Text Reconstruction Method of College English Textbooks from the Perspective of Language Images

Xiaojuan Peng

Guilin University of Electronic Technology, Guilin 541004, China

Correspondence should be addressed to Xiaojuan Peng; deliapeng0826@guet.edu.cn

Received 21 April 2022; Accepted 18 May 2022; Published 18 July 2022

Academic Editor: Lianhui Li

Copyright © 2022 Xiaojuan Peng. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

With the continuous development of social economy, English learning plays an increasingly important role in daily communication. However, the update speed of English textbooks is far lower than the development speed of English. How to reconstruct the text of English textbooks in colleges and universities to improve the learning effect of college English has become an urgent problem to be solved. In order to solve this problem more effectively, this paper proposes a text reconstruction method for college English textbooks from the perspective of language images. First, this paper proposes a text reconstruction network model for college English textbooks, which includes two self-network models, namely, a text feature extraction network and an image feature extraction network. Second, text designs an English text feature reconstruction network to fuse image features and text features to guide the generation of new English texts according to the generated emotions. Finally, through a large number of experiments, it is proved that the text reconstruction method of college English textbooks from the perspective of language images can effectively generate new texts of college English textbooks, enhance the emotional color of college textbooks, and improve the effect of English learning.

1. Introduction

English plays an important role in the daily affairs of internationalization, but learning English [1, 2] is often a difficult problem. Usually, English learning requires teachers to reconcept and design the content of the text [3, 4], improve students’ interest in learning in a reasonable way, turn boring learning into active learning, and deepen the understanding of the text content [5, 6]. The text reconstruction method of college English textbooks [7, 8] has gradually become an effective method to improve college students’ English learning. Through the reconstruction of a variety of English knowledge in English textbooks and the use of diversified text forms, college students can improve the English learning methods and skills [9, 10] mastered by college students and deepen their understanding of the text content. The method of text reconstruction of English textbooks is widely used in various learning tasks such as English learning in primary and secondary schools, professional English training, and English reading teaching.

The teaching method of English textbooks is relatively simple. Traditional learning methods [11] often fail to stimulate students’ desire to learn, limit students’ imagination and comprehension ability, and cannot improve students’ learning effect [12, 13]. Usually, teachers in colleges and universities will give a certain degree of guidance to study [14, 15] and reorganize [16, 17] the structure and content of college English textbooks, reconstruct the old-fashioned content in the text into lively cases, stimulate students’ imagination, and improve students’ learning effect. However, many teachers find some problems in the teaching process. First of all, each country has a certain degree of cultural difference [18, 19]. In order to improve students’ ability to master English and improve their learning effect, students need to constantly feel the emotions [20, 21] and foreign cultures reflected in the text. Secondly, each teacher’s teaching style [22, 23] is very different, and the understanding and reconstruction of the text content are very different. How to make students have a strong interest in the learning of English textbooks often requires teachers to pay a
Image emotional feature extraction is firstly to pre-process the input image, then use the convolutional neural network to extract the emotional information in the image, and then classify according to the extracted emotional information. Image emotion feature extraction is an indispensable link in neural network, and the effect of emotion feature extraction ultimately affects the effect of emotion classification. Convolutional neural networks can obtain higher-level and more abstract feature representations of images by directly extracting the features of images for learning, thus obtaining more essential features in images, and thus making deep learning features more accurate and general.

2.2. Text Feature Extraction. In recent years, text data has grown rapidly through the Internet, and a large amount of text data has been continuously accumulated. These massive data contain a lot of valuable information. Natural language processing is the main method to solve these pieces of text feature information. It uses the text to train a classifier model and then uses the trained network model to divide and supplement the new text.

Sentiment recognition of text content is also called sentiment analysis. It mainly divides the text into a variety of emotion types according to the meaning and emotional information expressed by the text, which is a multisentiment classification problem. By analyzing and researching these text data, important feature information is extracted, which can be used to analyze the public’s attention to hot topics and emotional tendencies, which provides important research ideas for correctly guiding the direction of social public opinion.

With the continuous advancement and development of technology, multimodal data is still a very challenging task at present. In image and text multimodal data, the information contained in text and image is generally complementary. Compared with single-modal data of text or images, multimodal data contains more comprehensive information and can better display and explain the emotional characteristics used. First, the amount of information contained in data of different modalities is often different. Sentiment analysis of multimodal data requires effective understanding and extraction of emotional features of multimodal data. Compared with the traditional single-modal sentiment method, the multimodal sentiment analysis task needs to combine the effective information of multiple modalities to extract the feature interaction between the modal information and the multimodal information in a reasonable way.

2.3. English Text Reconstruction. It is often difficult to improve the English ability of college students through traditional English teaching methods, which are not conducive to students’ good grasp of English syntax knowledge and enhance the effect of English learning. Therefore, on the basis of traditional college English textbook learning combined with English practical teaching activities, many teachers have proposed methods of text reconstruction of college English textbooks from the levels of syntactic
analysis, text sorting, title prediction, and textual analysis to improve English learning. The learning effect has a certain reference value.

English text prediction is a high-level idea of text reconstruction in English textbook learning. It requires readers to predict the unknown content of English textbooks based on existing information and personal understanding of textbook content. Mind map refers to presenting the text content in a more specific, vivid, and hierarchical organization based on the thematic events in English textbooks, using keywords, images, color changes, and other main branches. Based on the mind map of text content, it can help students have a deeper understanding of the content of English textbooks, splicing fragmentary knowledge, grasp the key points in learning, and understand English textbooks from a macrolevel. Reconstruction of text content is a common way of learning, which means that learners, from the perspective of the author, combine the knowledge and content they have learned and rewrite the text content, trends, trends, and so on on the basis of understanding and learning English textbooks. Reconstruction can deepen the understanding of the text content of the textbook and realize the process of learning, accumulating knowledge, and transforming the text content of college English textbooks. Text reconstruction can rearrange and understand the content of English textbooks, improve students’ interest in learning, and deepen their understanding of the text in textbooks.

3. Methods of Text Reconstruction in College English Textbooks

In order to more effectively obtain the feature information of each modality and the interaction of feature information between modalities, this paper adopts a feature fusion model based on standard transformer structure to fuse the extracted image and text features and use it to generate emotional new text message in Figure 1. First, this paper proposes a text feature extraction network, which uses the LSTM network structure to encode temporally continuous text feature information, then uses the attention mechanism to obtain the text feature information that this paper pays more attention to, and finally generates text feature vectors for input into college English textbooks for text reconstruction in the network. Second, this paper proposes an image feature extraction network to input images from textbooks into the network structure. Its main function is to extract human emotional information through images, input the generated image feature vector into the text reconstruction network of college English textbooks, and add more emotional features to the newly generated text information. With two subnetworks and a backbone network, the text reconstruction method of college English textbooks proposed in this paper can generate new content with specific emotions through a generative method. Deep learning methods can generate various types of text reconstruction content, increase students’ interest in learning, and greatly reduce the workload of school teachers.

3.1. Text Feature Extraction Network. The text feature extraction network proposed in this paper is improved on the basis of Att-CNN-BiGRU, which mainly consists of three aspects: text vectorization, text feature extraction, and feature vector generation, as shown in Figure 2. The first part is the vectorization of text features. The main function is to map each word in the input text to a vector representation, and the text vector is represented by the word vector.

Text vectorization is to map the text into a vector that can be recognized and processed by a computer. It is mainly to avoid the problems of one-hot encoding vector sparseness and dimension disaster. Generally speaking, text vectors are mainly composed of word vectors and position vectors. The word vector is to convert the words in the text into vectors that can express semantic information. Considering the meaning of the words and the influence between words, the semantic similarity of words is calculated. The position vector is to convert the position features of the words in the text into vectors and make it clear that the word in the sentence is the trigger word. The position feature is defined as the relative distance between the current word and the candidate word, which represents the relative position of the current word in the sentence.

\[ S = \{w_1, w_2, \ldots, w_n\}, \]
\[ X = \{x_1, x_2, \ldots, x_m\}, \]
\[ x_i = [e_i, p_i], \quad d = d_w + d_p, \]

where \( S \) represents the sentence, \( w_i \) represents the \( i \)-th value in the sentence, and \( n \) represents the length of the sentence. A similar method is used in this paper to map the word \( w_i \) to a real-valued vector \( x_i \); usually, a sentence consists of multiple values.

The text feature extraction network model proposed in this paper is used to extract text features. Text features are divided into lexical-level features and sentence-level features. The lexical-level feature is to extract the background knowledge of the word, including the word’s part of speech, semantic information, emotional information, and classification information. Sentence-level features are the contextual information of the entire sentence in the text, such as grammatical features, degree of association, and emotional expression. Convolution operations are used to extract lexical features in text, and BiLSTM and attention mechanism are used to extract sentence-level features.

Convolutional neural networks can only extract local feature information within the convolution window and cannot perform feature extraction and association on textual context information. In order to fully consider the lexical-level features and sentence-level features of each word and enhance the degree of association between words, this paper introduces a self-attention mechanism to obtain lexical-level features more comprehensively while avoiding lexical position information in the pooling operation of loss.
where $T$ represents the transpose operation of the matrix and dividing by $\sqrt{d_k}$ is to prevent the inner product result from being too large in the experiment. The self-attention mechanism essentially reencodes the input matrix into a new matrix after considering the global feature information through the convolution operation of the matrix. The self-attention mechanism assigns different weights to different vocabulary-level features and considers the global information of the vocabulary and the relationship between related associations so as to obtain the corresponding vocabulary features or weights.

\[ f_{att}(Q, K, V) = \text{softmax} \left( \frac{QK^T}{\sqrt{d_k}} \right)V, \]  
\[ Q \in \mathbb{R}^{n \times d_k}, \quad K \in \mathbb{R}^{n \times d_k}, \quad V \in \mathbb{R}^{n \times d_v}, \]  

3.2. Image Feature Extraction Network. The AlexNet neural network is an 8-layer convolutional neural network with three hidden layers in Figure 3. Its basic structure includes an input layer, 5 convolutional layers, 3 pooling layers, 2 fully connected layers, and an output layer. The hidden layer consists of convolutional layers and pooling layers. Convolutional neural networks can obtain multidimensional features of images through convolution layers, and different

![Figure 1: Structure diagram of the text reconstruction network of college English textbooks.](image1)

![Figure 2: Network structure diagram of text feature extraction.](image2)
convolution kernels can extract different image features. For emotional images of faces, convolution kernels of different scales can be used for feature extraction, and feature information of different scales can be obtained. In this paper, various types of convolution kernels are used to obtain richer image features so that the feature information of the input data can be more accurately expressed.

The sizes of the multiscale convolution kernels in this paper are $1 \times 1$, $3 \times 3$, and $5 \times 5$, which extract features from the input image data, respectively, and add a BN layer at the end of the convolution layer to improve the multiscale convolution network. Finally, the obtained multidimensional expression features are connected and fused. This feature connection method ensures the richness of features. Among them, the $1 \times 1$ convolution kernel can organize information across channels, improve the expressive ability of the network model, and greatly enhance the nonlinear features while keeping the feature scale unchanged. At the same time, the channel can also be adjusted, the pixels on different channels are linearly combined, and then the nonlinear operation is performed to reduce the dimension and the number of parameters.

After cross-layer information fusion, the edge information is also enhanced while retaining the rich high-dimensional features of the feature map. In this paper, the first and second pooling layers of the network are connected to the fully connected layer, and the low-level feature information and high-level feature information are fused as the input of the softmax layer. It can effectively express the feature information of input data, alleviate the problem of gradient disappearance and explosion, enhance the propagation of features in different dimensions, and effectively utilize multidimensional feature information. Cross-layer connection fuses low-level features and high-level features to more accurately describe the feature information of the input data. However, too many parameters are generated in this way, which will cause parameter explosion and overfitting. There are many redundant features, which will reduce the training speed of the network and affect the accuracy of recognition. Therefore, a global average pooling layer is added at the end of the network model to average each feature map, and finally, the result is input to the softmax layer.

3.3. LSTM Neural Network. Because neural networks such as RNN retain previous data information when processing time-series-based data, the previous data will have less and less effect on the model with the subsequent data input; that is, there is a problem of long-distance dependence. In addition, some unimportant data will also be retained by neural network models such as RNN, resulting in data redundancy. In order to solve the above problems, this paper introduces the LSTM network structure. It has the characteristics of maintaining long-term memory and has good performance in processing time-series-based data. The network model structure is shown in Figure 4.

Equation (3) represents the forget gate of LSTM calculated from the input data and the hidden state. Equation (4) indicates that the input data and the weight of the forgetting gate are subjected to linear operation, and the output of the forgetting gate is obtained to represent the memory level of the long-term memory state. Equations (5) and (6) represent the input gate part of the LSTM.

\[
C = \text{sigmoid}(W_1 x_{t-1} + h_{t-1} + b_1),
\]

\[
C' = \text{sigmoid}(W_1 x_{t-1} + h_{t-1} + b_1) \times c_t,
\]

\[
C'' = \text{sigmoid}(W_2 x_{t-1} + h_{t-1} + b_2),
\]

\[
C''' = \text{sigmoid}(W_3 x_{t-1} + h_{t-1} + b_3).
\]

Through the calculation of the above formula, the LSTM network model can complete the update of the existing long-term memory elements, as shown in (7) and (8). Finally, the output gate, LSTM network model is improved. The LSTM network model comprehensively considers the influence of two aspects of the current long-term memory and the current input data elements in the output, as shown in formulas (9) and (10). Finally, the result of the long-term memory is activated by the tanh function, which finally represents the size and positive or negative of the actual output of the LSTM.
4. Experimental Results and Analysis

4.1. Experimental Setup. In terms of text feature extraction and processing, this paper uses the ACE2005 English prediction library as training data. It mainly annotates information such as event mentions, event trigger words, and event elements and is widely used in feature extraction tasks in English. It divides the dataset into training, validation, and test sets in a roughly 8:1:1 ratio. At the same time, the precision rate, recall rate, and F1 value are selected as important indicators for model performance evaluation.

In terms of image datasets, this paper uses the COCO large-scale dataset to train and test the network model. The main purpose of image training is to add emotional information to the newly generated text.

The text uses Adam optimizer to optimize the parameters of the neural network model. The initial learning rate is $10^{-3}$, which is gradually reduced to $10^{-4}$ and $10^{-5}$. By adjusting the learning rate, the loss function value of the network model is continuously reduced, and the prediction accuracy is continuously improved until a stable equilibrium state is reached.

There are many types of human emotions, and different datasets have different types. In order to prove the effectiveness of the proposed text reconstruction method for college English textbooks, this paper screened these emotion types. Both LSTM and GRU units are capable of processing sequences of events. In general, the LSTM prediction effect is better, and the GRU unit is actually more efficient. In this paper, more attention is paid to the generation effect of the text content, so the LSTM unit is adopted.

4.2. Evaluation Indicators. The loss function value is an important indicator for evaluating the actual performance of the model. When training a neural network, in order to improve the prediction effect of the network model, the loss function value is often reduced to a very low level in order to obtain more ideal network model parameters. The loss function is an important reference index to measure the performance of the neural network. Usually, the smaller the loss function value in the test set, the better the function model. For the text reconstruction task of college English textbooks, we use the MSE mean square loss function as the loss function in this paper to fit the regression problem of the predicted value.

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2.$$  \hspace{1cm} (11)

Text or images based on time series usually contain hidden relationships. Through the training of neural network models, latent features between continuous data can be discovered to realize data prediction. In this paper, the MSE loss function is used to continuously adjust the parameters of the network model in reverse to improve the prediction effect of the network model.

In addition, in order to evaluate the prediction effect of the network model, this paper also adopts the MAE mean absolute error as the evaluation index of the model. MAE represents the mean absolute value of the error between the predicted value and the true value. Compared with MSE, MAE can directly reflect the difference with the original data. The smaller the MAE of the network model test data, the better the prediction result of the model.

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i|.$$ \hspace{1cm} (12)

4.3. Comparison and Analysis of Experimental Results. As shown in Figure 5, A–E represent the number of emotion categories in the image dataset in the subnetwork, which are angry, natural, fearful, happy, and sad, respectively. Correspondingly, the percentages in the figure represent the proportion of each type in the image dataset. From the figure, we can clearly see that the number of types B is the largest and the number of types E is the least. Natural and normal images make up the largest proportion of our dataset, and sad images make up the least. Image emotional feature extraction network is mainly trained on this dataset to extract emotional features in images.

Figure 6 shows the distribution of sentiment features in the text dataset. This paper mainly divides text sentiment features into two categories: one is positive sentiment, and
the other is negative sentiment. Positive emotions include joy, enthusiasm, and confidence, and negative emotions include anger, fear, and sadness. In the figure, we use green to represent positive sentiment in the text and yellow to represent negative sentiment in the text dataset. On the whole, the demarcation line between negative emotion and positive emotion is relatively clear, which is suitable for model training and testing.

Figure 7 shows the results of network model optimization learning. The two axes in the plane represent the model parameter settings of the image feature extraction and text feature extraction subnetworks. From the figure, we can see that the parameters of the network model can be optimized by setting some parameters in the image and text feature extraction subnetworks.

Figure 8 shows the learning effect of the image network model. The lighter the color value, the better the prediction effect of the model. From the above figure, we can clearly see that the network model in this paper has a certain accuracy in distinguishing emotion categories, which meets the requirements of image feature extraction.

Figure 9 shows the learning effect of the text feature extraction network. The deeper the red value of the color value, the worse the prediction effect, and the darker the blue value, the better the prediction effect. The text feature extraction network in this paper expands the category of emotion and has been able to classify 10 emotion types more accurately.

Figure 10 shows the learning effect of the backbone network model. The x-axis and y-axis, respectively, represent the number of iterations of the subnetwork model, which is the number of times of training in different data sets. The z-axis represents the effect of learning the backbone network model together with the two subnetwork models. From the
Figure 9: Prediction effect of text feature network model.

Figure 10: The learning effect of the backbone network model.

Figure 11: Comparison of various neural network models.
In order to improve the learning effect of college English, this paper proposes a text reconstruction method for college English textbooks from the perspective of language images. In this paper, the text feature extraction network and image feature extraction network are used to obtain text features and image features, respectively. Then, this paper designs a multimodal feature fusion network and a college English text generation network, which mainly guides the reconstruction of new college English texts according to the emotional features in image texts, generates new texts of college English textbooks with emotions, and improves college English learning effect. Through a large number of experiments, it is proved that the reconstruction method of college English textbooks based on the perspective of language and images can improve the learning ability of college students and reduce the teaching burden of college teachers. The neural network model adopted by the text can handle multimodal data and can be extended to more data types in future work, such as audio data and video data.

Data Availability

The dataset can be obtained from the corresponding author upon request.

Conflicts of Interest

The author declares that there are no conflicts of interest.

References

[16] H. Gong, Y. Sun, and X. Feng, “Tablegpt: few-shot table-to-text generation with table structure reconstruction and


