Report of Deep Learning for Natural Language Processing

Shixiang Li BY2203043

lishixiang@buaa.edu.cn

Abstract

This study investigates the information entropy of both Chinese and English corpora at the character and word levels. By calculating the frequency distributions of characters and words, we compute the corresponding entropy values and their averages. Additionally, visual analyses of top 10 frequency distributions are presented. This report explores the differences in information entropy between Chinese and English at different segmentation levels.

Introduction

Information entropy, introduced by Claude Shannon [1], measures the uncertainty or unpredictability of information in a dataset. In natural language processing (NLP), entropy quantifies the amount of information contained in a text by analyzing the frequency of linguistic units such as characters and words [1]. Chinese, a logographic language, presents unique challenges due to its lack of explicit word boundaries and extensive character set. English, on the other hand, is an alphabetic language with clear word boundaries. This report aims to calculate and compare the information entropy at both the character and word levels using a large Chinese corpus (wiki_zh_2019) and an English corpus (Gutenberg Corpus). Additionally, we visualize the frequency distributions and entropy values to better understand their patterns.

Methodology

Data Preprocessing

The corpora used in this study are the wiki_zh_2019 dataset for Chinese and the Gutenberg Corpus for English. The preprocessing steps include:

- Chinese Corpus: Extracting Chinese characters using regular expressions, tokenizing words with Jieba [3], and removing non-Chinese symbols and whitespace.

- English Corpus: Tokenizing words and characters, converting text to lowercase, and removing punctuation and whitespace.

Frequency Calculation

- Each character's frequency is calculated by counting occurrences in both corpora.
- Word Level: Word frequencies are computed using Jieba for Chinese and simple whitespace-based tokenization for English.

Entropy Calculation

The information entropy H is calculated using the formula:

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

where $P(x_i)$ is the probability of character or word x_i. The average entropy is calculated by taking the mean of individual entropies:

$$H_{avg} = \frac{1}{n}\sum_{i=1}^{n} H(x_i)$$

Visualization

We use matplotlib to visualize:

- Character and Word Frequency Distributions: Bar charts showing the most frequent characters and words for both languages.
- Entropy Distributions: Histograms illustrating the distribution of entropy values for characters and words in Chinese and English.

Experimental Studies

Chinese Wikipedia

We calculated the information entropy of Chinese Wikipedia in units of characters and words, and the results are shown in Table I. At the same time, we plotted the top ten characters/words with the highest frequency, as shown in Fig 1.

Table I Wikipedia information entropy calculation

Average information entropy (Ch)	Average information entropy (Word)
9.5881	12.1443

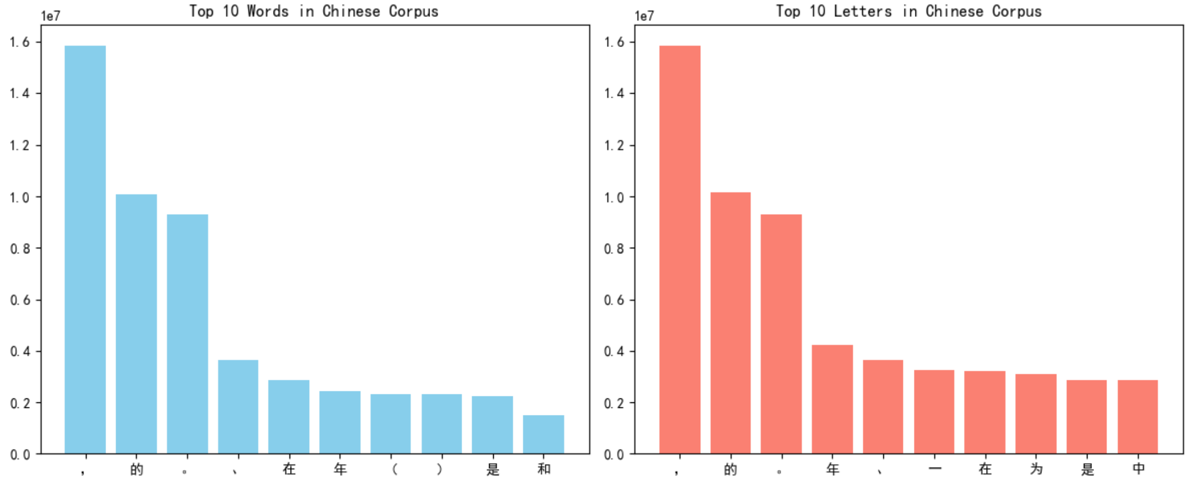

Fig 1 Top ten most frequent words in Wikipedia

Gutenberg Corpus from NLTK

Similarly, we segment the Gutenberg English corpus by letters and words to calculate the corresponding average information entropy. The results are shown in Table II and Fig 2.

Table II Gutenberg information entropy calculation

Average information entropy (Ch)	Average information entropy (Word)
4.3908	7.6777

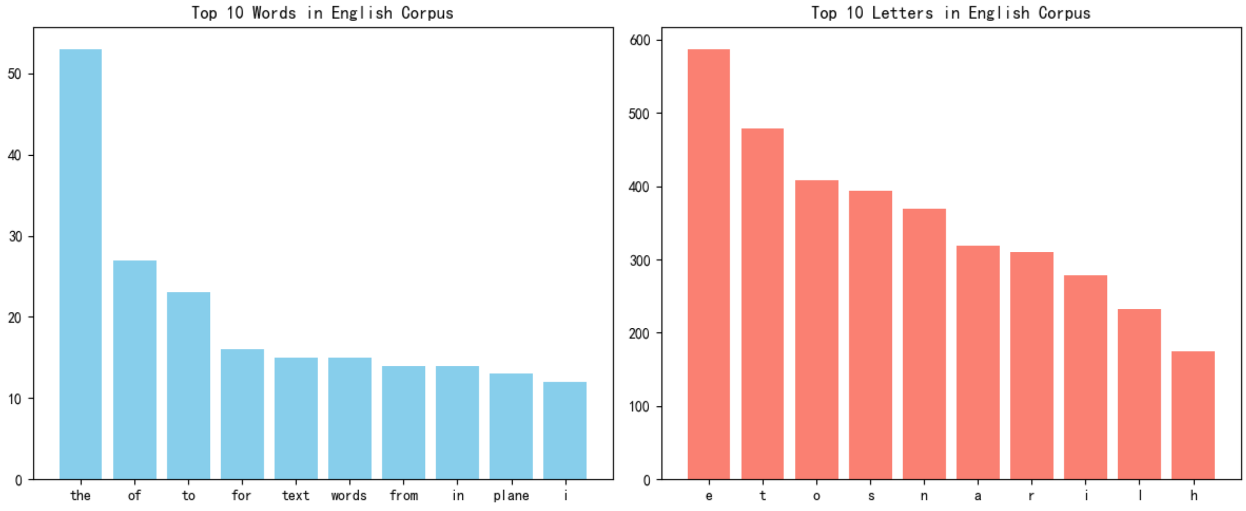

Fig 2 Top ten most frequent words in Gutenberg

We observed that word-level entropy tends to be higher than character-level entropy for both Chinese and English. Additionally, English text exhibited lower character entropy due to its smaller alphabet set, while Chinese showed greater variation in both character and word entropy.

Conclusion

This study quantified and compared information entropy at both the character and word levels in Chinese and English corpora. The results indicate that word-level entropy surpasses character-level entropy in both languages, reflecting the increased complexity of word combinations. Furthermore, the Chinese corpus demonstrated higher character entropy due to its extensive set of logographic characters, while English showed more uniform word-level entropy. These findings offer valuable insights for NLP tasks such as language modeling, text compression, and information retrieval.

References

[1] Shannon C E. A mathematical theory of communication[J]. The Bell system technical journal, 1948, 27(3): 379-423.

[2] Brown P F, Della Pietra S A, Della Pietra V J, et al. An estimate of an upper bound for the entropy of English[J]. Computational Linguistics, 1992, 18(1): 31-40.

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