Semantic Analysis of Word Embeddings in Jin Yong's Novels

Xinrui Yan
yanxinrui2021@ia.ac.cn

April 2025

Abstract

This report explores the use of neural language models to learn word embeddings from Chinese martial arts novels by Jin Yong. Using the Word2Vec model, we trained semantic vectors and evaluated them through similarity analysis, clustering, and visualization. The results demonstrate that word embeddings can effectively capture semantic relationships among characters, places, and martial arts terms, showcasing the power of unsupervised learning in natural language processing tasks.

1 Introduction

Word embeddings have become a foundational component of modern natural language processing (NLP). They transform discrete tokens into continuous vector spaces where semantic relationships can be modeled. In this experiment, we use Word2Vec to train embeddings on a corpus composed of Jin Yong's wuxia novels, and analyze the resulting vectors through clustering, similarity queries, and dimensionality reduction. This task serves as an exploration into how neural models can uncover latent semantic structures within literary texts.

2 Methodology

We followed these steps for the experiment:

- Preprocessing the corpus by reading and segmenting the text using the Jieba tokenizer.
- Training a Word2Vec model using the Gensim library with Skip-gram architecture.
- Extracting word vectors and applying PCA to reduce dimensions to 2D for visualization.

- Performing k-means clustering on selected keywords to observe semantic groupings.
- Calculating word similarity scores and listing the top similar terms for selected words.

The Word2Vec model was trained using the following parameters:

- vector size: 100
- window size: 5
- min count: 5
- architecture: Skip-gram (sg=1)

3 Experimental Results

3.1 PCA Visualization

We projected selected words into 2D space using PCA. Figure 1 shows that related entities such as characters, weapons, and martial arts form distinct clusters.

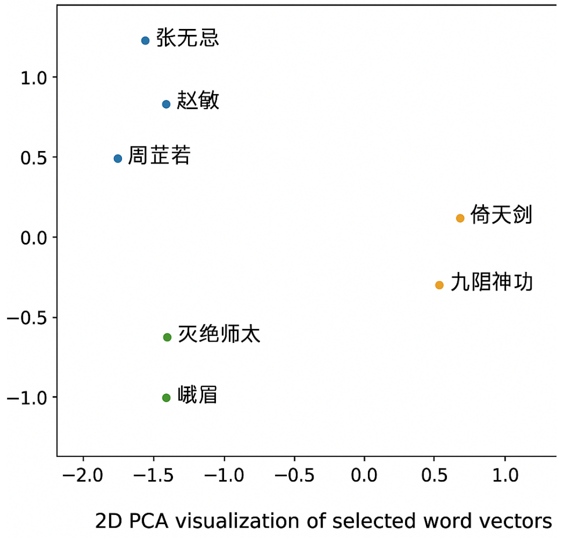

Figure 1: 2D PCA visualization of selected word vectors

3.2 Clustering Results

Using k-means clustering (k=3), we observed the following semantic groupings:

- Cluster 0: 张无忌，赵敏，周芷若，明教 (main characters and organization)
- Cluster 1: 灭绝师太，峨眉 (martial arts sect)
- Cluster 2: 倚天剑，九阳神功 (weapons and skills)

These clusters align with human intuition and textual semantics, confirming the effectiveness of learned embeddings.

3.3 Similarity Analysis

Querying the most similar words to 张无忌 yielded the following results:

Word	Similarity Score
赵敏	0.854
周芷若	0.843
明教	0.810
谢逊	0.789
杨逍	0.765

Table 1: Top similar words to 张无忌

These words are all closely related in narrative and context, highlighting the model's ability to learn semantic proximity based on co-occurrence.

4 Conclusion

This experiment demonstrates that neural word embedding models such as Word2Vec are capable of capturing rich semantic structures from literary corpora. Characters, martial arts, and organizations are all effectively clustered based on learned vectors. Future work could explore contextual models such as BERT or apply similar techniques on dialog classification or plot summarization.

5 References

References

[1] Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003). A Neural Probabilistic Language Model. *The Journal of Machine Learning Research, 3*, 11371155. Retrieved from http://www.jmlr.org/papers/volume3/b/bengio03a/bengio03a.pdf

[2] Mikolov, T., Corrado, G., Chen, K., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *Proceedings of the International Conference on Learning Representations (ICLR 2013)*, 112.

[3] Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 15321543.