

基于 CRFs 的汉语基本块识别

杨田, 黄德根, 李丽双

(大连理工大学 计算机科学与工程系, 辽宁 大连 116024)

Huangdg@dlut.edu.cn

摘要: 采用基于 CRFs 的分布式分类器与冲突处理策略来实现对汉语基本块的识别。将不同组块进行分组, 然后基于 CRFs 建立分组的基本块识别模型, 对基本块进行识别; 最后综合考虑各组基本块识别的 F 值及不同类型基本块的数量两种因素确定权值大小, 处理基本块的类型冲突, 从而提高基本块的整体识别结果。与单独基于 CRFs 的方法相比缩短了建模时间。在 CIPS-ParsEval-2009 的语料上进行了实验, 评测结果的 F 值达到 92.11。评测后对实验进行改进, 改变特征模版、增加并列结构处理和减小断句的长度后, F 值达到 93.06%。

关键词: 分布式; 冲突处理; 条件随机域 (CRFs); 中文信息处理; 基本块识别

Chinese Base Chunking Based on CRFs

Yang Tian, Huang De-gen, Li Li-shuang

(Department of Computer Science and Engineering, Dalian University of Technology, Liaoning 116024, China)

Huangdg@dlut.edu.cn

Abstract: This paper proposes a distributed and collision dispose strategy approach to Chinese basic chunking. First the paper gives an overview of Chinese basic chunking, and give a method for Chinese basic chunking based on a distributed strategy. The main idea is, divide the different chunks into different groups, and then build CRFs model respectively. Finally, a method is described to ascertain the chunks' PRI to dispose the collision chunks according to the F-measure values and the number of different chunks. By using test data from CIPS-ParsEval-2009, the method achieves the best results with F-measure of 92.11% in the best case compared with others method. After the evaluation, we do further experiment and get the results with F-measure of 93.06%.

Keywords: Distributed; Collision Dispose; Conditional Random Fields (CRFs); Chinese information processing; basic chunking

1 引言

基本组块识别是自然语言浅层句法分析^[1]的重要任务之一, 被普遍应用于机器翻译、搜索引擎等领域。基本块识别将句子分解成比较小的单元, 其分析结果可以解决大部分的局部歧义问题, 这样有利于将句法分析的难度降低, 为下一步的语法句法分析打下良好的基础。

中文组块识别的方法大致分为基于规则的方法^[2]和基于统计的方法^[3]。而其中基于统计的方法是目下组块识别中较常用的方法。它将组块识别问题转化为组块分类问题。

在大规模的语料处理中, 基于统计的方法与基于规则的方法相比较优点有:

(1) 基于统计的方法完备性强。基于规则的方法, 因制定规则的人的思想不同, 所看问题

的角度不同, 规则就会不同, 造成结果的不完备性。

(2) 基于统计的方法一致性强。基于规则的方法, 当想添加新规则时要与原有规则进行协调, 容易产生冲突。

(3) 基于统计的方法覆盖面广, 可更好的覆盖复杂的语言结构。

统计方法的缺点是需要大规模的语料库支持。随着科技的发展, 大规模语料库的构建成为可能, 这为统计方法奠定了语料基础。

一般的组块识别方法是将所有类型组块整体进行识别或者将组块的类型与边界分开识别, 但都是将组块的类型作为整体进行识别。而本文采用的是分类识别的方法, 通过组块类型分类器将不同的组块分别抽取出来进行处理, 构建 CRFs 识别模型; 然后利用冲突处理技

术将有类型冲突的组块进行处理；最后结合并列结构处理完成基本块标注。将组块识别问题转换成多个不同分类组块识别的问题，通过分类器分组后，每个分组里包含的组块类别减少，这对于提高组块的识别精确度有利，同时也减少了构建识别模型的时间。

2 汉语基本块的定义与表示

2.1 汉语基本块的定义

中文组块与英文 chunk 类似，是一种非重叠、非递归、全覆盖、成分边界不交叉的语言单位^[4]。即任何一种类型的组块内部都不包含其它类型的组块，但可以包含和它本身同一类型的组块，且组块是不互相交叉的。

本文中用到的基本块标注体系：对于每个基本块使用两个标记的组合^[5]，句法标记和关系标记。两种标记对基本块的外部句法表现和内部词汇进行完整描述，使之能方便地与句子中的其他成分相结合形成更大的句法单位。表 1 列出本文目前所使用的主要句法标记和关系标记。此句法标记和关系标记完全采用中文信息学会句法评测 CIPS-ParsEval-2009 中所使用的标准。

表 1 基本块句法标记和关系标记描述集合

句法标记	描述	关系标记	描述
np	名词块	ZX	右角依存
vp	动词块	LN	链式关联
ap	形容词块	LH	并列关系
mp	数量块	PO	述宾关系
sp	空间块	SB	述补关系
tp	时间块	AD	附加关系
dp	副词块	SG	单词语块
mbar	数词块	CD	重叠关系

2.2 中文基本块的表示

本文采用 BIO 标注集合来对组块进行标注。用 B-X 表示 X 类型组块的起始词，用 I-X 表示 X 类型组块中除起始词以外的其它词，O 表示不属于任何组块的其它词。

输入文本为经过正确词语切分和词性标注的汉语句子。例如：

执法/vN 部门/n 是/vC 反/v 腐败/a 斗争/vN 、/wD 搞好/v 廉政/vN 建设/vN 的/uJDE 重点/n 部门/n 之一/rN 。/wE

输出为基本块自动标注结果，例如：

[np-ZX 执法/vN 部门/n] [vp-SG 是/vC] [np-ZX 反/v 腐败/a 斗争/vN] 、/wD [vp-SG 搞好/v] [np-ZX 廉政/vN 建设/vN] 的/uJDE [np-ZX 重点/n 部门/n] [np-SG 之一/rN] 。/wE

3 基于 CRFs 的基本块识别

基本块的标注问题可以转化为序列标注问题^[6]。而基于 CRFs 的机器学习模型可以添加特征向量，它不仅解决了 HMM 的独立性假设问题，同时也解决了 ME 的标记偏置问题，并能充分利用上下文信息。

基于 CRFs 的基本块识别是在给定训练序列的前提下，定义所有类别标记的一个联合概率分布。CRFs 能够依据类别分布属性对数据进行建模。是一种很好的序列标注器^[7-8]。

本文基本块标注系统的工作流程图如图 1 所示。

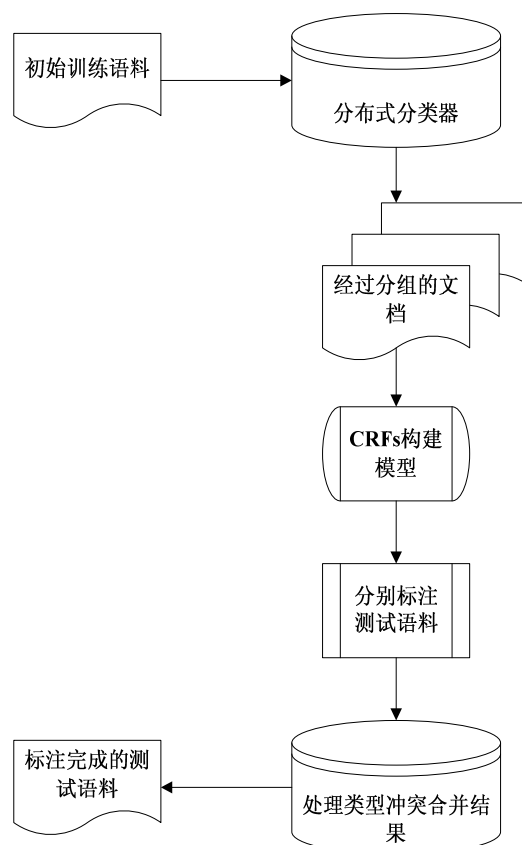


图 1 系统基本块识别工作流程图

3.1 分类器分组方法

同以往的集中式策略不同，基于 CRFs 的分布式分类器将基本块识别问题分解成为不同类别的基本块识别问题，从而降低了基本块识别的难度，缩短了构建 CRFs 识别模型的时间。本

文所采用的分布式分类器是将 8 种类型的组块分组，然后为每一个分组建立 CRFs 识别模型，进而识别测试语料中该分组的组块类型。基本块类型分类器对基本块进行分组的原则是：

(1) 使用不同特征模版进行训练，F 值有相同增减变化（同增同减）的组块类型划归为一组。以表 2 中第一分组 np、ap、dp 为例（其它分组相同），在不同模版的标注情况如图 2 所示（横坐标为不同模版，纵坐标是 F 值，单位是%）。在不同模版中 3 种组块的增减变化是相同的，所以将它们划归为一组。

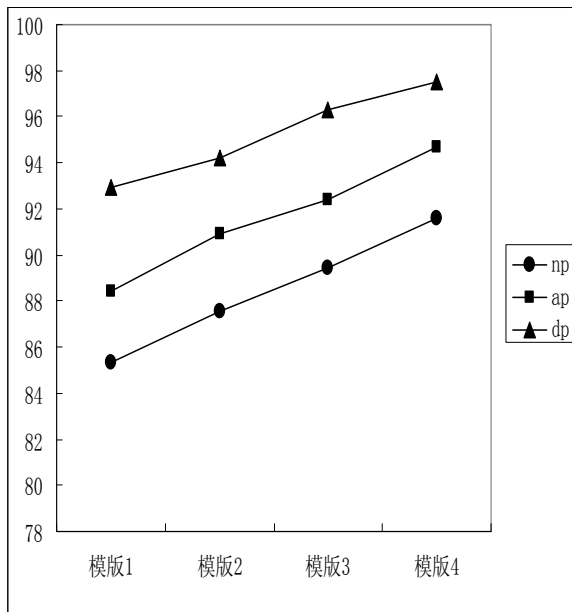


图 2 使用不同模版的基本块 F 值增减变化

(2) 本身容易产生混淆的组块类型不划归为一组。本文的基本块分组方式如表 2 所示：

表 2 基本块分组方式

组号	类型
1	np 、 ap 、 dp
2	vp
3	mp 、 mbar
4	sp 、 tp

3.2 特征选取

CRFs 模型特征集合的选取恰当与否会对识别结果造成影响。本文通过实验以及对基本块的分析，根据当前词的上下文环境确定特征集合，并根据影响当前基本块标注的几种因素确

定特征向量空间。特征的选取主要考虑的因素：

(1) 词的特征信息：当前位置词以及其前两个词和后两个词的词特征信息，即 $W_{-2}, W_{-1}, W_0, W_1, W_2$ ，特征窗口大小为 5。

(2) 词性特征信息：当前词以及前后两个窗口的词性特征信息，特征窗口大小也为 5，即 $P_{-2}, P_{-1}, P_0, P_1, P_2$ 。

基本块类型分类器的具体特征选取如下面 2 组表格所示。其中表 3 为本次评测中用到的模版，所有组块分组共用同一个特征模版。

表 3 评测所用模版

序号	模版	模版意义
1	$W_{-2}, W_{-1}, W_0, W_1, W_2$	当前词以及前后两个词
2	$P_{-2}, P_{-1}, P_0, P_1, P_2$	当前词以及前后两词词性
3	$P_0/P_1, P_{-1}/P_0, P_1/P_1, P_0/P_2$	词性组合
4	$W_1/P_0, W_{-1}/P_0$	词与词性组合

表 4 评测后改进模版

序号	模版	模版意义
1	$W_{-2}, W_{-1}, W_0, W_1, W_2$	当前词以及前后两个词
2	$P_{-2}, P_{-1}, P_0, P_1, P_2$	当前词词性及前后两词词性
3	$P_0/P_1, P_{-1}/P_0, P_1/P_1, P_0/P_2$	词性组合
4	$W_1/P_0, W_{-1}/P_0$	词与词性组合
5	$P_0/W_1/P_{-1}, P_{-1}/P_1/P_2, P_0/P_1/P_2$	词与词性组合
6	$P_0/P_{-1}/W_{-2}, P_0/P_{-1}/W_{-1}, P_0/P_{-1}/W_0$	词与词性组合

在此次评测结束后，对特征模版的选取进行了改进。考虑词性与词的关系，以及上下文环境，增加了复合特征模版的数量，从而进一步提高了结果。特征模版如表 4 所示。

3.3 基本块类型冲突处理策略

使用分类器对基本块进行分类抽取并用 CRFs 构建识别模型对测试语料进行标注，并没有完成对基本块的整体标注工作。还需要将各个分组的标注结果通过聚合器进行合并处理，完成整体基本块标注工作。但在此过程中会遇到基本块的类型冲突。如以“冲突 / 处理”为例，在不同的分组中可以标注为“冲突 B-np / 处理 I-np”或“冲突 B-np / 处理 B-vp”，这样在合并结果时就会产生类型冲突。因此就需要聚合器通过某种策略来对有类型冲突的基本块进行处理，确定一种最终结果，舍弃其它的结果。本文采用的是赋给不同分组一个权值，通过权值的高低来处理类型冲突的策略。

具体策略：综合考虑不同分组基本块标注的 F 值大小和不同分组基本块的数量两种因素来确定权值。在聚合器处理类型冲突的过程中，遇到类型冲突时将采用权值大（即优先级高）的基本块标注结果作为最终标注结果。本文基本块分组权值的大小顺序为：

np、ap、dp > vp > mp、mbar > sp、tp

除以上介绍的几种因素外，断句的长短以及名词组块的并列结构处理也都是影响最终标注结果的重要因素。

4 实验结果与分析

4.1 实验数据集及评测说明

本文采用的实验数据集合为中文信息学会句法评测 CIPS-ParsEval-2009 所提供的训练语料以及测试语料：清华句法树库 TCT，约 48 万词规模的新闻学术类语料库，侧重对规范书面语描述文件的分析处理。初始训练语料（2.89M）为 433 篇文章，其中主要以新闻语料（376 篇）为主，测试语料（714K）。本文首先使用 CRFs 对经过分组的训练语料进行训练，然后在分组标注的结果进行合并与冲突处理。

按照中文信息学会句法评测测试方案对基本块的标注结果进行分析与评价。评价指标：基本块分析的准确率(Precision)、召回率(Recall)和 F-1 测度 (F-1 measure)。

4.2 实验结果分析

在基本块的识别过程中，不同的方法会对结果造成影响，如不同特征模版、断句方式以及并列结构处理。在表 5、表 6 中给出了同一特征模版（表 3）不同方法的标注结果，表 5 是

同训练方法的基本块边界与类型总体识别结果；表 6 是不同训练方法的基本块边界、类型和关系标记的总体识别结果。两组表格包括了本次评测所提交的实验结果以及评测结束后所做的改进实验的标注结果。

在本次评测中所提交结果的权值大小以 F 值大小确定。评测后所做实验的优先级综合考虑 F 值与组块数量两种因素。

表 5 chunk boundary + chunk type 识别结果不同方法总体比较表

	方法	P	R	F
1	评测提交结果	92.48	91.74	92.11
2	权值优先级改变	92.61	92.21	92.41
3	加入并列结构处理	92.93	92.19	92.56
4	以逗号为断句标准	92.89	92.31	92.60

从表 5 中可以看出在相同模版的情况下权值优先级的改变对最终整体的 F 值影响较大。方法 1 的优先级顺序为：mp、mbar > vp > np、ap > sp、tp；以下三种方法的优先级顺序为：np、ap、dp > vp > mp、mbar > sp、tp。实验结果表明合并以及冲突处理在基本块标注过程中对结果的影响很大。

表 6 chunk boundary + chunk type + relation type 不同方法总体比较表

	方法	P	R	F
1	评测提交结果	91.31	90.58	90.94
2	权值优先级改变	91.49	91.09	91.29
3	加入并列结构处理	91.76	91.03	91.40
4	以逗号为断句标准	91.84	91.26	91.55

在评测后通过实验得出，不同特征模版的选取对基本块标注的结果会产生很大的影响。图 3 为使用评测后改进模版（表 4）并加入改进方法的标注结果与评测提交结果的特征模版（表 3）所做的对比实验。表 7 为不同类型组块的识别结果。

通过图 3 与表 7 可以看出在方法相同时，特征模版的选取成为关键。标注结果较好的模

版为表 4 中给出的模版，其 F 值达到 93.06%。但是由于处理基本块的类型冲突过程中，权值大小顺序对结果的影响，在制定相关策略时会造成实验结果的不完善。

表 7 不同模版各组块的 P、R、F 值对比

组块	评测结果			试验后改进结果		
	P	R	F	P	R	F
dp	97.37	97.16	97.27	97.13	97.90	97.51
mp	98.12	96.19	97.14	98.49	96.49	97.48
ap	95.90	92.59	94.22	95.89	93.46	94.66
vp	92.35	93.46	92.90	92.58	93.77	93.17
np	92.39	90.83	91.60	92.92	91.75	92.33
tp	91.86	89.27	90.54	92.56	89.04	90.77
mbar	88.60	90.18	89.38	92.17	89.29	90.70
sp	80.00	78.60	79.29	82.17	75.29	78.58
all	92.48	91.74	92.11	93.33	92.80	93.06

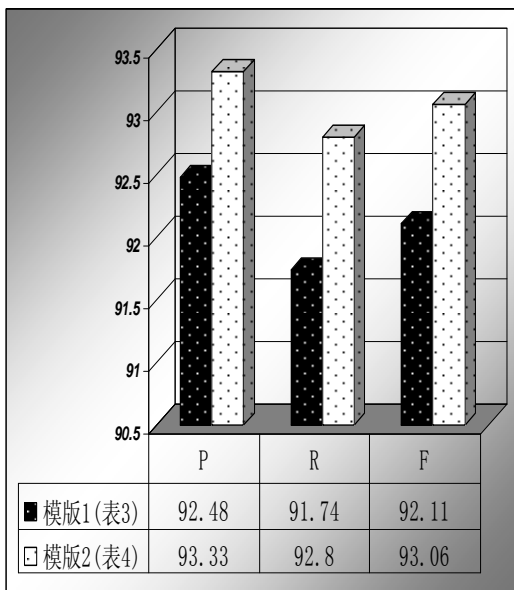


图 3 不同模版对比实验结果

实验结果的不完善主要表现在确定分组权值大小顺序策略时的不完善。如处理类型冲突合并结果时，基本块 sp 的识别。如图 4 所示（横坐标为不同模版，纵坐标是 F 值，单位是 %），在整体结果的 F 值上升的情况下，sp 会有下降的波动。但对整体结果影响不大的原因在于 sp 块数量较少。发生这种情况是因为在确

定权值优先级时，不能找到同时使整体结果的 F 值上升，又能使 sp 块的 F 值上升的权值优先级顺序，只能找到能使整体效果提高的权值优先级顺序。这使得基本块标注的结果出现不完善的现象。

在今后的工作中将会把重点放在这些数量较少的基本块上面，找出更加全面的规则使实验结果趋于完善，而不只是提高整体识别结果。

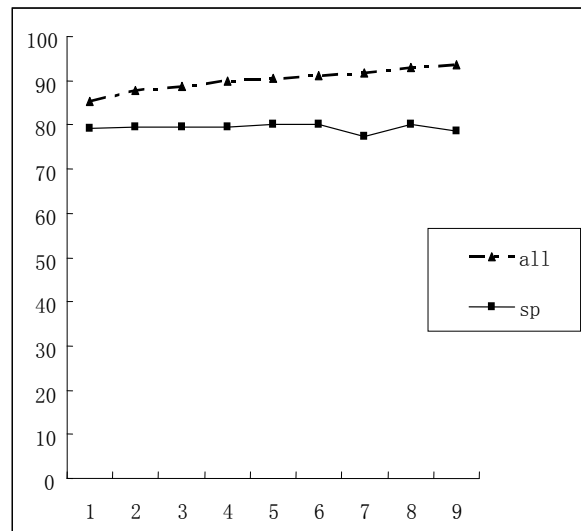


图 4 sp 块与整体标注结果比较

5 结论

本文将汉语基本块识别分为两个步骤，即通过分类器将组块进行分布式分组，结合 CRFs 构建不同分组的标注模型，这样可以减少训练时间，降低特征集合的复杂程度；通过聚合器合并各个分组结果并对基本块类型冲突进行处理，聚合器内部是按照权值大小顺序进行处理。权值大的组块类型作为最终结果，权值大小依据各分组组块的 F 值和各分组组块数量两种因素综合确定。实验表明，该方法提高了基本块识别的精确率、召回率以及 F 值，结果好于以往整体标注的结果。下一步，将借鉴更多的分词特征应用于分组策略中，增强识别结果的的正确性，进一步提高分组策略与冲突处理策略的性能。

参考文献：

- [1] Abney SP. 1991. Parsing by chunks [C]. Steven P, Abney, Carol Tenny. Principle-Based Parsing. Pages 257-278.
- [2] 詹卫东.面向中文信息处理的现代汉语短语结构规则研究 [D].北京：北京大学，1999.

- [3] Chen Wenliang, Zhang Yujie, Isahara Hitoshi. 2006. An empirical study of Chinese chunking [C]. *Coling-ACL2006(Poster Session)*, Sydney. Pages 97-104.
- [4] Erik F, Tjong Kim Sang, Sabine Buchholz. 2000. Introduction to the CoNLL-200 shared task : chunking [C]. *CoNLL-2000 and LLL-2000* . Lisbon. Pages 127-132.
- [5] 周强. 汉语基本块描述体系 [J]. 中文信息学报. 2007, (3).
- [6] Ramshaw L, Marcus M. 1995. Text chunking using transformation-based learning [C]. *Proceedings of the Third ACL workshop on Very Large Corpora*, Boston. Pages 82-94.
- [7] Sha F, Pereira F. 2003. Shallow parsing with Conditional random fields [C]. *Proceedings of Human Language Technology / North American chapter of the Association for Computational Linguistics annual meeting*, Edmonton. Pages 213-220.
- [8] Tan YM, Yao TS, chen Q, etc. 2005. Applying conditional random fields to Chinese shallow parsing [C]. *Proceedings of CICLing-2005*, Mexico. Pages 167-176.