Report of Deep Learning for Natural Langauge Processing

张永易
2394822700@qq.com

Abstract

本实验旨在使用实际的中文金庸小说语料库数据，验证著名的 Zipf's Law，并计算以词和字为基本单位的中文文本的平均信息熵。通过这些计算，旨在揭示语言的内在统计规律以及语言的信息复杂度。

Introduction

语言作为一种复杂而有序的社会现象，包含着丰富的统计规律和信息结构。对于语言学研究、信息论应用以及自然语言处理技术的发展，理解和量化这些规律至关重要。本实验的重点是探索两个关键的语言统计特性：Zipf's Law（齐夫定律）和信息熵。我们将通过使用实际的中文语料库进行深入研究，以揭示这些特性的内在机制和对语言的重要影响。

Zipf's Law 是由美国语言学家 George Kingsley Zipf 于 1935 年提出的经验定律，用于描述自然语言中词汇分布的非均匀性。该定律表明，在大规模文本中，单词（或在汉字书写体系中的字符）的使用频率与其在频率表中的排名成反比的幂律关系。换句话说，排名第 n 的单词（或汉字）的频率约为排名第 1 的单词（或汉字）频率的 1/n。这一定律揭示了语言使用的经济原则和人类认知的限制，它在不同语言和文本类型中普遍存在，是语言统计学的重要基础。

在语言学中，信息熵用于衡量语言表达的平均信息复杂度，即一个语言系统中词语或字符序列的平均不确定性。信息熵越大，意味着语言系统在表达信息时的平均选择自由度越高，语言的复杂度和多样性越显著。对于特定语言（如中文），验证这一概念仍然具有学术价值，因为不同的语言系统可能存在特定的词汇使用模式和频率分布特征。对中文语料库进行验证有助于深入理解汉语词汇使用的独特性，并为语言模型构建、信息检索算法优化等应用提供更准确的语言学基础。

分别计算词级和字级的信息熵有助于从不同粒度层面揭示汉语的信息组织特性。词级信息熵反映了词汇组合的复杂程度和词汇表征信息的能力，而字级信息熵则直接反映了单个汉字作为语言基本单元的信息承载能力。通过比较两者，可以了解汉语在词汇层面与字符层面的信息分布差异，这对于理解汉字的表意功能、词法构造以及汉语信息处理算法的设计具有重要的指导意义。综上所述，本实验报告旨在通过严谨的数据分析和理论探讨，揭示 Zipf's Law 在中文语料库中的表现以及汉语在词、字两个层次上的信息熵特征，从而增进对汉语内在规律和信息复杂性的认识，为语言学研究、信息论应用及自然语言处理技术提供理论依据和实践指导。

Methodology

一、Zipf's Law 验证
Zipf's Law 是语言学和信息科学中的一项重要统计规律，它指出在一个自然语言文本中，单词（或汉字）的出现频率与其在频率表中的排名大致呈幂律关系。换句话说，排名第 r 的单词（或汉字）出现的频率约为排名第 1 的单词（或汉字）频率的 1/r 左右。为验证 Zipf's Law 在中文语料中的适用性，首先需要对大型中文语料库进行预处理，包括分词（以词为单位验证时）、去除停用词和标点符号等，以获得纯净的语言单元频数数据。然后，对处理后的数据进行排序，并计算每个位置的单词（或汉字）频率。通过绘制频率与排名的双对数图，可以观察是否呈现出典型的直线趋势，以验证 Zipf's Law 的适用性。同时，可以利用幂律模型进行拟合，以量化中文语料库中词汇分布的幂律特征。二、信息熵计算
二、信息熵计算
信息熵是衡量信息不确定性和复杂度的关键指标，被广泛应用于评估语言的内在结构和多样性。在本实验中，我们分别计算了以词和字为单位的中文文本的信息熵。通过计算词级信息熵，我们可以了解词汇组合的复杂程度和词汇表征信息的能力。而通过计算字级信息熵，我们可以直接了解单个汉字作为语言基本单元的信息承载能力。
为了计算不同 n-gram 下的词和字为单位的中文文本的信息熵，首先需要对文本进行分词（以词为单位）或字符切分（以字为单位）。然后，根据不同的 n 值，将文本划分为相应的 n-gram 序列。
对于词级信息熵的计算，我们统计每个 n-gram 序列的出现频率，并计算其概率分布。然后，使用概率分布计算词级信息熵，可以使用以下公式：
词级信息熵 = -∑(P(w) * log2(P(w)))
其中，P(w)表示词的概率，∑表示对所有词进行求和的操作。
对于字级信息熵的计算，我们统计每个 n-gram 序列（以字为单位）的出现频率，并计算其概率分布。然后，使用概率分布计算字级信息熵，可以使用以下公式：
字级信息熵 = -∑(P(c) * log2(P(c)))
其中，P(c)表示字的概率，∑表示对所有字进行求和的操作。
通过分别计算 1-gram、2-gram、3-gram 和 4-gram 下的词和字级信息熵，可以得到不同粒度下的信息熵值。

el.

Experimental Studies

本实验旨在通过实际的中文语料库数据，一方面验证 Zipf's Law 在汉语环境中的适用性，通过绘制频率-排名双对数图和估计 Zipf 指数，直观展现并量化汉语词汇分布的幂律特征；另一方面，计算词级和字级的信息熵，通过利用 1-gram、2-gram、3-gram、4-gram 模型计算中文平均信息熵。对于 x-gram 模型，将语料库中的文本转换为 x-gram 列表，并统计 x-gram 的频率。同时，计算 x-gram 中相同句首词的频率统计。根据频率计算词单位和字单位的平均信息熵。

最后，输出实验结果，包括 1-gram、2-gram、3-gram、4-gram 模型下的词库总词数、不同词的个数、出现频率前 10 的词语，以及词单位和字单位的平均信息熵。以量化汉语文本的平均信息复杂度。实验采用标准的文本预处理步骤，包括分词、停用词过滤等，确保数据的准确性和代表性。所得结果将为深入理解汉语的统计规律和信息结构提供实证支持，同时也为相关领域的研究和应用提供有价值的参考数据、实验报告将总结 Zipf's Law 在中文语料库中的验证结果，包括直观的图形展示和 Zipf 指数估计值，分析其对中文语言特性的解释力。

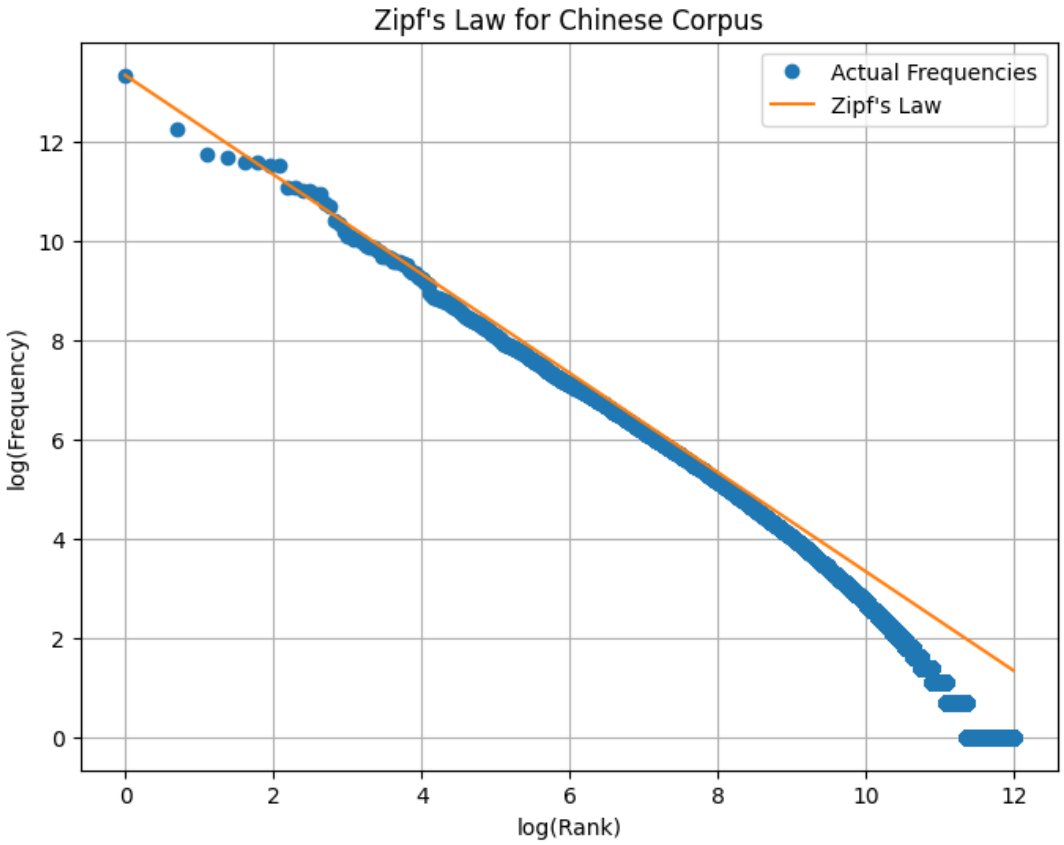

Figure 1：Zip's Law 验证实验结果和理论值

Table 1: n-gram 字、词信息熵

信息熵	1-gram	2-gram	3-gram	4-gram
词单位平均信息熵	12.168	6.944	2.303	2.780
字单位平均信息熵	9.539	8.130	8.605	8.856

Conclusions

通过本次实验，我们期望不仅验证 Zipf's Law 在中文语境下的普适性，还能够通过信息熵的定量计算，深入理解中文语言系统的内在规律与复杂特性

References

[1] Zenchang Qin and Lao Wang (2023)，How to learn deep learning? Journal of Paper Writing, Vol. 3: 23: pp. 1-12.
[2] ….