Report of Deep Learning for Natural Langauge Processing

Yiming Xu
1953999172@qq.com

Abstract

本文对金庸的 16 篇小说进行了语料分析，验证了齐夫定律（Zip's law），计算了中文一元、二元、三元字\词平均信息熵。

Introduction

齐夫定律（Zip's law），由美国语言学家乔治·齐夫在 20 世纪 40 年代提出，是描述语言中词汇分布的一种规律，指出在任何给定语料库中，词的频率与其频率排名成反比。这一发现揭示了语言使用中的一种普遍现象：少数高频词汇在交流中被频繁使用，而大多数词汇使用频率低。齐夫定律的普适性不仅对语言学有重要影响，也在信息理论等多个领域内有着广泛应用。

选用中文语料库验证齐夫定律，旨在探索这一定律在不同语言体系中的适用性。中文的独特性，如书写系统和语法结构，提供了一个独特的视角来观察和理解齐夫定律。此外，这一研究有助于深入分析中文的语言特性如何影响词汇的分布与使用，对提高中文文本处理技术和语言模型的准确性及效率具有实际意义。

此外，引入信息熵概念，量化语言的不确定性，可以进一步理解齐夫定律在中文语料库中的表现。信息熵反映了语言的多样性和复杂性，而齐夫定律所揭示的词频分布特性对语言的信息熵有直接影响。分析中文语料库中的词频分布与信息熵之间的关系，不仅能够深化我们对中文交流效率和表达丰富性平衡的理解，也为比较不同语言的信息处理效率提供了新的视角。

Methodology

Part1：验证齐夫定律

验证齐夫定律需要对文本中的词和词频进行统计。具体方法如下：首先对文本进行预处理，删除所有的隐藏符号、非中文字符和标点符号；接着使用jieba库，对每个txt文件中的文本进行分词；然后统计每个词出现的次数并绘图观察。

在 txt 文件中，在正文之中夹杂着大量空格、换行符、标点符号等，如图 1 所示，这些会影响分词的结果。因此再进行分词之前，首先对文本进行预处理，只保留文本。

Part2：计算信息熵

对文本信息而言，若信息有n种单元$x1,x2,\cdots,xn$，对应发生的概率为$P(xi)$。各种单元彼此独立，所以文本信息的平均不确定性应当为单个单元不确定性的统计平均值，称为信息熵。离散随机变量X的熵值H定义如下：

$$H(X)=\sum_i P(x_i)I(x_i)=-\sum_i P(x_i)\log_b^{P(x_i)}$$

其中，P为X的概率质量函数，E为期望函数，$I(X)$为X的信息量。 其中，b 取 2 的时候，熵的单位是 bit；b 取自然常数 e 时，熵的单位是 nat；b 取 10 时，熵的单位是 Hart。本文中 b 取 2。

M1：一元语言模型

一元语言模型中，每个字\词出现的概率与其它字\词无关。此时，字\词信息熵的计算公式为：

$$H(X)=-\sum_{x\in X} P(x)\log_2^{P(x)}$$

其中 $P(x)$ 可近似等于每个字或词在语料库中出现的频率。

M2：二元语言模型

假设语料中的字\词出现具有马尔科夫性，其出现概率只与前一个字\词有关，此时为二元语言模型。当前字\词x与前字\词y组成了二元组(x,y)，其信息熵计算公式为：

$$H(X|Y) = -\sum_{x\in X, y\in Y} P(x,y)\log_2^{P(x|y)}$$

其中联合概率$P(x,y)$可近似等于每个二元词组在语料库中出现的频率，条件概率 $P(x|y)$ 可近似等于每个二元组在语料库中出现的频数与以该二元组的第一个词为词首的二元组出现的频数的比值。

M3：三元语言模型

在二元语言模型基础上，假设字\词出现的概率与前两个字\词有关，此时为三元语言模型，可以理解为单词z正好出现在二元组(x,y)之后，组成了三元组(x,y,z)，其信息熵计算公式为：

$$H(X|Y,Z) = -\sum_{x\in X, y\in Y, z\in Z} P(x,y,z)\log_2^{P(x|y,z)}$$

其中联合概率 $P(x,y,z)$ 可近似等于每个三元词组在语料库中出现的频率，条件概率 $P(x|y,z)$ 可近似等于每个三元词组在语料库中出现的频数与以该三元词组的前两个词为词首的三元词组的频数的比值。

Experimental Studies

Part1：验证齐夫定律

对 16 部作品分别进行分词统计之后，得到结果在图 1 中。从可视化结果来看，我们的发现完全符合齐夫定律的预期。

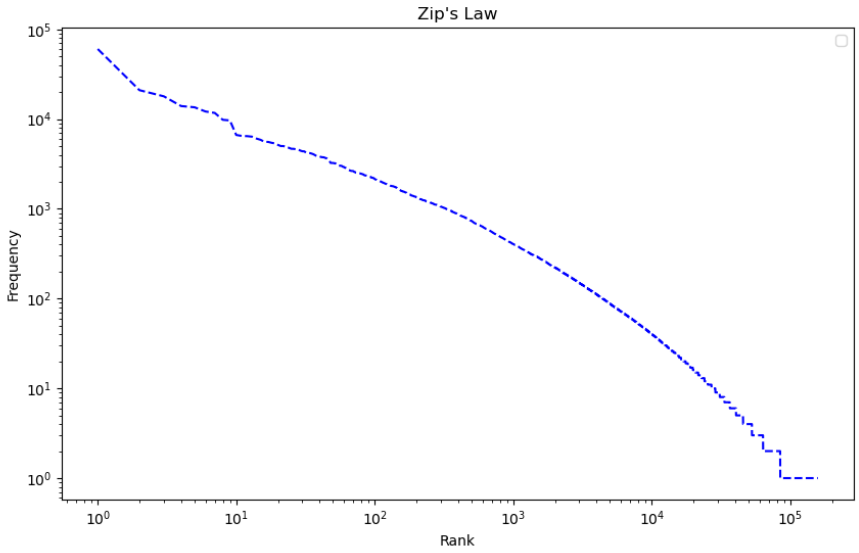

图 1：金庸作品词频统计

Part2：计算信息熵

对 16 个 txt 文件计算一元、二元、三元字和词的信息熵，结果在表 1 中。

表 2 金庸作品字\词信息熵计算结果

分词模型	字信息熵	词信息熵
M1	9.950816	13.587993
M2	7.030511	6.528536
M3	3.497624	1.177056

Conclusions

本研究通过计算中文文本的一元、二元、和三元信息熵，深入探讨了中文语言的信息分布特性及其独特性。结果表明，随着模型复杂度的增加，中文信息熵显著下降，从一元模型的 13.587993 比特/词到三元模型的 1.177056 比特/词，这一变化反映了中文在词与词组合时的高度规律性和上下文约束性。与英文相比，中文展现出更高的单字信息量和更快的信息熵下降速度，凸显了其作为一种高度结构化语言的特点。

这一发现对于理解中文的语言结构、优化中文文本处理技术具有重要意义。它不仅验证了中文信息处理的独特效率，也为未来基于中文的自然语言处理技术的发展提供了重要的理论支持和实证基础。通过本研究，我们得以更清晰地认识到在设计和应用中文处理算法时，需要充分考虑到语言的这些独特性质，以实现更高效和精准的信息理解与传递。

References

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