一种基于 N 元语法分布的语言模型自适应方法

尹继豪 姜志国 樊孝忠

(北京航空航天大学 宇航学院，北京 100191) (北京理工大学 计算机科学技术学院，北京 100081)

摘 要：N 元语法分布能恰当地描述语料库的特性，为了有效利用普通领域训练数据，提出一种基于 N 元语法分布的语言模型自适应方法. 该方法定义一个小的领域内的高质量种子集和一个大的普通领域的质量不稳定的训练集，将训练集的 N 元语法分布自适应到和种子集的 N 元语法分布相似，以更好地进行特定领域单词识别. 实验结果表明，基于 N 元语法分布的语言模型自适应方法可以使单词困惑度和词错误率分别比传统的简单插值法降低 11.1%和 6.9%.

关 键 词：N 元语法分布；种子集；训练集；自适应

中图分类号：TP 391

文献标识码：A **文 章 编 号**：1001-5965(2008)11-1276-04

Statistical language model adaptation based on *N*-gram distribution

Yin Jihao Jiang Zhiguo

(School of Astronautics, Beijing University of Aeronautics and Astronautics, Beijing 100191, China)

Fan Xiaozhong

(School of Computer Science and Technology, Beijing Institute of Technology, Beijing, 100081, China)

Abstract: N-gram distribution can represent the characters of corpus correctly. So an approach was proposed for statistical language modeling adaptation, which is based on N-gram distribution. Given a large set of out-of-task training data, called training set, and a small set of task-special training data, called seed set, one statistical language modeling was adapted towards the special domain by adjusting the N-gram distribution in the training set to that in the seed set. The experiment results show prominent improvements over conventional methods. Compared with the simple interpolation method, the perplexity and word error rate decreases 11.1% and 6.9% respectively.

Key words: N-gram distribution; seed set; training set; adaptation

统计语言模型(成功地应用于很多领域，如语音识别[1]和信息抽取[2]，尤其是 N 元语法模型在这些领域中使用相当广泛，并取得了良好的效果. 但是统计语言模型的建立基于以下前提：①具备大量的训练数据；②训练数据和测试数据来源于相同的领域[3]. 训练数据训练得到的语言模型的质量高低主要取决于训练数据的大小和质量，也就是说，如果不具备大量的训练数据，或者训练数据与测试数据不匹配，就不能得到精确的统计语言模型.

传统的统计语言模型自适应技术是这样做的：将一个通用的、训练充分的模型和一个特定领域的、训练不充分的模型通过某种方式组合成一个新的模型. 组合的方法一般有两种：插值法[4]和最大熵法[5]. 插值法比较常用，它最大的优点是易于实现，计算效率高，其缺点是难以保证模型的完整性，并且很难达到最优的插值结果. 最大熵法的优点是能够达到较好的插值结果，但其明显的缺点是计算量大、计算效率低.

N 元语法分布能恰当地描述训练数据的特

收稿日期：2007-11-29

基金项目：教育部博士点基金资助项目(20050007023)

作者简介：尹继豪(1980—)，男，河南叶县人，讲师，yjh@buaa.edu.cn.

性,因此,本文提出基于 N 元语法分布的语言模型自适应方法.该方法定义一个小的领域内的高质量训练数据集(种子集)和一个大的领域外的质量不稳定的训练数据集(训练集),其基本思路是将训练集的 N 元语法分布自适应到和种子集的 N 元语法分布相似.

1　常用的语言模型自适应算法

在以下的公式中,$P(x)$表示事件 x 的概率,$C(x)$表示事件 x 的出现次数.下标 c 表示通用(common)模型,对应从大规模的语料中训练出来的模型;下标 a 表示自适应(adaptation)模型,对应从小规模特定领域语料中训练出来的模型;无下标表示自适应后的最终模型.w 表示当前单元,h 表示历史单元.本文使用的通用语言模型和自适应语言模型都是二元语法模型 $p(w_n|w_{n-1})$,因为由二元语法很容易推广到 N 元语法.

1.1　简单插值算法

通常的插值算法是概率意义上的插值,即:

$$P(w\mid h)=\lambda P_{\mathrm{c}}(w\mid h)+(1-\lambda)P_{\mathrm{a}}(w\mid h) \tag{1}$$

为叙述方便,采用的是计数意义上的插值,即:

$$C(h,w)=C_{\mathrm{c}}(h,w)+\alpha C_{\mathrm{a}}(h,w) \tag{2}$$

事实上,它们的本质是相同的,这是因为

$$P(w\mid h)=\frac{C(h,w)}{C(h)}=\frac{C_{\mathrm{c}}(h,w)+\alpha C_{\mathrm{a}}(h,w)}{C_{\mathrm{c}}(h)+\alpha C_{\mathrm{a}}(h)}=\frac{\dfrac{C_{\mathrm{c}}(h,w)}{C_{\mathrm{c}}(h)}+\alpha\dfrac{C_{\mathrm{a}}(h,w)}{C_{\mathrm{a}}(h)}\times\dfrac{C_{\mathrm{a}}(h)}{C_{\mathrm{c}}(h)}}{1+\alpha\dfrac{C_{\mathrm{a}}(h)}{C_{\mathrm{c}}(h)}} \tag{3}$$

如果令

$$\frac{C_{\mathrm{a}}(h)}{C_{\mathrm{c}}(h)}=A(h)$$

则式(3)可以化为

$$P(w\mid h)=\frac{1}{1+\alpha\times A(h)}P_{\mathrm{c}}(w\mid h)+\frac{\alpha\times A(h)}{1+\alpha\times A(h)}P_{\mathrm{a}}(w\mid h) \tag{4}$$

显然,令

$$\lambda=\frac{1}{1+\alpha\times A(h)}$$

则式(1)和式(4)等价,但式(4)却可以针对不同的 h,对 λ 进行更精细的调整.

需要说明的是,α 是预先设定的自适应插值参数,α 越大,自适应模型起的作用越大.$A(h)$是和两个语料库相关的参数,它是 h 在自适应语料中出现次数与在通用语料中出现次数的比值,反映了 h 在两种语料中的重要性的不同,$A(h)$越大,则自适应模型起的作用越大.

1.2　考虑 Katz 平滑的插值算法

数据稀疏问题是 N 元语法所特有的,因为它们不能使用长距离的上下文,总是倾向于过低地估计那些在训练语料库中不是彼此邻近出现的符号串的概率.不过,可以使用一些技术给那些"零概率的 N 元语法"指派非零概率.这种给某些零概率和低概率的 N 元语法重新赋值并给它们指派非零的工作称为"平滑"[6].在此,采用已有的、著名的 Katz 平滑方法[7]:

$$P_{\mathrm{Katz}}(w_i\mid w_{i-1})=\begin{cases}C(w_{i-1},w_i)/C(w_{i-1}) & \text{if } C(w_{i-1},w_i)>r_{\mathrm{T}}\\ d_r C(w_{i-1},w_i)/C(w_{i-1}) & \text{if } 0<C(w_{i-1},w_i)\leqslant r_{\mathrm{T}}\\ \alpha(w_{i-1})P_{\mathrm{Katz}}(w_i) & \text{if } C(w_{i-1},w_i)=0\end{cases} \tag{5}$$

式中,r_{T} 是用于折扣的阈值,$\alpha(w_{i-1})$和 d_{r} 是平滑参数.

采用了 Katz 平滑方法,实际上出现次数小于 r_{T} 的单元会用于折扣.其意义在于表明出现次数较少的单元在统计上的可信度较低,所以把它们折扣给出现次数为 0 的单元;而出现次数较多的单元在统计上可信度较高,因此它们的统计结果保持不变.

由于自适应语料相对通用语料较少,所以要想让自适应模型发挥出理想的作用,式(2)和式(4)中的经验参数 α 的取值一般都大于 1.而对于出现次数较少(小于 r_{T})的单元,这样的 α 就存在一定的问题了.因为,这些可信度低的单元,如果被加一个较大的权 α,则概率有可能被人为地、过度地增大,而对于可信度低的单元,一般不应该给予较大概率.如果其概率过大,将导致其它单元的概率值不能领先,造成多个错误;如果让其概率较小,最坏的情况只是导致该单元概率该领先的时候没有领先,只影响到自身.因此,从直观上讲,对于此类可信度较低的单元,其概率宁可过小,不可过大[8].

所以,应对式(2)做相应的修改:

$$C(h,m)=\begin{cases}C_{\mathrm{c}}(h,m)+\alpha C_{\mathrm{a}}(h,m)\\ \quad\text{if } C_{\mathrm{c}}(h,m)+\beta C_{\mathrm{a}}(h,m)>r_{\mathrm{T}}\\ C_{\mathrm{c}}(h,m)+\beta C_{\mathrm{a}}(h,m)\\ \quad\text{if } C_{\mathrm{c}}(h,m)+\beta C_{\mathrm{a}}(h,m)\leqslant r_{\mathrm{T}}\end{cases} \tag{6}$$

在式(6)中,β 要远小于 α.

2　N 元语法分布的语言模型自适应

2.1 语言模型自适应算法

词条分布模型在信息抽取领域被广泛地使用[9]. 定义 $p_i(k)$ 为单词 w_i 在一篇文档中出现 k 次的概率,在二元分布模型中,给出一个概率模型来计算单词对 (w_{i-1},w_i) 在一篇新文档中出现的概率 $P(w_{i-1},w_i)$.

描述词条分布效果比较好的模型是 Katz 提出的 K 重混合模型[10],它是一个简单的分布模型,因此使用 K 重混合模型建立二元分布模型. 在 K 重混合模型中,单词 w_i 在一篇文档中出现 k 次的概率为

$$P_i(k)=(1-\eta)\delta_{k,0}+\frac{\eta}{\theta+1}\left(\frac{\theta}{\theta+1}\right)^k \tag{7}$$

式中,若 $k=0$,则 $\delta_{k,0}=0$,否则 $\delta_{k,0}=0$. η 和 θ 可以由观测均值 λ 和观测倒排文档频度 IDF(Inverse Document Frequency)计算:

$$\lambda=\frac{N_{\mathrm{cf}}}{N} \tag{8}$$

$$f_{\mathrm{IDF}}=\log_2\frac{N}{N_{\mathrm{df}}} \tag{9}$$

$$\theta=\lambda\times 2^{f_{\mathrm{IDF}}}-1=\frac{N_{\mathrm{cf}}-N_{\mathrm{df}}}{N_{\mathrm{df}}} \tag{10}$$

$$\eta=\frac{\lambda}{\theta} \tag{11}$$

式中,N_{cf}(collection frequency,收集频度)是所有文档中单词 w_i 出现的总次数;N_{df}(document frequency,文档频度)是出现单词 w_i 的文档数;N 是总的文档数目. 参数 θ 是每个文档中出现次数大于 1 的词条数的均值,$\frac{\theta}{\theta+1}=\frac{N_{\mathrm{cf}}-N_{\mathrm{df}}}{N_{\mathrm{cf}}}$ 实际上决定了 $\frac{p_i(k+1)}{p_i(k)}$ 的大小. 例如,如果这个比值为 0.1,则表示出现该词条一次的文档数是出现两次的文档数的 10 倍. 如果两词条有相同的 θ,表示它们有相同的收集频度和文档频度之间的比率. 参数 η 表示词条的绝对出现频度信息.

二元语法分布模型是 K 重混合模型的一个直接应用,通过式(12)计算单词对 (w_{i-1},w_i) 在一篇文档中出现的频率:

$$P(w_{i-1},w_i)=1-\sum_{k=1}^{K}p_i(k) \tag{12}$$

式中,K 的取值范围为 1～2. $p_i(l)$ 由式(7)得到,式中 η 和 θ 由式(8)～式(11)求得.

N 元语法分布很好地表征了语料库的特征,能恰当地描述特定领域训练数据的特性. 因此,文中有效地利用普通领域训练数据,提出一种基于 N 元语法分布的语言模型自适应方法.

自适应方法不是合并训练集和种子集的二元语法模型,而是直接合并二元计数 $C(w_1,w_2)$ 和自适应权重 $W(w_1,w_2)$,合并式子如下:

$$C(w_1,w_2)=\sum_i W_i(w_1,w_2)\times C_i(w_1,w_2) \tag{13}$$

式中,$W_i(w_1,w_2)$ 是第 i 个训练集的自适应权重,通过式(14)计算:

$$W_i(w_i,w_2)=\log_2\left(\frac{p(w_1,w_2)}{p_i(w_1,w_2)}\right)^{\alpha} \tag{14}$$

式中,α 是自适应系数;$p(w_1,w_2)$ 是二元语法 (w_1,w_2) 在种子集中的出现频率;$p_i(w_1,w_2)$ 是二元语法 (w_1,w_2) 在第 i 个训练集中的出现频率,可以由式(15)计算:

$$p_i(w_1,w_2)=\frac{C_i(w_1,w_2)}{\sum_{w_1,w_2}C_i(w_1,w_2)} \tag{15}$$

基于 N 元语法分布的模型自适应方法的关键在于自适应系数 α 的选取.

2.2 自适应系数选取

使用交叉验证方法确定自适应系数,随机把特定领域训练数据分割成同样大小的 5 个部分,每一次,选择其中一部分作为测试集,其余 4 部分合并后作为种子集. 然后使 α 从 0～10,每隔 1 取一次值,并且通过式(13)～式(15)计算二元语法计数. 最终估计一系列二元语法模型.

图 1 给出了当 α 取不同值时,种子集困惑度和测试集困惑度大小的变化曲线. 随着 α 的增长,种子集困惑度 p_{plx} 降低,测试集困惑度先降低然后急剧升高. 因此,选择当测试集困惑度将要升高时的 α 值,称这个点的 α 值为 α 关键值. 本实验中,α 关键值取值为 3.5.

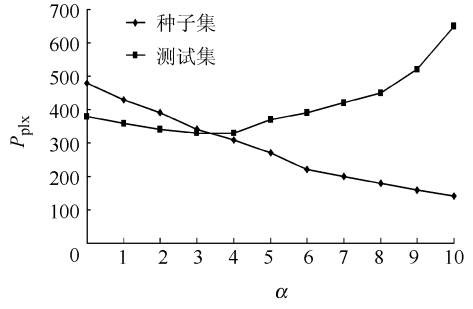

图 1 种子集和测试集 P_{plx} 变化曲线

3 实验与分析

3.1 评测标准和数据集

P_{plx} 和词错误率 WER(Word Error Rate)是评测语言模型最常用的两个指标. P_{plx} 是语言模型预计的词的候选个数,P_{plx} 越低语言模型越好;语言模型最终的评估标准应该是词的错误率,词

的错误率越低,认为语言模型性能越好.

本实验采用 P_{plx} 和 WER 作为评测标准. 选用的训练集是《人民日报》1998 年 1～6 月人工标注的语料,包括大约 13 000 000 个汉字. 特定领域训练数据和测试数据均来源于"新华网简体中文版-网评频道",特定领域训练数据包括大约 3 100 000 个汉字,测试数据包括大约 1 200 000 个汉字.

3.2　实验结果与分析

表 1 给出了基于 N 元语法分布的语言模型自适应方法的实验结果. 以新华网简体中文版-网评频道 XHC(Xinhuanet Comments)语料训练的语言模型测试结果为基准,与其它方法的实验结果进行比较.

表 1　不同方法下 P_{plx} 和词错误率比较

训练集	方法	P_{plx}	P_{plx} 减小的分率/%	词错误率/%	词错误减小的分率/%
XHC	基准	536.72	—	6.91	—
XHC+PD	SI	348.37	35.09	6.23	9.84
XHC+PD	KSI	315.68	41.18	5.97	13.60
XHC+PD	AND	309.71	42.30	5.80	16.06

第 3 行是使用简单插值法 *SI*(*Simple Interpolation*),把 *XHC* 训练得到的语言模型和由普通领域数据人民日报语料 *PD*(*People Daily*)训练得到的模型合并后的实验结果;第 4 行是考虑 *Katz* 平滑的插值算法 *KSI* 得到的实验结果;第 5 行是使用基于 N 元语法分布的语言模型自适应方法 *AND*(*LM Adaptation Based on* N-*grams Distribution*)得到的实验结果.

通过以上实验,可以得到如下结论:①考虑 *Katz* 平滑的插值算法可以在一定程度上降低单词困惑度和词错误率. ②基于 N 元语法分布的模型自适应方法可以使系统识别效果达到最优,单词困惑度和词错误率比传统的简单插值算法分别降低了 11.1% 和 6.9%.

4　结 束 语

本文针对普通领域训练语料和特定领域训练语料存在的差异,提出基于 N 元语法分布的语言模型自适应算法. 该算法不是合并训练集和种子集的 N 元语法模型,而是合并 N 元计数和自适应权重. N 元语法分布很好地表征了语料库的特征,能恰当地描述特定领域训练数据的特性,因此,可以将训练语料的 N 元语法分布自适应到和种子集的 N 元语法分布相似,以便更好地进行特定领域单词的识别.

需要说明的是,本文在自适应系数选取过程中,如果领域发生了变化,则需重新进行模型训练确定自适应系数,这在应用中是很不方便的,如何找到更好的方法来解决这个问题是后继研究的重点之一. 另外,综合考虑 *Katz* 平滑和基于 N 元语法分布的模型自适应方法,可能会进一步提高系统的识别效果,这也将是后继研究的重点.

参考文献(References)

[1] Jelinek F. Self-organized language modeling for speech recognition [C]// Readings in Speech Recognition. Morgan-Kaufmann, CA: San Mateo, 1990: 450－506

[2] Miller D, Leek T, Schwartz R M. A hidden Markov model information retrieval system[C]//Proceedings of 22nd International Conference on Research and Development in Information Retrieval. USA: Berkeley, 1999: 214－221

[3] Kim W. Language model adaptation for automatic speech recognition and statistical machine translation[D]. USA: Department of Computer Science, Johns Hopkins University, 2004

[4] Iyer R, Ostendorf M, Gish H. Using out-of-domain data to improve in-domain language models[J]. IEEE Signal Processing Letters, 1997, 4(8): 221－223

[5] Rosenfeld R. A maximum entropy approach to adaptive statistical language model[J]. Computer Speech & Language, 1996, 10: 187－228

[6] Jurafsky D, Martin J H. Speech and language processing: an introduction to natural language processing, computational linguistics and speech recognition[M]. USA: Prentice Hall, 2000

[7] Katz S M. Estimation of probabilities from sparse data for the language model component of a speech recognizer[J]. IEEE Transaction Acoustics, Speech and Signal Processing, 1987, 35(3): 400－401

[8] Wu Genqing, Zheng Fang, Wu Wenhu. Improved Katz smoothing for language modeling in speech recognition[C]// International Conference on Spoken Language Processing 2002. USA: Colorado, 2002: 925－928

[9] Manning C D, Sch tze H. Foundation of statistical natural language processing[M]. USA: The MIT Press, 1999

[10] Katz S M. Distribution of content words and phrases in text and language modeling[J]. Natural Language Engineering, 1996, 2: 15－59

(责任编辑:张　嵘)