation results of the model, i.e., Count(n-gram), was calculated. The smaller number was taken as the final matching times, and the relevant calculation formula is:

$$\text{Count}_{\text{clip}}(n\text{-gram}) = \min\{\text{Count}(n\text{-gram}),\text{maxrefcount}(n\text{-gram})\}. \tag{12}$$

(2) After obtaining $\text{Count}_{\text{clip}}(n\text{-gram})$, the BLEU value was calculated, and the formula is:

$$\text{BLEU} = \text{BP} \times \exp\left[\left(\sum_{n=1}^{N} w_n \log p_n\right)\right], \tag{13}$$

$$p_n = \frac{\sum_{C\in\{\text{candidates}\}}\sum_{n\text{-gram}\in C}\text{Count}_{\text{clip}}(n\text{-gram})}{\sum_{C'\in\{\text{candidates}\}}\sum_{n\text{-gram}'\in C'}\text{Count}_{\text{clip}}(n\text{-gram}')}, \tag{14}$$

$$\text{BP} = \begin{cases} 1, & \text{if } c > r, \\ e^{(1-r/c)}, & \text{if } c \le r, \end{cases} \tag{15}$$

where BP refers to a penalty factor, w_n refers to the weight of n-gram word, p_n refers to the score of precision, c refers to the length of the target text obtained by the mode, and r refers to the length of the target text for reference.

Perplexity [19]: it is one of the criteria for testing the performance of a model. In translation results of the model, the larger the probability of every word was, the more accurate the word was. The trained model was tested to obtain the final translation result. The probability was calculated. If there are N words, the calculation formula of perplexity is:

$$\text{PP}(T) = p(w_1 w_2 \cdots w_N)^{1/N} = \sqrt[N]{\prod_{n=1}^{N}\frac{1}{p(w_n|w_1 w_2 \cdots w_{n-1})}}, \tag{16}$$

where $p(w_i)$ stands for the translation probability of word w_i.

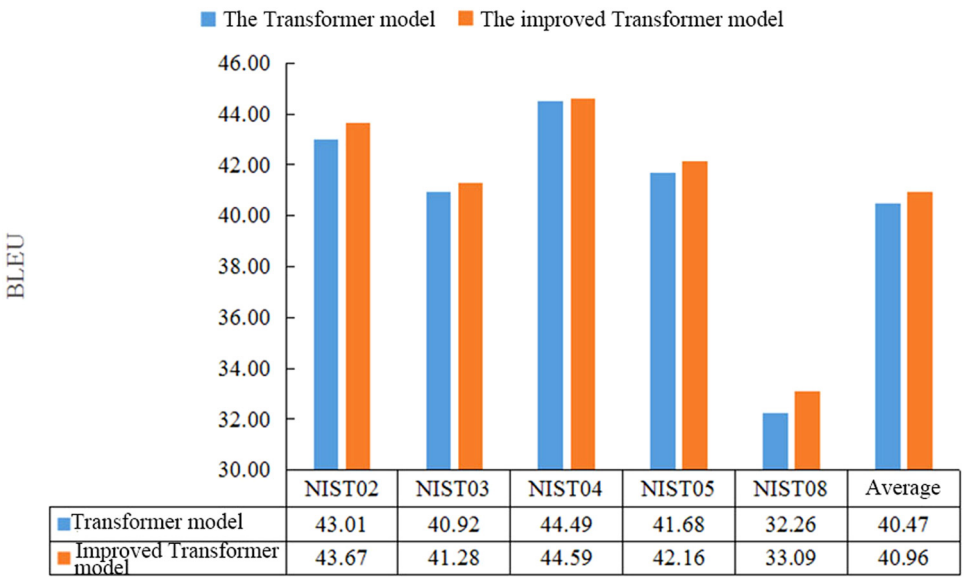
4.3 Translation results

The BLEU values of the two models are shown in Figure 2.

It was seen from Figure 2 that the BLEU value of the improved Transformer model had some improvement; the BLEU value of the improved Transformer model improved by 0.66 in NIST02, 0.36 in NIST03, 0.1 in NIST04, and 0.48 in NIST05, and NIST08 improved the most. The BLEU value of the dataset was 32.26 in the Transformer model and 33.09 in the improved Transformer model, i.e., the value improved by 0.83. The average BLEU value of the Transformer model was 40.47, while the average BLEU value of the improved Transformer model was 40.96, i.e., there was an improvement of 0.49. The above results verified the reliability of the improved Transformer model in improving the translation quality.

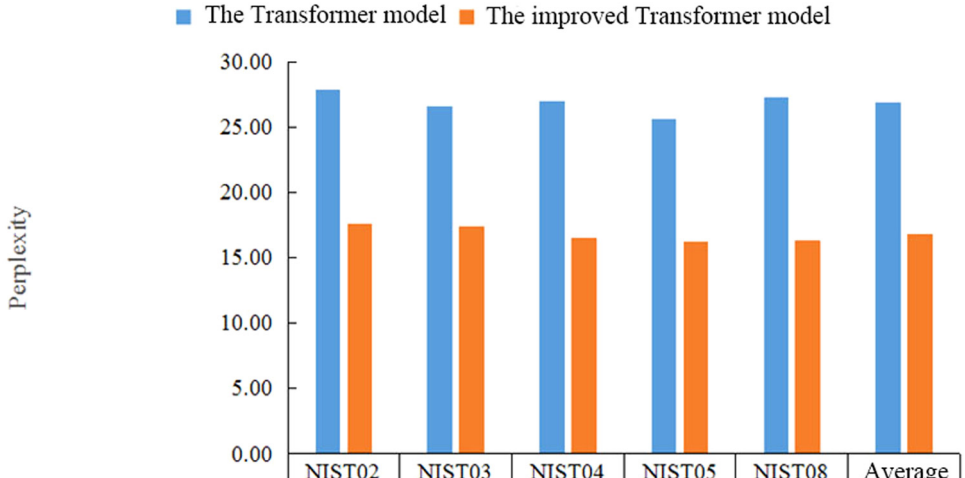
Then, the perplexity values of the two models were compared, and the results are shown in Figure 3.

It was seen from Figure 3 that the perplexity value of the improved Transformer model was significantly reduced. Specifically, when using the improved Transformer model, the perplexity values of NIST02 reduced by 10.27, NIST03 by 9.16, NIST04 by 10.47, NIST05 by 9.45, and NIST08 by 10.98. The average perplexity value of the Transformer model was 26.89, while the average perplexity value of the improved Transformer model was 16.83, which was reduced by 10.06. The above results indicated that the probability



	NIST02	NIST03	NIST04	NIST05	NIST08	Average
Transformer model	43.01	40.92	44.49	41.68	32.26	40.47
Improved Transformer model	43.67	41.28	44.59	42.16	33.09	40.96

Figure 2: Comparison of BLEU values between different models.



	NIST02	NIST03	NIST04	NIST05	NIST08	Average
Transformer model	27.89	26.58	27.01	25.66	27.32	26.89
Improved Transformer model	17.62	17.42	16.54	16.21	16.34	16.83

Figure 3: Comparison of perplexity values between different models.

of the model selecting the correct translation results somewhat increased, thus the model obtained results closer to the human translation.

The computational speed of the two models was compared, and the results are shown in Figure 4.

It was seen from Figure 4 that the computational speed of the Transformer model was 27,850 words per second, while the computational speed of the improved Transformer model was 25,890 words per second, which was only 7.04% lower than that of the Transformer model, indicating that the improvement of the Transformer model by combining GAN did not have a particularly significant impact on its computational speed.

The translation results of the two models were compared and analyzed, as shown in Tables 2 and 3.

It was seen from Table 2 that the Transformer model translated “欧盟办事处” as “The European Union offices,” while the improved Transformer model translated it as “The EU office,” which was more accurate.

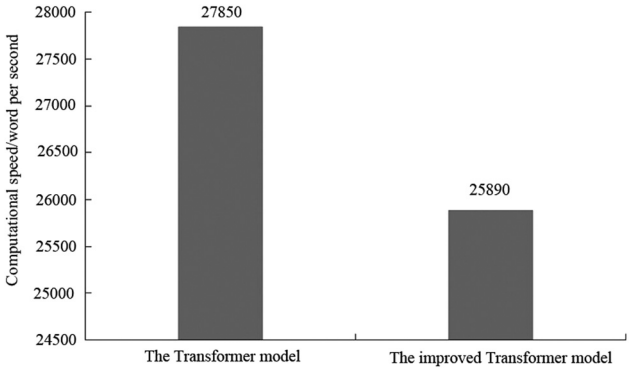


Figure 4: Comparison of computing speed between different models.

Table 2: Translation example 1

Source language	欧盟办事处与澳洲大使馆在同一建筑内。
Candidate translation 1	The EU mission is in the same building with the Australian embassy
Candidate translation 2	The European Union office and the Australian embassy are both located in the same building
Candidate translation 3	The European Union office is in the same building as the Australian embassy
Candidate translation 4	The EU office and the Australian embassy are housed in the same building
The translation of the Transformer model	The European Union offices with the Australian embassy in the same building
The translation of the improved Transformer model	The EU office is housed in the same building as the Australian Embassy

Table 3: Translation example 2

Source language	经过处理后的“中水”将率先在城市绿化浇灌中使用。
Candidate translation 1	The treated reclaimed water will first be used in city greenbelt irrigation
Candidate translation 2	The treated reclaimed water will be first used to irrigate urban greenery
Candidate translation 3	The treated middle-water will first be used in watering the trees in and around the city
Candidate translation 4	The treated reclaimed water will be first used in urban green area irrigation
The translation of the Transformer model	The treated intermediate water will be the first to be used in urban green irrigation
The translation of the improved Transformer model	The treated reclaimed water will be first used in urban green area irrigation

The reason for the above problem may be that the Transformer model did not consider “The EU office” as a proper noun, leading to a wrong translation.

In Chinese semantic, the term “中水” means recycled water. It was seen from Table 3 that the Transformer model translated “中水” directly as “intermediate water” without considering the specific semantics. In addition, the Transformer model translated “率先…” as “be the first to,” which was more inclined to the meaning of “the first time to,” the Transformer model translated it as “be first,” which was closer to the original meaning of the source language and the reference translations.

5 Discussion

Advances in artificial intelligence have led to the rapid development of NMT, which has become a new paradigm for MT [20]. Compared with statistical MT, NMT does not require steps such as word alignment and sequencing, and its translation process entirely relies on the self-learning of neural networks, which greatly reduces the complexity of the model and significantly improves translation performance. However, the current NMT has some problems, such as low-frequency words and unknown words [21]; therefore, the research on NMT is of great importance.

This article mainly focused on the Transformer model. In order to improve the translation effect, GAN was introduced to improve the Transformer model. Then, the translation performance of these two models was compared and analyzed on the LDC dataset. The comparison of BLEU and perplexity values showed that the improved Transformer model had larger BLEU values and smaller perplexity values, which indicated that the similarity between the results of model translation and human translation was high, i.e., the translation quality of the improved Transformer model was high. In addition, the comparison of the translation speed (Figure 4) suggested that the computational speed of the improved Transformer model only decreased by 7.04% compared with the traditional Transformer model, which indicated that the improved model did not increase the computational complexity to a great extent. Finally, the two translation examples (Tables 2 and 3) demonstrated that compared with the Transformer model, the translation results obtained by the improved Transformer model were more consistent with the meaning of the source language and matched better with the reference translations, which verified the reliability of its translation effect and its application feasibility in practice. At present, Chinese-English translation has a broad demand in many fields, and the improved Transformer model can provide simultaneous translation in multilingual communication in international conferences and provide services for cross-language retrieval in academic fields, which has great application values in scenarios such as foreign trade and overseas travel.

In this article, although some results have been obtained from the study of English translation based on the improved Transformer model, there are some shortcomings. For example, the amount of the experimental data was small and the translated sentences were short. In future work, experiments will be conducted on a larger dataset, and the performance of the improved Transformer model on English paragraph translation will be investigated.

6 Conclusion

This article improved the Transformer model by combining GAN and conducted experiments on the LDC dataset. The performance of the Transformer model and the improved Transformer model in English translation was compared. The results showed that:

1. The average BLEU value and perplexity value of the improved Transformer model were 40.96 and 16.83, respectively, which were superior to those of the Transformer model;
2. The computational speed of the improved Transformer model was 25,890 words, which only decreased by 7.04% compared to the Transformer model;
3. The translation results of the improved Transformer model were closer to the reference translations.

The experimental results verify that the improved Transformer model can improve translation quality effectively while ensuring computational speed. In order to realize better application in actual translation scenarios, the improved Transformer model will be further studied in the future to analyze its performance in translating long and complicated sentences and paragraphs.

Conflict of interest: Authors state no conflict of interest.

References

[1] Liu H, Zhang M, Fernández AP, Xie N, Li B, Liu Q. Role of language control during interbrain phase synchronization of cross-language communication. Neuropsychologia. 2019;131:316–24.
[2] Lumeras MA, Way A. On the complementarity between human translators and machine translation. Hermes. 2017;56:21.
[3] Liu HI, Chen WL. Re-transformer: a self-attention based model for machine translation. Proc Comput Sci. 2021;189:3–10.
[4] Lee J, Cho K, Hofmann T. Fully character-level neural machine translation without explicit segmentation. Trans Assoc Comput Linguist. 2017;5:365–78.
[5] Wu S, Zhang D, Zhang Z, Yang N, Li M, Zhou M. Dependency-to-dependency neural machine translation. IEEE/ACM T Audio Speech. 2018;26:2132–41.
[6] Choi H, Cho K, Bengio Y. Context-dependent word representation for neural machine translation. Comput Speech Lang. 2017;45:149–60.
[7] Hewavitharana S, Vogel S. Extracting parallel phrases from comparable data for machine translation. Nat Lang Eng. 2016;22:549–73.
[8] Sghaier MA, Zrigui M. Rule-based machine translation from tunisian dialect to modern standard Arabic. Proc Comput Sci. 2020;176:310–9.
[9] Germann U. Sampling phrase tables for the moses statistical machine translation system. Prague Bull Math Linguist. 2015;104:39–50.
[10] Luong MT. Addressing the rare word problem in neural machine translation. Bull Univ Agric Sci Vet Med Cluj-Napoca. 2015;27:82–6.
[11] Popel M, Bojar O. Training tips for the transformer model. Prague Bull Math Linguist. 2018;110:43–70.
[12] Lin F, Zhang C, Liu S, Ma H. A hierarchical structured multi-head attention network for multi-turn response generation. IEEE Access. 2020;8:46802–10.
[13] Wang HG, Li X, Zhang T. Generative adversarial network based novelty detection using minimized reconstruction error. Front Inf Tech El. 2018;01:119–28.
[14] Wolterink JM, Leiner T, Viergever MA, Isgum I. Generative adversarial networks for noise reduction in low-dose CT. IEEE T Med Imaging. 2017;36:2536–45.
[15] Creswell A, White T, Dumoulin V, Arulkumaran K, Sengupta B, Bharath AA. Generative adversarial networks: an overview. IEEE Signal Proc Mag. 2017;35:53–65.
[16] Wang KF, Gou C, Duan YJ, Lin YL, ZHeng XH, Wang FY. Generative adversarial networks: the state of the art and beyond. Acta Autom Sin. 2017;43:321–32.
[17] Ren Q, Su Y, Wu N. Research on Mongolian-Chinese machine translation based on the end-to-end neural network. Int J Wavel Multi. 2020;18:46–59.
[18] Shereen A, Mohamed A. A cascaded speech to Arabic sign language machine translator using adaptation. Int J Comput Appl. 2016;133:5–9.
[19] Brychcin T, Konopik M. Latent semantics in language models. Comput Speech Lang. 2015;33:88–108.
[20] Choi H, Cho K, Bengio Y. Fine-grained attention mechanism for neural machine translation. Neurocomputing. 2018;284:171–6.
[21] Hasigaowa, Wang S. Research on unknown words processing of Mongolian-Chinese neural machine translation based on semantic similarity. 2019 IEEE 4th International Conference on Computer and Communication Systems, ICCCS; 2019. p. 370–4.