N-gram 语言模型中的插值平滑技术研究

徐 望,王炳锡

(信息工程大学 信息工程学院,河南 郑州 450002)

摘要:本文研究了 N 元文法(N-gram)统计语言模型中的 4 种插值平滑算法,在中文语言模型中进行了应用,从语言模型复杂度的角度比较了该 4 种方法解决零概率问题的效率。

关键词:N-gram;复杂度;插值平滑

中图分类号:TN912.3 **文献标识码**:A **文章编号**:1671-0673(2002)04-0013-03

1 引言

研究表明,人类在进行自然语音识别时,不仅仅用人耳对语声进行捕捉和辨认,同时还利用了许多非声学的信息,诸如句法、语义、语境等方面的知识来进一步对话语做出识别和理解。显然建立适当的语言模型,与声学模型相结合,有利于提高现有语音识别系统的识别能力和性能。目前语言模型可分为两种:统计性语言模型和确定性语言模型。统计语言模型(SLM)没有利用建模对象语言本身的诸多特性如语法、结构、隐含意义等特性,它是将语言作为字符串进行统计分析,通过预测语言单元诸如字、句以及整个文本的出现的概率及相互关联的概率,来推测自然语言的规律性。本文研究和比较了 N 元文法(N-gram)统计语言模型的 4 种插值平滑算法,计算出了各种方法下中文语言模型的复杂度。

2 N-gram 语言模型

若把人类的自然语言看成为信息源,根据信息论,输出词串 $W = w_1, w_2, \cdots, w_n$ 所含的信息量为

$$H(W) = -\sum P(w_1, w_2, \cdots, w_n)\log P(w_1, w_2, \cdots, w_n)。 \quad (1)$$

如果信源是各态遍历的,则

$$H(W) = -\lim_{n\to\infty}\frac{1}{n}\log_2 P(w_1, w_2, \cdots, w_n) \approx -\frac{1}{n}\log_2 P(w_1, w_2, \cdots, w_n), \quad (2)$$

$$P(w_1, w_2, \cdots, w_n) = P(w_1)\,P(w_2|w_1)\,P(w_3|w_1, w_2)\cdots P(w_n|w_1, w_2, \cdots, w_{n-1}) = \prod_{i=1}^{n} P(w_i|w_{i-n+1}, \cdots, w_{i-1}) \quad (3)$$

$P(w_i|w_{i-n+1}, \cdots, w_{i-1})$ 表示在给定历史信息 $w_{i-n+1}, \cdots, w_{i-1}$ 的条件下,选取词 w_i 的概率。这种利用前 $N-1$ 个字来推测当前字的 Markov 模型就称为 N-gram(N 元文法)模型。在实际应用中,只考虑 0 个、1 个或 2 个历史信息,形成了 uigram 模型 $P(w_i)$,bigram 模型 $P(w_i|w_{i-1})$ 和 trigram 模型 $P(w_i|w_{i-1}, \cdots, w_{i-2})$。

一般用复杂度(perplexity)值的大小来评估统计语言模型的性能。复杂度 PP 定义为

$$PP = 2^{H(W)} = \prod_{i=1}^{n}[P(w_i|w_{i-n+1}, \cdots, w_{i-1})]^{-\frac{1}{n}}。 \quad (4)$$

PP 值反映了信源的熵的大小,表示对信源的不可知程度。直观上,复杂度可以理解为在给定的语言模型中,某个词后面可能接的词的平均数。显然复杂度越小,语言模型对上下文的约束能力就越强,模型就越好。因而语言模型的复杂度是评价语言模型好坏的一个准则。

3 N-gram 模型的平滑

在 N-gram 模型中,假设词 w 在语料库中出现的概率 $P(w)$ 符合二项分布规律,则当语料库容量 N 足够大时,我们可以用词出现的相对频率来近似概率,这就是 MLE 估计方法。

此时条件概率 $P(w_i|w_{i-n+1}, \cdots, w_{i-1})$ 可表示

收稿日期:2002-03-07

作者简介:徐望(1974-),男,江西南昌人,信息工程大学博士研究生,主要研究方向为语音信号处理。

为

$$P(w_i, w_{i-n+1}, \cdots, w_{i-1}) = \frac{C(w_i, w_{i-1}, \cdots, w_{i-n+1})}{C(w_{i-1}, \cdots, w_{i-n+1})}。 \quad (5)$$

其中，$C(w_i, w_{i-1}, \cdots, w_{i-n+1})$ 表示词串 $(w_i, w_{i-1}, \cdots, w_{i-n+1})$ 在训练语句中出现的次数。

(5) 式的概率估值方法简单实用，并可得到合理的估计。但当数据不能很好地适应模型时，这种估计方法也可能出问题。如实词 (content word) 倾向于“突发性”的出现，在语料库中的分布不能很好地符合二项分布规律；由于某些文章风格因素的作用，功能词 (function word) 可能也会偏离二项分布[1]。另外，由于统计数据的稀疏性，必然会出现一些语料库中不出现的情况，对此，MLE 方法将给出零概率的估计值，在语音识别中，如果 $P(w)$ 为 0，显然词 $w_1 w_2, \cdots, w_{n-1}$ 不能被观测到，因此将 (5) 式进行调整，使词序列的概率分布合理化，就称为平滑。

平滑的基本思想是将方程 (5) 中可见词 w_i 的概率值折扣，将该折扣值重新分布给不可见词。因此，平滑方法由概率值折扣的策略和折扣值的分布方法所决定。

假设某观测事件 k 的概率为

$$P(k) = \frac{r^*}{N} = \frac{r \cdot d_r}{N}, \quad (6)$$

r 为事件 k 的实际出现次数，r^* 为被折扣的出现次数，N 为总的事件出现次数，d_r 为折扣系数。

概率值折扣的方法一般有 3 种：Good Turing 折扣 (7)、绝对值折扣 (8) 及线性折扣 (9)。

$$d_r = \min\left\{1, \frac{(r+1) \cdot n_{r+1}}{r \cdot n_r}\right\}, \quad (7)$$

$$d_r = \frac{r - d}{r}, \qquad d = \frac{n_1}{n_2 + 2n_2}, \quad (8)$$

$$d_r = 1 - \alpha。 \quad (9)$$

n_r 为观测次数恰好为 r 次的事件次数，α 为常数值。

除此之外，加值平滑也是较早采用的一种概率值折扣方法，表达式为

$$P_{add}(w_i | w_{i-n+1}, \cdots, w_{i-1}) = \frac{C(w_i, w_{i-1}, \cdots, w_{i-n+1}) + \delta}{C(w_{i-1}, \cdots, w_{i-n+1}) + \delta |V|}。 \quad (10)$$

$|V|$ 为训练语句的词汇总数，δ 一般取值为 1。

从折扣值的分布策略来讲主要分两类[2]：back-off 模型和 interpolated (插值) 模型。

back-off 模型通用表达式为

$$P_{smooth}(w_i | w_{i-n+1}, \cdots, w_{i-1}) = \begin{cases} \alpha(w_i | w_{i-n+1}, \cdots, w_{i-1}) & \text{if } C(w_{i-n+1}, \cdots, w_i) > 0 \\ \gamma(w_{i-n+1}, \cdots, w_{i-1}) P_{smooth}(w_i | (w_{i-n+2}, \cdots, w_{i-1}) & \text{if } C(w_{i-n+1}, \cdots, w_i) = 0 \end{cases} \quad (11)$$

即如果 n-gram 有非零记数，则采取分布 $\alpha(w_i | w_{i-n+1}, \cdots, w_{i-1})$，否则回溯到更低阶的分布 $P_{smooth}(w_i | w_{i-n+2}, \cdots, w_{i-1})$，这里 $\gamma(w_{i-n+1}, \cdots, w_{i-1})$ 为保证条件概率求和为一的尺度系数。

插值模型通用表达式为

$$P_{smooth}(w_i | w_{i-n+1}, \cdots, w_{i-1}) = \lambda P_{ML}(w_i | w_{i-n+1}, \cdots, w_{i-1}) + (1 - \lambda) P_{smooth}(w_i | w_{i-n+2}, \cdots, w_{i-1}), \quad (12)$$

λ 为插值系数。

插值模型与 back-off 模型的主要区别在于：当求 n-gram 的概率值出现零记数 $(w_i, w_{i-1}, \cdots, w_{i-n+1})$ 时，插值模型利用了更低阶的概率分布信息，而 back-off 模型没有。文献 [3] 指出插值模型比 back-off 模型性能要好。插值模型又分为线性和非线性插值。

3.1 线性插值模型

考虑 bigram 模型，线性插值模型公式如下[4]：

$$P_{smooth}(w_i | w_{i-1}) = (1 - \lambda) \frac{N(w_i, w_{i-1})}{N(w_{i-1})} + \lambda P(w_i)。 \quad (13)$$

求语言模型熵最小化，可得到迭代公式

$$\lambda^{k+1} = \frac{n_1}{N} + \sum_{w_i w_{i-1}} \frac{N(w_i, w_{i-1})}{N} \frac{\lambda^{(k)} P(w_i)}{(1 - \lambda^{(k)}) \frac{N(w_i, w_{i-1})}{N} + \lambda^{(k)} P(w_i)}, \quad (14)$$

$\lambda^{(0)} = 0$。

λ 起到适当减少 uigram 概率值而增加 bigram 的观测次数的作用，文献 [5] 提出可表示为

$$\lambda = e^{-N(w_i, w_{i-1})}。 \quad (15)$$

即随着 bigram 观测次数的增多，加权系数指数减少。

指数型的线性平滑公式为

$$P_{smooth}(w_i | w_{i-1}) = (1 - e^{-N(w_i, w_{i-1})}) \frac{N(w_i, w_{i-1})}{N(w_{i-1})} + e^{-N(w_i, w_{i-1})} P(w_i)。 \quad (16)$$

3.2 非线性插值模型

非线性插值模型又称为 Kneser-Ney 插值模型[6]，其表达式为

$$P(w_i | w_{i-1}) = \frac{\max\{N(w_i, w_{i-1}) - d, 0\}}{N(w_{i-1})} + \frac{n_{>0}(\cdot, w_{i-1}) \cdot d}{N(w_{i-1})} \cdot \beta(w_i | w_{i-1})。 \quad (17)$$

其中 $n_{>0}(\cdot, w_{i-1}) = \sum_{w_i: N(w_i, w_{i-1})>0} 1$, $\beta(w_i | w_{i-1}) = \frac{n_{>0}(w_i, \cdot)}{\sum_{w_i} n_{>0}(w_i, \cdot)}$。

$\beta(w_i | w_{i-1})$ 又称为边际(marginal)分布。

Kneser 和 Ney 在文献[7]中提到用 $n_1(w_i, \cdot) = \sum_{w_{i-1}: N(w_i, w_{i-1})=1} 1$ 代替 $n_{>0}(w_i, \cdot)$,即通过只发生一次的事件的概率来估计不发生事件的概率,这时求出的 $\beta(w_i | w_{i-1})$ 称为单一(singleton)分布。

4　实验结果

4.1　实验数据

实验用训练语料采用 1998 年 1 月份的"人民日报",字典有 4400 多个字(去除标点符号),158904 句,一共 120 多万字次。测试语料摘录自 1995 年"人民日报",共 1862 句,1900 多个字,共 2 万多字次。

4.2　平滑算法实现

本文实现了第 3 节提到的 5 种平滑算法,并用计算了该它们求出的语言模型复杂度。具体实现时由(14)式得 $\lambda = 0.24$,另外(17)式中的 d 值采用(8)式近似。测试结果见表 1。

表 1　复杂度比较表

复杂度	Bigram 模型				
	加值平滑	线性插值(迭代型)	线性插值(指数型)	非线性插值(marginal)	非线性插值(singleton)
PP	492.5	163.9	157.5	156.0	154.9

结果表明:采用插值平滑后,语言模型复杂度较加值平滑大为降低,非线性插值得到的复杂度比线性的要低。

5　结论

统计语言模型是从大量的实际语言材料以自组织方式获取其中的语言结构信息的,因此采用不同领域类型的语料,以及保证语料具有足够的规模,是建立统计语言模型的重要基础。但由于语言材料的有限性,数据稀疏带来的零概率问题是无法避免的,因此平滑技术在统计语言模型中起着十分重要的作用。本文研究了 4 种平滑算法,依据 bigram 模型计算了它们的复杂度。实验结果表明,采用插值平滑能有效降低语言模型复杂度。由于语料容量限制,本文没有对 trigram 模型进行计算,但方法完全一样。

参考文献:

[1] 周强. 基于语料库和面向统计学的自然语言处理技术介绍[J]. 计算机科学,1995,22(4):36 - 40.

[2] Ney H, Essen U and Kneser R. On Structuring probabilistic dependencies in stochastic language modelling[J]. Computer Speech and Language, 1994, 8(1): 1 - 38.

[3] S Chen and J Goodman. An empirical study of smoothing techniques for language modeling[R]. Harvard University, 1998.

[4] F Jelinek, Mercer L R, Roukos S, Principles of lexical language modeling for speech recognition[A]. Advances in Speech Signal Processing[C]. New York: Marcel Dekker, 1991: 651 - 700.

[5] Wei Xu, Alex Rudnicky. Can Artificial Neural Networks Learn Language Models[A]. 6th International Conference on Spoken Language Processing (ICSLP2000)[C]. Beijing, 2000. M1 - 13.

[6] Woosung Kim, Sanjeev Khudanpur. Smoothing Issues in the Structured Language Model[A]. Proc. 7th European Conf on Speech Communication and Technology[C]. 2001, 1: 717 - 720.

[7] Kneser R and Ney H. Improved backing-off for m-gram language modeling[A]. Proc ICASSP'95[C]. 1995. 181 - 184.

Study of Interpolation Smoothing Techniques in N-Gram Language Modeling

XU Wang, WANG Bing-xi

(Institute of Information Engineering, Information Engineering University, Zhengzhou 450002, China)

Abstract: This paper describes interpolation smoothing techniques for statistical language models and their application to language model in China. From the point of view of perplexity of language model, this paper discusses the efficiency of four kinds of methods on probabilities of sparse date.

Key words: N-gram; perplexity; interpolation smoothing