ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.20, No.4, April 2009, pp.870−877
doi: 10.3724/SP.J.1001.2009.03233

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于大间隔方法的汉语组块分析*

周俊生[1,2+], 戴新宇[1], 陈家骏[1], 曲维光[2]

[1](南京大学 计算机软件新技术国家重点实验室,江苏 南京 210093)

[2](南京师范大学 计算机科学系,江苏 南京 210097)

Chinese Chunking with Large Margin Method

ZHOU Jun-Sheng[1,2+], DAI Xin-Yu[1], CHEN Jia-Jun[1], QU Wei-Guang[2]

[1](State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China)

[2](Department of Computer Science, Nanjing Normal University, Nanjing 210097, China)

+ Corresponding author: E-mail: zhoujs@nlp.nju.edu.cn, http://www.nju.edu.cn

Zhou JS, Dai XY, Chen JJ, Qu WG. Chinese chunking with large margin method. *Journal of Software*, 2009, 20(4):870−877. http://www.jos.org.cn/1000-9825/3233.htm

Abstract: Chinese chunking plays an important role in natural language processing. This paper presents a large margin method for Chinese chunking based on structural SVMs (support vector machines). First, a sequence labeling model and the formulation of the learning problem are introduced for Chinese chunking problem, and then the cutting plane algorithm is applied to efficiently approximate the optimal solution of the optimization problem. Finally, an improved *F*1 loss function is proposed to tackle Chinese chunking. The loss function can scale the *F*1 loss value to the length of the sentence to adjust the margin accordingly, leading to more effective constraint inequalities. Experiments are conducted on UPENN Chinese Treebank-4 (CTB4), and the hamming loss function is compared with the improved *F*1 loss function. The experimental results show that the training algorithm with the improved *F*1 loss function can achieve higher performance than the Hamming loss function. The overall *F*1 score of Chinese chunking obtained with this approach is 91.61%, which is higher than the performance produced by the state-of-the-art machine learning models, such as CRFs (conditional random fields) and SVMs models.
Key words: Chinese chunking; large margin; discriminative learning; loss function

摘 要: 汉语组块分析是中文信息处理领域中一项重要的子任务.在一种新的结构化 SVMs(support vector machines)模型的基础上,提出一种基于大间隔方法的汉语组块分析方法.首先,针对汉语组块分析问题设计了序列化标注模型;然后根据大间隔思想给出判别式的序列化标注函数的优化目标,并应用割平面算法实现对特征参数的近似优化训练.针对组块识别问题设计了一种改进的 *F*1 损失函数,使得 *F*1 损失值能够依据每个句子的实际长度进行相应的调整,从而能够引入更有效的约束不等式.通过在滨州中文树库 CTB4 数据集上的实验数据显示,基于改进

* Supported by the National Natural Science Foundation of China under Grant Nos.60673043, 60773173 (国家自然科学基金); the National High-Tech Research and Development Plan of China under Grant No.2006AA01Z143 (国家高技术研究发展计划(863)); the Natural Science Foundation of Jiangsu Province of China under Grant No.BK2006117 (江苏省自然科学基金); the Natural Science Foundation of Jiangsu Higher Education Institutions of China under Grant No.07KJB520057 (江苏省高校自然科学基金)

Received 2007-03-13; Accepted 2007-11-05

的 $F1$ 损失函数所产生的识别结果优于 Hamming 损失函数,各种类型组块识别的总的 $F1$ 值为 91.61%,优于 CRFs(conditional random fields)和 SVMs 方法.

关键词: 汉语组块分析;大间隔;判别式学习;损失函数

中图法分类号: TP391 **文献标识码**: A

汉语组块分析作为一种预处理手段,可以大大降低进行汉语短语划分和短语分析处理的复杂性,为进一步对句子的深层次分析提供了基础,使得句法分析任务在某种程度上得以简化.同时,组块分析可以应用到机器翻译、信息提取、信息检索和专有名词识别等多个领域,汉语组块分析的准确性将直接关系到后继文本分析与文本处理的正确性.

自从国际会议 Conll-2000 把组块识别作为共享任务提出之后,组块分析研究已经受到了广泛的关注,尤其是各种机器学习方法的应用取得了较好的效果.

在汉语的组块分析方面,文献[1,2]使用基于转换的机器学习方法(transformation based learning,简称 TBL)实现汉语的基本名词短语和组块识别,文献[3]实现了一个基于最大熵模型的组块标注器识别汉语中各种类型的组块,文献[4]提出了基于 SVMs(support vector machines)的汉语组块分析方法,文献[5]将条件随机场模型(conditional random fields,简称 CRFs)应用于汉语的组块识别问题,都取得了较好的效果.但他们或者使用了不同的语料,或者对于汉语组块的定义不尽相同,故而无法对不同的方法进行直接的比较.

近来,文献[6]基于同样的数据集(滨州中文树库的中文树库 CTB4),分别应用当前几种主流的模型和方法(包括 SVMs,CRFs,TBL 和基于记忆的学习方法 MBL(memory-based learning)等)对汉语的组块分析进行了实验性研究和比较,其实验结果显示,SVMs 和 CRFs 模型在汉语组块识别方面取得了最好的识别效果.文献[7,8]也分别表明,在英文的组块识别方面,SVMs 和 CRFs 也取得了目前领先的识别效果.其中在应用 SVMs 方法进行汉语组块分析时,需要建立多个(276 个)不同的分类器分别预测组块标记,最终的决策通过带权的投票决定.这种方法虽然取得了较好的汉语组块识别效果,但其计算复杂度和开销较大,而且这种传统的 SVMs 模型并不能充分利用序列化的结构特征.

在应用单个模型解决汉语组块分析问题时,CRFs 取得了最好的识别效果.CRFs 是一种判别式训练的概率无向图模型,它具有表达元素之间的长距离依赖性特征的能力以及较好地解决了标注偏置问题等特点[9].但 CRFs 模型采用极大似然的参数估计方法,且仅仅局限于 log-0/1 损失函数.

本文在一种新的结构化模型——结构化 SVM[10]的基础上,提出一种基于大间隔(margin)方法的汉语组块分析方法.我们首先针对汉语组块识别问题设计了序列化标注模型,然后根据大间隔思想给出判别式的序列化标注函数的优化目标,并应用割平面算法实现对特征参数的近似优化训练.针对组块分析问题,文中设计了一种改进的 $F1$ 损失函数.通过在 LDC 的 CTB4 数据集上的实验数据显示,本文所提出的汉语组块分析方法对各种类型组块识别的总的 $F1$ 值为 91.61%,整体效果优于其他汉语组块分析方法.

1 汉语的组块分析任务

根据 Abney[11]对组块的定义,组块是一种语法结构,是符合一定语法功能的非递归短语.每个组块都有一个中心词,组块内的所有成分都围绕该中心词展开,任何一种类型的组块内部不包含其他类型的组块.

在宾州中文树库 CTB4 的基础上[12],类似于文献[6],我们总共也定义了 12 种汉语组块类型:ADJP,ADVP,CLP,DNP,DP,DVP,LCP,LST,NP,PP,QP,VP.各种组块类型的具体含义见表 1.

下面给出一个文本组块的例子:

[NP 外商 投资 企业] [VP 成为] [NP 中国] [NP 外贸] [ADJP 重要] [NP 增长点]

当然也可以通过为组块加标记的方法来表示组块.我们也引入了 IOB 标注方法,即将句子中的每个词标注以组块类型和 IOB 标记的组合.其中,B 表示一个组块的开始,I 表示一个组块的内部,O 表示组块以外的其他位置,这样一共生成 25 个不同的标记,如 B-NP,I-NP,B-PP,I-PP 等等.于是上述的句子也可以表示如下:

外商 B-NP 投资 I-NP 企业 I-NP 成为 B-VP 中国 B-NP 外贸 B-NP 重要 B-ADJP 增长点 B-NP. O

这样,汉语的组块识别问题就转化成为一个序列化标注问题.

Tabel 1 Definition of Chinese chunks

表 1 汉语组块类型定义

Type	Definition	Example
ADJP	Adjective phrase	[ADJP 非排他性/JJ 和/CC 非歧视性/JJ]
ADVP	Adverbial phrase	[ADVP 积极/AD 、/PU 及时/AD]
CLP	Classifier phrase	[CLP 港元/M/ 与/CC 美元/M]
DNP	Deg phrase	[DNP 的/DEG]
DP	Determiner phrase	[DP 这些/DT]
DVP	DEV phrase	[DVP 平等/VA 和睦/VA 地/DEV]
LCP	Localizer phrase	[LCP 90年代/NT 初/LC]
LST	List marker	[LST (/PU 一/CD)/PU]
NP	Noun phrase	[NP 科技/NN 和/CC 教育/NN 事业/NN]
PP	Prepositional phrase	[PP 经过/P 了/AS]
QP	Quantifier phrase	[QP 四百九十/CD 克/M]
VP	Verb phrase	[VP 事先/AD 只/AD 准备/VV 了/AS]

2 汉语组块的序列化标注模型

将汉语的组块分析转化为一个序列化标注问题之后,我们的目标就是学习一个函数 f,用于从观察序列 $x=(x_1,x_2,\ldots,x_t,\ldots)$映射到一个同样长度的标号序列 $y=(y_1,y_2,\ldots,y_t,\ldots)$,其中,每个标号取自于一个标号集合$\Sigma$,即 $y_t\in\Sigma$.让 Y 表示所有可能的标号序列 y 的集合.构造函数 f 的一个重要任务就是需要学习一个基于输入/输出对的判别式(discriminant)函数 $F:X\times Y\to R$,通过对输出变量的最大化,实现对输出结果的预测.因此,目标函数的一般形式为

$$f(x)=\arg\max_{y\in Y}F(x,y;w) \tag{1}$$

其中,F 是基于输入/输出组合特征表示$\Phi(x,y)$的线性函数,即 $F(x,y;w)=\langle w,\Phi(x,y)\rangle$.而构造合适的参数判别式函数 F 必须要求设计一个映射Φ从每一个观察/标注序列对(x,y)中抽取相应的特征值.由于汉语组块分析问题中共包含了 25 个不同的标号,为了有效减少特征的数量,我们采用了不同于条件随机场中判别式特征表示方法,类似于文献[13],我们将特征表示$\Phi(x,y)$分解为两种类型的特征.这样在句子中的一个特定位置上的特征$\Phi(x,y;t)$可定义为 $\Phi(x,y;t)=\phi^{\beta}(x,y;t)+\varphi^{\alpha\beta}(y;t),\alpha,\beta\in\Sigma$.其中的第 1 个特征项是表示在位置 t 上取特定标注值时的观察值特征.若令$\psi(x^t)$为观察值 x^t 的特征表示,则该特征项可形式化定义为

$$\phi^{\beta}(x,y;t)=I(y^t=\beta)\psi(x^t),\ \beta\in\Sigma \tag{2}$$

I 是一个指示函数(indicator function).第 2 个特征项是表示在位置 t 时标注值之间的转移特征(假设存在一阶马尔可夫独立性),其形式化定义为

$$\varphi^{\alpha\beta}(y;t)=I(y^{t-1}=\alpha\wedge y^t=\beta),\ \alpha,\beta\in\Sigma \tag{3}$$

最后通过将在一个序列中的不同位置所抽取的特征累加在一起就构成整个序列的特征,如公式(4)所示:

$$\Phi(x,y)=\sum_{t=1}^{T}\Phi(x,y;t) \tag{4}$$

3 基于大间隔的判别式参数估计

传统的特征参数 w 的判别式训练方法是基于概率的极大似然方法.例如,条件随机场模型中定义了一个条件对数似然模型[9]:

$$P_{\Lambda}(s\mid o)=\frac{1}{Z_o}\exp\left(\sum_{t=1}^{T}\sum_{k=1}^{K}w_k f_k(s_{t-1},s_t,o,t)\right) \tag{5}$$

其中,$Z_0=\sum_s\exp(\sum_{t=1}^{T}\sum_k w_k f_k(s_{t-1},s_t,o,t))$ 是归一化因子.基于极大似然方法的特征参数获取模型为

$$\Lambda=\arg\max_{\Lambda} P_{\Lambda}(s|o).$$

其挑选模型的标准是模型与训练数据的拟合性,即模型要最大可能地解释训练集 O,而通常情况下,提供给学习器的数据只是目标语料的一小部分,于是,依据最大似然原则获取的模型参数虽然能够很好地解释训练语料中的数据,但是对训练语料以外的数据的解释能力很弱,从而会导致训练的过拟合(overfitting)问题.另外,极大似然的训练方法仅局限于log-0/1损失函数,而对于组块分析这样的结构化预测问题,简单的log-0/1损失显然不能更好地表达训练目标,因为它仅考虑所预测的词序列与正确标注的词序列是否完全一致,而不能表达两者之间不一致性的程度,如错误标注的比例,因而像Hamming损失这类用于序列化结构的损失函数则会具有更好的适应性.

3.1 基于大间隔的参数估计

对于单标号的二元分类问题,支持向量机提供了一种有效的基于最大间隔(margin)思想学习决策边界的方法,而使其具有良好的泛化性.但由于传统支持向量机的大间隔训练方法仅适用于单标号的二元分类问题,因而若使大间隔思想能够应用于组块分析模型的训练,则在第 2 节所阐述的序列化特征表示基础上,需要对基本的大间隔思想进行两步扩展:首先需要将其从基本的单标号二元分类情形扩展到一种自然的单标号多元分类情形,再进一步通过将每个序列化结构看成是一个独立的类别,从而将单标号多元分类扩展到多标号多元分类的情形,最终生成对应组块分析问题的优化目标函数和约束集,并应用割平面算法进行近似求解[10].

对于单标号的多元分类问题,Crammer 等人则在二元分类框架的基础上进行了扩展[14],通过在下列约束条件下最大化间隔γ:

$$\begin{aligned}&\max\ \gamma\\&\text{s.t.}\quad \|w\|\le 1;\ w\cdot(\Phi(x,\hat{y})-\Phi(x,y))\ge\gamma,\ \forall x\in S,\ \forall y\ne\hat{y}\end{aligned}\tag{6}$$

其中,$\hat{y}$ 表示正确的赋值.而在组块分析的情形下,由于对一个句子中的多个词需要同时进行标注,y 不再是单个的标号,因而需要对间隔的概念进行相应的扩展.设训练样本集为

$$D=\{(x_1,y_1),\ldots,(x_l,y_l)\}\in(X\times Y)^l,$$

定义训练样本 i 关于参数 w 的间隔为标注 y 和正确标注 y_i 的差:

$$w\cdot\Phi(x_i,y_i)-w\cdot\Phi(x_i,y)\tag{7}$$

公式(7)所表示的间隔值大小可以看成是在使用目标函数进行预测时,拒绝错误状态序列的信任度的一种量化表示,因此希望间隔值的大小能够随着标注 y 的错误程度而变化,y 的错误越严重,即损失函数 $L(x_i,y_i,y)$ 越大,则间隔值也越大.可以将这种设想表示成如下的一个优化问题:

$$\begin{aligned}&\max\ \gamma\\&\text{s.t.}\quad w\cdot\Phi(x_i,y_i)-w\cdot\Phi(x_i,y)\ge\gamma L_{i,y},\ \forall i,\ y\in Y\setminus y_i;\\&\qquad \|w\|^2\le 1,\end{aligned}\tag{8}$$

其中,$L_{i,y}=L(x_i,y_i,y)$.不过,对于组块分析问题的损失函数不宜再采用简单的0-1损失函数,而需要考虑多个标记的损失,如可定义损失函数为被错误预测的标记的比例.经过一个标准转换,可将公式(8)表示的极大化间隔问题转换为极小化权值问题,得到如公式(9)所示的二次规划问题:

$$\begin{aligned}&\min\ \frac{1}{2}\|w\|^2+C\sum_i\xi_i\\&\text{s.t.}\quad w\cdot\Phi(x_i,y_i)-w\cdot\Phi(x_i,y)\ge L_{i,y}-\xi_i,\ \forall i,y\in Y\setminus y_i;\end{aligned}\tag{9}$$

公式(9)中对第 i 个训练样本引进一个非负的松弛变量ξ_i,常量 C 则作为平衡间隔大小和样本错划程度的一个权重.由于对每个可能的序列化标注 y 均存在一个对应的约束,因而导致产生一个包含指数级数量约束的二次规划问题.但是考虑到大间隔问题的特殊结构,实际上在指数级数量的约束集合中只有非常少的一部分约束需要进行明确的检查,从而依据这个特性可以将指数级的约束数量减少为多项式数量级.割平面方法通过构造初始优化问题的一个嵌套序列的不断紧密的松弛[10],从而保证产生一个足够精确的近似解,因而是求解这类优化问题的有效方法,具体算法描述如下:

算法 1. 求解组块标注任务的近似优化算法.

Input: $(x_1,y_1),\ldots,(x_n,y_n),C,\varepsilon$

$K=\varnothing,w=0,\xi=0$

Output: w.

repeat

$K_{init}=K$

for i=1,…,n do

$\hat{y}=\arg\max_{y\in Y\setminus y_i}[L_{i,y}+w\cdot(\Phi(x,y)-\Phi(x,y_i))]$ with dynamic programming

if $w\cdot(\Phi(x,y_i)-\Phi(x,y))<L_{i,y}-\xi_i-\varepsilon$

$K=K\cup\{w\cdot((\Phi(x,y_i)-\Phi(x,y))\geq L_{i,y}-\xi_i-\varepsilon\}$

$(w,\xi)=\arg\min_{w,\xi}\frac{1}{2}\|w\|^2+C\sum_{i=1}^{n}\xi_i$ s.t. K

end if

end for

until $K=K_{init}$

算法开始时,约束集 K 为空.之后在每次迭代中,从可能的指数级数量的约束集合中选择出最违背的约束加入到约束集 K 中,不断地执行迭代,直到达到精确度值ε>0.可以证明[10],在该算法收敛之前,仅仅多项式数量的约束被加入到了约束集 K 中.

4 损失函数的设计

由于本文采用的大间隔的组块分析方法对损失函数不存在结构上可分解性的限制,因而可以应用不同的损失函数.对于组块分析这样的序列化标注问题,可以采用的损失函数主要有 Hamming 损失与 $F1$ 损失两种.其中,Hamming 损失函数的形式化定义如公式(10):

$$\ell^{Ham}(y,\hat{y})=\sum_{n=1}^{N}1[y_n\neq\hat{y}_n] \tag{10}$$

从公式(10)定义可以看出,Hamming 损失函数的优点是可以根据 Y 的结构进行分解.即对于任何的排列π,$\ell^{Ham}(y,\hat{y})=\ell^{Ham}(\pi\circ y,\pi\circ\hat{y})$,其中的$\pi$可以看成是序列结构上的一种组合操作.

另一种 $F1$ 损失函数的形式化定义如公式(11):

$$\ell^{F}(y,\hat{y})=1-\frac{2|y\cap\hat{y}|}{|y|+|\hat{y}|} \tag{11}$$

其中,集合的势与交集根据组块的具体情形解释如下:y 的势是指 y 中所确定的组块数量.分子中 y 与 $\hat{y}$ 交集的势则是 y 与 $\hat{y}$ 中相同的组块数量,即被正确识别的组块数量.由于相对于序列化结构的 Hamming 损失函数值,仅将纯粹的 $F1$ 值(0~1)作为损失函数值太小,不利于模型的判别式训练,因而我们引入一种改进的 $F1$ 损失函数:

$$\ell^{F}(y,\hat{y})=\left(1-\frac{2|y\cap\hat{y}|}{|y|+|\hat{y}|}\right)\times y.length \tag{12}$$

其中的 *y.length* 为放大系数,表示句子的长度.这里,我们并没有将放大系数设为一个固定的常数值,而是将其设置为一个动态值,这样使得 $F1$ 损失值能够依据每个句子的实际长度进行相应的调整,动态地增大间隔值,从而能够引入更有效的约束不等式.从公式(12)可以看出,$F1$ 损失函数无论对确定过多的组块还是过少的组块都将进行惩罚.当确定的组块数量过少时,通过分子进行惩罚;而当确定的组块数量过少时,通过分母进行惩罚.

5 实验与结果

我们的实验采用的是与文献[6]相同的训练集和测试集,它们来自于 LDC 的中文树库 CTB4,其中共包含 838 个文件.在实验中,我们使用前 728 个文件(FID 从 chtb001.fid 到 chtb899.fid)作为训练语料,后 110 个文件(FID 从 chtb900.fid 到 chtb1078.fid)作为测试语料,表 2 列出了训练数据与测试数据的统计数据.

测试的结果采取了常用的 3 个评测指标,即准确率(*P*)、召回率(*R*)和综合指标 *F* 值(*F*1)来评测组块识别的结果.其定义如下:

准确率 *P*=正确识别出的组块数/识别出的组块数×100%.

召回率 *R*=正确识别出的组块数/实际组块数×100%.

$$F=\frac{R\times P\times 2}{R+P}.$$

Tabel 2 Information of the CTB4 corpus

表 2 训练数据与测试数据的统计数据

	Training data	Test data
Num of files	728	110
Num of sentences	9 878	5 290
Num of words	238 906	165 862
Num of phrases	141 426	101 449

5.1 特征表示

如果把句子中每一个词的组块标注过程看作是一个事件,则应由当前词及其上下文环境来确定一个事件的特征集合,根据影响当前词组块标注的各种因素,我们主要利用词和词性信息定义特征空间如下:

(1) 词:当前词和前、后各 2 个词的 uni-gram 和 bi-gram;

(2) 词性:当前词的词性和前、后各 2 个词的词性的 uni-gram 和 bi-gram;

(3) 词和词性的组合:当前位置和前后各两个位置的词语与词性的组合.

由于大间隔方法在处理高维特征空间时仍能具有良好的泛化能力,并能够在具有多种特征组合的情况下进行训练.因此在实验中我们并没有如传统方法(像最大熵、条件随机场等)那样进行特征选择,而是直接对依据各个特征模板生成的所有特征集合进行训练.

5.2 不同损失函数之间的比较

基于同样的训练数据与测试数据集,我们分别应用了改进的 *F*1 损失与 Hamming 损失两种不同的损失函数进行优化训练与测试,表 3 给出了两种不同的损失函数的实验结果.从表 3 可以看出,对改进的 *F*1 损失函数进行优化的总体效果优于 Hamming 损失函数,这主要是因为对其中最重要的两种组块类型:名词短语(NP)和动词短语(VP)应用改进的 *F*1 损失函数进行识别的结果均优于 Hamming 损失函数.由于名词短语(NP)和动词短语(VP)相对于其他类型的短语而言,其单个短语可能具有更复杂的结构,因此往往由更多数量的词语构成,如名词短语[NP 污水/NN 处理/NN、/PU 能源/NN、/PU 交通/NN 等/ETC]由多达 7 个词构成,而 Hamming 损失函数仅统计整个句子中的标注错误程度,未能充分考虑标注错误与组块标注 *F*1 值之间的关联度.

Tabel 3 Performance comparison of different loss functions (ε=0.5, C=0.1)

表 3 不同损失函数的实验结果比较(ε=0.5,C=0.1)

	Improved *F*1 loss			Hamming loss		
	Precision	Recall	*F*1	Precision	Recall	*F*1
ADJP	83.95	86.97	85.43	82.75	86.70	84.68
ADVP	78.74	90.54	84.23	78.28	91.46	84.35
CLP	29.41	4.81	8.26	29.41	4.95	8.47
DNP	99.93	99.40	99.66	99.93	99.40	99.66
DP	99.88	99.52	99.70	99.88	99.52	99.70
DVP	99.22	98.45	98.83	99.22	99.61	99.41
LCP	99.96	99.74	99.85	99.93	99.74	99.84
LST	66.25	84.13	74.13	83.75	85.90	84.81
NP	90.01	91.33	90.66	89.87	91.21	90.53
PP	99.80	99.54	99.67	99.80	99.54	99.67
QP	96.91	97.12	97.01	96.81	97.25	97.03
VP	89.61	89.88	89.75	89.45	89.84	89.65
ALL	91.00	92.23	91.61	90.86	92.22	91.53

5.3 不同模型之间的比较

SVMs 和 CRFs 是目前用于解决汉语组块分析问题取得最好效果的两种模型与方法[6].为了验证本文中所提出的组块分析方法的实际效果,我们将其与 SVMs 和 CRFs 这两种方法进行实验比较,实验中我们和文献[6]采用了同样的基于单词和词性的特征表示.表 4 给出了 SVMs,CRFs 以及本文所提出的大间隔组块分析方法的组块分析结果.其中,前两种方法的结果来自于文献[6].表 4 中所给出的实验结果表明,基于大间隔方法对各种类型短语识别的 $F1$ 值均大于等于 SVMs 和 CRFs 的实验结果,且所有类型短语识别的整体结果也高于 SVMs 和 CRFs 模型.

从表 4 中的实验数据可以看出,本文所提出的大间隔方法的组块分析结果明显高于 CRFs 方法,$F1$ 值高出了 0.87%,使错误率减少了 9.4%,从而说明大间隔方法明显优于基于 CRFs 的组块分析方法,也验证了大间隔的训练算法比极大似然法具有更好的泛化性;与基于 SVMs 的组块分析方法相比,大间隔方法的优势则主要在于其可以充分利用句子的序列化结构特征.在同样的特征表示的前提下,大间隔方法的组块分析实验结果的 $F1$ 值比 SVMs 方法虽然仅高出了 0.15%,这主要是由于我们在当前的实验中对结构特征的利用采用的是一阶马尔可夫独立性假设,因而仅利用了非常有限的句子结构特征信息,但这样的实验比较结果也初步验证了在大间隔方法中能够利用句子序列中的结构特征所取得的效果.下一步的工作我们将引入二阶马尔可夫独立性假设,以更充分地利用句子中的序列化结构特征,组块分析效果也一定能够得到进一步提高.

另外从系统复杂性的角度来看,本文采用的大间隔组块分析方法也优于 SVMs 方法.因为 SVMs 方法在解决汉语组块分析问题时需要建立多个独立的 SVM 模型,而本文的大间隔方法与 CRFs 方法一样,只需建立单个的模型.

Tabel 4 Performance comparison of different approaches ($F1$)

表 4 不同方法的实验结果比较($F1$)

	SVMs	CRFs	Large margin
ADJP	84.45	84.55	85.43
ADVP	83.12	82.74	84.23
CLP	5.26	0.00	8.26
DNP	99.65	99.64	99.66
DP	99.70	99.40	99.70
DVP	96.77	92.89	98.83
LCP	99.85	99.85	99.85
LST	68.75	68.25	74.13
NP	90.54	89.79	90.66
PP	99.67	99.66	99.67
QP	96.73	96.53	97.01
VP	89.74	88.50	89.75
ALL	91.46	90.74	91.61

6 结束语

汉语中的组块分析是处于语句的分词标注和完整句法分析之间的一个步骤.组块分析问题的有效解决可以降低完整语法分析的复杂度,同时对信息抽取、指代消解等自然语言应用问题的解决具有重要意义.本文采用一种大间隔的方法实现汉语组块划分和识别的任务,它既具有 CRFs 模型的序列化建模能力,同时又具有 SVM 模型所具有的良好的泛化性的特点,而且不要求损失函数具有结构上的可分解性.本文针对中文组块分析问题设计了一种改进的 $F1$ 损失函数,使得 $F1$ 损失值能够依据每个句子的实际长度进行相应的调整,从而能够引入更有效的约束不等式.实验结果也表明,本文所提出的大间隔的汉语组块分析方法识别效果优于目前的其他组块分析方法.下一步我们将在模型中加入其他类型的上下文信息,如搭配信息、语义信息和共现信息等,以进一步提高汉语组块分析的效果.

References:

[1] Li HQ, Huang CN, Gao JF, Fan XZ. Chinese chunking with another type of spec. In: Proc. of the 3rd SIGHAN Workshop on Chinese Language Processing. 2004. 41−48. http://aclweb.org/anthology-new/W/W04/W04-1107.pdf

[2] Zhao J. A transform-based model for Chinese base noun phrase recognition. Journal of Chinese Information, 1999,13(2):1−7 (in Chinese with English abstract).

[3] Li SJ, Liu Q, Yang ZF. Chunking parsing with maximum entropy principle. Chinese Journal of Computers, 2003,26(12): 1722−1727 (in Chinese with English abstract).

[4] Li H, Zhu JB, Yao TS. SVM based Chinese text chunking. Journal of Chinese Information, 2004,18(2):1−7 (in Chinese with English abstract).

[5] Tan YM, Yao TS, Chen Q, Zhu JB. Applying conditional random fields to Chinese shallow parsing. In: Proc. of the CICLing. Berlin, Heidelberg: Springer-Verlag, 2005. 167−176.

[6] Chen WL, Zhang YJ, Hitoshi I. An empirical study of Chinese chunking. In: Proc. of the COLING/ACL 2006 Main Conf. Poster Sessions. Morristown: Association for Computational Linguistics, 2006. 97−104.

[7] Kudo T, Matsumoto Y. Chunking with support vector machines. In: Proc. of the NAACL. Morristown: Association for Computational Linguistics, 2001. 1−8.

[8] Sha F, Pereira F. Shallow parsing with conditional random fields. In: Proc. of the HLT-NAACL. Morristown: Association for Computational Linguistics, 2003. 134−141.

[9] Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proc. of the Int'l Conf. on Machine Learning (ICML). San Francisco: Morgan Kaufmann Publishers, 2001. 282−289.

[10] Tsochantaridis I, Joachims T, Hofmann T, Altun Y. Large margin methods for structured and interdependent output variables. Journal of Machine Learning Research, 2005,6(9):1453−1484.

[11] Abney S. Part of speech tagging and partial parsing. In: Church K, Young S, Bloothooft G, eds. Proc. of the Corpus-Based Methods in Language and Speech, An ELSNET Volume. Dordrecht : Kluwer Academic Publishers, 1996. 119−136.

[12] Xue NW, Xia F, Huang SZ, Kroch A. The bracketing guidelines for the penn Chinese treebank. Technical Report, University of Pennsylvania, 2000. http://www.cis.upenn.edu/~chinese/parseguide.3rd.ch.pdf

[13] Altun Y, Tsochantaridis I, Hofmann T. Hidden Markov support vector machines. In: Proc. of the Int'l Conf. on Machine Learning (ICML). Menlo Park: AAAI Press, 2003. 3−10.

[14] Crammer K, Singer Y. On the algorithmic implementation of multiclass kernel based vector machines. Journal of Machine Learning Research, 2001,2(5):265−292.

附中文参考文献:

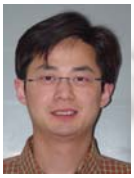
[2] 赵军.基于转换的汉语基本名词短语识别模型.中文信息学报,1999,13(2):1−7.

[3] 李素建,刘群,杨志峰.基于最大熵模型的组块分析.计算机学报,2003,26(12):1722−1727.

[4] 李珩,朱靖波,姚天顺.基于 SVM 的中文组块分析.中文信息学报,2004,18(2):1−7.

周俊生(1972－),男,安徽枞阳人,博士,讲师,主要研究领域为自然语言处理,信息抽取,机器学习.

陈家骏(1963－),男,博士,教授,博士生导师,CCF 会员,主要研究领域为自然语言处理,机器翻译,软件工程.



戴新宇(1979－),男,博士,讲师,CCF 会员,主要研究领域为机器翻译,信息检索.

曲维光(1964－),男,博士,副教授,主要研究领域为计算语言学,人工智能.