统计自然语言处理笔记

陈鸿峥

2019.11*

目录

1 简介

自然语言处理(natrual language processing, NLP)的主要内容：机器翻译、信息检索、自动文摘、观点挖掘、问答系统、信息抽取、文档分类、文字编辑和自动校对、语音识别、文语转换、语音合成、说话人识别/认同/验证。

定义 1 (熵). 概率分布为$p(x) = P(X = x)$，则熵$H(X)$为

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

并约定$0\log 0 = 0$，单位为二进制位比特*(bit)*

定义 2 (联合熵). X, Y为离散型随机变量$X, Y \sim p(x, y)$，联合熵定义为

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

定义 3 (条件熵).

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

*Build 20191125

定义 4 (相对熵(KL距离)). 两个概率分布$p(x)$和$q(x)$的相对熵定义为

$$D(p||q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$$

定义 5 (交叉熵). 若随机变量$X \sim p(x)$，$q(x)$用于近似$p(x)$的概率分布，则X与模型q的交叉熵定义为

$$H(X, q) = H(X) + D(p||q) = -\sum_{x} p(x) \log q(x)$$

其常用来衡量估计模型与真实概率分布之间的差异

定义 6 (互信息). 如果$(X, Y) \sim p(x, y)$，则X, Y之间的互信息$I(X; Y)$定义为

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

互信息值越大，表示两个汉字之间的结合越紧密，越有可能成词。

基于上下文分类的消歧方法：假设多义词w所处的上下文语境为C，若w的多个词义记为s_i，则可通过计算$\arg\max p(s_i \mid C)$确定w的词义。

由贝叶斯公式，并运用独立性假设

$$p(s_i \mid C) = \frac{p(s_i) \times p(C \mid s_i)}{p(C)} = p(s_i) \prod_{v_k \in C} p(v_k \mid s_i) \Big/ p(C)$$

因此只需求（可转换为对数加法运算）

$$\hat{s}_i = \arg\max_{s_i} [p(s_i)] \prod_{v_k \in C} p(v_k \mid s_i)$$

由统计数据可得

$$p(v_k \mid s_i) = \frac{N(v_k, s_i)}{N(s_i)}$$

$$p(s_i) = \frac{N(s_i)}{N(w)}$$

其中$N(s_i)$为训练数据中词w用于语义s_i时的次数，而$N(v_k, s_i)$为w用于语义s_i时词v_k出现在w的上下文中的次数，$N(w)$为w在训练数据中出现的总次数。

2 中文分词

最大匹配法：有一个词典，设定最大词长，做字符串匹配

输入：S1= “计算语言学课程是两个课时"

输出：S2= " "

设定最大词长MaxLen = 5

W1= 计算语言学

……

词 典
…
计算语言学
课程
课时
…

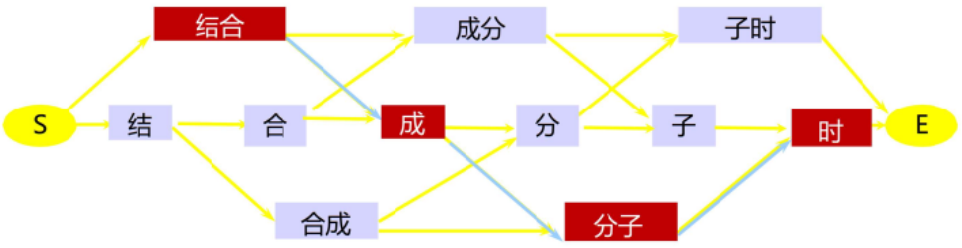
大规模真实语料中99%的词例（token）的长度都在5字以内 [1]

最优路径法：

- 选择一条词数最少的路径
- 半词法分词（加权）
- 最大概率法：在词图上选择词串概率最大的分词路径（动态规划）

$$\max(P(W_1 \mid S), P(W_2 \mid S))$$

- 看待汉语词语切分问题的新视角：词图上的最优路径求解问题



- 词图给出了一个字符串的全部切分可能性
- 分词任务：寻找一条起点S到终点E的最优路径

基于字序列标注的方法

- 分词可以看做是对字加“词位标记”的过程
- “人”的词位分类示例：

B	E	M	S
词首	词尾	词中	独立词
人们	古人	小人国	听 人 说

3 语言模型

3.1 统计语言模型

考虑语句的先验概率

$$p(s)=\prod_{i=1}^{m}p(w_i\mid w_1\cdots w_{i-1}),\ p(w_1\mid w_0)=p(w_1)$$

其中w_i可以是字、词、短语等，称为**统计基元**，通常用词代之。

为减少历史基元的个数，将$w_1w_2\cdots w_{i-1}$映射到等价类$S(w_1w_2\cdots w_{i-1})$，使等价类的数目远小于原来不同历史基元的数目，则有

$$p(w_i\mid w_1\cdots w_{i-1})=p(w_i\mid S(w_1\cdots w_{i-1}))$$

n元文法(n-gram)模型

- 当$n=1$时，出现在第i位上的基元w_i独立于历史，1元文法也被uni-gram或monogram
- $n=2$时，2-gram(bi-gram)称为1阶马尔可夫链
- $n=3$时，3-gram(tri-gram)称为2阶马尔可夫链，以此类推

实际操作加上句首<BOS>和句尾标记<EOS>。

❑ 举例：

给定句子：John read a book

增加标记：<BOS> John read a book <EOS>

Unigram: <BOS>, John, read, a, book, <EOS>

Bigram: (<BOS>John), (John read), (read a), (a book), (book <EOS>)

Trigram: (<BOS>John read), (John read a), (read a book), (a book <EOS>)

应用：

- 音字转换问题：给定拼音转为汉字串
- 汉语分词问题

对于n-gram，由最大似然估计求得

$$p(w_i \mid w_{i-n+1}^{i-1} = f(w_i \mid w_{i-n+1}^{i-1}) = \frac{c(w_{i-1+1})}{\sum_{w_i} c(w_{i-n+1}^{i})}$$

其中$\sum_{w_i} c(w_{i-n+1}^{i})$是历史串$w_{i-n+1}^{i-1}$在给定语料中出现的次数，即$c(w_{i-n+1}^{i-1})$。

为避免数据匮乏/稀疏导致的零概率问题，需要做数据平滑：调整最大似然估计的概率值，使零概率增值，使非零概率下调，消除零概率，改进模型的整体正确率。

- 加一法：

$$p(w_i \mid w_{i-1} = \frac{1 + c(w_{i-1}w_i)}{|V| + \sum_{w_i} c(w_{i-1}w_i)}$$

- 减值法/折扣法：将剩余概率量分配给未见概率

3.2 神经语言模型

词向量、RNN