Report of Deep Learning for Natural Langauge Processing

21374019 孙效宇
21374019@buaa.edu.cn

摘要（Abstract）

本报告探讨了中英文文本的平均信息熵，分别以字和词为基本单位计算信息熵。我们使用 NLTK 工具处理英文语料库 Gutenberg，采用 jieba 库处理中文语料库 wiki_zh。结果显示，中文的字信息熵和词信息熵均高于英文，体现了两种语言在结构和表达上的差异。

引言（Introduction）

信息熵是度量文本信息量的重要指标，反映了文本的不确定性程度。本研究旨在比较中英文文本的平均信息熵，揭示两种语言在字母（字符）和词汇层面的复杂性。通过计算中英文文本的字信息熵与词信息熵，我们可以更深入理解语言结构，并为自然语言处理（NLP）任务提供理论支持。

方法论（Methodology）

本研究旨在计算中英文文本的平均信息熵，分别以字符（字母或字）和词汇（单词或词）为单位。以下是具体的研究方法和步骤：

1.语料预处理：

中文语料库采用 wiki 语料库内容，读取文本内容后进行字符清洗，去除换行符、回车符、制表符、斜杠、引号等特殊符号，同时删除所有英文字母、标点符号和数字，仅保留汉字字符。然后对停用词进行过滤，加载老师提供的中文停用词表 cn_stopwords.txt，过滤掉常见但无实际意义的词语，如“的”、“是”、“和”等。由于中文分词不如英文分词直接，我们引入 jieba 分词库对对文本进行切分，这就完成了中文语料的预处理。

英文文本使用 nltk 库加载 Gutenberg 语料库，并选择莎士比亚的《Hamlet》作为实验文本，去除所有标点符号，仅保留字母和空格。提取文本中的所有字母，并统一转换为小写，以减少大小写对频率统计的影响。最后按照空格进行分词，得到单词列表统计。

2.中英文文本信息熵计算

根据前一步预处理统计出的英文文本字母频率和词频，利用香农公式（式 1）计算了字母信息熵和词信息熵。中文部分内容相似，根据统计出的字频和 jieba 分词的词频，利用香农公式进行计算，这里由于中文语料库较大（共 5.65 亿字符），为避免内存不足问题，我们将文本分块处理，每次处理 100 万字符。

$$H(X) = -\sum P(x_i) log_2 P(x_i) \tag{1}$$

英文实验结果：字母信息熵约为 4.1426 bits 每字，词信息熵：约为 9.3621bits 每词
中文实验结果：字信息熵约为 10.0692 每字，词信息熵约为 14.5775bits 每词

表 1 中英文字词平均信息熵对比

	Bits/Letter	Bits/Word
Chinese	10.0692	14.5775
English	4.1426	9.3621

此外，我们绘制了中英文字（字母）频率、词语（单词）频率长尾图和前 50 字频、词频数统计直方图如下所示：

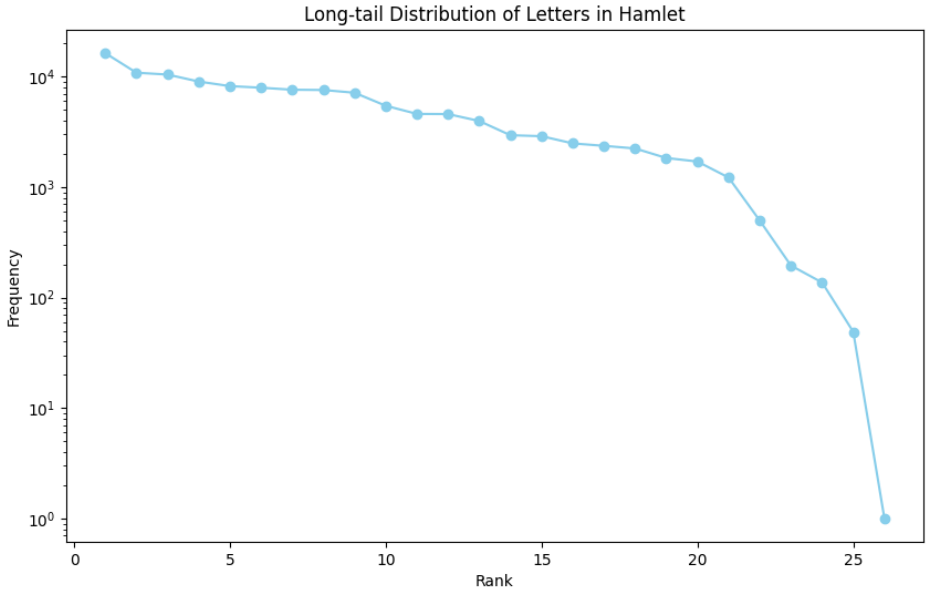

图 1 英文字母统计频率长尾图

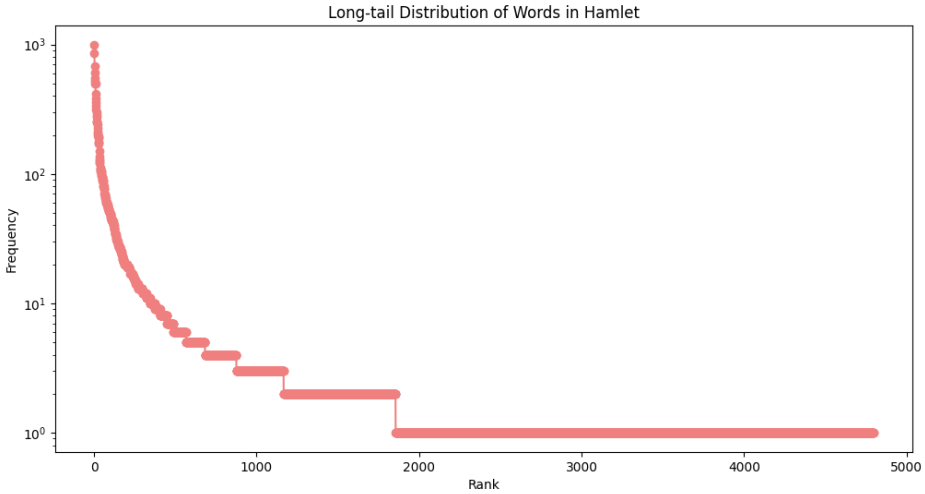

图 2 英文单词统计频率长尾图

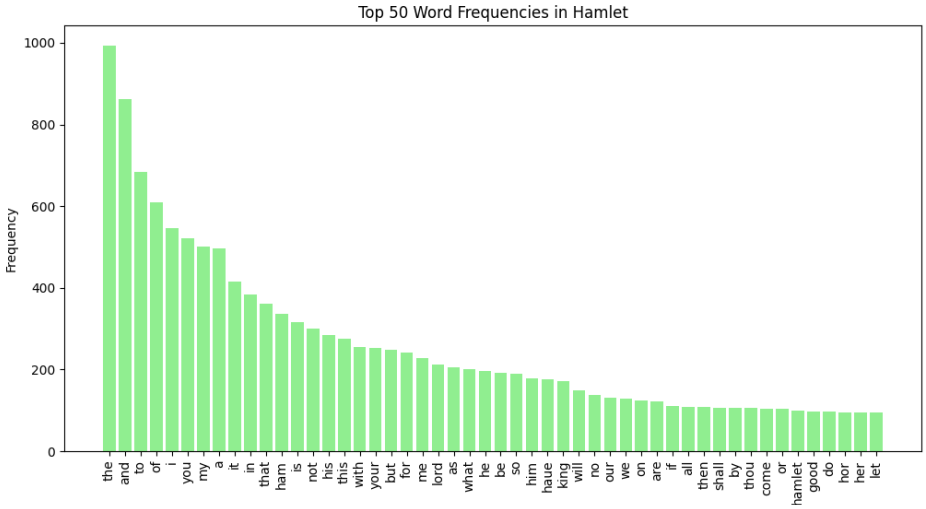

图 3 英文词频统计前 50 条形图

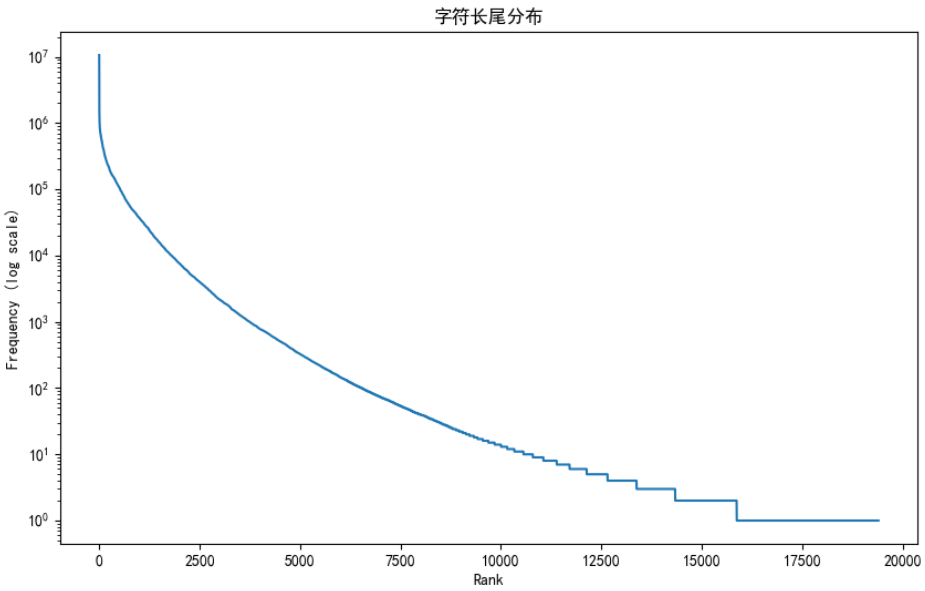

图 4 中文字频长尾图

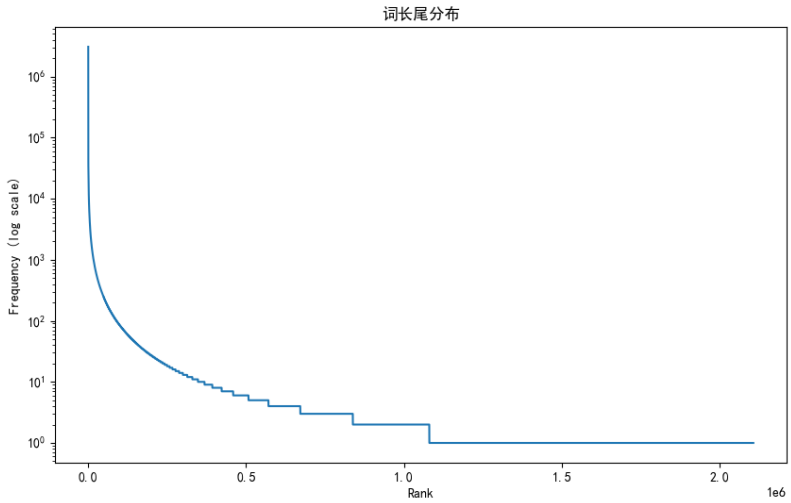

图 5 中文词语频数长尾图

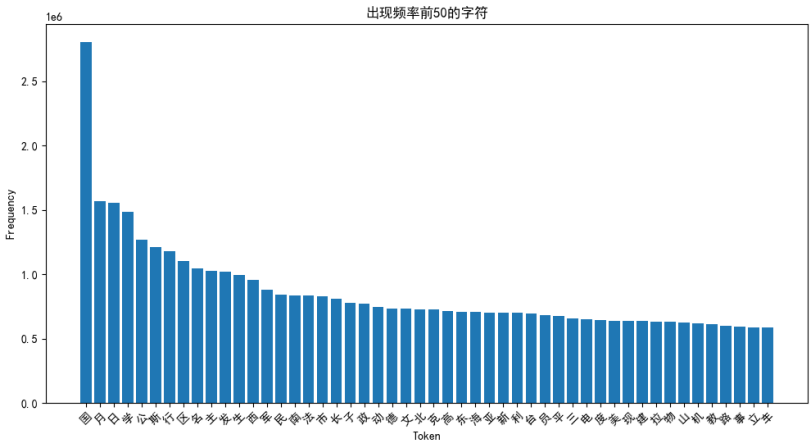

图 6 中文字频前 50 统计直方图

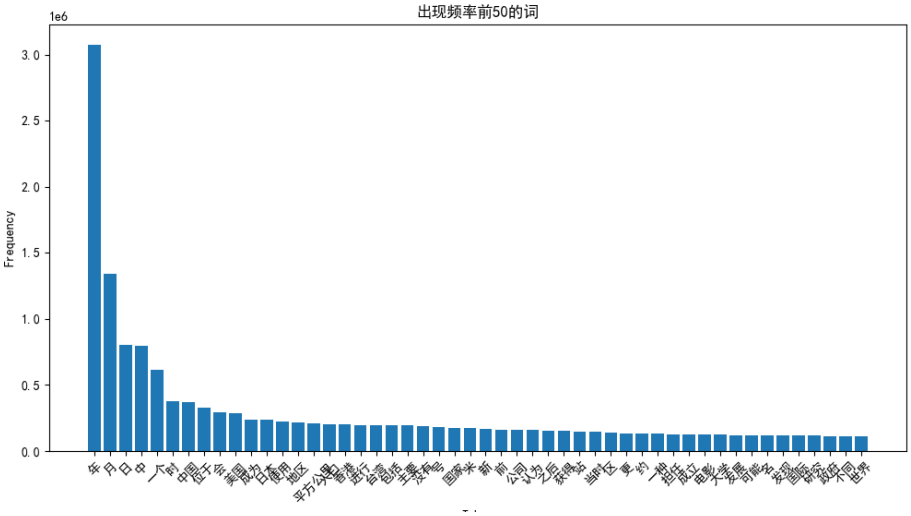

图 7 中文词频前 50 统计直方图

结论（Conclusion）

实验结果表明，中文的字信息熵显著高于英文，这是由于中文字符承载的信息量更大。而在词层面，中文词信息熵也高于英文，反映出中文词语组合的丰富性和复杂性。这些结果对自然语言处理任务如机器翻译、语言建模等具有重要参考价值。未来研究可进一步探讨不同语料库的影响，并尝试更高效的分词算法和计算方法，以提升实验的准确性和可扩展性。