Exploring the potential of data augmentation in poetry generation with small-scale corpora

Renxiang Huang

College of Liberal Arts & Sciences, University of Illinois Urbana-Champaign, Champaign, IL, 61820, USA

rh25@illinois.edu

Abstract. Poetry generation is a complex task in the field of natural language processing, especially when working with small datasets. Data augmentation techniques have been shown to be an effective way to improve the performance of deep learning models in various tasks, including image classification and speech recognition. Therefore, this study focuses on exploring the impact of four different data augmentation methods - Synonym Replacement, Random Insertion, Random Swap, and Random Deletion - on the performance of poetry generation with a small poetry dataset. The results of the study reveal that Random Insertion performed well in terms of Bilingual Evaluation Understudy (BLEU), Recall-Oriented Understudy for Gisting Evaluation (ROUGE), and manual evaluation when compared to other data augmentation techniques. Synonym Replacement performed poorly in all three evaluations. This study confirms the potential value of data augmentation technology in poetry generation tasks and provides innovative perspectives and directions for future research in this area. Data augmentation can be employed to help address the problem of limited data in poetry generation tasks and enhance the efficiency of deep learning models. Future research could focus on exploring more advanced data augmentation techniques and their impact on poetry generation tasks.

Keywords: Data Augmentation, Poetry Generation, Natural language Processing, Deep Learning.

1. Introduction

Poetry is an important literary genre with a long history. Poetry can be traced back to prehistoric times, when people wrote poems to record history, legends or sacrifices. Different from other literary genres, poetry often pays more attention to musicality, brevity, and rhetoric in addition to describing plots or characters. In recent years, as artificial intelligence and machine learning rapidly advanced, how to apply natural language processing technology to realize poetry generation has gradually become a popular research topic. However, because of the above-mentioned characteristics of poetry, how to make the generated poetry more natural and maintain rhythm is still a challenging problem.

Traditional poetry generation techniques mainly rely on large-scale training data. Large and diverse datasets can allow models to capture sufficient features [1]. However, due to the brevity of poems, the related datasets are very limited in size, which will lead to insufficient diversity and depth of generated poems. In addition, the small corpus will also lead to overfitting of the model during training, making it difficult for the model to generalize well [2]. This issue will result in the poor model performance and the generated poems quality. Therefore, solving the small corpus problem is a necessary and urgent task in the field of poetry generation.

In machine learning and deep learning tasks, data augmentation is a common method to expand data sets against weakness of the small original data sets. Data augmentation aims to generate new samples from existing data through transformations, thereby making the original data set larger and more diverse. This technique has already obtained impressive achievements in the field of computer vision. Although data augmentation increases the correlation between the data, this technology significantly reduces the risk of model overfitting [3]. However, in the domain of natural language processing, especially in the field of poetry generation, there is a lack of related research and applications. In recent years' researches, Wei and Zou points out that data augmentation has significantly improved the recognition accuracy of the model in text classification tasks [4]. This research indicated that data augmentation has great potential for application in natural language processing tasks. Therefore, this article aims to explore and verify the impact of data augmentation technology on the small poetry dataset, so as to improve the performance of the model and the quality of poetry generation.

This article will first apply different data augmentation techniques, including synonym replacement, random insertion, random swap, and random deletion, to a small poetry dataset for poetry generation, then evaluate the effectiveness of each method and analyse in detail its impact on the quality of generated poems.

2. Method

2.1. Dataset

The dataset comes from the "poetry" dataset published in hugging face [5]. The original dataset contains 573 poems with their author, title, age, and type. Among them, the "age" is divided into Renaissance and Modern, and the "type" is divided into Mythology & Folklore, Nature, and Love. In this article, only modern poems are kept in order to construct a small dataset for poetry generation. Then, the columns of author and title which are irrelevant to the poetry generation task are also removed. In addition, "type" is retained to test whether the content generated fit the type it belongs to. To make this dataset more suitable for the task of generation, the content of each poem is split at every newline character so that each row of the data set is a line of poetry instead of an entire poem. Therefore, there are a total of 4795 rows and two columns in the dataset used in this article. The examples of this dataset are shown in Table 1.

Table 1. Representative examples of the dataset.

content	type
when no soul may ever escape the eternal destiny of life	mythology & folklore
the rain beats down until the slow	nature
love god and keep his commandments	love

2.2. Data Augmentation

To overcome poor poetry generation quality on this small dataset, this article explores four common data augmentation methods: synonym replacement, random insertion, random swap, and random deletion.

Synonym replacement refers to selecting multiple words of the original sentence at random and replacing them with other words with similar or identical meanings. After synonym replacement, the sentence length remains the same as the original sentence and it is necessary to ensure that the new sample is semantically similar or identical to the original sentence. In this article, this method aims to search synonyms for 80% of the words in the original sentence and replace them if synonyms are found.

The synonym database “wordnet” from Princeton University is used as the source of the synonym replacement [6]. The database “wordnet” groups together words that have the same or similar meaning, including nouns, verbs, adjectives, and adverbs. The examples of this method are shown in Table 2.

Table 2. Examples generated by synonym replacement method.

content	type
when no soulfulness may ever scat the unending fate of aliveness	mythology & folklore
the rainwater beats down until the decelerate	nature
dearest immortal and restrain his commandments	love

Random insertion refers to choosing a word from a sentence at random and inserting it into a random position of the original sentence. This method makes the new sentence one word longer than the original sentence. The examples of this method are shown in Table 3.

Table 3. Examples generated by random insertion method.

content	type
when may no soul may ever escape the eternal destiny of life	mythology & folklore
the rain beats down beats until the slow	nature
love god and keep his and commandments	love

Random swap refers to randomly selecting two words in a sentence and swapping them to get a new sentence. The examples of this method are shown in Table 4.

Table 4. Examples generated by random swap method.

content	type
when no soul may destiny escape the eternal ever of life	mythology & folklore
slow rain beats down until the the	nature
his god and keep love commandments	love

Random deletion refers to randomly selecting some words in a sentence to delete them, and keeping the rest as a new sentence. In this article, for each word in the sentence, this method generates a random number between 0 and 1. If the random number is greater than p which is set to 0.8, then the word will be deleted. The examples of this method are shown in Table 5.

Table 5. Examples generated by random deletion method.

content	type
when no may ever escape the eternal of	mythology & folklore
the rain down the slow	nature
keep his commandments	love

2.3. Poetry Generation

After applying these four data augmentation methods on the original dataset, four new datasets will be obtained. The 80% of each dataset is allocated as the training dataset, and then the remaining 20% of the dataset is reserved as the testing dataset. The generating model selected is Long Short-Term Memory (LSTM). LSTM is an evolutionary version of Recurrent Neural Network (RNN) structure. LSTM is designed for processing sequence data and can capture long-term dependencies [7]. The results based on the original dataset are used as a benchmark to compare with the generated results after data augmentation.

To generate results, the "prompting strategy" is applied. The "prompting strategy" means using some prompts, such as words, as the contexts to guide the following text generation. The prompts should be clear but not too verbose; otherwise, it will affect the speed and quality of text generation. It is easy and potential to achieve text generation based on a "prompting strategy" [8]. First, one hundred different sentences in the testing dataset were chosen from each of three poetry types as the reference sentences. Then, only the first three words of each sentence were reserved as the prompts, and the other parts were dropped. These prompts will be transmitted into the model as inputs. Eventually, the model will generate the continuations based on these prompts.

2.4. Evaluation Indexes

Bilingual Evaluation Understudy (BLEU) and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) are applied as the quantitative evaluation scores in this article. These two methods calculate the degree of match between the model-generated results and the reference poems. Given the prompts, the continuations generated by the model will compare with the reference sentences. The BLEU score refers to the number of n-grams in the generated results that match with the reference sentences [9]. The ROUGE scores are divided into three criteria: ROUGE-1, ROUGE-2, and ROUGE-L [10]. The ROUGE-1 score refers to the percentage of overlap between unigrams in the generated results and the reference sentences. The ROUGE-2 score stands for the percentage of overlap between bigrams in the generated results and the reference sentences. The ROUGE-L score represents the percentage of overlap between the longest common subsequence in the generated results and the reference sentences. Although the BLEU and ROUGE are commonly applied to assess the quality of machine translation and text summarization, both methods still provide a measure of evaluating the superficial quality of a poem quickly and roughly.

Due to the characteristics of the poetry, manual evaluation is still an essential approach. Merely relying on quantitative scores is insufficient, the creativity of poetry requires the intervention of manual evaluation [11]. In this article, one prompt outside the dataset was set for each poetry type. Each prompt contains three words. For the Mythology & Folklore poetry, the prompt was set as "dragons breathe fire". For the Nature poetry, the prompt was set as "sunlight dapples leaves". For the Love poetry, the prompt was set as "love conquers all". The results from the models based on the different datasets compared to each other for integrity, coherence, emotion, and meaning. These four aspects in this article are defined as follows:

Integrity: Whether the generated sentence is complete.

Coherence: Whether the generated sentence is smooth and logical.

Emotion and meaning: Whether the generated sentence match the poetry type and express corresponding connotation.

3. Result

The results of the BLEU and ROUGE score are shown in Table 6. The average scores are obtained by taking the average of the BLEU scores calculated by 300 pairs of reference and generated sentences.

Table 6. Result measured by BLEU and ROUGE score.

	Original Dataset	Synonym Replacement	Random Insertion	Random Swap	Random Deletion
Average BLEU	0.2078	0.2050	0.4897	0.2722	0.2928
Average ROUGE-1	0.6276	0.6852	0.8469	0.7366	0.7681
Average ROUGE-2	0.4961	0.5705	0.7067	0.5439	0.6286
Average ROUGE-L	0.6276	0.6838	0.8457	0.7024	0.7666

For the original dataset, there are 20.78% of the n-grams in the generated results matching with the reference sentences. The 0.6276 of the ROUGE-1 score and the ROUGE-L score means that the results is acceptable at the unigrams and the longest common subsequence levels, but 0.4961 of the ROUGE-2 score means there is significant room for improvement at the bigrams level. These scores indicate the weak model performance based on the original dataset. For the Random Insertion method, the average BLEU score achieved 0.4897, which is the best among the scores of the original dataset and the other methods. And the three kinds of the ROUGE scores of the Random Insertion method are also the highest.

To obtain a general view of the comparison among the original dataset and four different data augmentation methods, the line chart of the scores is shown in the Figure 1.

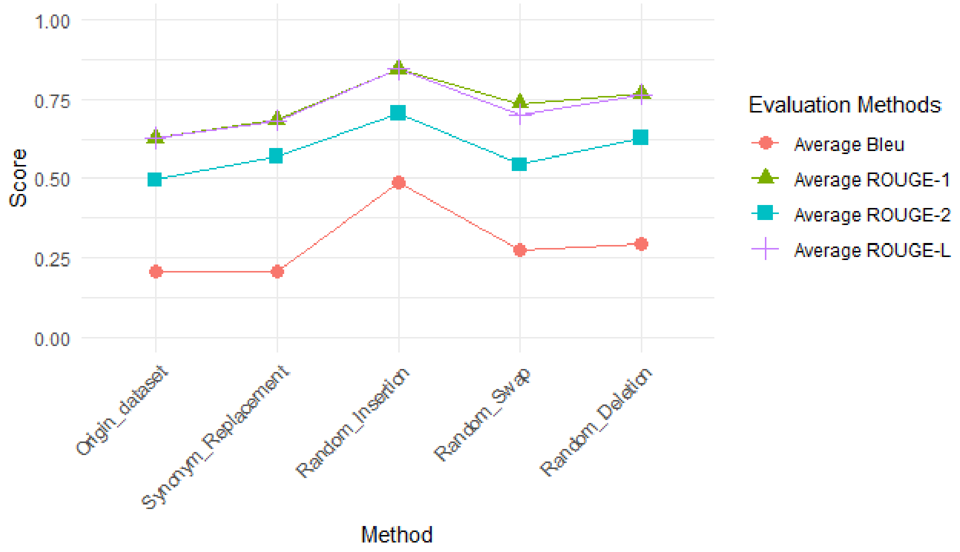

Figure 1. Result comparison of various data augmentation methods (Figure Credits: Orignal).

From the Figure 1, one can conclude that the four data augmentation methods have significantly improved the BLEU and ROUGE scores in comparison to the original data set. Although the BLEU score of the original dataset is close to that of the Synonym Replacement, the other scores of the Synonym Replacement still increased compared with the original dataset.

For the manual evaluation, the generated results given the prompts mentioned is Section Methods are shown in Table 7.

Table 7. Example of generated results.

	Original dataset	Synonym Replacement	Random Insertion	Random Swap	Random Deletion
Mythology & Folklore	dragons breathe fire of the moonlight against a earth and	dragons breathe fire now and time we in the level	dragons breathe fire dreams on the air turns and down	dragons breathe fire and my heart and think to be	dragons breathe fire was a man of a dying now
Nature	sunlight dapples leaves the air of a tobacco trance on	sunlight dapples leaves upon the paddle in the ocean earth	sunlight dapples leaves at the depth of the drippling tide	sunlight dapples leaves the on the we we feel my	sunlight dapples leaves too and and then the heart is
Love	love conquers all the year in a dark box and	love conquers all me too o mother who wrest my	love conquers all that men they should wake to sing	love conquers all that in substance and waterblobs save the	love conquers all the works and days of hands are

For the original dataset, these three results end with unnecessary prepositions or conjunctions, that are incomplete sentences. In terms of emotion and meaning, the first two sentences match relatively well with the mythology & folklore and natural styles respectively, but the third sentence reflects no content about love.

For the Synonym Replacement method, the generated sentences are better than the sentences based on the original dataset in terms of structural integrity. In terms of emotion and meaning, the generated sentence of nature type has begun to show the charm of poetry. There is a certain connection between “sunlight”, “paddle”, and the “ocean”. However, even if the first and third sentences are complete, their meanings are confusing.

For the Random Insertion method, the structural integrity and coherence of the three sentences had been strengthened. The emotion and meaning are also match better with the types of the poem belongs to. For example, the sentence “sunlight dapples leave at the depth of the drippling tide” paints a vivid picture, combining natural imagery of sunlight with rhythmic tides. The sentences of the remain two methods, Random Swap and Random Deletion, are neither complete nor convey the appropriate emotion and meaning, making it difficult to meet the standards of the poetry.

In summary, the performance of the Random Insertion method is the best data augmentation through the above experiment.

4. Discussion

This article tests four data augmentation methods and demonstrates the potential application value of data augmentation in the field of poetry generation by comparing the quantitative scores and manual evaluation of the four methods.

Data augmentation extends the small corpus by applying transformation to the original data, significantly improving the quality of poetry generation. Through the comparison of the results in the Section Result, it can be inferred that the Random Insertion method achieved the best augmentation results. The Random Insertion method retains the original intent of the poetries while enriching the original dataset. This method ensures that the core meanings of the original content are not distorted or lost. In contrast, Synonym Replacement has the worst results, which is worse than the original dataset in terms of average BLEU score. Considering the intrinsic musicality and rhetoric of poetry, although the words were replaced by their synonyms, this method still destroyed the rhythm and meaning of the poems, yielding unsatisfactory results by this method. Kobayashi proposed a novel data augmentation method similar to synonym replacement, named Contextual Augmentation [12]. This method replaces the original word with a new word predicted based on the context of the original word. The reasonable predictions can preserve a greater portion of the original meaning than simple synonym replacement. In subsequent research, it can be considered to narrow down the range of synonyms search based on the emotions and meanings of the poetry through innovative approaches, so that synonym replacement can retain the original intent to the greatest extent possible.

In the experiments, the performances of two data augmentation methods, Random Swap and Random Deletion, is between Random Insertion and Synonym Replacement. Random Swap introduced new context and structure by changing the order of words in a poem, which also slightly disrupted the original rhythm and meanings of the poems. And Random Deletion method helps the model extract the core meanings of the poems effectively by removing some words, but at the same time, it can also lead to the loss of some crucial information. One can focus on using more robust methods to expand the dataset in the future such as using adversarial generation to obtain more realistic poems [13]. Hence, under the situation of the limited data, a reasonable data augmentation strategy can improve the performance of the model significantly, but it also needs to be selected depends on specific application.

This article mainly explores the effectiveness of data augmentation technology, so the general LSTM model is chosen as the generative model instead of more advanced language models, such as Transformer or GPT [14,15]. For further research in the future, the application of more effective models coupled with reasonable data augmentation techniques can be considered. In addition, as mentioned in Section Method, the BLEU and ROUGE score mainly focus on the overlap of the vocabulary, even though these two criterions well measured the similarity of the generated poems to the original poems, they are not enough to fully assess the quality and emotion of the poems. Thus, the evaluation standard specific to poetry require further research in the future. For example, it is noteworthy that Oliveira et al. attempted to develop an automatic poetry evaluation algorithm to rate the content and rhythm of the generated poetry [16].

5. Conclusion

This study demonstrates the potential of data augmentation techniques to improve the generation of small poetry corpora. The results show that the Random Insertion method surpasses other methods in both BLEU and ROUGE scores, also in manual evaluation. In stark contrast, due to its inherent limitations, Synonym Replacement has difficulty maintaining the unique musicality and rhetoric of poetry, and therefore performs poorly. Meanwhile, the Random Swap and Random Deletion methods have their own advantages and challenges in different poetry types.

In addition, the model built in this study is the general LSTM model. Data augmentation techniques perform well on this base model. This finding indicate that combined with strategic data augmentation, the improvement effect of the other advanced generative models could be more significant and obtain more ingenious results. Moreover, the objective and quantitative criteria in this article rely on BLEU and ROUGE scores. These two scores have crucial gaps in assessing the quality and emotion of poetry. Hence, more advanced and refined evaluation criteria are required in future research to completely capture the complexity of poetic expression.

References

[1] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

[2] Li, J., Tang, T., Zhao, W. X., Nie, J. Y., & Wen, J. R. (2022). Pretrained language models for text generation: A survey. arXiv preprint arXiv:2201.05273.

[3] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 1-9.

[4] Wei, J., & Zou, K. (2019). Eda: Easy data augmentation techniques for boosting performance on text classification tasks. arXiv preprint arXiv:1901.11196.

[5] Merve Noyan. Poetry. URL: https://huggingface.co/datasets/merve/poetry. Last Accessed: 2023/09/17.

[6] WordNet Princeton University. URL: https://wordnet.princeton.edu/. Last accessed 2023/09/23.

[7] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

[8] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys, 55(9), 1-35.

[9] Yan, R. (2016). i, Poet: Automatic Poetry Composition through Recurrent Neural Networks with Iterative Polishing Schema. In IJCAI, 2238, 2244.

[10] Yan, R., Jiang, H., Lapata, M., Lin, S. D., Lv, X., & Li, X. (2013). I, poet: automatic chinese poetry composition through a generative summarization framework under constrained optimization. In Twenty-Third International Joint Conference on Artificial Intelligence, 2197-2203.

[11] Van de Cruys, T. (2020). Automatic poetry generation from prosaic text. In Proceedings of the 58th annual meeting of the association for computational linguistics, 2471-2480.

[12] Kobayashi, S. (2018). Contextual augmentation: Data augmentation by words with paradigmatic relations. arXiv preprint arXiv:1805.06201.

[13] Zhang, Y., Gan, Z., & Carin, L. (2016). Generating text via adversarial training. In NeurIPS workshop on Adversarial Training, 21, 21-32.

[14] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al (2017). Attention is all you need. Advances in neural information processing systems, 30, 1-9.

[15] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training, 1-12.

[16] Oliveira, H. G., Hervás, R., Díaz, A., & Gervás, P. (2017). Multilingual extension and evaluation of a poetry generator. Natural Language Engineering, 23(6), 929-967.