Generating Wuxia Novels with LSTM and Transformer Models

Xinrui Yan
yanxinrui2021@ia.ac.cn

March 2025

Abstract

This project explores the effectiveness of LSTM and Transformer-based models in generating Chinese wuxia (martial arts) novel text, using a corpus composed of classic works by Jin Yong. We train both models on the same dataset and compare their performance in generating stylistically consistent, semantically meaningful content. Our analysis covers perplexity scores, sample quality, and model behavior. Results suggest that while Transformers generate more coherent long-term dependencies, LSTMs show more stability on smaller datasets.

1 Introduction

Natural Language Processing (NLP) has witnessed significant advancements in recent years, particularly in the field of text generation. Among various genres of generated text, Chinese martial arts fiction (wuxia) presents a unique challenge due to its complex narrative structures, rich cultural references, and distinct writing style. This project focuses on generating wuxia-style novels using two popular neural architectures: Long Short-Term Memory (LSTM) and Transformer models. Using a dataset composed of 18 classical works by Jin Yong, we aim to train and compare these models in their ability to capture stylistic and semantic nuances of wuxia literature. While LSTM has been a traditional choice for sequence modeling, the Transformer architecture has emerged as the state-of-the-art for many generation tasks. By analyzing the outputs and training behaviors of both models, this study provides insights into the strengths and limitations of each approach in modeling long-range dependencies and narrative coherence in literary text generation.

2 Dataset and Preprocessing

The dataset consists of 18 plain-text files, each representing a complete novel by the renowned wuxia author Jin Yong. The files include:

The Deer and the Cauldron.txt, Demi-Gods and Semi-Devils.txt, The Smiling, Proud Wanderer.txt, Heaven Sword and Dragon Saber.txt, The Return of the Condor Heroes.txt, The Legend of the Condor Heroes.txt, The Book and the Sword.txt, Sword Stained with Royal Blood.txt, Flying Fox of Snowy Mountain.txt, Ode to Gallantry.txt, A Deadly Secret.txt, Fox Volant of the Snowy Mountain.txt, White Horse Neighs in the Western Wind.txt, The Thirty-Three Swordsmen.txt, The Blade of the Mandarin Ducks.txt, Sword of the Yue Maiden.txt.

Preprocessing steps included text normalization, segmentation, and vocabulary construction.

Text preprocessing involved:

- Unicode normalization and traditional-to-simplified Chinese conversion;
- Removing chapter headings and extraneous punctuation;
- Tokenization using the Jieba library;
- Building a character-level and word-level vocabulary.

3 Model Architectures

3.1 LSTM Model

A six-layer LSTM language model was implemented. Each layer consists of 512 hidden units, and the input embeddings are of size 256. Cross-entropy loss was used, and gradient clipping was applied to prevent exploding gradients.

3.2 Transformer Model

We used a vanilla Transformer encoder-decoder architecture. Configuration details:

- Number of layers: 6;
- Number of attention heads: 8;
- Embedding dimension: 512;
- Feed-forward size: 1024.

Positional encoding was applied to maintain token order.

4 Training Setup

The training strategies were used:

- **From-scratch training:** models trained solely on the Jin Yong corpus;

Both models were trained with the Adam optimizer, a learning rate of 1×10^{-4}, a batch size of 64, and up to 20 epochs.

5 Results and Analysis

5.1 Sample Outputs

To qualitatively evaluate the models, we generated text samples using both the trained LSTM and Transformer models. The following examples are produced after training on the Jin Yong corpus:

LSTM Output:

> 1. 张无忌说，六大派围攻光明顶魔卫道不，什么为了不一定，不知道，赵敏眼睛闪过一点狡猾，笑了一下，天下女子倾心她为什么这样。
>
> 2. 赵敏说，张公子心思透，兵符除非自己来，拿不想交，我赵敏看透他不，看不透心情这样，衣袖拂了下，眼睛很狡猾，结果自己不知道怎么办。
>
> 3. 周芷若说，师父我命不敢违心里张不能放下，泪光眼，低头说，来生见不想再见你，或者怎么都可以反正从未见过他。

Transformer Output:

> 1. 张无忌说，他是张无忌，明教的教主，而这个教主就应该心系天下，不能让中原武林再这样乱下去。环顾着，声音也不算大，压得大家都不能动。
>
> 2. 赵敏说，自己说不动心其实就是在骗他，她赵敏生在金帐王庭，所有的事她都不会随便做决定，看到张无忌时心里早就有些动摇，硬是装冷静说救明教就随便他一次。
>
> 3. 周芷若说，峨眉是修行的地方，她本该无情的，哪里会有那么多痛苦，素衣飘在风里，却痛的深深，风中转身，素衣飘动，张无忌这生恩怨清了，谁都忘了。

5.2 Quantitative Evaluation

To systematically assess model performance, we conducted a quantitative evaluation using two key metrics: Perplexity and BLEU score. These metrics respectively reflect the model' s ability in language modeling and sentence generation accuracy.

- **Perplexity:** LSTM = 45.6, Transformer = 38.2;
- **BLEU score (against true next-sentence):** LSTM = 0.165, Transformer = 0.190;

Perplexity Perplexity measures the model's uncertainty when predicting the next word in a sequence. A lower perplexity indicates higher predictive accuracy and better language modeling capabilities. In our experiments, the Transformer model achieved a perplexity of 38.2, significantly lower than LSTM' s 45.6. This

suggests that the Transformer is more effective in capturing contextual information, especially in texts with long-range dependencies, which are common in wuxia narratives.

BLEU Score BLEU (Bilingual Evaluation Understudy) score is a metric originally designed for machine translation, used here to assess the similarity between the model-generated next sentence and the ground-truth sentence from the corpus. The Transformer model scored 0.190, slightly outperforming the LSTM' s 0.165. Although the numerical gap is modest, it indicates a superior ability of the Transformer to maintain semantic coherence and structural alignment with the original text.

5.3 Advantages and Disadvantages

Model	**Advantages**	**Disadvantages**
LSTM	Simpler to train on small data; stable convergence	Poor at modeling long-term dependencies; prone to repetition
Transformer	Excellent long-range coherence and parallelism	Requires more data and compute; may overfit small corpora

Table 1: Comparison of LSTM and Transformer

6 Conclusion

This study investigated the performance of LSTM and Transformer models in the generation of Chinese wuxia text based on Jin Yong novels. While LSTM models are more efficient on small datasets, Transformer models produce significantly more coherent and contextually rich content. Future work may include incorporating reinforcement learning for better control over plot structure, and exploring diffusion or retrieval-augmented architectures for stylistic refinement.

7 References

References

[1] Vaswani, A., et al. "Attention is All You Need." *NeurIPS*, 2017.

[2] Greff, Klaus, et al. "LSTM: A search space odyssey." IEEE transactions on neural networks and learning systems 28.10 (2016): 2222-2232.

[3] Radford, A., et al. "Language Models are Unsupervised Multitask Learners." OpenAI, 2019.

[4] Han, Kai, et al. "A survey on vision transformer." IEEE transactions on pattern analysis and machine intelligence 45.1 (2022): 87-110.