ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.20, No.11, November 2009, pp.2915−2924
doi: 10.3724/SP.J.1001.2009.03435

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

头驱动句法分析中的直接插值平滑算法*

刘 水+, 李 生, 赵铁军, 刘鹏远

(哈尔滨工业大学 计算机科学与技术学院,黑龙江 哈尔滨 150001)

Directly Smooth Interpolation Algorithm in Head-Driven Parsing

LIU Shui+, LI Sheng, ZHAO Tie-Jun, LIU Peng-Yuan

(Institute of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

+ Corresponding author: E-mail: liushui@mtlab.hit.edu.cn

Liu S, Li S, Zhao TJ, Liu PY. Directly smooth interpolation algorithm in head-driven parsing. *Journal of Software*, 2009,20(11):2915−2924. http://www.jos.org.cn/1000-9825/3435.htm

Abstract: Based on the classical smoothing technology, this paper proposes a smoothing approach within head-driven parsing, which directly calculates interpolation weight from the average occurrences of event in the training sample and is proved by the statistic theory of errors. By using this approach and deriving zero-value assumption from other smoothing technologies, this paper proposes four smoothing algorithms for head-driven parsing. Experiments indicate that these four smoothing algorithms have higher performance than the Baseline algorithm and reduce the disturbing curve of the optimized parameter significantly, which prove the effectiveness of the proposed approach.

Key words: parsing; smoothing algorithm; interpolation smoothing; head-driven parsing

摘 要: 在头驱动句法分析模型下,基于经典插值平滑算法,提出了以统计空间中平均事件数为基础的直接插值平滑建模原则,并应用经典的误差理论分析了该原则的合理性.基于该原则并借鉴语言模型中其他插值平滑算法对模型的零点进行假设的方法,在头驱动句法分析模型下,重新构造了 4 种平滑算法.实验数据显示,新平滑算法在高于经典平滑算法性能的同时,显著降低了自由参数的扰动程度,从实验的角度证明了该平滑建模原则的有效性.

关键词: 句法分析;平滑算法;插值平滑;头驱动句法分析

中图法分类号: TP18 **文献标识码**: A

句法分析中主要面临的问题是结构歧义的问题[1,2].在句法分析中,结构歧义分为两种:由于语义歧义造成的结构歧义和由于句法分析算法造成的分析歧义.由于目前的句法分析技术都是在单句句法分析范畴内,因此基本无法很好地解决语义歧义的问题.句法分析主要解决的是由于分析算法造成结构歧义问题:对于同一输入序列存在多个候选分析结果.

词汇化是解决结构歧义的一个重要的手段.词汇信息作为一种近似的语义信息具有很强的消歧能力,但是,

* Supported by the National Natural Science Foundation of China under Grant Nos.60736014, 60773069 (国家自然科学基金); the National High-Tech Research and Development Plan of China under Grant No.2006AA010108 (国家高技术研究发展计划(863))

Received 2008-05-20; Accepted 2008-08-11

在基于统计的极大似然模型中,几乎所有词汇化模型都存在数据稀疏问题,所以,解决数据稀疏问题是词汇化句法分析模型中的一个重要问题.

本文涉及的数据稀疏问题是在统计事件存在多个属性的前提下发生的,产生数据稀疏的原因与统计数据的属性数目息息相关:统计数据属性数目越大,数据稀疏情况越显著.鉴于此,本文用如下形式表示统计事件:若统计事件有属性$x_1,x_2,\ldots,x_n$,则统计空间的事件用这些属性的笛卡尔乘积$x=x_1\times x_2\times\ldots\times x_n$表示.下面给出本文基于统计事件属性个数的统计空间维度的概念.

定义 1. 统计空间的维度$n=|x|$,其中$x=x_1\times x_2\times\ldots\times x_n$,$n=|x|$为事件的属性个数.

回退平滑(backoff)和插值平滑(interpolation)是两种主要的平滑技术[3].回退平滑应用折扣(discount)在计算统计空间的零概率事件的概率值时回退到维度较低的统计空间概率,为了满足统计空间的归一化条件,当统计事件稀疏程度不显著时(一些回退算法并不对所有统计事件进行折扣)应用折扣技术进行重新统计.

插值平滑[4]采用另一种方法来解决数据稀疏问题,平滑算法由两部分组成,形如公式(1).

$$\tilde{p}^n=\lambda\hat{p}^n+(1-\lambda)\tilde{p}^{n'} \tag{1}$$

其中,n,n'为极大似然模型下统计事件的维度,$\tilde{p}^n$为插值平滑算法下n维事件的概率估计,$\hat{p}^n$为训练空间中n维事件的极大似然统计,且$0\le\lambda\le1$,$n>n'$.

插值平滑把数据稀疏问题看成一种泛化的问题,在插值平滑框架下的高维概率都由训练空间的高维度事件的似然概率和低维事件概率估计(插值平滑可以是个多极递归的过程)的加权平均得到.插值平滑概率模型的形式简单,训练过程也不复杂,是 NLP(nature language processing)领域中一种得到广泛应用的平滑技术.

本文第 1 节首先介绍条件似然空间理论及在条件似然空间下的一些平滑算法,然后介绍在条件似然空间下的头驱动句法分析概率模型,最后介绍在头驱动句法分析概率模型下的平滑算法.第 2 节对第 1 节介绍的头驱动句法分析平滑算法进行分析,并引入平均事件数建模原则,根据该原则进行平滑算法建模.第 3 节根据两种经典的统计理论分别对第 2 节提出的建模原则进行分析.第 4 节介绍实验设置、实验过程及结果分析.第 5 节给出本文的结论和本文成果的展望.

1 插值平滑算法及头驱动句法分析模型

1.1 条件极大似然估计(MCLE)空间下的插值平滑算法

在句法分析、词性标注等NLP领域的研究课题中,许多研究方法最终都转化为使条件概率模型$p^\theta(x|y)$在隐变量θ下取得最大值的问题.由于极大似然分布(也称作联合概率)$p(x,y)$具有收敛性,条件极大似然分布(也称作条件概率)$p^\theta(x|y)$也具有收敛性[5].与定义 1 相似,如下给出条件极大似然估计中维度的定义:

定义 2. 统计空间的维度$d=|x|+|y|=n+m$,其中$x=x_1\times x_2\times\ldots\times x_n$,$|x|$为事件的属性个数,$y=Y_1\times Y_2\times\ldots\times Y_m$,$|y|$为条件极大似然统计中条件的属性数目.

图 1 是极大似然估计(maximum likelihood estimate,简称MLE)和条件极大似然估计(MCLE)的统计空间,可以看出,条件极大似然下的概率不是在整个统计空间Ω中统计得到的,而是从条件概率的边际分布$p(y_i)$在统计空间中划定的一个子区间Ω_i中统计得到的.

插值平滑算法的核心是插值权值λ的求法.直接法和似然优化法是计算插值权值的两种主要方法[6]:似然优化通过对开发集语料的似然优化得到插值权值,直接法根据概率空间的统计参数直接建摸计算插值权值.

似然优化法反映了所有似然统计空间的平均情况:统计空间中所有平滑统计概率的计算都使用这个平均权重.而在条件极大似然模型中,每个平滑统计概率都是从相对独立的条件子空间Ω_i得到的,应用似然优化法求得权重并没有区分每个条件子空间Ω_i的权重差异.

直接插值通过统计空间的参数直接计算插值权重.我们认为,直接插值平滑算法对于权值的求法都应遵循以下原则:高维权值与条件似然统计空间大小成正比,当高维度概率统计空间的大小趋近于无穷大时,高维度插值权值λ趋近于 1,且高维权值满足$0\le\lambda\le1$.下文把这个建模原则称作直接插值原则.

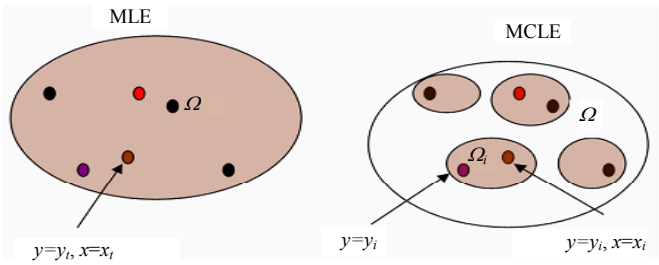

Fig.1 MLE and MCLE distribution

图 1 极大似然分布和条件极大似然分布

插值平滑算法[7](以下简称Witen-Bell平滑)和插值平滑算法[8](以下简称Bikel平滑)是两种在自然语言处理领域应用比较广泛的直接插值模型.

Witen-Bell 平滑为插值权值与统计空间的样本种类和样本大小建立一种量化关系,形如后文的公式(3).首先,公式(2)给出对于统计条件极大似然统计概率空间样本种类数 $D(X)$的概念:

$$D(X)=|Y(X)| \tag{2}$$

其中,X表示条件极大似然统计中统计事件的集合,且$X=\{x|x\in\Omega_i\}$,$Y(x)=\{x|Count(x,y)>0\}$表示条件似然空间中统计事件数目大于 0 的事件的集合,x表示条件极大似然统计中的统计事件,y表示条件极大似然统计空间的边际分布条件属性(如前文所述,x,y均为向量),$|Y(X)|$表示集合$Y(X)$的事件数目.

$$\lambda=\frac{|X|}{|X|+D(X)} \tag{3}$$

其中,$|X|$表示条件极大似然空间中的统计事件数目.

Bikel 插值平滑算法(如公式(4))在插值平滑公式中引入优化变量.

$$\lambda=\frac{|X|}{|X|+C\times D(X)} \tag{4}$$

其中,C 为常数,通过对系统性能的反馈优化而获得.文献[6]中介绍了另外一些直接插值的高维权值λ的求法:

$$\lambda=\frac{|X|}{|X|+C} \tag{5}$$

$$\lambda=\frac{|X|}{|X|+C\times OneCount(y'(X))} \tag{6}$$

其中,$OneCount(Y'(X))=|Y'(X)|$表示在统计空间中只出现过 1 次的事件个数,$Y'(X)=\{x|count(x,y)=1\}$表示在统计空间只出现过 1 次的事件集合,C 为优化常数.显然,在式(3)~式(6)中,统计空间大小$|X|$与高维权值λ具有相同的增减性,当$|X|$趋近于无穷时,插值权值λ均趋近于 1.

1.2 头驱动句法分析

在句法分析中,解码过程中的规则匹配是影响句法分析算法性能的重要因素.基于依存关系的规则匹配[9−11]和基于短语结构的规则匹配[12−14]是句法规则匹配的主要方法.

头驱动句法模型[15]在依存关系下引入短语的结构信息,在一定程度上平衡了依存关系分析和短语结构分析的优缺点:使句法分析模型在具有可扩展性的同时也很好地识别了短语边界,词汇化信息的融入也提高了句法分析的消歧能力.

1.2.1 头驱动句法分析的概率模型

头驱动句法分析的的概率计算如图 2 所示.头驱动句法分析以核心节点为界把句法规则分成左、右两个部分.按照这种依存结构的划分,一个短语规则可以用公式(7)表示:一条短语规则的概率是核心节点生成父节点的条件概率与其左、右两边所有依存关系的条件概率的乘积.

$$p(H \mid P,w,t)\times\prod_{i=1}^{m}p(L_i(lt_i,lw_i),c,p \mid P,H,w,t,\Delta)\times\prod_{i=1}^{n}p(L_i(lt_i,lw_i),c,p \mid P,H,w,t,\Delta) \tag{7}$$

其中,H表示头节点的非终结符,P代表父节点的非终结符,w代表头节点的词,t代表头节点的词性,$L_i(lt_i,lw_i)$代表距离头节点$H(h)$左边第i个位置的依存节点的非终结符、词性和词,c和p分别代表并列符号、标点符号,Δ代表距离变量.

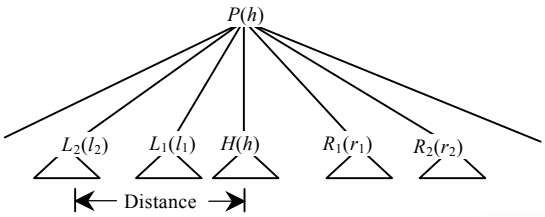

Fig.2 Dependency relation in head-driven

图 2 头驱动中的依存关系

1.2.2 头驱动句法分析模型中的平滑算法

为了降低依存条件似然概率估计的维度,根据贝叶斯过程对左、右依存项分别进行拆解,其中,对左依存项的拆解如公式(8)所示.

$$p(L_i(lt_i,lw_i),c,p \mid P,H,w,t,\Delta)=p(L_i(lt_i),c,p \mid P,H,w,t,\Delta)\times p(lw_i \mid L_i(lt_i),c,p,P,H,w,t,\Delta) \tag{8}$$

其中,$L_i(lt_i)$代表依存节点的非终结符和词性.

拆解后的概率项中仍存在词汇信息这个造成数据稀疏的主要因素,仅依赖于贝叶斯过程进行的数据平滑显然是不够的,所以,头驱动句法分析系统应用了 3 级直接插值平滑方法进一步解决数据稀疏问题.

根据公式(1),递归 3 次计算插值概率,最终得到依存概率.实际上,应用公式(1)进行的 3 级插值平滑最终转化为计算式(9):

$$\tilde{p}^{n3}=\lambda_3\hat{p}^{n3}+(1-\lambda_3)\lambda_2\hat{p}^{n2}+(1-\lambda_3)(1-\lambda_2)\lambda_1\hat{p}^{n1}+(1-\lambda_3)(1-\lambda_2)(1-\lambda_1)\hat{p}^{n0} \tag{9}$$

其中 $\tilde{p}^{n0}$ 取 10^{-19},且 $0\le\lambda_1,\lambda_2,\lambda_3\le1$,$n_1\le n_2\le n_3$,各级直接插值平滑的维参见文献[6].

2 头驱动句法分析模型下的直接插值平滑算法

文献[6]实现的头驱动句法分析模型应用的是前文提到的 Bikel 平滑(如公式(4)).根据这个公式,直接插值的权值只与两个变量有关系:统计空间大小和统计空间中事件的种类.本文将式(4)进行如下的处理:

$$\lambda=\frac{|X|/D(X)}{|X|/D(X)+C} \tag{10}$$

定义 3. 统计空间中每个事件的平均发生数目 $\tilde{n}=|X|/D(X)$,其中,$|X|$表示统计空间的样本数,$D(X)$表示统计空间的样本种类数(由公式(2)定义).

这样,通过定义 3,公式(4)等价于下式:

$$\lambda=\frac{\tilde{n}}{\tilde{n}+C} \tag{11}$$

从式(11)可以看出,实际上 Bikel 平滑只与统计空间的 1 个参数相关联,直接插值权值可以通过对这个参数的建模进行估计.对式(11)进行求导,得到:

$$\frac{\mathrm{d}\lambda}{\mathrm{d}\tilde{n}}=\frac{c}{(\tilde{n}+c)^2}>0 \tag{12}$$

从式(12)可以得出这样的结论:Bikel 直接插值高维权值只与空间平均事件数有关,并与其成正比.下文将符合这个结论的直接插值算法称为符合平均事件直接插值平滑建模原则的平滑算法(在文献[16]中也提到了空间平均事件数的概念,并且认为空间平均事件数与统计空间的熵存在一种联系).根据前面的讨论,Bikel 平滑是

一种符合直接插值建模原则的平滑算法.为了检验本文提出的直接插值建模原则,本文遵循平均事件直接插值平滑建模原则,对头驱动句法分析模型中的直接插值平滑进行了重新建模.

本文首先尝试了与公式(11)幂次相同的插值算法 smooth-1:

$$\lambda = 1 - \frac{1}{C \times \tilde{n}},\ C \geq \frac{1}{\tilde{n}} \tag{13}$$

其中,C 为优化常数.

对直接插值算法进行的建模无法直接观察到所有训练空间的情况,所以若用公式(13)确定某个C,直接插值模型无法对训练空间Ω中所有的子空间Ω_i总满足 $0 \leq \lambda \leq 1$.造成这种情况的原因是,一些条件似然统计空间Ω_i中的平均样本数目太少,无法满足 $C \geq \frac{1}{\tilde{n}}$.针对这种情况,本文采取了统一归 0 的处理:认为无法满足式插值权值限制($0 \leq \lambda \leq 1$)的统计空间不具有统计意义,直接将插值高维权值λ置为 0,使插值平滑估计直接由下一级的插值估计得到.这种通过样本空间中的参数对模型进行置 0 处理的方法,在其他平滑算法中[17]都有体现,下文将这种方法称为直接插值零点假设.

为了使插值权值对样本中参数的变化更敏感,本文直接对公式(13)的右侧平方,得到 smooth-2 平滑算法:

$$\lambda = \left(1 - \frac{1}{C \times \tilde{n}}\right)^2,\quad C \geq \frac{1}{\tilde{n}} \tag{14}$$

当然,基于空间平均事件数原则,可以对公式(13)右部尝试更高幂次的模型.此时,当平均事件数 $\tilde{n}$ 大于某个常数时,随着幂次的升高,高维权值对平均事件数 $\tilde{n}$ 的敏感程度(模型函数对 $\tilde{n}$ 的偏导数)将会升高.这就为优化常数 C 的选取带来更大的训练代价:为了使系统性能(本文应用句法分析的 F 测度衡量)最优,在训练优化常数 C 时需要选取更小的步长.基于这样的考虑,本文没有尝试幂次更高的直接插值模型.

公式(14)所计算出的权值符合插值权值限制($0 \leq \lambda \leq 1$),所以,与公式(13)相比,公式(14)实际上在一定程度上削弱了高维权值(虽然优化常数 C 可以在一定程度上抵消这种削弱).提高公式(14)对平均事件数 en 的敏感度是建立在牺牲高维权值的基础上的.如果可以在提高敏感程度的同时又不损失权值,可能对提高系统性能有所帮助.基于这样的考虑,本文构造了下面的 exp 平滑算法:

$$\lambda = 1 - \alpha^{\tilde{n}},\quad 0 \leq \alpha \leq 1 \tag{15}$$

其中,α为优化常数.与对公式(13)进行的零点假设相似,本文对公式(14)引入一个简单的直接插值零点假设,得到 exp-z 平滑算法:

$$\lambda = 1 - \alpha^{\tilde{n}-1},\quad 0 \leq \alpha \leq 1 \tag{16}$$

在公式(16)中,如果统计空间的平均事件数等于 1,那么高维权值被置为 0,为了减少训练代价,本文没有进一步尝试更高的零点值.

3 基于统计误差分析的直接插值研究及分析

我们认为,高维度统计空间的统计误差(由统计样本大小造成的误差)是高维统计概率需要应用直接插值平滑进行重新估计的主要原因.高维似然概率越准确,高维概率在平滑算法中获得的权值应越高,当高维度的条件似然统计空间大小趋近于正无穷时,高维度条件似然概率不需要插值平滑就可以得到准确的似然统计概率.实际上,可以把对插值空间的建模看作对条件似然统计空间误差的建模.下文将根据这个假设对直接插值模型进行做进一步的分析讨论.

在概率统计领域,一些经典定理都对统计空间的样本误差进行过讨论,如伯努力大数定理、契比晓夫大数定理.下面将应用这两个经典的误差统计理论对统计空间的误差上限进行分析.

公式(17)为伯努力大数定理,该定理证明统计概率与实际概率的差的绝对值大于ε的概率存在一个上限.

$$p\left(\left|\frac{x_i}{n} - p_i\right| \geq \varepsilon\right) \leq \frac{p_i(1-p_i)}{n\varepsilon^2} \tag{17}$$

其中,$OneCount(Y'(X))=|Y'(X)|$表示在统计空间中只出现过 1 次的事件个数,$Y'(X)=\{x|count(x,y)=1\}$表示在统计空间只出现过 1 次的事件集合,C 为优化常数.显然,在式(6)~式(10)中,统计空间大小$|X|$与高维权值λ成正比,当$|X|$趋近于无穷时,插值权值λ均趋近于 1.

为了得到整个统计空间的统计特性,对式(17)在统计空间中所有统计事件上求和:

$$\sum_{i=1}^{m} p\left(\left|\frac{x_i}{N}-p_i\right|\geq\varepsilon\right)\leq\sum_{i=1}^{m}\frac{p_i(1-p_i)}{n\varepsilon^2} \tag{18}$$

其中,m表示统计空间的事件种类数$D(X)$,且p_i满足归一化条件:

$$\sum_{i=1}^{m} p_i=1 \tag{19}$$

将式(19)代入式(18)右边,得到:

$$\sum_{i=1}^{m}\frac{p_i(1-p_i)}{n\varepsilon^2}=\frac{1-\sum_{i=1}^{m}p_i^2}{n\varepsilon^2} \tag{20}$$

而 $\sum_{i=1}^{m}p_i^2$ 与归一条件(式(19))可以确定一个最小值:

$$令\ F=\sum_{i=1}^{m}p_i^2-\lambda\sum_{i=1}^{m}p_i \tag{21}$$

当 F 取得极值时,$\frac{\partial F}{\partial p_i}=0$, 则 $p_i=\frac{\lambda}{2}$, 带入式(19)得到 $p_i=\frac{1}{m}$,则此时式(20)取得最大值:

$$\sum_{i=1}^{m}\frac{p_i(1-p_i)}{n\varepsilon^2}=\frac{1-\sum_{i=1}^{m}p_i^2}{n\varepsilon^2}\leq\frac{1-\frac{1}{m}}{n\varepsilon^2} \tag{22}$$

根据公式(17),得到:

$$\sum_{i=1}^{m} p\left(\left|\frac{x_i}{N}-p_i\right|\geq\varepsilon\right)\leq\sum_{i=1}^{m}\frac{p_i(1-p_i)}{n\varepsilon^2}\leq\frac{m-1}{mn\varepsilon^2} \tag{23}$$

本文根据上面的推导对插值权值进行建模,遵循直接插值原则,将式(23)不等式的最右侧与插值权值λ建立下面的关系:

$$1-\lambda=\frac{m-1}{mn\varepsilon^2} \tag{24}$$

显然,公式(24)要满足插值权值限制 $0\leq\lambda\leq1$,则

$$0\leq\frac{m-1}{mn\varepsilon^2}\leq1 \tag{25}$$

实际上,式(24)并没有产生很好的平滑效果.产生这样的结果的主要原因是,公式(23)是对误差上限的一个粗略分析,这个粗略分析无法达到实际应用中平滑算法的需求.但是,式(23)同样反映了统计空间的一些统计特性:统计误差不仅与统计空间事件数相关,也与统计空间的样本种类数相关.根据前面的定义 3,空间平均样本数是统计空间事件数与样本种类数的比值.这在一定程度上证明了直接插值建模原则的合理性.

与上面基于伯努力大数定理对误差的建模相似,本文同样基于契比晓夫大数定理对误差上限估计进行了建模.契比晓夫大数定理证明了对于相互独立且具有相同概率分布函数的随机变量,存在如公式(26)的一个误差上限.

$$p\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i-\frac{1}{n}\sum_{i=1}^{n}E(\xi_i)\right|\geq\varepsilon\right\}\leq\frac{C}{\varepsilon^2 n} \tag{26}$$

其中,ξ_i表示随机变量,ε表示误差范围,C表示ξ_i的方差上限,n表示随机变量个数,$E(\xi_i)$表示随机变量ξ_i的数学期望.令

$$\xi=\begin{cases}0, & 当x=x_i\\ 1, & 当x\neq x_i\end{cases} \tag{27}$$

其中,$x=x_i$表示统计事件x属于事件种类i,则式(26)表示当统计空间大小为n时,统计空间事件概率与该概率的数学期望在一定误差范围内存在一个概率上限.显然,此时方差上限$C\leq 1$.

与上文基于伯努力大数定理对整个统计空间误差的分析类似,将式(26)在整个统计空间中累加,得到下式:

$$\sum_{i=1}^{m}p\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i-\frac{1}{n}\sum_{i=1}^{n}E(\xi_i)\right|\geq\varepsilon\right\}\leq\sum_{i=1}^{m}\frac{C}{\varepsilon^2 n} \tag{28}$$

其中,m 表示统计样本空间事件的种类数 $D(X)$,n 表示统计空间大小$|X|$,得到:

$$\sum_{i=1}^{m}p\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i-\frac{1}{n}\sum_{i=1}^{n}E(\xi_i)\right|\geq\varepsilon\right\}\leq\frac{mC}{\varepsilon^2 n} \tag{29}$$

根据定义 3,得到:

$$\sum_{i=1}^{m}p\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i-\frac{1}{n}\sum_{i=1}^{n}E(\xi_i)\right|\geq\varepsilon\right\}\leq\frac{C}{\varepsilon^2\tilde{n}} \tag{30}$$

令 $C'=\dfrac{\varepsilon^2}{C}$,则式(30)与下式等价:

$$\sum_{i=1}^{m}p\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i-\frac{1}{n}\sum_{i=1}^{n}E(\xi_i)\right|\geq\varepsilon\right\}\leq\frac{1}{C'\times\tilde{n}} \tag{31}$$

根据上文的推导对插值权值进行建模,遵循直接插值原则,将式(31)不等式的右侧与插值权值λ,建立下面的关系:

$$1-\lambda=\frac{1}{C'\times\tilde{n}} \tag{32}$$

式(32)得到与上文基于直接插值原则的插值模型式(13)的等价形式,并可以得出这样的结论:平均事件平滑插值建模原则是符合基于契比晓夫大数定理的误差分析的.以上讨论从误差分析的角度验证了平均事件平滑插值建模原则的合理性.

4 实验设置及结果分析

4.1 实验设置

本文提出的建模方法是基于头驱动句法分析模型下的平滑方法.由此,本文选用句法分析领域标准测试集 CTB 5.0 来验证平滑算法的有效性:根据以往研究的惯例[18],我们把第 301 篇~第 325 篇这 25 篇文章作为调试集,把第 271 篇~第 300 篇这 30 篇文章作为测试集,把其余的 835 篇文章(大约 18 000 句)作为训练集.

为了验证本文提出的平均事件数直接插值原则的有效性,我们以 Bikel 平滑(公式(4))为 Baseline 与前面基于平均事件数直接插值原则构造的 4 种插值平滑算法进行了比较,这 4 种平滑算法分别是:基于公式(13)的 smooth-1 平滑算法、基于公式(14)的 smooth-2 平滑算法、基于公式(15)的 exp 平滑算法和基于公式(16)的 exp-z 平滑算法.实验过程如下:

① 使用测试集训练句法分析模型;

② 在步骤 1 的句法分析模型下对开发集进行句法分析,根据分析结果观测平滑算法优化参数区间VAR_{span}并根据优化参数区间(粗略地)选择步长;

③ 训练集以 10 000 句为起点,每次添加 1 000 句重新训练句法分析模型;

④ 根据步骤 3 得到的句法分析模型在优化参数范围内对测试集进行句法分析,得到优化参数的最优值及对应的 F 测度;

⑤ 如果训练集已经达到最大(18 000 句)则退出,否则返回步骤 3.

本文的实验目的在于检验前面所提出的插值方法的有效性.我们认为,本文所涉及的平滑算法有效性体现在如下两个方面:

(1) 有效的平滑算法可以使句法分析系统达到一定的性能.

(2) 有效的平滑算法的优化参数对训练规模的变化不应十分敏感,即应具有一定的泛化性.

如何选取优化区间的步长,也是实验中值得考虑的一个部分.在选取区间步长时,应根据句法分析模型对参数的敏感程度来选择步长,基于这个考虑,本文给出如下参数来衡量句法分析模型对优化参数的敏感度.

定义 4. F测度平均变化率$\Delta f=(f_{\max}-f_{\min})/N_{step}$.其中,$f_{\max}$为在所有训练规模下系统取得最高$f$值,$f_{\min}$为在系统取得最优F测度的训练规模下最低的$f$值,$N_{step}$为优化区间的步数.我们认为,F测度平均变化率与算法逼近区间最大值的概率成反比,具有最小F测度平均变化率的算法最容易接近于最大f值点.根据上面关于有效性(2)的定义,本文定义了一个衡量平滑算法关于训练规模变化敏感程度的变量,即参数扰动率.

定义 5. 优化参数扰动率$\Delta C=(VAR-VAR_{avg})/VAR_{avg}$.其中,$VAR$为平滑算法的优化参数在某个训练规模下的最优值,$VAR_{avg}$为所有训练规模(本实验规模为 10 000 句~18 000 句)VAR的平均值.

4.2 实验结果及讨论

图 3 为优化参数扰动表,其中,纵轴为优化参数扰动率.可以看出,5 种扰动曲线中 Baseline 算法的扰动最剧烈,smooth-2 的曲线比 smooth-1 更平稳,exp 与 exp-z 的扰动曲线比其他 3 条曲线都平稳,其中,exp-z 是所有 5 条曲线中扰动最不明显的.

从图 4 可以看出,由于训练规模问题,所有平滑算法都没有使句法分析系统的 f 曲线呈现明显的收敛趋势.在公平选择步长的基础上,Baseline 算法、smooth-1 最高 F 测度很接近,smooth-2 在 12 000 句以后的表现要优于 smooth-1,exp 和 exp-z 的 f 值明显要高于其他 3 种平滑算法.

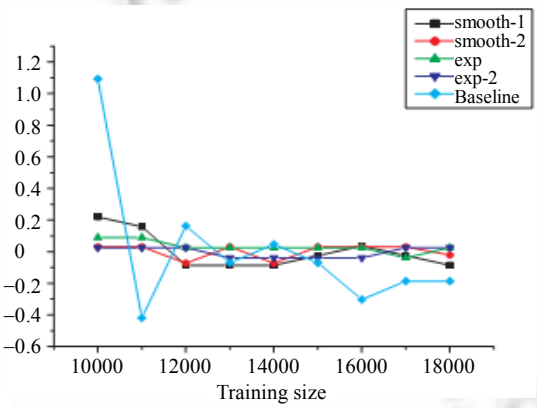

Fig.3 Disturbing curve of optimized parameters

图 3 优化参数扰动曲线

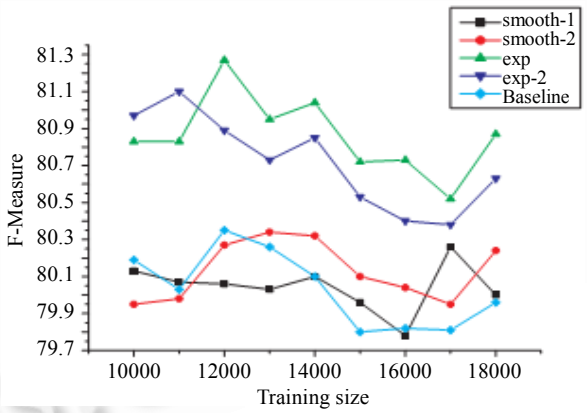

Fig.4 F-Measure in different training sizes

图 4 不同训练规模下句法分析 F 测度

表 1 为各平滑算法优化参数的训练参数表,其中,VAR_{span}为各平滑算法训练优化参数选取的参数区间,VAR_{opt}^{10000}为训练规模为 10 000 句时的优化参数最优值,σ^2为算法在各规模下优化参数与均值差的平方和.

我们认为,产生上述实验现象的因素主要有以下几个原因:

- Baseline 与 smooth-1 是幂次相同的模型,所以,二者具有接近的 f 最大值.
- 当平均事件数大于某个值时,smooth-2 比 smooth-1 对于平均事件数具有更好的敏感度,所以在 12 000 句以后,smooth-2 有更高的 f 值.
- Baseline 模型由于对平均事件数具有更好的敏感度不够而导致自由参数的扰动.自由参数的扰动是对平滑模型的一种补偿.

- 由于 smooth-2 是 smooth-1 的高阶算法,smooth-2 具有平稳的优化参数扰动曲线.
- exp 与 exp-z 对于平均事件数具有最好的敏感度,所以有更高的 f 值.
- exp 加入零点假设后,exp-z 的 f 值并没有提高反而有略微的降低,但是降低了平滑参数的扰动.

Table 1 Optimized parameters of head-driven parsing in different training sizes

表 1 句法分析模型下各训练规模下的优化参数

	smooth-1	smooth-2	exp	exp-z	Baseline
VAR_{span}	[0.5,1]	[0.5,1]	(0,1]	(0,1]	[1,10]
$f_{\max}-f_{\min}$ (%)	1.07	2.37	8.61	8.38	1.03
N_{step}	20	20	40	40	40
Δf	0.053 5	0.118 5	0.215 3	0.209 5	0.025 8
VAR_{opt}^{10000}	1.00	1.00	0.85	0.80	9.00
VAR_{opt}^{11000}	0.95	1.00	0.85	0.80	2.50
VAR_{opt}^{12000}	0.75	0.80	0.80	0.80	5.00
VAR_{opt}^{13000}	1.00	1.00	0.85	0.80	9.00
VAR_{opt}^{14000}	0.75	0.90	0.80	0.75	4.50
VAR_{opt}^{15000}	0.80	1.00	0.80	0.75	4.00
VAR_{opt}^{16000}	0.85	1.00	0.80	0.75	3.00
VAR_{opt}^{17000}	0.80	1.00	0.75	0.80	3.50
VAR_{opt}^{18000}	0.75	0.95	0.80	0.80	3.50
σ^2	0.090 0	0.038 9	0.008 9	0.005 0	47.888 9

5 结论及展望

本文在分析经典平滑算法的基础上,提出一种直接插值平滑算法建模原则,并从误差分析的角度分析了该原则的合理性.本文基于该原则在头驱动句法分析模型下构造了 4 种直接插值平滑算法,从实验数据上分析,这 4 种算法在与 Baseline 算法达到同水平的平滑效果(exp 算法和 exp-z 算法使头驱动句法分析模型的 F 测度曲线有明显的提升)的同时,具有更好的优化参数稳定度,从实验的角度验证了该原则的有效性.

本文提出的平滑算法建模原则的泛化性,有待于其他平滑算法的应用领域的继续验证.同时,研究者可以基于该原则尝试更多既有的经典概率模型(比如经典的泊松分布),而不仅仅局限于本文提出的 4 种平滑算法.基于该原则,其他研究者一定会构造出更多、更好的平滑算法.

致谢 在此,我们向对本文的工作给予支持和建议的实验中心和老师,尤其是李晗静老师、徐志明老师、曹海龙博士,以及哈尔滨工业大学机器翻译实验室的其他老师和同学表示感谢.

References:

[1] Allen J. Natural Language Understanding. Redwood: Benjamin Cummings Publishing Company, 1987.

[2] Magerman DM. Natural language parsing as statistical pattern recognition [Ph.D. Thesis]. Stanford: Stanford University, 1994.

[3] Chen SF, Goodman J. An empirical study of smoothing techniques for language modeling. In: Proc. of the 34th Annual Meeting on Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 1996. 310−318.

[4] Frederick J, Mercer RL. Interpolated estimation of Markov source parameters from sparse data. In: Gelsema ES, Kanal LN, eds. Proc. of the Workshop on Pattern Recognition in Practice. New York: Institute of Electrical and Electronics, 1980. 381−397.

[5] Johnson M. Joint and conditional estimation of tagging and parsing models. In: Proc. of the 39th Annual Meeting of the Association of Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2001. 322−329.

[6] Collins M. Head driven statistical models for natural language parsing [Ph.D. Thesis]. Philadelphia: University of Pennsylvania, 1999.

[7] Witten IH, Bell TC. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. IEEE Trans. on Information Theory, 1991,37(4):1085−1094.

[8] Bikel DM, Miller S, Schwartz R, Weischedel R. Nymble: A high-performance learning name-finder. In: Grishman R, ed. Proc. of the 5th Conf. on Applied Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 1997. 194−201.

[9] Eisner JM. Three new probabilistic models for dependency parsing: An exploration. In: Tsujii J, ed. Proc. of the 16th Int'l Conf. on Computational Linguistics. 1996. 340−345.

[10] Collins M. Three generative lexicalized models for statistical parsing. In: Proc. of the Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 1997. 16−23.

[11] Hall K. *k*-Best spanning tree parsing. In: Proc. of the 45th Annual Meeting of the ACL. Stroudsburg: Association for Computational Linguistics, 2007. 392−399.

[12] Klein D, Manning CD. Accurate unlexicalized parsing. In: Proc. of the 41st Annual Meeting of the Association of Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2003. 423−430.

[13] Dreyer M, Eisner J. Better informed training of latent syntactic features. In: Grishman R, ed. Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg: Association for Computational Linguistics, 2006. 317−326.

[14] Bod R. Is the end of supervised parsing in sight? In: Proc. of the Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2007. 400−407.

[15] Collins M. Head-Driven statistical models for natural language parsing. In: Proc. of the Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2003. 590−637.

[16] Bikel DM. On the parameter space of generative lexicalized statistical parsing models [Ph.D. Thesis]. Philadelphia: University of Pennsylvania, 2004.

[17] Kneser R, Ney H. Improved backing-off for *M*-gram language modeling. In: Proc. of the IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. 1995. 181−184.

[18] Bikel DM, Chiang D. Two statistical parsing models applied to Chinese Treebank. In: Proc. of the 2nd Chinese Language Processing Workshop. Stroudsburg: Association for Computational Linguistics, 2000. 1−6.

刘水(1981－),男,黑龙江哈尔滨人,博士生,主要研究领域为句法分析,机器翻译.

赵铁军(1962－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为自然语言处理,机器翻译.

李生(1943－),男,教授,博士生导师,CCF 会员,主要研究领域为人工智能,自然语言处理.

刘鹏远(1974－),男,博士,讲师,主要研究领域为词义消歧,机器翻译.