Bidirectional Transformer with Phonetic Embedding for Various Formatted Rhyming Sentence Generation for Poetry or Lyric Completion from Given Keywords in Chinese

Stanford CS224N Custom Project

Jiabin Wang
Department of Computer Science
Stanford University
dovejb@stanford.edu

Abstract

A Bidirectional Transformer model with phonetic embedding is proposed for generating rhyming sentences in various formats, suitable for poetry or lyric completion based on given keywords in Chinese. Given keywords can be used for different types of acrostic poems(such as "cangtoushi/藏头诗", "cangweishi/藏尾诗", "xiangqianshi/镶嵌诗" in Chinese), or phrases to convey the main topic. Therefore, it is essential to incorporate contexts from both directions in the Bidirectional Transformer model, instead of relying solely on the previous context as in GPT. Moreover, conventional models only take into account semantics, whereas our approach utilizes both semantics and phonetics to improve the rhyming performance of the generated poetry. As a result, our model can produce high-quality poems that excel in both meaning and rhyme, regardless of the position of the given keywords within the poem. Furthermore, it demonstrates superior proficiency in utilizing out-of-vocabulary words. Lastly, due to the natural advantage of the Bidirectional Transformer model, our generated outputs adhere strictly to the desired format specified by the user.

1 Key Information to include

- Mentor: Siyan Li
- External Collaborators (if you have any): None
- Sharing project: None

2 Introduction

There have been numerous attempts to employ GPT-based models for the generation of Chinese poetry. Though these models are capable of producing high-quality poems, users have limited control over the output. As a result, the generated poem may not align with the user's preferences. As an example, users may want to incorporate specific phrases into the poem, such as "happy new year(新年快乐)", "may you prosper(恭喜发财)", or "I love you(我喜欢你)". Alternatively, users may choose to include a specific phrase as the final sentence to convey the main topic, such as the slogan "My life is under my control, not that of a higher power(我命由我不由天)". Traditional GPT-based models do not perform well when it comes to incorporating arbitrary-positioned keywords into the generated text. Another important attribute of poetry is the format. Various forms of Chinese poetry follow specific formats, for instance, the Five-Character Jueju (五言绝句) adheres to a "5-5-5-5" structure, the Seven-Character Lvshi (七言律诗) follows a "7-7-7-7,7-7-7-7" pattern, Xijiangyue in

Songci (宋词-西江月) follows a "6-6-7-6,6-6-7-6" structure, and Xiangjianhuan in Songci (宋词-相见欢) follows a "6-3-9,3-3-3-9" pattern. When it comes to song lyrics, there are countless formats we should consider. GPT-based models are unable to generate sentences accurately in arbitrary but rigid formats. These are the limitations of previous GPT-based models, which make them unsuitable for actual usage. In contrast, our bidirectional model approach allows users to define arbitrary formats with keywords in any position. This helps composers in their actual creative work.
In addition to bidirectional, we also utilize phonetic embedding to depict the phonetic characteristics of tokens. In theory, this approach can facilitate the usage of out-of-vocabulary tokens for the purpose of composing poetry. The Internet's evolution has led to an influx of new words. However, traditional models are unable to incorporate them for creating rhyming sentences as they have not been encountered for training before. Even when forcefully added into a poem, they fail to produce satisfactory results in terms of rhyming. However, through the utilization of phonetic embedding in practice, it is possible to assign a semantic embedding value to new words by linking them to the embeddings of related synonyms and antonyms. Phonetic embedding can also be assigned as the pronunciation of these words is already known. This enables us to fully incorporate any new words for generation, without the need for re-training, let alone there might be very few rhyming corpus with such words for train.
As there are no appropriate bidirectional models in the previous research, we cannot compare the impact of phonetic embedding directly. Therefore, we have prepared two bidirectional models: one without phonetic embedding as a baseline and the other one with phonetic embedding. We will demonstrate the impact using specific examples and multiple evaluation methods.

3 Related Work

We reviewed the GPT-based Chinese poetry generation(Liao et al., 2019) and one for rigid-format poetry generation(Shi, 2021). As stated in the introduction, the GPT-based approach has limitations when it comes to practical usage, and the latter one didn't make use of phonetic characteristics for out-of-vocabulary tokens. I will not mention it again here.

4 Approach

4.1 Phonetic Token Extraction(Pinyin)

Each Chinese character is associated with its own pinyin, some of which may have multiple variants. We divide pinyin into three components: initial, final, and tone, and each component has its own corresponding vocabulary, as shown in Figure 1. For each token, our model will retrieve the appropriate embeddings for the token, initial, final, and tone from their respective vocabularies, and use them as inputs for the transformer. It is important to note that some finals in Chinese pinyin are compound, and only the rhyming part should be considered as the final token.

4.2 BERT(Devlin et al., 2018) with Phonetic Embedding and Rhyming Position Embedding

Each input token is associated with four types of embeddings: a token embedding that represents its semantic characteristics, as well as initial, final, and tone embeddings that capture its phonetic characteristics. Additionally, we employ a rhyming position embedding to indicate the relative relationship between different positions in the minimal sentence elements (i.e., the parts that contain no symbols). The index of the symbol is 1, while the index of the last token is 2, and so on, in reverse order. As Figure 2 illustrates. The baseline model does not have these embeddings.

4.3 Pretrain: BERT Default MLM

To pretrain our model, we employ the default BERT Masked Language Model approach. This involves retaining 85% of the tokens as is, and modifying the remaining 15% of tokens as follows: 80% are replaced with "<mask>", 10% are replaced with random tokens, and the remaining 10% are unchanged but appear in the label.

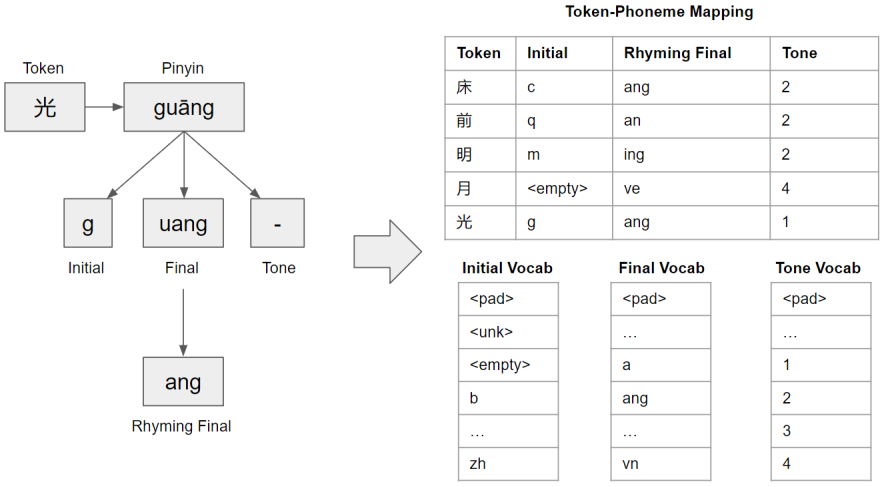

Figure 1: Example of Phonetic Token Extraction(Pinyin)

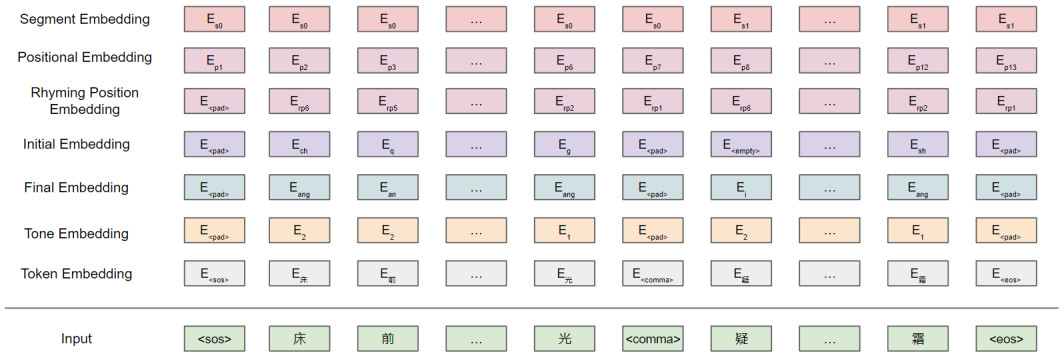

Figure 2: Embeddings of BERT including Phonetic Embeddings and Rhyming Position Embedding

4.4 Finetune: Dense Masked Position LM

To fine-tune our model, we utilize a large density of masked positions. For each input, we first randomly select a mask probability between 10% and 90%, and use this probability to determine how many positions will be masked. Each masked token serves as the output label. This fine-tuning method simulates a bidirectional creation process.

4.5 Generation: Multi-Position Beam Search

Unlike GPT-based models, which only predict the next token (word) at a single position, our bidirectional model, which uses a large density of masked positions, can predict all masked positions simultaneously. Therefore, we employ a multi-position beam search to generate the best possible result. In one step, one position may have multiple candidates, and candidates may also appear in other positions. Only top k candidates with highest score will be kept, until all the masks are filled up.

5 Experiments

5.1 Data

We use Chinese ancient poetry dataset `https://github.com/hlthu/Chinese-Poetry-Dataset` as train dataset. At present, we only choose poems with uniform sentence lengths. Furthermore, if a poem consists of more than four sentences, it will be divided into multiple inputs. Each input will contain only four sentences. See table 5.1. And Table 5.1 is some input and label samples.

Cell Len	Quantity	Example
3	255	一日日，一时时。龙门老，心自知
4	9269	一切境界，病眼倒见。但静意根，空慧自现
5	249626	鹏程三万里，别酒一千钟。好景当三月，春光上国浓
6	2664	虽自九天分派，不与万李同林。步处雷惊电绕，空余翰墨窥寻
7	246245	黄鸟无声叶满枝，闲吟想到洛城时。惜逢金谷三春尽，恨拜铜楼一月迟
total	508059	

Table 1: Dataset inspect

5.2 Evaluation method

Table 5.2 lists all metrics for evaluation. Rhyme Score is based on Chinese Jueju Rhyming Rules, the rules in Chinese are " 首句押韵的：仄起平收、平起平收，只有第三句不押韵。首句不押韵的：仄起仄收、平起仄收。1、3 句不押韵，2、4 句押韵。". If a poem satisfies either of the rule, it receives full score 1.0. And there are lower score level for worse rhyming quality, from 0.9 to 0.4. If a poem does not rhyme at all, the score is 0. See table 5.2 for details.

Rhyme Diversity is the percentage of finals that all the generated poems uses for rhyme among all finals.
Rhyme Token Diversity is the percentage of tokens that a poems use for rhyming among all tokens.
Token Position Diversity has 2 parts: sentence position diversity and intra-sentence position diversity. The final value is the average of the two. For example, if a token only appears in the 1st sentence cell, the sentence position diversity is 0.25. And if a token only appears the last and 2nd last position inside a sentence cell, the intra-sentence position diversity is 2/7.
About Distinct-N, please refer to (Li et al., 2015). All these metrics are automatic and quantitative.

Stage	Sample
pretrain	野水分微 __，巢禽惊稳栖。前村应曙色，依约数 ____ ————白，————。————色，————声鸡
pretrain	迳草细 __ 合，溪流 __ 不喧。幽怀未能惬，城郭已朝暾 ————将—，——深——。————，————
pretrain	一丸岂虑封函谷，__ 骑 ____ 饮 ____。好立功名 __ 竹素，莫教空 __ 霍嫖姚 ————————，千—无由—渭桥。————标——，————说————
pretrain	幸自东南好，西 __ 未可忙。茶 __ 橄榄味，酒借蛤蜊 __ ————，—归————。—添————，————香
finetune	__ 水 __ 微 __，_____ 稳 __。前 ____ 曙色，____ 数 __ 鸡 野—分—白，巢禽惊—栖。—村应——，依约—声—
finetune	__ 短 _____，_____。__ 山常入梦，何日到吾庐 晷—全疏客，窗晴好对书。故————，————
finetune	扶 __ 起 __，__ 海扬尘。孰能高蹈，独洁 __ 身 —摇—鹏，四————。————，——其—
finetune	衰年 _____，__ 贼 _____。落叶 _____，空 ____ 夜声 ——踏险行，五—太无情。——滑霜路，—山生——

Table 2: Samples of input data and label

Rule No.	Score	Rhyming Rule	Ping-Ze Rule
1	1.0	1-2-4 rhyme. 3 do not rhyme.	Ze-Ping, Ping-Ping
2	1.0	2-4 rhyme. 1 and 3 do not rhyme with them.	Ze-Ze, Ping-Ze
3	0.9	1-2-4 rhyme. 3 do not rhyme.	Do not meet #1
4	0.8	2-4 rhyme. 1 and 3 do not rhyme with them.	Do not meet #2
5	0.6	All rhyme, or 2-3-4 rhyme.	Any
6	0.5	1-2 rhyme, 3-4 rhyme. The 2 groups do not rhyme.	Any
7	0.4	Only 3-4 rhyme.	Any
8	0	None of above.	Any

Table 3: Rhyme rules in detail

Metric	Description
Average Rhyme Score	Rhyming quality, max 1
Rhyme Diversity	How many different finals can be used for rhyme
Rhyme Token Diversity	How many token can be used for rhyme
Token Position Diversity	A token can be appear in every position, or tend to be only some fixed position
Token Diversity	The total tokens that are used for generating poems
Distinct-1	Diversity(Dinstinct-Ngram) N=1
Distinct-2	Diversity(Dinstinct-Ngram) N=2
Distinct-3	Diversity(Dinstinct-Ngram) N=3
Distinct-4	Diversity(Dinstinct-Ngram) N=4

Table 4: Metrics that are used to evaluate

5.3 Experimental details

All the experiments is under Mini-BERT model, sequence_length=40, hidden_size=256, num_hidden_layers=4, num_attention_heads=4. Vocab_size=9718.
In pretrain stage, batch_size=256, num_epoch=50, learning_rate=1e-3.
In funetune stage, batch_size=256, num_epoch=150, learning_rate=1e-4.
The final loss of Baseline is 3.794, the one of Phonetic Embedding is 3.7857.

Besides the model pretraining and finetuning, we have 3 main tasks to evaluate the model, decribed in Table 5.3.

5.4 Results

The result of first task is shown in Table 5.3 and Figure 3. As for task 2 and task 3, we only care about Average Rhyme Score, and it's show in Table 5.4.
We have selected some incredibly high-quality poems, as well as some good examples of Cangtoushi and Xiangqianshi. Please refer to Table 5.4 for their review.

No.	Name	Purpose	Method
1	Uniform Poem	Evaluate the basic ability of model	Use the first character of the 4 sentences of every poem in the corpus to generate
2	Non-Uniform Poem	Evaluate the generalization ability for various-length poems	Define various lengths format, and use the previous method to generate
3	Out-of-vocabulary	Evaluate the generalization ability of phonetic embedding	Make a new token into vocabulary, assign semantic embedding and phonetic embedding to generate

Table 5: Definition of 3 main tasks

Metric	Corpus	Baseline	Phonetic Embedding
Average Rhyme Score	0.556	0.558	0.553
Rhyme Diversity	1.0	0.952	0.952
Rhyme Token Diversity	0.697	0.191	0.190
Token Position Diversity	0.844	0.645	0.644
Token Diversity	1.0	0.701	0.703
Distinct-1	0.909	0.921	0.922
Distinct-2	0.962	0.961	0.961
Distinct-3	0.926	0.923	0.923
Distinct-4	0.890	0.884	0.884

Table 6: Task 1 result

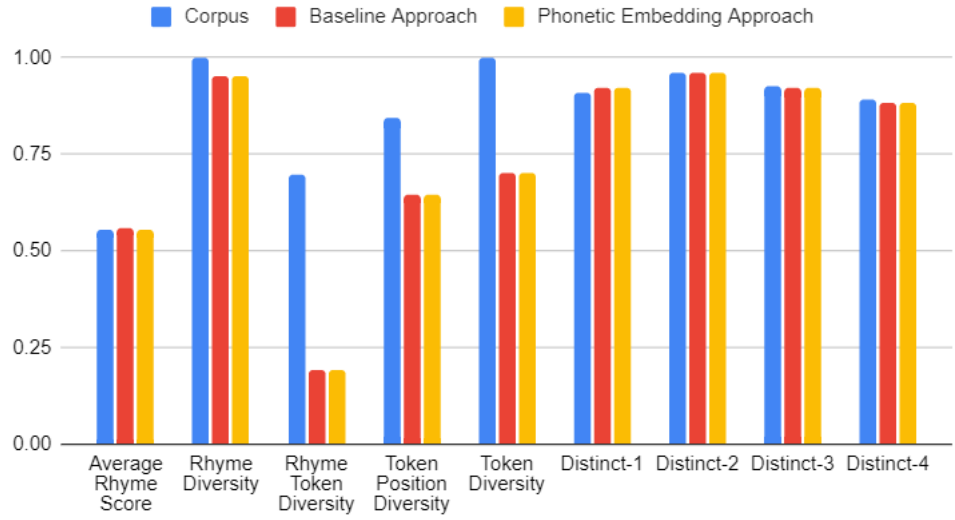

Figure 3: Result 1: Comparison between Corpus, Baseline and Phonetic Embedding

There are also some excellent generated poems with both high-quality meaning and rhyming. To display their meanings, we have included the English version in the appendix.

6 Analysis

Based on the evaluation of Task 1 shown in Table 5.3, it can be observed that both the baseline and phonetic embedding approaches have achieved good results in comparison to the corpus. Except the diversity related to tokens, the main metrics are quite similar to those of the corpus. This indicates that the bidirectional model has successfully captured the majority of poetry's characteristics, enabling it to produce excellent new poems as a language model (See Table 5.4 for some example of human evaluated high-quality poems).
Regarding the Phonetic Embedding approach, the evaluation of Task 2 and 3 presented in Table 5.4 demonstrates its significant superiority over the baseline approach. The baseline approach can hardly generate rhyming sentences in such cases. However, the phonetic embedding approach can achieve rhyming to a certain extent. This highlights the high potential of Phonetic Embeddings. Though, when compared to the average rhyming score of the corpus and Task 1, the resulting score is still too low. We believe that this is due to the loss of some information from both the semantic

Task	Baseline	Phonetic	Improvement over baseline
2	0.078	0.203	260.3%
3	0.0495	0.104	210.1%

Table 7: Task 2 and 3 result of rhyming score

Category	Poem	Poem
Human Evaluated High-Quality	每忆何人访旧游 海边独上最高楼 风烟目断无鸿雁 万里相思一点愁	刘阮同游颍水滨 柴门不掩隔埃尘 闲来白发思归客 每到青山忆故人
Human Evaluated High-Quality	峒岭东边望白云 四时风雨乱缤纷 山头花色朝阳见 谷口泉声到处闻	隔岸云山空碧 临水桃李深红 杨柳一般夜雨 芙蓉千里春风
Multiple applications of Cangtou(藏头), Cangwei(藏尾) and Xiangqian(镶嵌)	新恭若夫子 年少喜婴孩 快哉白发翁 乐此万钱财 (新年快乐 + 恭喜发财)	我爱清阴密叶 喜闻细雨新文 欢意🄵花皎洁 你心蜂蝶缤纷 (我喜欢你 + 叶文洁)
Out-of-vocabulary rhyming example using the letter 'A'	车马往来三十里 门前杨柳如流水 禾头互见若夫差 一点天青山🄵A ('A' rhymes with 水)	丁宁玉帛若夫子 见缀珠玑走神鬼 和了新诗日暄妍 地近画屏山🄵A ('A' rhymes with 鬼)
Non-uniform poem rhyming example	临济对面一笑 选佛比肩可知 吾祖西南大夫师 无是天下之事	一点虚空世界 光明普照丛林 森罗万古无限今 亘然不二知音

Table 8: Examples of generated poems in different categories

and phonetic embedding when they are simply added together in the out-of-vocabulary case. Additionally, for the Non-uniform case, there were no such poems in the corpus at all for the model to learn. To address these issues, we are considering the following improvement plans:

1. Augment the data into various length series to increase the model's exposure to diverse poem structures.
2. Employ a new algorithm or architecture to combine semantic embedding and phonetic embedding in a more effective way rather than simply adding them together.
3. Expand the training corpus by incorporating Songci (a type of Chinese poetry), Yuanqu (a form of Chinese opera), and other forms of poetry.

7 Conclusion

We propose a bidirectional poetry generation model that employs phonetic embedding to enhance its generalization ability, with the aim of facilitating artistic production. Through the design of two training methods with BERT and three evaluation tasks, we have demonstrated the exceptional ability of the bidirectional model to generate high-quality poems, including various forms of acrostic Chinese poetry(Cangtoushi, Cangweishi, Xiangqianshi). Additionally, it has shown great potential for creating poetry with out-of-vocabulary tokens or in various non-uniform length series, although further refinement is necessary.
For future work, we will make use of more kinds of datasets, employ data augmentation, and we believe it will help a lot to make it a bidirectional encoder-decoder architecture with more well-designed training procedure.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models.

Yi Liao, Yasheng Wang, Qun Liu, and Xin Jiang. 2019. Gpt-based generation for classical chinese poetry.

Piji Li Haisong Zhang Xiaojiang Liu Shuming Shi. 2021. Rigid formats controlled text generation.

A Appendix

HIGH-QUALITY GENERATED POEMS

Common - Landscape - Rhyming Score 0.8

峒岭东边望白云
一时风雨乱缤纷
山头花色朝阳见
谷口泉声到处闻

From Dongling's east, I gaze up at the cloud-filled sky,
And suddenly the wind and rain rush in, their colors awry.
Amidst the flowers atop the mountain, the sun appears,
And at the valley's entrance, the sound of spring water cheers.

Common - Longing, Love - Rhyming Score 0.9

Whenever I yearn for her, I visit the places we roamed,
By the shore, I climb the highest building, lost and alone.
Sea breeze and mist shroud the beach, wild geese unseen in the distance,
My cherished one is thousands of miles away, my heart heavy with absence.

每忆何人访旧游
海边独上最高楼
风烟目断无鸿雁
万里相思一点愁

Common - Landscape - Rhyming Score 0.8

隔岸云山空碧
烟水桃李深红
杨柳一般夜雨
芙蓉千里春风

Green hills beyond the coast, white clouds, skies like jasper bright,
Peaches and plums bloom crimson at the water's edge in delight.
Soft night rain falls like willows outside the window's glass,
Lotus flowers sway in the spring breeze on the pond, such a sight to amass.

Acrostic - Lonely - Rhyming Score 0.9
Keywords: **乌兹，永远的神。**(A meme of Chinese LOL competition)
Keyword Meaning: UZI, you are the eternal God in my heart.

In the night so long, crows sing a song,
Full moon beams, bright all night long.
Wild geese take flight, in the dark of night,
White clouds drift by, a leisurely sight.
Mountain climb's path, not hard to find,
The peak's in sight, within your mind.
Yet my restless soul, like the vast sea,
No place to hold, forever free.

乌啼夜**永**明月满
雁飞天**远**白云闲
兹山标**的**非难事
沧海精**神**杳无间

Acrostic - Love - Rhyming Score 0.8
Keywords: **我喜欢你，叶文洁。**(I love you, Wenjie Ye)

我爱清阴密**叶**
喜闻细雨斯**文**
欢意繁花皎**洁**
你心蜂蝶缤纷

I love the shade of leaves that's cool and dense,
And I love the sound of a light rain, that's gentle and intense.
My mood is eager, bright like blooming flowers,
While your heart flutters like elusive bees and butterflies in showers.