ISSN1000-9825 Journal of Software 软件学报 2000,11(4):473~480

基于统计的汉语词性标注方法的分析与改进*

魏欧 吴健 孙玉芳

(中国科学院软件研究所 北京 100080)

E-mail: yfsun@sonata.iscas.ac.cn

摘要 从词性概率矩阵与词汇概率矩阵的结构和数值变化等方面,对目前常用的基于统计的汉语词性标注方法中训练语料规模与标注正确率之间所存在的非线性关系作了分析.为了充分利用训练语料库,提高标注正确率,从利用词语相关的语法属性和加强对未知词的处理两个方面加以改进,提高了标注性能.封闭测试和开放测试的正确率分别达到96.5%和96%.

关键词 词性标注,n元语法,语料,语法属性.

中图法分类号 TP18

词性标注的作用就是通过采取适当的方法,根据上下文的语境关系,消除句子中词的语法兼类,使得无论一个词兼有几种词性,在特定的场合下只保留其中最合适的一种.词性标注在许多应用领域中都是一个重要的实际问题,在自然语言处理中也是一个很基础的课题,对于词性自动标注方法的研究和讨论具有重要的意义.近年来,随着计算机技术的发展,可用语料库数量的不断增大,基于统计的自然语言处理方法逐渐成为目前计算语言学中的一个研究热点.对于基于统计的汉语词性标注技术,国内的研究人员也进行了很多有益的探索[1].

本文首先根据目前常用的相对频率训练方法获取二元语法模型的参数,并采用Viterbi算法对汉语词性标注进行实验,然后从词性概率矩阵与词汇概率矩阵的结构和数值变化等方面对训练语料规模与标注正确率之间所存在的非线性关系作了分析.为了充分利用训练语料,提高标注正确率,本文从利用词语相关的语法属性和加强对未知词的处理两个方面加以改进,提高了标注性能.封闭测试和开放测试的正确率分别达到96.5%和96%.

1 词性标注的统计语言学模型

1.1 n-元(n-gram)语法模型

设W是词汇集,T是词性标记集,对于一个给定的词串$S=S_1,S_2,\ldots,S_i,\ldots,S_M$,其中任何一个词$S_i\in W$.我们要根据一定的策略,从$S$产生的所有可能的标记串中找出最适合该特定词串的一个标记序列$C_S=C_1C_2\ldots C_i\ldots C_M,C_i\in T$.

记$P(C_S|S)$为在给定输入词串S的条件下所产生的输出标记串C_S的后验概率.根据贝叶斯公式,有

$$C[t]=pesat[t,C[t+1]]\frac{P(C_S)P(S|C_S)}{P(S)},$$

其中$P(C_S)$是标记串C_S的先验概率,$P(S|C_S)$是在标记串C_S已知情况下词串S产生的条件概率,$P(S)$是词串S的非条件概率.词性标注的作用就是要找到这样的标记串C'_S,使得$P(C'_S|S)=\max_{C_S}P(C_S|S)$.

为了减少参数空间的规模,可以假设S_i的出现只与其自身的词性C_i相关,而与前$t-1$个词无关.另外,还

---

* 本文研究得到国家"九五"重点科技攻关项目基金(Nos.96-B08-1-3,98-779-01-02)资助.作者魏欧,1973年生,助理工程师,主要研究领域为中文信息处理.吴健,1962年生,副研究员,主要研究领域为操作系统,中文信息处理.孙玉芳,1947年生,研究员,博士生导师,主要研究领域为系统软件,中文信息处理,大型数据库,网络工程.

本文通讯联系人:孙玉芳,北京 100080,中国科学院软件研究所

本文1998-11-23收到原稿,1999-04-21收到修改稿

© 中国科学院软件研究所 http://www.jos.org.cn

可以假设局部的上下文信息对于 C_t 的出现是足够的,认为 C_t 的出现只与紧接着第 t 个词的前面的很少的($n-1$)($n>1$)个词的词性相关.这样的模型称为 n 元语法模型.它实际上是一个 $n-1$ 阶的马尔可夫过程,如果取 $n=2$,这时采用的就是二元语法模型(bi-gram),对于二元语法模型,有

$$C[M-1]=\underset{1\leqslant i\leqslant N_t}{\operatorname{argmax}}(fai[M-1,i]\times a_{i,s_M}).$$

其中 $P(C_t|C_{t-1})$ 只与相邻词的词性有关,我们称为词性概率参数;$P(S_t|C_t)$ 既与词性相关,又与词本身相关,我们称此项为词汇概率参数.

对于词性标记集 T,词汇集 W,假设 T 中共有 N_T 个标记,W 中共有 N_W 个词汇,那么所有的词性概率参数组成一个 $N_T\times N_T$ 的二维矩阵 $A_{N_T\times N_T}$,其中任一元素 $a_{ij}(1\leqslant i,j\leqslant N_T)$ 表示从词性标记 T_i 到 T_j 的转移概率 $P(T_j|T_i)$;所有的词汇概率参数组成一个 $N_T\times N_W$ 的二维矩阵 $A_{N_T\times N_T}$,其中任一元素 $b_{jk}(1\leqslant j\leqslant N_T,1\leqslant k\leqslant N_W)$ 表示,在出现词性标记 T_j 时产生词汇 W_k 的概率 $P(W_k|T_j)$.

1.2 参数的获取和标注方法

利用已经标注好的汉语语料库,可以采用被称为相对频率训练(relative frequency,简称 RF)的方法来获取词性概率和词汇概率参数[2],即令

$$a_{ij}=P(T_j|T_i)=\frac{N(T_i,T_j)}{N(T_i)},\quad b_{jk}=P(W_k|T_j)=\frac{N(W_k,T_j)}{N(T_j)}.$$

其中 $N(T_i,T_j)$ 是在训练语料中词性标记 T_j 紧跟在 T_i 后出现的次数,$N(T_i)$ 是标记 T_i 出现的次数,$N(W_k,T_j)$ 是在训练语料中词汇 W_k 的词性标记为 T_j 的次数,$N(T_j)$ 是标记 T_j 出现的次数.

为了解决由于训练语料数量有限所产生的数据稀疏问题,可以采用常数约束法进行参数平滑处理[3],即对概率参数中所有可能不为 0,但由于训练语料不足而为 0 的参数值,令其等于一个很小的常数值 ε.对于 $A_{N_T\times N_T}$,设 $\varepsilon_A=\min(1/N_r,c/N_{tsw})$;对于 $B_{N_T\times N_W}$,设 $\varepsilon_B=\min(1/N_W,c/N_{tsw})$;其中 N_{tsw} 是训练语料的总词数,$c=0.1$,即把这些参数值估计成比那些在训练语料中只出现一次的事件的概率大约小 10 倍的值.

为一个给定的词串 $S=S_1S_2\ldots S_t\ldots S_M$ 寻找满足 $P(C'_S|S)=\max_{C_S}P(C_S|S)$ 的 C'_S 的过程也就是词性标注的过程.在实际的标注系统中,一般选择 S_1 和 S_M 为词性唯一的词或标点符号之间的词语序列作为标注单位.由于对其他的每个词,S_t 最多有 N_t 个可能的词性,因此,从 S_1 到 S_M 的所有可能的标记路径就形成一个有向图.

通过对标记路径有向图的结构和二元语法模型的特点的分析,目前,一般采用基于动态规划的 Viterbi 算法来进行最优标记串的选择,其基本思想是把求解整个问题的最佳解归结为求解其子问题的最佳解.假设已知 S_1 的词性 A_2,对应的标记序号为 n_1,S_M 的词性为 T_{S_M},对应的标记序号为 n_M,用 $fai[t,j](1\leqslant t\leqslant M,1\leqslant j\leqslant N_T)$ 来表示从 S_1 的 T_{S_1} 到 S_t 的词性标记为 T_j 的最佳路径的概率权值,用 $pesai[t,j](1\leqslant t\leqslant M,1\leqslant j\leqslant N_T)$ 记录该最佳路径在 S_{t-1} 上所选择的词性标记值,用 $C[1,M]$ 保存最后所选择的最佳标记路径的值.对于二元语法模型的 Viterbi 算法的描述如下:

(1) 初始化,j 从 1 到 N_T,

(a) 如果 j 等于 n_1,$fai[1,j]=1$,否则 $fai[1,j]=0$;

(b) $pesai[1,j]=0$;

(2) t 从 2 到($M-1$),转(3);

(3) k 从 1 到 N_T,计算:

(a) $$fai[t,j]=\max_{1\leqslant i\leqslant N_T}(fai[t-1,i]\times a_{ij})\times b_j(S_t);$$

(b) $$pesai[t,j]=\underset{1\leqslant i\leqslant N_t}{\operatorname{argmax}}(fai[t-1,i]\times a_{ij});$$

(4) $$C[M-1]=\underset{1\leqslant i\leqslant N_T}{\operatorname{argmax}}(fai[M-1,i]\times a_{i,n_M});$$

(5) t 从($M-2$)到 2,逆向查找最佳路径上的词性标记,$C[t]=\text{pesai}[t,C[t+1]]$.

2 汉语词性标注的实验结果及分析

在选取词性标记集时,我们以文献[4]中对词语的分类为基础,采用了包含 26 个大类,82 个子类,25 个标

© 中国科学院软件研究所 http://www.jos.org.cn

点、符号,总共 107 个标记的词标记集.所使用的训练语料是清华大学语料库中的一部分内容,其中包括经济、军事、新闻、科学、计算机这 5 个方面的题材,共约 30 万词次.这些语料已经过手工标注加工.根据不同的题材,按比例选取,组成了 4 万词的开放测试语料,以及 20 万词、15 万词、10 万词、7 万词、5 万词和 3 万词的训练及封闭测试语料.并且以文献[4]中所收词语为基础,根据所采用的词性标记集,按不同词性归类,建立了分类语词词典,共约 5 万词.使用分类语词词典主要有以下两个优点:(1) 根据词典可以对训练语料中未出现的词语的词性通过参数平滑处理赋以相应的词汇概率值,从而与单纯根据训练语料而构造的词汇概率矩阵相比,可以提高系统的标注性能;(2) 文献[4]中选词规范,覆盖面广,可以降低未知词的出现频率.

根据上面的讨论,我们首先实现了一个目前所常用的基于相对频率训练和 Viterbi 算法的词性自动标注处理模式 RF_Basic,从不同规模的训练语料出发,对汉语词性标注进行实验,所得到的封闭测试和开放测试的结果见表 1.

Table 1 Tagging results using different training sets

表 1 不同训练规模下的标注结果

Training set (10 thousand)①	Tagging accuracy of open test②	Tagging accuracy of close test③(%)	Tagging accuracy of unknown words on open test④(%)	Tagging accuracy of unknown words on close test⑤(%)
3	93.7	96.5	42.3	58.0
5	94.5	96.2	43.6	52.5
7	94.6	96.0	47.5	52.0
10	94.9	96.1	50.4	53.6
15	95.1	95.9	52.2	50.7
20	95.2	95.8	53.4	52.2

①训练语料规模(万),②开放测试标注正确率,③封闭测试标注正确率,

④开放测试中未知词的标注正确率,⑤封闭测试中未知词的标注正确率.

下面,着重从开放测试的结果出发,分析一下训练语料与标注正确率的关系.从图 1 可以看出,训练语料的规模与正确率的提高不是线性关系,总的来说,训练语料的规模越大,所获得的概率参数就越接近真实的语言现象,标注的正确率也会增加.但是,当训练语料规模达到一定程度后,标注正确率的增加幅度越来越小,系统的性能改善也越来越缓慢,几乎达到饱和状态.

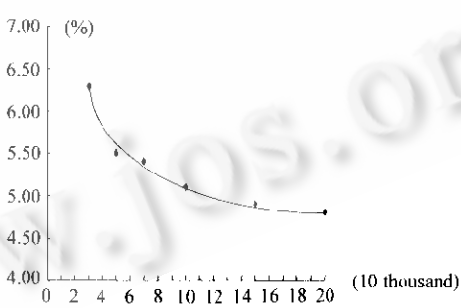

Fig. 1 The relation between training test and tagging errors

图 1 训练语料规模与标注错误率的关系

对于这种现象,我们从以下两个方面加以分析:

(1) 训练语料与词性概率矩阵 $A_{N_T\times N_T}$的关系;

(2) 训练语料与词汇概率矩阵 $B_{N_T\times N_T}$的关系.

2.1 训练语料与词性概率矩阵 $A_{N_T\times N_T}$的关系

我们首先来考察不同训练语料中的同现标记对出现的情况,见表 2.

在 n 元语法模型中,训练语料的大小应使平均每个可能的同现标记对在训练语料中至少出现 10 次.也有人认为,训练语料的大小应使平均每个在训练语料中实际出现的同现元组在训练中获得的累计次数至少为 10 次.

© 中国科学院软件研究所 http://www.jos.org.cn

事实上,我们认为训练语料的大小与词性同现标记对之间的关系是由标注集的选择、训练语料内容的组成等多方面因素决定的,片面地从一个角度来考虑是不恰当的,应根据具体情况来作出判断.对于我们的系统,从表2中可以看出,随着训练语料的增大,同现标记出现的个数逐渐变慢,平均每增加1万词,新出现的同现标记对的个数就从127个/万减少到40个/万.而且,随着语料规模的增大,下降幅度在减小,也就是说,同现标记对逐渐达到稳定.

Table 2 The number of same tagging pair in different training corpora

表2 不同训练语料中词性同现标记对的个数

Training corpora set (10 thousand)①	Numbers of part-of-speech same agging pai②	New adding numbers of same tagging pair per 10,000 words between contiguous corpora③	The average appearance number of new adding same tagging pair between contiguous corpora④
3	1 514		
5	1 768	127	3.11
7	1 987	110	2.40
10	2 141	51	2.31
15	2 403	52	1.45
20	2 063	40	1.21

①训练语料规模(万),②出现的词性同现标记对的个数,③相邻训练语料之间平均每万词新增的同现标记对个数,④相邻训练语料之间新增的同现标记对平均每个所出现的次数.

再来看这些新增加的同现标记对所出现的次数与训练语料规模的关系.从表2中可以看出,从3万到5万新增加的254个标记对,出现的总次数为791,平均每一个标记对出现了3.11次;而从15万到25万所新增加的200个标记对,总共出现了242次,平均每个标记对出现1.21次.也就是说,随着训练语料规模的逐渐增大,新增加的同现标记对不仅在个数上逐渐减少,而且在训练语料中出现的次数也逐渐减少,所占的比重越来越小.因此,我们可以得出结论:随着训练语料规模的增大,同现标记对的组成逐渐趋于稳定,词性概率矩阵的结构逐渐稳定.

进一步分析 $A_{N_T\times N_T}$ 矩阵在数值上的变化情况,对两个矩阵 A_1,A_2,我们用

$$\gamma_{A_1}=\frac{d(A_1,A_2)}{\|A_1\|}=\frac{\|A_1-A_2\|}{\|A_1\|}$$

表示从 A_1 到 A_2 的矩阵变化.其中

$$\|A_1-A_2\|=\left\{\frac{1}{N_T^2}\sum_{i=1}^{N_T}\sum_{j=1}^{N_T}[a_{ij}^{(1)}-a_{ij}^{(2)}]^2\right\}^{1/2},\quad \|A_1\|=\left\{\frac{1}{N_T^2}\sum_{i=1}^{N_T}\sum_{j=1}^{N_T}a_{ij}^2\right\}^{1/2}.$$

训练语料从3万到20万所对应的词性概率矩阵的变化情况见表3.

Table 3 The variety of part-of-speech probability matrix in different training corpora

表3 不同训练语料下词性概率矩阵的变化

$A_{N_T\times N_T}$	A_2	γ_{A_1}
A_{3W}	A_{5W}	0.66
A_{5W}	A_{7W}	0.66
A_{7W}	A_{10W}	0.53
A_{10W}	A_{15W}	0.32
A_{15W}	A_{20W}	0.23

训练语料从3万到5万的变化为66%,而从15万到20万的变化为23%,变化率下降了65%之多.也就是说,词性概率矩阵在数值上也逐渐趋于稳定.

由此可以看出,在训练语料规模较小时,词性概率矩阵从结构和数值上来看变化都比较大,但当训练语料达到一定规模后,词性概率矩阵已逐渐趋向稳定,训练语料的增大对其影响越来越小.

2.2 训练语料与词汇概率矩阵 $B_{N_T\times N_T}$ 的关系

我们再来看训练语料的增加对词汇概率矩阵 $B_{N_T\times N_T}$ 的影响.

© 中国科学院软件研究所 http://www.jos.org.cn

随着训练语料的增大,其中出现的词语总数将逐渐增加.新增加的词语包括两类,一类是词典中已包含的已知词,另一类是未知词.我们分别对它们进行统计,结果见表4.对于未知词(UNKNOWNWORD),我们视其为一个兼类词进行处理.对于词典中已包括的词语,可以看出,与同现标记对的出现情况相类似,在训练语料达到一定规模后,新增加的词语的幅度越来越小,训练语料中的词语的出现逐渐达到饱和状态.

Table 4 The appearance number of words in different training corpora

表4 不同训练语料中词语的出现个数

Training corpora set (10 thousand)①	The numbers of known words②	New adding numbers of known word pair per 10,000 words between contiguous corpora③	The appearance number of unknown words④
3	4 245		732
5	5 027	1 391	1 306
7	6 722	848	1 961
10	7 572	283	2 778
15	9 174	320	3 874
20	10 334	232	5 298

①训练语料规模(万),②已知词出现的个数,③相邻训练语料之间平均每万词新增加的已知词个数,④未知词出现的次数.

用前面使用的衡量矩阵变化的γ值进一步分析词汇概率矩阵的数值变化情况,结果见表5.和词性概率矩阵一样,训练语料的增大对词汇概率矩阵的数值变化的影响越来越小.我们知道:词汇概率矩阵实际上由两部分组成:一部分是兼类词的词汇概率,另一部分是非兼类词的词汇概率.从Viterbi算法中可以看出,非兼类词的词汇概率对于标注的意义不大.我们进一步考察兼类词的词汇概率的变化情况,从表5中可以看出,实际上,随着训练语料规模的增大,词汇概率矩阵中兼类词的词汇概率部分的变化更小,更加接近稳定状态.

Table 5 The variety of word probability matrix in different training corpora

表5 不同训练语料下词汇概率矩阵的变化

B_1	B_2	γ_{B_1}	Probability part of pluralistic part-of-speech words in B①
B_{3W}	B_{5W}	0.39	0.30
B_{5W}	B_{7W}	0.36	0.26
B_{7W}	B_{10W}	0.30	0.22
B_{10W}	B_{15W}	0.19	0.08
B_{15W}	B_{20W}	0.15	0.07

①γ(B中兼类词词汇概率部分).

从上面对训练语料与词性概率矩阵和词汇概率矩阵的分析可知,当训练语料规模在某一范围内时,训练后所得到的相邻的统计概率模型有较大的变化,系统的标注正确率也有较大的提高.但当训练语料的规模达到一定程度后,相邻模型的就越来越接近,变化的幅度也越来越小,模型趋于稳定.因此,系统的标注正确率的提高就非常缓慢,错误率的下降越来越小.从而使训练语料的规模大小与标注正确率的提高之间呈现出非线性关系.

3 对训练标注方法的改进

从上面的分析可以看出,当训练语料的规模达到一定程度后,仅仅通过扩充训练语料规模的大小来提高标注正确率是不适当的.因此,研究如何在训练语料的规模不变的情况下尽可能地提高标注正确率,就是一件很有意义的事情.我们在这方面作了一些尝试,从以下两个方面对原来的训练标注方法作了一些改进,试图充分利用已有的训练语料来改善标注性能.

3.1 根据词语的属性对有关的词性间的组合予以优先处理

我们知道,词类划分的依据是词的语法功能,但是词的语法功能只是一个词的语法属性的一个方面,仅仅从这一个方面来分析认识一个词是不够的.根据词类划分词语虽然简洁、清晰,信息密度大,但是属于同一词类的各个具体的词语的语法属性还是有差别的,只有从词语的语法属性的多个角度出发,才能比较全面地认识一个词.我们试图利用词语的有关语法属性来帮助对词语的词性的判断.在文献[4]中,不仅对词语按照词类进行了

© 中国科学院软件研究所 http://www.jos.org.cn

分类,而且还按类对每个词语的语法属性给予了详细描述,这就使我们有了正确的词语语法属性信息作为依据.

我们来分析一种产生错误标注的情况,"vgo(动词不带宾语)+n(名词)"与"vgn(动词带名词宾语)—n(名词)".在"新时期的领导干部要认真学习管理科学"与"作为一名班主任,她的任务是教育管理学生"这两句话中,"管理"一词前面的词的词性都是vgn,而后面所跟的词都是名词.但是,在第1句中,"管理"修饰的是"科学",其词性是vgo,而在第2句中,"学生"是"管理"的宾语,"管理"的词性应为vgn.因此,如果用原来的标注方法,肯定就会有一个被标注错.

在文献[4]中,对于名词词语有一项描述该名词词语能否受动词直接修饰(不带"的")构成定中结构的"前动"属性.在上述例子中,"科学"是属于可以受动词直接修饰的名词词语,而"学生"则属于不能受动词直接修饰的名词词语.我们还注意到,在文献[4]的动词库中,也有一条用于描述动词词语后面是否可以直接修饰名词以构成定中结构的"后名"属性.而"管理"一词恰恰具有这样的属性.这样,有理由相信,在第1句中,"管理"作为vgo修饰"科学"的概率值应该大于其在第2句中作为vgo修饰"学生"的概率.这样处理也有助于避免一些兼有动词词性的名词被错标成动词.由于动词与名词的兼类是汉语兼类现象中比重最大的一种情况,解决好这个问题对于提高标注正确率会很有好处.

因此,根据文献[4]中所提供的这样的属性,我们把原来名词中的两个子类:无量名词(不受任何量词修饰,nf)和一般名词(受名量词修饰,nr)按"前动"属性再分成无量名词_前动可(nfqdk)、无量名词_前动否(nfqdf)、一般名词_前动可(nrqdk)、一般名词_前动否(nrqdf)这4个小类,把动词不带宾语(vgo)再按"后名"属性分成动词不带宾语_后名可(vgohmk)和动词不带宾语_后名否(vgohmf)两个小类.在Viterbi算法中,当计算 $fai[t,j]=\max_{1\leqslant i\leqslant N_T}(fai[t-1,i]\times a_{ij})\times b_j(S_t)$ 时,我们增加如下的处理:

对于 a_{ij},若 T_i 是vgo,且 T_j 是nf或者nr,那么判断是否有 S_{t-1} 属于vgohmk,且 S_t 属于nfqdk或nrqdk;若是,则用 $a_{ij}\cdot\tau$ 替换 a_{ij} 进行运算,τ 是一个优先因子,用于放大vgohmk与nfqdk或者vgohmk与nrqdk之间的优先组合的概率关系,τ 的值可以通过对实验结果的比较来选取.

需要注意的是,采用这种方法后,对于某些具有歧义的"动词+名词"的结构,可能会降低它们的标注正确率.例如,对于"学习文件",它既可能是"学习"作为不带宾语的动词修饰"文件"构成定中结构,也可能是"文件"作为"学习"的宾语而构成"述宾结构".由于"学习"具有"后名"属性,"文件"也具有"前动"属性,所以,"学习"将更有可能被标注成vgo.但是,从最后的实验结果来看,利用词语的这种语法属性辅助进行词性标注,还是可以比较有效地提高标注正确率的.在以后的工作中,我们也将会继续这方面的分析,使这种方法更加完善.

3.2 改进与未知词相关的词性概率的估计

在本系统中,根据对语料库的统计,一般会有2~3%的词语属于未知词.从表1中可以看出,在开放测试下对未知词的标注正确率大约仅为50%.对于专业性较强的真实文本,未知词的出现率可能会更高,成为影响系统标注正确率的一个重要因素,因此,必须采取更加有效的手段对未知词进行标注.我们的做法是加强与未知词相关的词性概率的估计.

对于词串 $S=S_1S_2\ldots S_t\ldots S_M$,假设 S_t 是未知词,在计算从 S_{t-1} 到 (S_t,T_j) 的最佳路径的概率权值时,原来的做法是:

$$fai[t,j]=\max_{1\leqslant i\leqslant N_T}(fai[t-1,i]\times a_{ij})\times b_j(S_t).$$

现在,当 S_t 是未知词时,用 $P(T(UW)=T_j|T_i-(UW))$(UW 表示未知词UNKOWNWORD)替换原式中的 $a_{ij}=P(T_j|T_i)$ 再进行计算,$P(T(UW)=T_j|T_i-(UW))$ 是标记 T_i 后面出现的未知词的词性标记为 T_j 的概率值.由于将未知词看成是一个特殊的兼类词,因而用针对它的词性转移概率代替一般的词性转移概率值所对应的语言现象更少,可以更加准确地反映出不同词性对未知词所产生的影响,因此对于预测未知词的词性也会更加有效.

在计算 $P(T(UW)=T_j|T_i-(UW))$ 时,考虑到数据稀疏所带来的影响,采用插值估计法[3]进行参数平滑处理,对 $P(T(UW)=T_j|T_i-(UW))$ 的计算如下:

$$P(T(UW)=T_j|T_i-(UW))=\lambda_1P(T(UW)=T_j|T_i-(UW))+\lambda_2P(T(UW)=T_j).$$

© 中国科学院软件研究所 http://www.jos.org.cn

其中 $$P(T(UW)=T_j|T_i-(UW))=\frac{N(T_i-(UW),T(UW)=T_j)}{N(T_i-(UW))},$$

$$P(T(UW))=T_j=\frac{N(T(UW))=T_j}{N(UW)}.$$

$N(T_i-(UW),T(UW)=T_j)$是在训练语料中标记T_i后面出现的未知词的词性标记T_j的次数,$N(T_i-(UW))$标记T_i后面紧跟着出现未知词的次数,$N(T(UW)=T_j)$是未知词的标记为T_j的次数,$N(UW)$是训练语料中所有未知词的个数.

以上是从两个方面对原有的训练标注方法所进行的改进.我们的目的就是要充分利用已有的训练语料,改进统计训练和标注方法,从更细的角度来统计参数,更加准确地反映出语言现象中所存在的词语的语法功能的概率分布规律.改进后的方法没有对原来的训练和标注算法的时间或空间复杂度产生较大的影响,同样地,利用原来的训练和测试语料进行实验后所得到的结果见表6.

Table 6 Part-of-Speech tagging result after improving

表6 改进后的词性标注实验结果

Training corpora set (10 thousand)①	Tagging accuracy of open test②(%)	Tagging accuracy of close test③(%)	Tagging accuracy of unknown words on open test④(%)	Tagging accuracy of unknown words on close test⑤(%)
3	95.0	97.2	58.2	68.9
5	95.6	97.0	55.6	61.3
7	95.8	96.8	63.1	62.1
10	96.0	96.8	66.7	65.2
15	96.0	96.6	64.4	61.1
20	96.1	96.5	65.5	65.3

①训练语料规模(万),②开放测试标注正确率,③封闭测试标注正确率,

④开放测试中未知词的标注正确率,⑤封闭测试中未知词的标注正确率.

从表6中可以看出,与原来的方法相比,改进后在开放测试下,标注错误率下降了20%,未知词的错误率下降了18%;在封闭测试下,标注错误率下降了25%,未知词的错误率下降了22%.基于3万词的训练语料所获得的开放测试的正确率几乎达到了原来基于10万词的效果,有效地提高了标注的正确率.

4 结束语

本文从目前常用的基于统计的汉语词性标注方法出发,对不同训练语料规模下的实验结果,从词性概率矩阵与词汇概率矩阵的结构和数值变化等方面对训练语料规模与标注正确率之间所存在的非线性关系作了分析,并对其加以改进,得到一个增强的处理模式RF_Enhanced,有效地提高了自动标注的正确率.

结合工作中的体会,我们认为,对于汉语词性标注,可以综合运用词语的其他已知属性辅助词性标注.词语所包含的信息可以说是一个多维空间,不同属性之间会产生相互的影响.在进行词性标注时,如果只考虑词性与词性、词性与词语之间的关系是不全面的,词语的其他属性(不仅仅是语法属性)也可以用来辅助判断词性.

另外,可以通过对标注错误现象的统计分析,对一些易错的特殊词性或者词,根据它们的语法功能的特点,相应地加强对与它们有关的概率信息的统计,作特殊处理.甚至可以对一些现象进行总结,制订成规则,用于标注处理.事实上,我们认为,基于规则的标注方法与基于统计的方法两者之间并不矛盾,规则其实就是人们对一些发生的概率值比较大的语言现象的总结.对于基于统计的方法,由于受统计模型本身所固有的缺点和语料中其他语言现象的影响,训练后所得到的概率参数并不一定能够有效地反映出语言本身的一些确定性的语法特点,而利用规则却可以弥补这方面的缺陷.

在以后的工作中,我们将继续对这些问题进行研究,希望能充分利用语言学知识,结合统计分析方法,使标注性能获得更进一步的提高.

参考文献

1 Zhuo Qiang. Chinese corpus tagging using rule techniques and statistics techniques. Journal of Chinese Information

© 中国科学院软件研究所 http://www.jos.org.cn

Processing, 1995,9(2):1~10
(周强.规则与统计相结合的汉语词类标注方法.中文信息学报,1995,9(2):1~10)
2 Bernard Merialdo. Tagging English text with a probabilistic model. Computational Linguistics, 1994,20(2):155~171
3 Zhuo Qiang. Corpus-Based and statistics-oriented natural language processing techniques. Computer Science, 1995,22(4):36~40
(周强.基于语料库和面向统计学的自然语言处理技术介绍.计算机科学,1995,22(4):36~40)
4 Yu Shi-wen, Zhu Xue-feng, Wang Hui *et al*. The Detail of Modern Chinese Syntax Information Dictionary. Beijing: Tsinghua University Press, 1998
(俞士汶,朱学峰,王惠等.现代汉语语法信息词典详解.北京:清华大学出版社,1998)

Analysis and Improvement of Statistics-Based Chinese Part-of-Speech Tagging

WEI Ou WU Jian SUN Yu-fang

(*Institute of Software The Chinese Academy of Sciences Beijing* 100080)

Abstract In this paper, a popular statistics-based training and tagging method for Chinese texts is studied, and the nonlinear relation between training set and tagging accuracy is analyzed from the aspects of the structure and numerical value of the matrix of transition probabilities and the matrix of symbol probabilities. In order to make use of training corpus sufficiently and get the higher tagging accuracy, the training and tagging method is improved from two aspects: using other grammatical attributes of words, and strengthening the processing of unknown words. With the improved method, open test and close test showed that the overall accuracies are about 96.5% and 96% respectively.

Key words Part-of-Speech tagging, *n*-gram, corpus, grammatical attribute.

© 中国科学院软件研究所 http://www.jos.org.cn