Report of Deep Learning for Natural Language Processing

Chenggang Xie SY2406119

xiechenggang@buaa.edu.cn

Abstract

Information entropy, as a key concept, is of great significance for a deep understanding of the information structure and characteristics in text data. This research report focuses on the scientific method of calculating information entropy from the detailed perspectives of words and letters in the rich and diverse Chinese and English corpora. In the process of research, not only the data processing method is elaborated in detail, but also the differences in the information entropy of Chinese and English corpora are compared and analyzed in an all-round way.

Introduction

The important concept of information entropy[1] was first proposed by in an in-depth academic study. In essence, it is a subtle concept used to accurately measure the uncertainty or unpredictability of information in a data set. In the field of natural language processing (NLP), the concept of entropy has been given a more specific and practical connotation. It mainly achieves high-precision quantification of the amount of information contained in the text by conducting a comprehensive and detailed analysis of the frequency of various language units such as characters and words. The analysis of text data entropy can reflect the complexity of the data to a certain extent.

Methodology

In this study, we used a bilingual corpus for comparative analysis: the Chinese corpus selected the 2019 Chinese Wikipedia dataset (wiki_zh_2019), and the English corpus selected the Gutenberg corpus. For the processing of the Chinese corpus, we designed a rigorous processing flow: first, the system traverses all document files and uses regular expression technology to accurately extract Chinese characters. To ensure the accuracy of the information entropy statistics of Chinese words, all non-Chinese characters are strictly removed; secondly, the Jieba word segmentation tool is used for Chinese word segmentation. In order to deal with the memory shortage problem that may occur in large-scale data processing, we designed a block processing mechanism to optimize memory usage efficiency through batch processing.

Compared with the Chinese corpus, the processing flow of the English corpus is simpler: first, the text is standardized using the text processing method built into the nltk toolkit, including removing punctuation marks, line breaks and other non-essential characters; second, through the normalization processing steps, redundant spaces are deleted and all characters are uniformly converted to lowercase; finally, based on the characteristics of English text, the space segmentation method is directly used for word segmentation to ensure processing efficiency and accuracy. This differentiated processing strategy fully considers the characteristics of Chinese and English languages and the scale characteristics of the corpus, laying a reliable data foundation for subsequent comparative studies.

We use the following formula to calculate entropy. The main calculation method is to use the statistically obtained frequency to calculate the sum in order, and the entropy results by character or phrase can be obtained.

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

Experimental Studies

We calculated the information entropy of the Chinese Wikipedia dataset in units of characters and words, and the specific results are shown in Table 1. Through this analysis, we can deeply understand the information distribution characteristics of Chinese text at the character and vocabulary levels. Table 1 shows the information entropy values at different granularities, which provides an important quantitative basis for subsequent Chinese language model optimization and information processing.

Table 1 Chinese information entropy calculation

Chars	Words
9.85437	13.355

Similarly, we process the English corpus by letters and words to calculate the corresponding average information entropy results as shown in Table 2.

Table *2* English information entropy calculation

Chars	Words
4.158	9.728

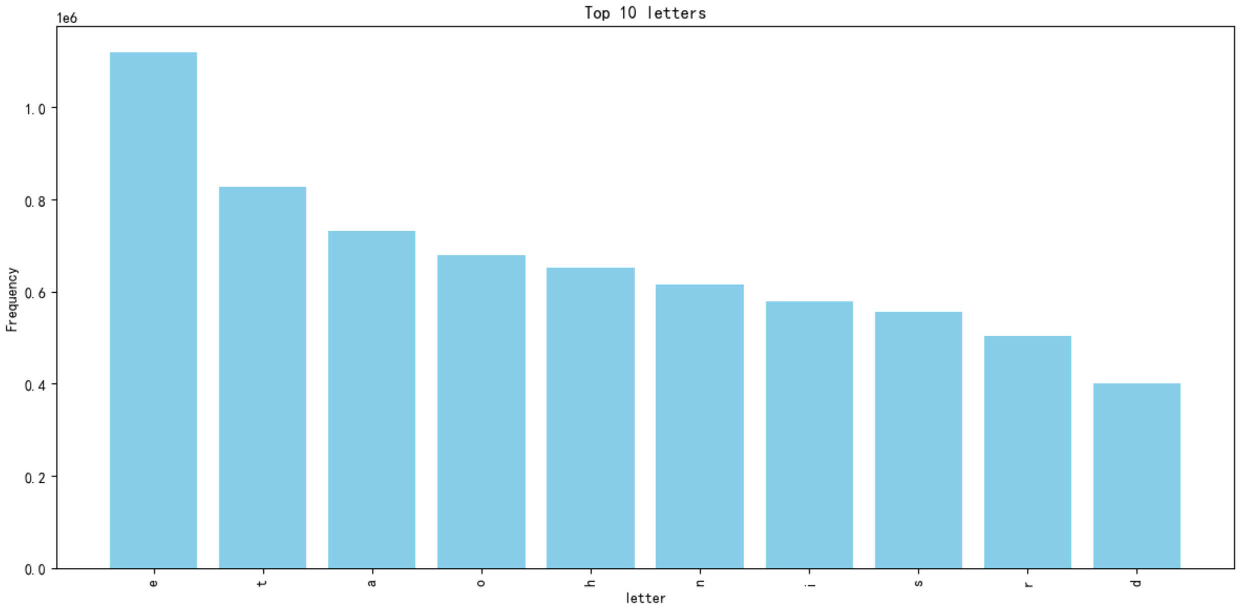

Fig 1 Top ten frequent chars in English

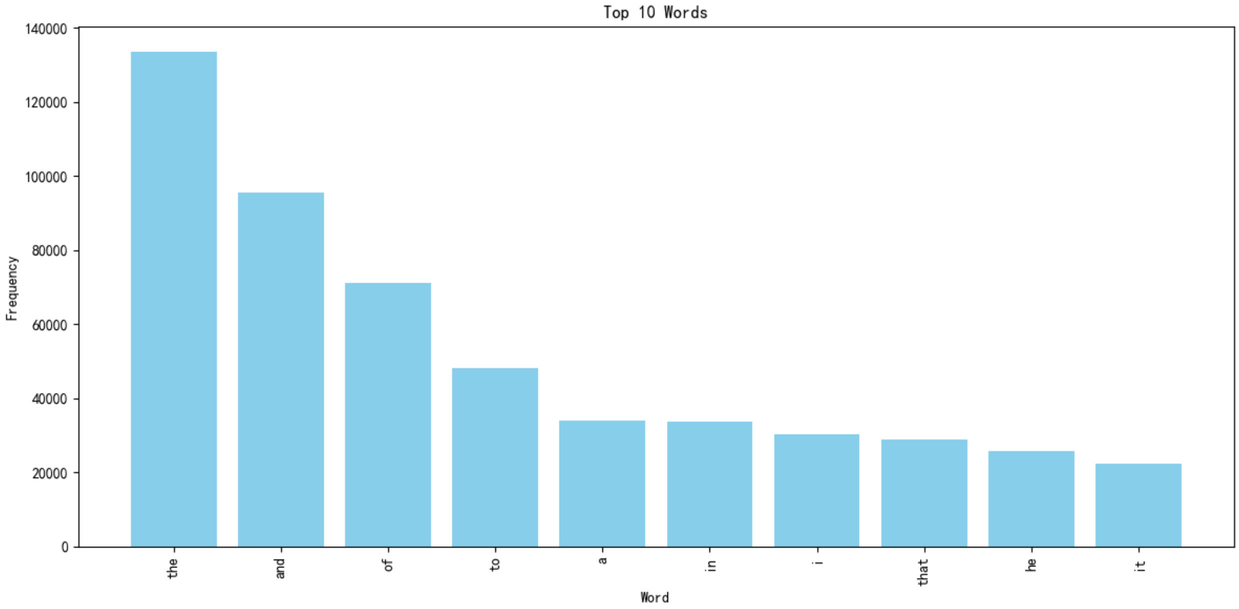

Fig 2 Top ten frequent words in English

The following important observations can be made from the experimental results: First, in Chinese text, the information entropy at the word level is generally higher than the information entropy at the character level, which reflects the richness and diversity of Chinese vocabulary in semantic expression. Second, due to the relatively small English alphabet set (containing only 26 letters), English text shows lower information entropy at the character level; while Chinese shows greater entropy changes at both the character and word levels, which reflects the complexity of Chinese at the level of glyphs and vocabulary. In addition, in English text, the high-frequency appearance of simple words (such as "the") significantly reduces the overall information entropy. The presence of such high-frequency words is an important manifestation of the statistical characteristics of English text.

Conclusion

The report found that the information entropy at the word level in both Chinese and English is higher than that at the character level, reflecting the diversity and complexity of word combinations. Due to its large morpheme character system, Chinese has a higher entropy value at the character level, while English has a more balanced entropy distribution at the word level. These differences reveal the essential characteristics of information distribution in Chinese and English.

References

[1] Volkenstein M V. Entropy and information[M]. Springer Science & Business Media, 2009.