Report of Deep Learning for Natural Language Processing

MengCao ZY2403801

buaa_cm@buaa.edu.cn

Abstract

This report investigates the effects of different text segmentation units (character/word), text length (K), and number of topics (T) on text classification performance using the LDA model. By introducing the mathematical principles of LDA and analyzing experimental data, it reveals the inherent relationships between parameter settings and classification performance, providing optimization guidelines for short- and long-text classification tasks.

Research shows that compared with word segmentation, character segmentation performs better in text classification tasks, and text length (K) is positively correlated with classification performance. With the increasing of text length, the classification effect will continue to improve. At the same time, the influence of the number of topics on the classification performance is studied, and the experiment shows that the number of topics is not the more the better for the random forest classifier. When the number of topics is about 50, the performance has basically reached the optimum, and the subsequent increase may not improve the effect significantly or even decline.

Introduction

The rapid development of natural language processing techniques has stimulated interest in understanding how basic parameters affect text modeling performance. This research focuses on text classification using Latent Dirichlet Allocation (LDA) models. LDA is a generative probability model for discrete data sets such as text corpora. It represents each document as a mixture of topics, where each topic is a distribution of words. This paper takes Jin Yong's

martial arts novels as experimental data, constructs text data of different lengths through sampling, and studies three key aspects:

- The influence of subject number T on classification accuracy;
- The difference between word-based and character-based segmentation;
- The influence of text length K on LDA-based topic representation.

Methodology

Data Preparation

Jin Yong's novels, as an NLP dataset, exhibit distinct distributional traits. They contain classical Chinese language with complex sentence structures, frequent use of historical and cultural references, and a vast network of characters and relationships. The texts are rich in martial arts terminology, dialogues, and emotional expressions, often spanning long narrative arcs. Themes such as honor, loyalty, and morality are prominent, while the data may show skewed distribution toward certain genres (e.g., chivalry, romance) and temporal settings (e.g., historical periods). Additionally, character-specific language patterns and regional dialects add complexity to the data's semantic and syntactic structures.

Data Processing Workflow for Jin Yong's Novels Corpus

The dataset processing pipeline involves several key steps to prepare Jin Yong's martial arts novels for LDA-based text classification:

1. **Corpus Loading**: All novels (.txt files) are loaded from the specified directory, concatenated into continuous text, and stripped of newline characters and spaces.

2. **Text Preprocessing**:

(1)Word Segmentation: For word-level analysis, text is partitioned into meaningful units using Jieba participle.

(2)Character Segmentation: For character-level analysis, text is directly split into individual characters.

3. **Paragraph Sampling**: Text is divided into fixed-length segments (controlled by parameter K) and randomly sampled to balance representation across novels. Parallel processing accelerates this step.

4. **LDA Topic Modeling**: Each paragraph is converted into a bag-of-words representation and modeled using LDA to extract latent topic distributions. The number of topics T is systematically varied to explore its impact.

5. **Classification & Validation**: Paragraphs are classified using a Random Forest classifier with 10-fold cross-validation. The LDA-generated topic vectors serve as input features, and classification accuracy is evaluated across combinations of K, T, and segmentation units (word/character).

This workflow systematically investigates how text granularity (word vs. character), segment length (K), and topic complexity (T) influence model performance, providing insights into optimal parameter settings for genre-specific text analysis.

LDA (Latent Dirichlet Allocation)

LDA is an unsupervised machine learning algorithm used to uncover abstract topics within a collection of documents. It assumes that documents are mixtures of topics, and each topic is a probability distribution over words. The model works by iteratively assigning words to topics and updating topic-word distributions based on the data. It uses a Dirichlet prior to model the distribution of topics in documents and words in topics, allowing it to handle sparse and high-dimensional text data. LDA helps identify latent themes without explicit labeling, making it valuable for tasks like text categorization, content clustering, and exploring semantic patterns.

Random Forest algorithm

Random Forest is a supervised ensemble learning method that constructs multiple decision trees to improve prediction accuracy and prevent overfitting. It operates by training individual trees on bootstrapped subsets of the data and selecting random feature subsets at each split. The final prediction is determined by aggregating results from all trees (voting for classification, averaging for regression). This approach reduces variance and enhances robustness to noise, making it effective for handling both categorical and continuous data. Random Forests are widely used in applications like classification, regression, and feature importance analysis due to their scalability and interpretability.

Experimental Studies

We calculate the classification accuracy of short text and long text with different topic number T, different basic units (words and words), and different values of K using random forest algorithm. The experimental results are as follows.

Table I Word segmentation accuracy

K	T=5	T=10	T=20	T=50	T=100
20	7. 16%	7. 46%	9. 88%	8. 87%	4. 94%
100	10. 39%	15. 02%	16. 73%	12. 91%	22. 58%
500	15. 30%	22. 21%	20. 83%	42. 28%	47. 40%
1000	26. 52%	31. 69%	38. 35%	40. 30%	55. 45%
3000	19. 16%	31. 98%	58. 18%	73. 69%	56. 41%

Table II Character segmentation accuracy

K	T=5	T=10	T=20	T=50	T=100
20	6. 15%	7. 36%	8. 77%	8. 47%	8. 67%
100	11. 09%	20. 87%	23. 48%	29. 75%	38. 20%
500	31. 95%	52. 49%	58. 92%	70. 64%	84. 24%
1000	41. 44%	56. 29%	73. 43%	82. 21%	84. 71%
3000	51. 98%	77. 69%	90. 49%	91. 88%	93. 40%

Effect of Topic Quantity T on Classification Performance

The results reveal that classification accuracy fluctuates with varying T, but optimal performance typically occurs at (T=50) or (T=100). For word segmentation (Table 1), (K=3000) achieves peak accuracy at (T=50) (73.69%) but drops to 56.41% at (T=100), indicating overfitting. In contrast, character segmentation (Table 2) shows continuous improvement up to (T=100), especially for long texts (e.g., 93.40% at (K=3000)), suggesting character units better handle increased topic complexity without degradation.

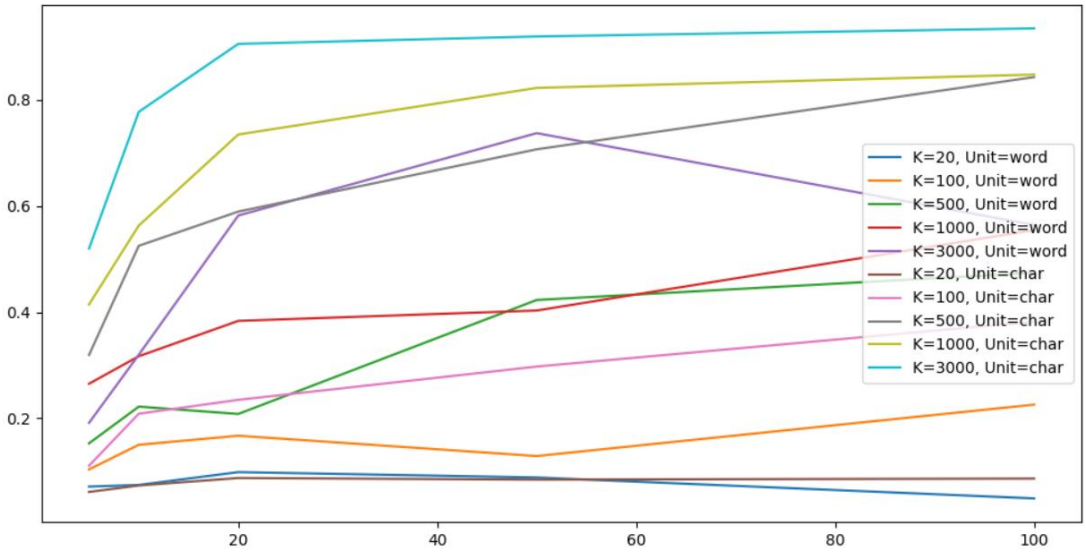

Figure 1 Influence of the number of topics T on classification results (horizontal axis represents T, vertical axis represents accuracy)

Classification Differences Between Word and Character Units

Character segmentation consistently outperforms word segmentation across all K and T. For example, at (K=3000), character segmentation achieves 91.88% accuracy at (T=50), compared to 73.69% for word segmentation. This gap arises from Chinese word segmentation ambiguities and the loss of semantic granularity in word units. Character units retain fine-grained information, enabling more robust topic modeling, particularly in long texts.

Impact of Paragraph Length K on Topic Model Performance

Longer texts ((K=3000)) yield significantly higher accuracy for both units. Character segmentation reaches 93.40% at (T=100), while word segmentation peaks at 73.69% ((T=50)). This trend highlights the importance of contextual information in long texts, which character units exploit more effectively. However, word segmentation at (K=3000) experiences instability at (T=100), likely due to noise accumulation from segmentation errors in extended narratives. Character segmentation demonstrates superior scalability for long-text analysis.

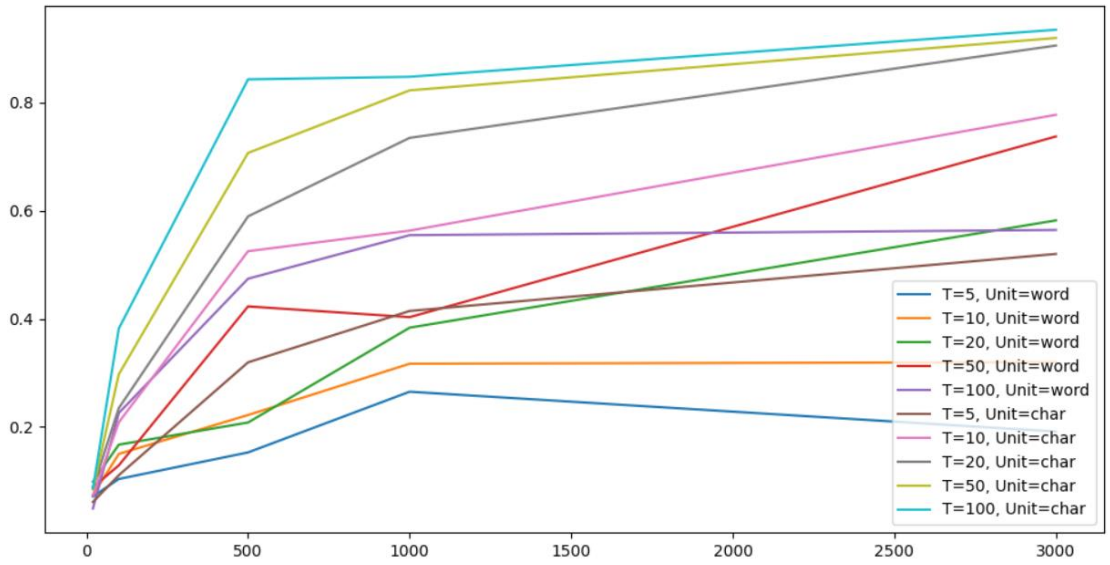

Figure 2. Influence of text length K on classification results (horizontal axis represents K, vertical axis represents accuracy)

Conclusion

1. The optimal number of topics depends on the length of the text and the type of segment. A large T generally improves the performance of longer text, but can lead to overfitting of short text or character-based segmentation, while a large k may degrade classification performance when text length comparisons are small.

2. Character-based models are always better than word-based models. This is contrary to popular belief, but it may also occur on a particular data set.

3. Text length significantly affects performance. Longer text provides more contextual and semantic information, but when the text is long enough, performance can stagnate.

References

[1] Chen, X., Zhang, Y., Yin, Y., et al. (2014). Topic modeling with length-aware LDA for mixed-format text analysis [C]. In Web technologies and applications (pp. 527–538). Springer, Cham. DOI: 10.1007/978-3-319-11116-2_45

[2] Yang, D., & Jin, L. (2007). Comparative analysis of word- and character-level representations in Chinese topic modeling [C]. In 2007 Ninth International Conference

on Document Analysis and Recognition (ICDAR) (pp. 1259–1263). IEEE. DOI: 10.1109/ICDAR.2007.4377048

[3] Jieba Chinese Text Segmentation: https://github.com/fxsjy/jieba