PoemBERT: A Dynamic Masking Content and Ratio Based Semantic Language Model For Chinese Poem Generation

COLING 2025 • Abu Dhabi

The 31st International Conference on Computational Linguistics

Chihan Huang, Xiaobo Shen*

Nanjing University of Science and Technology

huangchihan@njust.edu.cn, njust.shenxiaobo@gmail.com

Abstract

Ancient Chinese poetry stands as a crucial treasure in Chinese culture. To address the absence of pre-trained models for ancient poetry, we introduced PoemBERT, a BERT-based model utilizing a corpus of classical Chinese poetry. Recognizing the unique emotional depth and linguistic precision of poetry, we incorporated sentiment and pinyin embeddings into the model, enhancing its sensitivity to emotional information and addressing challenges posed by the phenomenon of multiple pronunciations for the same Chinese character. Additionally, we proposed Character Importance-based masking and dynamic masking strategies, significantly augmenting the model's capability to extract imagery-related features and handle poetry-specific information. Fine-tuning our PoemBERT model on various downstream tasks, including poem generation and sentiment classification, resulted in state-of-the-art performance in both automatic and manual evaluations. We provided explanations for the selection of the dynamic masking rate strategy and proposed a solution to the issue of a small dataset size.

1. Introduction

CHINESE poetry, rich in history, literary sophistication, and emotional depth, presents challenges for machine learning due to its complex rules of rhyme, meter, and symbolic language. Models need to capture prosody, decode metaphors, and maintain coherence to generate content with depth and accuracy.

Recent advancements in automated poetry generation include LSTM networks, unsupervised machine translation, and self-attention mechanisms. The Mask Language Model (MLM), known for its strong representation and adaptability, uses selective token prediction, where balancing mask rate and content is key to generating poetry that respects traditional linguistic structures.

2. Contributions

1. We proposed a BERT-based model for four ancient Chinese downstream tasks. With comprehensive evaluation and ablation, our PoemBERT reaches SOTA on all tasks.
2. Sentiment embedding and pinyin embedding which circumvents the constraints imposed by interwined morphemes are fused for better character representation.
3. We proposed novel Character Importance, for adding intricate elemental information embedded in single Chinese character to generate highly informative masks.
4. We for the first time explained why early high and later low masking rate is better for model performance and generalization, and adopted dynamic masking rate, both automatic and human evaluation convey appealing results.

3. Methodology

FIGURE 1 is the overview of the proposed PoemBERT model architecture. It can be seen that our model is consist of two main steps: (1) Training a BERT sentiment classification model, thus obtaining the sentiment embeddings of each character. (2) For each character in a sentence, its character embeddings, sentiment embeddings and pinyin embeddings with dimension d are first concatenated to dimension $3d$ and then mapped to fusion embeddings with dimension d through a fully connected layer.

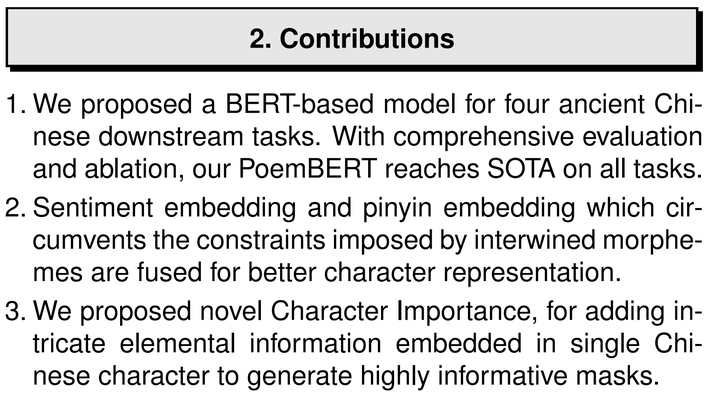

Figura 1: *The overview of PoemBERT model architecture. A BERT sentiment classifier is first trained to obtain the sentiment embeddings of each character. Thus, the fusion embeddings is consist of character embeddings, sentiment embeddings and pinyin embeddings, they are concatenated and mapped to the original dimension through a fully connected layer. '之' in the sentence is a tune word without actual meaning.*

3.1 Embeddings

We developed a sentiment classification model based on BERT to capture nuanced emotional tones, using the [CLS] token for sentiment embedding, which enriches the emotional depth of PoemBERT-generated poetry. To enhance character representation, Pinyin embeddings were integrated using CNNs to model phonetic patterns, with sequences fixed at a length of 8 for consistency. Sentiment, Pinyin, and character embeddings are concatenated, passed through a fully connected layer for fusion, and combined with position embeddings before inputting into the BERT model.

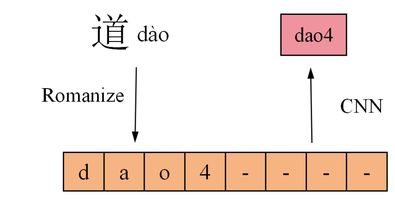

Figura 2: *The overview of pinyin embedding process. Take '道' as an example, we first romanize the pinyin of it to derive a list of 8, then apply CNN with a width of 2 followed by max-pooling to obtain the pinyin embedding output.*

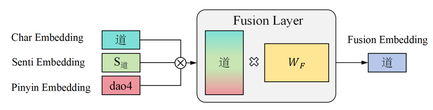

Figura 3: *The process of fusion layer. $\otimes$ represents concatenation, which means concatenate the character embedding, sentiment embedding and pinyin embedding together. $\times$ represents matrix multiplication, we multiple the concatenated matrix with the learnable parameter matrix W_F to derive fusion embedding.*

3.2 Character Importance based masking

PoemMask enhances training efficiency by prioritizing the masking of words with higher importance, measured using Character Importance derived from PMI (Pointwise Mutual Information). This approach focuses on masking less predictable words to improve learning. To address the computational burden of identifying optimal masked words, PoemMask selects a random subset of candidates and ranks them by Character Importance, reducing complexity to $O(kn)$ while minimizing overfitting risks.

3.3 Dynamic masking ratio

We propose a dynamic masking ratio for BERT training, starting with a higher rate to enhance contextual understanding and gradually reducing it to minimize noise and better simulate real-world conditions. Three strategies - linear, cosine, and elliptical decay - are used to optimize this process.

4. Results

4.1 Poem generation

We assessed PoemBERT's poetry generation with two tasks: generating three lines from one input line and generating two lines based on the previous two, emphasizing coherence and contextuality. PoemBERT achieved state-of-the-art results, excelling in both automatic metrics like cosine similarity and human evaluations, highlighting its superior ability to handle the nuanced nature of classical Chinese poetry.

Table 1: *Results on poem generation task. Sim. in $1 \rightarrow 3$ represents the cosine similarity between the first line and the subse-quent three lines of the poem, while $2 \rightarrow 2$ denotes the cosine similarity between the first two lines and the latter two lines of the poem. Con., Flu., Mea., and Poe. respec-tively stand for Consistency, Fluency, Meaningfulness, and Poeticness, which are the four metrics evaluated through human assessment.*

Task	Model	Automatic evaluation		Human evaluation			
		BLUE-4	Sim.	Con.	Flu.	Mea.	Poe
$1 \rightarrow 3$	BERT-base	21.63	0.413	1.87	2.44	1.26	1.45
	AnchiBERT	22.10	0.477	2.47	2.91	2.03	2.17
	QA-MLM	22.87	0.502	2.88	2.96	2.46	2.50
	mT5	22.92	0.508	2.91	3.10	2.53	2.52
	LLaMa	23.42	0.523	2.97	3.22	2.67	2.56
	GPT4	23.44	0.521	2.99	3.25	2.68	2.55
	ChatGLM	23.85	0.542	3.09	3.39	2.79	2.61
	PoemBERT	**24.91**	**0.569**	**3.17**	**3.53**	**2.96**	**2.68**
$2 \rightarrow 2$	BERT-base	29.82	0.507	2.07	2.23	1.99	1.83
	AnchiBERT	30.08	0.558	2.54	2.97	2.45	2.62
	QA-MLM	31.34	0.642	2.67	3.01	2.53	2.89
	mT5	31.42	0.654	2.59	3.30	2.54	2.88
	LLaMa	31.93	0.661	2.82	3.26	2.67	2.94
	GPT4	32.05	0.672	2.92	3.24	2.83	2.87
	ChatGLM	32.97	0.701	3.10	3.41	2.84	3.12
	PoemBERT	**33.47**	**0.717**	**3.35**	**3.68**	**3.22**	**3.45**

4.2 Poem-modern chinese translation

PoemBERT achieved SOTA performance in poetry theme classification, with a BLEU-4 score of 22.76 and a cosine similarity of 0.672, surpassing ChatGLM and GPT4. This highlights its superiority in handling classical Chinese poetry and text translation tasks.

Table 1: *Results on poem-modern chinese translation task. Here Sim. represents the cosine similarity between poem and modern chinese.*

Model	BLEU-4	Sim.
Transformer	18.33	0.445
BERT-base	21.45	0.497
GPT2	21.60	0.489
mT5	21.97	0.523
LLaMa	22.13	0.562
GPT4	22.52	0.617
ChatGLM	22.65	0.649
PoemBERT	**22.76**	**0.672**

4.3 Sentiment classification

PoemBERT achieved an F_1 score of 87.76% in poetry sentiment classification, outperforming baseline models. Data augmentation significantly boosted its performance from 65.74% to 84.24%, addressing limitations of the original dataset.

Table 2: *Results on sentiment classification task.*

Model	F_1/%
BERT-base	43.49
mT5	55.81
LLaMa	57.33
GPT4	60.29
ERNIE	62.88
SA-Model	63.58
ChatGLM	64.01
PoemBERT	**65.74**
PoemBERT+Data augmentation	**84.24**

4.4 Theme classification

PoemBERT achieved 86.36% accuracy in poetry theme classification, surpassing ChatGLM by 2.41% and AnchiBERT by 4.06%, showcasing the effectiveness of pinyin and sentiment embeddings despite the small dataset size.

Table 3: *Results on theme classification task.*

Model	Accuracy/%
Transformer	61.33
GPT2	68.42
BERT-base	75.31
mT5	76.16
LLaMa	78.54
GPT4	80.48
AnchiBERT	82.30
ChatGLM	83.95
PoemBERT	**86.36**

5. Conclusion

We introduce PoemBERT, a BERT-based model for Chinese poetry that integrates sentiment and pinyin embeddings with advanced masking strategies. It achieves outstanding results in poem generation, translation, and classification tasks, improving F_1 scores by nearly 20% through data augmentation.