Word Vectors of Jin Yong Novels

Sanfeng Xin
by2410222@buaa.edu.cn

Abstract

Word embedding models play a pivotal role in natural language processing by mapping words into low-dimensional dense vectors that capture semantic relationships. This study applies two widely used word embedding methods—Word2Vec and GloVe—to a corpus consisting of Jin Yong’ s sixteen martial arts novels, aiming to construct semantic representations of vocabulary within the wuxia (martial arts fiction) context. For evaluation, I employed word similarity measurements to compare the models’ abilities. Preliminary results indicate that Word2Vec is more effective.

Introduction

Word embedding techniques such as Word2Vec and GloVe have become foundational tools in natural language processing (NLP), enabling effective semantic representation of textual data. By encoding words as dense vectors in a continuous space, these models capture syntactic and semantic relationships that are crucial for downstream tasks like classification, clustering, and information retrieval.

In this study, we investigate the effectiveness of Word2Vec and GloVe in learning word embeddings from a corpus of Jin Yong’ s sixteen wuxia novels.Our approach includes corpus preprocessing tailored to Chinese text, model training and evaluation.

Methodology

Data Preparation

The dataset for this study consists of Jin Yong’s sixteen martial arts novels, which serve as the textual corpus for topic modeling and classification. To ensure consistency and reliability in text processing, the following preprocessing steps were applied:

1.Remove useless content: The text was cleaned to remove any irrelevant content, such as advertisements at beginning and the end of the texts and blank characters. This step ensures that the analysis focuses solely on the narrative content of the novels.

2.Some novels have line break errors, and we have performed normalization.

3.Conversion from Traditional to Simplified Characters: The OpenCC [4] library was utilized to convert all traditional Chinese characters into their simplified counterparts. This step standardizes the script across the corpus, mitigating variations arising from orthographic differences (e.g., treating ” 國” and ” 国” as equivalent), which is crucial for maintaining consistency in character-level and word-level analyses.

4.Punctuation marks, special symbols, and numerical values unrelated to the content were removed.

5.Character-level analysis: Treating each Chinese character as an independent unit.

Word-level analysis: Using a Chinese word segmentation tool named jieba [1] to tokenize the text into meaningful words.

5.Removal of Stop Words: Frequently occurring stop words (e.g., ” 的,” ” 是,” ” 在”) were removed from the tokenized text.

Word2Vec

Word2Vec is a predictive model introduced by Mikolov et al. [2], based on the idea that words appearing in similar contexts tend to have similar meanings.Word2Vec uses a shallow neural network and optimizes its embeddings via stochastic gradient descent, typically with negative sampling or hierarchical softmax. The model is particularly effective at capturing local syntactic and semantic patterns from sequential co-occurrence information.

GloVe

GloVe (Global Vectors for Word Representation), developed by Pennington et al. [3], is a count-based model that leverages global word co-occurrence statistics. It constructs a word–word co-occurrence matrix from the entire corpus and learns embeddings by factorizing this matrix through a weighted least-squares objective. Unlike Word2Vec, which focuses on predicting context, GloVe explicitly encodes how frequently word pairs appear together in the corpus. This often results in improved global semantic consistency, especially for rare words or long-range dependencies.

Conclusions

A key observation emerged from our analysis of the character ” 令狐冲”, a central figure in The Smiling, Proud Wanderer. When querying for similar words based on cosine similarity, the Word2Vec model successfully retrieved names of other prominent characters such as “周伯通” and ” 张无忌”, reflecting a strong capacity for modeling narrative and character co-occurrence relationships. In contrast, the GloVe model tended to return other things (e.g., ” 寒气” ,” 姑姑”,” 时刻”), suggesting that it captured broader stylistic or descriptive associations rather than fine-grained character-level semantics.

This result highlights the advantage of Word2Vec’s local context modeling in capturing character-centric semantic relationships in narrative texts, particularly where proximity and co-occurrence within dialogue and action sequences are significant. While GloVe remains useful for identifying thematic or stylistic elements, Word2Vec may be better suited for downstream tasks involving entity relations, character networks, or narrative structure in literary corpora.

References

References

[1] fxsjy. Jieba chinese text segmentation. https://github.com/fxsjy/jieba. Accessed: 2025-03-10.

[2] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[3] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[4] Xue Yang, Xiaowei Zhao, Gwan Tjio, et al. Opencc–an open benchmark data set for corpus callosum segmentation and evaluation. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 3020–3024. IEEE, 2020.