Report of Deep Learning for Natural Langauge Processing

刘新宇
LXYbhu@buaa.edu.cn

Abstract

本报告通过阅读Entropy_of_English_PeterBrown，参考此文章来计算中文的平均信息熵。数据集的链接为 https://share.weiyun.com/5zGPyJX，内容为金庸先生的16本小说，采用一元、二元、三元模型来计算数据集的信息熵。

Methodology

信息熵的概念最早由香农（1916-2001）于1948年借鉴热力学中的“热熵”的概念提出，旨在表示信息的不确定性。熵值越大，则信息的不确定程度越大。其数学公式如式（1）

$$H(x) = E[-\log(P(x))] = -\sum_{x \in X} P(x)\log(P(x)) \quad (1)$$

式（1）中的对数通常以 2 为底，单位为比特。

信息量度量的是一个具体事件发生了所带来的信息，而熵则是在结果出来之前对可能产生的信息量的期望——考虑该随机变量的所有可能取值，即所有可能发生事件所带来的信息量的期望。对于任何一个事件，通常来说，它的不确定性越大，那么其信息熵也越大。越是不确定的事件，如果我们得到了一条有关的信息，那么信息量就会很大，即信息熵与信息的价值成正相关。

语言模型（language model，LM）在自然语言处理中占有重要的地位，它是对自然语言的建模，其任务是预测一个句子在语言中出现的概率。语言模型的定义是：给定语言序列$w_1, w_2, w_3, \cdots, w_n$，语言模型就是计算该序列的概率，即$P(w_1, w_2, w_3, \cdots, w_n)$。从机器学习的角度来看，语言模型是对语句的概率分布的建模，通俗的解释即判断一个语言序列是否是正常语句。

在统计学模型横行NLP的时代，语言模型任务中最常使用是N-gram语言模型。为了简化$P(w_n|w_1, w_2, w_3, \cdots, w_{n-1})$的计算，我们引入一阶马尔可夫假设：每个词只依赖前一$P(w_n|w_1, w_2, w_3, \cdots, w_{n-1}) \approx P(w_n|w_{n-1})$，我们也可以引入二阶马尔科夫假设：每个词依赖前两个词$P(w_n|w_1, w_2, w_3, \cdots, w_{n-1}) \approx P(w_n|w_{n-2}, w_{n-1})$，有了马尔可夫假设，就可以方便地计算条件概率。以N=3的tri-gram语言模型为例，它使用二阶马尔可夫假设，$P(w_n|w_1, w_2, w_3, \cdots, w_{n-1}) \approx P(w_n|w_{n-2}, w_{n-1})$，对于$P(w_n|w_{n-2}, w_{n-1})$，可以得到它的概率值$P(w_n|w_{n-2}, w_{n-1}) = \frac{\text{count}(w_{n-2}, w_{n-1}, w_n)}{\text{count}(w_{n-2}, w_{n-1})}$，其中count(*)表示在*在训练集中出现的次数。所以，N-gram语言模型有两个要点：（1）使用N-1阶马尔可夫假设简化后验概率，提高模型的泛化能力；（2）使用数数法计算后验概率，蕴含着最大似然估计的思想。

Experimental Studies

数据集为金庸先生的16本小说，其中包含了乱码与无用或重复的中英文符号，因此需要对该实验数据集进行预处理：（1）删除所有的隐藏符号；（2）删除所有的非中文字符；（3）不考虑上下文关系的前提下删去所有的标点符号这里的预处理用jieba进行分词。jieba是python中的一个中文分词库，在本实验中以精确模式进行分词。分词后，还对停用词（stop words）进行了删除。对于按照字来统计信息熵的情况，没有做任何的删除处理。

根据上一节中的N元模型公式，计算信息熵，若n=1，直接计算一元模型进而得到信息熵；对于n>=2，分别计算N元词频模型和N-1元词频模型，计算条件信息熵，实验结果如表（1）与表（2）：

表1 信息熵（字）

小说	1-gram	2-gram	3-gram
三十三剑客图	10.0503	3.4441	0.7263
书剑恩仇录	9.8583	4.6374	1.8821
侠客行	9.5035	4.4237	1.8183
倚天屠龙记	9.7706	5.0116	2.1675
天龙八部	9.8887	5.0866	2.1783
射雕英雄传	9.8168	4.9754	2.1798
白马啸西风	9.1982	3.0427	1.3212
碧血剑	9.8010	4.7115	1.7905
神雕侠侣	9.6743	5.1705	2.1787
笑傲江湖	9.6102	4.9143	2.1251
越女剑	8.6164	2.4000	0.9969
连城诀	9.5919	4.1782	1.5627
雪山飞狐	9.5337	3.8860	1.3864
飞狐外传	9.6787	4.5618	1.8642
鸳鸯刀	9.2049	2.7862	0.9888
鹿鼎记	9.7428	4.9523	2.2457
ALL	10.0916	7.0873	3.2491

表2 信息熵（词）

小说	1-gram	2-gram	3-gram
三十三剑客图	10.0503	3.4441	0.7263
书剑恩仇录	9.8583	4.6374	1.8821
侠客行	9.5035	4.4237	1.8183
倚天屠龙记	9.7706	5.0116	2.1675
天龙八部	9.8887	5.0866	2.1783
射雕英雄传	9.8168	4.9754	2.1798
白马啸西风	9.1982	3.0427	1.3212
碧血剑	9.8010	4.7115	1.7905
神雕侠侣	9.6743	5.1705	2.1787
笑傲江湖	9.6102	4.9143	2.1251
越女剑	8.6164	2.4000	0.9969
连城诀	9.5919	4.1782	1.5627
雪山飞狐	9.5337	3.8860	1.3864
飞狐外传	9.6787	4.5618	1.8642
鸳鸯刀	9.2049	2.7862	0.9888
鹿鼎记	9.7428	4.9523	2.2457
ALL	10.0916	7.0873	3.2491

Conclusions

对比1-gram、2-gram、3-gram三种语言模型得到的结果可以看到，N取值越大，即考虑前后文关系的长度越大，不同词出现的个数越多，这是因为长度的增加也增加了由字组合成词的组合个数，所以会出现更多不同的词。

而随着N取值变大，文本的信息熵则越小，这是因为N取值越大，通过分词后得到的文本中词组的分布就越简单，N越大使得固定的词数量越多，固定的词能减少由字或者短词打乱文章的机会，使得文章变得更加有序，减少了由字组成词和组成句的不确定性，也即减少了文本的信息熵，符合我们的实际认知。

通过这次作业，也让我对信息熵有了更多的理解。信息熵是消除不确定性所需信息量的度量，也即未知事件可能含有的信息量。在自然语言处理中，信

息熵只反映内容的随机性（不确定性）和编码情况，与内容本身无关，而随机变量的信息熵大小是客观的，又是主观的，与观测者的观测粒度有关。

References

[1] Brown P F, Della Pietra S A, Della Pietra V J, et al. An estimate of an upper bound for the entropy of English[J]. Computational Linguistics, 1992, 18(1): 31-40.