统计自然语言处理中的线性插值平滑技术

张敬芝 高 强 耿 桦 潘金贵

(南京大学计算机软件新技术国家重点实验室 南京 210093)

摘 要 统计自然语言处理中,一个很复杂的问题是数据稀疏问题。主要有两种平滑方法解决:回退法和线性插值法。本文分析和比较了几种典型的线性插值方法,着重研究了它们所引发的词性聚类倾向。在此基础上,给出了2种改进的平滑方法。实验表明,改进的方法比原来的方法有更出色的平滑效果。

关键词 统计语言模型,数据稀疏问题,平滑技术,回退法,线性插值法,n-gram

Linear Interpolated Methods in Statistical Natural Language Processing

ZHANG Jing-Zhi GAO Qiang GENG Hua PAN Jin-Gui

(State Key Laboratory for Novel Software Technology of Nanjing University, Nanjing 210093)

Abstract One of the complicated problems in statistical natural language processing is the data-sparseness problem. There are mainly two kinds of smoothing technologies to solve it, backing-off models and linear interpolated models. This article compares several typical linear interpolated methods, and focuses on studying the relationship between the smoothing parameters and the parts of speech. Besides, two improved methods are proposed. Our experiment results show that both of them outperform original ones.

Keywords Statistical language model, Data sparse problem, Smoothing technology, Backing- off methods, Linear interpolated methods, N-gram

1 引言

统计自然语言处理技术认为,语言模型就是该语言的单词序列空间上的某种概率分布,它反映了任何一个单词序列S成为该语言中的一个句子的可能性[1]。迄今为止,应用最广泛的是n-gram模型。该模型基于马尔可夫假设:第 n 个词的出现,由前面已经出现的($n-1$)个词决定。所以n-gram模型是($n-1$)阶的马尔可夫模型。

考虑单词序列 $S=w_1w_2\cdots w_l$,其发生概率为:$p(S)=\prod_{i=1}^{l} p(w_i|w_{i-n+1}\cdots w_{i-1})$

常用的是 $n=2$ 和 $n=3$ 的情况,分别又称为Bigram和Trigram模型。概率的计算采用最大似然估计(Maximum Likelihood Estimation,MLE):

$$p(w_i|w_{i-n+1}^{i-1})\approx p_{ML}(w_i|w_{i-n+1}^{i-1})=\frac{c(w_{i-n+1}^{i})}{\sum_{w_i}c(w_{i-n+1}^{i})}=\frac{c(w_{i-n+1}^{i})}{c(w_{i-n+1}^{i-1})}$$

MLE基于二项分布假设,是以训练语料库中单词序列实际出现的频率来估计它在语言中的真实概率。其不足是,过分依赖于训练集。一般来说,训练集越大,根据该方法得到的结果越符合经验值。但语料库的容量总是有限的,对于那些在训练语料中没有出现的单词序列,MLE方法将简单地赋予零概率。造成的后果是,在训练集中未出现的n-gram一旦在测试集中出现,测试集的交叉熵将无穷大,使得模型完全不可用。另一方面,我们希望通过增加马尔可夫模型的阶数来提高模型的准确性,但实际情况常常相反。随着 n 的增大,零概率纵向传递、累积的效果也就越明显,导致统计数据的可靠性进一步下降。比如,计算同一个单词序列 S 的概率,使用二元模型 $p(S)=\prod_{i=1}^{l} p(w_i|w_{i-1})$ 比使用四元模型 $p(S)=\prod_{i=1}^{l} p(w_i|w_{i-3}w_{i-2}w_{i-1})$ 产生零概率的风险小得多。这也是Bigram和Trigram模型比其他更高阶模型更适用的原因之一。

零概率问题又称数据稀疏问题,是统计自然语言处理中最普遍的问题之一。为了使得语言模型可用,必须确保那些在训练集中未出现的n-gram也获得一个非零概率。而如何分配这些概率正是平滑技术所要解决的。

已经产生了许多有效的平滑技术,主要归于两类:回退法和线性插值法。Chen和Goodman的研究表明,前者在大的训练语料下会有更好的表现;而后者在训练语料较小的情况下占优势。

本文使用Bigram模型比较了统计自然语言处理中的几种典型的线性插值平滑技术,并在此基础上给出了2种新的平滑方法。

2 线性插值平滑技术

线性插值平滑的基本思想是:如果没有足够的训练数据来准确地估计高阶模型,则可以使用更低阶的模型来对高阶模型进行插值,因为低阶模型受数据稀疏问题的影响更小一些。

插值模型有多种不同的实现版本,本文中提到的包括早期的additive方法,以及后来出现的Jelinek-Mercer方法和Witten-Bell方法。Chen和Goodman在比较和总结前人方法

张敬芝,高 强,耿 桦 硕士研究生,主要研究方向为自然语言处理,Web信息检索和抽取;潘金贵 教授,博士生导师,研究方向为多媒体信息处理。

的基础上给出了另一个更好的方案。

2.1 Additive 方法

Additive 方法最初由 Laplace 于 1814 年提出，通过假设每个事件比实际出现次数多 1 次(adding one)来避免零概率。其后，Lidstone 等人对该方法进行了泛化，假设事件比实际出现次数多 $\delta(0<\delta\leqslant 1)$：

$$p_{add}(w_i|w_{i-n+1}\cdots w_{i-1})=\frac{\delta+c(w_{i-n+1}\cdots w_{i-1}w_i)}{|V|\delta+\sum_{w_i}c(w_{i-n+1}\cdots w_{i-1}w_i)}=\frac{\delta+c(w_{i-n+1}\cdots w_{i-1}w_i)}{|V|\delta+c(w_{i-n+1}\cdots w_{i-1})} \quad (1)$$

其中，V(vocabulary)是根据训练语料库得到的总词集。

Additive 平滑方法是一种最简单的插值。实际上，它并没有考虑低阶模型的特点，只是 MLE 和统一的先验概率之间的线性插值。该方法已被证明为效果很差[2]，在实际处理中很少使用。

2.2 Jelinek-Mercer 方法

首次将低阶模型考虑在内的插值技术由 Jelinek 和 Mercer(1980)提出，而更完善的模型则来自 Brown 等人(1992)：

$$p_{interp}(w_i|w_{i-n+1}^{i-1})=\lambda(w_{i-n+1}^{i-1})p_{ML}(w_i|w_{i-n+1}^{i-1})+(1-\lambda(w_{i-n+1}^{i-1}))p_{interp}(w_i|w_{i-n+2}^{i-1}) \quad (2)$$

后面介绍的几种插值方法都使用这一形式。区别仅在于，不同的插值方法对应不同的参数集合 $\lambda(w_{i-n+1}^{i-1})$。

参数 $\lambda(w_{i-n+1}^{i-1})$的计算是插值模型的最重要的工作。因为对每一个历史数据 w_{i-n+1}^{i-1}都训练一个 $\lambda_{w_{i-n+1}^{i-1}}$会恶化数据稀疏问题，Jelinek 和 Mercer 等人建议应该依据某种策略将 w_{i-n+1}^{i-1}分桶，每个桶内的参数 $\lambda_{w_{i-n+1}^{i-1}}$取同样的值。

他们提出的方法是根据 $c(w_{i-n+1}^{i-1})$把 $\lambda_{w_{i-n+1}^{i-1}}$划分到等价类。这样，所有具有相同频率的历史都绑定了一个相同的权值。

2.3 Witten-Bell 方法

令 $N_{1+}(w_{i-n+1}^{i-1}\bullet)$表示跟在历史 w_{i-n+1}^{i-1}后的 word 种类数，即 $N_{1+}(w_{i-n+1}^{i-1}\bullet)=|\{w_i:c(w_{i-n+1}^{i-1}w_i)>0\}|$

$N_{1+}(w_{i-n+1}^{i-1}\bullet)$越小，表明历史 w_{i-n+1}^{i-1}对其后跟随的单词的选择性越强，MLE 对于已观测数据的统计也就越可靠，故模型中相应的权值 $\lambda_{w_{i-n+1}^{i-1}}$也该设置得越大；反之，若历史 w_{i-n+1}^{i-1}后跟随的单词很分散，表明事件发生的随机性很大，MLE 对于已观测数据的统计也就不可靠，应该给予较小的权值 $\lambda_{w_{i-n+1}^{i-1}}$。Witten 和 Bell 建议，参数应满足：

$1-\lambda_{w_{i-n+1}^{i-1}}=\frac{N_{1+}(w_{i-n+1}^{i-1}\bullet)}{N_{1+}(w_{i-n+1}^{i-1}\bullet)+c(w_{i-n+1}^{i-1})}$可见，该方法实际上用一个新事件出现的概率来代替训练语料中未出现的事件的概率。

2.4 Chen-Goodman 方法

在比较和总结了前人方法的基础之上，Chen 和 Goodman (1996)提出了被他们自己称之为 average-count 的新方法。此方法使用$\frac{c(w_{i-n+1}^{i-1})}{|w_i:c(w_{i-n+1}^{i-1})>0|}$来划分参数 $\lambda_{w_{i-n+1}^{i-1}}$的空间。实验证明这种划分依据比前面的方法能更好地刻画“稀疏”的概念。

3 我们的工作

3.1 Variance 方法

以上几种方法的区别主要源于对“稀疏”问题认识的不同。Chen 和 Goodman 的方法比其他方法更接近问题的本质，但仍然存在不足。下面的例子可以很好地说明这一点。

训练语料库中单词 h_1 和 h_2 均出现 9 次，各自的分布如下：

$c(h_1w_1)=5, c(h_1w_2)=2,$

$c(h_1w_3)=1, c(h_1w_4)=1$

$p(w_1|h_1)=0.5556, p(w_2|h_1)=0.2222,$

$p(w_3|h_1)=0.1111, p(w_4|h_1)=0.1111$

$c(h_2w_1)=2, c(h_2w_2)=2,$

$c(h_2w_3)=2, c(h_2w_4)=3$

$p(w_1|h_2)=0.2222, p(w_2|h_2)=0.2222$

$p(w_3|h_2)=0.2222, p(w_4|h_2)=0.3333$

按照 Jelinek-Mercer 以及 Chen-Goodman 方法，这 2 个历史数据的稀疏程度相当，所以应该有相同的参数。但根据观察我们可以发现，这两种历史的局部空间的概率分布相差很大。h_1 后面单词的分布不均匀，对词的选择性强；h_2 后面单词的分布比较均匀，对词的选择性弱。分布的不同导致这两种历史的稀疏程度也不同。我们倾向于认为选择性强的历史，MLE 的估计越接近真实，应该给予更大的权重。

我们的方法是根据方差(variance)$\frac{\sum_{w_i}(c(w_{i-n+1}^{i})-\bar{c})^2}{c(w_{i-n+1}^{i-1})}$(其中，$\bar{c}$ 是 Chen-Goodman 方法的依据)来划分 λ 空间。因为方差能够很好地反映历史数据局部空间的分布情况，而此分布是影响数据稀疏问题的重要因素之一。

3.2 New Additive 方法

将词性因素考虑在内，我们对传统的 Additive 方法作了如下的修正：

$$p_{add}(w_i|w_{i-n+1}\cdots w_{i-1})=\frac{\delta_{V(w_i)}+c(w_{i-n+1}\cdots w_{i-1}w_i)}{|V|\delta_{V(w_i)}+c(w_{i-n+1}\cdots w_{i-1})} \quad (3)$$

定义 $\delta_{V(w_i)}=\frac{|V(w_i)|}{|V|}$；$V(w_i)$和 V 分别是单词 w_i 所在的词集(按词性划分)和训练语料的总词集。通过以下的实验证明，新的 Additive 方法比原来的方法有明显的改进。

4 实验

我们使用了 ICE-HK 的英文语料库，文本总长度超过 1000,000 个词次，覆盖 36,000 种词形(注：我们未对单词做词干化处理，视 operate, operating, operated 为 3 种不同的词形)。实验之前，先将语料库分成三部分：随机选择大约 100,000 词作为留存数据，训练参数 λ；同样大小的集合做测试数据；剩下的全部用作训练集。λ 的计算采用 Baum-Welch 迭代算法。

出于方便，对于训练集的词性标注，我们只使用了一个包含 12 种词性的粗糙词性集 V={adjective(1), adverb (2), conjunction(3), cardinal numeral(4), determiner(5), noun (6), ordinal(7), pronoun(8), preposition(9), verb(10), interjection(11), alphabetical(12)}。落在这个集合之外的其他词性的词不在我们此次实验的考虑范围之内。

4.1 数据分析

通过实验我们发现，Witten-Bell 和 Chen-Goodman 方法都有一定的词性聚类倾向，暗示不同词性的单词其周围局部空间的分布不同。而 Jelinek-Mercer 方法则没有这种倾向，

它对于单词的划分带有很大的盲目性。比如同样是冠词，a与 the 在语料库中出现的频率分别是 82011 次和 74926 次，而 an 只出现了 3000 多次，与前两个词的词频段相隔很远。但这三个词是同一类词，其后面所跟的单词的分布很相似。

使用词频分类，只有频率比较高的一些词（依次如冠词、短介词、代词、系动词等）反映了一定的词性聚类的倾向，即词性自动按照词频聚类（表 1）。

表 1　ICE-HK 语料库的部分词频信息

序号	单词	词频
1	a	82011
2	the	74926
3	to	37612
4	and	35123
5	of	31748
6	in	25384
7	you	24101
8	it	20849
…		

但随着词频的下降，这种方法导致的聚类效果越来越差。词频在 1000 次以下，多种词性的单词混乱地聚集在同一个词频段上（图 1）。与之相反，对低频词的聚类，Chen-Goodman 的方法却好得多（图 3）。后面的 4 张图均是对词频在 100～1000（主要的词频段）的共 1663 种不同词形的词性聚类。其中，纵坐标表示词性，图中的每个点表示一种词形。

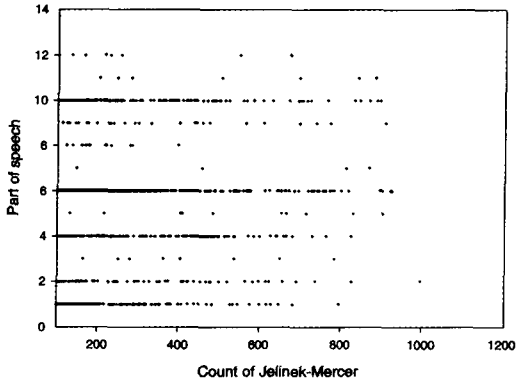

图 1　Jelinek-Mercer 方法的词性聚类

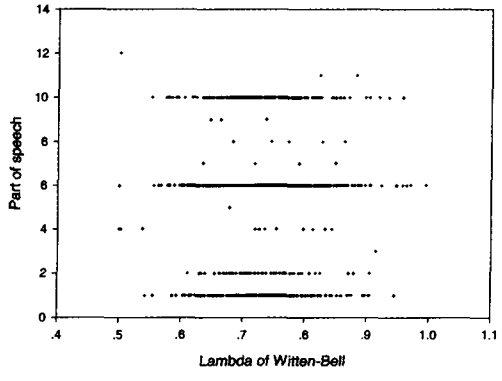

图 2　Witten-Bell 方法的词性聚类

从词性聚类的角度来看，variance 方法比其他 3 种方法的聚类效果更好。在图 4 中，conjunction(3)，determiner(5)，alphabetical(12) 被约束在了一个 variance 段上；而 adverb(2)，cardinal numeral(4)，ordinal(7)，pronoun(8)，preposition(9) 也均呈现了比较明晰的归类划分；即使像数量庞大的 adjective(1)，noun(6)，verb(10) 类词，其分布也有了不同程度的收敛。

我们还注意到，尽管 Witten-Bell 方法的聚类效果不是很理想，但它给出了 λ 参数分布的大体形态。几乎对所有词性的词集，其 λ 都近似满足平均值在 0.725 左右的正态分布（图 2）。这也说明按照词性聚类的方法来平滑 MLE 的统计数据是合理可行的，因为不同词性的词集隐含潜在一致的折扣率。而传统的平滑方法不考虑词性因素，将所有词性的词集并在一起统一折扣，容易破坏这些词集间原有的比例关系。这恰恰是造成原来的 Additive 方法不可用的原因。按照这种思路，我们还改进了原有的 Additive 方法（见 3.2 节）。

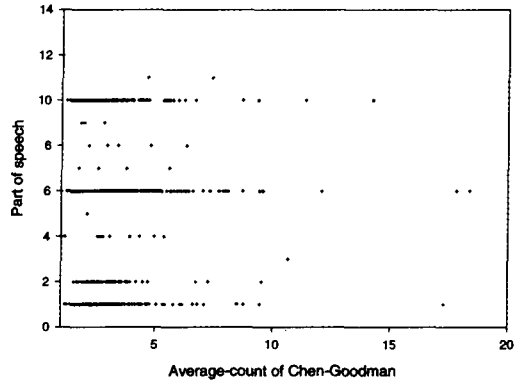

图 3　Chen-Goodman 方法的词性聚类

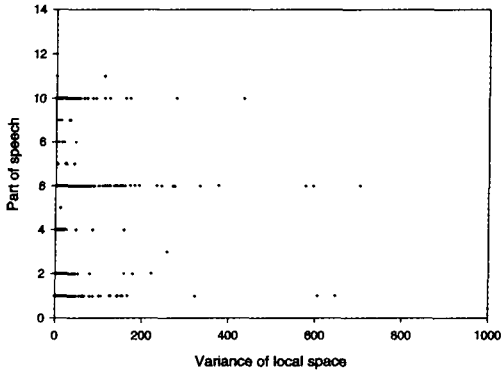

图 4　variance 方法的词性聚类

4.1　实验结果

我们使用交叉熵（cross-entropy）和复杂度（perplexity）对模型进行评估。交叉熵或复杂度越小，说明模型的适应性就越好，平滑产生的效果更出色。表 2 是上述不同方法在同一测试集下得到的交叉熵和复杂度。

表 2　不同方法的交叉熵和复杂度

方法	交叉熵	复杂度
Additive	13.53	11828.7
New Additive *	9.01	515.6
Jelinek-Mercer	8.22	298.2
Witten-Bell	8.09	272.5
Chen-Goodman	7.93	243.9
Variance *	7.91	240.5

注意到 2 种改进的方法 New Additive 与 Variance 都比原来的方法有更好的适应性。

结束语　本文比较了统计自然语言处理中的几种典型的

（下转第 244 页）

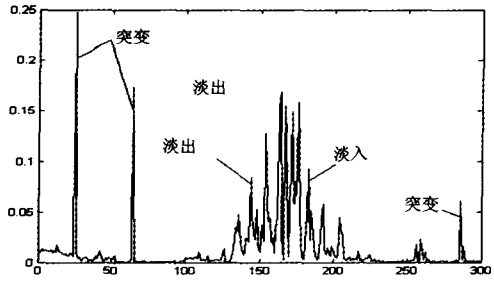

图 1 本算法的检测结果

4 实验

镜头边界检测结果的评价包括查全率(Recall)和查准率(Precision)两个指标,它们的定义分别为:

查全率=正确检测数/(正确检测数+漏判数)

查准率=正确检测数/(正确检测数+错判数) (11)

这两个指标往往是一对矛盾:仅仅为了提高查全率,通常会导致误检,使得查准率下降;而仅仅为了提高查准率,通常会导致漏检,使得查全率下降。因此,查全率和查准率都高才能说明检测方法好。

在实验中,选用了五个具有不同特点的视频片段组成本系统的实验数据集,包括:电影《简爱》中聚会的片段(JE)、儿童片《绿野仙踪》中庆祝坏女巫死亡的片段(LY)、体育比赛中"排球比赛"的结尾片段(PQ)、晚间新闻中的反映贫困山区教育片段(XW)和 MTV《赞酒歌》中歌伴舞片段(ZJ)。视频序列从 4 千多帧到 2 万多帧,每帧为 352×240 像素。

表 1 实验结果

视频片段	帧数	突变	渐变	本算法(%)		VideoAnnEx 算法(%)	
				查准率	查全率	查准率	查全率
JE	18644	76	6	89.9	92.2	95.1	80.3
LY	11740	38	3	90.5	94.2	94.6	82.7
PQ	4576	23	0	92.6	95.5	98.6	74.7
XW	4650	16	0	88.4	93.3	91.9	98.3
ZJ	25532	44	3	85.7	95.8	93.9	83.3

表 1 给出了本算法与 VideoAnnEx 系统中镜头边界检测算法[5]在突变和渐变镜头边界检测中的实验对比结果。从实验结果可以看出,文[5]算法虽然对运动和闪光灯场景具有较高的鲁棒性,但同时取得了较低的查全率。从视频结构分析的角度来看,由于镜头边界检测是场景分析的基础,在场景分析中可以有效地合并过分割的镜头,因此,查全率相对于查准率更为重要。本文方法既取得了较好的查准率,同时也取得了较高的查全率。

图 2 中的实验结果取自于电影《简爱》中 300 视频帧,图 2(a)给出的是基于颜色特征的图像熵差值的变化的曲线,检测出包含 5 个切变,其中错检 2 个,漏检 1 个。图 2(b)给出的是利用本算法融合了颜色特征和纹理特征的图像熵差值变化的曲线,检测出 4 个切变,与人工检测的结果完全一致,因此避免了镜头内物体运动或光线突然发生变化时而发生的误检。

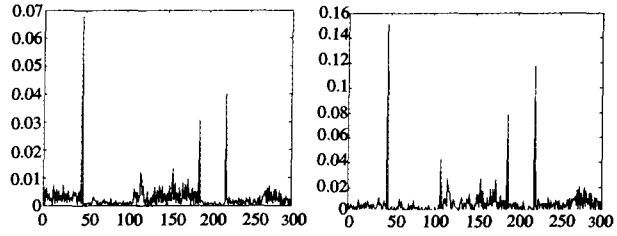

图 2 图像熵差值变化曲线

结论 文中介绍了典型的镜头边界检测的基本方法,并针对其不足提出了一种基于信息熵的自适应阈值的镜头边界检测算法,首先利用小波变换提取了帧图像的颜色特征和纹理特征,然后利用信息熵的方法来计算帧间差,由于融合了颜色特征和纹理特征,从而避免了镜头内物体运动或光线突然发生变化时而发生的误检,并根据滑动窗口中差值的分布来动态计算局部阈值,该方法能较好地检测出镜头突变,对渐变镜头也能达到检测的目标,下一步可以对分割出来的镜头进行语义上的分析聚类,形成高层的视频结构,以提供视频基于语义的检索。

参 考 文 献

1 Zhang H J, Kankanhalli A, Smoliar S W. Automatic partitioning of full-motion video. ACM Multimedia Systems, 1993, 1(1): 10～28

2 Rubner Y. Perceptual metrics for image database navigation: [PhD Thesis]. Stanford University, May 1999

3 Cheng W G, Xu D, Jiang Y W, Lang C Y. Information theoretic metrics in shot boundary detection. In: Proc. of IIH-MSP, LNCS 3683, Melbourne, Australia, Sept. 2005. 388～394

4 Bescos J, Cisneros G, et al. A unified model for techniques on video-shot transition detection. IEEE Trans. on Multimedia, 2005, 7(2): 293～307

5 Amir A, Berg M, et al. IBM research TRECVID-2003 video retrieval system. In: TRECVID Workshop, Washington D. C. USA, Nov. 2003

(上接第 225 页)

插值平滑技术。在实验中我们发现,Chen-Goodman 方法比其他方法优越的原因在于该方法所引发的词性聚类倾向比较明显。词性因素可能是划分模型参数的重要因子。基于这种启发,我们对各种方法导致的词性聚类作了对比分析,并在此基础上,给出了 2 种改进的平滑方法。实验结果表明,这 2 种方法比原来的方法有相对更好的适应性。

参 考 文 献

1 Chen S F, Goodman J. An Empirical Study of Smoothing Techniques for Language Modeling: [Technical Report TR-10-98]. Computer Science Group, Harvard University, 1998

2 Gale W A, Church K W. What's wrong with adding one? In: N. Oostdijk, P. de Haan, eds. Corpus-Based Research into Language. Rodolpi, Amsterdam, 1994

3 Gale W A, Sampson G. Good-Turing frequency estimation without tears. In: Journal of Quantitative Linguistics, 1995, 2(3): 217～237

4 Church K W, Gale W A. A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. Computer Speech and Language, 1991, 5(1): 19～54

5 Chen S F. Building probabilistic models for natural language. Harvard University, Cambridge, MA, 1996

6 Jelinek F, Mercer E L. Interpolated Estimation of Markov Source Parameters from Sparse Data. In: D. Gelsema and L. Kanal, eds. Pattern Recognition in Practice. North-Holland, 1980

7 Katz S. Estimation of probabilities from sparse data for the language model component of a speech recognizer. IEEE ASSP, 1997, 35 (3): 400～401

8 Kneser R, Ney H. Improved Backing-off for M-Gram Language Modeling. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1995, 1: 181～184

9 Ney H, Essen U. On smoothing Techniques for Bigram-based Natural Language Modelling. In: Proceedings of the IEEE 1991 International Conference on Acoustic, Speech, and Signal Processing, Toranto, 1991. 251～258

10 Manning C D, Schutze H. 统计自然语言处理基础. 苑春法, 等译. 电子工业出版社, 2005