深度学习与自然语言处理第一次作业报告

张甫成
sy2206303@buaa.edu.cn

摘要

本文是深度学习与自然语言处理第一次作业的实验报告，分别将金庸小说全集看做字序列和词序列，并使用 n-gram 模型对进行建模，求出了 n=1 时模型的熵，作为语料库熵的估计，并通过比较得出结论：词的信息量高于字的信息量。

绪论

信息熵（information entropy）是信息论的基本概念。描述信息源各可能事件发生的不确定性。20 世纪 40 年代，香农（C.E.Shannon）借鉴了热力学的概念，把信息中排除了冗余后的平均信息量称为“信息熵”，并给出了计算信息熵的数学表达式。信息熵的提出解决了对信息的量化度量问题。[1]

对于一个中文语料库而言，使用不同的切分方法（字、词）进行建模可以计算得到不同的信息熵。本文使用不同的切分方法获得 token 序列，并用 n-gram 模型对金庸小说全集进行建模，计算其信息熵。

方法

我们首先介绍分割语料库的方法，然后介绍语言模型的构造方法，最后给出模型条件熵的计算方法。

字序列和词序列的获取方法

对于字序列，我们将语料库视作字序列，然后删除标点符号。对于词序列，我们使用 jieba 库对语料库进行分词，并删除停用词。

n-gram 模型的构造

N-gram 模型统计文本的 token 序列中不同 N 元组的数量，并试图通过前 N-1 个 token 预测第 N 个 token 的概率，通过循环生成整个序列。当 N=3 时，N-gram 模型的形式化定义如公式(1)所示[3]。本文中，我们对[3]的模型进行了简化，将公式(1)中的 M 视为概率 token 序列出现的概率或条件概率。

$$M_{token}(t_1 t_2 \dots t_n) = M_{token}(t_1 t_2) \prod_{i=3}^{n} M_{token}(t_i \mid t_{i-2} t_{i-1}) \tag{1}$$

模型条件熵的计算

可以通过计算上一节中 N-gram 模型的条件熵来估计语料库的信息熵上界。

语料库 X 的熵表达式如公式(2)所示，其中 P 表示语料库 P 的概率分布，E_P 表示求期望的算子。

$$H(X) \equiv H(P) \equiv -E_P \log P(X_0 \mid X_{-1}, X_{-2}, \dots) \tag{2}$$

N-gram 模型是 P 的近似，其条件熵的形式化表达如公式(3)所示。可以证明，该条件熵是语料库的熵的上界的估计，模型越准确，熵的上界估计越准确。[3]本文中，我们计算了 n=1

是模型的熵。

$$H(P,M) = \lim_{n\to\infty} -E_P \log M(X_0 \mid X_{-1}, X_{-2}, \ldots, X_{-n}) \tag{3}$$

实验结果

经过计算可知，字模型的熵约为 9.5，词模型的熵约为 13.1。

```
字模型的熵 9.492270902462609
Building prefix dict from the default dictionary ...
Loading model from cache /tmp/jieba.cache
Loading model cost 0.461 seconds.
Prefix dict has been built successfully.
词模型的熵 13.125313357446846
```

Figure 1：this is obviosuly overfit

结论

我们分别用字模型和词模型对金庸小说全集进行建模，通过计算得到了模型的信息熵作为中文信息熵的上界估计，得到了词的信息量高于字的信息量的结论。

参考文献

[1] 百度百科. 信息熵[EB/OL]. 北京：百度百科, 2022. https://baike.baidu.com/item/信息熵/7302318.

[2] 忆臻. 通俗理解条件熵[EB/OL]. 知乎, 2017-05-22. https://zhuanlan.zhihu.com/p/26551798.

[3] Mori S , Yamaji O . An Estimate of an Upper Bound for the Entropy of Japanese[J]. Ipsj Journal, 1997, 38:2191-2199.