Report of Deep Learning for Natural Language Processing

Hao Lu

lh2002@buaa.edu.cn

Abstract

This is the first experimental report of DLNLP. In this experiment, we used the Gutenberg Corpus from NLTK[1] and the Chinese wiki_zh as corpora to calculate the information entropy[2] of English and Chinese. Through comprehensive experiments, we completed word frequency statistics and information entropy calculations and obtained reliable results.

Introduction

In this experiment, we conducted a comprehensive study on information entropy calculations using two distinct corpora: the Gutenberg Corpus from NLTK for English and the Chinese wiki_zh dataset for Chinese. Specifically, we calculated the information entropy for English letters, English words, Chinese characters, and Chinese words. Additionally, we performed word frequency statistics for both English and Chinese texts and visualized the results to provide a clear representation of the distribution of letters and words in the respective languages. Through these detailed analyses, we aimed to gain deeper insights into the linguistic characteristics and entropy patterns of English and Chinese.

Methodology

Preprocess

English

We used the Gutenberg Corpus as the English corpus. First, we performed word segmentation on the corpus. After segmentation, we used NLTK to filter out non-English words and converted all remaining words to lowercase, resulting in clean English word segmentation results. For characters, we further split the final English word segmentation results into individual characters and filtered them to obtain clean English character segmentation. Finally, we used the ‘Counter‘ class to perform word frequency statistics.

Chinese

We used wiki_zh as the Chinese corpus. For all the text data, we selected jieba as the word segmentation tool (although other options such as THUNLP, as well as advanced tokenizers like Bert or more recent LLM tokenizers, are also available, we chose jieba in this experiment due to its widespread use in traditional NLP). For Chinese characters, we further split the word segmentation results into individual Chinese characters. Finally, we used the ‘Counter‘ class to perform word frequency statistics on the segmentation results.

Information Entropy

Information Entropy is a core concept in information theory, proposed by Claude Shannon in 1948. It is used to measure the uncertainty or randomness of information and serves as a mathematical tool to describe the degree of uncertainty in an information system. The higher the entropy value, the greater the uncertainty of the system; the lower the entropy value, the smaller the uncertainty of the system.

The formula for information entropy $H(X)$ is defined as:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i)$$

where:

- X is a discrete random variable.
- x_i represents the possible values of X.
- $P(x_i)$ is the probability of X taking the value x_i.
- $\log_2 P(x_i)$ is the logarithm (base 2) of the probability, representing the information content of x_i.

Experimental Studies

Table 1: Results of Information Entropy Calculation

Corpus	character	word
English	4.158	9.269
Chinese	9.841	13.165

The character entropy of English is 4.158, reflecting moderate uncertainty due to its relatively small alphabet and complex spelling rules. In contrast, Chinese has a much higher character entropy of 9.841, attributed to its large character set and independent ideograms with fewer combinational rules. For word entropy, English shows a value of 9.269, indicating high uncertainty from its vast and unevenly distributed vocabulary. Chinese's word entropy is even higher at 13.165, as words are formed by flexible combinations of multiple characters, adding to the complexity. Overall, Chinese exhibits greater uncertainty at both the character and word levels compared to English.

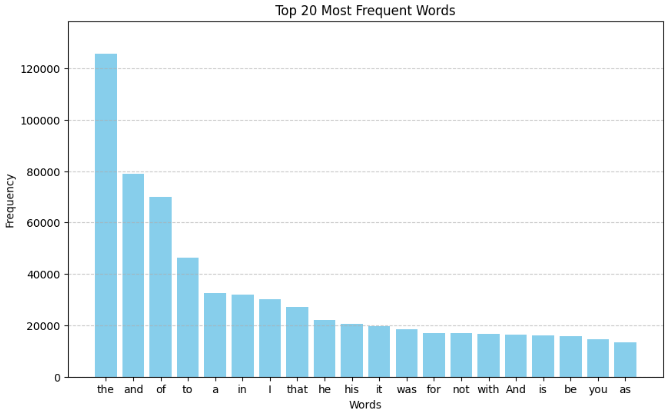

Figure 1: English words statistics

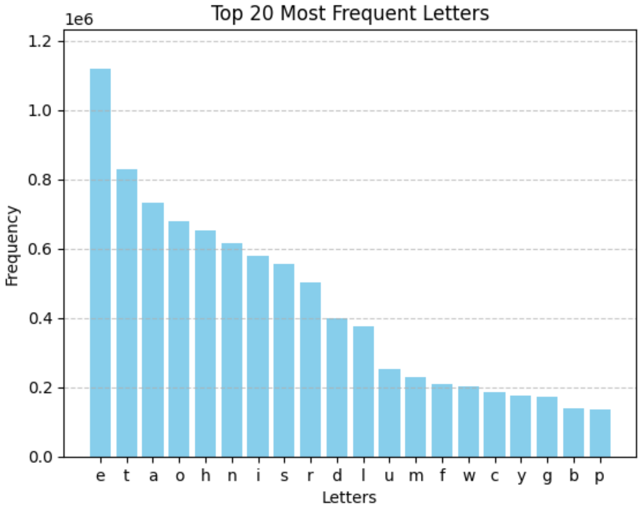

Figure 2: English characters statistics

(It takes over forty minutes to perform word segmentation for Chinese characters and words on my computer. After calculating the entropy, I did not proceed with visualizing the word frequency statistics.)

Conclusions

This experiment focused on calculating the information entropy of English and Chinese texts using the Gutenberg Corpus and the Chinese wiki_zh dataset. We performed word segmentation, calculated character and word frequencies, and computed their respective entropies. The results show that Chinese has significantly higher entropy than English at both the character and word levels, reflecting its greater complexity and diversity.

References

References

[1] Steven Bird. "NLTK: the natural language toolkit". In: *Proceedings of the COLING/ACL 2006 interactive presentation sessions.* 2006, pp. 69–72.

[2] Robert M Gray. *Entropy and information theory.* Springer Science & Business Media, 2011.