**Report of Deep Learning for Natural Langauge Processing**

武学谦

[sy2303705@buaa.edu.cn](mailto:sy2303705@buaa.edu.cn)

**Abstract**

本实验旨在计算英文与中文在字符和词汇级别的平均信息熵，以量化两种语言的信息密度差异。基于NLTK的Gutenberg英文语料库和中文维基百科词条数据，实验结果表明：英文字符级平均信息熵显著高于中文，而中文词级平均信息熵更高。这一发现可为自然语言处理任务中的模型设计与优化提供理论支持

**Introduction**

信息熵是衡量语言信息不确定性的核心指标，由香农在通信理论中提出，后被广泛应用于自然语言处理领域‌。英文与中文因语言结构差异（如汉字表意特性与英文字母组合特性），其信息分布模式存在显著不同。现有研究多聚焦单一语言熵值分析，但对跨语言对比的探索较少‌。本研究通过计算字符级与词级信息熵，揭示中英文在信息压缩潜力、模型训练复杂度等方面的差异，为多语言大模型设计提供数据支持。

**Methodology**

**‌**英文语料库‌：采用NLTK的Gutenberg语料库，包含18部经典文学作品，通过正则表达式去除标点与数字，保留纯文本单词与字符序列‌。

‌中文语料库‌：使用维基百科中文词条，过滤非中文字符，过滤中文标点符号，经Jieba分词工具进行分词处理。

统计每个字符出现频率，计算每个字符的出现概率

通过概率计算每个字符的信息熵并累加

最后计算平均信息熵：

其中为字符总数。

**Experimental Studies**

经编程计算得到中英文字和词的总信息熵与平均信息熵，如表1所示，可以看出，在总信息熵方面，中文在词级和字符级都比英文的总信息熵要高。在平均信息熵方面，中文的词级平均信息熵和字符级平均信息熵更小。

表1 中英文字和词的总信息熵与平均信息熵

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 语言 | 总信息熵（词） | 平均信息熵（词） | 总信息熵（字） | 平均信息熵（字） |
| 英文 | 9.8605 | 0.0001878 | 4.0995 | 0.1577 |
| 中文 | 12.9221 | 0.00009386 | 9.8029 | 0.001673 |

**Conclusions**

中文字符级平均信息熵较低，可能与汉字高频字符（如“的”、“是”）在语料库中的集中分布相关‌。这一结果部分验证了中文书面文字通过高频字复用降低冗余度的特性，但其总信息熵仍因汉字基数庞大（超6万字）而维持较高水平‌。   
 中文词级平均信息熵更低，反映了词汇组合的高度灵活性。中文词级平均熵仍低于英文，但字符级与词级总信息熵更大，反映汉字系统通过多层级组合（单字→复合词）实现信息密度的非线性叠加‌  
 ‌中文字符与词级总信息熵更大，印证了其书面文字的高信息密度优势‌，但也暴露了实际应用中的矛盾：高熵值提升信息传递效率，但现有计算机字库（如GB18030标准仅支持27,484字）难以覆盖全部汉字，导致生僻字处理困境。‌模型设计中仍需平衡字符嵌入压缩（应对低字符熵）与词表动态扩展（应对高词级熵）的需求‌。

‌