基于最大熵模型的组块分析

李素建 刘 群 杨志峰
(北京大学计算语言学研究所 北京 100871)

摘 要 采用最大熵模型实现中文组块分析的任务.首先明确了中文组块的定义,并且列出了模型中所有的组块类型和组块标注符号.组块划分和识别的过程可以转化为对于每一个词语赋予一个组块标注符号的过程,我们可以把它作为一个分类问题根据最大熵模型来解决.最大熵模型的关键是如何选取有效的特征,文中给出了相关的特征选择过程和算法.最后给出了系统实现和实验结果.

关键词 组块分析;语法分析;最大熵原理;浅层分析
中图法分类号 TP391

Chunk Parsing with Maximum Entropy Principle

LI Su-Jian LIU Qun YANG Zhi-Feng
(*Institute of Computational Linguistics, Peking University, Beijing* 100871)

Abstract This paper proposes to use Maximum Entropy (ME) model to conduct Chinese chunk parsing. First we define Chinese chunks and list all chunk categories and tags used in the model. Thus the process of chunking can be regarded as a classification problem which trains from the corpus with chunk tags and POS tags. The focus of ME model is how to select useful features. Then, the procedure and algorithms of feature selection is introduced. At last we test the model, and experimental results are given.

Keywords chunk parsing; syntactic parsing; maximum entropy principle; partial parsing

1 引 言

句法分析是自然语言处理中的重点和难点,虽然经过几十年的研究和发展,仍是现在的一个瓶颈问题.因此目前通常采用"分而治之"的原则,降低完全句法分析的难度,进行部分的句法分析,也称作组块分析.它是和完全句法分析相对的.完全句法分析着眼于充分分析整个句子的语法特点,最大限度地揭示句子所反映的主题内容.而组块分析只限于把句子解析成较小的单元,而不揭示这些单元之间的句法关系.组块分析目前逐步受到重视,国际会议 Conll-2000 把它作为共享任务提出[1].Steven[2]总结了英语中的各种基本组块(名词组块、形容词组块、动词组块等).而目前对于中文组块的研究主要侧重于最长名词短语、基本名词短语等的研究[3,4],系统的汉语组块及其划分的研究还很少.

在自然语言处理中有不少统计建模的例子,由于最大熵模型的简洁、通用和易于移植,目前在对英语的处理中经常采用该技术[5~8].汉语中词性标注和短语边界识别多使用 HMM 的统计模型[3,9],还未见有论文或资料谈到使用最大熵的方法.本文结合汉语语言的特点,实现了一个基于最大熵模型的组块标注器获得汉语中各种类型的组块,并采用宾

收稿日期:2002-04-08;修改稿收到日期:2003-06-04.本课题得到国家"九七三"重点基础研究发展规划项目(G1998030504-01,G1998030507-4)资助.**李素建**,女,1975 年生,博士,主要研究方向为自然语言处理、知识挖掘、机器翻译.E-mail:lisujian@pku.edu.cn.
刘 群,男,1966 年生,博士,副研究员,主要研究领域为机器翻译、自然语言处理、人工智能.**杨志峰**,男,1975 年生,博士,主要研究方向为信息检索、自然语言处理、知识挖掘.

州大学的真实语料进行训练和测试，组块的召回率达到 90.6%，精确率达到 91.9%.

本文中，第 2 节是对中文组块的定义，并列出各种组块类型和组块标注符号；第 3 节简要介绍组块分析采用的最大熵模型；第 4 节阐述最大熵模型建立特征集合的过程；第 5 节描述如何进行模型测试，并给出实验结果；最后对全文进行总结.

2　组块分析的任务

2.1　组块及其类型的定义

首先明确本文中组块的定义以及组块分析的任务及目标. 我们借用了 Abney[10] 对英语组块的定义，为汉语中的组块定义如下.

定义 1.　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，并围绕该中心词展开，以中心词作为组块的开始或结束. 任何一种类型的组块内部不包含其它类型的组块.

这里的定义与 Abney 的英文组块定义有两点不同：(1) 本文中组块是构成语句的最小句法功能单位，不能包含其他的组块，所有组块都位于同一个层次上，各种组块类型是平等的；而在 Abney 的定义中，组块是分层次的，高层次的组块由低层次的组块构成；(2) Abney 的组块定义中，组块中心词只作为组块的结束，中心词后的从属成分另起一个组块，而本文的定义中，组块的中心词也可以作为组块的开始. 例如组块"拿 到"的中心词是"拿"，"到"是处于中心词后的从属部分.

根据以上对组块的定义，结合汉语的特点，我们定义了 12 种组块类型，如表 1 所示.

表 1　组块类型

组块类型	组块描述	组块类型	组块描述
ADJC	形容词组块	NC	名词组块
ADVC	副词组块	PC	介词组块
DNC	"的"字组块	QC	量词组块
DVC	"地"字组块	VCC①	动词组块
LCC	方位组块	NOC	非组块
LST	列举标示组块	O	分割组块的标点符号

2.2　组块标注

组块分析的任务包括组块的划分和识别，在实际系统中通过组块标注来实现，对每一个词语赋予一个组块标注符号，这样组块分析就可以被看作分类问题来解决. Church[11] 和 Ramshaw[12] 采用左右括号的标注方法达到组块分析的目的. Conll-2000 会议上则由组块类型和边界标志共同组成组块标注的符号[1].

采用左右括号进行标注，在获得组块后还存在着组块识别的问题，并且只有边界作为组块标注符号，则分类类型太少，上下文特征过于分散，分类效果不好；同时根据已经标注的语料进行统计，平均每个汉语组块包括 2.76 个词，则有 72% 的词位于组块边界位置，因此没有必要把组块边界标志用来组成组块标注符号. 所以本文中主要把组块类型作为组块标注符号.

前面提到的 12 种组块类型，前 10 种是具体的组块类型，不能归类到这十种组块类型的则赋值为"非组块(NOC)"，这样可以保证句中任何一个词都可以归入某种组块类型中去. 标点符号，单独赋予一个组块类型"O". 当一个词属于这 12 种的任一种时，则赋予与组块类型一致的组块标注符号. 由于两个名词组块或动词组块经常会出现相邻的情况，为了分离开两个组块，对于第二个名词组块开头的词赋予一个"NC$ "的组块标注符号，第二个动词组块开头的词赋予一个"VC$ "的组块标注符号. 这样共定义了 14 种组块标注符号，语句中任何一个词都可以被赋予一个适当的组块标注符号，且属于 12 种组块类型的某一种.

3　最大熵模型框架

最大熵模型是一个比较成熟的统计模型，适合于分类问题的解决. 最大熵框架的计算模型不依赖语言模型，独立于特定的任务. 这里我们再简单回顾一下最大熵框架的原理. 进行组块分析时，我们选取的训练数据以每一个词作为一个事件. 假设有一个样本集合为 $\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$，每一个 $x_i(1\leq i\leq N)$ 表示一个将被组块标注的词的上下文，$y_i(1\leq i\leq N)$ 表示该词被组块标注的结果. 利用最大熵框架模型得出在特征限制下最优的概率分布，即概率值为 $p(y|x)$. 根据最大熵原理，概率值 $p(y|x)$ 的取值符合下面的指数模型：

$$p(y\mid x)=Z_\lambda(x)\exp\Big(\sum_i\lambda_i f_i(x,y)\Big)$$
$$Z_\lambda(x)=\frac{1}{\sum_y\exp\Big(\sum_i\lambda_i f_i(x,y)\Big)} \tag{1}$$

① 由于存在词性标注 VC 表示系动词，故采用 VCC 表示动词组块.

这里 f_i 即我们上面所说的特征，它是一个二值函数，描述某一个特定的事实. λ_i 指示了特征 f_i 对于模型的重要程度. $Z_\lambda(x)$ 在 x 一定的情况下为一范化常数. 公式(1)使模型由求概率值转化为求参数值 λ_i，一般的估计方法是 Darroch 和 Ratcliff[13]的通用迭代算法(Generalized Iterative Scaling，GIS)，用来得到具有最大熵分布的所有参数值 λ_i. Pietra 等[14]则描述了一个改进的迭代算法. 具体算法可参考相应文献.

4 最大熵模型的特征表示

最大熵模型的关键在于如何针对特定的任务为模型选取特征集合. 采用简单的特征表示复杂的语言现象，承认已有的可观察到的事实，不做任何独立性假设. 这些观察到的事实表示为最大熵模型的特征集合.

由于是对每一个词进行组块标注，每一个词的组块标注过程都被看作是一个事件，因此由当前词及它的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，定义特征空间为：

(1)词性信息. 当前词及其前后各两个词的词性；

(2)词. 当前词前后对当前词组块标注造成影响的一些具有特定用法的词语；

(3)组块标注信息. 只考虑当前词前面一个词的组块标注信息；

(4) 词音节数. 考虑当前词及其前后各一个词的音节数.

根据这个特征空间，我们定义了模型中的模板，如表 2 所示，在这个表中每个模板只考虑一种因素，我们称之为原子模板. 原子模板也可以看作是对于当前上下文的一个特征函数. 原子模板 3 比较特殊，表示当前组块标注的模板，它的取值是针对当前上下文环境要输出的函数值，是最终要求的结果.

表 2 原子特征模板

序号	原子模板(函数)	模板意义
1	*CurPOSTag*	当前词性标注
2	*CurWord*	当前词
3	*DEFAULT*	当前词和前一个词的组块标注
4	*ChunkTag*−1	
5	*POSTag*−1	前后词的词性标注
6	*POSTag*−2	
7	*POSTag*+1	
8	*POSTag*+2	
9	*Word*−1	前后特定的词
10	*Word*−2	
11	*Word*+1	
12	*Word*+2	
13	*CurRhythm*	当前词的音节数目
14	*Rhythm*−1	前后词的音节数目
15	*Rhythm*+1	

当特征函数取特定值时，则该模板被实例化，得到具体的特征. 对于模板的取值，词性标注采纳宾州大学中文树库中的词性标注集合. 对于特殊的词，我们首先拟定了一个词表，主要是列出一些虚词，如“的”、“地”、“得”、“了”、“和”、“与”、“在”等，此外一些具有特殊用法的词，如“把”、“将”、“被”、“为”等. 组块的标注结果则为前面提到的 14 种可能组块标注符号. 词音节数的取值为三种情况：单音节、双音节和多音节. 由原子模板实例化后，并结合当前组块标注的值(即模板 *DEFAULT* 的取值). 当模板的取值确定后就可以产生一个特征，这个特征称作为原子特征.

下面我们对原子特征的格式进行一下定义，每一个原子特征由三部分构成：

(1)第一部分是下划线“_”前的部分，为上下文的特征函数，如 $word+1$，$POSTag-2$ 等，表示特征空间要考虑的因素；

(2)第二部分是下划线和等号中间的部分，例如“的”、“NT”，表示特征函数的取值；

(3)第三部分为输出值，即表示当前词进行组块标注的结果.

在得到一个特征后，该特征可以表示为二值特征函数的形式.

例如，由原子模板 1 可以得到一个特征 *CurPOSTag_NN=NC*，表示为二值特征函数：

$$f_j(x,y)=\begin{cases}1, & CurPOSTag(x)=NN \text{ 且 } y=NC\\ 0, & \text{否则}\end{cases} \tag{2}$$

$CurPOSTag(x)$ 可以看作模板 *CurPOSTag* 在上下文中的函数表示，y 实质上是模板 *DEFAULT* 的取值，表示当前输出的结果.

表 3 复合特征模板

序号	复合模板
1	*CurPOSTag*，*POSTag*+1
2	*POSTag*−1，*CurPosTag*
3	*POSTag*−1，*CurPOSTag*，*POSTag*+1
4	*POSTag*+1，*ChunkTag*−1，*CurPOSTag*
5	*ChunkTag*−1，*CurPOSTag*，*POSTag*−1

（续　表）

序号	复合模板
6	*PosTag*−1,*PosTag*+1
7	*CurPosTag*,*ChunkTag*−1
8	*CurWord*,*PosTag*+2
9	*Word*+1,*CurPosTag*
10	*Word*−1,*CurPosTag*
11	*CurPosTag*,*Word*+1,*PosTag*−1
12	*PosTag*−1,*PosTag*+1,*PosTag*+2
13	*CurPosTag*,*PosTag*+1,*PosTag*+2
14	*CurPosTag*,*PosTag*+2
15	*CurPosTag*,*PosTag*−1,*Word*−2
16	*CurPosTag*,*PosTag*−1,*Word*+1
17	*CurPosTag*,*PosTag*−1,*CurWord*
18	*PosTag*−1,*Rhythm*−1_1,*CurRhythm*−1
19	*Rhythm*−1,*ChunkTag*−1,*CurRhythm*
20	*Rhythm*−1,*POSTag*−1,*CurPOSTag*

由于在上下文中,仅仅用原子特征不足以表示上下文中的一些现象.通过对表 2 中各种原子模板进行组合,构成一些复合特征模板来表示更复杂的上下文环境,如表 3 所示.原子特征模板和各种复合特征模板共同构成了模型的所有特征模板,共有 35 种模板类型.同样,对于复合特征模板,也是首先对各个原子模板通过实例化,对模板函数取值后,可能会输出某种组块标注,从而产生一个特征,为复合特征.复合特征表示为二值特征函数的形式与原子特征相似,只是在取值时需要满足的条件变多.

例如:复合特征 *CurPOSTag*_*CD*,*POSTag*+1_*M*=*QC* 表示为

$$f_j(x,y)=\begin{cases}1, & y=QC \text{ 且 } CurPOSTag(x)=CD,\\ & POSTag+1(x)=M\\ 0, & \text{否则}\end{cases} \tag{3}$$

另外,最大熵模型的一个优势在于选取特征灵活,可以方便地把一些跨距离的特征加入到模型中.前面通过模板自动获取特征,对于模板的设定都选在当前词左右不超过两个词的距离.然而对当前词发生影响的因素可能不在这个距离范围之内,所以对于跨距离的语言现象,我们规定了 4 个特征函数,如表 4 所示.

表 4　跨距离特征函数

序号	特征函数	含义
1	*PrevPOS*	位于当前词所在句中,并在其前任一位置的词性标注
2	*PrevWord*	位于当前词所在句中,并在其前任一位置的特定词
3	*NextPOS*	位于当前词所在句中,并在其后任一位置的词性标注
4	*NextWord*	位于当前词所在句中,并在其后任一位置的特定词

通常我们结合跨距离特征函数和其他原子模板,通过观察语料来手工编写一些典型的特征,并纳入到规则集合中,我们把这些特征称为混合特征.混合特征也可以表示为二值特征函数的形式.

例如:*CurPOSTag*_*LC*,*PrevPOS*_*P*=*PC* 表示为

$$f_j(x,y)=\begin{cases}1, & y=PC \text{ 且 } CurPOSTag(x)=LC\\ & PrevPOS(x)=P\\ 0, & \text{否则}\end{cases} \tag{4}$$

在实际文本中,例如“在/*P*…中/*LC*”是一个介词组块,对于词“中”,组块标注为 PC,则该混合特征的二值函数取值为 1.

根据特征模板可以自动从语料中得到一个数量庞大的特征集合,然而并非所有特征都适合引入到最大熵模型中去.Pietra[14] 对自然语言处理中随机域的特征选取进行了描述,根据特征的信息增益作为是否引入的衡量标准.对于要处理的问题,特征所含的信息量越大,该特征就越适合引入到模型中.通过模板得到的特征构成候补特征集合,然后从中选取对模型最为有用的特征.由于篇幅有限,对于特征引入算法不再赘述.

5　系统实现和实验结果

根据公式(1),在完成最大熵模型的参数估计后,可以得到模型的概率分布.由模型的概率分布以及词的上下文环境可以得到词被赋予某种组块标注符号的概率值.

如图 1 所示,给定词序列 $W=w_1w_2\cdots w_n$ 为一个新语句,令 $F=F_1F_2\cdots F_n$,$F_i(1\leq i\leq n)$ 为第 i 个词的特征向量表示.因此组块标注的问题,可以视为在给定词序列及上下文特征的情况下,搜寻组块标注序列 $C=c_1,c_2,\cdots,c_n$,使得 $P(C|W,F)$ 最大.由于每个词的特征向量表示中已经包含了全部的上下文特征,那么各个词进行组块标注的事件是相互独立的.显然词信息也被包含在特征向量中,因此只要求得 $P(C|F)$ 最大即可.即

$$C=\arg\max_C P(C\mid F)=\arg\max_C\prod_{i=1}^{n}P(c_i\mid F_i) \tag{5}$$

输入:	w_1/F_1	w_2/F_2	w_3/F_3	…	w_n/F_n
输出:	c_1	c_2	c_3	…	c_n

图 1　模型输入和输出

在组块标注时，当前词的组块标注要受上文中词组块标注结果的影响，而任一个词都有被标注为14种组块标注符号的可能性，如果采用穷举法得到所有可能的概率值，则数据量为 14^n，因此我们采用动态优化方法得到一个最优概率值，采用如下公式：

$$\begin{cases} c_1 = \arg\max_{1\le i\le 14} P(c_{1i} \mid F_1) \\ c_2 = \arg\max_{1\le i\le 14} P(c_{2i} \mid F_2(c_1)) \\ \cdots \\ c_n = \arg\max_{1\le i\le 14} P(c_{ni} \mid F_n(c_{n-1}, c_{n-2}, \cdots, c_1)) \end{cases} \tag{6}$$

利用公式(6)，对一个语句从左向右进行组块标注. 根据当前词上文的组块标注信息，动态获取当前词的特征向量. 计算当前具有最大概率值的组块标注符号，从左到右按照顺序依次完成每一个词的组块标注过程.

进行组块划分和识别的语料来自于宾州中文树库，共有 4185 个语句，大约由 10 万个汉词组成. 每个词含有词性标注信息、组块标注信息. 同时把这个数据集合分为两部分：前面 8 万个词作为训练集合，其余 2 万个词作为测试集合来评测组块分析的效果. 评测标准包括组块标注的准确率、各种类型组块的召回率和精确率、召回率和精确率的综合评价指标为 $F_{\beta=1}$. 几种评价函数定义具体如下：

A = 具有正确组块标注符号的词语数目，

B = 全部词语数目，

C = 标注正确的 XC 组块的数目，

D = 应当标注为 XC 组块的数目，

E = 标注为 XC 的组块数目，

$$\text{标注准确率} = \frac{A}{B},$$

$$XC \text{ 组块召回率} = \frac{C}{D},$$

$$XC \text{ 组块精确率} = \frac{C}{E},$$

XC 组块

$$F_{\beta=1} = \frac{(\beta^2+1)\times XC\text{ 召回率}\times XC\text{ 精确率}}{\beta^2\times XC\text{ 召回率} + XC\text{ 精确率}} = \frac{2\times XC\text{ 召回率}\times XC\text{ 精确率}}{XC\text{ 召回率} + XC\text{ 精确率}} \tag{7}$$

XC 表示某一种组块类型，例如 NC 表示名词组块，VCC 表示动词组块等等. 在这些评价函数中，标注准确率影响着各种组块的召回率和精确率，每个词被组块标注的准确率越高，各个组块的召回率和精确率也就越高. 也就是说，只有每个词标注正确，才更有可能得到正确的组块. 最终特征集合包含 1500 多个特征，标注准确率可以达到 93.34%. 同时表 5 中列出了一些组块类型的标注结果.

表 5 最大熵模型进行组块划分的结果

组块类型	召回率	精确率	$F_{\beta=1}$
NC	0.94	0.82	0.88
VCC	0.93	0.92	0.925
ADJC	0.9	0.87	0.63
ADVC	0.75	1.00	0.86
PC	0.976	0.93	0.945
QC	0.95	1.00	0.97
DNC	1.00	1.00	1.00
DVC	1.00	1.00	1.00
LCC	0.71	0.73	0.70
所有组块	0.906	0.919	0.912

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. 它对于一些未登录词和常用语的识别有很好的效果，可以降低分词标注中的错误. 同时组块分析降低了完整语法分析的复杂度. 本文采用最大熵模型建模实现组块划分和识别的任务，比一般的统计模型能获取到更丰富的不受限文本特征，诸如可以灵活地把一些跨距离的特征加入到模型中去. 这样就结合了统计模型(例 HMM)和规则方法(例 TBL)的优点，和 TBL 方法相似，都选用大量的特征作支持；同时利用统计模型的优势，估计对每个词赋予组块标注符号的概率分布. 目前我们的实验结果比当前的各种系统并没有特别显著的提高，存在着很多原因，如训练语料库规模小和特征集不够大等. 基于当前的状况和任务，该模型还需要进一步地改进和完善.

致谢 中国科学院计算技术研究所白硕研究员对本文的工作给予了悉心指导；中国科学院计算技术研究所软件室的李继峰、张浩、张华平等同学对本文的完成提出了很多有益的建议，在此一并表示感谢.

参 考 文 献

1 Erik F, Tjong Kim Sang, Buchholz S. Introduction to the CoNLL-2000 Shared Task: Chunking. In: Proceedings of CoNLL-2000 and LLL-2000, Lisbon, Portugal, 2000. 127～132

2 Steven A. Parsing by Chunks. In: Berwick, Abney, Tenny eds. Principle-Based Parsing: Kluwer Academic Publishers, 1991. 257～278

3 Zhou Qiang, Sun Mao-Song, Huang Chang-Ning. Automatic identification of Chinese maximal noun phrases. Journal of Software, 2000, 11(2): 195～201(in Chinese)
(周　强,孙茂松,黄昌宁.汉语最长名词短语的自动识别.软件学报,2000,11(2):195～201)

4 Zhao Jun, Huang Chang-Ning. Recognition model of Chinese BaseNP based on transformation. Journal of Chinese Information Processing, 1999, 13(2): 1～7(in Chinese)
(赵　军,黄昌宁.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)

5 Ratnaparkhi A. A maximum entropy model for part-of-speech tagging. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 1996

6 Ratnaparkhi A. A simple introduction to maximum entropy models for natural language processing. Institute for Research in Cognitive Science, University of Pennsylvania: Technical Report 9708, 1997

7 Berger A, Pietra S D, Pietra V D. A maximum entropy approach to natural language processing. Computational Linguistics, 1996,22(1):39～71

8 Skut, Wojciech, Thorsten Brants. A maximum entropy partial parser for unrestricted text. In: Proceedings of the 6th Workshop on Very Large Corpora, Montreal, Canada, 1998. 143～151

9 Zhou Qiang. Chinese POS tagging method with rules and statistics combined. Journal of Chinese Information Processing, 1995, 9(3): 1～10(in Chinese)
(周　强.规则和统计相结合的汉语词类标注方法.中文信息学报,1995,9(3):1～10)

10 Abney S. Part-of-speech tagging and partial parsing. In: Church K, Young S, Bloothooft G eds. Corpus-Based Methods in Language and Speech, An ELSNET volume, Dordrecht: Kluwer Academic Publishers, 1996. 119～136

11 Church K W. A stochastic parts program and noun phrase parser for unrestricted text. In: Proceedings of the 2nd Conference on Applied Natural Language Processing, Texas, USA, 1988. 136～143

12 Ramshaw L A, Marcus M P. Text chunking using transformation-based learning. In: Proceedings of ACL Third Workshop on Very Large Corpora, Cambridge, USA, 1995. 82～94

13 Darroch J N, Ratcliff D. Generalized iterative scaling for log-linear models. Annals of Mathematical Statistics, 1972,43(5): 1470～1480

14 Pietra S D, Pietra V D, Lafferty J. Inducing features of random fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19(4):380～393

LI Su-Jian, born in 1975, Ph. D., Postdoctor. Her research interests include natural language processing, computational linguistics, information extraction.

LIU Qun, born in 1966, associate professor. His research interests include machine translation, natural language processing, artificial intelligence.

YANG Zhi-Feng, born in 1975, Ph. D.. His research interests include information retrieval, natural language processing, knowledge discovery.