中心词驱动句法分析中的平滑技术

袁里驰
(江西财经大学信息学院数据与知识工程江西省重点实验室,江西南昌 330013)

摘 要: 解决数据稀疏问题是中心词驱动句法分析中的一个重要问题,基于词类的统计语言模型是解决统计模型数据稀疏问题的重要方法.本文在分析经典平滑算法的基础上,提出一种基于语义依存信息和互信息的词聚类算法,并利用绝对权重差分方法构造了一种可变长语言模型,即根据历史词对当前词预测所作的贡献不同,n 值的大小也随之变化.进而提出了一种基于语义类和可变长模型的中心词驱动句法分析改进模型,既增强了句法分析模型的消歧能力,又解决了严重的数据稀疏问题.改进模型性能有了明显的提高,精确率和召回率分别为 84.53% 和 82.41%,综合指标 F 值比 Collins 的中心词驱动句法分析模型提高了 2.02 个百分点.

关键词: 句法分析模型; 平滑算法; 中心词驱动句法分析; 聚类算法
中图分类号: TP391.1 **文献标识码:** A **文章编号:** 0372-2112 (2013)07-1337-06
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2013.07.015

Smooth Technologies in Head-Driven Parsing

YUAN Li-chi
(*School of Information Technology, Jiangxi University of Finance & Economics, Nanchang, Jiangxi* 330013, *China*)

Abstract: Solving the data sparseness problem is an important problem about head-driven parsing, cluster-based statistic language model is an important method to solve the problem of sparse data. Based on the analysis of the classical smoothing technology, this paper proposes a word clustering algorithm by utilizing mutual information and semantic dependency, and an absolute weighted difference method was presented and was used to construct vari-gram language model which has good predictable ability, then proposes an improved head-driven parsing model based on word cluster and vari-gram model. Experiments are conducted for the refined statistical parser, it achieves 84.53% precision and 82.41% recall, F measure is improved 2.02% comparing with the head-driven parsing model introduced by Collins.

Key words: parsing model; smoothing algorithm; head-driven parsing; clustering algorithm

1 引言

在信息化社会中,语言信息处理的技术水平和每年所处理的信息总量已成为衡量一个国家现代化水平的重要标志之一.并且随着计算机硬件的快速发展,计算机越来越广泛地进入到我们的日常生活中来,计算机与自然语言相结合的领域也越来越广阔.自然语言理解作为语言信息处理技术的一个高层次的重要方向,一直是人工智能界所关注的核心课题之一.句法分析在自然语言处理领域中具有十分重要的地位,同时它也是公认的一个研究难题.句子分析上接篇章理解,下联词汇分析,起着承上启下的作用.词汇分析是基础,句子分析是中心,篇章理解是最终目的.那么,一旦得到了句子成分的计算机表示,无论是应用于句群划分、篇章理解,还是机器翻译、机器释义、人机对话或是情报检索等等方面,都有着实际意义.

自上世纪 90 年代以来,随着语料资源的获取变得容易,基于统计的方法[1~6]开始在自然语言处理领域成为主流.这种方法采用统计学的处理技术从大规模语料库中获取语言分析所需的知识,放弃人工干预,减少对语言学家的依赖.其基本思想是:(1)使用语料库作为唯一的信息源,所有的知识(除了统计模型的构造方法)都是从语料库中获得;(2)语言知识在统计意义上被解释,所有参数都是通过统计处理从语料库中自动学习的.

基于统计的方法具有效率高、鲁棒性好的优点,大量的实验已经证明了该方法的优越性.目前,统计方法已经被句法分析的研究者普遍采用.为进行统计句法分析,首先要遵循某一语法体系,根据该体系的语法确定

收稿日期:2012-05-22;修回日期:2012-09-27;责任编辑:郭游
基金项目:国家自然科学基金(No.61262035);江西省自然科学基金(No.2010GZS0072,20122BAB201033);江西省教育厅科技项目(No.GJJ12271)

语法树的表示形式.目前,在句法分析中使用比较广泛的有短语结构语法和依存语法.

当前短语结构句法分析普遍基于概率上下文无关文法(Probabilistic Context Free Grammar,PCFG).在早期研究工作中,基于上下文无关文法的短语结构句法分析方法直接从人工标注的树库中读取文法规则,并以相对频率作为规则的概率[7].这类方法实现简单,但是先前的研究工作表明这种方法的性能并不理想.其主要原因在于上下文无关文法中的独立性假设,而这些独立性假设在实际情况中往往并不成立.有两种不同的思路可以放宽上下文无关文法中的独立性假设:

(1)对句法树中原有的非终结符标记进行重新标注——拆分或者合并.这个研究方向的出发点在于现有的句法类别体系中,有些非终结符标记导致过强的独立性假设,而有些标记却使独立性假设过弱.理论上,更加合理的数据标注可以使上下文无关文法中的独立性假设相应地变得合理.

(2)在文法规则中引入词汇的信息,即在句法树的每个非终结符节点上标注词汇信息,利用词汇信息放宽上下文无关文法的独立性假设.Magerman 最先开展了这个方向的研究工作,论证了词汇信息的有效性[8].Charniak 和 Collins 随后分别推进了这一方向的研究[9~12].

Collins 提出的中心词驱动句法分析模型[12]是当前句法分析的主流模型,其基本思想就是在上下文无关文法规则中引入词汇化信息和短语的中心词信息,这两种信息的引入,增强了句法分析模型的消歧能力,然而却不可避免地带来了严重的数据稀疏问题.所以,解决数据稀疏问题是中心词驱动句法分析中的一个重要问题.回退平滑(back-off)和插值平滑(interpolation)是两种主要的平滑技术,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.上述两种平滑技术未能将语言学知识与数学模型结合起来,因而其平滑效果仍有待进一步提高.基于词类的统计语言模型是解决统计模型数据稀疏问题的另一重要方法,本文利用互信息定义词相似度,基于相似度,提出了一种自下而上的分层聚类算法.结合插值平滑技术,基于词类的统计语言模型将语言学知识融入数学模型,成功解决了中心词驱动句法分析模型引入词汇信息所带来的数据稀疏问题.

2 中心词驱动句法分析模型及其改进

2.1 中心词驱动句法分析模型的基本原理

中心词驱动句法分析模型是最具有代表性的词汇化模型.为了发挥词汇信息的作用,中心词驱动模型为文法规则中的每一个非终结符(none terminal)都引入核心词/词性信息.由于引入词汇信息,将不可避免地出现严重的稀疏问题.为了缓解这个问题,中心词驱动模型把每一条文法规则的右手侧分解为三大部分,分别为:一个中心成分;若干个在中心左边的修饰成分;若干个在中心右边的修饰成分.可以写成如下形式:

$$P(ht,hw)-L_m(lt_m,lw_m)\cdots L_1(lt_1,lw_1)H(ht,hw)\\ R_1(rt_1,rw_1)\cdots R_n(rt_n,rw_n) \tag{1}$$

其中,P 为非终结符,H 表示中心成分,L_1 表示左边修饰成分,R_1 表示右边修饰成分.hw,lw,rw 均是成分的核心词,ht,lt,rt 分别是它们的词性.在 Collins 提出的中心词驱动句法分析模型中,进一步假设,首先由 P 产生核心成分 H,然后以 H 为中心分别独立地产生左右两边的所有修饰成分.这样,形如式(1)的文法规则的概率为:

$$P_h(H\mid P,h)\cdot\prod_{i=1}^{m+1}P_i(L_i(lt_i,lw_i),c,p\mid H,P,h,\Delta_l(i-1))\cdot\prod_{i=1}^{n+1}P_i(R_i(rt_i,rw_i),c,p\mid H,P,h,\Delta_r(i-1)) \tag{2}$$

其中,c 和 p 分别代表并列符号、标点符号,L_{m+1} 和 R_{n+1} 分别为左右两边的停止符号,$\Delta_l(i-1)$ 为距离函数,补偿结构信息的缺失.距离信息考虑了三种情况:(1)该成份前是否有成份;(2)该成份前是否出现动词;(3)该成份前是否出现有标点符号.

Charniak 提出了另一种中心词驱动句法分析统计模型,每一条词汇化文法规则的计算分为两个步骤:首先生成整个上下文无关规则,然后填入词汇化的词语.这样,形如式(1)的文法规则的概率为:

$$\begin{aligned}&P(L_n(l_n)\cdots L_1(l_1)H(h)R_1(r_1)\cdots R_m(r_m)\mid P,h)\\&=P(L_n\cdots L_1HR_1\cdots R_m\mid P,h)\times\prod_{i=1}^{n}P_l(l_i\mid L_i,P,h)\\&\quad\times\prod_{j=1}^{m}P_r(r_j\mid R_j,P,h)\end{aligned} \tag{3}$$

2.2 中心词驱动句法分析模型的改进

在我们的中心词驱动句法分析模型中,以 H 为中心产生左右两边的所有修饰成分,但这些修饰成分不是相互独立地.这样,形如式(1)的文法规则的概率为:

$$P_h(H\mid P,h)\prod_{i=1}^{m+1}P_i(L_i(lt_i,lw_i),c,p\mid L_{i-1}(lt_{i-1},lw_{i-1}),\cdots,L_1(lt_1,lw_1),P,H,h)\prod_{i=1}^{n+1}P_i(R_i(rt_i,rw_i),c,p\mid R_{i-1}(rt_{i-1},rw_{i-1}),\cdots,R_1(rt_1,rw_1),P,H,h) \tag{4}$$

其中,L_{m+1} 和 R_{n+1} 分别为左右两边的停止符号,其他符号的表示同上文一致.假定 lt_i,lw_i 仅与 L_i、lw_{i-1},…,lw_1 和 P,H,h 有关,L_i 仅与 L_{i-1},…,L_1 和 P,H,h 有关,则有:

$$P_i(L_i(lt_i,lw_i),c,p\mid L_{i-1}(lt_{i-1},lw_{i-1}),\cdots,L_1(lt_1,lw_1),$$

$$P, H, h) =$$
$$P(lt_i, lw_i \mid L_i, lw_{i-1}, \cdots, lw_1, P, H, h,) \cdot P(L_i, c, p \mid L_{i-1}, \cdots, L_1, P, H, h) \quad (5)$$

进一步假定 lt_i, lw_i 仅与 L_i 和 P, H, h 有关,$L_{i-1}, \cdots, L_1$ 和 $P, H, h, \Delta_l(i-1)$关于 L_i 条件独立,则式(5)中的概率有:

$$P(lt_i, lw_i \mid L_i, lw_{i-1}, \cdots, lw_1, P, H, h)$$
$$= P(lt_i \mid lw_i) \cdot P(lw_i \mid L_i, lw_{i-1}, \cdots, lw_1, P, H, h)$$
$$\approx P(lt_i \mid lw_i) \cdot P(lw_i \mid L_i, P, H, h) \quad (6)$$

$$P(L_i \mid L_{i-1}, \cdots, L_1, P, H, h) \approx \frac{P(L_i, c, p \mid L_{i-1}, \cdots, L_1) \cdot P(L_i, c, p \mid P, H, h, \Delta_l(i-1))}{P(L_i, c, p)} \quad (7)$$

再假定 $L_i, L_{i-1}, \cdots, L_1$ 是阶数为 k 的马尔可夫链,则有

$$P(L_i, c, p \mid L_{i-1}, \cdots, L_1, P, H, h) \approx \frac{P(L_i, c, p \mid L_{i-1}, \cdots, L_{i-k+1}) \cdot P(L_i, c, p \mid P, H, h, \Delta_l(i-1))}{P(L_i, c, p)} \quad (8)$$

3　插值平滑算法

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,用如下形式[13]表示统计事件:若统计事件有属性 $x_1, x_2, \cdots, x_n$,则统计空间的事件用这些属性的笛卡尔乘积 $x = x_1 \times x_2 \times \cdots \times x_n$ 表示,统计空间的维度 $n = |x|$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[14].回退平滑应用折扣(discount)在计算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[15]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,形如式(9).

$$\tilde{p}^n = \lambda \hat{p}^n + (1-\lambda) \tilde{p}^{n'} \quad (9)$$

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 $0 \leqslant \lambda \leqslant 1, n > n'$.

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 领域中一种得到广泛应用的平滑技术.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[11]: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.

插值平滑算法[16](以下简称 Witen-Bell 平滑)和插值平滑算法[17](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,形如后文的式(11).首先,式(10)给出对于统计条件极大似然统计概率空间样本种类数 $D(\boldsymbol{X})$的概念:

$$D(\boldsymbol{X}) = |Y(\boldsymbol{X})| \quad (10)$$

其中,$\boldsymbol{X}$ 表示条件极大似然统计中统计事件的集合,且 $\boldsymbol{Y}(\boldsymbol{X}) = \{\boldsymbol{x} \mid \mathrm{Count}(\boldsymbol{x}, \boldsymbol{y}) > 0\}$表示条件似然空间中统计事件数目大于 0 的事件的集合,$\boldsymbol{x}$ 表示条件极大似然统计中的统计事件,$\boldsymbol{y}$ 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,$\boldsymbol{x}$, $\boldsymbol{y}$ 均为向量),$|\boldsymbol{Y}(\boldsymbol{X})|$表示集合 $\boldsymbol{Y}(\boldsymbol{X})$的事件数目.

$$\lambda = \frac{|\boldsymbol{X}|}{|\boldsymbol{X}| + D(\boldsymbol{X})} \quad (11)$$

其中,$|\boldsymbol{X}|$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(如式(12))在插值平滑公式中引入优化变量:

$$\lambda = \frac{|\boldsymbol{X}|}{|\boldsymbol{X}| + C \times D(\boldsymbol{X})} \quad (12)$$

其中,C 为常数,通过对系统性能的反馈优化而获得.

4　基于类的中心词驱动句法分析模型

4.1　基于词相似度的分层聚类算法

聚类算法[18]有很多种,但可归结为两种基本类型:层次聚类与非层次聚类.非层次聚类只是简单的包括了每类的数量,类与类之间的关系不确定.层次聚类的每一个节点是其父节点的一个子类,叶节点对应的是类别中每个单独的对象,常用算法有:自下向上与自上向下(凝聚与分裂).

传统的统计聚类方法通常基于贪婪原则,以语料的似然函数或困惑度作为判别函数.这种传统方法的主要缺点是聚类速度慢,初值对结果影响大,易陷入局部最优.而我们提出的分层聚类算法[19]基于词的相似度,因此我们首先要找到一种可靠的,适于计算的词与词间相似度的定量标准.基于语料库的统计方法通常认为一个词的意义与其所处的上下文中出现的其它词有关,也即语言环境有关.如果两个词在语料库中所处的语言环境总是非常相似,我们就可以认为这两个词彼此之间非常相似[20].

假定词 w_1 与词 w_2 相似,则可推断这两个词与其

它词的互信息也是相似的,现在我们可以定义两个词 w_1, w_2 之间的相似度如下:

$$\mathrm{sim}(w_1, w_2) = \frac{\sum_w P(w)[\min(I(w, w_1), I(w, w_2)) + \min(I(w_1, w), I(w_2, w))]}{\sum_w P(w)[\max(I(w, w_1), I(w, w_2)) + \max(I(w_1, w), I(w_2, w))]} \tag{13}$$

其中 $I(w_i, w_j)$为相邻词对 w_i, w_j 之间的互信息:

$$I(w_i, w_j) = \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \tag{14}$$

这里 $p(w_i), p(w_j)$分别为词 w_i 和 w_j 在训练语料中出现的概率,$p(w_i, w_j)$是联合概率,由式(13)知,w_1, w_2 与它们的左右近邻之间互信息差别越小,两词的相似度也越高,因此这种定义是合理的.

基于词相似度,词类 C_1, C_2 之间的相似度定义如下:

$$\mathrm{sim}(C_1, C_2) = \frac{\sum_{w_i \in C_1, w \in C_2} C(w_i)\, C(w_j)\, \mathrm{sim}(w_i, w_j)}{\sum_{w \in C_1} C(w_i) \sum_{w_j \in C_2} C(w_j)} \tag{15}$$

其中 $C(w_i), C(w_j)$分别表示词 w_i 与 w_j 在语料中出现的数量.

其次,我们可以根据语义依存关系和语法特性对词聚类[21].设 w_1, w_2 是具有语义依存关系 rel 的词对,用三元组(w_1, rel, w_2)表示词对和它们之间的依存关系.则词对(w_1, w_2)在依存关系 rel 下的互信息定义为:

$$I_{rel}(w_1, w_2) = \log \frac{p(w_1, w_2 \mid rel)}{p(w_1 \mid rel)\, p(w_2 \mid rel)} \tag{16}$$

其中 $p(w_1, w_2 \mid rel) = \dfrac{p(w_1, rel, w_2)}{p(rel)}$

这里计算要用到的概率使用极大似然估计(Maximum Likeihood Estimation)的方法统计:

$$p(w_1, rel, w_2) = \frac{\mathrm{Count}(w_1, rel, w_2)}{\mathrm{Count}(*, *, *)} \tag{17a}$$

$$p(w_1 \mid rel) = \frac{\mathrm{Count}(w_1, rel, *)}{\mathrm{Count}(*, rel, *)} \tag{17b}$$

$$p(w_2 \mid rel) = \frac{\mathrm{Count}(*, rel, w_2)}{\mathrm{Count}(*, rel, *)} \tag{17c}$$

$$p(rel) = \frac{\mathrm{Count}(*, rel, *)}{\mathrm{Count}(*, *, *)} \tag{17d}$$

其中 * 表示可能的词或依存关系,因而有

$$I_{rel}(w_1, w_2) = \log \frac{\mathrm{Count}(w_1, rel, w_2)\mathrm{Count}(*, rel, *)}{\mathrm{Count}(w_1, rel, *)\, p(*, rel, w_2)} \tag{18}$$

定义 1 词对 w_1, w_2 在依存关系 rel 下的相似度定义为

$$\mathrm{sim}_{rel}(w_1, w_2) = \frac{\sum_w P(w)\min(I_{rel}(w, w_1), I_{rel}(w, w_2))}{\sum_w P(w)\max(I_{rel}(w, w_1), I_{rel}(w, w_2))} \tag{19}$$

定义 2 词对 w_1, w_2 之间的相似度则定义为:

$$\mathrm{sim}(w_1, w_2) = \sum_{rel} p(rel)\, \mathrm{sim}_{rel}(w_1, w_2) \tag{20}$$

整个聚类算法的流程如下所示:

a.计算词对之间的相似度.

b.初始化,词表中的每个词各代表一类,共 N 类,(N 为词表中词的数量).

c.找出具有最大相似度的两个词类,将这两个词类合并成一个新的词类.

d.计算刚合并词类与其它词类的相似度.

e.检查是否达到结束条件(词类之间最大相似度小于某个预先决定的门槛值,或是词类的数目达到了要求),是,程序结束;否则,转 *c*.

4.2 基于语义类的中心词驱动句法分析模型

设 $C(h)$表示中心词 h 所在的词类,$C(lw_i)$表示词 lw_i 所在的词类,则式(6)、(8)中的有关概率可以用下面的插值平滑方法计算:

$$P(lw_i \mid L_i, P, H, h) \approx \lambda P(lw_i \mid L_i, P, H, h) + (1-\lambda) P(C(lw_i) \mid L_i, P, H, C(h)) \tag{21}$$

$$P(L_i, c, p \mid P, H, h, \Delta_l(i-1)) \approx \lambda P(L_i, c, p \mid P, H, h, \Delta_l(i-1)) + (1-\lambda) P(L_i, c, p \mid P, H, C(h), \Delta_l(i-1)) \tag{22}$$

其中 $0 \leqslant \lambda \leqslant 1$ 为平滑参数.

5 基于可变长模型的中心词驱动句法分析平滑模型

基于类的 n-gram 模型已经被证明是一种能有效解决基于词的模型所存在的数据稀疏问题的方法,但该方法牺牲了一部分预测能力.为解决这一问题,人们提出可变长语言模型,即根据历史词对当前词预测所作的贡献不同,n 值的大小也随之变化.本文提出了一种绝对权重差分方法,并用这种方法针对式(8)中的概率 $P(L_i, c, p \mid L_{i-1}, \cdots, L_{i-k+1})$构造了一种可变长语言模型.

设 L 表示当前修饰成分,L_g 表示对应修饰成分的历史(其中 g 是历史的长度);L_{new}表示在 L_g 前的扩展历史.现在我们用绝对权重差分方法测量用 $P(\cdot \mid L_g, L_{\mathrm{new}})$替代分布 $P(\cdot \mid L_g)$后的变化:

$$\Delta_{\mathrm{diff}} = \sum_C [\mathrm{Num}(L, L_g, L_{\mathrm{new}}) - D(\mathrm{Num}(L, L_g, L_{\mathrm{new}}))] \times \left| \log P(L \mid L_g, L_{\mathrm{new}}) - \log P(L \mid L_g) \right| \tag{23}$$

其中 $\mathrm{Num}(L, L_g, L_{\mathrm{new}})$表示序列 L, L_g, L_{new}在训练语料中出现的次数,$D(\mathrm{Num}(L, L_g, L_{\mathrm{new}}))$是折扣函数.

构造基于类的可变长语言模型的算法如下:

Step 1 初始化:$g = -1$

Step 2　$g = g + 1$

Step 3　增加:对第 $g-1$ 层所有的节点,向第 g 层扩展,即将训练语料中出现的$(g+1)$元词类增加到文法树中已存在的 g 元词类后面.

Step 4　剪除:对第 g 层每一个新增加的叶结点计算 Δ_{diff},如果 Δ_{diff}的值小于预先确定的门槛值,则去掉该节点.

Step 5　结束:如果第 g 层新增节点全部被裁掉或达到规定 的 层数,结束;否则,转步骤 2.

6　实验结果

试验数据取自宾州中文树库(CHTB)5.0 版本,大部分取材于新华社新闻,Sinorama 新闻杂志以及香港新闻.CTB 是由语言数据联盟(LDC)公开发布的一个语料库,为汉语句法分析研究提供了一个公共的训练、测试平台.该树库包含了 507222 个词,824983 个汉字,18782 个句子,有 890 个数据文件.为了在训练集、开发集和测试集中平衡各种语料来源,我们将语料分割如下:我们将文件 301 ~ 320、611 ~ 630 作为调试集,将文件 271 ~ 300、631 ~ 660 作为测试集,其余的 790 个文件作为训练集,训练集以 390 个文件为起点,每次添加 100 个文件重新训练句法分析模型.本文的所有实验中,模型的参数都是从训练集中采用极大似然法估计出来的.

测试的结果采取了常用的 4 个评测指标,即准确率 P、召回率 R、综合指标 F 值和交叉括号 CB.其定义如下:

精确率(Precision)用来衡量句法分析系统所分析的所有成份中正确的成份的比例.

召回率(Recall)用来衡量句法分析系统分析出的所有正确成份在实际成份中的比例.

综合指标:$F = (P \times R \times 2)/(P + R)$.

交叉括号 CB:给出了在一棵树中与其他树的成分边界交叉的成分数目的平均数.

实验中采用的句法分析 Baseline 系统是 Collins 基于 Bikel 平滑模型实现的句法分析器.表 1 列出了 baseline 系统和改进模型的句法分析实验结果.

表 1　句法分析实验结果

模型	准确率%	召回率%	F%	交叉括号
Baseline	82.76	80.17	81.44	2.05
改进模型	84.53	82.41	83.46	1.84

从表 1 可以看出:由于利用语言学知识对规则概率计算中进行了重新分解,并利用语义依存信息和互信息对此聚类,成功解决了严重影响句法分析系统性能的数据稀疏问题,改进模型的准确率 P、召回率 R、综合指标 F 值、交叉括号比 Collins 的中心词驱动句法分析模型有了明显的提高.

从图 1 可以看出,在不同训练规模上,改进模型的综合指标 F 值明显优于 Baseline 系统,而且与 Collins 的 Bikel 平滑算法相比较,基于词类的句法分析模型使句法分析系统的 F 曲线呈现明显的收敛趋势.

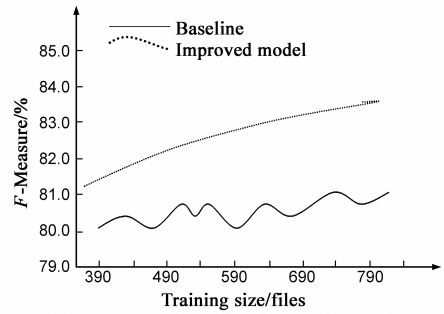

图1　不同训练规模下句法分析F测度

7　结论

本文在分析经典平滑算法的基础上,提出一种基于语义依存信息和互信息的词聚类算法,并利用绝对权重差分方法构造了一种可变长语言模型,进而提出了一种基于语义类和可变长模型的中心词驱动句法分析改进模型.数据稀疏问题是一个严重影响句法分析系统性能的重要因素,改进模型采用了基于语义类和可变长模型的平滑技术,成功解决了数据稀疏问题,大大提高了系统性能,精确率和召回率分别为 84.53% 和 82.41%,综合指标 F 值比 Collins 的中心词驱动句法分析模型提高了 2.02 个百分点.

本文提出的中心词驱动句法分析改进模型,有待进一步改进,在规则的分解及概率计算中,融入更多的语法、语义、语用等特征知识.改进模型对提高中文句法分析的性能效果明显,其在英文句法分析中的应用有待进一步研究.

参考文献

[1] Jesus Vilares, Miguel A Alonso, Manuel Vilares. Extraction of complex index terms in non-English IR: A shallow parsing based approach[J]. Information Processing and Management, 2008, 44(4): 1517 – 1537.

[2] 代印唐,吴承荣,等.层级分类概率句法分析[J].软件学报,2011,22(2):245 – 257.
DAI Yin-Tang, WU Cheng-Rong, et al. Hierarchically classified probabilistic grammar parsing[J]. Journal of Software, 2011, 22(2): 245 – 257. (in Chinese)

[3] Aviran S, Siegel P H, Wolf J K. Optimal parsing trees for run-length coding of biased data[J]. IEEE Transaction on Information Theory, 2008, 54(2): 841 – 849.

[4] ZHOU De-yu, HE Yu-lan. Discriminative training of the hidden vectors state model for semantic parsing[J]. IEEE Transaction on Knowledge and Data Engineering, 2009, 21(1): 66 – 77.

[5] 孙昂,江铭虎,贺一帆,等.基于句法分析和答案分类的中文问答系统[J].电子学报,2008,36(5):833 – 839.
SUN Ang,JIANG Ming-hu,HE Yi-fan,et al.Chinese question answering based on syntax analysis and answer classification [J].Acta Electronica Sinica,2008,36(5):833 – 839.(in Chinese)

[6] 陈毅恒,秦兵,等.基于 ontology 抽取优化初始选择的检索结果聚类[J].电子学报,2008,36(12A):166 – 171.
CHEN Yi-heng,QIN Bing,et al.Search result clustering based on centroid optimization by ontology extraction[J].Acta Electronica Sinica,2008,36(12A):166 – 171.(in Chinese)

[7] Daniel Jurafsky,James H.Martin.Speech and Language Processing[M].New Jersey:Prentice Hall,2009.210 – 265.

[8] David M Magerman.Statistical decision-tree models for parsing [A].Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics[C].Cambridge,1995.276 – 283.

[9] Eugene Charniak.Statistical parsing with a context-free grammar and word statistics[A].Proceedings of the 14th National Conference on Artificial Intelligence[C].Menlo Park,1997.598 – 603.

[10] Eugene Charniak.A maximum-entropy-inspired parser[A].Proceedings of the First Conference of the North American Chapter of the Association for Computational Linguistics[C].Seattle,2000.132 – 139.

[11] Collins M.Head-Driven Statistical Models for Natural Language Parsing[D].Pennsylvania:The University of Pennsylvania,1999.65 – 78.

[12] Collins M.Head-driven statistical models for natural language parsing[J].Computational Linguistics,2003,29(4):589 – 637.

[13] 刘水,李生,赵铁军等.头驱动句法分析中的直接插值平滑算法[J].软件学报,2009,20(11):2915 – 2924.
LIU Shui,LI Sheng,ZHAO Tie-Jun,et al.Directly smooth interpolation algorithm in head-driven parsing[J].Journal of Software,2009,20(11):2915 – 2924.(in Chinese)

[14] Chen S F,Goodman J.An empirical study of smoothing techniques for language modeling[A].Proceedings of the 34th Annual Meeting on Association for Computational Linguistics [C].Stroudsburg:Association for Computational Linguistics,1996.310 – 318.

[15] Frederick J,Mercer RL.Interpolated estimation of Markov source parameters from sparse data[A].Proceedings of the Workshop on Pattern Recognition in Practice[C].New York:Institute of Electrical and Electronics,1980.381 – 397.

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

[17] Bikel DM,Miller S,et al.Nymble:A high-performance learning name-finder[A].Proceedings of the 5th Conf on Applied Natural Language Processing[C].Stroudsburg:Association for Computational Linguistics,1997.194 – 201.

[18] Gao Jian-feng,Goodman J,Miao Jiang-bo.The use of clustering techniques for language model-application to Asian language[J].Computational Linguistics and Chinese Language Processing,2001,6(1):27 – 60.

[19] 袁里驰.基于相似度的词聚类算法和可变长语言模型[J].小型微型计算机系统,2009,30(5):912 – 915.
YUAN Li-chi.Word clustering based on similarity and varigram language model[J].Journal of Chinese Computer Systems,2009,30(5):912 – 915.(in Chinese)

[20] Lee L.Similarity-Based Approaches to Natural Language Processing[D].Cambridge,MA:Harvard University,1997.25 – 87.

[21] 袁里驰.基于词聚类的依存句法分析[J].中南大学学报:自然科学版,2011,42(7):2023 – 2027.
YUAN Li-chi.Dependency language paring model based on word clustering[J].Journal of Central South University:Natural Science,2011,42(7):2023 – 2027.(in Chinese)

作者简介

袁里驰 男,1973 年 5 月出生于湖南邵阳,江西财经大学信息管理学院副教授,硕士生导师.研究方向为自然语言处理.
E-mail:yuanlichi@sohu.com