DOI:10.16751/j.cnki.hbkj.2005.03.023
第25卷第3期 咸宁学院学报 Vol.25,No.3
2005年6月 Journal of Xianning College Jun.2005
文章编号:1006-5342(2005)03-0079-04

基于统计的自然语言处理模型*

戴文华,焦翠珍,徐 斌

(咸宁学院 计算机系,湖北 咸宁 437005)

摘 要:基于统计的自然语言处理模型采用统计方法进行自然语言建模.实际应用过程中可根据具体情况在多种模型中选择适当的模型.本文简要介绍了N-gram模型和最大熵模型,并给出了几种参数估计和数据平滑方法,为自然语言建模提供了一定的参考.

关键词:自然语言处理;N—gram模型;最大熵模型;数据平滑
中图分类号:TP309 文献标识码:A

0 引言

自然语言处理就是研究如何让计算机理解并生成人们日常生活中所使用的自然语言,从而建立起人与计算机之间的密切联系,使其能高效地进行信息传递和认知活动.自然语言处理时经常遇到的问题有分词、词性标注、语法分析、句法分析和语义分析等,这些自然语言中的问题都可以使用一些基于规则的语言分析方法进行处理.但对基于规则的系统来说,需要将专家的领域知识融入各种规则中,并且该方法随着规则库的增大效率明显下降.直到目前还没能出现一种比较完善的表示自然语言的规则系统.

随着计算机技术及Internet的迅速发展,大量联机语料库随之出现,计算机处理能力也大幅度提高,人们自然地想到利用统计方法对这些语料及新生成的语言进行分析处理.由于语料库具有信息量大、领域广、真实及实时性强等特点,我们可通过对语料库进行深层加工、统计和学习,获取大规模真实语料中的语言知识,这就是所谓的基于统计的自然语言处理.基于统计的自然语言处理模型使用分布函数来表示词、词组及句子等自然语言基本单位,它描述了自然语言的基于统计的生成和处理规则.

自然语言可以在人与人之间实现信息传输,信息传输的一端为发送者,另一端是接收者.信息源不断发送确定字符集 V 中的字符,字符的发送遵从一定的统计规则.接收者对接收的字符序列具有一定的先验知识.当发送的字符为独立非同分布的情况下,假定 w_i 表示 V 中的第 i 个字符,该字符被发送出去的概率是 $P(w_i)$,则字符流消息所携带的平均信息量为[1]:

$$H=-\frac{1}{|V|}[\sum_{i=1}^{|V|}P(w_i)\log P(w_i)] \quad (1)$$

其中 H 称为熵.如果发送的字符为非独立的,则前后发送的字符相互约束,因此一个字符流消息 $W=w_1w_2\cdots w_N$ 所携带的信息量为:

$$H=-\sum P(w_1,w_2,\cdots,w_N)\log(w_1,w_2,\cdots,w_N) \quad (2)$$

其中 $P(w_1,w_2,\cdots,w_N)$ 为字符流消息 $W=w_1w_2\cdots w_N$ 发送出去的概率.

如果把该模型用于描述自然语言交流,则字符表示语言中的单词或词组,字符集 V 表示这种语言的词汇表.

基于统计的自然语言处理模型有许多,模型中使用的参数估计与数据平滑方法也有多种.本文将主要介绍N-gram模型和最大熵模型,并给出几种参数估计和数据平滑处理方法.

1 N-gram模型

根据公式(2)可以看出,各词串的信息量由组成该词串的各个单词的联合概率决定.从语言的角度来看,$P(w_1,w_2,\cdots,w_N)$就是某种语言按次序产生出词串 $w_1w_2\cdots w_N$ 的概率,其大小反映了该

* 收稿日期:2004-12-08

词串在本语言中的使用情况，小的概率表明该词串在本语言中很少使用，大的概率表明该词串在本语言中经常出现.

我们将 $P(w_1, w_2, \cdots, w_N)$分解成条件概率的形式[2, 3]：

$$P(W)=P(w_1, w_2, \cdots, w_N)$$
$$=\prod_{n=1}^{N}P(w_n|w_1, w_2, \cdots, w_{n-1})$$
$$=P(w_1)\prod_{n=2}^{N}P(w_n|w_1, w_2, \cdots, w_{n-1}) \quad (3)$$

其中 $P(w_1)$为单词 w_1 的先验概率，可以通过统计大量语料库中该单词出现的频率获得，条件概率 $P(w_n|w_1, w_2, \cdots, w_{n-1})$表示已知前面 $n-1$ 个单词为$(w_1, w_2, \cdots, w_{n-1})$时，下一个单词为 w_n 的概率，计算这个概率就是进行统计预测，即已知前面若干个单词，预测下一个单词可能是什么.

为了进行有效预测，我们必须采用 Markov 假设，即假定单词 w_n 出现的概率只与其前面 $n-1$ 个单词有关. 满足该假设的自然语言模型称为 N－gram 模型，其中参数 n 称为模型的阶数，其取值决定了模型的精度和复杂性. n 的值越大，单词之间的依赖关系表述越准确，但这种准确性的提高需要增加参数估计量，参数估计问题也比低阶模型更为复杂，从而降低估计值的可靠性，对预测性能起到负面影响.

适当选取参数值 n，既可保证模型相对精确度，也能使模型不至过于复杂. 在实际应用中常使用 1 阶、2 阶和 3 阶等低阶模型，分别称为 Unigram 模型、Bigram 模型和 Trigram 模型，下面分别作简要介绍[4]：

(1)Unigram 模型

在 Unigram 模型中，我们认为单词与单词之间相互独立，建模时只需要考虑当前单词本身的概率，不必考虑单词所对应的上下文环境，此时公式(3)简化为：

$$P(W)=P(w_1, w_2, \cdots, w_N)=\prod_{n=1}^{N}P(w_n) \quad (4)$$

这是一种最为简单且易于实现的模型，但实际应用价值不高，很少被独立使用. Unigram 模型常与 Bigram 模型和 Trigram 模型配合使用，以实现平滑功能.

(2)Bigram 模型

在 Bigram 模型中，假定某个单词在句子中出现的概率条件依赖于其前一个单词，在这种假设下公式(3)可变为：

$$P(W)=P(w_1, w_2, \cdots, w_N)$$
$$=P(w_1)\prod_{n=2}^{N}P(w_n|w_{n-1}) \quad (5)$$

其中单词 w_n 的出现仅与 w_{n-1} 有关，由概率 $P(w_n|w_{n-1})$决定. 尽管在此模型中仅仅考虑了很少一部分上下文信息，但对下一个单词的预测具有较强的约束. 概率 $P(w_n|w_{n-1})$的估计有多种方法，通常采用相对频率法，通过统计大量训练样本中词对(w_{n-1}, w_n)的出现次数来估计条件概率.

(3)Trigram 模型

在 Trigram 模型中，假定某个单词在句子中出现的概率与其前面出现的两个单词有关，在此条件下公式(3)可简化为：

$$P(W)=P(w_1, w_2, \cdots, w_N)$$
$$=P(w_1)P(w_2|w_1)\prod_{n=3}^{N}P(w_n|w_{n-1}, w_{n-2}) \quad (6)$$

其中单词 w_n 的出现与 w_{n-1}和 w_{n-2}有关，由概率 $P(w_n|w_{n-1}, w_{n-2})$决定. 概率 $P(w_n|w_{n-1}, w_{n-2})$的估计方法与 Bigram 模型类似.

2 最大熵模型

最大熵模型又称指数模型，可以解决数据碎化问题，是自然语言处理研究的热点问题，它将统计语言问题看作一个求解受限概率分布问题，能够较好地包容各种约束信息.

假设给定训练语料 T，X 为上下文集，Y 为结果集，建立统计语言模型就是要在给定上下文 x 的条件下，估计出现单词 y 的条件概率 $P(y|x)$. 为了描述上下文 x 与单词 y 之间的关系，我们可引入二值特征函数 $f(x, y)$来表示特征(x, y)的有无. 特征函数 $f(x, y)$定义如下[5, 6]：

$$f(x, y)=\begin{cases}1 & (x, y)\in R\\ 0 & (x, y)\notin R\end{cases} \quad (7)$$

其中 R 为特征值集合.

利用给定的训练语料 T，可计算出特征函数 f 对于经验概率分布 $P(x, y)$的数学期望：

$$P(f)=\sum_{(x, y)}P(x, y)f(x, y) \quad (8)$$

根据此公式，能将任何先验知识表示为恰当的二值特征函数 $f(x, y)$的估计值，为了将有用的特征纳入到模型中，可以通过增加约束使模型满足相应特征的期望值来实现. 相对于概率分布 $P(y|x)$，特征函数 $f(x, y)$的期望表示为：

$$P(f)=\sum_{(x, y)}P(x, y)f(x, y) \quad (9)$$

由于 $P(x, y)=P(x)P(y|x)$，令 $P(x)=P(x)$，$P(x)$是 x 在训练样本中的观测概率，则限定所求模型的概率为在样本中观察到的事件的概

率.若 f 对模型有用，则令 $P(f)=\tilde{P}(f)$，由(8)和(9)两式可推得：

$$\sum_{(x,y)}\tilde{P}(x)P(y|x)f(x,y)=\sum_{(x,y)}\tilde{P}(x,y)f(x,y) \tag{10}$$

该式为约束等式，在该式约束下的模型 $P(y|x)$包含了特征函数 f.至此基于样本数据，对模型施加了单个特征约束.假设有 N 个特征函数 $f_i(i=1,2,\cdots,N)$表示有关特征，随机模型应该满足所有 N 个特征的约束.满足约束条件的模型很多，我们的目标是产生在约束集下具有最均匀分布的模型，即找到约束条件下具有最大熵的模型，约束条件下的熵可表示为：

$$H(P)=-\sum_{(x,y)}\tilde{P}(x)P(y|x)\log P(y|x) \tag{11}$$

其中 $0\leqslant H(P)\leqslant\log|y|$.

3　参数估计和数据平滑

在 N-gram 模型中，利用语料数据中的词汇同现的相对频率即可得到条件概率的极大似然估计：

$$P(w_i|w_1,w_2,\cdots,w_{n-1})=\frac{N(w_1,w_2,\cdots,w_{n-1},w_n)}{N(w_1,w_2,\cdots,w_{n-1})} \tag{12}$$

其中 $N(w_1,w_2,\cdots,w_{n-1},w_n)$是在训练语料中词串$(w_1,w_2,\cdots,w_{n-1},w_n)$出现的频次.从式中可以看到，没有在训练语料中出现的词串其估计量为零.因此在计算包含该子串的某个词串的概率时，即使该词串中包含其它具有较高概率的子串，整个词串的出现概率也为零.

对于一个确定的训练语料，即使规模相当大，也会有大量的词串没有同时出现，这就不可避免大量估计值为零的条件概率的出现，这就是所谓的数据稀疏问题.这种情况大大削弱了模型描述能力和后处理能力.数据平滑技术通过调整概率分布的取值，使低概率(包括零概率)被调高，高概率被调低，从而避免了零概率的出现，因此能解决数据稀疏问题.与此同时数据平滑使模型参数概率分布更加均匀，概率的计算更加精确.

数据平滑技术主要有 Katz 平滑、Kneser-Ney 平滑、Jelinek-Mercer 平滑、Church-Gale 平滑、插值平滑和加法平滑等.下面简单介绍大多数数据平滑的核心技术 Good-Turing 估计和实际应用中使用较广泛的 Katz 平滑技术.

3.1　Good-Turing 估计

Good-Turing 估计是许多数据平滑技术的核心.Good-Turing 估计用于 N-gram 模型时，对于模型中出现 r 次的事件，我们应假设它的出现次数为：

$$r^*=(r+1)\frac{n_{r+1}}{n_r} \tag{13}$$

其中 n_r 代表训练集中实际出现 r 次的事件个数.对于模型中出现次数为 r 次的事件，其条件概率则可标准化为：

$$P_{GT}(\alpha)=r^*/N \tag{14}$$

其中 N 为模型中所有 n 元对的总数，α 表示训练集中实际出现 r 次的事件.从公式(13)可知当 n_r 为零时，不能使用 Good-Turing 估计，必须对 n_r 进行平滑.

3.2　Katz 平滑

Katz 平滑[7,8]是实际应用中使用非常广泛的一种数据平滑技术，该技术在 Good-Turing 估计的基础上，采用低阶 N-gram 模型对高阶 N-gram 模型的插值方法扩展了 Good-Turing 的平滑方法.

与插值平滑不同，Katz 平滑将每个 N-gram 模型表示为 M-gram 模型的非线性组合，n、m 分别为 N-gram 模型和 M-gram 模型的阶数，且 $m=1,2,\cdots,n$.

对每个 M-gram 模型，由回退概率 β_m 表示由 M-gram 模型回退到(M-1)-gram 模型的概率，也可理解为在学习语料集中，给定上下文(M-1)-gram，M-gram 不出现的概率，因此存在递推公式：

$$\begin{aligned}&P(w_n|w_1^{n-1})=P_{GT}(w_n|w_1^{n-1})+\beta_m P(w_n|w_2^{n-1})\\&P(w_n|w_2^{n-1})=P_{GT}(w_n|w_2^{n-1})+\beta_{m-1}P(w_n|w_3^{n-1})\\&\cdots\cdots\end{aligned} \tag{15}$$

其中 w_i^j 表示词串$(w_i,w_{i+1},\cdots,w_{j-1},w_j)$，$P_{GT}$表示采用 Good-Turing 估计确定的概率值，给定训练语料，它可以被预先计算出来.因此确定回退概率是该方法的主要任务，包含两种情况：

(1)如果一个 M-gram 出现在训练语料集中，即 $c(w_1^m)>0$，$c(w_i^j)$表示语料集中词串 w_i^j 的出现次数，则回退概率应为 $\beta_m=0$；

(2)如果一个 M-gram 没有出现在训练语料中，即 $c(w_1^m)=0$，则按照下式计算回退概率：

$$\beta_m=\sum_{w_1^m:c(w_1^m)>0}P_{GT}(w_m|w_1^{m-1})\Big/\Big[1-\sum_{w_1^m:c(w_1^m)>0}P_{GT}(w_m|w_2^{m-1})\Big] \tag{16}$$

4　模型的评价

模型的性能可以通过其在语言处理系统的最

终表现来评估,但是一般语言处理系统通常涉及多种处理任务,各种任务相互影响,具有很大的复杂性.评价模型性能最直观的指标是在测试语料集上计算的信息熵.熵的值越大,则学习模型与真实模型的差别也越大,表明模型的效果越差[1].

由前面的介绍可知,一段文本的熵如公式(1)所示,其中 $P(w_i)$ 为 w_i 的真实分布,但 $P(w_i)$ 是未知的,只能用语言模型的估计值代替,因此常采用如下公式代表模型的熵:

$$H=-\frac{1}{|V|}\sum_{i=1}^{|V|}\log P(w_i\mid w_{i-1}) \tag{17}$$

另一个常用的评价指标是模型的复杂度 PP,定义如下:

$$PP=2^H=2^{-\frac{1}{|V|}\sum_{i=1}^{|V|}\log P(w_i\mid w_{i-1})} \tag{18}$$

复杂度是对模型选择下一个单词的范围大小的度量,复杂度越大,识别器的难度就越大,模型越复杂.

从复杂度的定义我们可以知道,语言模型的复杂度依赖于用于评估它的语言数据.在训练语料上具有小的复杂度只表明语言模型对训练语料具有较好的逼近能力,但并不能保证在测试集上一定有较小值.如果在训练集上复杂度很小,但在测试集上复杂度较大,则说明语言模型的推广能力较差,称该现象为语言模型被过训练.一般情况下,越复杂的语言模型,由于逼近能力较强,比较容易出现过训练的问题.

反之,由于复杂度依赖于语言数据,因此也可利用同一个语言模型在不同的测试集上的复杂度来评估语言数据的复杂度.

对两种语言模型进行比较时,为了保证不同语言模型比较的客观性,必须要求两个模型在相同的训练集上训练,并在相同的测试集上测试.

5 结束语

通过对目前自然语言处理模型的分析,我们发现仅仅使用基于规则的语言模型进行自然语言处理根本不可能得到比较理想的效果.同样,如果单独采用基于统计的语言模型,也不能达到较为高效的目的.我们必须将两者合理的结合起来,并利用一些先进的优化技术,才能发挥两类模型各自的优势,提高自然语言处理的能力.

参考文献:

[1] 王小捷,常小宝.自然语言处理技术基础[M].北京:北京邮电大学出版社,2002.83~91.

[2] 盛骤,谢式千等.概率论与数理统计[M].北京:高等教育出版社,2001.18~31.

[3] Rob Callan. Artificial Intelligence[M]. 北京:电子工业出版社,2004.297~327.

[4] 张健,李素建等. N-gram 统计模型在机器翻译系统中的应用[J].计算机工程与应用,2002,(8).

[5] 徐延勇,周献中等.基于最大熵模型的汉语句子分析[J].电子学报,2003,(11):31.

[6] 李素建,刘群等.语言信息处理技术中的最大熵模型方法[J].计算机科学,2002,(7):29.

[7] S. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer[J]. IEEE Transactions on Acoustics, Speech and Signal Processing, 1987,(3):35.

[8] 陶志荣. N-gram 语言模型的 Katz 平滑技术[J].电子计算机,2002,(2).

The Natural Language Processing Models Based on Statistics

DAI Wen-hua, JIAO Cui-zhen, XU Bin

(Department of Computer, Xianning College, Xianning 437005, China)

Abstract: The natural language processing models based on statistics adopts the statistical method to setting up the natural language models. In actual applications, we can select the proper models from a lot of models according to the specific situation. In this paper, we introduce N-gram Model and Maximum Entropy Model. At the same time, we give a few methods of parameters estimate and data smoothing, which can provide some reference for setting up the natural language models.

Key words: Natural language processing; N-gram model; Maximum entropy model; Data smoothing