Report of Deep Learning for Natural Langauge Processing

JiaXiang Yu

Zy2403124@buaa.edu.cn

Abstract

本研究通过计算中文维基百科与莎士比亚英文语料库的字符级和词汇级信息熵，揭示了中英文语言系统的统计特性差异。实验结果表明，中文维基百科的字符级熵为 9.855 比特/字符，词汇级熵为 14.3536 比特/词，显著高于莎士比亚英文语料库的 4.15 比特/字符和 9.73 比特/词，印证了汉字系统的高信息密度特性。两类语料均严格遵循 Zipf 定律，高频词分布体现了功能词主导的语言共性，而熵值差异反映了英语冗余结构与中文紧凑表达的对比，为语言建模和跨语言 NLP 任务提供了量化依据。。

Introduction

自然语言具有高度复杂性和不确定性，这种特性使得建模语言规律成为自然语言处理（NLP）的核心挑战。信息论中的信息熵（Shannon, 1948）为量化语言的不确定性提供了数学框架，其通过概率分布衡量随机事件的“信息量”。研究表明，英语的词级熵通常为 8-12 比特，而汉语因词汇的高信息密度呈现更低的熵值。尽管传统 n-gram 模型基于熵理论构建了语言概率模型，但长距离依赖和数据稀疏性问题仍限制了对真实语言分布的准确估计。近年来，深度学习模型虽显著提升了语言生成能力，但其对熵的隐式建模机制尚未完全明晰。

本文所使用的语料库分别是维基百科 2019 语料库和莎士比亚语料库。其中，维基百科语料库是基于全球最大在线百科全书——维基百科（Wikipedia）构建的大规模多语言文本资源。截至 2023 年，维基百科涵盖超过 300 种语言版本，总词量超过 60 亿（英文版独占约 45 亿词条），内容覆盖科学、历史、文化、技术等几乎所有人类知识领域。其结构化数据（如超链接、分类标签、跨语言对照）与非结构化文本的结合，使其成为自然语言处理（NLP）领域最广泛使用的语料库之一。

莎士比亚语料库是基于英国文学巨匠威廉·莎士比亚经典戏剧作品构建的高质量文学文本资源。该语料库收录了《凯撒大帝》《哈姆雷特》《麦克白》三部核心悲剧，总词量逾 30 万，涵盖权力、复仇、道德等人性主题的深度探讨。其剧

本文本完整保留早期现代英语的诗体结构、角色独白及舞台指示，同时蕴含丰富的戏剧元数据。文学叙事与非结构化语言特征的结合，使其成为古典文学计算分析、风格生成模型训练的首选语料库之一。

Methodology

本文的任务是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。

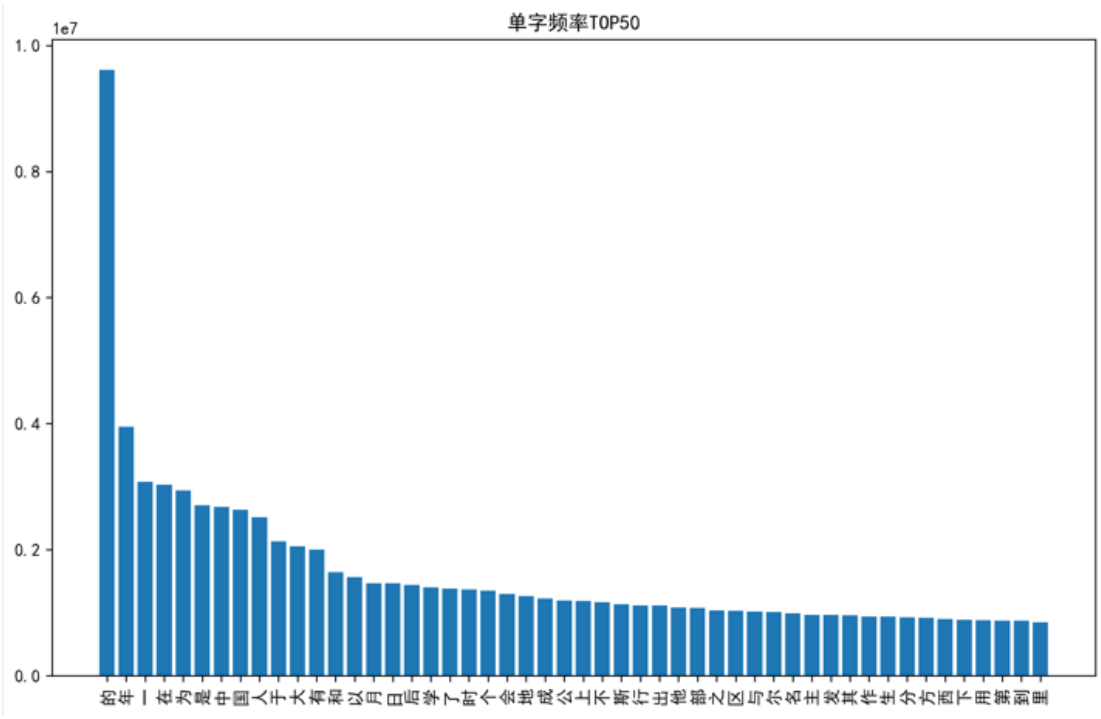
Step1: 数据预处理

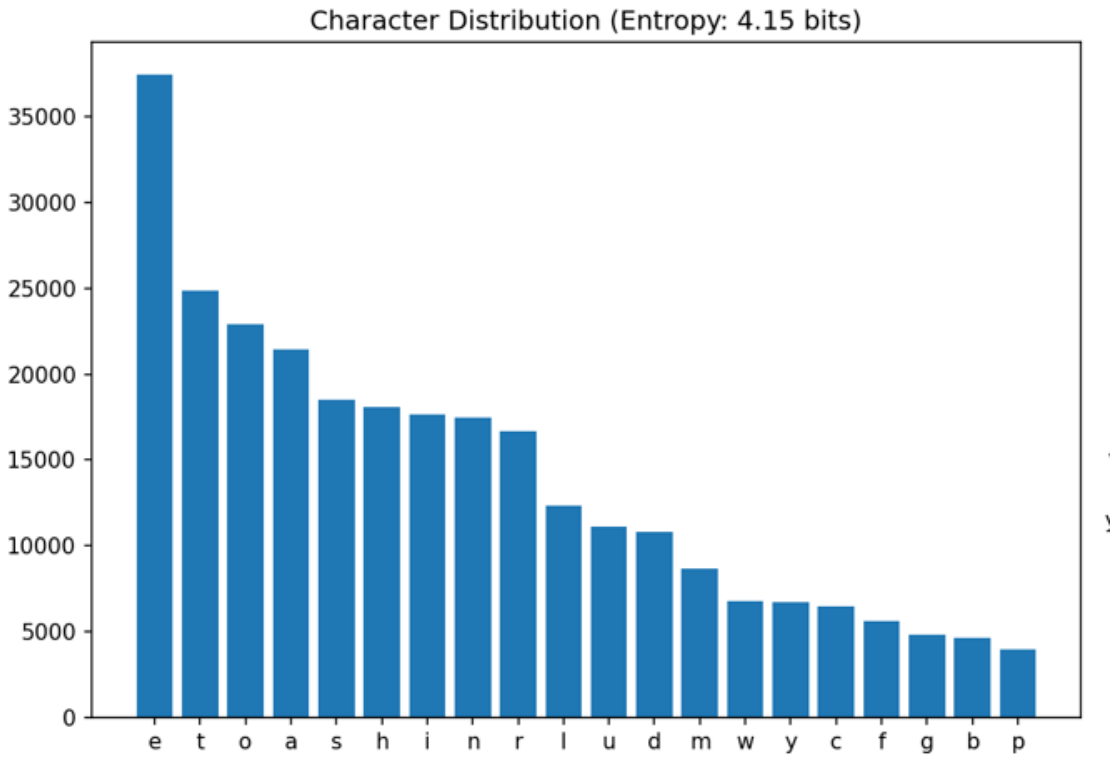
本研究针对莎士比亚戏剧语料库与中文维基百科语料库分别设计预处理流程。对于莎士比亚戏剧文本，首先通过NLTK库加载原始剧本文件，使用正则表达式合并连续空行并统一转换为小写字母，随后移除所有非字母字符（包括标点、数字及特殊符号），仅保留基本拉丁字母和空格以构建纯净字符序列，同时采用正则化提取长度≥2的词汇用于词级统计。

针对中文维基百科，从分布式存储文件中逐文档读取文本，利用正则表达式过滤非汉字字符，通过jieba分词工具对清洗后的文本进行精确分词，并移除空白符与孤立标点。两类语料库均进行低频词过滤和词形归一化处理，最终生成标准化的字符与词汇序列，为后续信息熵计算及分布分析提供基础数据。预处理过程中，莎士比亚文本额外处理了早期现代英语的拼写变体，而中文文本则通过句号切分语句以支持句子级熵值计算。

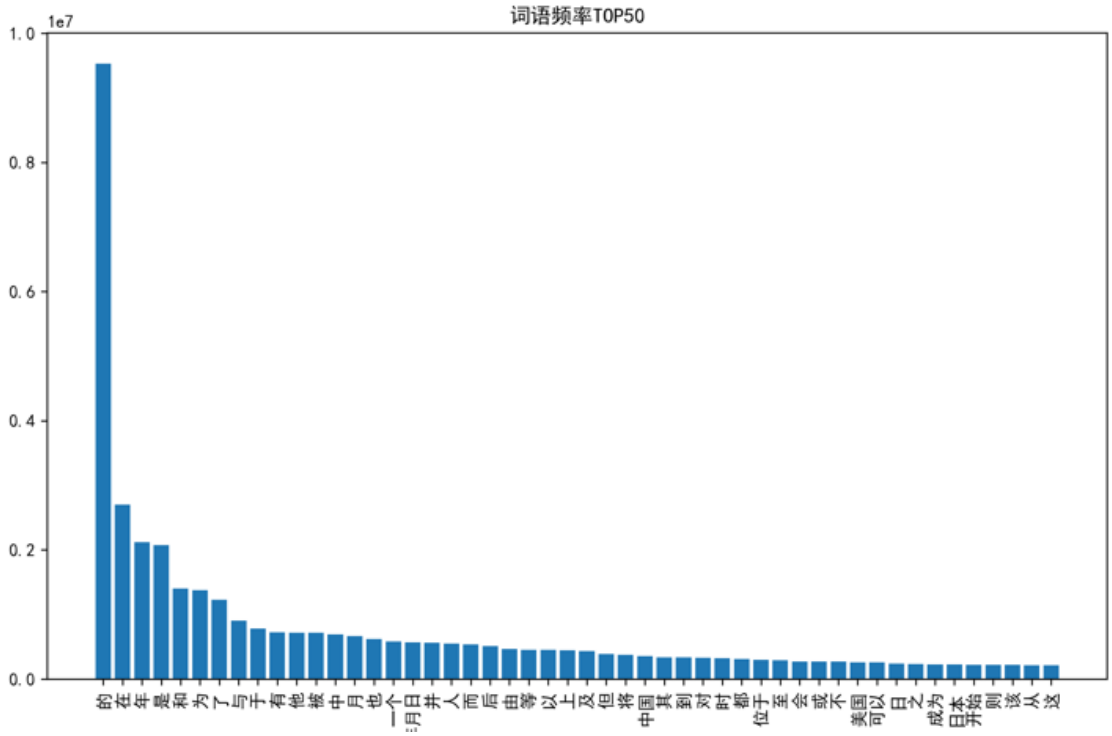
Step2: 统计分布分析

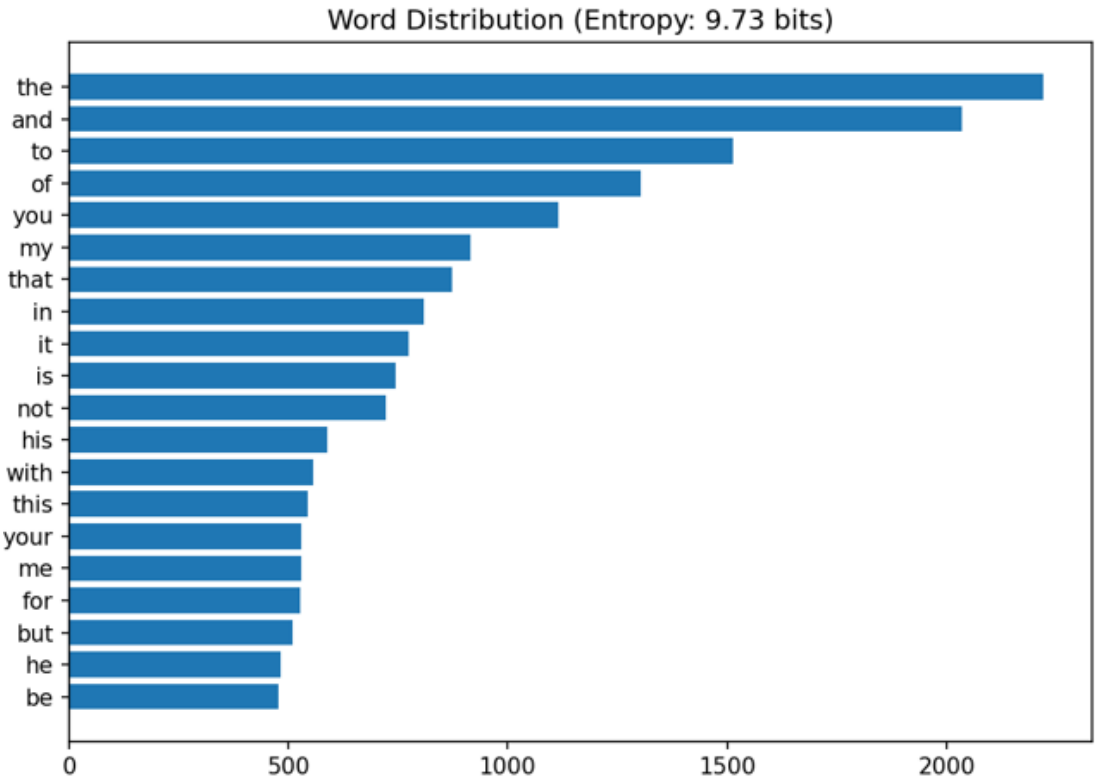
在本次自然语言处理作业中，我对中文 Wiki 百科 2019 数据集和莎士比亚著作进行了统计分析。首先，从字符分布来看，莎士比亚文本中的字母 "e" 最为常见，频次显著高于其他字符，紧随其后的为 "t"、"o"、"a" 等字母，这符合英文语言中常见的频率分布模式。而在中文 Wiki 数据集中，最频繁的字符是 "的" 和 "一"，这反映了中文中虚词和常用词的高出现频率。这两种语言的字符频率差异凸显了中英文在基本构造上的显著不同。下图直观展示了两组数据的字符分布情况：



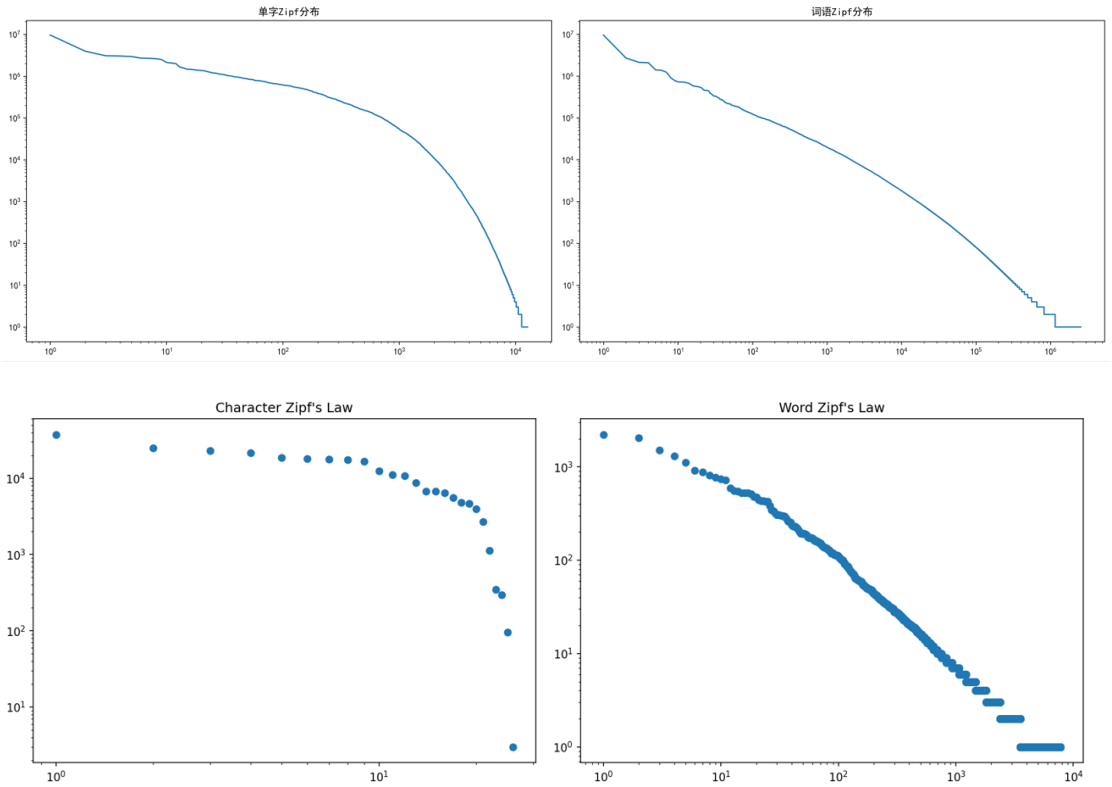


在词频统计方面，莎士比亚文本中 "the"、"and"、"to" 等词频最高，表现出英文中功能词主导的特性；而中文 Wiki 数据集中，类似 "的"、"是" 等高频词同样属于虚词类，进一步展现了两种语言在语言结构上的相似性。具体的词频分布如下图所示：





在 Zipf 定律的分析中，图表清楚地展现了两种数据集都遵循典型的幂律分布。无论是字符还是词汇，其频率都与其排名呈现出较明显的负幂次关系，进一步印证了 Zipf 定律在自然语言中的普适性。具体的 Zipf 定律曲线如下所示：



这种分布规律对自然语言处理有着重要意义，尤其在模型优化、数据压缩和词汇稀疏性问题的处理中具有参考价值。

Experimental Studies

对英文，莎士比亚语料库中的多部英文文学作品，分别计算字符级和词汇级信息熵。对中文，使用中文维基百科语料库，计算字级和词级信息熵。得到的计算结果如表1所示。

表 1 中英文字词平均信息熵

	字符/字母	单词/词语
维基百科	9.855bits/char	14.3536/word
莎士比亚	4.15bits/char	9.73bots/char

实验通过对中文维基百科与莎士比亚英文语料库的信息熵计算，揭示了两类语言的显著统计差异与共性。中文维基百科的字符级信息熵为 9.855 比特/字符，词汇级信息熵为 14.3536 比特/词，其单字与词语的高熵值印证了汉字系统的信息密度优势；莎士比亚英文语料库的字符级熵为 4.15 比特/字符，词汇级熵为 9.73 比特/词，低频值反映英语依赖冗余结构降低理解门槛。高频词分布进一步体现语言文化差异。两类语料均严格遵循 Zipf 定律，验证了自然语言的普遍统计规律

性。

Conclusion

中英文在信息熵分布上呈现显著差异，中文凭借单字高信息熵展现更强的语义承载能力，而英语通过冗余结构降低理解复杂度。两类语言高频词均以功能词为主，且严格遵循 Zipf 定律的幂律分布，验证了自然语言统计规律的普适性。研究结果为语言模型优化、跨语言生成任务及信息压缩算法设计提供了理论支撑，同时揭示了深度学习模型隐式建模语言统计特征的内在潜力。