基于条件随机场(CRFs)的中文词性标注方法

洪铭材 张 阔 唐 杰 李涓子
(清华大学计算机系知识工程组 北京 100084)

摘 要 本文提出一种基于CRFs模型的中文词性标注方法。该方法利用CRFs模型能够添加任意特征的优点,在使用词的上下文信息的同时,针对兼类词和未登录词添加了新的统计特征。在《人民日报》1月份语料库上进行的封闭测试和开放测试中,该方法的标注准确率分别为98.56%和96.60%。

关键词 词性标注,条件随机场,维特比解码

A Chinese Part-of-speech Tagging Approach Using Conditional Random Fields

HONG Ming-Cai ZHANG Kuo TANG Jie LI Juan-Zi
(Knowledge Engineering Lab, DCST, Tsinghua Univ., Beijing 100084)

Abstract This paper presents a new approach to part-of-speech (POS) tagging for Chinese texts using conditional random fields (CRFs). To take advantage of the ability of using arbitrary features as input in CRFs, not only contexts of words are exploited, but also are new statistical features adopted for multiple-category and out-of-vocabulary words. Closed and open tests conducted on People Daily dataset obtain POS tagging accuracies of 98.56% and 96.60%, respectively.

Keywords Part-of-speech tagging, Conditional random fields (CRFs), Viterbi decoding

1 引言

词性标注是自然语言处理的重要内容之一,是其他信息处理技术的基础,被广泛地应用于机器翻译、文字识别、语音识别、信息检索等领域。目前基于统计的词性标注方法得到了广泛的应用并取得了很好的效果。

在基于统计方法的词性标注中,对兼类词和未登录词的标注是两个需要解决的问题。对于兼类词,可以根据该词的上下文信息来确定该词在句子中的唯一词性。对于未登录词,能够获取关于该词的信息相对较少,可以根据词的上下文信息以及词的构词特点来确定其词性。在基于隐马尔可夫(HMM)模型的词性标注方法中,通常假设中心词的词性只与它前面的n个词有关,而与它后面的词无关。这个假设在词性标注任务中并不符合实际。最大熵模型(MEMM)能够充分利用词的上下文信息,但是存在着"label bias"[1]的弱点。

本文使用条件随机场(Conditional Random Fields, CRFs)[1]进行中文的词性标注。CRFs通过建立概率模型来进行序列数据的标注。与最大熵模型一样,CRFs是指数形式的模型,具有很强的推理能力,并且能够使用复杂、有重叠性和非独立的特征进行训练和推理。目前CRFs在信息抽取、命名实体识别、词组识别、语音句子边界识别等领域都表现出很好的性能。本文利用CRFs的特点,在进行词性标注时不但利用了词的上下文信息作为特征,而且充分利用了训练集的统计信息作为特征,为兼类词的标注提供了更多的特征信息。同时根据汉语的构词特点,使用词的后缀信息作为特征,在标注未登录词时起到了很好的效果。

本文使用PFR《人民日报》标注语料库作为实验数据。实验结果表明,基于CRFs的中文词性标注方法取得了很好的标注正确率,其封闭测试和开放测试的准确率分别为98.56%和96.60%,兼类词和未登录词的标注也取得了很好的结果。

文章的组织结构如下:第2节介绍中文词性标注的相关工作;第3节阐述CRFs理论及其训练方法;使用CRFs进行中文词性标注的方法将在第4节中介绍;第5节是实验结果和实验分析;最后是对本文的概括以及对未来工作的展望。

2 相关工作

词性是词的句法功能类别。在各种自然语言处理过程中,几乎都有一个词性标注的阶段。因此,词性标注的正确率将直接影响到后续的分析处理结果。基于其很高的重要性,词性标注一直是自然语言处理的重要内容。词性标注的方法大致可以分为3类:

①基于规则的方法。基于规则的方法是最早提出的词性标注方法,它手工编制包含繁杂的语法和/或语义信息的词典和规则系统。这种方法不仅费时费力,而且带有很大的主观性,难以保证规则的一致性。更大的问题是处理歧义长句、生词、不规范句子的能力非常脆弱,词性标注准确率不高。

②基于变换的方法。该方法由Eric Bill提出,用于标注英语的词性,其基本思想是利用一个带词性标注的语料库来例示实现设计好的模板,从一个已标注词性的语料库中统计每个词最可能的词性标记,然后用该标记标注训练语料库,称为初始标注,然后通过规则学习的方法获取新的规则。在进

洪铭材 硕士生,研究方向为信息抽取、信息检索;张 阔 博士生,研究方向为信息抽取、语义网络;唐 杰 博士生,研究方向为信息抽取、语义网络、信息集成;李涓子 副教授,研究方向为自然语言处理、语义网络。

行文本标注时，先对文本进行初始标注，然后按照规则获取的次序应用规则进行标注。该方法在英文词性标注上取得了很好的效果。其主要问题是学习规则的时间过长。针对这一问题，周明等提出一个快速学习算法，使训练速度大大提高[2]。

③基于统计的方法。基于统计的方法是目前应用最广泛的词性标注方法。白栓虎提出基于统计的汉语词性自动标注方法，使用二元语法模型和动态规划的方法进行汉语的词性标注。当前大部分汉语词性系统采用基于二元语法模型或三元语法模型的隐马尔可夫模型，通过 EM 训练的方法，给每个词和词性标签对分配一个联合概率，通过维特比解码的动态规划方法来获取一个句子对应的最佳的词性标注序列。隐马尔可夫模型的缺点是假设词的词性只与词本身和它前面的 n 个词有关，而与该词后面的词没有关系。这个假设与实际情况并不吻合。基于最大熵模型的词性标注方法，有效地利用了上下文信息，在一定的约束条件下可以得到与训练数据一致的概率分布，得到了很好标注效果。但是最大熵模型存在一种称为“label bias”问题的弱点[1]，在某些训练集上得到的模型可能会得到很差的标注效果。常见的基于统计的方法还有神经元网络、决策树、线性分离网络标注模型等。

3 CRFs 理论及其训练方法

3.1 CRFs 的定义

给定数据序列随机变量 X，CRFs C 定义了标注结果序列随机变量 Y 的条件概率分布 $p(Y|X)$，它通过训练的方法来使得条件概率 $p(Y|X)$ 最大。CRFs 是无向图模型，其最简单的形式是线性的 CRFs，即模型中各个节点之间构成线性结构。一个线性的 CRFs 对应于一个有限状态机，它非常适合于进行线性数据序列的标注。下面，如果不加说明，CRFs 均指线性的 CRFs。用 $x=(x_1, x_2, \cdots, x_n)$ 表示要进行标注的数据序列，$y=(y_1, y_2, \cdots, y_n)$ 表示对应的结果序列，并且假设 x，y 的长度相同。例如，x 可以表示一个中文句子 $x=$（彭，楚政，再，获，全国，民族，团结，进步，模范，称号。），y 则表示该句子中每个词的词性序列 $y=(nr, nr, d, v, n, n, a, v, n, n, w)$。

对于 (X, Y)，C 由局部特征向量 f 和对应的权重向量 λ 确定。每个局部特征可能是一个状态特征 $s(y,x,i)$，也可能是一个转移特征 $t(y,y',x,i)$，其中，y，y' 是标注的结果标签，x 是一个输入数据序列，i 是输入序列的某个位置。为了表示统一，用

$$s(y,y',x,i)=s(y',x,i)$$

$$s(y,x,i)=s(y_i,x,i)$$

$$t(y,x,i)=\begin{cases}t(y_{i-1},y_i,x,i) & i>1\\ 0 & i=0\end{cases}$$

来表示所有的状态特征 s 和转移特征 t。

对于输入数据序列 x 和标注结果序列 y，条件随机场 C 的全局特征表示为

$$F(y,x)\sum_i f(y,x,i) \tag{1}$$

其中 i 遍历输入数据序列的所有位置，$f(y,x,i)$ 表示在 i 位置时各个特征组成的特征向量。于是，CRFs 定义的条件概率分布为

$$p_\lambda(Y,X)=\frac{\exp[\lambda\cdot F(Y,X)]}{Z_\lambda(X)} \tag{2}$$

其中

$$Z_\lambda(X)=\sum_y\exp[\lambda\cdot F(y,x)] \tag{3}$$

给定一个输入数据序列 x，标注的目标就是找出其对应的最可能的标注结果序列 $\bar{y}$，即

$$\bar{y}=\arg\max_y p_\lambda(y|x) \tag{4}$$

由于 $Z_\lambda(X)$ 不依赖于 y，因此有 $\bar{y}=\arg\max_y p_\lambda(y|x)=\arg\max_y\lambda\cdot F(y,x)$。与隐马尔可夫模型相似，CRFs 使用维特比解码（Viterbi decoding）方法来得到最佳的标注结果序列。

CRFs 具有很强的推理能力，并且能够使用复杂、有重叠性和非独立的特征进行训练和推理，能够充分地利用上下文信息作为特征，还可以任意地添加其他外部特征，使得模型能够获取的信息非常丰富。同时，CRFs 解决了最大熵模型中的“label bias”问题。CRFs 与最大熵模型的本质区别是：最大熵模型在每个状态都有一个概率模型，在每个状态转移时都要进行归一化。如果某个状态只有一个后续状态，那么该状态到后续状态的跳转概率即为 1。这样，不管输入为任何内容，它都向该后续状态跳转。而 CRFs 是在所有的状态上建立一个统一的概率模型，这样在进行归一化时，即使某个状态只有一个后续状态，它到该后续状态的跳转概率也不会为 1，从而解决了“label bias”问题。因此，从理论上讲，CRFs 非常适用于中文的词性标注。

3.2 CRFs 的模型训练

CRFs 通过训练来使得条件概率 $p_\lambda(y|x)$ 最大。可以使用一种称为 log-likelihood[3] 的方法对条件随机场进行训练。给定一组固定的训练集合 $T=\{(x_k,y_k)\}_{k=1}^N$ 通过调整权重向量 λ 来最大化对数－可能性 L_λ

$$L_\lambda=\sum_k\log p_\lambda(y_k,|x_k)=\sum_k[\lambda\cdot F(y_k,x_k)-\log Z_\lambda(x_k)] \tag{5}$$

为了找出 L_λ 的最大值，对 L_λ 进行微分，得到

$$\nabla L_\lambda=\sum_k[F(y_k,x_k)-E_{p_\lambda(Y|x_k)}F(Y,x_k)] \tag{6}$$

也就是说，当全局特征向量的平均值等于其模型的数学期望时，L_λ 取得最大值。数学期望 $E_{p_\lambda(Y|x_k)}F(Y,x_k)$ 可以通过向前-向后算法（forward-backward algorithm）的变种快速地计算出来。对于某个给定的输入数据序列 x，定义其在位置 i 的转移矩阵如下：

$$M_i[y,y']=\exp\lambda\cdot f(y,y',x,i)$$

令 f 为局部特征，$f_i[y,y']=f(y,y',x,i)$，$F(y,x)=\sum_i f(y_{i-1},y_i,x,i)$，并令×表示实数与矩阵相乘，于是：

$$E_{p_\lambda(Y|x)}F(Y,x)=\sum_y p_\lambda(y|x)Fy,x)=\sum_i\frac{\alpha_{i-1}(f_i\times M_i)\beta_i^T}{Z_\lambda(x)} \tag{7}$$

$$Z_\lambda(x)=\alpha_n\cdot 1^T \tag{8}$$

其中 α_i，β_i 分别定义如下：

$$\alpha_i=\begin{cases}\alpha_{i-1}M_i & 0<i\leqslant n\\ 1 & i=0\end{cases},\beta_i^T=\begin{cases}M_{i+1}\beta_{i+1}^T & 1\leqslant i<n\\ 1 & i=n\end{cases}$$

因此，对于每个数据序列，可以通过一次向前扫描和一次向后扫描计算出 α_i，β_i，从而计算出数学期望。

4 词性标注

使用 CRFs 进行中文词性标注的过程就是给定一个中文句子 $x=(x_1, x_2, \cdots, x_n)$，通过维特比解码算法找出其对应

的词性标注结果序列 $y=(y_1, y_2, \cdots, y_n)$，使得条件概率 $p_\lambda(y|x)$ 最大。为了计算条件概率，本文利用词的上下文信息作为词的特征之一。相对于隐马尔可夫模型只能利用中心词的前 n 个词作为上下文信息的弱点，CRFs 能够同时使用中心词的前 n 个词和后 m 个词作为该词的上下文信息。这样，中心词的词性不仅与它前面的词有关，还与它后面的词有关，更加符合实际情况。在本文中，使用了中心词本身，中心词前一个词，中心词后一个词，以及它们之间的不同组合构成了 6 个不同的特征，每个特征的权值都设为 1。

CRFs 最大的优点之一就是它能够加入任意的特征作为输入。因此，为了向模型提供关于词更多的信息，充分利用训练集的统计信息和中文的构词特点，为每个词添加了新的统计特征。

通过对训练集的统计，可以得到训练集中每个词的词性。对于在训练集中只对应一种词性的词，可以认为它为非兼类词，为该词添加新的特征 T：T＝该词对应的词性，该特征的权值为 1；对于在训练集中对应多个词性（假设为 N）的词，则它一定为兼类词，统计出在训练集中该词对应每个不同词性出现的次数 C_i，$i=1,2,\cdots,N$。找出出现次数最多的词性 k，即 k 满足 $C_k=\max_i C_i$，新添加的特征 T 为 T＝词性 k，该特征的权重为 $\frac{C_k}{\sum_i C_i}$。

对于未登录词，由于其在训练集中能够获取的信息很少，因此可以考虑词的构词特点，利用该词的后缀信息作为新的特征。通过对中文语料的分析，发现词的构词特点与词的词性有一定的联系，例如“镇”、“县”和“市”等一般用在词尾构成地名，“所”、“院”等一般用在词尾构成机构名等。本文通过词的后缀最长匹配，得到未登录词最可能的词性 i，新添加的特征 T 为：T＝词性 i。由于该特征具有较大的不可靠性，因此赋予该特征的权值为 0.5。在词性标注时有一类词的词性为成语，因此通过互联网收集了一个成语列表。如果要进行标注的未登录词出现在成语列表中，则添加的特征 T 为：T＝成语，且该特征的权值为 1。

通过添加新的特征，得到了系统最终的词性标注特征模板，如表 1 所示。

表 1　词性标注特征模板

特征	说明	权值		
$W=W_0$	中心词	1		
$W=W_{-1}$	中心词的前一个词	1		
$W=W_1$	中心词的后一个词	1		
$W=W_{-1}W_0$	中心词的前一个词及其本身	1		
$W=W_0W_1$	中心词及其后一个词	1		
$W=W_{-1}W_0W_1$	中心词前一个词，本身和后一个词	1		
T＝(tag)	中心词可能的词性	非兼类词	兼类词	未登录词
		1	$\frac{C_k}{\sum_i C_i}$	0.5(词性是成语时为 1)

特征 T 利用训练集的统计信息和中文构词的特点，在确定中心词的词性时应当能起到很好的作用。为了检验特征 T 的作用，在实验中针对特征 T 进行了不同的实验。

5　实验结果和分析

5.1　数据描述与实验设置

本文采用 PFR《人民日报》标注语料库中 1 月份语料作为实验数据。PFR《人民日报》标注语料库以 1998 年《人民日报》语料为对象，由北京大学计算语言学研究所和富士通研究开发中心有限公司共同制作[4]。在实验中，将整个语料集随机分割为两个部分：一部分作为训练集以及封闭测试集，另一部分作为开放测试集。训练集/封闭测试集的句子数和词数分别是开放测试集句子数和词数的 2.42 倍和 2.37 倍。两个集合的具体信息如表 2 所示。

从表 2 可以看到，在训练集/封闭测试集，兼类词出现总次数占集合词数的 46.56%，在开放测试集中兼类词出现总次数则占集合词数的 39.45%。因此，对兼类词标注的结果好坏将直接影响到这个词性标注系统的性能。

表 2　训练集/封闭测试集与开放测试集描述

	训练集/封闭测试集	开放测试集
句子数	31,668	13,110
不同词个数	33,070	20,162
总词数	776,817	327,991
不同非兼类词数	29,415	18,126
非兼类词出现总次数	415,149	198,596
不同兼类词数	3,655	2,036
兼类词出现总次数	361,668	129,395
不同未登录词数	—	1,618
未登录词出现总次数	—	15,417

在开放测试集中，未登录词出现总次数占集合词数的 4.70%。由于训练集和测试集均来自《人民日报》，在实际应用时，未登录词的比例可能要高于这个比例，因此标注未登录词的结果将对系统的实用性具有很大的意义。

在进行词性标注时，根据规范[4]，给每个词标上 39 种不同词性中的一种。在训练集/封闭测试集和开放测试集中，不同词性的词出现的次数差别非常大。表 3 给出了训练集/封闭测试集与开放测试集中不同词性的词出现的次数情况。

从表 3 可以看出，训练集/封闭测试集与开放测试集在词性分布上基本上相同，说明了这两个集合具有一定的代表性，能够反映中文词性的分布特点。其中，名词、动词、标点符号、助词、副词、数词、介词、动名词和形容词等出现的比例最大，而叹词、语素、前接成分、后接成分和副语素等出现的比例非常低（出现次数仅为几十次）。

本文分别进行了封闭测试和开放测试两组实验，其中封闭测试使用的数据集与训练集完全一致。为了检验特征 T 的作用，本文进行了两组实验：第一组实验采用的是只包含词的上下文信息为特征的特征模板，称为特征模板 A；第二组实验使用第 4 节中的特征模板，称为特征模板 B。最后，本文比较了 HMMs、MEMMs 和 CRFs 在上述数据集上的标注效果。

为了评价实验结果，本文采用标注正确率对标注结果进行评价，其中又分为总标注正确率、非兼类词标注正确率、兼类词标注正确率和未登录词标注正确率。总标注正确率的定义是（所有正确标注的词个数 / 测试集中词个数）* 100%，其他正确率的定义与此相似。

表 3　训练集/封闭测试集与开放测试集按照词性统计结果

标注符号	词性名称	比例(%)		标注符号	词性名称	比例(%)	
		训练集	开放测试集			开放测试集	训练集
Ag	形语素	0.09	0.14	Ns	地名	2.46	2.58
A	形容词	3.17	3.04	Nt	机构团体	0.31	0.35
Ad	副形词	0.56	0.48	Nz	其他专名	0.30	0.41
An	名形词	0.24	0.30	O	拟声词	0.00	0.01
B	区别词	0.77	0.77	P	介词	3.64	3.54
C	连词	2.30	2.32	Q	量词	2.21	2.16
Dg	副语素	0.01	0.01	R	代词	2.94	2.89
D	副词	4.30	4.26	S	处所词	0.35	0.35
E	叹词	0.00	0.00	Tg	时语素	0.04	0.05
F	方位词	4.31	1.54	T	时间词	1.83	1.96
G	语素	0.00	0.00	U	助词	6.74	6.82
H	前接成分	0.00	0.00	Vg	动语素	0.16	0.18
I	成语	0.42	0.46	V	动词	16.94	16.22
J	简称略语	0.92	0.95	Vd	副动词	0.05	0.04
K	后接成分	0.00	0.00	Vn	名动词	3.97	3.57
L	习用语	0.56	0.55	W	标点符号	14.04	14.65
M	数词	3.81	3.58	X	非语素字	0.04	0.03
Ng	名语素	0.40	0.43	Y	语气词	0.16	0.19
N	名词	21.86	20.86	Z	状态词	0.11	0.15
Nr	人名	2.80	4.12				

5.2　CRFs 词性标注实验结果与分析

封闭测试和开放测试的实验结果如表 4 所示。

表 4　词性标注实验结果

	特征模板 A		特征模板 B	
	封闭测试	开放测试	封闭测试	开放测试
非兼类词标注正确率	98.31%	97.56%	99.20%	98.61%
兼类词标注正确率	96.03%	84.25%	97.82%	91.64%
未登录词标注正确率	—	76.95%	—	81.67%
总标注正确率	97.25%	92.31%	98.56%	96.60%

从表 4 实验结果可以得到以下几个结论：

①使用 CRFs 进行中文词性标注能够取得很好的标注效果。例如使用特征模板 B 时，在封闭测试中总标注正确率为 98.56%，而在开放测试中总标注正确率也达到 96.60%。这验证了第 3 节中的理论分析，即 CRFs 非常适合中文词性标注。

②非兼类词标注正确率高于兼类词标注正确率，特别是在进行开放测试中，使用特征模板 A 时前者要比后者高 13.31%，而使用特征模板 B 时前者比后者高 6.97%。同时也可以看出，使用特征模板 B 时，非兼类词和兼类词标注的正确率都比使用特征模板 A 时高，其中在开放测试中，兼类词标注正确率提高了 7.39%。这说明了特征模板 B 能够有效地提高词性标注的正确率，特别是兼类词标注的正确率。

③相对于特征模板 A，特征模板 B 能够提高未登录词标注正确率。使用特征模板 B 时，未登录词标注正确率比使用特征模板 A 时的正确率高 4.72%。

④未登录词标注正确率相对较低。由于系统能够从未登录词本身获取的信息比较少，导致了未登录词标注的正确率要远远低于在训练集中出现的词。该问题的可能解决方法是增大训练集的数据量，使得在实际应用时未登录词出现尽量少。

从结论 3、4 可以看出，含有特征 T 的特征模板 B 的标注正确率高于不包含特征 T 的特征模板 A，这说明了特征 T 在进行词性标注时有很好的预示作用，验证了第 4 节中提出的方法是有效的。

在实验中还统计了不同词性的词进行标注后的查准率、查全率和 F1 值。限于篇幅，本文仅选择列出其中部分词性在使用模板 B 进行标注后的详细结果，见表 5。

将表 3 与表 5 相对照可以发现，出现比例越高的词性，其 F1 值一般都越高。这是因为 CRFs 是基于统计的模型，对于出现比例越高的词性，统计的信息越全面，其 F1 一般也会越高。但是也存在着例外，例如标点符号的 F1 值到达了 100%，说明几乎所有的标点符号都标注正确。这是因为标点符号既不是兼类词，而且训练集中包含了所有的标点符号，它也不是未登录词，所以能获得很好的标注效果。从标注结果看，在集合中出现比例高于 3%的词性的 F1 值，在封闭测试中除了动名词之外都达到 96%以上，在开放测试中除了动名词和形容词之外都达到 94%以上，而这些词的个数之和分别占了封闭测试集和开放测试集中词总数的 78.47%和 76.54%，使系统能够获得很好的性能。

表 5　不同词性的标注结果

词性名称	封闭测试			开放测试		
	查准率%	查全率%	F1 值%	查准率%	查全率%	F1 值%
形语素	95.87	92.55	94.18	92.62	83.66	87.91
形容词	96.80	97.01	96.91	91.29	91.89	91.59
副语素	82.26	57.95	68.00	54.55	64.86	59.26
名词	99.26	99.38	99.32	98.31	98.65	98.48
动词	97.53	97.65	97.59	93.88	94.95	94.91
名动词	93.27	92.27	92.65	84.84	80.59	82.66
助词	99.81	99.91	99.86	99.39	99.75	99.57
标点符号	100.00	100.00	100.00	100.00	100.00	100.00

在两组实验中，占比例较大的动名词的标注效果都比较差，而动词相对于名词。标点符号和助词的效果也较差，这可以解释为动名词和动词之间比较容易混淆，导致了这两个词性的标注效果都有所下降。

（下转第 155 页）

音占 78.2%。

基于该商业语音识别系统及其接口，本文搭建了一个汉语语音识别纠错系统，通过构筑声韵混淆音矩阵和基于语句语义分析的纠错方法，可以对语音识别结果中的错误进行纠错处理。使用 216 句测试语句对纠错处理系统进行测试，其中有 95 句含有语音识别错误。纠错处理系统发现 74 句错误语句，错误发现率为 78%。在发现错误的语句中，纠错系统纠错后语句正确的有 53 句，纠错的正确率为 72%。

测试结果表明，纠错系统可以提高语音识别系统的识别正确率，尤其对由于错音引起的识别错误或语义搭配型错误纠错，效果较好。这主要是因为该商业语音识别系统采用了 Tri-gram 统计语言模型，这种语言模型将人类的语言感知过程作为一个黑箱来处理，用概率(或频度)的方法来模拟。如果遇到的词串是训练语料中出现过的、高频的，会给出很好的结果；否则，只能根据概率最佳给出结果。这种模型和方法即使在最好的情况下，也仅仅只是给出了词语紧邻的搭配知识，得到的数据不能全面反映词语间的语义关联知识。

N-gram 语言模型无法解决非常用词的数据稀疏问题。语句中词语之间的关联不是线性连续的关系，而是一种多层级的关系。N-gram 语言模型无法具体区分这些不同，都简单地用概率给予表达，无法给出符合概念联想脉络的词语，也无法判断给出的结果是否正确，只能得到一个概率最优的结果。

本纠错系统所依据的概念层次网络模型可以克服 N-gram 语言模型所固有的缺陷，揭示语句中词语之间的概念关联性知识，通过知识库和规则库对词语进行正确有效的语义约束。这种语义约束关系正好可以用来对语音识别系统产生的错误进行语义纠错。

结束语 语音识别系统所处理的语音信号具有两个特点：一个是存在环境噪声，而且环境噪声的声学特性与人的语音相近；第二个是不同的发音人语音具有较大的差异，即使是同一个人，在不同时间的语音都存在一定的差异。对于连续语音识别，由于发音的连贯性，音位和音位之间的连续平滑过渡必然造成识别中音与音、字与字、词与词之间的分割困难。目前，仅凭单纯的信号处理已经很难提高语音识别系统的正确率，语音识别正确率的提高越来越取决于语音识别的后期处理和语言理解模型。

基于 N-gram 的统计语言模型便于处理大规模的、经常出现的语言现象，但是对于语言深层次的语义约束等问题却难以解决。本文基于概念层次网络语言模型，提出了一种基于语句语义分析的语音识别纠错方法。实验表明，该方法在纠正因语义约束而产生的错误方面有很好的效果。如果能够把这种方法和语音识别系统相融合，将可以提高语音识别的正确率。关于如何根据上下文语境、利用超语句的更大范围的语言知识来对语音识别的错误进行纠正，则是本文进一步的研究方向。

参 考 文 献

1 黄曾阳. HNC(概念层次网络)理论[M]. 北京：清华大学出版社，1998
2 晋耀红. 基于 HNC 理论的句类分析系统的设计与实现[D]：[硕士学位论文]. 北京：中国科学院声学研究所，1998
3 孙伟峰. 基于句类的指代解析及其在语音识别中的应用[D]：[硕士学位论文]. 北京：中国科学院北京软件工程研制中心，2000
4 苗传江. HNC 句类知识研究[D]：[博士学位论文]. 北京：中国科学院声学研究所，2001
5 王轩，等. 语音识别中统计与规则结合的语言模型[J]. 自动化学报，1999，25(3)：309～315
6 关毅，等. 现代汉语计算语言模型中语言单位的频度-频级关系[J]. 中文信息学报，1999，13(2)：8～15
7 赵力，等. 汉语连续语音识别中语音处理和语言处理统合方法的研究[J]. 声学学报，2001，26(1)：73～78.

(上接第 151 页)

5.3 HMMs，MEMMs 和 CRFs 的比较

第 3 节从理论上说明了 CRFs 的优越性。为了验证这一结论，本文在相同的数据集上分别采用 HMMs，MEMMs 和 CRFs 进行了词性标注，所有的标注均遵照规范[4]。其中，HMMs 使用一阶的马尔科夫模型，MEMMs 和 CRFs 分别使用了模板 A 和模版 B 进行实验。开放测试实验的结果如表 6 所示。

表 6 HMMs，MEMMs 和 CRFs 词性标注结果比较

模型	HMMs	MEMMs		CRFs	
		模板 A	模板 B	模板 A	模板 B
总准确率	92.03%	91.71%	95.65%	92.31%	96.60%

由表 6 可以看出，当采用模板 A 时，标注的效果按照高到低的顺序依次为 CRFs，HMMs，MEMMs。这是因为 MEMMs 存在着 Label Bias 问题，因此效果最差，而 CRFs 效果最好。当采用模板 B 时，MEMMs 的效果好于 HMMs，而 CRFs 的效果比 MEMMs 好，这主要得益于 CRFs 和 MEMMs 都能够采用任意的特征来构造模板，从而为模型标注效果的提高提供了可能性和方便性。

结论和展望 本文使用 CRFs 对中文进行词性标注。除了使用词的上下文信息之外，对于兼类词，利用词在训练集中的统计信息，得到词最可能的词性，为特征模板添加新的特征；对于未登录词，利用中文词的构词特点以及成语词典列表生成新的特征。在实验中，对实验数据和实验结果进行了统计分析。实验结果表明，使用 CRFs 的中文词性标注方法能够取得很好的标注效果。在相同的条件下标注效果优于 HMMs 和 MEMMs。

目前，对未登录词的标注效果还不是非常理想，今后的工作将针对未登录词，深入地调研汉语的构词特点对词性的影响，以及不同词性的词各自的特点，使用更多的特征，进一步提高标注的正确率。

参 考 文 献

1 Lafferty J，McCallum A，Pereira F. Conditional Random Fields：Probabilistic Models for Segmenting and Labeling Sequence Data. In：Proceedings of the 18th International Conf on machine Learning，2001. 282～289
2 周明，吴进，黄昌宁. 用于词性标注的一种快速学习算法对 Brill 的基于变换算法的一项改进. 计算机学报，1998(4)：357～366
3 Sha F，Pereira F. Shallow Parsing with Conditional Random Fields. In：Proceedings of Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics (HLT-NAACL)，2003
4 现代汉语语料库加工规范-词语切分与词性标注. 北京大学计算语言学研究所，1999
5 白栓虎. 基于统计的汉语词性自动标注方法. 语文建设，1994(10)：38～40
6 Bai Shuanhu. An Integrated Model of Chinese Word Segmentation and Part-of Speech Tagging. In：Advanced and Applications on Computational Linguistics，Third National Computational Linguistics Meeting，Shanghai. Nov. 1995. 56～61
7 Bai S H，Xia，Y，Huang C N. Automatic Part-of-Speech Tagging System of Chinese：[Technical Report]. Beijing：Tsinghua University，1992