统计与规则并举的汉语词性自动标注算法*

张民 李生 赵铁军 张艳风

(哈尔滨工业大学计算机科学与工程系 150001)

E-mail: zm@mtlab.hit.edu.cn

摘要 本文提出并实现了一种基于定量统计分析优先的统计和规则并举的汉语词性自动标注算法.本算法引入置信区间的概念,优先采用高准确率的定量统计分析技术,然后利用规则标注剩余语料和校正部分统计标注错误.封闭和开放测试表明,在未考虑生词和汉语词错误切分的情况下,本算法的准确率为98.9%和98.1%.

关键词 汉语,词性标注,隐马尔可夫模型,规则,置信区间.

中图法分类号 TP391

词性标注是语料库加工的重要环节.目前,有关英语和其他西方语言的词性标注研究比较深入,有关汉语的研究相对薄弱,比较典型的标注算法有:(1) 基于规则的方法.[1,2]国外在70年代初主要采用这种方法,著名的TAGGIT系统,利用3 300条上下文规则,对100万词次的Brown语料库标注正确率达到77%.[1](2) 基于统计的方法.[3~5]80年代初,随着经验主义方法在计算语言学中的重新崛起,统计方法在语料库词性标注中又占据了主导地位.CLAWS标注系统对LOB语料库的标注正确率达到96%左右.[3](3) 混合策略.[5]国内北京大学计算语言学研究所提出了一种先规则、后统计的规则和统计相结合的标注算法,其准确率达到了96.6%.[6](4) 另外,有人用神经网络和遗传算法的方法进行词性标记.[3]这类文献很少.

哈尔滨工业大学从1992年开始,先后对基于规则和基于统计的方法进行了研究,标注正确率分别达到89%和96%.对错误标注结果进行分析,可以看出,无论是哪种标注算法都有其固有缺陷.概率标注方法总会抑制小概率事件的发生,况且这种方法还会受到标记集、长距离搭配上下文等的限制[7],因此,对汉语语料的标注很难突破96%的准确率.规则的方法本质上是说一种确定性的演绎推理方法,这就决定在自然语言处理中不可能具有很强的鲁棒性.我们曾试图引入非确定性推理方法,但并不是很成功.[2]

受文献[8]的启发,本文将其译文选择算法中所用置信区间的方法引入到词性标注中[8],提出了一种基于定量统计分析优先的统计和规则相结合的汉语词性自动标注算法.所谓"定量统计分析优先"就是通过引入置信区间的评价函数,优先采用统计方法,只对满足置信区间评价函数的词给出选择,对不满足的按评价值大小给出多个输出,然后通过规则进行标注,规则还能起到校正部分统计标注错误的作用.本算法的核心,也即与其它混合策略[6]不同之处是,运用置信区间评价函数,优先采用统计方法,做到统计和规则的结合.实验结果显示,置信区间评价函数可以保证在前端的纯统计方法具有一定的召回率(Recall)的同时,具有可界定的较高的消歧率(Precision).

本文第1节对汉语兼类词现象的分布特征进行了简单的描述.第2和第3节分别介绍了基于统计和规则的标注算法.第4节全面地论述了规则和统计的结合方法.最后对实验结果进行了详细的数据分析,并给出结论.

1 汉语的兼类词现象

由于自然语言处理的特殊性,汉语词的兼类现象错综复杂,其主要构成如下[2]:

(1) 形同音不同,如:好(hao(三声、形容词)、hao(四声、动词)).

(2) 同音同形但意义上毫无联系,如:会(开个会(名词)、会(动词)滑冰).

---

* 本文研究得到国家863高科技项目基金资助.作者张民,1970年生,博士生,主要研究领域为机器翻译和计算语言学.李生,1943年生,教授,主要研究领域为机器翻译,计算语言学和人工智能.赵铁军,1962年生,博士,副教授,主要研究领域为机器翻译,计算语言学和人工智能.张艳风,1975年生,本科生,主要研究领域为计算语言学.

本文通讯联系人:张民,哈尔滨150001,哈尔滨工业大学计算机科学与工程系

本文1996-08-21收到原稿,1997-03-20收到修改稿

(3) 具有典型意义的兼类词,如:典型(名词/形容词).

(4) 上述的组合,如:行(动词/形词/名词/量词).

由以上4种情况构成的兼类词,在汉语中普遍存在.为了研究兼类现象的静态和动态分布特征,我们对标注所用词典(也是机器翻译所用词典)和一个已标记好的13万词语料库进行统计结果如下.

表1　兼类现象的静态分布特征(对词典统计结果)

总词数		54 760
兼类种类		113
兼类词条数		3 680
兼类词占总词数的百分比		6.72
高频兼类占总兼类词百分比	名词/动词	36.1(1 331)
	adj./ad.	26.4(974)
	其他(112)	37.5(1 375)

表2　兼类现象的动态分布特征(对语料统计结果)

总词次	131 230
总词条	8 761
兼类词词次	30 972(23.6%)
兼类词词条	527
兼类词种类	78

从以上两表可看出,汉语兼类词的静态和动态分布特征差别很大.兼类词条数虽然不是很多,但在语料中出现的词次已不可忽视.另外,不同的兼类现象和不同的兼类词分布差别很大.例如,在113种兼类现象中,"名/动"和"形/副"兼类就占62.5%;在语料中,兼类词次达30 972次,却只出现527个不同的兼类词条.这说明在真实语料中,某些兼类词出现的频度极高(如过、好、得、没有等词).这些兼类现象出现的分布特征在某种程度上决定了消歧的策略.

2　统计标注算法

统计方法是一种非确定性的定量推理方法,在词性标注中它把句子中每个词及其词性的出现都看作是一个随机过程.HMM模型把词和词性的出现看作是一个向前依赖的条件概率事件,其模型参数是通过对大规模语料训练自动习得的.因此,它不仅可获得较好的一致性和很高的覆盖面,而且还可将一些不确定的知识客观定量化地描述出来.

一般说来,基于统计的标注算法包括VB(viterbi)和FB(forward-backward)算法.VB算法源于Bayes公式,反映了局部概率简化的条件下整体的最佳效果,因此,它对每个歧义词只给出唯一候选.这样,在准确率要求不很高或没有后处理的情况下,VB算法是很有效的.FB算法源于对HMM模型参数的迭代训练过程,反映了在整体限制条件下寻求局部最优概率分布,因此,该算法可按照可信度大小给出多个输出.在目前的文献中往往把具有最大概率值的标记或概率值本身或概率比值大于某一阈值的标记作为正确选择[3~5],这从概率角度讲是正确的.但如后文所述,在某些情况下,概率最大并不一定完全正确,而且一个合适的阈值也很难选择.另外,由于有规则的后处理,因此,本文首先采用FB算法,然后试图通过引入置信区间的概念[8],构造出一种基于"置信区间"的动态评价函数[8],对FB算法的多输出词性进行评价,仅对满足该评价函数的词给出最大概率的词性作为唯一正确选择,否则,按照概率值大小给出多个候选,留给规则进行处理.如后文所述,测试表明,置信区间评价函数可以保证统计消歧模块在具有一定召回率的同时,具有可界定的较高的消歧率(消歧率为94.8%时,召回率可达81.5%).

下面仅给出3-gram模型下FB算法的描述[4]:

令
$$F(t_{i-1},t_i)=\sum_{t_{i-2}}[F(t_{i-2},t_{i-1})*p(t_i/t_{i-1},t_{i-2})*p(w_{i-1}/t_{i-1})] \tag{1}$$

$$B(t_{i-1},t_i)=\sum_{t_{i+1}}[B(t_i,t_{i+1})*P(t_{i+1}/t_i,t_{i-1})*P(w_{i+1}/t_{i+1})] \tag{2}$$

则
$$\hat{\wp}(W)_i=\arg\max_{t_i}\sum_{t_{i-1}}\{F(t_{i-1},t_i)*B(t_{i-1},t_i)*p(w_i/t_i)\} \tag{3}$$

(3)式就是所谓的3-FB算法,其中t是词性标记,w是词条标记,(t_{i-1},t_i)在本文中被称为节点对.利用(1)(2)式,F值和B值可通过月递推算法对整个HMM模型遍历求得,再利用(3)式即可求出歧义词w的每个可选标记的非归一化概率值.本文使用3-FB算法作为纯概率标注算法,然后利用评价函数对其输出结果进行评价.

关于模型的训练(转移概率和发射概率的训练),包括模型复杂度的讨论(模型参数的估计)、RF(relative frequen-

cy)训练、ML(maximum likelihood)训练、数据稀疏的处理等见文献[4].

3 规则标注算法

统计方法的优点在于对不确定事件的定量描述.定量是基于概率的,因此其必然会掩盖小概率事件的发生.规则方法的优点在于根据上下文对确定事件的定性描述.有些统计方法无法解决的问题,利用规则却很容易解决,如,词"地"的"名/助"兼类问题.所以,最好的标注算法是统计和规则结合的方法.

本文采用传统的规则标注算法,将文献[2]中"利用线性结构知识"模块抽出并对其进行改造,应用于本系统.主要采用个性词规则和通用上下文两种规则形式,用来标注统计方法所无法确定的标记.本系统共有规则213条,其中个性规则174条,通用规则39条.本文所使用的规则,一是用来标注统计方法所无法保证正确标记的剩余语料,二是可以校正用统计方法标注的明显错误之处(见第5节),因此其上下文约束条件有其特殊性.有关详细讨论见文献[2].

4 统计和规则的结合(置信区间的引入)

如前文所述,比较好的标注算法应是统计和规则并举的算法,但问题的关键是如何做到统计和规则的结合.文献[6]给出一种先规则后统计的结合算法,取得了很好的效果,但该算法规则的作用域是非受限的,而且并没有研究统计的可信度.这样,规则和统计的作用域不明确.本文则通过研究统计的可信度,借鉴文献[8],引入置信区间的方法,构造出一种基于置信区间的评价函数,做到先统计后规则的统计和规则的并举.

为讨论方便,可设兼类词 w 的候选词性为 T_1,T_2,T_3,其对应概率的真实值为 t_1,t_2,t_3,FB算法计算出的概率值为 $\hat{t}_1,\hat{t}_2,\hat{t}_3$,利用公式(3)计算出的 $\varnothing(w)_{T_i}$ 为 n_1,n_2,n_3(令 $n_1>n_2>n_3$),$\hat{t}_i=\dfrac{\varnothing(w)_{T_i}}{\sum\varnothing(w)_{T_j}}$.若 $\hat{t}_1$ 与 $\hat{t}_2$ 相差很大时,选择 T_1 导致错误的可能性就很小;若 $\hat{t}_1$ 与 $\hat{t}_2$ 相差不大时,选择 T_1 导致错误的可能性就很大,如 $\hat{t}_1=0.50,\hat{t}_2=0.45$,很难说应该选择 T_1 还是 T_2.简单的阈值法肯定是不可取的.以 $\hat{t}_1/\hat{t}_2$ 是否大于阈值作为是否选择 T_1 的判定条件,比直接判断 T_1 的阈值更加合理.但这种判定条件仍然存在下面的问题:假设 $\hat{t}_1/\hat{t}_2=n_1/n_2=3$,一种情况是 $n_1\cong300,n_2\cong100$;另一种情况是 $n_1\cong3,n_2\cong1$,显然在前一种情况下选择 T_1 比在后一种情况下选择 T_1 更加可靠.评价算法必须能够反映出这种差别.为保证在一定的正确率下作出选择,需要研究选择的统计的可信度.也就是说,根据 n_1,n_2 计算出的 $\hat{t}_1,\hat{t}_2$ 只是 t_1,t_2 的近似值,我们必须估计出这种近似的误差,对 $\hat{t}_1/\hat{t}_2$ 进行修正,然后再对修正后的 $\hat{t}_1/\hat{t}_2$ 进行判别.

因为 $\ln(\hat{t}_1/\hat{t}_2)$ 比 $\hat{t}_1/\hat{t}_2$ 更快地逼近正态分布[8],可应用单边区间估计方法计算 $\ln(\hat{t}_1/\hat{t}_2)$ 的置信区间.设显著性水平为 α(Error Probability),即置信度为 $1-\alpha$,服从正态分布的随机变量 X 的置信区间为 $Z_\alpha*\sigma$,Z_α 为置信系数(Confidence Coefficient),可从统计表中直接查到,σ 为标准差,$\sigma\approx\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}$[8],若把 $\ln(\hat{t}_1/\hat{t}_2)$ 看作随机变量,则其置信区间为:

$$Z_\alpha*\sigma=Z_\alpha*\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \tag{4}$$

本文不加证明地给出最终的评价函数如下,有关证明请参见文献[4].

$$\ln\frac{n_1}{n_2}\geqslant\theta+Z_\alpha*\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \tag{5}$$

其中 θ 为经验值,可通过训练得到,本文取0.4.可以看出(5)式是一个动态阈值函数.设 $\beta=\ln\dfrac{n_1}{n_2}$,阈值 $L_\beta=\theta+Z_\alpha*\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}$.它共有3组参数:$\alpha,\theta$ 和 n_1,n_2.θ 是 β 的最低静态阈值,独立于训练语料的规模;α 反映了统计的显著性,当 α 变小时,可信度变大,Z_α 变大,阈值 L_β 变大,这和我们的直观理解一致;L_β 和 n_1,n_2(在某种程度上反映了语料的稀疏程度)是倒数关系,当 n_1,n_2 变大时,L_β 变小.因此,动态阈值函数(5)是在静态阈值 θ 的基础上,由可信度($1-\alpha$)和训练语料规模(n_1,n_2)共同决定的.

下面举例说明.对于前面 n_1/n_2 为300/100和3/1的情况,取 $\theta=0.4,\alpha=0.05$,则 $Z_\alpha=1.649,\beta=1.098,L_\beta(3/1)=0.4+1.649*(1/3+1)^{1/2}=2.304;L_\beta(300/100)=0.4+1.649*(1/300+1/100)^{1/2}=0.590$.因为 $\beta>L_\beta(300/100)$,所以可对其进行选择.同时,我们还可以看到,同样的 α 和 β,而 $L_\beta(300/100)$ 却明显小于 $L_\beta(3/1)$.

利用评价函数(5),应用统计方法,对满足(5)式的给定的 α(本文取0.05)作出选择,使可选择的结果达到一个可界定的较高的准确率($1-\alpha$),对剩余语料利用规则标注.这样,既保证统计标注具有较高的准确率,又使规则标注具有更强的针对性.

5　实验结果分析

依据上面所论述的思想,本文实现了一个统计与规则并举的汉语词性自动标注系统.系统的标记集共有42个词类标记,所用语料共13 000汉语句子,均已经过人工校对和人工标注.语料主要来自"DEAR"翻译工作站的例句库[9]、《现代汉语动词小词典》的例句、清华大学外语系的语料库.其中10 000句用来训练,3 000句用来测试.统计模块使用了RF训练和FB标记算法,利用评价函数(5)作出选择;规则模块使用213条通用和个性规则标注统计剩余的语料,并校正统计标注的部分错误结果.下面给出对系统的测试和评价.这些测试和评价指标未考虑未登录词的错误和切分错误.由于经过人工校对,不存在切分错误,对登录词均按名词处理.

5.1　统计测试

引入几个评价参数,一是召回率(Recall),评价(5)式的覆盖(Coverage)能力,定义为:

$$\text{召回率}=\frac{\text{(5)式作出选择的兼类词数}}{\text{测试文本中总的兼类词数}}\times 100\%$$

二是消歧率(Precision),评价(5)式的消歧能力,定义为:

$$\text{消歧率}=\frac{\text{(5)式作出正确选择的兼类词数}}{\text{(5)式作出选择的兼类词数}}\times 100\%$$

鉴于第1候选的消歧率(第1候选正确的兼类词数占整个兼类词数的比例)已达到83.8%,所以,α最大值取0.16(≈1−83.8%).表3为一组测试结果.

表3

α	Z_α	召回率(%)	消歧率(%)
0.16	0.999	92.2	85.7
0.13	1.128	90.1	87.8
0.10	1.282	86.8	89.7
0.05	1.649	81.5	94.8
0.01	2.329	72.9	98.1
0.00	3.900	72.7	98.3

从表中可以看出,消歧率基本约为$1-\alpha$,与上节所述相符,证明(5)式是正确的.当α取0.10以下时,消歧率已小于$1-\alpha$,进一步证明了有些大概率标记并不一定可靠.表中最后3行,当取0.05、0.01和0.00时,召回率已小于83.8%,说明这时(5)式仅能把第1候选正确的部分词选择出来.当α取0.01时,召回率已明显下降,但α取0.00时的召回率同α取0.01时相比几乎没有下降.召回率在α取0.05以下时下降幅度较大,所以在系统实现时,建议α取0.05,这时消歧率为94.8%,比第1候选消歧率(83.8%)高11个百分点,而召回率81.5%也是可以接受的,共有约19%的歧义词没有做出选择.

我们曾对简单的阈值法(即当$\hat{t}_1$大于某一阈值时选择T_1)进行测试,结果表明,当消歧率为94.8%时,其阈值高达0.93,而召回率仅为58.4%,远低于(5)式的召回率81.5%.以$\hat{t}_1/\hat{t}_2$的比值是否大于某一阈值为条件进行测试,当消歧率为94.8%时,其阈值为3.24,而召回率仅为67.1%,虽高于简单阈值法的召回率,但远低于(5)式的召回率,这进一步证明评价函数(5)是有效的.

此外,对评价函数的错误评价结果进行分析,其大部分错误可归结为:①模型上下文约束距离的限制;②标记集的分布特征不足,描述粒度不够或偏向于语义分类;③标记训练语料的稀疏.这些都为算法的进一步改进提供了方向.

5.2　规则校正和标注

通过对统计标注错误的观察发现,出错的主要原因是模型上下文描述距离的限制和标记集描述粒度不足;(5)式未作出选择的情况主要是n_1,n_2的值比较小或二者较接近(即概率接近),这部分原因是由语料稀疏造成的,但对于n_1,n_2比较接近的情况,(5)式是绝对无法作出选择的.

规则在某些情况下,可突破上下文约束距离的限制且与训练语料稀疏无关,而且规则可使用复杂特征集,不象一般的统计仅使用单一标记,因此规则的引入不仅可标注统计标注剩余的语料,还可校正部分统计标注错误.

例如,句1"我在(介词)学习英语的时候"和句2"我在(助词)学习英语".对3-gram,"在"的上下文环境相同["在"(介词:0.83;助词:0.14;动词:0.03)],均满足(5)式,因此,统计方法均把其选为介词.而规则对上述两种情况则可作出区分.基于此,规则可校正部分统计出错情况.句3"我一把屋子收拾好,就走.",其中"一"[数词:0.68;连词:0.31],不满足(5)式,统计方法无法对其选择.利用一条规则,"一"前一词且是句首词为人称代词,下一个分句句首词是"就",

则可把“一”和“就”选为连词.这样,规则可标注统计无法标注的剩余语料.

对剩余语料标注发现,如果全选择第1候选,则正确率为47.7%,而应用本节规则标注,则可达87.1%的正确率,提高了39.4个百分点.

6 结 论

封闭和开放测试表明,本算法的准确率可达98.9%和98.1%,证明本文提出的基于定量统计分析优先的统计和规则并举的算法是有效的.

对本算法的错误标注统计发现,最容易出错的是对连词和介词的标注.连词出错占整个出错情况的20.59%,而连词的频度仅为整个语料的1.98%,且在出错的连词中有79.41%是被标成介词.通常,区分连词和介词的一个标准是看其左右成份的关系.如,“我和他都是学生”,这个“和”是连词;而在“我和他谈了很久,感到很受启发”一句中,这个“和”表示对象关系,应标为介词,这从英语角度很好理解(I talk with him),其实,这种区分已是句子级的语义范畴.再比如,“处理好(副词)关系”和“养成好(形容词)习惯”中的“好”字更是需要引入细粒度的语义描述.这种语义的引入是本算法在规则部分需要加强的地方.

论述和测试表明,本文提出的这种基于定量统计分析优先的统计和规则相结合的自动标注算法是正确的.置信区间的引入,定量统计优先与定性规则的结合是有效的,这也是本文与其他混合策略的不同之处.对本算法,最主要的改进之处是细粒度的语义知识在规则部分的引入,其次还有语料稀疏的处理,改进标记集使其更具有分布特征等.

致谢 非常感谢张弛、安鲁新、马光庆、刘小虎、蔡萌同学对本文的帮助,同时非常感谢清华大学周明副教授对本文工作的指导.

参考文献

1 Brill E, Magerman D, Marcus M *et al*. Deducing linguistic structure from the statistics of large corpus. In: Proceedings of the DARPA Speech and Natural Language Workshop. Hidden Valley PA: Addison Wesley Longman Limited, USA, 1990. 275~282

2 Zhao Tie-jun, Mao Cheng-jiang, Zhang Min *et al*. Solving the ambiguity of Chinese POS in the CEMT-III system. Chinese Information Journal, 1994,7(4):52~59

3 Bernard Merialdo. Tagging English text with a probalistic model. Computational Linguistics, 1994,20(2):1~29

4 Zhang Chi. Research on the algorithm of POS tagging on Chinese corpus based on statistics [Bachelor Thesis]. Harbin Institute of Technology, 1996

5 Bai Shuan-hu. Research on the algorithm of POS tagging on Chinese corpus based on statistics [Master Thesis]. Tsinghua University, 1992

6 Zhou Qiang. An algorithm of tagging Chinese POS based on statistics and rule. Chinese Information Journal, 1996,9(3):1~9

7 Elliott Macklovith. Where the tagger falters. In: Proceedings of TMI-92, the 4th International Conference on Theoretical and Methodological Issues in Machine Translation. Pierre Isabelle, Bell Canada Publishing House, Canada, 1992. 113~126

8 Ido Dagan, Alon Itai. Word sense disambiguation using a second language monolingual corpus. Computational Linguistics, 1994,20(4):563~596.

9 Zhou Ming, Li Sheng *et al*. Dear: a translator's workstation. In: Proceedings of the NLPRS'95, Natural Language Processing Pacific Rim Symposium. KPChoi: Publishing House of Korean Advance Institute of Science and Technology, Korea, 1995. 388~397

Part of Speech Tagging Chinese Corpus Based on Statistics and Rules

ZHANG Min LI Sheng ZHAO Tie-jun ZHANG Yan-feng

(*Department of Computer Science and Engineering Harbin Institute of Technology Harbin* 150001)

Abstract This paper proposes an algorithm of automaticallytagging the POS(part of speech) of Chinese words which is based on integration of the statistical technique and the rule technique with the priority of the quantitative statistical analysis. The confidence intervals in the estimation of parameters is employed in the algorithm, and this makes the high-accuracy quantitative statistical technique as the top priority of tagging a corpus. Then the untagging part of the corpus is tagged in terms of rules, and some errors by statistics can be corrected by rules. Both closed and opened tests indicated that the accuracies of the algorithm are 98.9% and 98.1% respectively without consideration of both unknown words and segmentation errors.

Key words Chinese, part of speech tagging, hidden Markov model, rule, confidence intervals.

Class number TP391