

Received 26 July 2023, accepted 9 August 2023, date of publication 15 August 2023, date of current version 25 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3305397

## RESEARCH ARTICLE

# A Deep Learning-Based Intelligent Quality Detection Model for Machine Translation

MEIJUAN CHEN 

College of Foreign Languages, Wuchang Shouyi University, Wuhan, Hubei 430064, China

e-mail: 2019111009@wsyu.edu.cn

This work was supported by the Teaching Research Division, Wuchang Shouyi University, under the Project of Study on the Teaching Mode of Intercultural Communication Under Outcomes-Based Education under Project 2021Y06.


**ABSTRACT** With more and more active international connections, the complex scenes-aware machine translation has been a novel concern in the area of natural language processing. Although various machine translation methods have been proposed during the past few years, automatic and intelligent quality detection for translation results failed to receive sufficient attention. Actually, the real-time quality evaluation for machine translation results remains important, because it can facilitate constant debugging and optimization of machine translation products. Existing approaches mostly focused on the offline written contents rather than real-time extensive oral contents. To bridge current gap, a sentence-level machine translation quality estimation method is deployed in this paper. In particular, a specific recurrent neural network with double directions (Double-RNN) is proposed as the backbone network structure. The feature extraction process utilizes the Double-RNN translation model, which makes full use of a large amount of parallel corpus. The evaluations show that Double-RNN method proposed in this paper is the closest to the standard quality assessment, and thus can also evaluate the quality of Chinese and English translations more fairly.

**INDEX TERMS** Quality detection, deep learning, machine translation, complex scenes.

## I. INTRODUCTION

### A. BACKGROUND

After nearly half a century of development, machine translation has been able to meet people's communication needs with low price and extremely short response time [1]. With the continuous progress of speech recognition technology, machine translation has also no longer stayed in the field of translation, but expanded to the field of interpretation [2], [3]. As early as the end of 2016, Sogou was the first to demonstrate the simultaneous interpretation technology [4]. Following the speech, the machine translation system will automatically turn the words into Chinese text displayed on the big screen [5], and the English subtitles translated by the machine will appear below the Chinese after a very short interval [6], and Sogou officially announced that its speech recognition accuracy is 97% and the machine translation accuracy is 90% [7]. Xunfei launched the Xiaoyi Translation

The associate editor coordinating the review of this manuscript and approving it for publication was Xiaojie Su .

Translator in 2017, and during the demonstration, guests said a sentence in Chinese to the translator [8], [9], and about one second later, the machine would read out the corresponding English sentence [10]. The translation quality of the XiaoYi translator was high when demonstrating a simple conversation [11]. A few months later, Baidu applied the translation system at its Baidu World Congress, displaying English and Chinese subtitles in real time as the speaker spoke, and its voice recognition rate reached 95% according to Baidu's official statistics [12], [13].

### B. RESEARCH OBJECT

However, simultaneous interpretation is defined as: "A form of interpretation in which the interpreter uses one language (the incoming language) to express the content of the ideas expressed in another language (the source language) accurately in oral form at almost the same speed as the speaker of the source language" [14], [15]. The translation systems of Sogou and Baidu did not express the translated text in oral form, while the translation system of Xiaoyi did not continue

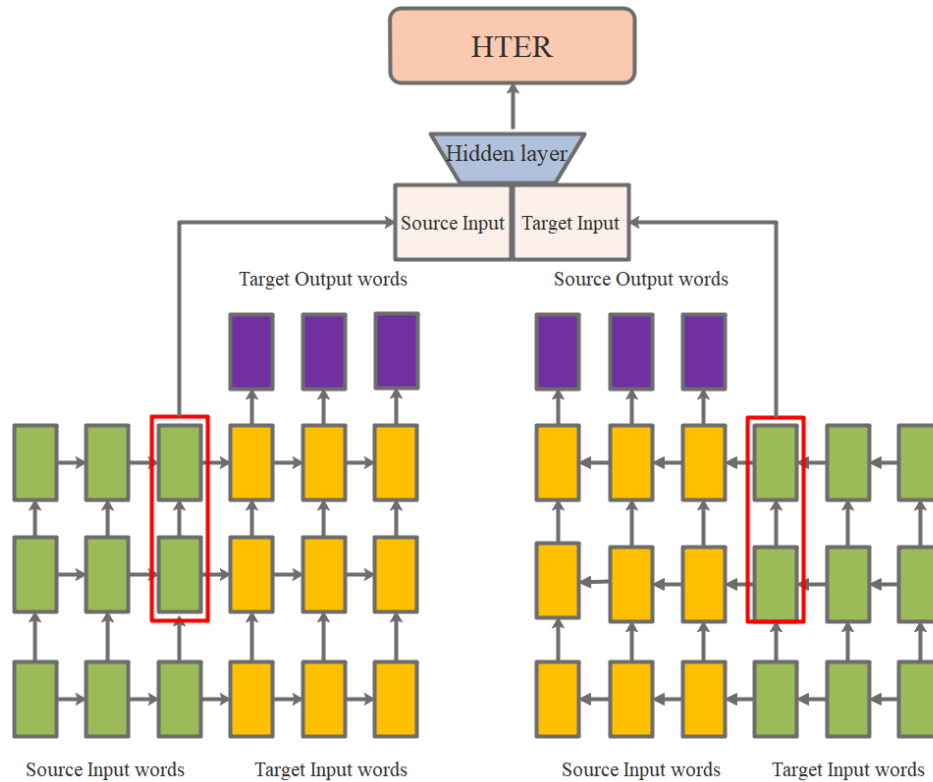


FIGURE 1. Illustration for workflow of the network structure utilized in this paper.

to speak when the speaker was making oral expressions, and thus did not achieve the requirement of near-simultaneous translation and source language [16], [17]. According to the definition of simultaneous interpretation, the products of the aforementioned companies are not true simultaneous interpreters, and cannot be evaluated by the measurement index of simultaneous interpreting [18]. However, what the aforementioned translation systems have in common is that the source language is recognized by speech, the text is generated, then the text is translated, and finally the translated language is output. Since text translation is the core of machine interpretation, and the speech recognition part is another technical issue which is well studied in the natural language processing [19], this paper only focuses on the text translation after speech recognition.

### C. MOTIVATION

This work is distinguished itself from existing researches from two aspects. For one thing, test contents in this paper are different from others. Although some papers have analyzed and categorized the translation quality of machine translation from the perspective of linguistics, these papers are mostly formal texts such as news, political books or foreign speeches. The test contents adopted in this paper, on the other hand, the actual contents of live Chinese speeches in order to meet the practical application of neural network machine

translation in the field of interpretation, only to eliminate the influence of speech recognition on the accuracy of neural network machine translation. For another, existing approaches were mostly established on the basis of machine learning and deep learning-based intelligent algorithms. Especially the neural network-based models play some important roles. Although neural network-based machine translation systems can achieve perfect results with little support from external linguistic knowledge, the combination of linguistic features and learning ability of neural networks has high potential [20].

### D. OUR CONTRIBUTIONS AND PAPER ORGANIZATION

In order to remedy current research gap, this work uses the solution thought of “manual detection plus diagnostic detection” to formulate a novel detection framework. Such solution is expected to make the evaluation scores more linguistically meaningful while quantifying the evaluation results. Therefore, this paper proposes a deep learning-based intelligent quality detection model for machine translation. In particular, a specific recurrent neural network model with double directions (Double-RNN) is developed as the backbone network. The Double-RNN utilizes a large amount of parallel corpus to construct a discriminative model for translation evaluation. It is trained from the perspective of

linguistic learning. Main contributions of this paper can be summed up as following points:

- The significance of automatic quality detection for machine translation is discussed and recognized.
- A deep learning-based intelligent quality detection model is proposed in this paper to realize above purpose.
- Some experiments are conducted on the basis of computer programming to assess performance of the proposal.

The remainder of this paper is planning to be organized as follows. For Section II, it is responsible for survey of related works. For Section III, it introduces mathematical description of technical method in this paper. For Section IV, it gives experimental setting and demonstrates results with analysis. For Section V, this paper is concluded and future direction is outlooked.

## II. RELATED WORK

From the perspective of research methods, the current evaluation methods for machine translation can be roughly divided into two kinds: manual evaluation and automatic evaluation [8]. As early as between 1992 and 1994, the U.S. Department of Defense Advanced Research Projects Agency (DARPA) had organized translation experts to evaluate the three machine translation systems of French-English, Japanese-English and Spanish-English with English as the conversion language from three perspectives of translation fidelity [8]. For the neural network-based machine translation methods which have not been introduced for a long time, there have been corresponding human evaluations [21]. For example, an automatic translation rating system called Test Translation Treasure held a human-machine translation evaluation activity in 2017 in cooperation with FT, in which the neural translation machine and the senior translator of FT translated several English sentences at the same time [22]. And each sentence yielded four answers, and then the audience was asked to choose one of the answers from them [23]. The audience is then asked to choose the answer they think is the human translation [24]. Such an evaluation system lacks quantitative evaluation metrics, and manual evaluation without any quality assessment framework is too subjective, so there is much room for improvement in the manual evaluation of neural network machine translation.

The second evaluation method is the automatic machine translation evaluation [8]. For example, the BLEU system is based on translation accuracy [30]. The METEOR system, on the other hand, is based on single-precision weighted summation averages and single-word recall, and uses a thesaurus. These systems tend to compare the similarity between the machine translation and the reference translation through mathematical calculations before scoring. Some scholars argue that these systems often can only give a measurement number, but cannot explain the meaning of this number [31], and therefore cannot identify the corresponding problem areas to help translation systems progress.

Although there has been some research related to quality assessment of machine translation, existing ones actually dealt with this issue by identifying errors according to the specific rules. We have also listed some typical ones in Table 1 to facilitate reading. It can be seen from descriptions that existing methods can just realize objective discrimination, rather than real subjective assessment. Therefore, it is of significance to make corresponding exploration in this work. In summary, machine translation has received great progress in recent years. However, the intelligent assessment for machine translation quality still remains a challenging issue. The manual assessment for machine translation quality cannot be suitable for large business amount in era of big data. To deal with such challenge, this paper investigates the deep learning-based intelligent quality detection techniques for machine translation.

## III. METHODOLOGY

### A. PREDICTED TARGETS

Sentence-level machine translation quality estimation requires inputting the source utterance and the corresponding machine translated utterance, and then outputting an estimate of the quality of that translation [32]. Currently, the predicted target “quality” mentioned here is generally chosen. HTER (Human-targeted Translation Edit Rate), which is an improvement of TER (Translation Edit Rate), is used [33], [34]. If there is a machine translated translation (hypothesis) and several corresponding reference translations (references), TER is defined as the minimum number of operations required to completely transform machine translated translation into one of the reference translations, and then divided by the average length of the reference translation, as shown in Equation 1.

$$TER = \frac{insertions + dels + subs + shifts}{referencewords} \quad (1)$$

The definition refers to operations, including delete, insert, replace, and move. Because the TER definition requires the minimum number of edits, it is actually the shortest editing distance between the machine translated translation and the reference translation that is most similar to it. The shortest edit distance between the machine translated translation and its closest reference translation. The process of generating the reference translations required for HTER is roughly as follows: the human annotator is first given a machine translated sentence (hypothesis) and one or more untargeted reference translations, which are the references used in the calculation of TER. The human annotator is then asked to modify the machine translated statement until the sentence is fluent and has the same meaning as the untargeted reference translation, and the modified sentence is the human targeted reference translation [35]. In this way, the reference translation needed to calculate the HTER is obtained. The final calculation of HTER The calculation process of HTER is basically the same as that of TER, which is to use the targeted reference

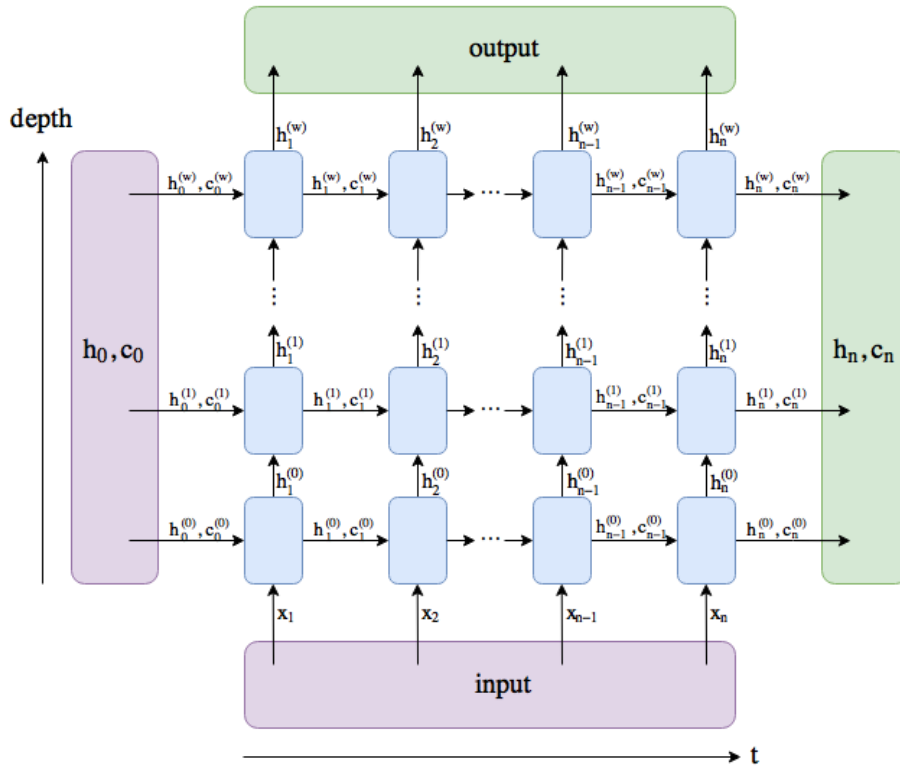


FIGURE 2. The sketch map for structure of the multi-layer RNN.

TABLE 1. Comparison of some existing research works with respect to quality assessment of machine translation.

Reference	Proposed by	Brief description
[25]	Ângela Costa et al.	A linguistic rules-based error detection system is proposed for machine translation.
[26]	Arle Lommel	An evaluation metric for machine translation quality is developed in this work.
[27]	Juncal Gutiérrez-Artacho et al.	The linguistic rules between English and Spanish are investigated to analyze machine translation errors.
[28]	Mireia Farrús et al.	A novel evaluation metric for machine translation quality is proposed from several linguistic levels.
[29]	Marta R. Costa-jussà et al.	Linguistics implicitly is introduced to develop a new quality measurement method.

translation as reference, to calculate the TER of The TER of the machine translated translation.

**B. MODEL**

In this section, the feature extraction algorithm that incorporates translation knowledge is described in detail. The feature extraction process utilizes a neural network machine translation model, which makes full use of a large amount of parallel corpus, and the extracted features are fused with translation knowledge. Then, after the above features are extracted, they are fed into a simple Quality Evaluation (QE) model, and finally the prediction result HTER is output. Here, a single hidden layer forward neural network is chosen as the structure of the QE model [36]. The whole model consists of two parts, the first part is a machine translation model of two neural networks with opposite translation directions. The second part is the QE model, which outputs the quality HTER

of the final machine translation, and the input is the feature vector extracted from the source and target utterances, in this case the coding vector obtained from the two neural network machine translation models, and the feature vector can also contain features extracted by other means. The feature vectors can also contain features extracted by other means.

The overall model structure is shown in Figure 1. There are two machine translation models with opposite translation directions, one for translating the source language to the target language and the other for translating the target language to the source language. As is shown in Figure 2, the two RNN models have identical structures and share word vector parameters. The Double-RNN model with translation direction from source language to target language is introduced as an example below. The source utterance  $X = x_1, x_2 \dots x_s, x_i(1 \leq i \leq S)$  is the word embedding encoding of the words in the source utterance, and  $s$  is the length of the

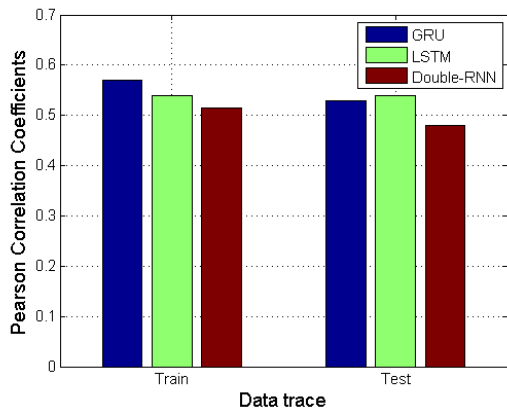


FIGURE 3. Comparison among experimental methods of detection effect with respect to Pearson Correlation Coefficient (when English is translated into Chinese).

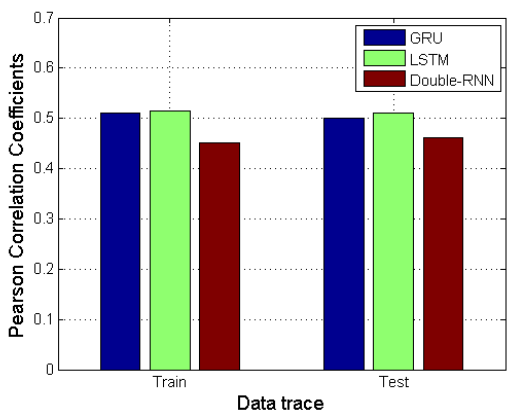


FIGURE 4. Comparison among experimental methods of detection effect with respect to Pearson Correlation Coefficient (when Chinese is translated into English).

source utterance. The target utterance  $Y = y_1, y_2, \dots, y_S$ ,  $y_j (1 \leq j \leq T)$  is the word embedding encoding of the words in the target utterance, and  $T$  is the length of the target utterance. The neural network models used by both the encoder and decoder are variants of recurrent neural networks GRU and LSTM. The function of the encoder is to encode the source utterance  $X$  into a fixed vector  $C$ , and then the decoder decodes  $C$  to obtain the target-end utterance  $Y$ . The whole Double-RNN model can be expressed as  $P(Y|X; \Theta)$ , and this conditional probability can be decomposed by the multiplicative law of probability as shown as:

$$P(Y|X, \theta) = \prod_{j=1}^T p(y_j|x, y_1, \dots, y_{j-1}; \theta) \quad (2)$$

The encoder is mainly composed of GRU or LSTM, and the initial hidden states are all zero vectors. At each step of the computation, the word of the step is first mapped to the corresponding word vector  $x$ , and then used as input along with the hidden state of the previous step for the current step. The computation process of each step is related to the chosen network structure, and GRU and LSTM are chosen as

TABLE 2. Performance comparison of several experimental methods when English is translated into Chinese.

Method	Train	Test
GRU	0.51	0.50
LSTM	0.51	0.51
Double-RNN	0.46	0.45

the encoder (decoder) networks in this project. The decoder decodes the encoding vector  $C$  of the source utterance. The neural network model used is the same as the encoder (GRU or LSTM), and the initial hidden state is  $C, C$  theoretically contains all the relevant information in the source utterance used to translate the target-side utterance. The final output of each step is the probability distribution of all words in the word list for that step. In the evaluation phase, the input is the word vector of the predicted words from the previous step, and the input during training is the word vector of the words corresponding to the target utterance in the previous step. The formula of the hidden state  $h_t$  in step  $t$  is similar to that of the encoder part. The input of the target word probability distribution in step  $t$  is the hidden state of the step, and then the model uses a single hidden layer forward neural network, and the activation function of the hidden layer is the tanh function, and the final output layer is normalized by the softmax function, as shown in equation 3:

$$p(y_i|x, y_1, \dots, y_{i-1}; \theta) = \text{soft max}(W_{o2} \tanh(W_{o1}h_t + b_1)) \quad (3)$$

The input of the QE model is the feature vector  $V$ , the feature vector  $V$  is the splicing of the source-sentence coding vector  $C_S$  and the target-sentence coding vector  $[C_S : C_T]$ . The model uses a single hidden layer forward neural network with weights  $W_1$  and  $W_2$  and bias vectors  $b_1$  and  $b_2$ . relu is used as the activation function for the hidden layer, and sigmoid is used as the activation function for the output layer because the scores from 0 to 1 are to be output. The final prediction formula for hter is as follows:

$$\alpha_1 = \text{relu}(W_1V + b_1)) \quad (4)$$

$$\text{hter} = \text{sigmoid}(W_2\alpha_1 + b_2)) \quad (5)$$

## IV. EXPERIMENT AND EVALUATION

### A. EXPERIMENTAL SETUP

In this paper, in order to test the prediction effect of features extracted with Double-RNN models incorporating translation knowledge, two sets of experiments were conducted on two QE datasets with different orientations and domains, and the final QE models used to output HTER were both forward neural networks. Among them, the input features used in the first set of experiments are the word vector average features of the source and target utterances of the experiments. Because the features incorporating translation knowledge are an improvement of the feature extraction method of directly finding the word vector averages of utterance words, the results of

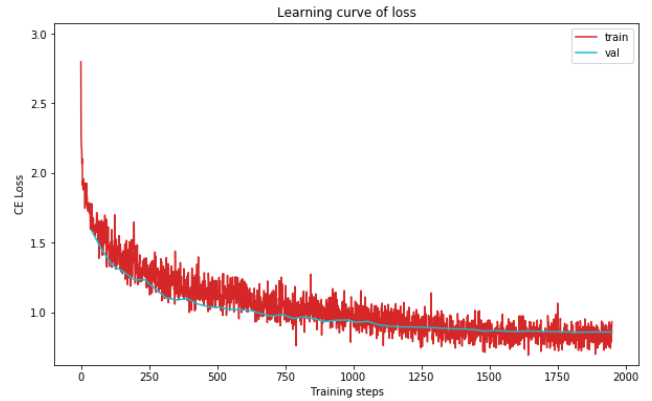
**TABLE 3. Performance comparison of several experimental methods when Chinese is translated into English.**

Method	Train	Test
GRU	0.57	0.53
LSTM	0.55	0.54
Double-RNN	0.51	0.48

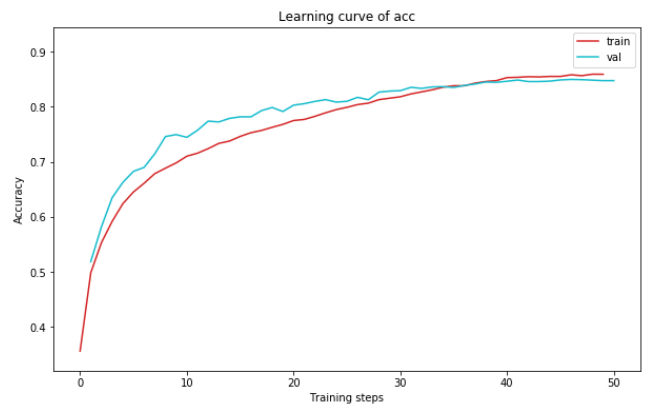
this experiment are used in this section as a comparison to see whether the features incorporating translation knowledge can improve the Pearson correlation coefficients between the predicted and true HTERs.

The features used in the second and third group of experiments are both features fused with translation knowledge and extracted with the Double-RNN model. The difference is that the recurrent neural network structure of both the encoder and decoder of the Double-RNN model used in the second group of experiments is LSTM, while the recurrent neural network structure used in the third group of experiments is GRU. The parameter settings for the second and third sets of experiments are as follows: The word list size is 74000 for the source-side language setting and 74000 for the target-side language setting, OOV (out of vocabulary) was mapped to the special token UNK. The dimension of the word vector was set to 512. The dimension of the word vector is set to 512, and the number of hidden layer neurons of the recurrent neural network (including GRU and LSTM) is 1024. The optimization algorithm for training the neural network model uses the maximum length of the statements in the training RNN model is set to 55, and the learning rate is set to 64. The maximum length of a statement for training NMT model is set to 55, the learning rate is 3e-4, and the loss function is Cross Entropy. The learning rate of the QE model is 5e-5, and the loss function is MSE (mean squared error). The data used in the experiments are described as follows: The dataset used for training Double-RNN was obtained from WMT 2017 Shared Task: Machine Translation of News, with a total of three million sentence pairs. Because the structure of the neural network machine translation model is relatively simple, the entire corpus is not used for training. Instead, 90w sentence pairs are randomly selected from all the three million sentence pairs, and 2w sentence pairs from the corresponding QE dataset (the source utterance plus the translation after being manually post-edit) are added to form the parallel corpus for training the Double-RNN model required in this paper. Before the formal training, the corpus was pre-processed (i.e. tokenize and truecase )by using the tools in Moses.

Pearson Correlation Coefficients (PCC) is the main evaluation index for assessing the scores of sentence-level machine translation quality estimates. The Pearson Correlation Coefficients range from -1 to 1. The larger the absolute value of the correlation coefficient, the stronger the correlation between the two variables: the closer the correlation



**FIGURE 5. Changing tendency for the loss of Double-RNN method.**



**FIGURE 6. Changing tendency for accuracy of Double-RNN method.**

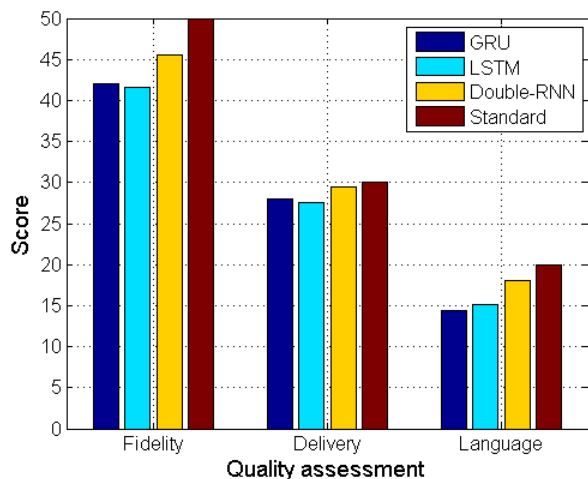
coefficient is to 1 or -1, the stronger the correlation, and conversely, the closer the correlation coefficient is to 0, the weaker the correlation. Here, since the HTER is estimated, the closer the Pearson correlation coefficient between the output of the model and the true HTER is expected to be to 1, the better. The Pearson correlation coefficients of the two variables X and Y are calculated as shown below.

$$r(X, Y) = \frac{cov(X, Y)}{\sqrt{Var[X]Var[Y]}} \tag{6}$$

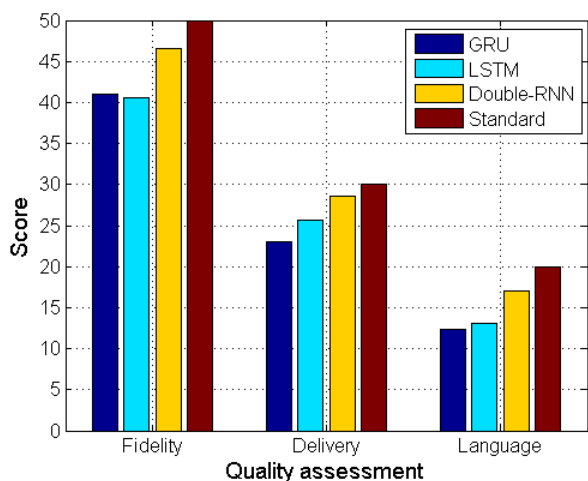
where the numerator Cov(X,Y) is the covariance of variables X and Y, and Var[X] and Var[Y] in the denominator are the variances of variables X and Y, respectively.

**B. EVALUATION AND ANALYSIS**

Table 2 shows the Pearson correlation coefficients between the predicted and true values of the translation quality HTER for the experimental results on the Chinese to English oriented dataset. The Train column in the table shows the results of the model on the development set, and the Test column shows the results on the test set. For comparison with the method proposed in this chapter, the Embedding average row of the table shows the Pearson correlation coefficients between the prediction results and the true values of the



**FIGURE 7.** Quality assessment results between English-Chinese translation with respect to several experimental methods (when English is translated into Chinese).



**FIGURE 8.** Quality assessment results between English-Chinese translation with respect to several experimental methods (when Chinese is translated into English).

model with word vector averages for words characterized as utterances in the Basic Experiments chapter.

The LSTM and GRU show the Pearson correlation coefficients between the prediction results of the Sentence-level QE model and the true HTER using a neural network machine translation model that extracts features incorporating translation knowledge, and the LSTM and GRU in parentheses refer to the recurrent neural network chosen for the encoder and decoder of the neural network machine translation model used, respectively. structure is LSTM or GRU. Table 3 shows the Pearson correlation coefficients between the predicted and true values of the translation quality HTER for the experimental results on the QE dataset in the English to Chinese direction. The major experimental results in Table 2 and Table 3 are also demonstrated in Figure 3 and Figure 4. For the former, it demonstrates the results when 60% of data are

used for training. For the latter, it demonstrates the results when 70% of data are used for training.

The final experimental results on both datasets show that this nonlinear transformation is more predictive of the quality of machine translation than the direct averaging of the word vectors. Moreover, both the word vector features and the features proposed in this topic that incorporate translation knowledge, are not related to specific language pairs, and therefore, compared with some methods of manual feature extraction, they are more generalized. The trends of loss and accuracy of Double-RNN on the dataset are given in Figures 5 and 6, respectively, from which it can be seen that the loss gradually decreases and the accuracy gradually increases as the number of training rounds increases, and the final accuracy stabilizes at 84%, and the results show that the accuracy of Double-RNN proposed in this paper for Chinese and English translation remains relatively high. In addition, the evaluation of translation quality HTER is also more accurate.

According to previous research of the interpretation quality assessment, it can be found that fidelity, delivery and language are three major aspects that people concerned about when assessing the quality of interpretation. Therefore, the following was comprehensive comparison in these three aspects. As can be seen in Figures 7 and 8, the Double-RNN method proposed in this paper is the closest to the standard quality assessment, and thus can also evaluate the quality of Chinese and English translations more fairly.

### C. DISCUSSION

It is really true that RNN consumes a large amount of memory and computing resources when processing large-scale data. In order to establish a robust model for automatic quality evaluation, large amount of corpus data are required as support. Especially in era of large models, massive training data are the foundation for providing intelligent services. It is believed that large models which were pretrained via massive data can have proper performance in many scenarios. Actually, our exploration is still in the beginning stage, we just make evaluation under limited computing resources and on data with limited scale. Our proposal can have proper performance on current experimental data. In the future, we will explore to develop more effective quality evaluation methods under large-scale data operations. If possible, we would like to try to explore suitable large models for this purpose.

It is also noted that Transformer remains prevalent in some typical semantic analysis tasks in recent years. It utilizes the self-attention mechanism to realize representative learning for sequences. It generally has some advantages: better parallel computing efficiency, contextual semantic analysis ability, and better generalization ability. However, there is still no mature Transformer-based methods that are used for translation quality evaluation. And it cannot have quite ideal performance in our experimental process. Therefore, we finally choose to use RNN as the backbone structure to construction specific quality evaluation models. In the

future, we will introduce more samples for training, and try to explore the specific large models on the basis of current work.

## V. CONCLUSION

This work is developed towards the quality evaluation of machine translation results. Compared with existing related works, it has two aspects of distinct points. Firstly, its research objects are beyond the written contents, and are focused on the real-time stream of oral contents. Secondly, it is combined deep neural network with the learning of linguistic features. Under such assumptions, this paper proposes a Double-RNN structure as the backbone network for the investigated purpose. In addition to the theoretical methodology, this work also makes some evaluation experiments. The results show that proposed Double-RNN method is the closest to the standard quality assessment, and thus can also evaluate the quality of Chinese and English translations more fairly.

In all, the intelligent evaluation for machine translation quality is still a novel study in the area of natural language processing. How to make the intelligent algorithms more effective and practical while playing subjective roles, still remains a challenge. The development of deep learning is providing more insight for many areas. Therefore, we would like to make exploration on the basis of this study in the future, and search for more reliable solution towards intelligent detection of machine translation quality.

## REFERENCES

- [1] K. You, G. Qiu, and Y. Gu, "An efficient lightweight neural network using BiLSTM-SCN-CBAM with PCA-ICEEMDAN for diagnosing rolling bearing faults," *Meas. Sci. Technol.*, vol. 34, no. 9, Sep. 2023, Art. no. 094001.
- [2] K. You, G. Qiu, and Y. Gu, "Rolling bearing fault diagnosis using hybrid neural network with principal component analysis," *Sensors*, vol. 22, no. 22, p. 8906, Nov. 2022.
- [3] K. You and H. Liu, "Research on optimization of control parameters of gravity shaking table," *Sci. Rep.*, vol. 13, no. 1, p. 1133, Jan. 2023.
- [4] Y. Keshun and L. Huizhong, "Intelligent deployment solution for tabling adapting deep learning," *IEEE Access*, vol. 11, pp. 22201–22208, 2023.
- [5] Z. Zhou, X. Dong, Z. Li, K. Yu, C. Ding, and Y. Yang, "Spatio-temporal feature encoding for traffic accident detection in VANET environment," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 19772–19781, Oct. 2022.
- [6] C. Chen, Z. Liao, Y. Ju, C. He, K. Yu, and S. Wan, "Hierarchical domain-based multicontroller deployment strategy in SDN-enabled space-air-ground integrated network," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 58, no. 6, pp. 4864–4879, Dec. 2022.
- [7] Y. Li, H. Ma, L. Wang, S. Mao, and G. Wang, "Optimized content caching and user association for edge computing in densely deployed heterogeneous networks," *IEEE Trans. Mobile Comput.*, vol. 21, no. 6, pp. 2130–2142, Jun. 2022.
- [8] H. Ren, X. Mao, W. Ma, J. Wang, and L. Wang, "An English-Chinese machine translation and evaluation method for geographical names," *ISPRS Int. J. Geo-Inf.*, vol. 9, no. 3, p. 139, Feb. 2020.
- [9] Z. Guo, C. Tang, W. Niu, Y. Fu, T. Wu, H. Xia, and H. Tang, "Fine-grained recommendation mechanism to curb astroturfing in crowdsourcing systems," *IEEE Access*, vol. 5, pp. 15529–15541, 2017.
- [10] L. Zhao, Z. Bi, A. Hawbani, K. Yu, Y. Zhang, and M. Guizani, "ELITE: An intelligent digital twin-based hierarchical routing scheme for software-defined vehicular networks," *IEEE Trans. Mobile Comput.*, vol. 22, no. 9, pp. 5231–5247, Sep. 2022.
- [11] X. He, "Evaluation of machine translation quality based on neural network and its application on foreign language education," in *Proc. 3rd Int. Conf. Artif. Intell. Adv. Manuf.*, Oct. 2021, pp. 1395–1399.
- [12] Z. Guo, K. Yu, A. Jolfaei, G. Li, F. Ding, and A. Beheshti, "Mixed graph neural network-based fake news detection for sustainable vehicular social networks," *IEEE Trans. Intell. Transp. Syst.*, early access, Jul. 7, 2022, doi: 10.1109/TITS.2022.3185013.
- [13] S. Xia, Z. Yao, Y. Li, and S. Mao, "Online distributed offloading and computing resource management with energy harvesting for heterogeneous MEC-enabled IoT," *IEEE Trans. Wireless Commun.*, vol. 20, no. 10, pp. 6743–6757, Oct. 2021.
- [14] Z. Yang, T. Hirasawa, M. Komachi, and N. Okazaki, "Why videos do not guide translations in video-guided machine translation? An empirical evaluation of video-guided machine translation dataset," *J. Inf. Process.*, vol. 30, pp. 388–396, Jan. 2022.
- [15] Z. Guo, K. Yu, Z. Lv, K. R. Choo, P. Shi, and J. J. P. C. Rodrigues, "Deep federated learning enhanced secure POI microservices for cyber-physical systems," *IEEE Wireless Commun.*, vol. 29, no. 2, pp. 22–29, Apr. 2022.
- [16] V. Macketanz, E. Avramidis, S. Manakhimova, and S. Möller, "Linguistic evaluation for the 2021 state-of-the-art machine translation systems for German to English and English to German," in *Proc. 6th Conf. Mach. Transl.*, L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussà, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. Jimeno-Yepes, P. Koehn, T. Kocmi, A. Martins, M. Morishita, and C. Monz, Eds., 2021, pp. 1059–1073.
- [17] M. Stefánik, V. Novotný, and P. Sojka, "Regressive ensemble for machine translation quality evaluation," in *Proc. 6th Conf. Mach. Transl.*, L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussà, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. Jimeno-Yepes, P. Koehn, T. Kocmi, A. Martins, M. Morishita, and C. Monz, Eds., 2021, pp. 1041–1048.
- [18] Y. Lu, L. Yang, S. X. Yang, Q. Hua, A. K. Sangaiah, T. Guo, and K. Yu, "An intelligent deterministic scheduling method for ultralow latency communication in edge enabled industrial Internet of Things," *IEEE Trans. Ind. Informat.*, vol. 19, no. 2, pp. 1756–1767, Feb. 2023.
- [19] N. Goyal, C. Gao, V. Chaudhary, P.-J. Chen, G. Wenzek, D. Ju, S. Krishnan, M. Ranzato, F. Guzmán, and A. Fan, "The flores-101 evaluation benchmark for low-resource and multilingual machine translation," *Trans. Assoc. Comput. Linguistics*, vol. 10, pp. 522–538, May 2022.
- [20] E. Comelles and J. Atserias, "VERTA: A linguistic approach to automatic machine translation evaluation," *Lang. Resour. Eval.*, vol. 53, no. 1, pp. 57–86, Mar. 2019.
- [21] M. Liu, H. Zhang, and G. Wu, "Fine grained human evaluation for English-to-Chinese machine translation: A case study on scientific text," 2021, *arXiv:2110.14766*.
- [22] K. Mrinalini, P. Vijayalakshmi, and T. Nagarajan, "SBSim: A sentence-BERT similarity-based evaluation metric for Indian language neural machine translation systems," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 1396–1406, 2022.
- [23] S. Maruf, F. Saleh, and G. Haffari, "A survey on document-level neural machine translation: Methods and evaluation," *ACM Comput. Surv.*, vol. 54, no. 2, pp. 45:1–45:36, 2021.
- [24] R. Bawden, E. Bilinski, T. Laverge, and S. Rosset, "DiaBLA: A corpus of bilingual spontaneous written dialogues for machine translation," *Lang. Resour. Eval.*, vol. 55, no. 3, pp. 635–660, Sep. 2021.
- [25] Â. Costa, W. Ling, T. Luís, R. Correia, and L. Coheur, "A linguistically motivated taxonomy for machine translation error analysis," *Mach. Transl.*, vol. 29, no. 2, pp. 127–161, 2015.
- [26] A. Lommel, "Metrics for translation quality assessment: A case for standardising error typologies," *Transl. Quality Assessment, Princ. Pract.*, vol. 10, pp. 109–127, Jan. 2018.
- [27] J. Gutiérrez-Artacho, M. Olvera-Lobo, and I. Rivera-Trigueros, "Hybrid machine translation oriented to cross-language information retrieval: English-Spanish error analysis," in *New Knowledge in Information Systems and Technologies*, vol. 930. Cham, Switzerland: Springer, 2019, pp. 185–194.
- [28] M. Farrús, M. R. Costa-Jussà, J. B. Mariño, and J. A. R. Fonollosa, "Linguistic-based evaluation criteria to identify statistical machine translation errors," in *Proc. 14th Annu. Conf. Eur. Assoc. Mach. Transl.*, Mar. 2010, pp. 1–12.



- [29] M. R. Costa-jussa and M. Farrus, "Towards human linguistic machine translation evaluation," *Digit. Scholarship Humanities*, vol. 30, no. 2, pp. 157–166, Jun. 2015.
- [30] S. Tripathi and V. Kansal, "Machine translation evaluation: Unveiling the role of dense sentence vector embedding for morphologically rich language," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 34, no. 1, Jan. 2020, Art. no. 2059001.
- [31] L. Tingting and X. Mengyu, "Analysis and evaluation on the quality of news text machine translation based on neural network," *Multimedia Tools Appl.*, vol. 79, nos. 23–24, pp. 17015–17026, Jun. 2020.
- [32] M. Li and M. Wang, "Optimizing automatic evaluation of machine translation with the ListMLE approach," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 18, no. 1, pp. 1–18, Mar. 2019.
- [33] S. Marzouk and S. Hansen-Schirra, "Evaluation of the impact of controlled language on neural machine translation compared to other MT architectures," *Mach. Transl.*, vol. 33, nos. 1–2, pp. 179–203, Jun. 2019.
- [34] H. Yu, W. Xu, S. Lin, and Q. Liu, "Machine translation evaluation metric based on dependency parsing model," *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, vol. 18, no. 4, pp. 44:1–44:15, 2019.
- [35] R. Haque, S. Penkale, and A. Way, "TermFinder: Log-likelihood comparison and phrase-based statistical machine translation models for bilingual terminology extraction," *Lang. Resour. Eval.*, vol. 52, no. 2, pp. 365–400, Jun. 2018.
- [36] D. Kouremenos, K. Ntalianis, and S. Kollias, "A novel rule based machine translation scheme from Greek to Greek sign language: Production of different types of large corpora and language models evaluation," *Comput. Speech Lang.*, vol. 51, pp. 110–135, Sep. 2018.



**MEIJUAN CHEN** was born in Zaozhuang, Shandong, China, in 1983. She received the B.A. degree in English from the Central South University of Forestry and Technology, Changsha, China, in 2006, and the M.A. degree in English from Wuhan University, Wuhan, Hubei, in 2010. She is currently teaching in the Foreign Languages Department, Wuchang Shouyi University. Her research interests include English lexicology, cross-cultural communication, and English pedagogy.

• • •