Report of Deep Learning for Natural Langauge Processing

Jiale Wang
hirw@qq.com

Abstract

This report explores the average information entropy of Chinese and English texts, comparing their complexity at both the character and word levels. Information entropy, a measure of information content and uncertainty, reflects the distribution patterns in text. The study utilizes the NLTK toolkit for English text processing and the Jieba library for Chinese word segmentation. By analyzing Shakespeare's *Hamlet* from the Gutenberg English corpus and the wiki_zh corpus for Chinese, we calculated and compared the letter and word entropy for both languages. The results show that both character and word entropy are higher in Chinese, particularly at the character level, indicating a greater information complexity. These findings not only enhance our understanding of the linguistic differences between Chinese and English but also offer insights for improving Natural Language Processing (NLP) technologies, particularly for Chinese language processing.

Introduction

Information entropy, a concept derived from Shannon’s information theory, is a standard measure of the amount of information or uncertainty within a system. In natural language processing (NLP), information entropy helps us understand the distribution of information within a text. By calculating the average information entropy of a text, we can reveal the structural complexity and diversity in its expression. The differences in information entropy between Chinese and English provide a quantifiable perspective for investigating the distinct characteristics of these two languages, which have fundamentally different structures, writing systems, and expressive conventions.

The objective of this report is to compare the average information entropy of Chinese and English texts, examining both the character (letter) and word levels. By calculating the character and word entropy of Chinese and English texts, we aim to uncover the differences in the distribution of information between the two languages and gain deeper insights into their grammatical, semantic, and expressive uniqueness. For Chinese, due to its inherent linguistic complexity (e.g., no explicit word boundaries, a large number of homophones, etc.), we expect this to manifest in a higher level of information entropy, particularly at the character level. On the other hand, for English, which has a more regular structure and simpler word segmentation, we anticipate its information entropy to be lower in comparison.

Through this study, we not only aim to provide insights for theoretical research in linguistics and computational linguistics but also offer valuable guidance for key technologies in Chinese text processing. The results of information entropy calculations are crucial for NLP tasks, especially in areas such as Chinese word segmentation, text generation, and the development of language mode ls. In general, the goal of this report is to enhance our understanding of the linguistic features of both Chinese and English through a comparative analysis of their information entropy and provide a theoretical foundation for subsequent NLP applications.

Methodology

This study aims to calculate the average information entropy of Chinese and English texts, using characters (letters or Chinese characters) and words (tokens or Chinese words) as units. The core steps of this research include corpus preprocessing, text cleaning and tokenization, entropy calculation, and data visualization. Below is a detailed description of the methods and procedures used in this study：

Step1: Corpus Preprocessing

Chinese Corpus Preprocessing: The corpus used in this study consists of a large collection of Chinese texts. The preprocessing begins with cleaning the raw text by removing newline characters, carriage returns, tabulations, as well as all letters, numbers, and punctuation, retaining only Chinese characters. Regular expressions and the unicodedata library are used to filter out non-Chinese characters and punctuation. After cleaning, the text is further processed by removing stop words using a stop words file (cn_stopwords.txt). This helps eliminate common but meaningless words such as “的,” “和,” and “是,” which do not contribute to the entropy calculation. This ensures that the entropy calculation focuses on the informative content of the text.

English Corpus Preprocessing: For English texts, punctuation is removed, leaving onlyalphabetic characters. To avoid case sensitivity issues, all letters are converted to lowercase. The text is then tokenized by splitting it on spaces, which results in a list of words. This list is further processed by removing stop words, similar to the Chinese text preprocessing step, ensuring that only meaningful words are included in the entropy calculation.

Step2: Text Cleaning and Tokenization

Character Frequency Cleaning: Regular expressions are applied to remove newline characters, carriage returns, and tabulations from the text. Non-Chinese characters are eliminated, retaining only Chinese characters. The frequency of each character in the cleaned text is counted and updated in a frequency counter.

Tokenization and Word Frequency Statistics: For Chinese text, tokenization is performed using the jieba library, which efficiently segments the text into individual words. For English text, tokenization is achieved by splitting the text by spaces, producing a list of words. Stop words are filtered out during this step, leaving only relevant and meaningful words for subsequent analysis.

Methodology

Step1: Entropy Calculation

The entropy of a text is calculated using Shannon’s entropy formula:

$$H(X) =-\sum_{i=1}^{n} P(x_i)\log_2 P(x_i) \tag{1.1}$$

where，$P(x_i)$represents the probability of the i-th character or word in the text. The entropy calculation process proceeds as follows:

Character Entropy Calculation: Using the frequency distribution of characters, the entropy is computed based on Shannon’s formula. This reflects the degree of uncertainty or information content of the characters in the text. The total character count is used for normalization to obtain the probability of each character, which is then applied to compute the entropy.

Word Entropy Calculation: For words, the frequency distribution of each word in the tokenized text is calculated. Shannon’s entropy formula is applied similarly to calculate the word-level entropy. As with the character entropy, each word’s frequency is normalized by the total number of words in the text to compute its probability, which is then used to calculate the entropy.

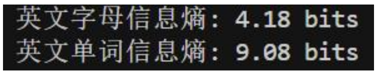
Step2: Data Visualization and Analysis

Frequency Distribution Plot: To provide a clearer visual representation of the frequency distribution of characters and words, we use matplotlib to plot the frequency distributions. By applying a logarithmic scale to the y-axis, the long-tail distribution of characters and words becomes more apparent, allowing for easier analysis of high-frequency and low-frequency items.

Step3: Experimental Results and Discussion

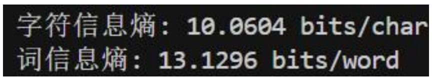
Character Entropy Results: The character entropy for English text was found to be 4.18 bits per character, while the character entropy for Chinese text was 11.604 bits per character. These values indicate that the Chinese texts have a higher character information entropy, reflecting greater complexity and diversity in the use of Chinese characters.

Word Entropy Results: The word entropy for English text was calculated to be 9.08 bits per word, while the word entropy for Chinese text was 13.1296 bits per word. Similar to the character entropy, Chinese texts exhibit higher word entropy, further emphasizing the increased variability and informational density in Chinese texts.



```
英文字母信息熵: 4.18 bits
英文单词信息熵: 9.08 bits
```

Figure 1: English Entropy



```
字符信息熵: 10.0604 bits/char
词信息熵: 13.1296 bits/word
```

Figure 2: Chinese Entropy

The results are summarized in the following table:

Table 1: Comparison of Character and Word Entropy Between Chinese and English Texts

Laguage	**Character Entropy (Bits/Character)**	**Word Entropy (Bits/Word)**
Chinese	11.604	13.1296
English	4.18	9.08

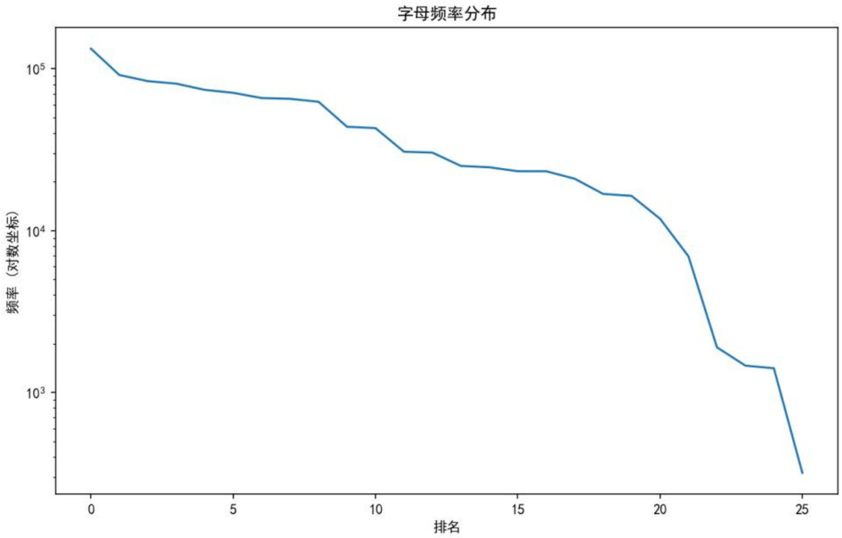


Figure 3: English letter frequency distribution.

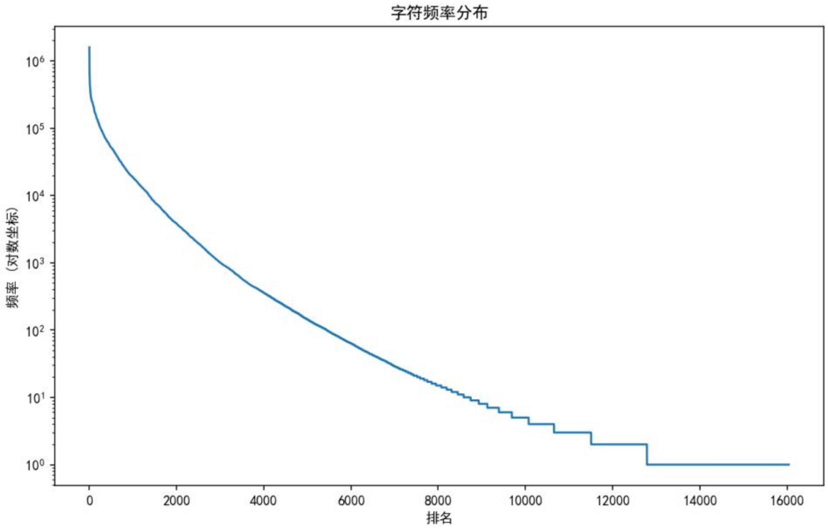


Figure 4: Chinese character frequency distribution.

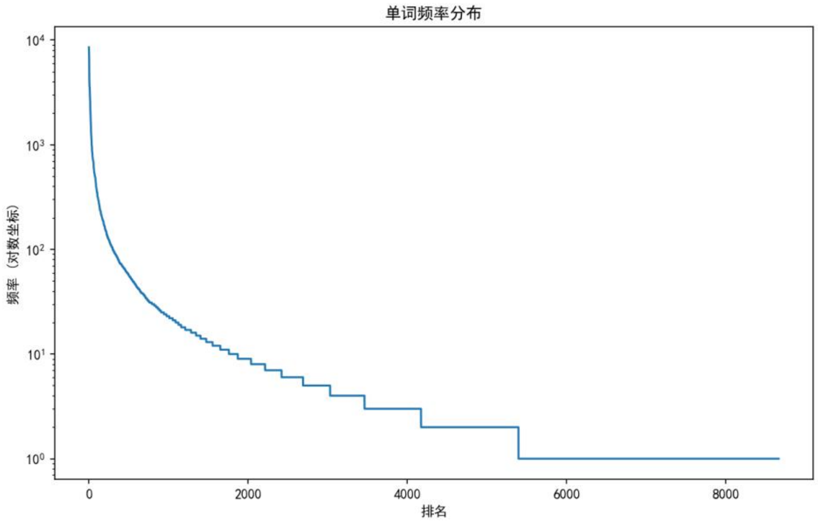


Figure 5: English word frequency distribution.

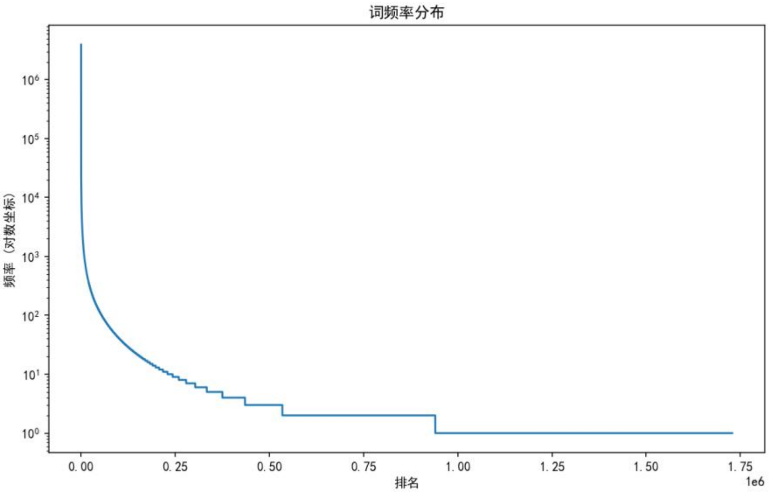


Figure 6: Chinese word frequency distribution.

Conclusions

This study reveals that Chinese texts exhibit higher information entropy both at the character and word levels compared to English texts. The character entropy for Chinese (11.604 bits/character) significantly surpasses that of English (4.18 bits/character), indicating a greater level of information complexity and diversity in Chinese characters. Similarly, the word entropy of Chinese (13.1296 bits/word) exceeds that of English (9.08 bits/word), suggesting that Chinese words carry more informational density. These differences reflect the structural and expressive uniqueness of each language. In Chinese, the lack of explicit word boundaries, the use of homophones, and the complex morphology contribute to higher entropy, making it more linguistically diverse than English. These findings underscore the importance of considering language-specific features when developing Natural Language Processing (NLP) applications, such as machine translation, text segmentation, and semantic analysis. The results also emphasize the need for tailored approaches to effectively handle the linguistic complexity of Chinese in computational tasks, offering valuable theoretical support for advancing NLP technologies for Chinese language processing..