深度学习与自然语言处理第一次报告

郭子龙

1216199336@qq.com

摘要

本文以金庸的小说作为中文语料库，去除语料库中的中文停词和常见符号后，通过jieba库实现词语分割，进而计算以字为单位的信息熵、以词为单位的一元信息熵，二元信息熵和三元信息熵。得出的结果为，以字为单位的信息熵要小于以词为单位的信息熵，同时三元信息熵要小于一元信息熵。

1. 介绍

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。具有三个基本性质：

1、单调性，发生概率越高的事件，其携带的信息量越低；

2、非负性，信息熵可以看作为一种广度量，非负性是一种合理的必然；

3、累加性，即多随机事件同时发生存在的总不确定性的量度是可以表示为各事件不确定性的量度的和，这也是广度量的一种体现。

本文的主要工作包括：

1. 文件读取和处理；以金庸小说为语料库，实现对语料库的预处理，删去无明显意义的停词和标点符号。
2. 语料库的划分，以jieba库和字为单位对语料库进行分割，统计和处理。
3. 分别计算以字为单位、以词为单位的一元模型、二元模型和三元模型下中文的信息熵。

2. 方法

M1: 数据读取与预处理模块

第一部分对文本进行读取，首先删去原文中的标点符号，转化为回车符，这一步骤在应对多标点符号相邻的时候，会导致有很多空行出现，对结果无影响但

影响文本的美观性，在之后的工作中会加以改进。随后依据停词语料库删去文本中相关停词。第一步骤仍不够细化，例如不同的标点应该有不同的处理方式，文本中包含着一些特殊的网站和无意义的广告文字，在处理之初没有删去。

M2: 语言模型的构建

根据信息香农定义的随机变量 X 的信息熵 H 的公式：

$$H(X)=E[I(X)]=E[-\ln(P(X))]$$

其中 $P(X)$ 指 X 的概率质量函数，E 为期望函数，$I(X)$ 为信息量，是一个随机变量，当样本有限时，其又可表示为：

$$H(X)=\sum_i P(x_i)I(x_i)=-\sum_i P(x_i)\log_b P(x_i)$$

本文中b取2时，熵的单位为bit，条件熵的定义如下：

$$H(X\mid Y)=-\sum_{i,j} p(x_i,y_i)\log_b \frac{p(x_i,y_i)}{p(y_i)}$$

进而有信息熵的一元模型：

$$H(X)=-\sum_{x\in X} p(x)\log_2 p(x)$$

二元模型：

$$H(X\mid Y)=-\sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2 p(x\mid y)$$

三元模型：

$$H(X\mid Y,Z)=-\sum_{x\in X}\sum_{y\in Y}\sum_{z\in Z} p(x,y,z)\log_2 p(x\mid y,z)$$

M3: 信息熵的计算与结果输出

在以上结果的基础上对四类情境下中文的信息熵进行计算，并统计最高频率出现的一元，二元和三元词组进行输出。

3. 实验结果

3.1基于字分割的中文信息熵

基于字分割的中文一元信息熵计算结果为9.961比特，统计所得总字数为4452220，其中出现频率最高的字的列表如下：

字	出现次数
道	110847
中	48093
说	47016
子	44590

手	41497
见	35022

3.2基于词分割的中文信息熵

基于词分割的中文一元信息熵计算结果为13.623比特，统计所得总词数为2495677，平均词语长度为1.784字/词，其中出现频率最高的词语的列表如下：

词	出现次数
道	54183
说	19345
便	16729
见	16717
中	15037
说道	13156
听	12099
笑	9210
时	9132
想	8723
甚	8332
没	7540
出	6438
武功	6283
令狐	6225

3.3基于词分割的二元模型中文信息熵

基于词语分割的二元模型的中文信息熵计算结果为：5.599比特，小于一元模型的信息熵，出现频率最高的二元词组如下表所示：

二元词组	出现次数
('笑','道')	4655
('低声','道')	1403
('忽','听')	1326
('突然','间')	1228
('句','话')	1088
('令狐','道')	963

(’甚’,’麽’)	821
(’微笑’,’道’)	739
(’骂’,’道’)	693
(’件’,’事’)	654
(’便’,’道’)	642
(’张忌’,’道’)	632
(’微微’,’笑’)	628
(’忙’,’道’)	600
(’冷笑’,’道’)	590
(’少林’,’派’)	590
(’摇头’,’道’)	575
(’点头’,’道’)	526
(’陈家洛’,’道’)	514

3.3基于词分割的三元模型中文信息熵

基于词语分割的三元模型的中文信息熵计算结果为：0.875比特，小于三元模型的信息熵，出现频率最高的三元词组如下表所示：

三元词组	次数
(('韦小宝', '笑'), '道')	369
(('五岳', '剑'), '派')	253
(('令狐', '笑'), '道')	166
(('怦怦', '乱'), '跳')	159
(('句', '话'), '说')	159
(('出', '意料'), '外')	120
(('洪七公', '笑'), '道')	112
(('五岳', '派'), '掌门')	98
(('心中', '怦怦'), '乱')	92
(('话', '未'), '说完')	89
(('听', '脚步'), '声响')	86
(('呛', '啷'), '啷')	84
(('拍手', '笑'), '道')	83
(('恨', '恨'), '道')	80

(('转身', '便'), '走')	78
(('康熙', '笑'), '道')	78
(('忽', '听'), '远处')	77
(('嗤', '嗤'), '声响')	67
(('恒山', '派'), '掌门')	67
(('欧阳锋', '笑'), '道')	66
(('杨', '逍'), '道')	65
(('嗤', '嗤'), '嗤')	64

4.总结

相较于英文单字母而言，中文的单字信息熵更高，可以传递的信息也就越多，以词为单位的中文信息熵相较于单字而言更高，随着词组元数的增加，信息熵也逐渐降低，实验结果表明中文传递信息的效率要更高。除此之外，中文本文的压缩率在各种语言的比较中，总是最低的，也可以从侧面佐证这一结果。

参考文献

[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.

[2] Lin J. Divergence measures based on the Shannon entropy[J]. IEEE Transactions on Information theory, 1991, 37(1): 145-151.