Report of Deep Learning for Natural Langauge Processing

Zengchang Qin
zengchang.qin@gmail.com

摘要

本文参考[1]中基于频率统计计算香农熵的方法，用于同时分析中文与英文文本的信息特性。针对中文文本，通过预处理、分词、停用词过滤和字符 / 词频统计，计算了中文字符及词的信息熵，并绘制了频率长尾分布图；针对英文文本，则利用NLTK自带的 Gutenberg 语料库提取莎士比亚《哈姆雷特》的文本，通过正则表达式预处理、字母和单词统计，计算并通过图像展示了英文的熵分布。

简介

香农信息熵作为最基本的信息熵概念，可以通过统计文本中各符号出现的概率来计算各符号所含有的信息量，即通过其消除的不确定性来评估各符号的信息含量多少；其计算方法会根据语言种类变化而变化：中文文本没有天然的词边界，需要借助分词工具（如 jieba库）和停止词表（）来提高统计的准确性。

对于英文文本，单词或字母之间天然存在空格分隔，处理相对简单；而中文文本没有天然的词边界，因此需要借助分词工具（如jieba）和停用词表来提高统计的准确性。本报告将中英文文本的预处理与熵计算方法进行整合，通过对两种语言分别进行清洗、统计及可视化，展示文本内部信息分布的长尾效应，同时计算出不同层次（字母 / 字符和单词 / 词）的平均香农熵。报告中所采用的方法是基于频率分布的香农熵公式：

$$H = -\sum_i p_i \log_2 p_i$$

虽然 Brown 等人的论文《An Estimate of an Upper Bound for the Entropy of English》讨论了英文熵的上界估计，并采用了更复杂的N_gram模型及预测方法，本报告中的方法属于基础的统计熵估计，在其思想基础上利用频率统计对信息量进行量化。

理论方法

M1: Brown Model

本文的数学基础如下：对于一个平稳随机过程 $X = \{..., X_{-2}, X_{-1}, X_0, X_1, X_2, ...\}$ 上的熵易得为：

$$H(X) = H(P) = -E_P[\log P(X_0 \mid X_{-1}, X_{-2}, ...)]$$

即对所有可能的过去的和将来的符号取期望值，得到平均每个符号的信息量，反映了在知道上下文信息后当前符号的不确定性。利用链式法则可得到联合概率：

$$P(X_1,X_2,...,X_n)=\prod_{i=1}^{n}P(X_i\mid X_1,...,X_{i-1}).$$

取对数并除以n得到：

$$H(P)=\lim_{n\to\infty}-\frac{1}{n}E_p[\log P(X_1,X_2,...,X_n)]$$

可见当样本足够长时，每个符号的平均编码长度趋于稳定，得到熵的实际数值。当真实分布P不可知时，构造模型M近似P，定义两者交叉熵为：

$$H(P,M)=-E_p[\log M(X_0\mid X_{-1},X_{-2},...)]$$

极限形式为：

$$H(P,M)=\lim_{n\to\infty}-\frac{1}{n}E_p[\log M(X_0\mid X_{-1},X_{-2},...)]$$

根据信息论原理，对于任意唯一译码的编码方案，平均编码长度满足

$$E_p[I(X_1,X_2,...,X_n)]\approx -E_p[\log P(X_1,X_2,...,X_n)]$$

利用模型M编码时可得：

$$I_M(X_1,X_2,...,X_n)\approx -\log M(X_1,X_2,...,X_n)$$

所以得到模型M的交叉熵为：

$$H(P,M)=\lim_{n\to\infty}\frac{1}{n}I_M(X_1,X_2,...,X_n)$$

实验过程

代码的整体方法可分为以下几步：

1. 文本预处理

中文部分代码从本地语料库（目录 wiki_zh）中递归读取所有文件内容。利用正则表达式清洗文本，移除换行符、空格、斜杠、引号、英文字母、标点、数字及等号，以便进行字符统计。对于分词，采用轻度清洗（保留部分标点信息以辅助分词后使用 jieba 进行中文分词，并利用停用词列表过滤无效词项。

英文部分使用 NLTK 的 Gutenberg 语料库提取《Hamlet》的全文。通过正则表达式仅保留字母和空格，去除标点符号和其他非字母字符，以确保字母和单词的统计准确性。

2. 频率统计

利用 Python 的 Counter 对文本中的字符 / 字母和词 / 单词进行统计。分别统计中文字符、中文词、英文字母以及英文单词的出现频率。

3. 香农熵计算

根据频率统计结果，使用香农熵计算出中文字符熵、中文词熵、英文字母熵和英文单词熵。

4. 数据可视化

分别绘制中文和英文部分的长尾分布图，采用对数刻度展示各元素的频次衰减特性。绘制出现频率前 50 项的直方图，以直观展示高频元素分布情况。

4. 实验结果

表 1: 中英文字词平均信息熵对比

	Bits/Letter	Bits/Word
中文	10.0692	14.5775
英文	4.1426	9.3621

以上为中英文自此平均信息熵量对比，以下分别为

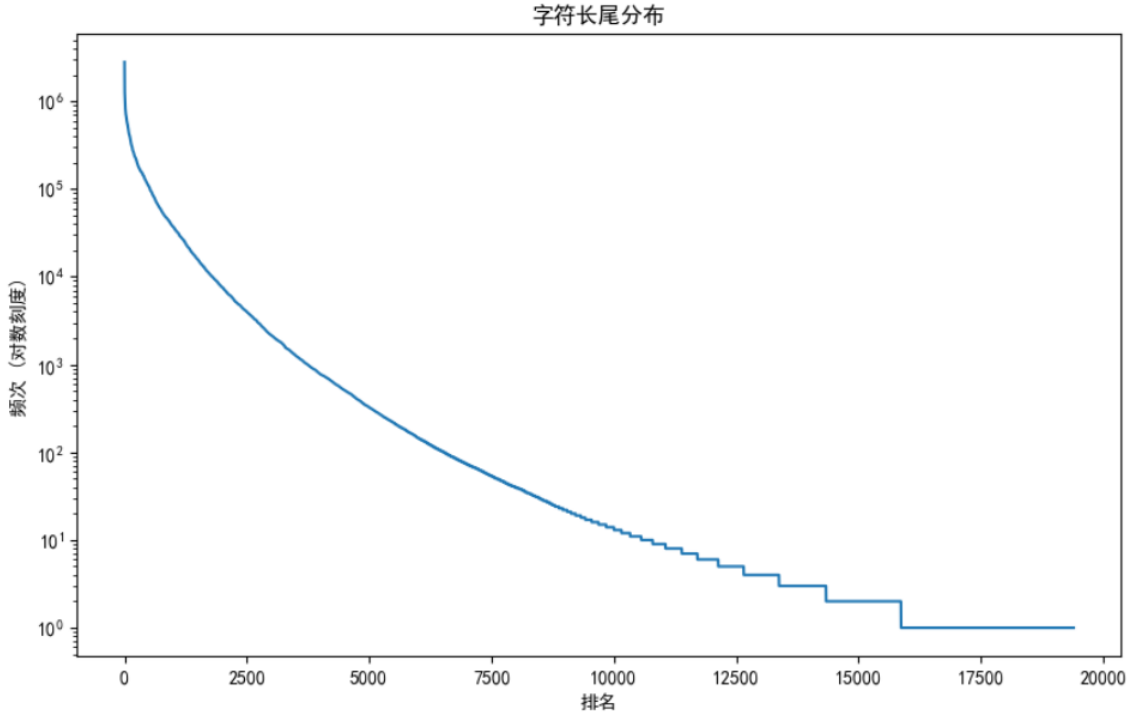

图 1：中文字频长尾图

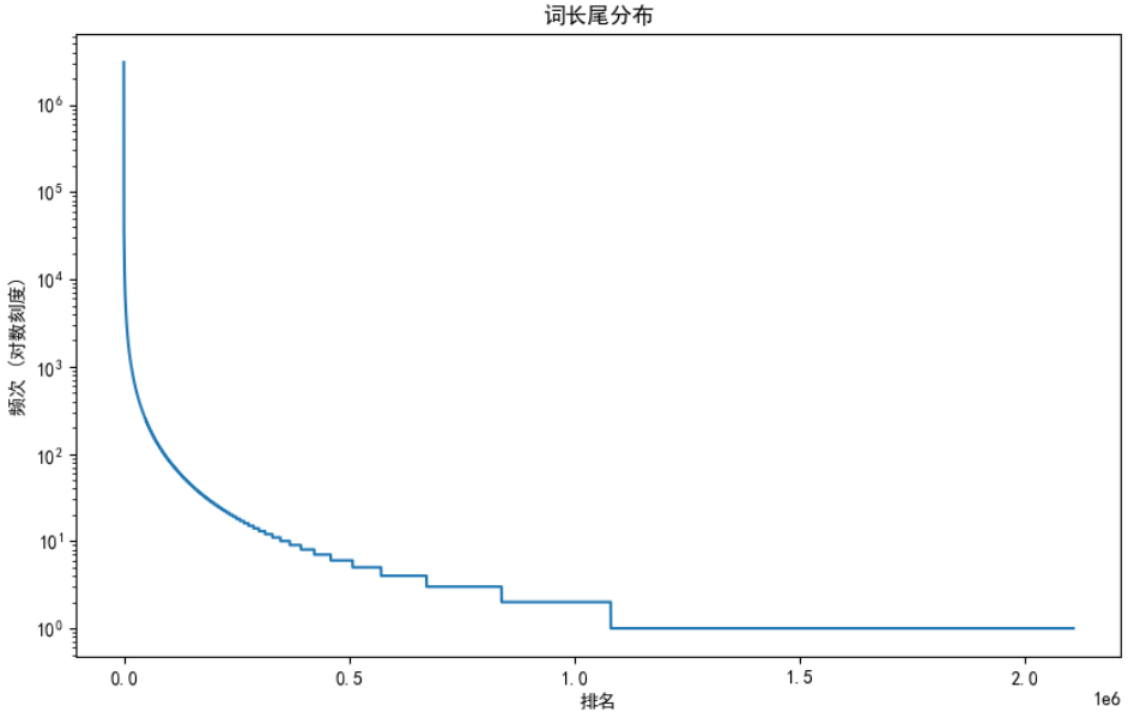

图 1：中文词频长尾图

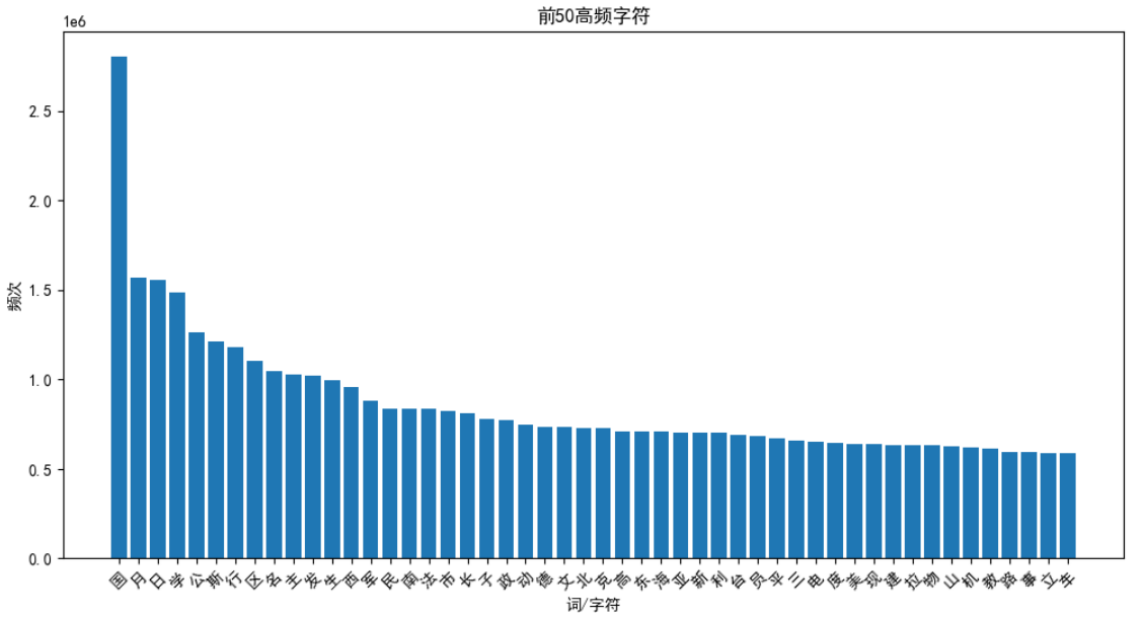

图 3：中文字频前50统计直方图

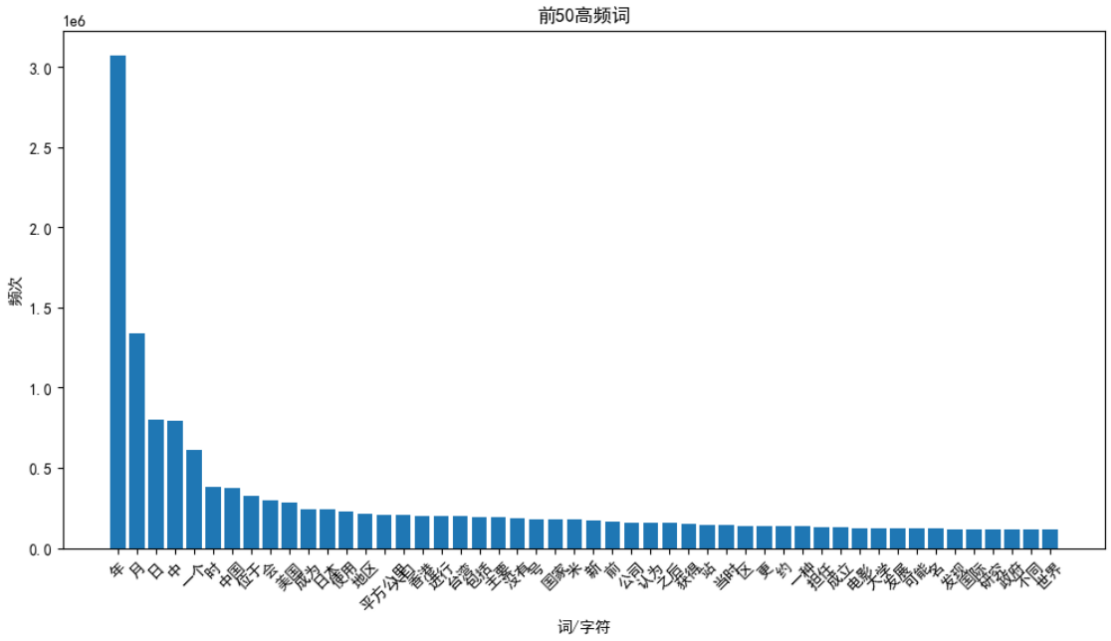

图 4：中文字词频前50统计直方图

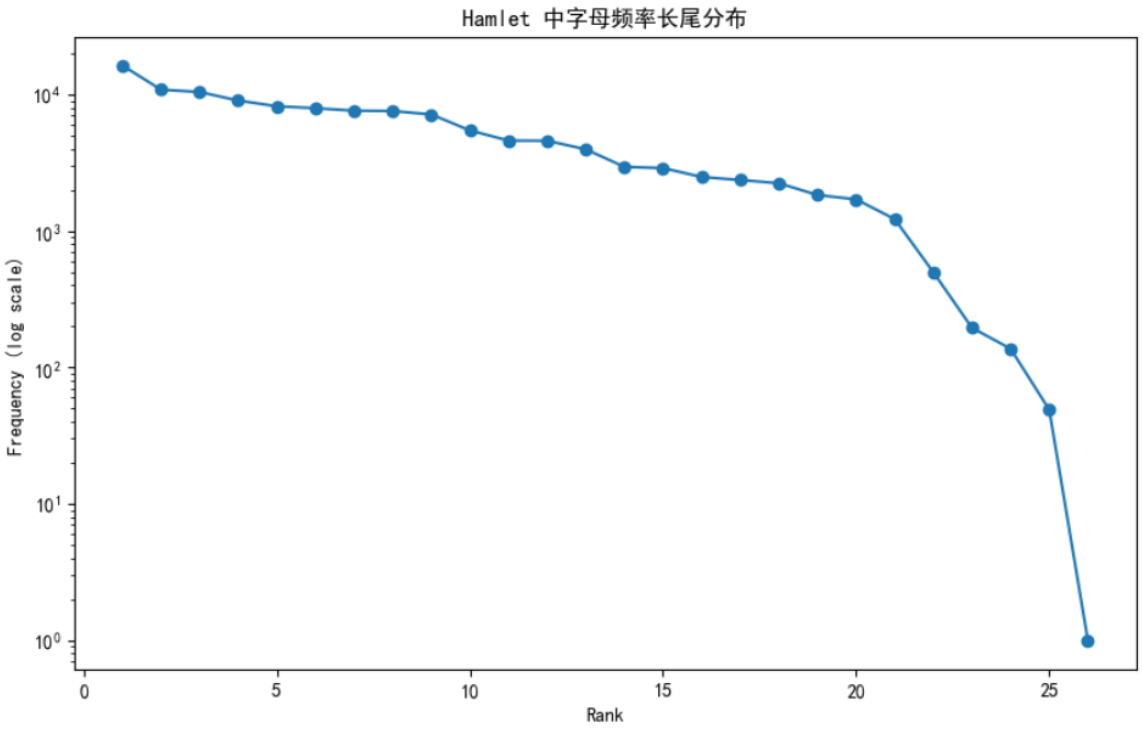

图 5：英文字母统计频率长尾图

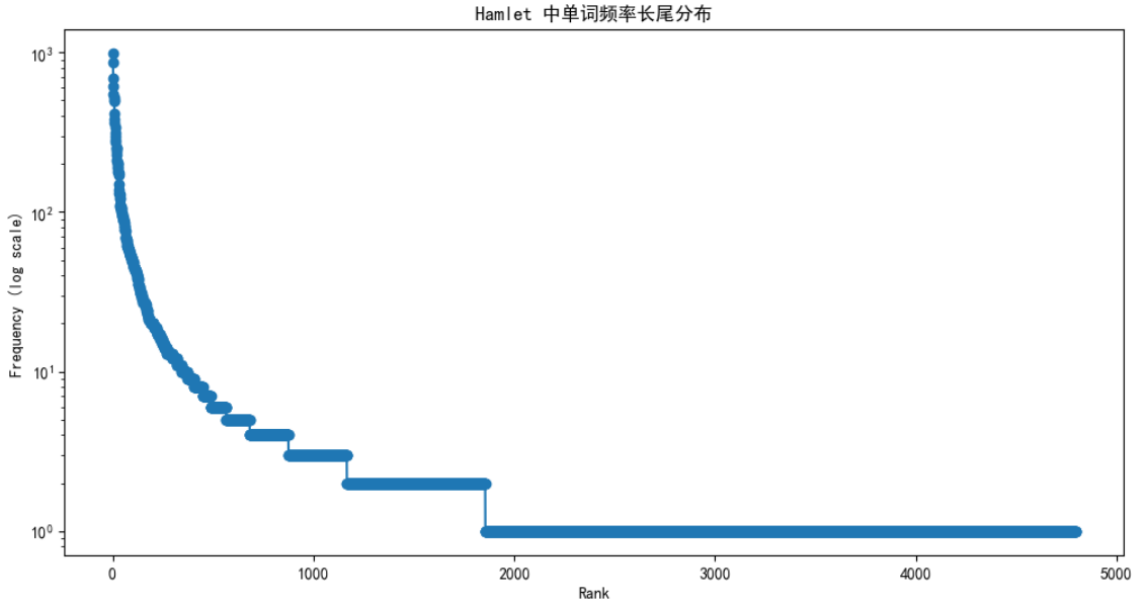

图 6：英文单词统计频率长尾图

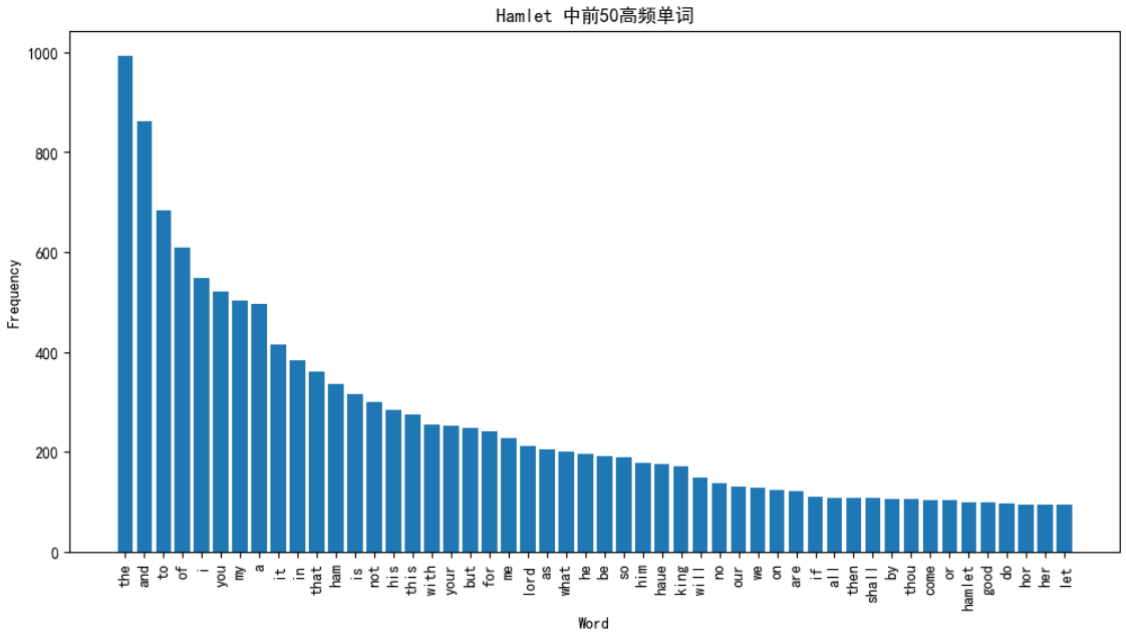

图 6：英文词频统计前50条形图

结论

实验结果表明，无论是中文还是英文，其单词出现频率较高的的频率大约为次一位单词的频率的二倍，即词频长尾图都会呈现指数下降趋势，且根据信息熵结果可以看出，中文无论是字还是词，平均信息熵都要远高于英文，反映出以语素构字的分析语，在信息含量上要远高于使用拼音文字的分析化的屈折语，具有语言学上重要参考价值，未来可以从形态学上相似语种中挑选其他语言进行多次实验，提升实验结果的普适性。

References

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

[2] P. F. Brown, V. J. D. Pietra, R. L. Mercer, S. A. D. Pietra和J. C. Lai, 《An Estimate of an Upper Bound for the Entropy of English》, Comput. Linguist., 卷 18, 期 1.