Proceedings of CONF-SEML 2025 Symposium: Machine Learning Theory and Applications
DOI: 10.54254/2755-2721/160/2025.TJ23487

Chinese Punctuation Restoration Based on Transformer Model

Yizhen He

Shanghai Jiao Tong University, Shanghai, China
h1zn666@sjtu.edu.cn

Abstract: This paper explores the application of Chinese punctuation restoration technology utilizing the pretrained Transformer model XLM-RoBERTa-Base. With the increasing use of digital communication and development of automatic speech recognition, the absence of punctuation can lead to problems like misunderstandings, thus highlighting the importance of effective punctuation restoration technology. This paper focuses on Chinese punctuation restoration with data collected from the IWSLT2012 dataset, where punctuation signs are replaced with designated labels for preprocessing. The Transformer model is fine-tuned on this dataset, focusing on restoring punctuation tasks. After the training, the model's performance is evaluated using metrics including loss, accuracy, PR curves, and confusion matrices. The results indicate that, compared to existing models, the model outperforms in restoring Period and Question. However, the capacity of restoring Comma fails to be enhanced and remains at a moderate level. Also, the punctuation restoration of Chinese faces a performance drop compared to English but improvement compared to Bangla, with same model architecture. Over all, our findings demonstrate the applicability of pretrained Transformer models in Chinese punctuation restoration and suggest avenues for future improvements, particularly in enhancing comma restoration.

Keywords: Punctuation Restoration, Transformer, Chinese, Fine Tuning

1. Introduction

In recent years, the frequent use of digital text has stimulated the need for advanced natural language processing (NLP) like punctuation restoration, as punctuation plays important roles in conveying meanings as well as structures of sentences. In daily life, some people get used to not using punctuation when communicating with others on the Internet, which may lead to misunderstanding. Also, though some of automatic speech recognition models can produce punctuated text [1], most of them produce text without punctuation, which may be hard for computers to process. Thus, Punctuation Restoration technology is of great use. The Transformer Model, which has shown strong capabilities in NLP tasks with its attention mechanisms, is adopted to implement punctuation restoration. A few studies do the work on the Chinese language, like exploring the performance of BERT-CNN-RNN [2].

Thus, this paper aims to develop Chinese Punctuation Restoration technology using a pretrained RoBERTa model with a BiLSTM layer. Data containing sentences with punctuation is collected from the IWSLT2012 dataset [3] and then preprocessed by replacing punctuation signs with labels like COMMA, PERIOD and QUESTION. The preprocessed data is used to fine-tune the Transformer

DOI: 10.54254/2755-2721/160/2025.TJ23487

Model, based on the pretrained XLM-RoBERTa-Base model, which adapts to Chinese. By evaluating the performance of the model, including the PR curve, confusion matrix, and also specific case analysis, the applicability of the pretrained Transformer model in the Chinese language can be verified, as well as the fine-tuning capability in punctuation restoration. Our findings will contribute to the broader field of NLP and provide valuable insights for future research and applications of pretrained models.

2. Related work

Punctuation restoration plays an important role in automatic speech recognition due to unpunctuated text generated by the model. Also, in our daily life, sometimes a lack of punctuation may lead to misunderstanding. This issue particularly happens in languages like Chinese, where the absence of punctuation can lead to ambiguity in sentence structure and comprehension.

Recent research has developed several advanced models to restore punctuation and shown promising results. For instance, a study introduces a method using two-stage RNN equipped with long short-term memory units [4]. In the first stage, the model learns textual features from a large text corpus, and then in the second stage, it is trained with pause duration on the basis of textual features learning and adapted in the speech domain. The testing results show that it reduces error of punctuation restoration up to 16.9% compared with a decision tree [5]. What's more, another study introduces a method using a bidirectional recurrent neural network (RNN) with an attention mechanism, which takes use of long-term context in two directions to improve the performance [6]. Also, the incorporation of the attention mechanism enables the model to gain useful information from the input. The model has demonstrated its effectiveness in English as well as Estonian. Several researcheres take advantage of transfer learning. Nagy et al. utilizes the BERT model to automatically restore punctuation in English and Hungarian [7]. Yi et al. constructs an adversarial learning strategy to restore punctuation, taking use of pre-trained BERT [8]. What's more, there are some models focusing on specific topics, like developing technology for restoring punctuation in medical field [9]. As for Chinese Punctuation Restoration, Zhang et al. developed a BERT-CNN-RNN model, which combines the encoder part of Transformer, CNN, and RNN [2], getting an accuracy of 69.1% on the IWSLT2012 dataset. [10] explore the performance of a model with bidirectional GRU and attention mechanism on small amount of data.

In this paper, the transformer model of Alam et al., equipped with a BiLSTM layer as well as an output layer, is adopted [11]. The model is of good performance on English and Bangla, but its applicability hasn't been verified on the Chinese language. Thus, this paper will explore the performance of the model on the Chinese IWSLT2012 dataset.

DOI: 10.54254/2755-2721/160/2025.TJ23487

3. Introduction to the Transformer Model

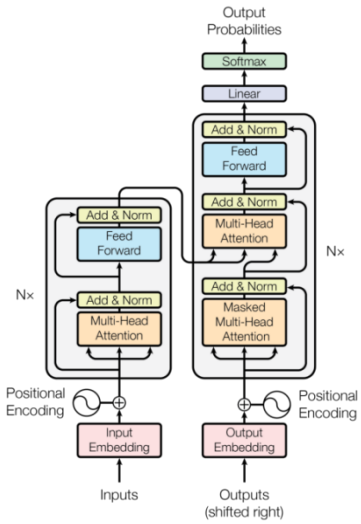

Figure 1: The Transformer Model architecture [12]

Transformer, as is introduced in the paper "Attention is all you need," utilizes an attention mechanism, which contains query, key, and value, mapping them to an output [12]. The output is computed by getting weights with a function of key and query and then calculating the weighted sum of the value. Also, Transformer also takes advantage of the structure of the encoder-decoder. The architecture is shown in Figure 1.

The left encoder part contains N layers of a multi-head self-attention mechanism and a fully connected feed-forward network. Residual connection is adopted between each two sub-layers, as is layer normalization. Like the encoder, the decoder also consists of N layers. Crucially, it incorporates a multi-head self-attention mechanism over the encoder's output, building upon the fundamental structure of the encoder. It is worth mentioning that, as is shown in the figure, the self-attention is masked to ensure that it can only make predictions on the basis of known output before its position. Tokens are transformed into vectors by learned embeddings and then processed by positional encoding to show the information of relative and absolute position. After the tokens are input into the encoder and the decoder, the output of the decoder is fed into a learned linear layer and softmax function to predict the probabilities of the next token, making the prediction. That's how the model works.

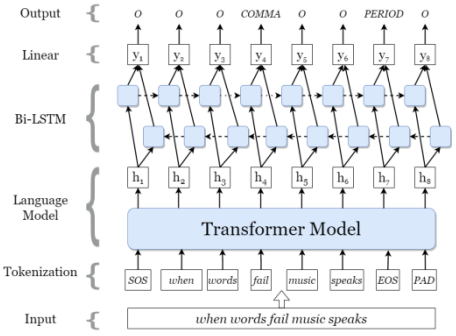

Figure 2: RoBERTa-BiLSTM Model architecture [11]

In NLP project, transfer learning, which takes advantage of pretrained model, are widely used to save time as well as gain good performance. RoBERTa, called Robustly Optimized BERT, is

DOI: 10.54254/2755-2721/160/2025.TJ23487

developed by Facebook AI, which adopts dynamic masking, cancels the use of NSP (Next Sentence Predict) and enlarges the batch size [13]. It contains the embedding, encoder of Transformer and fully-connected layers. The pretraining work focuses on MLM (Masked Language Modeling). In this paper, pretrained Transformer Model XLM-RoBERTa-Base is adopted. XLM-RoBERTa is a multilingual version of the RoBERTa model, which is self-supervised pretrained on 2.5TB of filtered CommonCrawl dataset [14], able to handle over 100 languages, including Chinese.

The total architecture of the model adopted in this paper is a RoBERTa pre-trained Transformer model followed by a BiLSTM layer. In Figure 2, first the input sentence is transformed into tokens by tokenization. For each token, an n-dimension vector will be output from the pretrained Transformer model (i.e., h1, h2, ... in the figure). Then the vector is put into a BiLSTM layer, which combines both forward and backward LSTM (Long Short-Term Memory) to fully take use of contextual information [15]. The outputs from the BiLSTM layer are input to a fully connected layer with linear as the activation function. Finally, 4 output neurons, corresponding to COMMA, PERIOD, QUESTION, and O (character) tokens, are output (i.e., recognizing whether there should be punctuation at the place). For example, in figure 2, the sentence "when words fail music speaks" has no punctuation. With the sentence as the input, the model outputs O, O, O, COMMA, O, O, PERIOD, O, which means that there is a comma after "fail" and a period after "speaks." Then the whole sentence becomes "when words fail, music speaks."

4. Experiment setup

4.1. Dataset and preprocessing

IWSLT2012 Chinese dataset [3] is used in this paper, which contains TED talks transcripts. More details about the dataset are shown in table 1.

Table 1: IWSLT dataset details

Label	Train and Dev Set	Test Set
Characters	3655539	150859
COMMA	132959	5210
PERIOD	135600	5146
QUESTION	10811	515

However, the punctuation marks in dataset collected are in the form of symbols rather than in the form of labels, which means that, in the dataset the labels are ",", "。" and "?" instead of "COMMA," "PERIOD" and "QUESTION". Hence, the first step is to transform the symbols into labels.

4.2. Training details

The model uses pretrained XLM-RoBERTa-Base in the Transformer library of HuggingFace. Also, augmentation method put forward by Alam et al. is adopted to represent typical errors made by automatic speech recognition models [11]. In detail, three types of errors, insertion, substitution, and deletion, are simulated. Token modify probability $\alpha = 0.15$, substitution probability $\alpha_{sub} = 0.4$, deletion probability $\alpha_{del} = 0.4$ and insertion probability $\alpha_{ins} = 0.2$ are chosen [11]. The results of Alam et al. presents that the augmentation also contributes to better performance of the model on normal data. The model using model-specific tokenizer is trained on 3070ti Laptop for 20 epochs with unfrozen RoBERTa layer. For each epoch, only when the model's performance on validation set improves, the weights change will be adopted. As for the training parameters, the maximum length of

DOI: 10.54254/2755-2721/160/2025.TJ23487

sequence is set to 256, batch size is set to 8 and learning rate is set to 1e-5 and Adam optimizer is adopted.

4.3. Evaluation metrics

Performance on training, validation and test set is used to evaluate the model. Precision, Recall and F1 score are the evaluation metrics, which are shown in the following formulas.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{2}$$

$$\text{F1} = \frac{2*\text{P}*\text{R}}{\text{P+R}} \tag{3}$$

Table 2: Supplementary information of TP, FP, TN, FN

	Predict True	Predict False
Actual True	TP	FN
Actual False	FP	TN

5. Results

5.1. Training results

The loss and accuracy of each epoch are shown in Table 3, containing both training and validation results.

Table 3: Loss and accuracy every epoch

Epoch	Training loss	Training accuracy	Validation loss	Validation accuracy
1	0.1070	0.9395	0.0656	0.9535
2	0.0723	0.9508	0.0604	0.9563
3	0.0671	0.9535	0.0590	0.9568
4	0.0639	0.9551	0.0580	0.9576
5	0.0618	0.9564	0.0576	0.9580
6	0.0598	0.9575	0.0568	0.9585
7	0.0580	0.9587	0.0564	0.9587
8	0.0564	0.9596	0.0579	0.9579
9	0.0550	0.9605	0.0571	0.9586
10	0.0537	0.9615	0.0573	0.9592

As is shown above, the performance of the model improves to some extent.

5.2. Testing results

Confusion matrices evaluated on validation and test set are presented to visualize the training results:

On validation set:

DOI: 10.54254/2755-2721/160/2025.TJ23487

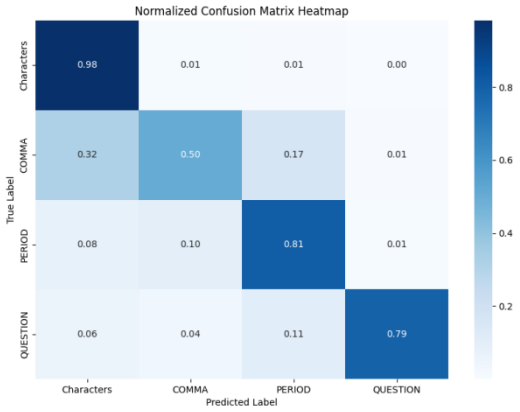

Figure 3: Normalized confusion matrix on validation set

On testing set:

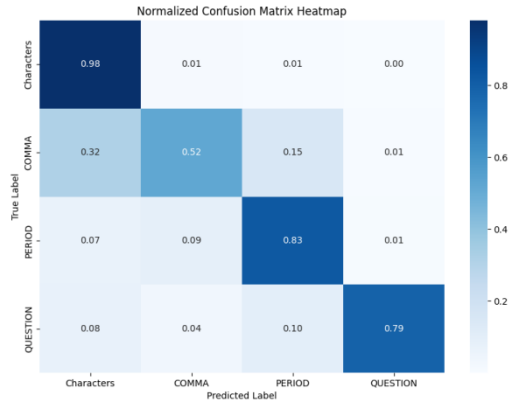

Figure 4: Normalized confusion matrix on test set

It's easy to draw a conclusion that PERIOD and QUESTION gain good precision and recall while COMMA fails to be restored accurately.

In addition, the performance of the model in this paper is compared with an RNN-based model as well as Zhang et al. 's BERT-CNN-RNN model [1, 3] on Chinese punctuation restoration.

Table 4: Comparison of performance of different models on Chinese punctuation restoration

Language Chinese	COMMA			PERIOD			QUESTION			Overall		
Model Name	Pr.	Re.	F1	Pr.	Re.	F1	Pr.	Re.	F1	Pr.	Re.	F1
BiLSTM-Attention	46.6	25.8	33.2	59.9	65.7	62.6	75.6	55.4	64.0	60.7	49.0	53.3
BERT-CNN-RNN	52.0	60.1	55.8	68.4	83.5	75.2	78.8	73.9	76.3	66.4	72.5	69.1
RoBERTa-BiLSTM	55.8	52.4	54.0	69.6	82.7	75.6	77.2	79.0	78.1	64.1	68.0	66.0

Table 4 shows the results of the restoration of COMMA, PERIOD, and QUESTION, which indicates that the RoBERTa-BiLSTM model outperforms the existing models in some aspects.

Compared with the RNN-based model, the RoBERTa-BiLSTM model outperforms in all aspects, which proves the stronger capacity of the Transformer model over a single attention mechanism. What's more, compared with Zhang et al. 's model, the precision of COMMA, PERIOD restoring, and the recall of QUESTION are better, which indicates the strengths of the pretrained RoBERTa model.

DOI: 10.54254/2755-2721/160/2025.TJ23487

The performance of the RoBERTa-BiLSTM model on different languages, like English, Bangla, and Chinese are also compared below. It is worth mentioning that the model with pretrained XLM-RoBERTa-Base is tested on the IWSLT2011 English Ref. test set and model with pretrained XLM-RoBERTa-Large is tested on the manually transcribed Bangla data set. The results are from Alam et al. 's paper [11].

Table 5: Comparation of performance of RoBERTa-BiLSTM model on different languages

RoBERTa-BiLSTM	COMMA			PERIOD			QUESTION			Overall		
Language	Pr.	Re.	F1	Pr.	Re.	F1	Pr.	Re.	F1	Pr.	Re.	F1
English	75.1	70.5	72.7	81.2	89.3	85.1	71.7	82.6	76.8	78.1	79.9	79.0
Bangla	43.3	37.3	40.1	76.5	82.6	79.4	53.0	56.5	54.7	68.3	70.8	69.5
Chinese	55.8	52.4	54.0	69.6	82.7	75.6	77.2	79.0	78.1	64.1	68.0	66.0

As is shown in table 5, the restoration of Chinese faces a performance drop of about 20% for COMMA and a bit of a performance drop for PERIOD and QUESTION compared to the restoration of English. As compared to Bangla, the restoration of COMMA in Chinese is about 10% more precise, and that of QUESTION is 24% more precise. However, the restoration of PERIOD faces a 7% performance drop.

5.3. Case showing

Unpunctuated text in Chinese "但这个例子则不同它是一张来自国会图书馆的扫描图片拥有 3 亿个像素然而浏览它并没有什么区别因为限制系统性能的唯一因素是你所使用的屏幕的像素数" is input to the model and it outputs: "但这个例子则不同。它是一张来自国会图书馆的扫描图片，拥有 3 亿个像素。然而浏览它并没有什么区别。因为限制系统性能的唯一因素只是你所使用的屏幕的像素数。" It can be obtained from the results that there is no problem with the segmentation of pauses based on sentence meaning. Totally, the model successfully restores the punctuation. However, sometimes the model fails to restore COMMA but restores PERIOD.

6. Conclusion

This paper presents the applicability of pretrained Transformer models in Chinese punctuation restoration. We trained the model on laptop 3070ti for 20 epochs, getting an overall precision of 64.1% and a recall rate of 68.0%. The results show that, totally, the model can accurately restore punctuation. However, the precision of restoration for a comma fails to achieve an ideal level, which sometimes is confused with a period. Compared to other models, the RoBERTa-BiLSTM outperforms in precision of restoring commas and periods, and recall of restoring questions. The case of inputting unpunctuated Chinese sentences also indicates the flaws in restoring commas. What's more, compared to the performance of RoBERTa-BiLSTM on other languages like English and Bangla, that on Chinese differs.

To sum up, the pretrained Transformer model is of great applicability in Chinese punctuation restoration. XLM-RoBERTa-Base has good accuracy in the topic even before the work of fine-tuning. Also, through fine-tuning, the performance of the pretrained model is improved to some extent, especially in the accuracy of restoring period and question.

In the future, the architecture of Transformer can be modified, or other mechanisms can be adopted to better restore commas; for example, for words followed by commas, the process of backpropagation can be adjusted to help the model gain a more profound understanding of the existence of commas. Also, there is a limit of computing capability and VRAM on the laptop 3070ti.

DOI: 10.54254/2755-2721/160/2025.TJ23487

If possible, training the model on a 4090 or better device can better present the performance of the model, as it can be trained with larger parameters and for more epochs and get closer to the best performance it can achieve.

References

[1] *Nozaki, J., Kawahara, T., Ishizuka, K., & Hashimoto, T. (2022). End-to-end speech-to-punctuated-text recognition. arXiv preprint arXiv:2207.03169.*

[2] *Zhang, Z., Liu, J., Chi, L., & Chen, X. (2020, December). Word-level BERT-CNN-RNN model for Chinese punctuation restoration. In 2020 IEEE 6th International Conference on Computer and Communications (ICCC) (pp. 1629-1633). IEEE.*

[3] *IWSLT Home Page. 2012. International Workshop on Spoken Language Translation (IWSLT). https://iwslt.org/*

[4] *Tilk, O., & Alumäe, T. (2015, September). LSTM for punctuation restoration in speech transcripts. In Interspeech (pp. 683-687).*

[5] *Kolář, J., Švec, J., & Psutka, J. (2004). Automatic punctuation annotation in Czech broadcast news speech.*

[6] *Tilk, O., & Alumäe, T. (2016, September). Bidirectional Recurrent Neural Network with Attention Mechanism for Punctuation Restoration. In Interspeech (Vol. 3, p. 9).*

[7] *Nagy, A., Bial, B., & Ács, J. (2021). Automatic punctuation restoration with bert models. arXiv preprint arXiv:2101.07343.*

[8] *Yi, J., Tao, J., Bai, Y., Tian, Z., & Fan, C. (2020). Adversarial transfer learning for punctuation restoration. arXiv preprint arXiv:2004.00248.*

[9] *Ling, T., Lai, Y., Chen, L., Huang, S., & Liu, Y. (2023). A small and fast BERT for Chinese medical punctuation restoration. arXiv preprint arXiv:2308.12568.*

[10] *Chao, Y. C., & Chang, C. H. (2020, December). Automatic Punctuation Restoration for corpus in Traditional Chinese Language using Deep Learning. In 2020 International Conference on Technologies and Applications of Artificial Intelligence (TAAI) (pp. 91-96). IEEE.*

[11] *Alam, T., Khan, A., & Alam, F. (2020, November). Punctuation restoration using transformer models for high-and low-resource languages. In Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020) (pp. 132-142).*

[12] *Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.*

[13] *Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.*

[14] *Common Crawl Home Page. 2007. Common Crawl. https://commoncrawl.org/*

[15] *Siami-Namini, S., Tavakoli, N., & Namin, A. S. (2019, December). The performance of LSTM and BiLSTM in forecasting time series. In 2019 IEEE International conference on big data (Big Data) (pp. 3285-3292). IEEE.*