

# 基于 CRFs 的级联中文组块识别\*

程勇 孙承杰 刘远超 刘秉权

(哈尔滨工业大学计算机学院智能技术与自然语言处理实验室)

{ycheng, cjsun, lyc, liubq}@insun.hit.edu.cn

**摘要:** 本文介绍了一种基于 CRFs 的级联组块识别系统, 该系统把组块识别任务分成边界识别和组块分类两个阶段来进行。该系统参加了 CIPS-ParsEval-2009 评测的基本块识别测试, 封闭测试的结果为: 组块的“边界+成分”识别的 F1 值为 91.29%, “边界+成分+类型”识别的 F1 值为 90.13%。在对参加评测的结果进行分析后, 将类型的信息引入边界识别的过程中, 最终使得整体识别效果得到了提升, 改进后的模型测试的结果为: 组块的“边界+成分”识别的 F1 值为 92.91%, “边界+成分+类型”识别的 F1 值为 91.93%。

**关键词:** 中文组块分析; 边界识别; 类型识别; 条件随机场

## Cascade identification of Chinese chunks based on CRFs

Yong Cheng, Chengjie Sun, Yuanchao Liu, Bingquan Liu

(ITNLP lab, School of Computer, Harbin Institute of Technology)

{ycheng, cjsun, lyc, liubq}@insun.hit.edu.cn

**Abstract:** This paper describes a Cascaded chunk identification system based on CRFs. In the system, the task of chunk identification is completed through two steps: chunk boundary detection and chunk type classification. The system participated in the chunk track of CIPS-ParsEval-2009 open evaluation. In close test, the system achieve an F1-measure of 91.29% for chunk “boundary+type” identification, and an F1-measure 90.13% for chunk “boundary+type+relation” recognition. Through the analysis of evaluation results, we proposed to introduce type information into the process of boundary identification, and the overall performance has been enhanced. The modified system can achieve an F1-measure of 92.91% for chunk “boundary+type” identification and 91.93% for “boundary+type+relation” recognition.

**Key words:** Chinese chunking; boundary identification; type identification; conational random fields

### 1. 引言

作为一种重要的浅层句法分析技术[1], 组块分析可以通过对完整分析问题的合理任务分解, 大大降低自动分析的处理难度, 在自然语言处理领域的信息抽取、问答系统、文本挖掘等应用系统研究都可以发挥重要作用。英文组块识别在 CoNLL2000 的规范和评测平台上已有了较成熟的研究成果, 而中文组块的研究仍存在不少亟待解决的问题。中文组块识别大致可以分为基于规则的方法和基于统计[2]的方法。对于基于统计的组块识别来说, 组块边界的识别性能常常决定识别的整体性能。现在的常用的组块识别方法主要有两种。一

---

\*国家自然科学基金面上资助项目(60673019, 60673037); 国家 863 计划资助项目(2007AA01Z172)

种是对于组块的边界和类型识别同时进行，此种方法因为融入了较多的特征，对边界的识别率较高，但常常会出现对处于同一组块中的词的类型识别不一致的情况，影响了对组块的类型标记。二是将边界识别和类型识别分开[3]，此种方法克服了方法一的缺点，但在边界识别的过程未能充分利用组块类型和关系所提供的信息，使得边界识别效果较差，从而进一步影响了后面的类型识别，使得识别的整体效果较差。本文采取条件随机域模型（CRFs）将以上两种方法进行折中，首先进行边界和类型识别同时进行，即对每个词进行多分类。但只从其识别的结果中提取边界信息。然后对已进行边界识别的组块进行类型识别。实验表明，采取此种方法在保证了边界识别率的基础上提高了类型识别率，从而使得组块识别的整体性能得到提升。

## 2. 中文基本块的定义和表示

### 2.1 中文基本块的定义

汉语基本块[4]主要描述句子中直接相邻的、以名词、动词、形容词等实义词为中心聚合形成具有特定语义内容的词语序列，其中一般不包括各种功能词，包括连词、叹词、语气词、助词、标点符号等。它们大多由 1-3 个词语组成，通过不同的外部内容表现和内部聚合关系形成特殊的拓扑结构体，成为汉语的字/词进入组块成句过程的基础和出发点。

在此次评测中，对基本块的标记由成分标记和关系标记两部分组成，即表示为<成分标记>-<关系标记>。对于标记的说明如表 1 所示：

表 1 基本块的标记表示

成分标记	标记说明	关系标记	标记说明
np	名词块	ZX	右角依存结构
vp	动词块	PO	述宾关系结构
ap	形容词块	SB	述补关系结构
mp	数量块	LH	并列关系结构
sp	空间块	LN	链式关联结构
tp	时间块	SG	单词语块
dp	副词块		

### 2.2 基本块的表示

基本块的表示即编码问题，常见的有 BIO 模型，BIO 表示模型在 CoNLL2000 上提出，用 B-表示 X 类型组块的起始词，I-X 表示 X 类型的组块中非起始词之外的其他词，O 表示不属于任何组块的词。这种表示方法存在一个问题：识别结果不一致问题。如 B-X 标记的下一个词可能为 I-Y, X 和 Y 为不同类型的组块。而在本次评测中由于标记由成分标记和关系标记两部分组成，组块的类别数比较多。使得这个问题更加严重，影响了整体的识别性

能。

对于以上的问题，有人提出了简单的级联组块识别方法，即将边界识别和类型识别分开，先对每个词进行简单的 2 分类，用“1”表示该词右边界非组块边界，用“0”表示该词右边界同时也是组块边界。再对识别好边界的组块进行类型识别。如下例所示：

中国/nS 传统/a 医学/n 的/uJDE 发生/vN 发展/vN 及/cC 学术/n 特点/n

边界识别：[中国 /nS 1] [传统 /a 1] [医学 /n 0] [的 /uJDE 0] [发生 /vN 1] [发展 /vN 0] [及/cC 0] [学术 /n 1] [特点 /n 0]

类型识别：[np-ZX 中国/nS 传统/a 医学/n ] 的/uJDE [np-LH 发生/vN 发展/vN ] 及/cC [np-ZX 学术/n 特点/n ]

级联组块识别也存在自己的问题，即在第一步进行边界识别的过程，因为未能充分利用语料给出的组块类型信息，而实际中类型信息对于边界的识别可能有促进作用，最终影响了整体的识别效果。

本文在分析了以上两种方法的优点和缺点和基础上，对两种方法的优点进行了结合。即首先进行多分类，对组块中的每个词识别成 B-X, I-X 的形式，但只提取其中关于边界的信息，相当于只进行了边界识别，然后在此基础上在进行类型识别。如下例所示：

中国/nS 传统/a 医学/n 的/uJDE 发生/vN 发展/vN 及/cC 学术/n 特点/n

首次识别：[中国 /nS B-np-ZX] [传统 /a I-np-ZX] [医学 /n I-np-ZX] [的 /uJDE O] [发生 /vN B-np-LH] [发展 /vN I-np-LH] [及 /cC O] [学术 /n B-np-ZX] [特点 /n I-np-ZX]

提取边界信息：[中国 /nS B] [传统 /a I] [医学 /n I] [的 /uJDE O] [发生 /vN B] [发展 /vN I] [及 /cC O] [学术 /n B] [特点 /n I]

类型识别：[np-ZX 中国/nS 传统/a 医学/n ] 的/uJDE [np-LH 发生/vN 发展/vN ] 及/cC [np-ZX 学术/n 特点/n ]

### 3. CRFs 模型及特征选择

#### 3.1 条件随机场

本文运用条件随机场来实现组块识别的 2 个子任务。CRFs 是一种能保证损失函数(loss function) 收敛到全局最优的算法，在用于命名实体识别、词性标注、组块识别[5-7]等实验中，CRF 显示了良好的性能，克服了标记偏置的问题 (label bias)，实验中使用了 CRF++ 工具包(<http://chasen.org/~taku/software/CRF++>)。

#### 3.2 特征选择

为了寻找更好的级联组块识别方法，本文尝试了四种不同的策略。

策略一：将边界识别和类型识别、成分识别结合到一起进行识别。

策略二：将边界识别与类型识别分开，在边界识别中，将每个词简单地识别成 1 或者 0 两类。再对类型和成分进行识别。

策略三：将边界识别与类型识别分开，在边界识别中，将每个词先识别成类似于 B-X(代

表成分标记), 再从中只提取边界的信息, 再对类型和成分进行识别。

策略四: 将边界识别和类型识别分开, 在边界识别中, 将每个词先识别成类似 **B-X-Y** (X 代表成分标记, Y 代表关系标记), 再从中只提取边界的信息, 再对类型和成分进行识别。

四种策略都以组块 [np-ZX 中国 /nS 传统 /a 医学/n] 为范例进行说明。

### 1) 策略一

训练数据的格式如下所示:

中国 nS B-np-ZX

传统 a I-np-ZX

医学 n I-np-ZX

该策略用到的模板如表 2 所示。

表 2 用于边界识别的模板

00:%x[-2,0]	07:%x[-2,1]	14:%x[0,1]/%x[1,1]
01:%x[-1,0]	08:%x[-1,1]	15:%x[1,1]/%x[2,1]
02:%x[0,0]	09:%x[0,1]	16:%x[-2,1]/%x[-1,1]/%x[0,1]
03:%x[1,0]	10:%x[1,1]	17:%x[-1,1]/%x[0,1]/%x[1,1]
04:%x[2,0]	11:%x[2,1]	18:%x[0,1]/%x[1,1]/%x[2,1]
05:%x[-1,0]/%x[0,0]	12:%x[-2,1]/%x[-1,1]	19:%x[-2,0]/%x[-1,0]/%x[0,0]
06:%x[0,0]/%x[1,0]	13:%x[-1,1]/%x[0,1]	20:%x[-1,0]/%x[0,0]/%x[1,0]

上述模板的格式用 %x[row,col] 来表示, 其中 %x 代表当前词汇, row 代表对于当前词汇的相对位置, col 代表列的绝对位置。

### 2) 策略二

对于边界识别的训练数据格式如下:

中国 nS 1

传统 a 1

医学 n 0

对应的特征模板如表 2。

进行第二步类型识别的训练数据格式如下:

中国-传统-医学 nS-a-n nS n 3 中国 传统 医学 np-ZX

从左到右的特征依次是: 组块中的词汇连接 (用“-”连接), 组块中的词性链接 (用“-”链接), 组块中第一个词的词性, 组块中最后一个词的词性, 组块中的第一个词, 组块中的第二个词, 组块中的最后一个词, 该组块对应的类别。类型识别所采用的特征模板如表 3 所示。

表 3 用于组块类型识别的模板

00:%x[-1,2]	07:%x[1,2]
01:%x[-1,3]	08:%x[1,3]
02:%x[0,0]	09:%x[0,2]/%x[0,3]

03:%x[0,1]	10:%x[0,5]/%x[0,6]
04:%x[0,2]	11:%x[0,5]/%x[0,7]
05:%x[0,3]	12:%x[-1,3]/%x[0,2]
06:%x[0,4]	13:%x[0,3]/%x[1,2]

### 3) 策略三

边界识别的训练数据格式如下所示：

中国 nS B-np

传统 a I-np

医学 n I-np

特征模板如表 2 所示。

类型识别的训练数据格式如下所示：

中国-传统-医学 nS-a-n nS n 3 中国 传统 医学 np-ZX

特征模板如表 3 所示。

### 4) 策略四

边界识别的训练数据格式如下所示：

中国 nS B-np-ZX

传统 a I-np-ZX

医学 n I-np-ZX

特征模板如表 2 所示。

类型识别的训练数据如下所示：

中国-传统-医学 nS-a-n nS n 3 中国 传统 医学 np-ZX

特征模板如表 3 所示。

## 4. 实验分析

### 4.1 实验数据

本次实验进行两次比较，第一次比较是参加评测的模型和改进后的模型进行比较，利用的数据是主办方发给的训练数据，将此数据分成 5 份，4 份用于训练，1 份用来测试。第二次比较利用主办方发给的训练数据和测试数据来对上文提到的四种策略进行比较。

### 4.2 实验结果及分析

表 4 是参加测试模型和改进后的模型的 F1 值的比较。从中可以看出，改进后的模型（采用策略 4 的模型）可以显著提高组块识别的性能。改进后的结果接近了所有参赛系统的最好结果。

表 5 是采用不同策略的模型的 F1 值的比较。通过策略 2, 3 和 4 的比较可以得出：在边界识别中引入组块的类型信息会反过来促进对边界的识别准确率，而引入的信息越多，识别的准确率提高的越多。而通过方法 1 与方法 4 的比较可以说明，在边界识别的准确率

相当的情况下通过把类型识别分出去，以级联的方式进行识别，可以提高整体的识别率。

表 4 利用原始训练数据进行比较

模型	Boundary (%)	boundary+type (%)	boundary+type+relation (%)
参加测试模型 (策略 2)	92.58	92.33	91.38
改进后模型 (策略 4)	94.10	93.88	92.69

表 5 利用分发的训练数据和测试数据进行比较

模型	boundary (%)	boundary+type (%)	boundary+type+relation (%)
策略 1	93.09	92.85	91.75
策略 2	91.48	91.29	90.13
策略 3	93.05	92.88	91.78
策略 4	93.09	92.91	91.83

## 5. 总结

本文将中文组块识别分解为 2 步实现，即组块的边界识别和类型识别，但这两步并非独立的，在第一步边界识别的过程中利用了类型识别的信息，提升了边界识别的准确率，从而最终使得整体的识别率得到了提高。实践表明，组块边界的识别性能在很大程度上影响组块识别的整体性能，而充分利用类型信息可以提高对组块边界的识别率。

## 6. 参考文献

- [1]Abney S P. Principle based parsing: computation and psycholinguistics. Dordrecht: Kluwer Academic Publishers, 1991: 257-258.
- [2]李衍, 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.
- [3]秦颖, 王小捷, 钟义信. 级联中文组块识别. 北京邮电大学学报, 2008, 31(1):14-18.
- [4]周强. 汉语基本块描述体系. 中文信息学报, 2007, 21(3): 21-28.
- [5]黄德根, 于静. 分布式策略 CRFs 相结合识别汉语组块. 中文信息学报, 2009, 23(1): 16-23.
- [6] 孙广路, 王晓龙, 关毅. 基于词聚类特征的统计中文组块分析模型. 电子学报. 2008, 36(12): 2450-2453
- [7]Tan Y M, Yao T S, Chen Q. Applying conditional random fields to Chinese shallow parsing. Proceedings of CICLing-2005. Mexico, 2005:167-176.