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

何明璞

1113490780@qq.com

**Abstract**

本研究基于NLTK Gutenberg英文语料库和维基百科中文语料库，分别计算了两种语言以字母和以词为单位的平均信息熵，定量比较了英文和中文的信息熵特征。实验结果显示，无论是以词还是以字母（字）为单位，中文的总信息熵都显著高于英文，这反映了中文语言系统的复杂性和高不确定性‌。但是中文的平均信息熵则在两种情况下都低于英文。本报告验证了信息熵作为语言复杂度量化指标的有效性，并为跨语言自然语言处理模型优化提供数据支持。‌

**Introduction**

信息熵是衡量语言不确定性和信息密度的核心指标。在自然语言处理中，信息熵可用于优化分词算法、评估语言模型性能，并为机器翻译等任务提供理论依据‌。本研究选取英文（NLTK Gutenberg）和中文（维基百科）两大代表性语料库，通过统一方法计算信息熵，揭示语言结构的本质差异‌。

**Methodology**

1. 语料库与预处理

**英文语料**‌：NLTK Gutenberg语料库包含18部经典英文文学作品，经去除非字母字符、移除标点后保留纯文本单词。

‌**中文语料**‌：维基百科中文词条经去除非中文字符，保留约500万字连续语料。在对语料进行分词时使用jieba分词，以提高分词正确率。‌

1. 信息熵计算

首先计算每个单词/字母出现的概率，然后使用以下公式计算总信息熵：

随后计算平均信息熵，公式如下：

**Experimental Studies**

经过计算得到中文和英文的词和字母（字）的平均信息熵，如表1所示：

表1：中英文词和字母（字）的总信息熵及平均信息熵

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Language | Entropy(word) | Avg\_entropy(word) | Entropy(letter) | Avg\_entropy(letter) |
| English | 9.8604 | 0.0001878 | 4.0995 | 0.1576 |
| Chinese | 12.9221 | 0.00009386 | 9.8029 | 0.0016 |

从表一中可以看出，在总信息熵方面，无论是词还是字母（字），中文的信息熵都要高于英语。但是中文的平均信息熵则在两种情况下都低于英文。

**Conclusions**

根据以上实验，可以得出以下结论：

1. 跨语言信息熵差异验证：

总信息熵对比‌：中文在词和字母（字）层面的总信息熵均显著高于英文。这一差异源于中文单字多义性、复合词组合复杂度更高，而英文依赖有限字母的线性组合传递语义‌。

‌平均信息熵对比‌：中文的平均信息熵无论是在词还是在字母（字）层面都低于英文的平均信息熵。在字母（字）层面，出现这种情况可能的原因是：英文只有26个字母，而中文包含约7000个常用汉字‌。在词的层面，可能的原因是上下文强依赖性‌：中文词语的语义高度依赖邻近字词（如成语固定搭配），降低了单个单元的不确定性‌。

‌2. 对NLP模型优化的启示：‌

‌中文处理需强化上下文建模‌：中文高信息熵要求模型配备更强的注意力机制，以捕捉跨字符的长距离语义关联‌。

‌英文处理的词表优化策略‌：英文信息熵较低，可通过子词切分平衡词表规模与语义粒度，而字符级模型需更多关注高频字母组合的统计规律‌。

本实验为跨语言NLP任务提供了量化依据，未来可扩展至低资源语言研究及多模态信息熵联合分析。