# 中文语料库信息熵的计算


## 摘要

在自然语言处理中，信息熵描述无限长文本中一个词元的平均信息量。我们知道交叉熵是信息熵的一个上界，可以使用交叉熵来估计信息熵。本文对指定的中文语料库，使用 Jieba 分词器分词，建立三元语法，并估计其字和词的信息熵。



**关键字**：自然语言，中文，信息熵，三元语法模型


## 一、语言模型

在自然语言处理(NLP，Nature Language Process)中，词元（token）是文本的基本单位。在中文语料库中，词元可以是字也可以是词。将文本数据映射为词元，将词元看作一系列离散的观测值。

语言模型（LM， Language Model），是一串词元序列的概率分布，其作用是为长度为$m$的词元序列$t_1t_2\dots t_m$确定一个概率分布$P(t_1t_2\dots t_m)$。

对于一个词元序列$t_1t_2\dots t_{m-1}$，预测这个词元序列的后续词元$t_m$，就是求概率$P(t_m|t_1t_2\dots t_{m-1})$的有效估计。可以采用频数估计概率，即$P(t_m|t_1t_2\dots t_{m-1})\approx\frac{c(t_1t_2\dots t_m)}{c(t_1t_2\dots t_{m-1})}$，其中$c(t_1t_2\dots t_m)$表示词元$t_1t_2\dots t_m$在语料库中的频数。但当词元序列$t_1t_2\dots t_m$较长时，其在语料库中出现的频率非常低，导致概率$P(t_m|t_1t_2\dots t_{m-1})$的估算非常困难。

研究人员提出假设任意一个词元$t_i$出现的概率只同它前面的$\tau$个词$t_{i-\tau}\dots t_{i-1}$有关，即只将某个固定长度$\tau$的词元序列$t_{m-\tau}\dots t_{m-1}$作为输入，估计$P(t_m|t_{m-\tau}\dots t_{m-1})$的概率，这种假设称为马尔可夫假设。其对应的模型被称为$\tau$阶马尔可夫模型，其对应的涉及$n$个变量的概率公式的统计语言模型被称为$n$元语言模型（n-gram）。一元语言模型认为每个词元出现的概率与其他词元无关，$n$元语言模型认为每个词元出现的概率只与其前$n-1$个词元相关。

对于词元序列$t_1t_2t_3t_4$，其一元语言模型的概率为

$$P(t_1t_2t_3t_4)=P(t_1)P(t_2)P(t_3)P(t_4)$$

其二元语言模型的概率为

$$P(t_1t_2t_3t_4)=P(t_1)P(t_2|t_1)P(t_3|t_2)P(t_4|t_3)$$

其三元语言模型的概率为

$$P(t_1t_2t_3t_4)=P(t_1)P(t_2|t_1)P(t_3|t_1t_2)P(t_4|t_2t_3)$$

## 二、文本数据预处理

三元语言模型是基于词元的，所以需要对中文文本语料库进行分词。本文采用 Jieba 中文分词工具对句子进行分词，其中非中文字符使用<unk>替代，标点符号使用<space>替代。

## 三、信息熵

信息熵，描述信息的混乱程度。对于给定的随机变量$X=\{X_1,X_2,\dots,X_n\}$，且事件$X_i$的概率为$p_i$，则随机变量$X$的信息熵

$$H(X)=E[-\log p_i]=\sum_{i=1}^{n}p_i(-\log p_i)=-\sum_{i=1}^{n}p_i\log p_i$$

在自然语言处理领域，信息熵描述无限长文本中一个词元的平均信息量，或对文本编码所需的比特位。它表征了自然语言的复杂性，如果很容易预测下一个词元，那么这个词元就很容易压缩。

然而自然语言文本的真实概率$P(X)$分布是无法获得的，只能得到其非真实分布$M(X)$。在自然语言处理中，当训练数据足够大时，我们可以假设训练数据所服从的分布是真实数据的分布。从文章 “An Estimate of an Upper Bound for the Entropy of English” [1]中，可以类比出中文语言的信息熵$H(chinese)\leq H_{token}(sentence)$(这里为简化模型，只考虑词元 token 的信息熵)，因此可以使用$H_{token}(sentence)$作为中文语言信息熵的近似估计。

对于三元语法模型，由词元$t_1t_2\dots t_n$组成的句子 sentence，其概率$M(t_1t_2\dots t_n)=M(t_1t_2)\prod_{i=3}^{i=n}M(t_i|t_{i-2}t_{i-1})$，而$M(t_i|t_{i-2}t_{i-1})=\frac{M(t_{i-2}t_{i-1}t_i)}{M(t_{i-2}t_{i-1})}$。因此中文信息熵的近似估计

$$H_{token}(sentence)=-\frac{1}{3}\sum M(t_1t_2t_3)\log M(t_3|t_1t_2)=-\frac{1}{3}\sum M(t_1t_2t_3)\log_2\frac{M(t_1t_2t_3)}{M(t_1t_2)}$$

其中信息熵的单位为比特/词元。

三元语法模型需要统计每个二元词元组在语料库中出现的频数，得到二元模型词元频表，作为计算条件概率$M(t_i|t_{i-2}t_{i-1})$的分母$M(t_{i-2}t_{i-1})$，并且需要统计每个三元词元组在语料库中出现的频数，得到三元模型词元频表。将字和词分别作为词元计算信息熵。

## 四、实验结果

在三元语法模型中使用的语料库字数为 7272471 字，词元个数为 5372295 词，计算得到基于字的三元语法模型的中文信息熵为 1.135 比特/字，基于词的三元语法模型的中文信息熵为 1.133 比特/词元。

参考文献

[1] An Estimate of an Upper Bound for the Entropy of English

[2] Cross Entropy of Neural Language Models at Infinity—A New Bound of the Entropy Rate