极大似然优化EM算法的汉语分词认知模型

赵 越[1,2,3],李 红[1,3]
(1.辽宁师范大学 脑与认知神经科学研究中心,辽宁 大连 116029;2.辽宁师范大学 文学院,辽宁 大连 116029;
3.深圳大学 心理与社会学院,广东 深圳 518060)

摘 要:针对标准EM算法在汉语分词的应用中还存在收敛性能不好、分词准确性不高的问题,本文提出了一种基于极大似然估计规则优化EM算法的汉语分词认知模型,首先使用当前词的概率值计算每个可能切分的可能性,对切分可能性进行"归一化"处理,并对每种切分进行词计数,然后针对标准EM算法得到的估计值只能保证收敛到似然函数的一个稳定点,并不能使其保证收敛到全局最大值点或者局部最大值点的问题,采用极大似然估计规则对其进行优化,从而可以使用非线性最优化中的有效方法进行求解达到加速收敛的目的。仿真试验结果表明,本文提出的基于极大似然估计规则优化EM算法的汉语分词认知模型收敛性能更好,且在汉语分词的精确性较高。

关键词:EM算法;汉语分词认知;收敛性优化;极大似然估计规则;归一化处理
DOI:10.13774/j.cnki.kjtb.2016.04.040
中图分类号:TP391.1 **文献标识码**:A **文章编号**:1001-7119(2016)04-0178-04

Chinese Word Segmentation Cognitive Model Based on Maximum Likelihood Optimization EM Algorithm

Zhao Yue[1,2,3], *Li Hong*[1,3]
(1.Brain and Cognitive Neuroscience Reserch Center, Liaoning Normal University, Dalian Liaoning 116029, China;
2.School of Chinese Language and Literature, Liaoning Normal University, Dalian Liaoning 116029, China;
3.College of Psychology and Sociology, Shenzhen University, Shenzhen Guangdong 518060, China)

Abstract: In view of bad convergence and inaccurate word segmentation of standard EM algorithm in Chinese words segmentation, this paper put forward a cognitive model based on optimized EM algorithm by maximum likelihood estimation rule. Firstly, it uses the probability of current word to calculate the possibility of each possible segmentation and normalize them. Each segmentation is counted by words. Standard EM algorithm cannot make sure converging to a stable point of likelihood function, and converging to a global or local maximum point. Therefore, the maximum likelihood estimation rule is adopted to optimize it so as to use an effective method in nonlinear optimization and accelerate the convergence. the simulation experiments show that the optimized EM algorithm by maximum likelihood estimation rule has better convergence performance in the Chinese words cognitive model and more accurate in the words segmentation.

Keywords: EM algorithm; Chinese words cognition; convergence optimization; maximum likelihood estimation rule; normalization

收稿日期:2014-12-14
基金项目:辽宁师范大学青年科研项目。
作者简介:赵越(1984-),女,汉族,辽宁,讲师,硕士,语言认知发展。

0　引言

随着信息技术的不断发展,计算机能够存储可读的文字信息[1]。计算机需要通过提取关键知识达到中文信息小型化的目的,即把中文信息按照词语进行划分,然后通过计算机的词频统计、文本分类和知识发现获得用户提交的文字信息。由此可以看出词语的切分是汉语文本处理的关键技术[2]。汉语分词在搜索技术领域中非常重要,并且在机器释义、机器翻译、句群划分、篇章理解、人机对话和情报检索等计算机运用中有着重要意义[3]。

汉语分词备受国内外学者的关注,并且在研究中也取得了一定的效果。Xue等学者提出四种标签,分别是"LL","RR","MM"以及"LR",他们通过研究字符与词的位置关系总结了这四种标记。按照这四种标签可以有效的得出分词结果[4]。Peng等学者在这四种标签的基础上进行了优化,在特征模板的基础上结合了领域词典,并将最大熵模型由条件随机域(CRF, conditional random field)模型代替,这样的分词方法是目前最为理想的,在原有的基础上获得了质的飞跃[5]。Ge等学者利用EM算法以及迭代算法,从生语料中不断学习词频改善分词结果[6]。Peng等学者相继提出了建立在EM算法之上的的无监督分词,通过用户信息与结果的比对,过滤分词结果,随后Peng等学者利用分层的方式对无监督分词做了研究,通过EM算法在生语料中学习获得词的片段,并进一步获得完整的词[7]。邻接字变化数(accessor variety)这个重要指标也是Peng等学者提出的,它可以计算出字串成词的概率,配合动态规划算法获得分词结果[8]。分支信息熵(branching entropy)的定义是由Jin等学者提出的,词的边界通过它在句中的极值点来计算[9]。Chen等学者在PageRank算法中获得启发建立了WordRank算法,它通过词类别成网页计算字串成词的概率,然后根据最优化方法检索出最好的分词结果[10]。Magistry等学者将分支信息熵进行规范,用规范化后的变化率计算字串成词的概率,并通过维特比解码方式对汉语分词进行转换求解[11]。

本文通过研究网球运动员目标跟踪的特性,并基于极大似然估计规则,提出了优化EM算法的汉语分词认知模型,对原算法的收敛性能进行改进,以此来加强汉语分词结果的可靠性。

1　基于EM算法的汉语分词模型

EM算法的实质是迭代,是用来处理后验分布的众数,它是DLR提出的算法。基于EM算法的汉语分词模型的详细步骤如下:

(1)对于还未做出处理的句子,通过计算得出句中所有可切分词语的概率,使用"归一化"的方法将可切分的词语定义为"尾数",使所有尾数相加和为1;并在每一次切分后进行计数,也可以理解为"尾数"与词数的相加。

(2)通过词数更新词的概率;

(3)重复(1)(2)步直至收敛。

一种简便的计数方法在EM分词算法中被使用,以标点符号作为分割点,两个标点符号间的文本即为句子。例如,句子长为n,句子切分成词语的方式就有2^{n-1}种。2^{n-1}种切分方式中仅一种是正确的,因此要根据切分的概率对每种可能性进行估算。"软计数"这种计数方法,即某种切分方式的概率为p_l,则$\frac{p_l}{\sum_{j=1}^{2^{n-1}} p_j}$表示每个词增加的词数。

"软计数"的实现过程如下:

输入的句子定义为$CC_1,...,CC_n$,句子中的每个词语定义为w_l,计数增加量定义为$S_i^{left}p(w_i)S_i^{right}/\alpha$,其中:

(1)w_i左侧的子字符串可能切分的概率和定义为S_i^{left},同理,右侧的子字符串可能切分的概率和定义为S_i^{right};

(2)w_i的概率值定义为$p(w_i)$;

(3)定义一个归一化常数α,它是句子所有可能性切分的概率和,即与S_{n+1}^{left}相等。

将S_i^{left},S_i^{right}通过动态程序进行计算。则S_i^{left}的递归函数如公式(1):

$$S_i^{left}=\begin{cases}1 & i=1\\ p(w_i) & i=2\\ \sum p(CC_j...CC_{i-1})S_j^{left} & i>2\end{cases} \tag{1}$$

首先计算$S_i^{left}(i=1,...,n)$,从左到右扫描解得$\alpha=S_{n+1}^{left}$。再从右向左扫描计算$S_i^{right}(i=n,...,1)$,以此获得每个词的数量。

本文建立在以上算法的基础上,通过吉布斯

不等式对EM算法进行收敛性的研究，不等式如(2)：

$$\log x \leqslant x-1,(x>0) \tag{2}$$

当且仅当 $x=1$ 时，等号成立。EM算法的行收敛性结果如图1所示。

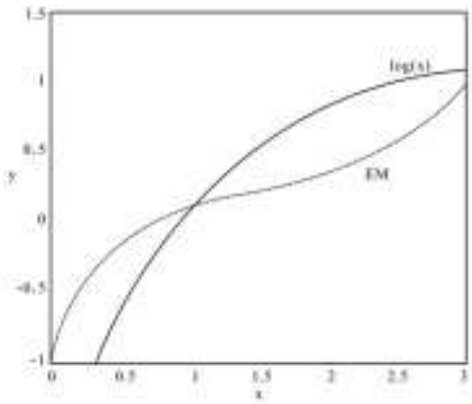

图1 EM算法的行收敛性结果

Fig.1 Line EM algorithm convergence results

从图中得出，EM算法计算的估计值在似然函数稳定点收敛，并非一定收敛在似然函数最大值点。

2 基于极大似然优化的EM算法

2.1 极大似然估计规则

极大似然估计方法是目前处理点估计问题有效且常用的手段之一，即使用多个估计值计算实际值的方法。

参数的点估计即由某个数值对参数进行估算。点估计过程如下：定义随机变量 X 的概率密度函数为 $f(x|\theta)$，并假设其已知，定义需估算的参数为 θ。定义 $X_1,X_2,...,X_n$ 为 X 的已知样本，则 $x_1,x_2,...,x_n$ 是 $X_1,X_2,...,X_n$ 的对应样本值。点估计问题的本质即建立合适的统计量 $\theta(X_1,X_2,...,X_n)$，用已知估计值 $\theta(x_1,x_2,...,x_n)$ 当作未知参数 θ 的一个近似估计值。定义 $\hat{\theta}(X_1,X_2,...,X_n)$ 为 θ 的估计量，则 $\hat{\theta}(x_1,x_2,...,x_n)$ 是 θ 的估计值，并简单记为 $\hat{\theta}$。

当 $x_1,x_2,...,x_n$ 相互独立，并且其联合条件概率密度函数的分布如公式(3)：

$$\begin{aligned} f(x_1,x_2,...,x_n|\theta)&=f(x_1|\theta)f(x_2|\theta)..f(x_n|\theta)\\ &=\prod_{i=1}^{n}f(x_i|\theta)\end{aligned} \tag{3}$$

如果 $x=[x_1,x_2,...,x_n]$，公式(3)简化得公式(4)：

$$f(x|\theta)=\prod_{i=1}^{n}f(x_i|\theta) \tag{4}$$

将公式(3)与(4)作为参数 θ 的似然函数，并记作 $L(\theta|x)$。如公式(5)：

$$L(\theta|x)=f(x|\theta)=\prod_{i=1}^{n}f(x_i|\theta) \tag{5}$$

定义对数似然函数 $l(\theta|x)$，则得到公式(6)：

$$\begin{aligned} l(\theta|x)&=\ln L(\theta|x)=\ln f(x|\theta)=\ln\prod_{i=1}^{n}f(x_i|\theta)\\ &=\sum_{i=1}^{n}\ln\prod_{i=1}^{n}f(x_i|\theta)\end{aligned} \tag{6}$$

选取参数 θ 的估计值

$$\hat{\theta}=(\hat{\theta}_1,\hat{\theta}_2,...,\hat{\theta}_k) \tag{7}$$

使似然函数 $L(\theta|x)$ 为最大值。记作：

$$L(x_1x_2,...,x_n;\hat{\theta})=\max_{\theta\in\Theta}L(\theta|x)=\max_{\theta\in\Theta}l(\theta|x) \tag{8}$$

其中：Θ 为参数 θ 的取值范围。并称 $\hat{\theta}$ 为 θ 的极大似然估计值。

由此可知，求出似然函数 $L(\theta)$ 的最大值即可得到总体参数 θ 的极大似然估计值 $\hat{\theta}$。通过微积分可知，如果似然函数 $L(\theta)$ 在 θ 上有连续偏导数，那么极大似然估计值 $\hat{\theta}$ 可由公式(9)求解：

$$\frac{\partial L(\theta)}{\partial\theta_i}=0,i=1,2,...,k \tag{9}$$

又因为 $l(\theta)=\ln L(\theta)$ 与 $L(\theta)$ 同时得到最大值，所以等价替换后由公式(9)可求 $\hat{\theta}$。

$$\frac{\partial l(\theta)}{\partial\theta_i}=0,i=1,2,...,k \tag{10}$$

2.2 基于极大似然优化的EM分词算法

EM汉语分词算法将 θ_0 定义为 θ 的初始值，再进行以下两个步骤的计算：

(1)将完备数据对数似然函数 $\ln f(x,y|\theta)$ 在 Y 的条件下求期望，消除 Y，即得公式(11)：

$$Q(\theta,\theta_k)=E[\ln f(x,y|\theta)|x,\theta_k] \tag{11}$$

(2)求 $\theta_{k+1}\in\Theta$，使 $Q(\theta,\theta_k)$ 极大化，即得(12)：

$$Q(\theta_{k+1}|\theta_k,X)=\max_{\theta}Q(\theta|\theta_k,X) \tag{12}$$

上式中的 X 是观测数据集。

$\theta_k\rightarrow\theta_{k+1}$ 即为一次迭代，如此以来，只需将步骤(1)和(2)反复进行迭代运算，直至 $\|\theta_{k+1}-\theta_k\|$ 或者 $\|Q(\theta_{k+1},\theta_k)-Q(\theta_k,\theta_k)\|$ 完全消失，则停止运算。

在步骤(1)中求 θ_{k+1}，可通过方程组(13)求得答案：

$$\frac{\partial Q(\theta,\theta_k)}{\partial\theta}=0 \tag{13}$$

为了提升EM算法的收敛速度,在极大似然估计规则的基础上对其进行改进:

$$\left.\frac{\partial Q(\theta,\theta_k)}{\partial\theta}\right|_{\theta=\theta_k}=g(\theta_k) \tag{14}$$

公式(14)中的 $g(\theta_k)$ 表示在 θ_k 处不完全数据对数似然函数 $L(\theta)$ 的梯度:

$$g(\theta_k)=\nabla L(\theta)\big|_{\theta=\theta_k} \tag{15}$$

因此问题可以转变成求 θ_*,使 $g(\theta_*)=0$,并且提升收敛速度可以通过非线性最优化中的有效方法进行求解。

3　算法性能仿真

本文通过仿真实验对前面提出的优化算法进行验证,以此来证明算法的有效性。第一步,对建立在极大似然估计规则上的EM算法进行收敛性分析,如图2所示:

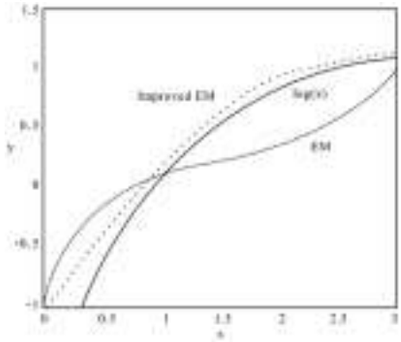

图2 改进算法的收敛性分析

Fig.2 Improved convergence analysis algorithms

第二步,在汉语分词中运用建立在极大似然估计规则上的EM算法;第三步,分析其效果,将结果进行比对,如图3所示:

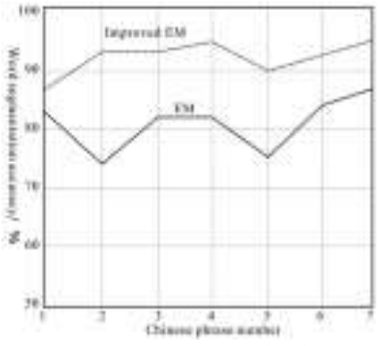

图3 汉语分词的准确性比较

Fig.3 To compare the accuracy of Chinese word segmentation

从图3中可以得知,优化后的EM算法在仿真结果中汉语分词的准确性明显高于原EM算法,证明建立在极大似然估计规则上的EM算法收敛性较高。

4　总结

汉语分词即通过提前设定的分类方式将未分类的文本经过相应的规则进行自动分类。这之中包括了计算机语言学、人工智能、信息学、数据挖掘等多个专业的运用,它是自然语言处理的关键领域。本文提出了一种建立在极大似然估计规则之上的优化EM算法汉语分词认知模型,它提高了EM算法的收敛性,使汉语分词更加准确有效,并且在仿真实验结果中得到验证。

参考文献:

[1] Turdi Tohti.Semantics-based Feature Extraction and Its Application in Uyghur Text Classification[J].Journal of Chinese Information Processing,2014,28(4):140-144.

[2] Zhang B Y.Chinese word segmentation algorithm based on pair coding[J].Journal of Nanjing University of Science and Technology(Nature Science),2014,4:526-530.

[3] Zhang J.Research of the Word Segmentation for Chinese Patent Claims[J].New Technology of Library and Information Service,2014,9:91-98.

[4] 项炜.基于词频学习和动态词频更新的自动分词系统设计[J].计算机应用与软件,2014,31(5):106-109.

[5] Qian Z Y.Research on Automatic Word Segmentation and Pos Tagging for Chu Ci Based on HMM[J].Library and Information Service,2014,58(4):105-110.

[6] 颜端武.基于N—gram复合分词的领域概念自动获取方法研究[J].情报理论与实践,2014,37(2):122-126.

[7] 战学刚.基于TF统计和语法分析的关键词提取算法[J].计算机应用与软件,2014,31(1):47-49.

[8] Guan X J.Academic Paper Translation System Based on Different Word Segmentation Frame[J].Journal of Xiamen University(Natural Science),2013,52(6):781-786.

[9] Yang W C.Research of an improved algorithm for Chinese word segmentation dictionary based on double-array Trie-tree[J].Computer Engineering & Science,2013,35(9):127-131.

[10] Cao Z Q.Unsupervised Chinese Word Segmentation Based on HDP and Mutual Information Getting together [J].Journal of Chinese Information Processing,2013,27(6):1-5.

[11] Liang S.Improvement of Chinese Word Segmentation Based on Combination Method[J].Journal of Nanjing University of Posts and Telecommunications,2013,33(6):112-117.