The First Report of Deep Learning for Natural Language Processing

Yue Zhao
zy2303607@buaa.edu.cn

Abstract

This paper explores the application of Zipf's Law and the concept of information entropy within the context of the Chinese language, employing both word-based and character-based N-Gram models. Zipf's Law, reflecting the inverse relationship between word frequency and rank, is validated through empirical analysis of a Chinese corpus. Additionally, the study delves into calculating the average information entropy for Chinese, illustrating the impact of context length (N) on the diversity of word combinations and the simplicity of text structure.

Introduction

Zipf's Law, proposed by the American scholar G.K. Zipf in the 1940s, reveals an inverse relationship between the frequency of word usage and its rank in natural languages: the most frequently used words appear with high frequency, while the frequency of other words decreases as their rank increases. This law is not only applicable in the field of linguistics but also prevalent in various domains such as urban population and corporate size, reflecting universal principles of information organization and social structure. The study of Zipf's Law offers significant insights into understanding the efficiency of language, distribution of information, and the organization of complex systems.

Entropy, broadly defined, measures the state of certain material systems, indicating the degree to which certain states of material systems may occur. It has also been employed metaphorically in social sciences to describe the extent of certain states in human societies. The concept of entropy was introduced by the German physicist Rudolf Clausius in 1865. Initially, it was used to describe one of the material state parameters related to "energy degradation" and has found extensive application in thermodynamics. However, at that time, entropy was merely a physical quantity that could be determined by changes in heat, and its essence was not well explained. It was not until the development of a series of scientific theories, such as statistical physics and information theory, that the nature of entropy began to be clarified. Namely, the essence of entropy is the "intrinsic level of disorder" within a system. It plays a significant role in fields such as control theory, probability theory, number theory, astrophysics, and life sciences, and has derived more specific definitions in different disciplines.

Methodology

M1: Zipf's Law

Zipf's Law is a law of word frequency distribution. It can be articulated as follows: If one compiles the frequency of each word's occurrence in a lengthy text, arranges them in a descending order with high-frequency words at the beginning and low-frequency words following, and assigns natural numbers as rank indices to these words, such that the word with the highest frequency is assigned rank 1, the next highest frequency rank 2, and so on, until the word with the lowest frequency is assigned rank D. Letting f represent frequency and r represent the rank index, the relationship is as shown in equation (1).

$$f \times r = C \tag{1}$$

Where C is a constant.

M2: Average Information Entropy of Chinese

The concept of information entropy was first introduced by Claude Shannon (1916-2001) in 1948, drawing on the concept of "thermal entropy" from thermodynamics, aimed at representing the uncertainty of information. The higher the entropy value, the greater the degree of uncertainty in the information.

According to [1], for text $X = \{...X_{-2}, X_{-1}, X_0, X_1, X_2, ...\}$, its definition of information entropy is given by Equation (2).

$$H(X) \equiv H(P) \equiv -E_P \log P(X_0 \left| X_{-1}, X_{-2}, ...\right.) \tag{2}$$

According to the Law of Large Numbers, when the sample size is sufficiently large, the probability of occurrence of words, bigrams, and trigrams approximates their frequency of occurrence.

Thus, the information entropy formula for the unigram model is given by Equation (3),

$$H(X) = -\sum_{x \in X} P(x) \log P(x) \tag{3}$$

where $P(x)$ can be approximated by the frequency of each word in the corpus.

The information entropy formula for the bigram model is given by Equation (4),

$$H(X \left| Y \right.) = -\sum_{x \in X, y \in Y} P(x, y) \log P(x \left| y \right.) \tag{4}$$

where the joint probability $P(x, y)$ can be approximated as the frequency of occurrence of each bigram in the corpus, and the conditional probability $P(x \left| y \right.)$ can be approximated as the ratio of the frequency of occurrence of each bigram in the corpus to the frequency of bigrams starting with the first word of the bigram.

The information entropy formula for the trigram model is given by Equation (5),

$$H(X|Y,Z) = -\sum_{x\in X, y\in Y, z\in Z} P(x,y,z)\log P(x|y,z) \tag{5}$$

where the joint probability $P(x,y,z)$ can be approximated as the frequency of occurrence of each trigram in the corpus, and the conditional probability $P(x|y,z)$ can be approximated as the ratio of the frequency of occurrence of each trigram in the corpus to the frequency of trigrams starting with the first two words of the trigram.

Experimental Studies

E1: Verify Zipf's Law

Based on the aforementioned principle and the available data, code is developed to utilize a Chinese corpus for the validation of Zipf's Law, yielding Figures 1 as results.

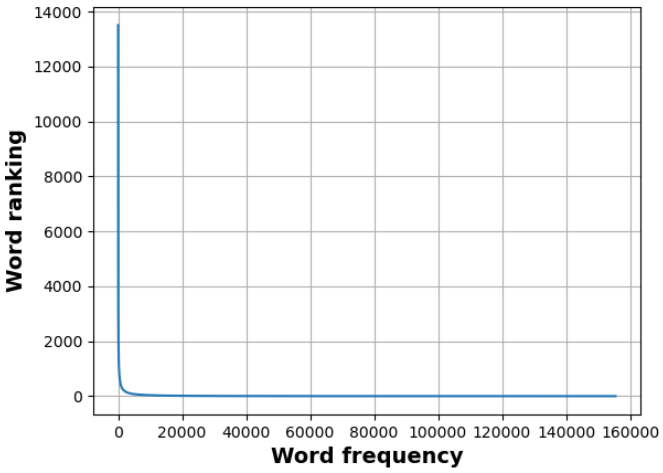

Figure 1: The curve of Word ranking-Word frequency

In Figure 1, we observe that due to the extensive range and dramatic variations in the data, it is challenging to intuitively determine whether the relationship between rank and frequency conforms to Zipf's Law.

To address this issue, we take the logarithm of both sides of Equation (1), resulting in Equation (6),

$$\log f = \log k - \log r \tag{6}$$

where theoretically, the relationship between rank and frequency should present a linear correlation.

Redrawing the graph according to Equation (6) yields Figure 2, where the curve approximates a straight line, indicating a near-linear relationship between rank and frequency, thereby validating Zipf's Law.

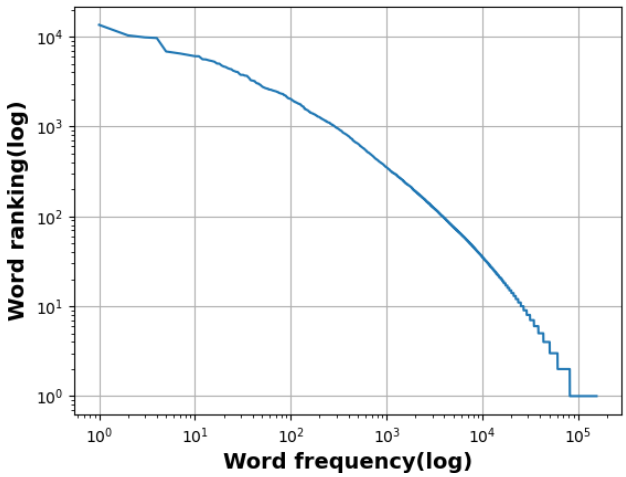

Figure 2: The logarithmic curve of Word ranking-Word frequency

E2: Calculate the Average Information Entropy of Chinese

Preprocessing of the Chinese corpus: The corpus in the database is in txt format, which includes portions of code and irrelevant symbols. Initially, it is necessary to preprocess the data, which involves removing hidden characters (such as newline characters, page breaks, etc.); eliminating irrelevant information and characters obtained through web crawling; and deleting punctuation marks from the text, as punctuation does not contribute meaningfully to the information in a Chinese corpus.

Average Information Entropy of Chinese by word basis: Following the preprocessing of the Chinese corpus, code is written based on the principles introduced in Section M2 to calculate the average information entropy of Chinese under the N-Gram model. The statistical language model employed in this paper is based on words, necessitating the segmentation of Chinese sentences into words. For this purpose, the jieba segmentation system—a Chinese word segmentation system for Python—is utilized to tokenize the sentences, with the results summarized in Table 1.

Table 1: Average Information Entropy of Chinese by word basis

Rank	Combined-gram, Frequency		
	1-Gram	2-Gram	3-Gram
1	'的', 115604	'道你', 5738	'只听得', 1613
2	'了', 104527	'叫道', 5009	'忽听得', 1140
3	'他', 64712	'道我', 4953	'站起身来', 731
4	'是', 64457	'笑道', 4271	'哼了一声', 580
5	'道', 58623	'听得', 4202	'笑道你', 572
6	'我', 57483	'都是', 3905	'吃了一惊', 539
7	'你', 56679	'了他', 3638	'啊的一声', 523
8	'在', 43691	'他的', 3497	'点了点头', 505
9	'也', 32606	'也是', 3201	'说到这里', 476
10	'这', 32199	'的一声', 3102	'了他 s 的', 459
Total number of combined-grams	4267805	4208528	4149591
The number of different combined-grams	172209	1940075	3448842
Entropy (bits per word)	12.180902382994336	6.932954107965453	2.2941584704204274

From Table 1, comparing the results obtained from the 1-Gram, 2-Gram, and 3-Gram language models, it is evident that the larger the value of N, meaning the greater the length of context considered, the more numerous the distinct words that appear (corresponding to the “The number of different combined-grams” in the table). This is attributed to the fact that as the length increases, so does the number of combinations of characters forming words, resulting in a greater variety of distinct words.

Furthermore, a comparison among the three models reveals that as the value of N increases, the information entropy of the text decreases. This is reasoned to be due to the fact that, with larger values of N, the distribution of word groups in the text obtained after segmentation becomes simpler. Upon analysis, it is found that with larger N, the number of possible fixed words that can be formed is fewer. Fixed words reduce the chances of characters or short words disrupting the coherence of the article, thereby making the article more ordered. The uncertainty involved in forming words from characters and sentences from words decreases, consequently lowering the text's information entropy.

Average Information Entropy of Chinese by character basis: Furthermore, by segmenting Chinese sentences into characters as units, the average information entropy of Chinese under the N-Gram model is recalculated, with the results summarized in Table 2.

Table 2: Average Information Entropy of Chinese by character basis

Rank	Combined-gram, Frequency		
	1-Gram	2-Gram	3-Gram
1	'一', 139396	'说道', 13525	'韦小宝', 9803
2	'不', 134149	'了一', 12174	'令狐冲', 5889
3	'的', 121668	'一个', 10571	'张无忌', 4645
4	'是', 112707	'自己', 10319	'的一声', 3478
5	'了', 111916	'道你', 10262	'袁承志', 3037
6	'道', 111055	'小宝', 9942	'小宝道', 2417
7	'人', 84302	'韦小', 9856	'陈家洛', 2115
8	'他', 73573	' 也不', 9304	'小龙女', 2081
9	'这', 68993	'道我', 8472	'石破天', 1818
10	'我', 67000	'笑道', 8140	'不由得', 1803
Total number of combined-grams	7299807	7240491	7181175
The number of different combined-grams	5782	731519	3173250
Entropy (bits per character)	9.53836558846301	6.715805200021205	3.9380548669875464

Compared to Table 1, under the same underlying principles, Table 2 exhibits a greater number of word combinations for each N-Gram model. This increase is attributable to the fact that, in this instance, each word in the combinations consists of only a single character. Consequently, the length of word combinations derived from segmentation by characters does not exceed that of combinations derived from segmentation by words, resulting in a higher overall count. For this reason, segmenting text into combinations of characters is less simple and efficient than using combinations of words. This disadvantage becomes more pronounced with the increase of N. Hence, the difference in information entropy for the 3-Gram model between Tables 2 and 1 is significant.

Conclusions

The investigation confirms Zipf's Law's applicability to the Chinese corpus, with graphical representations substantiating the expected linear relationship between word rank and frequency on a logarithmic scale. The study further reveals that as the length of context considered increases (higher N values in N-Gram models), not only does the variety of distinct word combinations augment, but also the information entropy of the text decreases, indicating a transition towards a more ordered and predictable structure. This reduction in entropy is more pronounced when analyzing the text on a character basis, due to the increased combination possibilities offered by individual characters. These findings underscore the significant impact of context length on the distribution and predictability of language, providing valuable insights into the efficiency and organization of linguistic information, as well as offering implications for the fields of computational linguistics, information theory, and language modeling.

References

[1] Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., Lai, J. C., & Mercer, R. L. (1992). An estimate of an upper bound for the entropy of English. *Computational Linguistics*, *18*(1), 31-40.