Research and Implementation of Chinese Couplet Generation System With Attention Based Transformer Mechanism

Yufeng Wang, *Member, IEEE*, Jiang Zhang, Bo Zhang, and Qun Jin, *Senior Member, IEEE*

***Abstract*— Couplet is a unique art form in Chinese traditional culture. The development of deep neural network (DNN) technology makes it possible for computers to automatically generate couplets. Especially, Transformer is a DNN-based "Encoder–Decoder" framework, and widely used in natural language processing (NLP). However, the existed Transformer mechanism cannot fully exploit the essential linguistic knowledge in Chinese, including the special format and requirements of Chinese couplets. Therefore, this article adapts the Transformer mechanism to generate meaningful Chinese couplets. Specifically, the contributions of our work are threefold. First, considering the fact that the words in the corresponding positions of the antecedent clause and the subsequent clause in a Chinese couplet always have same part-of-speech (pos, i.e., word class), pos information is intentionally added into the Transformer to improve the accuracy of the conceived couplet. Second, to deal with the large number of unregistered and low-frequency words in Chinese couplet, a specific unregistered/low-frequency word processing mechanism (UWP) is designed and combined with the Transformer model. Third, to further improve the coherence of couplets, we incorporate the polish mechanisms (PMs) into Transformer model. In terms of three evaluation criteria including bilingual evaluation understudy (BLEU), perplexity, and human evaluation, the experimental results demonstrate the effectiveness of our designed Chinese couplet generation system.**

***Index Terms*— Deep neural network (DNN) based Transformer mechanism, part-of-speech features, polish-up mechanism, unregistered and low-frequency words.**

I. Introduction

TRADITIONAL couplet (namely "对联" in Chinese) is a literary genre originated from the balanced sentences in ancient Chinese poetry. Normally, a couplet consists of the antecedent clause and the subsequent clause. The corresponding characters in the same position of the two sentences are subject to certain constraints of semantic relevance. Usually, it is difficult or even impossible to standardize these rules using logical expressions.

Manuscript received October 30, 2020; revised March 2, 2021; accepted March 31, 2021. This work was supported by the QingLan Project of Jiangsu Province. *(Corresponding author: Yufeng Wang.)*

Yufeng Wang is with College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210049, China (e-mail: wfwang1974@gmail.com).

Jiang Zhang is with Wuxi Municipal Public Security Bureau, Wuxi 214002, China (e-mail: 15651633187@163.com).

Bo Zhang is with the College of Science, Nanjing University of Posts and Telecommunications, Nanjing 210049, China (e-mail: zhangb@njupt.edu.cn).

Qun Jin is with the Department of Human Informatics and Cognitive Sciences, Faculty of Human Sciences, Waseda University, Tokyo 8050, Japan (e-mail: jin@waseda.jp).

Digital Object Identifier 10.1109/TCSS.2021.3072153

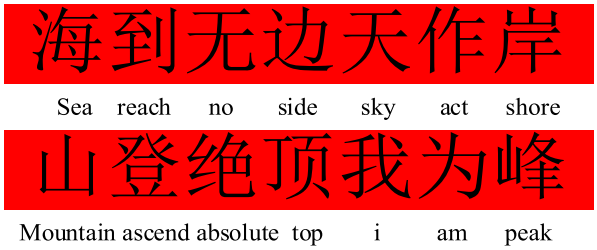


Fig. 1. Example of Chinese couplet. Note that English translation of each Chinese character is given at the bottom.

Fig. 1 is a famous couplet created by Chinese national hero Zexu Lin. Obviously, there is a correspondence between the words in the same positions of the two sentences: "sea" is paired with "mountain," "shore" is mapped into "peak." Creating an excellent couplet as Fig. 1 requires professional knowledge and rich literary skills. In the field of natural language processing (NLP), it is difficult for computers to process couplets because of their rich meanings. Recently, the development of deep neural network (DNN) technology makes it possible for computers to generate couplets based on given conditions. Using computers to generate couplets automatically help common people create couplets for fun or other purposes. Especially, the method of couplet generation using DNN is mainly based on the structure of "Encoder–Decoder" [1], in which the Encoder compresses all the information in the antecedent clause into the context vector, and then the Decoder generates the subsequent clause based on the context vector. And Transformer is a full attention mechanism based Encoder–Decoder structure. More details about Encoder–Decoder structure and Transformer mechanism will be described in Section II.

Targeting at special format and requirements of Chinese couplets, in our article, the following three aspects are designed to make Transformer mechanism suitable for generating Chinese couplets.

First, in Chinese couplets, usually, the part-of-speech (pos, i.e., word class) in the corresponding positions in the antecedent clause and the subsequent clause should be the same. The existed DNN models cannot explicitly take this part-of-speech information into account. This article proposes a method to combine the pos with the word vector to meet the specific requirements of the couplets.

Second, word dictionary is a prerequisite for NLP models, and a dictionary with a huge size will significantly burden these models [2]. Moreover, in Chinese couplets, there exist

many unregistered words and rare words. This article utilizes the similarity calculation of the word vector to replace the unregistered words and low-frequency words with the most similar words in the dictionary, which can reduce the size of the dictionary, increase the speed of the model training, and improve the quality of couplet generated.

In the end, inspired by the way poets revise their works, this article puts forward a polish mechanism (PM) to further improve the quality of the generated couplets. Instead of the traditional method of single-pass generation, the PM takes the initial subsequent clause generated by the adapted Transformer model as a "draft" and generates the true subsequent clause based on it.

The rest of the article is organized as follows. Section II describes the knowledge of Transformer model and the existing works in the field of Chinese couplet generation. The overall framework and main functional modules of the proposed couplet generation framework are provided in Section III. The experimental settings and evaluation criteria are given in Section IV. The experimental results and their analysis are presented in Section V. Finally, we briefly conclude this article.

II. Related Work

The couplet is a unique art form in China. There are few researches focused on the generation of Chinese couplet. However, there are many studies focusing on the generation of poetry. It is feasible to treat couplets as two-line poetry. For computers, the generation of poetry or couplets is essentially the same, in which corresponding sentences are generated under given conditions. It should be noted that there is a difference between them: poetry is not as symmetrical as couplets. The method of poetry generation is not completely applicable to the generation of couplets.

In literature, the means to automatically generate poetry in different languages are as follows:

1) Poetry generation based on template. In this method, some words are removed from an existing poem or couplet, and then words with similar meanings are selected from the dictionary to replace them to produce new poems. A typical system is Wishful Automatic Spanish Poet (WASP), a Spanish poetry generation system [3]. The poetry generated using this method is very hard-shelled, far away from flexible.
2) The method based on reasoning. A Typical poem generation system belonging to this category is Automatic Spanish Poetry Expert and Rewriting Application (ASPERA), which retrieves existing lines and adjust the contents according to the information described by users [4]. The method only considers the superficial semantic information and does not meet the requirements of poetic connotation.
3) Evolutionary algorithm. The method continuously optimizes the generated verse through the selective evolution, until it satisfies a constraint [5].
4) Statistical machine translation. The basic idea is to construct a statistical translation model through analysis of a large number of parallel corpus [6]. Zhou *et al.* [7] extended this method to generate poems of four lines in Chinese classical poetry. The drawback of the method is that the resulting poems are often difficult to understand when combined together.
5) Deep learning techniques. In order to overcome the limitation of the above methods, the method of poetry generation based on neural network has attracted the attention of researchers. Specifically, a Chinese poem generation method named iPoet based on "Encoder–Decoder" neural network framework was proposed in [8] and [9].

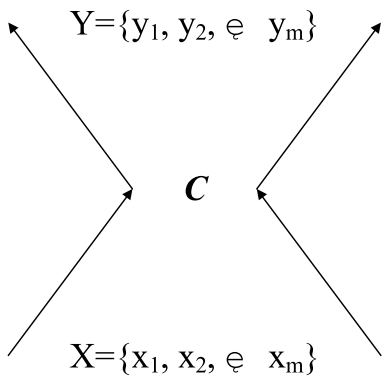


Fig. 2. Schematic Encoder–Decoder structure.

The schematic of Encoder–Decoder framework is shown as Fig. 2. The basic idea of generating couplets with "Encoder–Decoder" framework lies in that: the Encoder of the model represents the antecedent clause as a context vector, and then the Decoder outputs the subsequent clause based on the context vector. In traditional, the Encoder and Decoder are composed of recurrent neural networks (RNNs). Formally, the problem can be described as follows. Given the antecedent clause $X = \{x_1, x_2, \ldots, x_m, x_i \in V$, in which x_i is a character and V is the dictionary of the model, the language model maps the X into the context vector C. The model generates the subsequent clause $Y = \{y_1, y_2, \ldots, y_m\}$ according to the context vector C, where $y_i \in V$.

Based on iPoet, the authors designed an improved Chinese couplet generation system [1], which integrates the traditional "Encoder–Decoder" framework with the attention mechanism.

The attention mechanism was proposed by Bahdanau *et al.* [11]. It originally refers to the selective attention mechanism of human vision. In the process of applying the traditional "Encoder–Decoder" framework, the Encoder will compress all the information of the source sentence into a single context vector, which will result in information loss. Instead of a single context vector C, the "Encoder–Decoder" model with attention mechanism compresses the information of the antecedent clause into a context vector sequence $\boldsymbol{C} = \{c_1, c_2, \ldots, c_m$, in which $c_i \in \boldsymbol{C}$ contains information on different parts of the antecedent clause. Usually, there is a certain connection between the words corresponding to the positions of the antecedent clause and the subsequent clause, i.e., x_i and y_i. The attention mechanism allows the Decoder to dynamically assign different weights to different parts of the subsequent clause when

generating each word of the subsequent clause. Therefore, in a sense, the attention mechanism can be regarded as an alignment model, and is popularly used in the field of poetry generation.

In practice, the "Encoder–Decoder" framework with attention mechanism was used to study the generation of Song lyrics [15]. The antagonistic generation network in image processing was applied to the text generation task [19]. However, all of these schemes use RNN as the basic neural network unit, which not only severely limits the parallelism of model computation, but also hinders the performance. Building a model entirely based on the attention mechanism can be helpful to solve this problem. Specifically, the attention mechanism is applied into the automatic generation of the acrostic couplet in [12]. Self-attention mechanism is used to improve the performance of language representation model in [13], which is future used to deal with the understanding and generation of the ancient Chinese [14].

Transformer is a special case of Encoder–Decoder framework. Different from the traditional Encoder–Decoder networks, It should be note that, Instead of using traditional RNN or convolutional neural network (CNN) for Encoder and Decoder components, Transformer entirely uses the attention mechanism (including self-attention attention) to build the neural network [16]. While, the Encoder–Decoder combined with attention schemes just use the attention mechanism to connect Encoder and Decoder. The traditional Encoder–Decoder is entirely based on RNNs or CNNs. It may not fully capture some useful latent relationships between sentences, and its processing speed is slow.

In summary, Encoder–Decoder framework and attention mechanisms have become the most important method to generate poetry or couplets. However, those models do not explicitly utilize the prior knowledge of linguistics. This article proposes three strategies to improve Transformer based Chinese couplet generation model.

III. Chinese Couplet Generation System

A. Model of Transformer

Considering the sequential nature of RNN limits the speed of model, this article uses Transformer to generate couplets. The Transformer is also based on "Encoder–Decoder" framework. As described above, the distinguished difference between Transformer and Encoder–Decoder lies in that the former is made entirely of attention mechanisms. The parallelism of the attention mechanism not only increases the speed of the model, but can adaptively capture the semantic relationships among clauses.

The structure of Transformer is shown in Fig. 3. The left side is the "Encoder" and the right side is the "Decoder." $N\times$ means that the Encoder and Decoder are stacked by N-layers. Each layer has two sublayers. The first is a multihead self-attention mechanism, and the second is a simple, fully connected feed-forward network. Transformer adopts a residual connection around each of the two sublayers, followed by the layer normalization, i.e., Add&Norm block in Fig. 3. That is, the output of each sublayer is *LayerNorm(x + Sublayer(x))*,

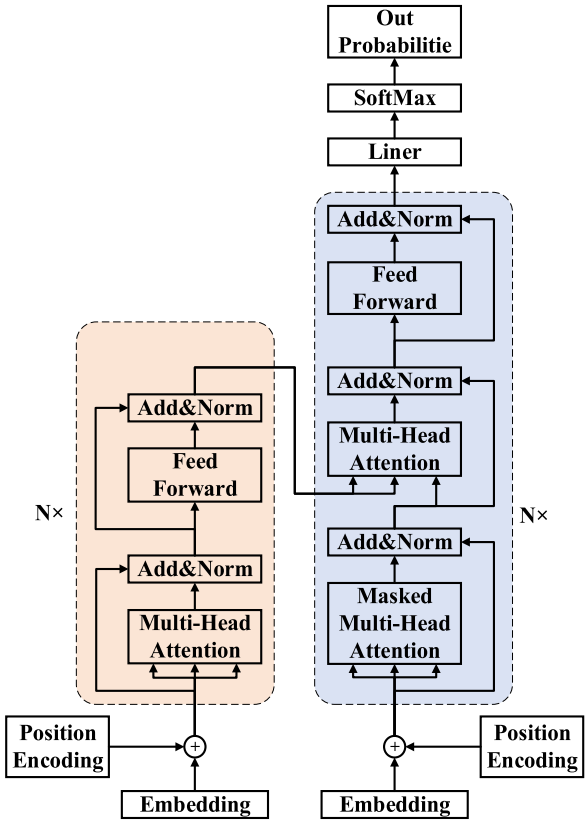


Fig. 3. Illustration of the structure of Transformer.

where *Sublayer(x)* is the function implemented by the sublayer itself, and x is the input to the sublayer. For each layer in Decoder, in addition to the two sublayers, a third sublayer is inserted, which performs multihead attention over the output of the Encoder stack.

The multihead self-attention is a superposition of multiple self-attention mechanisms. An attention function can be described as mapping a query and a set of *key-value* pairs to an output, where the *query*, *keys*, *values*, and output are all vectors. The *output* is computed as a weighted sum of the *values*, where the weight assigned to each *value* is computed by a compatibility function of the *query* with the corresponding *key*. In practice, we compute the attention function on a set of *queries* simultaneously, packed together into a matrix Q. The *keys* and *values* are also packed together into matrices K and V. Then, the output matrix can be represented as the following equation:

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

where d_k is the dimension of the *queries* and *keys*. The multihead self-attention is actually a connection of several self-attentions. It can be described as the following terms:

$$\begin{aligned}\text{MutiHead}(Q, K, V) &= \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^o \\ \text{head}_i &= \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right)\end{aligned}$$

where W is neural network parameters. The objective of multihead self-attention is to establish the dependencies

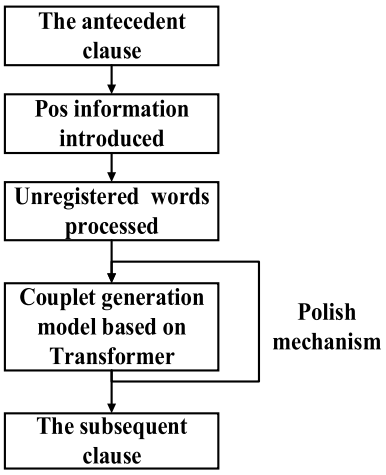


Fig. 4. Flowchart of the designed Chinese couplet generation system.

between each word in a sentence. It captures the internal structure of the sentence and generates the context vectors for each word in the couplets.

The fully connected feed forward neural network integrates the context vector of the source text sequence generated by multihead self-attention with the current word information, to generate the hidden state sequence containing all the information of the couplets. It consists of two linear transformations and a rectified linear unit (ReLU) function.

Before using Transformer to generate couplets, the Chinese characters should be converted into word vectors. Our work simply adopts the existing word vector modules such as WordToVector to embed the word vector based on the corpus in the training set. The position encoding is after the embedding of the Encoder and Decoder. It alleviates the defect that the attention mechanism cannot capture location information. The value of positional embedding is directly superimposed on the word vector, so that the position information of each token and its semantic information (embedding) are fully integrated.

In all, the Encoder outputs the context vector containing all the information about the antecedent clause. The Decoder outputs the probability distribution of each word according to the context vector.

This article adopts Transformer to automatically generate couplets. The framework of our proposed Chinese couplet generation system is shown in Fig. 4, in which the specific functions of each module are described in Sections III-B–III-D.

In detail, we made the following improvements to the traditional Transformer framework. First, the pos information is integrated into the Transformer model. Second, the unregistered/low-frequency words processing mechanism (UWP) is designed to deal with the issue of unregistered and rare words often occurred in Chinese couplets. Third, we incorporate the PMs into Transformer model to improve the coherence of couplets.

B. Pos Information

Recently, researchers have introduced the prior information of words into the language model [17]. The purpose of this

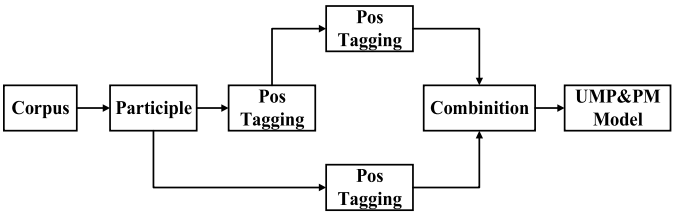


Fig. 5. Process of introducing pos information.

method is to improve the ability of models to learn language knowledge. Based on this idea, a method is designed to enrich the information of the word vector through combining pos information. In detail, this article incorporates the pos of words in the corresponding positions of the antecedent clause and the subsequent clause, which can help learn about the symmetric relationship in Chinese couplet, and improve the accuracy of couplet generation system. The detailed process of model training is shown in Fig. 5.

The specific steps are as follows:

1) Participle. Different from ordinary written or spoken language, the couplet corpus contains many ancient Chinese words. It is difficult to use the word segmentation tool to segment it. Therefore, "participle" component simply regards each character as a word.
2) Tag pos information. This component separates the pos information of corpus from the original corpus.
3) Word vector and pos vector. The word vectors of couplets are trained from the corpus. The corresponding pos vector is obtained from the extracted pos information.
4) Combination. The word vector and pos vector are combined in a certain way. The combination methods include direct addition and splicing. Direct addition simply adds the pos vector and the word vector with the same dimension. Splicing combines the pos vector with word vector through end-to-end connection, that is, the word vector comes before the pos vector.
5) Finally, the combined word vectors are used to take part in the model training.

C. Unregistered Word and Low Frequency Word Processing Mechanism

The problem of unregistered words has always been an important part of NLP. The unregistered words mean the words that are not in the dictionary. The existing solutions include the character based and subword segmentation based methods [18]. The former slices a text by characters. The latter subdivides words into more granular subword. However, neither is suitable for Chinese NLP tasks, since the total number of Chinese characters is close to one hundred thousand, but only a few thousand are in daily use. Most Chinese characters are used infrequently in daily life. Therefore, it is difficult to control the size of the dictionary in this case. Larger size of dictionary can cover more words, but increase the time required for model training and prediction. It is shown that that the training time of Decoder is linearly correlated with the size of dictionary, or even above [2]. Moreover, the dictionary with small size always simply denotes the unregistered words

in the corpus as "UNK" (unknown), which results in the loss of some semantic information in couplets.

This article proposes a method to replace the unknown words and low frequency words. The method exploits the similarity calculation of word vectors: Seeking the high frequency words similar to the unknown words/low-frequency words to replace them. The method can eliminate the low-frequency words from the target dictionary and reduce the search space of the model. At the same time, this mechanism can prevent the unregistered word from being replaced with "UNK" and keep the semantic integrity of the sentence.

Each Chinese character can be represented as a numerical vector (i.e., word vector). There exist several typical approaches to measure the similarity of two vectors in a vector space, including cosine similarity, and Euclidean Distance, and so on. Compared with other measurements, cosine similarity can obtain the bounded similarity value for any number of vector dimensions, and widely used in text comparison, and meanwhile its computational complexity of cosine similarity is low. In a result, our work simply selects cosine similarity as the measurement.

Assume that w_l is an unregistered or low-frequency word, and wh is a high-frequency word in the dictionary, respectively, denoted as $w_l = (x_{l1}, x_{l2}, \ldots, x_{ln})$ and $w_h = (x_{h1}, x_{h2}, \ldots, x_{hn})$, the similarity between w_l and w_h is defined as the following equation:

$$\text{Similarity}(w_l, w_h) = \frac{\sum_{k=1}^{n} x_{lk} x_{hk}}{\sqrt{\sum_{k=1}^{n} x_{lk}^2} \sqrt{\sum_{k=1}^{n} x_{hk}^2}} \tag{2}$$

where n is the number of dimensions in numerical word vector.

The words whose frequency is below a certain value in the dictionary are defined as low-frequency words. The value needs to be determined through experiments. For the low-frequency word and the unregistered word w_l, the word-vector similarity calculation method is used to obtain the word w_h with the highest similarity, which then is use to replace the low-frequency word w_l.

D. Polish-Up Mechanism

Similar to the repetitive process of a poet composing a poem, this article adds a PM to the couplet generation model, which is completely based on the attention mechanism. The Decoder in the Transformer consists of several identical layers stacked on top of each other. Technically, the proposed PM returns the output vector of the last layer in the Decoder back to the first layer of the Decoder. That is, the vectors of subsequent clause generated by the Decoder are calculated by self-attention and context-attention again. The self-attention means the computation of the attention mechanism inside the sentence, and the context-attention means the computation of the attention mechanism between the antecedent clause and the subsequent clause. The corresponding characters in the same position of the two sentences are subject to certain constrain. From the macroscopic viewpoint, the purpose of self-attention calculation is to enhance the coherence of the couplet, and the purpose of context-attention calculation is to enhance the connection between the antecedent clause and the subsequent clause. In brief, attention mechanism is used to polish the hidden state vector of Decoder.

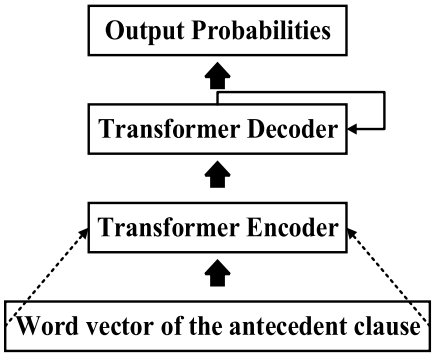


Fig. 6. Schematic of PM.

A simple sketch of the PM is shown as Fig. 6. After the hidden state sequence of the subsequent clause is obtained, it is sent to the input of the Decoder again. Several self-attention and context-attention computations were carried out to obtain the final probability distribution of the subsequent clause. For the iterations, the stopping criterion is set as follows. If the cosine similarity of two sentences is higher than 0.5, these two sentences can be considered similar [1] and the iteration processing will stop. In polish-up mechanism we proposed, the cosine similarity between the subsequent clause before and after polish-up mechanism is always greater than 0.85. In practice, we empirically set the threshold as one time of polishing, which means two iterations in all.

IV. Experiments Settings and Performance Evaluation

A. Evaluation Metrics

To evaluate the quality of the automatically generated couplets, it is necessary to carry out from multiple dimensions and require the professional knowledge of the judgment. This article adopts the metrics of combining human evaluation and machine evaluation to evaluate the generated couplets.

1) Bilingual Evaluation Understudy: Is a measurement for evaluating the quality of machine-translated text, i.e., the subsequent clauses, denoted as candidates in our work, through compare with the actual clauses, denoted as references. The key idea in bilingual evaluation understudy (BLEU) is to compare n-grams of the candidate with the n-grams of the reference and count the number of matches. n-gram represents a set of phrases with a length of n words.

First, the n-gram matching probability is defined as the following equation:

$$P_n = \frac{\sum_{C \in \{\text{Candidates}\}} \sum_{n-\text{gram} \in C} \text{Count}_{\text{clip}}(n-\text{gram})}{\sum_{C \in \{\text{Candidates}\}} \sum_{n-\text{gram} \in C} \text{Count}(n-\text{gram})} \tag{3}$$

where $\text{Count}(n - \text{gram})$ represents the times of the specific n-gram occurring in the clause C, and $\text{Count}_{\text{clip}}(n - \text{gram})$ represents the minimal times of the special n-gram simultaneously occurring in both the clause C and its corresponding

reference. Then, BLEU is defined the following equation:

$$\text{BLEU} = BP \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{4}$$

where $w_n = 1/N$ and N is the maximum number of elements in the N-tuple. BP is the penalty factor. It is used to prevent the situation where the translated result is shorter than the actual answer. In our work, since the length of the generated clause is same as that of the reference answer, thus $BP = 1$.

BLEU is widely used in machine translation and other fields. In a sense, couplet generation can be regarded as one special form of machine translation. Therefore, we use the BLEU to measure the accuracy of the couplet generation. The BLEU score ranged from 0 to 1, and the higher the BLEU score is, the better the performance of the model is. BLEU is 1 only if the generated subsequent clause is exactly the same as the real subsequent clause. The advantage of the BLEU is checking how close computer-generated couplets are to human-written couplets.

2) Perplexity: For language models, the most common metric is perplexity. A language model is a conditional probability model used to calculate the probability of a sentence. Given a sentence $S = W_1, W_2, \ldots, W_k$, including k words, its probability can be expressed as

$$\begin{aligned} P(S) &= P(W_1, W_2, \ldots, W_k) \\ &= p(W_1)P(W_2|W_1)\ldots P(W_k|W_1, W_2, \ldots, W_{k-1}). \end{aligned}$$

Then, the perplexity can be defined as the following equation:

$$\text{Perplexity} = P(w_1 w_2, \ldots, w_N)^{-1/N}. \tag{5}$$

Given the first $(k-1)$ words of a sentence, the language model is wanted to give the probability distribution of the kth word. The basic idea of perplexity is that the language model with a higher probability value for the sentences in the test set is better. After the training of the language model, the sentences in the test set are all normal sentences, so the higher the probability of the trained model on the test set, the better performance of the model. Actually, we can learn from the definition that perplexity calculates the geometric average of the reciprocal probability of each word.

3) Human Evaluation: In addition to the automation evaluation metrics, we also adopt the human evaluation to evaluate the couplet generation system. In detail, the couplets generated by the system are manually evaluated in terms of syntactic relationship relation and semantics. From the perspective of syntactic relationship, the evaluator considers whether the antecedent clause and the subsequent clause are of equal length and the corresponding positional words conform to the antithesis relation. From the perspective of semantics, the evaluator considers whether the antecedent clause and subsequent clause are semantically meaningful and coherent. Table I above shows the specific criteria, each of which is set to a maximum of five points: The higher the score, the better performance of the model. In our experiments, 50 groups of results were randomly selected from the test results, and 20 evaluators were asked to rate the different results. Most of the 20 evaluators were college students. There are 12 postgraduate students majoring in engineering in Nanjing University of Posts and Telecommunications. Six of them are academic students of Nanjing Forestry University, majoring in art. There are two Chinese teachers in middle school. The final result was the average score of 20 evaluators.

TABLE I
HUMAN EVALUATION CRITERIA

Evaluation criteria	The specification	Score range
Syntactic relation	Antithetical or not	1-5
Semantics	Fluency or not	1-5
	The theme is consistent or not	
The overall effect	The overall effect	1-5

B. Data Sets and Experimental Setups

1) Data Sets: The study on the generation of Chinese couplets using neural networks needs a large amount of couplet corpus. The data used in the experiment is from a blog (https://github.com/wb14123/couplet-data set)". The data set contains 770 491 Chinese couplets, including 740 032 couplets with more than four words. 4000 sentences were randomly selected as the test set, 2000 sentences as the verification set, and the remaining 734 032 as the training set.

2) Hyperparameters and Setups: Word embedding is a prerequisite step in using computers to generate couplets. Word embedding maps a word to real-valued vector. The commonly used representation methods include one-hot representation and distributed representation. The one-hot representation method maps the word to a 1-D vector. The word vectors of any two words are independent of each other. The distributed word vector representation method maps the word into a fixed length vector and the distance between the word vectors reflects their similarity. Because the extremely large dimension of one-hot representation may cause dimension disaster, in our model, a 128-dimensional distributed word vector representation method is selected through experiments. The number of layers of Encoder and Decoder in Transformer is set to six layers, and the number of units fully connected to feed forward neural network is 1024. The number of headers in the multihead attention mechanism is set to 8, the size of Batch size is 128, and the probability of the node being dropout is set to 0.1. The cross entropy is used as the loss function of neural network training to describe the distance between the predicted probability distribution and the real probability distribution. The model uses Stochastic Gradient Descent as the optimizer. When the accuracy of the model reaches a certain standard on the verification set, the model stops training. In practice, the training converges after six or seven epochs. The specific values of the parameters are those adopted by the model with better performance in the experiment.

C. Benchmark Schemes for Comparison

In order to verify the effectiveness of attention mechanism on couplet generation task, this article reimplements several

TABLE II
PERFORMANCE OF ENCODER–DECODER, ENCODER–DECODER WITH ATTENTION, TRANSFORMER MODELS

Algorithm	Perplexity	BLEU	Human Evaluation		
			Syntactic	Semantics	Overall
Encoder-Decoder	83.62	0.238	3.39	3.14	3.27
Encoder-Decoder with Attention	79.53	0.243	3.46	3.54	3.5
Transformer	73.43*	0.272*	3.67*	3.77*	3.72*

TABLE III
RESULTS OF ADDING POS CHARACTERISTICS EXPERIMENTS

Algorithm	Way of combination	Perplexity	BLEU
Transformer	/	72.73	0.272
Transformer+Pos	additive	70.71	0.311
Transformer+Pos	64+64 Pos vector + word vector	72.17	0.285
Transformer+Pos	32+96 Pos vector + word vector	70.22*	0.331*
Transformer+Pos	16+112 Pos vector + word vector	71.12	0.294

existing couplet generations models for comparison. For fairness, all models use the same approach to preprocess data.

1) Encoder–Decoder. A couplet generation model based on the basic Encoder–Decoder framework is implemented. Both Encoder and Decoder use long short-term memory (LSTM) as the neural network unit. LSTM is an improvement model of RNN and be used commonly. Each memory unit of LSTM can be used to store information. The Encoder reads the antecedent clause in sequence and outputs the context vector containing all the information of the subsequent clause. The Decoder outputs the subsequent clause according to the context vector.
2) Encoder–Decoder with Attention. The traditional "Encoder–Decoder" model has the shortcoming: The context vector cannot represent the entire input sequence. Experiments given in 10] demonstrate that the performance of the Encoder–Decoder declines rapidly as the input sentence length increases. Therefore, this article reimplements the Encoder–Decoder and attention mechanism-based Chinese couplet generation scheme. The Encoder–Decoder model combined with attention maps the antecedent clause to a context vector sequence $\boldsymbol{C} = \{c_1, c_2, \ldots, c_i\}$. The attention mechanism gives the model the ability to dynamically select a portion of the input vector.
3) Transformer. We also reimplement the Transformer to generate couplets.
4) Transformer + Pos + UWP + PM (i.e., Transformer + Full). Based on the Transformer model, pos information, UWP and PM improvement ways are added, and their performances are given to illustrate the advantage of these strategies.

V. EXPERIMENTS AND ANALYSIS

A. Experiments of Various Models

Table II shows the overall performance of the first three schemes described in the above section, i.e., Encoder–Decoder, Encoder–Decoder combined with Attention, and Transformer.

The data representation with asterisk is optimal. It is clear that performance of the Encoder–Decoder combined with Attention is better than that of the Encoder–Decoder. And the Transformer model, which is based entirely on the attention mechanism, performs best on every metric. In summary, we can conclude that the attention mechanism can improve the performance of the model on couplet generation task.

B. Experiment of Adding Pos Information

To introduce pos information into couplet generation model, it is necessary to study the combination of pos vector and word vector. By using word2vec module, the pos vector is obtained from the pos information, and the word vector is obtained from the original corpus. According to the experiment, the total dimension of the word vector is 128. The results are shown in Table III.

The results of experiments show that the combination of pos information can improve the model obviously. However, different ways of combination have different effects on the model. The model performs best when the dimension of the word vector is 96 and the dimension of the pos vector is 32. Compared with the basic Transformer model, the BLEU score on the test set was improved by 0.059 and the confusion was reduced by 2.51. And the way of addition can also improve the model to some extent. However, other splicing methods do not significantly improve the model. The main reason may be that when the dimension of pos vector is too large, the semantics carried by the word vector itself may be compressed. When the dimension of pos vector is too small, the introduced pos information feature will be insufficient.

C. Experiment of Low Frequency Word Processing Mechanism

In this section, we demonstrate the effectiveness of using high frequency words to replace unregistered words and low frequency words through experiments. Moreover, we design experiments to compare the effects of different substitution ratios on the model.

As shown in Table IV, when the word whose frequency equals 1 is replaced, the BLEU score of the model increased by 0.004. Although the BLEU improvement was not so much, the obtained dictionary was reduced by 1437 words, about 16% smaller. When the replacement threshold of word frequency is set as less than or equal to three and less than or equal to five, the dictionary size can be further reduced, but the BLEU of the model is slightly reduced. The possible reason is that when the high-frequency words are used to

TABLE IV
RESULTS OF THE SUBSTITUTION EXPERIMENTS OF UNRECORDED WORDS AND LOW-FREQUENCY WORDS

Algorithm	Word frequency	BLEU
Transformer	/	0.272
Transformer (UWP)	frequency≤1	0.276*
Transformer (UWP)	frequency≤3	0.271
Transformer (UWP)	frequency≤5	0.270

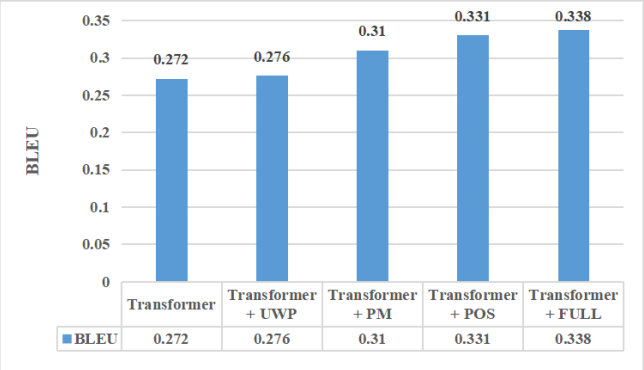


Fig. 7. Performance comparison of three improvement strategies in terms of BLEU.

replace the low-frequency words, the introduced noise has an un-neglectable negative effect on the original meanings, when the replacement ratio is too large.

Note that the process of words substitution really increases the computation overhead of our proposed system. The data used by the experiments have a large size, which contains 770 491 Chinese couplets. According to the statistics, the words whose frequency equal to 1 is only 0.01% in the training set. Assume that the input corpus contains $|S|$ words in total, and the size of the original dictionary is $|V|$. The increase in computation is $|S||V| \times 10^{-4}$. Replacing the low frequency words with frequency one will lead to the size of dictionary was reduced to $0.84|V|$, and the dictionary size was reduced by 16%. The reduction in computation is $0.16|V|^2$. Therefore, the amount of computation brought by the replacement process is negligible.

D. Experiment of PM

There are three improvement strategies in our proposed Transformer-based Chinese couplet generation: 1) introducing Pos information (POS); 2) unregistered word or low frequency word processing (UWP); and 3) PM. This article first compares the improvement effects of three strategies in step-by-step way. Finally, these three strategies are combined together on the model.

The experimental results are shown as Figs. 7 and 8, in terms of BLEU and perplexity, respectively. Obviously, our proposed three strategies can incrementally improve the quality of the generated couplets. Compared with the basic Transformer, our scheme name Transformer + Full improves the perplexity from 74.43 to 68.10, and the BLEU from

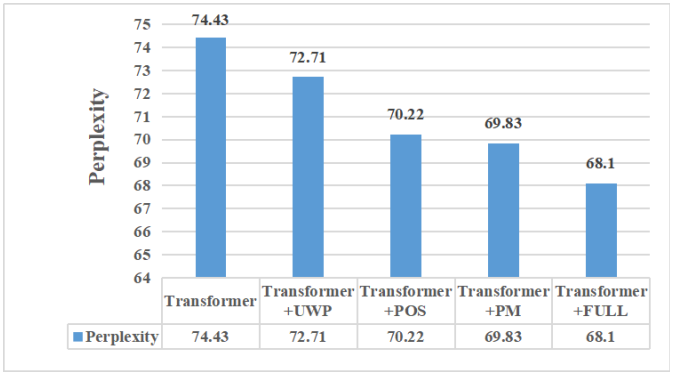


Fig. 8. Performance comparison of three improvement strategies in terms of perplexity.

TABLE V
PERFORMANCE COMPARISON OF TRANSFORMER AND TRANSFORMER + FULL THROUGH HUMAN EVALUATION

Algorithm	Human Evaluation		
	Syntactic relationship	Semantics	Overall
Transformer	3.67	3.77	3.72
Transformer+Full	3.92	3.97	3.95

0.272 to 0.338. That is, the metric of perplexity is reduced by 8.5%, and BLEU increased by 24.26%. To a certain degree, it is a significant improvement. It also should be noticed that the BLEU of the Transformer model with PM is lower than that of the Transformer model with pos information. However, the perplexity of Transformer model with PM is better than that of Transformer model combined with pos information. The reason may be that the PM improves the semantic coherence of the generated subsequent clause. In a result, it performs better on the perplexity. Pos information enables the model to learn the antithesis relationship between the corresponding positions of the antecedent clause and the subsequent clause. And it has a slightly higher BLEU score.

Due to high manual cost of human evaluation, we compare the basic Transformer model (Transformer) and the model combined with three strategies (Transformer + Full), through human evaluation. The results are given in Table V. Obviously, through the manually subjective evaluation, in terms of both syntactic and semantic points, our designed Transformer + Full system is viewed as better and more acceptable than pure Transformer system.

VI. CONCLUSION

The Chinese couplet generation is a difficult task in the NLP field. Instead of adopting the traditional RNN-based Encoder–Decoder framework to generate Chinese couplets, we improve the complete attention mechanism-based Transformer model to generate Chinese couplets that are more coherent and more accurate in syntactic format and semantic meaning.

The results of experiments show that the three strategies proposed by the article can further increase the performance of the language model to generate Chinese couplets.

REFERENCES

[1] R. Yan, C.-T. Li, X. Hu, and M. Zhang, "Chinese couplet generation with neural network structures," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, 2016, pp. 2347–2357.

[2] S. Jean, K. Cho, R. Memisevic, and Y. Bengio, "On using very large target vocabulary for neural machine translation," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process. (Long Papers)*, vol. 1, 2015, pp. 1–10.

[3] P. Gervás, "Wasp: Evaluation of different strategies for the automatic generation of Spanish verse," in *Proc. AISB Symp. Creative Cultural Aspects AI*, 2000, pp. 93–100.

[4] P. Gervás, "An expert system for the composition of formal Spanish poetry," in *Applications and Innovations in Intelligent Systems VIII*, A. Macintosh, M. Moulton, and F. Coenen, Eds. London, U.K.: Springer, 2001, pp. 19–32.

[5] V. Kempe, R. Levy, and C. Graci, "Neural networks as fitness evaluators in genetic algorithms: Simulating human creativity," in *Proc. Annu. Meeting Cogn. Sci. Soc.*, 2001, pp. 1–2.

[6] E. Greene, T. Bodrumlu, and K. Knight, "Automatic analysis of rhythmic poetry with applications to generation and translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2000, pp. 524–533.

[7] L. Jiang and M. Zhou, "Generating Chinese couplets using a statistical MT approach," in *Proc. 22nd Int. Conf. Comput. Linguistics COLING*, 2008, pp. 43–52.

[8] R. Yan, "i, Poet: Automatic poetry composition through recurrent neural networks with iterative polishing schema," in *Proc. 25th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2016, pp. 2238–2244.

[9] N. Kalchbrenner and P. Blunsom, "Recurrent continuous translation models," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2013, pp. 1700–1709.

[10] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder–decoder approaches," in *Proc. SSST-8, 8th Workshop Syntax, Semantics Struct. Stat. Transl.*, 2014, pp. 103–111.

[11] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, 2015.

[12] H. Fan, J. Wang, B. Zhuang, S. Wang, and J. Xiao, "Automatic acrostic couplet generation with three-stage neural network pipelines," in *Proc. Pacific Rim Int. Conf. Artif. Intell. (PRICAI)*, 2019, pp. 314–324.

[13] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[14] H. Tian, K. Yang, D. Liu, and J. Lv, "AnchiBERT: A pre-trained model for ancient Chinese language understanding and generation," 2020, *arXiv:2009.11473*. [Online]. Available: http://arxiv.org/abs/2009.11473

[15] Q. Wang *et al.*, "Chinese song iambics generation with neural attention-based model," in *Proc. 25th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2016, pp. 2943–2949.

[16] A. Vaswani *et al.*, "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 5998–6008.

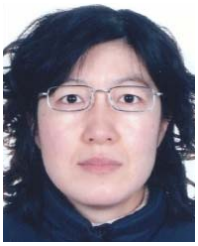
[17] W. Chen, E. Matusov, S. Khadivi, and J.-T. Peter, "Guided alignment training for topic-aware neural machine translation," 2016, *arXiv:1607.01628*. [Online]. Available: http://arxiv.org/abs/1607.01628

[18] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, 2016, pp. 1715–1725.

[19] L. Yu, W. Zhang, J. Wang, and Y. Yu, "SeqGAN: Sequence generative adversarial nets with policy gradient," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 2852–2858.

Yufeng Wang (Member, IEEE) acted as a Expert Researcher with the National Institute of Information and Communications Technology (NICT), Tokyo, Japan, from March 2008 to April 2011. He is a Guest Researcher with the Advanced Research Center for Human Sciences, Waseda University, Tokyo. He is currently a Professor with the Nanjing University of Posts and Telecommunications, Nanjing, China. His research interests focus on cyber–physical–social systems, data sciences, and so on.

Jiang Zhang received the master's degree with a major in telecommunications and information engineering from the Nanjing University of Posts and Telecommunications (NUPT), Nanjing, China, in 2013. He is currently a Police Technician with Wuxi Municipal Public Security Bureau, Wuxi, China. His main research interest are neural networks and natural language process.



Bo Zhang is currently a Senior Experimentalist with the College of Science, Nanjing University of Posts and Telecommunications, Nanjing, China. Her research interests focus on theoretical physics and complex networks, and so on.

Qun Jin (Senior Member, IEEE) is currently a Professor with the Networked Information Systems Laboratory, Department of Human Informatics and Cognitive Sciences, Faculty of Human Sciences, Waseda University, Tokyo, Japan. He has been extensively engaged in research works in the fields of computer science, information systems, and human informatics. His recent research interests include cover human-centric ubiquitous computing, behavior and cognitive informatics, big data, personal analytics and individual modeling, intelligence computing, blockchain, cyber security, cyber-enabled applications in healthcare, and computing for well-being.