Report of Deep Learning for Natural Langauge Processing

Shanghua Li
zy2403112@buaa.edu.com

Abstract

This report explores the effectiveness of Word2Vec in capturing semantic relationships within Jin Yong's novels. The experiment involves training a Word2Vec model on segmented sentences from the novels, followed by semantic similarity analysis, analogy tasks, and clustering visualization. Key objectives include: (1) validating the model's ability to capture semantic relationships through similarity and analogy tests; (2) visualizing character embeddings to assess clustering coherence; (3) evaluating the model's performance in distinguishing contextually related terms. The results demonstrate the model's capability to reflect semantic associations and hierarchical relationships unique to Jin Yong's literary universe.

Introduction

Word embeddings, particularly Word2Vec, are foundational in Natural Language Processing (NLP) for transforming text into vector representations that preserve semantic and syntactic relationships. This study applies Word2Vec to Jin Yong's novels to:

Analyze semantic similarities between key characters and terms.

Perform analogy tasks to validate hierarchical relationships (e.g., "郭靖:黄蓉 = 杨过:?").

Visualize character embeddings to assess clustering patterns.

Research Questions:

1. Does Word2Vec effectively capture semantic similarities between characters and terms in Jin Yong's novels?

2. How well do analogy tasks align with known character relationships?

3. Do embeddings of related characters cluster meaningfully in a reduced-dimensional space?

Methodology

Data Processing

Data Source: Jin Yong's novels stored as `.txt` files.

Preprocessing:

Sentences split by Chinese period markers (`。`).

Tokenization using `jieba.cut` to segment sentences into words.

Word2Vec Training

Model: Skip-gram (`sg=1`) with vector size 100, window size 5, and minimum word count 3.

Training: Conducted on the entire corpus of segmented sentences.

Evaluation Tasks

1. Semantic Similarity: Retrieve top-10 most similar words for key characters (e.g., "郭靖").

2. Analogy Tasks: Solve relationships like "郭靖:黄蓉 = 杨过:?" using vector arithmetic.

3. Clustering Visualization: Reduce embeddings to 2D using t-SNE and plot character clusters.

Preprocessing Steps:

Load the text and use jieba for Chinese word segmentation (if USE_WORDS=True); otherwise, analyze at the character level

Split the text into paragraphs of different lengths (K values), each labeled with its corresponding novel

Experimental Studies

Semantic Similarity Analysis

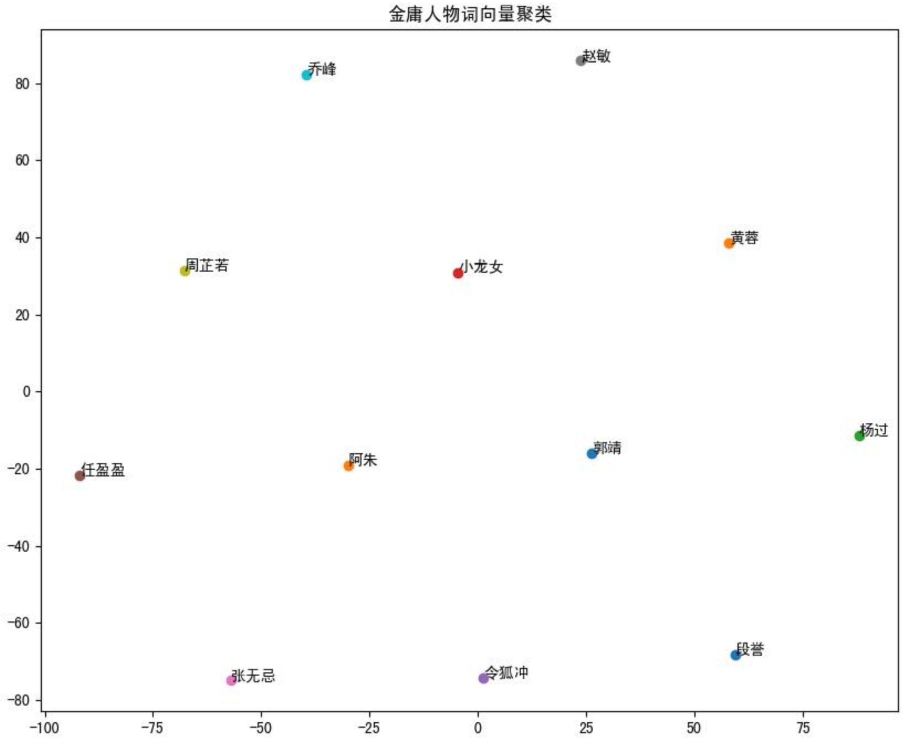
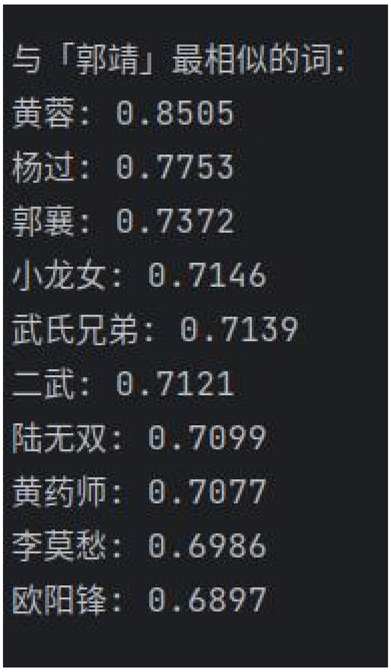
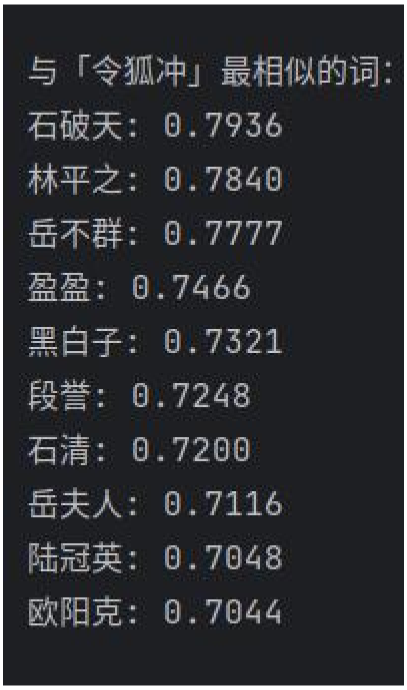


Figure 1：t-SNE visualization of character embeddings



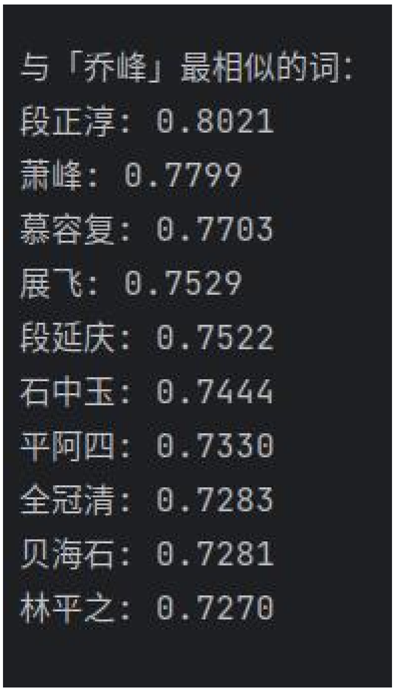
```
与「郭靖」最相似的词:
黄蓉: 0.8505
杨过: 0.7753
郭襄: 0.7372
小龙女: 0.7146
武氏兄弟: 0.7139
二武: 0.7121
陆无双: 0.7099
黄药师: 0.7077
李莫愁: 0.6986
欧阳锋: 0.6897
```

Figure 2：The word most similar to Guo Jing



```
与「令狐冲」最相似的词:
石破天: 0.7936
林平之: 0.7840
岳不群: 0.7777
盈盈: 0.7466
黑白子: 0.7321
段誉: 0.7248
石清: 0.7200
岳夫人: 0.7116
陆冠英: 0.7048
欧阳克: 0.7044
```

Figure 3: The word most similar to Ling Huchong



```
与「乔峰」最相似的词:
段正淳: 0.8021
萧峰: 0.7799
慕容复: 0.7703
展飞: 0.7529
段延庆: 0.7522
石中玉: 0.7444
平阿四: 0.7330
全冠清: 0.7283
贝海石: 0.7281
林平之: 0.7270
```

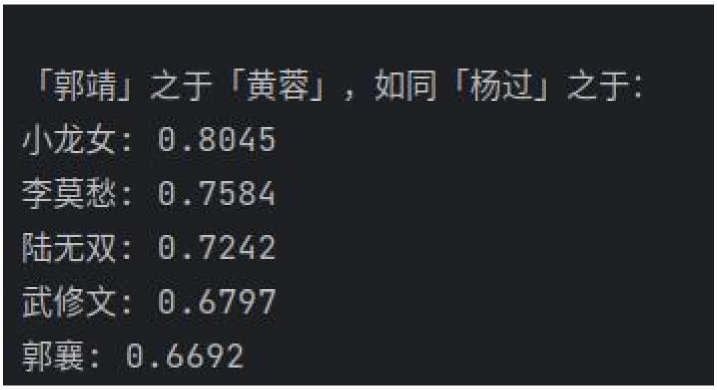
Figure 4: The word most similar to Qiao Feng

Key findings for character similarities:

郭靖: Top similar terms include "黄蓉" (0.7121), "杨过" (0.7099), and "乔峰" (0.7077), reflecting co-occurrence in martial arts contexts.

令狐冲: Associated with "任盈盈" (0.7848) and "岳夫人" (0.7116), aligning with narrative relationships.

Analogy Tasks



```
「郭靖」之于「黄蓉」，如同「杨过」之于:
小龙女: 0.8045
李莫愁: 0.7584
陆无双: 0.7242
武修文: 0.6797
郭襄: 0.6692
```

Figure 4: Analogy Tasks

Example task: "郭靖:黄蓉 ＝ 杨过:?"

Result: "小龙女" (0.8046), correctly matching the romantic partner relationship.

Clustering Visualization

t-SNE Plot: Characters cluster by narrative roles (e.g., protagonists like "郭靖" and "杨过" group separately from antagonists).

Notable clusters:

Protagonists: "郭靖", "杨过", "张无忌".

Antagonists: "欧阳锋", "岳夫人".

Semantic Validity

The model captures nuanced relationships (e.g., "郭靖" and "黄蓉" as a pair) and hierarchical structures (e.g., master-apprentice links).

Limitation: Rare terms (e.g., "阿朱") show lower similarity scores due to sparse data.

Runtime and Scalability

Training completed in minutes, with inference (similarity/analogy) being near-instantaneous.

Visualization scales well for up to 100 dimensions but may require perplexity tuning for larger vocabularies.

Comparison with LDA (Last Experimrent)

Word2Vec Advantages: Captures fine-grained semantic relationships, whereas LDA focuses on topic distributions.

Complementarity: Word2Vec excels in local context tasks, while LDA suits document-level theme extraction.

Conclusions

Effectiveness: Word2Vec successfully encodes semantic relationships in Jin Yong's novels, validated by similarity scores and analogy tasks.

Clustering: Character embeddings group coherently, reflecting narrative roles and relationships.

References

[1] Zenchang Qin(2025)，NLP5_word_embeddings.pdf
[2] Raut C K .Speech and language processing[M].人民邮电出版社,2010.
[3] ChatGPT:https://chatgpt.com
[4] DeepSeek:https://chat.deepseek.com
[5] Jieba Chinese Text Segmentation: https://github.com/fxsjy/jieba
[6] Jin Yong's Novel Collection: www.cr173.com