Research Article

Mengyang Qin*

A study on automatic correction of English grammar errors based on deep learning

https://doi.org/10.1515/jisys-2022-0052
received February 24, 2022; accepted April 20, 2022

Abstract: Grammatical error correction (GEC) is an important element in language learning. In this article, based on deep learning, the application of the Transformer model in GEC was briefly introduced. Then, in order to improve the performance of the model on GEC, it was optimized by a generative adversarial network (GAN). Experiments were conducted on two data sets. It was found that the performance of the GAN-combined Transformer model was significantly improved compared to the Transformer model. The $F_{0.5}$ value of the optimized model was 53.87 on CoNIL-2014, which was 2.69 larger than the Transformer model; the generalized language evaluation understanding (GLEU) value of the optimized model was 61.77 on JFLEG, which was 8.81 larger than that of the Transformer model. The optimized model also had a favorable correction performance in an actual English essay. The experimental results verify the reliability of the GAN-combined Transformer model on automatic English GEC, suggesting that the model can be further promoted and applied in practice.

Keywords: deep learning, grammatical error correction, transformer model, English essay

1 Introduction

In English learning, grammar is one of the key elements [1]. In order to improve grammar, learners often do many writing exercises. In the specific teaching process, the correction of writing exercises takes much time and is burdensome for both learners and teachers; therefore, the efficiency of grammar learning can be significantly improved if automatic English grammatical error correction (GEC) can be implemented. With the improvement of technology, GEC has also been widely studied [2]. Solyman et al. [3] studied the Arabic GEC model, introduced a multi-convolutional layer model containing an attention mechanism, and found through experiments that the method obtained high accuracy. Park et al. [4] analyzed correction and over-correction problems in GEC and pointed out that the current GEC model might make unnecessary changes to correct sentences. Acheampong and Tian [5] improved the seq2seq model using a neural cascade strategy and found through experiments that the method was more effective in correcting grammatical errors in a low-resource model. Liu and Liu [6] suggested training the GEC model with unlabeled data, used an attention-based neural network, and verified the advantages of the method through experiments. Lin et al. [7] regarded GEC as a multi-classification task, integrated different language embedding models and deep learning models to correct ten part-of-speech errors in Indonesian texts, and found that the average $F_{0.5}$ of the model reached 0.551, suggesting a good performance. Zhou and Liu [8] established a basic model for GEC based on the classification model and found through experiments that the model could constantly improve accuracy and correction efficiency in the learning process. Facing with the problem of grammatical errors in Chinese, Wang et al. [9] established word vectors using the Bidirectional Encoder

* **Corresponding author: Mengyang Qin,** College of Tourism Management, Henan Vocational College of Agriculture, No. 38, Qingnianxi Road, Zhongmu County, Zhengzhou, Henan 451450, China, e-mail: uym7do@126.com

Representation from Transformers (BERT) model, designed and implemented the BERT BILSTM CRF-based Chinese grammatical error detection model, and found that the model was feasible and had a high accuracy. This article studied automatic English GEC based on deep learning, designed a method that combined the Transformer model with a Generative Adversarial Network (GAN), and verified the effectiveness of the method through experiments on data sets. This article makes a contribution to further improving GEC.

2 Deep learning models

Natural language has strong flexibility and uncertainty. English, as a language, has an extensive vocabulary and complex grammar. It is difficult to correct English grammar. Automatic English GEC [10] can not only reduce teachers' burden and save their resources but also help learners to get feedback on grammar learning faster. Deep learning has been successfully applied in image processing [11] and speech recognition [12], and its application in natural language processing (NLP) has been continuously researched [13].

Deep learning obtains more feature information for data classification and prediction through continuously learning from automatically extracted data features, including convolutional neural networks (CNN) [14] and recurrent neural networks (RNN) [15]. In NLP, RNN is one of the most common models. Since RNN can only implement sequential computation, its parallel computation capability is poor; thus, the Transformer model [16] has emerged. This article realized automatic English GEC based on the Transformer model.

The Transformer model follows the encoder–decoder framework. The encoder and decoder both consist of six identical layers stacked together. Every layer in the encoder includes two sub-layers, multi-head attention and feed-forward, and the decoder includes a sub-layer in addition to these two sub-layers. Multi-head attention is performed on the output of the encoder. The most important element in the model is the self-attention mechanism, which makes every word have three vectors: query, key, and value. The calculation formula of attention is written as:

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

where d_k is the dimensional value of the vector, 512.

With different attention heads, different information can be obtained. Then, these attention heads are combined to get the output of multi-head attention. The relevant calculation formulas are written as:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_n)W^o, \tag{2}$$

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

The formula for the feed-forward network (FFN) is written as:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2, \tag{4}$$

where W_1, W_2, b_1, and b_2 are training parameters.

The Transformer model implements the position encoding of the sequence through sin and cos functions. The calculation formulas are written as:

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

$$\text{PE}(\text{pos}, 2i + 1) = \cos(\text{pos}/10,000^{2i/d_{\text{model}}}), \tag{6}$$

where pos is the feature location and i is the word dimension.

Automatically correcting English grammatical errors with the Transformer model is considered as a translation task, i.e., "translating" grammatically incorrect sentences into grammatically correct ones. In the training process, maximum likelihood estimation is adopted to maximize the likelihood of the model on the training data S. The computational formula is written as:

$$\alpha = \text{argmax} \sum_{(x,y)\in S} \log p(y|x;\alpha). \tag{7}$$

In order to further improve the performance of the Transformer model on GEC, this article optimized the Transformer model with GAN.

3 GEC approach combined with GAN

GAN [17] consists of two independent networks: a generator and a discriminator. The training of a model is achieved by adversarial learning of the above two networks. The generator used in this article is the Transformer model, and the discriminator is a CNN-based classification model. An "error-corrected" sentence is written as (X, Y). Let the parameter of generator G be θ, the error sentence at the source side be

$$x = (x_1, x_2, \ldots, x_m), \tag{8}$$

the corrected sentence at the target end be

$$y = (y_1, y, \ldots, y_n), \tag{9}$$

and the corrected sentence generated by the generator be

$$y' = (y_1', y', \ldots, y_n'). \tag{10}$$

When the generator outputs a corrected sentence, it is used as input to the discriminator along with the incorrect sentence at the source side, and the discriminator calculates the probability that the sentence is manually labeled and feeds it back to the generator as a reward. The goal of the whole adversarial learning is to obtain the maximum desired reward. The process of GAN adversarial training is as follows:

(1) The generator is pre-trained on (X, Y) using maximum likelihood estimation.
(2) Taking (X, Y) as the positive sample, a negative sample (X, Y') is established using the generator to pre-train the discriminator.
(3) Generator updating: subset $(X_{\text{batch}}, Y_{\text{batch}})$ is sampled from (X, Y). Source-side error sentence X_{batch} is sampled by the generator to obtainY'_{batch}. Then, a Monte Carlo search is performed on every position of Y'_{batch} to calculate the respective reward values. Then, the generator is updated through the policy gradient method.
(4) Discriminator updating: subset $(X_{\text{batch}}, Y_{\text{batch}})$ is sampled from (X, Y), and negative sample $(X_{\text{batch}}, Y'_{\text{batch}})$ is established to pre-train the discriminator.

4 Experimental analysis

4.1 Experimental setup

The Transformer model was realized by the open-source tensor2tensor. There were eight heads in the multi-headed attention. As to the FNN, the dimension of the hidden layer was 2,048, the gelu function was used as the activation function, dropout was 0.1, and lr was 0.01. The Adam optimizer was used. The batch_size was 20, and the epoch number was 40. The maximum length of the sentence was 50. The beam size was 8. The length penalty parameter was 0.6. The training ended when the model had no effect improvement for three consecutive epochs on the development set. In CNN, the word vector dimension was 300, the size of the convolutional window was 3×3, the pooling window was 2×2, the first convolutional layer was mapped with 128 features, the second convolutional layer was mapped with 256 features, and the dimension of the hidden layer was 128. For adversarial training, the RMSprop algorithm was used [18], the initial

learning rate was set as 0.0003, and the batch_size was 128. When training the discriminator, 5,000 samples were randomly sampled as positive samples, and the corresponding negative samples were created.

4.2 Experimental data set

Training sets: ① Lang-8: it is a corpus established by extracting from the social network Lang-8 and has multilingual versions, as shown in Table 1. The English version is extracted as the training set.

② CLC FCE: It includes 1,244 exam scripts written based on test papers of candidates who participated

Table 1: Lang-8 data set

Language	Quantity
English	1,069,549
Japanese	925,588
Mandarin	136,203
Korea	93,955
Spanish	51,829
French	58,918
German	37,886

in exams of English for Speakers of Other Languages, containing original texts, labels, error comments, etc. There are 80 types of grammatical errors and 1.36 million data.

③ NUCLE: It includes about 1,400 papers written by National University of Singapore undergraduates, which have been annotated and corrected by professional English teachers. The percentage of incorrect sentences in the corpus is 42.4%, involving 28 types of grammatical errors.

Test sets: ① CoNIL-2014: It is a standard data set of GEC, which includes 1,312 sentences and 28 types of grammatical errors. The evaluation index is $F_{0.5}$.

② JFLEG: It is a common data set of GEC, which can help correct grammatical errors and produce more fluent language. It includes 747 difficult sentences. The evaluation indexes are GLEU and $F_{0.5}$.

4.3 Evaluation indicators

(1) $F_{0.5}$: F refers to the weighted harmonic mean of precision and recall rate; $F_{0.5}$ refers to that the precision is twice as important as recall rate. The samples are divided according to the confusion matrix shown in Table 2.

Then, the precision (P) of the model is

Table 2: Confusion matrix

		Model output results	
		Positive sample	Negative sample
Manual labeling results	Positive sample	TP	TN
	Negative sample	FP	FN

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{11}$$

and the recall rate (R) is

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{12}$$

The calculation formula of $F_{0.5}$ is

$$F_{0.5} = \frac{(0.5^2 + 1) \times P \times R}{0.5^2 \times P \times R}. \tag{13}$$

(2) MaxMatch (M^2): In GEC, an error sentence may have multiple results after corrections. If every answer is considered as a branch of a node, then the whole process of error correction can be regarded as a connected graph, and the evaluation of the error correction result can be regarded as the evaluation of the graph, i.e., scoring the edge of the node that can reach the final correct answer. For source-side sentence:

$$S = \{s_1, s_2, \ldots, s_n\}, \tag{14}$$

let the corrected sentence output by the model be

$$E = \{e_1, e_2, \ldots, e_n\}, \tag{15}$$

and the manually labeled sentence be

$$G = \{g_1, g_2, \ldots, g_n\}. \tag{16}$$

The specific calculation formulas are as follows:

$$P = \frac{\sum_{i=1}^{n} |e_i \cap g_i|}{\sum_{i=1}^{n} |e_i|}, \tag{17}$$

$$R = \frac{\sum_{i=1}^{n} |e_i \cap g_i|}{\sum_{i=1}^{n} |g_i|}, \tag{18}$$

$$F_{0.5} = \frac{(0.5^2 + 1) \times P \times R}{0.5^2 \times P \times R}. \tag{19}$$

(3) GLEU: In a translation task, BLEU is based on the similarity between sentences. The specific calculation formulas are as follows:

$$\mathrm{BLEU} = \mathrm{BP} \times \exp\left(\sum_{n=1}^{N} w_n \times \log P_n\right). \tag{20}$$

$$\mathrm{BP} = \begin{cases} 1, & \text{if} \quad c > r, \\ \mathrm{e}^{\frac{1-r}{c}}, & \text{if} \quad c \le r, \end{cases} \tag{21}$$

where P_n is the precision rate of n-gram level words, BP is the length parameter, and c and r are the lengths of reference translation and machine translation, respectively. However, unlike machine translation, in the GEC task, an untranslated word may not be wrong, and changes can be made to the words in the source sentence; therefore, in order to evaluate GEC, BLEU is modified to GLEU, and the specific calculation formulas are as follows:

$$\mathrm{Count}_B(n\text{-gram}) = \sum_{n\text{-gram}' \in B} d(n\text{-gram}, n\text{-gram}'), \tag{22}$$

$$\mathrm{GLEU}(C, R, S) = P \times \exp\left(\sum_{n=1}^{N} w_n \times \log P'_n\right), \tag{23}$$

$$P'_n = \frac{\sum_{n\text{-gram}\in G}\mathrm{Count}_{R\setminus S}(n\text{-gram}) - \gamma(\mathrm{Count}_{R\setminus S}(n\text{-gram})) + \mathrm{Count}_R(n\text{-gram})}{\sum_{n\text{-gram}'\in G'}\mathrm{Count}_S(n\text{-gram}') + \sum_{n\text{-gram}\in R\setminus S}\mathrm{Count}_{R\setminus S}(n\text{-gram})}, \tag{24}$$

where C, R, and S refer to the candidate set, reference set, and source-end sentences, P'_n is the accuracy rate after correction, and γ is the penalty rate, which is used to penalize incorrect answers that appear in the source sentence but not in the reference set.

4.4 Experimental results

The experimental results on CoNIL-2014 are shown in Figure 1.

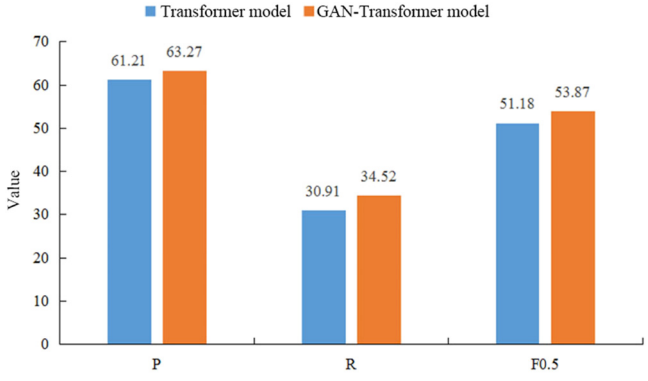

Figure 1: GEC results of the model on CoNIL-2014.

It is seen from Figure 1 that the P-value, R-value, and $F_{0.5}$ of the basic Transformer model were 61.21, 30.91, and 51.18, respectively; the P-value of the GAN-combined Transformer model was 63.27, which was 2.06 larger than that of the Transformer model; the R-value of the optimized model was 34.52, which was 3.61 larger than that of the Transformer model; and the $F_{0.5}$ of the optimized model was 53.87, which was 2.69 larger than that of the Transformer model. The experimental results showed that the performance of the Transformer model on the GEC task was improved to some extent after combining GAN, which suggested that the improvement was effective.

The experimental results on JFLEG are shown in Figure 2.

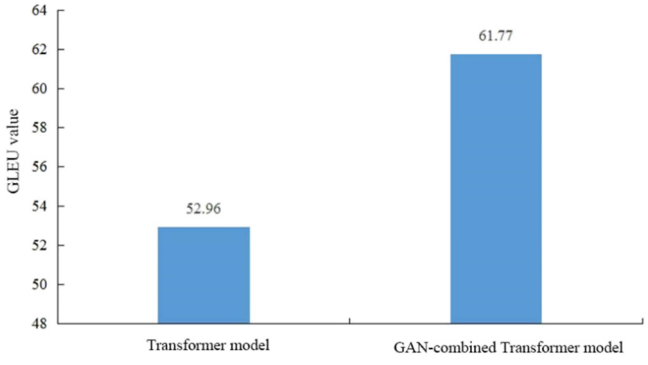

Figure 2: GEC results of the model on JFLEG.

It is seen from Figure 2 that on the JFLEG data set, the GLEU value of the Transformer model was 52.96, and the GLEU value of the GAN-combined Transformer model was 61.77, which was 8.81 larger than the Transformer model, indicating that the similarity between the results obtained by the GAN-combined Transformer model and the manually annotated reference set was higher. The results suggested that the GAN-combined Transformer model was superior.

In order to further understand the application possibilities of the GAN-combined Transformer model, a student's English essay was used as an example. The grammatical errors in the essay were automatically corrected using the model designed in this article, and the original text is as follows.

Technology is all around us, changing the way our live. The emergence of ships, trains, planes, etc. has made transportation more convenience. No matter where you want to go, as long as you find a suitable means of transportation, you can get to there quickly. On the same time, the development of science and technology has promoted the birth of computer, which has further narrowed the distant between people, allowing people to communicate to distant friends without leaving home, and can also realize online shopping, entertainment, work and so on. The development of technology have enabled us to live a better life.

The correction results of the model are shown in Table 3.

Table 3: English essay correction example

Source-side sentence	Mode output
Technology is all around us, changing the way **our** live	Technology is all around us, changing the way **we** live
The emergence of ships, trains, planes, etc. has made transportation more **convenience**	The emergence of ships, trains, planes, etc. has made transportation more **convenient**
No matter where you want to go, as long as you find a suitable means of transportation, you can get **to** there quickly	No matter where you want to go, as long as you find a suitable means of transportation, you can get **to** there quickly
On the same time, the development of science and technology has promoted the birth of **computer**	**At** the same time, the development of science and technology has promoted the birth of **computers**
allowing people to communicate **to** distant friends without leaving home	allowing people to communicate **with** distant friends without leaving home
The development of technology **have** enabled us to live a better life	The development of technology **has** enabled us to live a better life

It is seen from Table 3 that the GAN-combined Transformer model corrected tense errors, pronoun errors, and singular and plural errors, but some errors were not detected, for example, in the sentence "which has further narrowed the distant between people," the word "distant" should be corrected to "distance." The above results suggested that the GAN-combined Transformer model had some shortcomings compared with manual annotation, which need to be improved in future research to further improve its performance in GEC.

5 Discussion

With the development of globalization, the number of English learners is also increasing, which puts pressure on English teaching. Grammar, as an important part of English learning, requires a lot of practice. Relying entirely on teachers' manual corrections for grammar will consume a lot of time and energy. Therefore, automatic GEC has received more and more research. Deep learning has a wide range of applications in machine translation. This article designed an automatic GEC method based on deep learning.

First, on the CoNIL-2014 data set, the P-value, R-value, and $F_{0.5}$ of the traditional Transformer model were 61.21, 30.91, and 51.18, respectively, while the GAN-combined Transformer model had a P-value of

63.27, which was 3.32% higher than the Transformer model; a R-value of 34.52, which was 11.7% higher than the Transformer model; and a $F_{0.5}$ value of 53.87, which was 5.26% higher than the Transformer model. It was concluded that the GAN-combined Transformer model had significantly improved translation performance, indicating that the addition of adversarial learning greatly improved the performance of the model for GEC. Then, on the JFLEG data set, the GLEU value of the GAN-combined Transformer model was 16.64% higher than the Transformer model (61.77 vs 52.96). The above results suggested that the GAN-combined Transformer model had a better performance in GEC, further suggesting its reliability.

Finally, it was seen from the performance of the GAN-combined Transformer model in practical correction tasks that it could not only find out grammatical errors but also effectively corrected the mistakes in tenses and pronouns. The model can be applied in practice to reduce teachers' pressure in correcting students' homework and help teachers respond faster to students' situation.

6 Conclusion

In this article, based on deep learning, the Transformer model was combined with GAN to study GEC. The experimental analysis found that the GAN-combined Transformer model had a good performance on both CoNIL-2014 and JFLEG data sets and had better $F_{0.5}$ and GLEU values than the traditional Transformer model. The results on the actual English essay correction also showed that the model was effective in automatically correcting grammatical errors in the English essay. The GAN-combined Transformer model can be promoted and applied in practice.

Conflict of interest: Author states no conflict of interest.

References

[1] Hos R, Kekec M. Unpacking the discrepancy between learner and teacher beliefs: what should be the role of grammar in language classes? Eur Educ Res J. 2015;4:70–6.
[2] Madi N, Al-Khalifa HS. Grammatical error checking systems: a review of approaches and emerging directions. 2018 Thirteenth International Conference on Digital Information Management (ICDIM); 2018. p. 142–7.
[3] Solyman A, Wang Z, Tao Q. Proposed model for Arabic grammar error correction based on convolutional neural network. 2019 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE); 2019. p. 1–6.
[4] Park C, Yang Y, Lee C, Lim H. Comparison of the evaluation metrics for neural grammatical error correction with over-correction. IEEE Access. 2020;8:106264–72.
[5] Acheampong KN, Tian W. Toward end-to-end neural cascading strategies for grammatical error correction. 2019 IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering (ISKE); 2019. p. 1265–72.
[6] Liu ZR, Liu Y. Exploiting unlabeled data for neural grammatical error detection. J Computer Sci Technol. 2017;032:758–67.
[7] Lin N, Chen B, Lin X, Wattanachote K, Jiang S. A framework for Indonesian grammar error correction. ACM Trans Asian Low-Resource Lang Inf Process. 2021;20:1–12.
[8] Zhou S, Liu W. English grammar error correction algorithm based on classification model. Complexity. 2021;2021:1–11.
[9] Wang H, Zhang YJ, Sun XM. Chinese grammatical error diagnosis based on sequence tagging methods. J Phys Conf Series. 2021;1948:012027 (7pp).
[10] Solyman A, Wang Z, Tao Q, Elhag AAM, Toseef M, Aleibeid Z. Synthetic data with neural machine translation for automatic correction in Arabic grammar. Egypt Inform J. 2020;22:303–15.
[11] Won YS, Han DG, Jap D, Bhasin S, Park JY. Non-profiled side-channel attack based on deep learning using picture trace. IEEE Access. 2021;9:22480–92.
[12] Fayek HM, Lech M, Cavedon L. Evaluating deep learning architectures for speech emotion recognition. Neural Netw. 2017;92:60–8.
[13] Tsuruoka Y. Deep learning and natural language processing. Brain Nerve. 2019;71:45–55.
[14] Milletari F, Ahmadi SA, Kroll C, Plate A, Rozanski VE, Maiostre J, et al. Hough-CNN: deep learning for segmentation of deep brain regions in MRI and ultrasound. Computer Vis Image Underst. 2017;164:92–102.

[15] Shi H, Xu M, Li R. Deep learning for household load forecasting – a novel pooling deep RNN. IEEE Trans Smart Grid. 2018;9:5271–80.
[16] Wu H, Shen GQ, Lin X, Li M, Li CZ. A transformer-based deep learning model for recognizing communication-oriented entities from patents of ICT in construction. Autom Constr. 2021;125:103608.
[17] Wolterink JM, Leiner T, Viergever MA, Isgum I. Generative adversarial networks for noise reduction in low-dose CT. IEEE Trans Med Imaging. 2017;36:2536–45.
[18] Tieleman T, Hinton G. RMSProp: divide the gradient by a running average of its recent magnitude. COURSERA: Neural Netw Mach Learning. 2012;4:26–31.