

汉语组块分析算法*

周 强

智能技术与系统国家重点实验室
清华大学计算机科学与技术系, 北京 100084

摘要: 本文提出了一种高效的汉语组块分析算法。它通过采用基于规则的有限状态成分组分析和基于统计的词界块界定预测相结合的处理策略以及多个有限状态成分组转换器相互配合的处理机制, 在对真实文本的汉语句子的组块自动识别实验中取得了较好的处理效果。

关键词: 词界块, 成分组, 有限状态模型, 界定预测模型, 部分分析。

The Chunk Parsing Algorithm for Chinese Language

Zhou Qiang

State Key Laboratory of Intelligent Technology and Systems
Dept. of Computer Science and Technology, Tsinghua University, Beijing 100084
zhouq@sl000e.cs.tsinghua.edu.cn

ABSTRACT: In this paper, we proposed an efficient Chinese chunk parsing algorithm. Based on the following processing strategies and schemes: 1) To combine rule-based finite-state constituent parser with statistics-based word boundary predictor, 2) To use finite-state transducers for constituent group identification, It obtained better performance in the experiments of the automatic chunk identification on Chinese real texts.

KEYWORDS: Word Boundary Block, Constituent Group, Finite-State Model, Boundary Prediction Model, Partial Parsing

一、引言

句法分析是自然语言理解的基础。针对完整的句法分析方法在分析大规模真实文本中遇到的困难, 许多研究人员开始尝试着把一个完整的句法分析问题分解为几个易于处理的子问题, 以逐步降低完整句法分析的难度, 提高分析效率。在这方面, 一个很成功的例子是将词性标注(Part-Of-Speech Tagging)从句法分析中分离出来。通过利用局部语境信息进行基于规则或基于统计的词类排歧, 目前的大部分词性标注工具对真实文本的标注正确率都达到了 96%以上, 为进一步进行句法分析打下了很好的基础。

针对汉语的特点, 我们提出了一个句法描述能力介于线性词语/词性标记序列和完整句法树表示之间的浅层句法知识描述体系: 组块分析体系。它通过引入词界块和成分组概念, 将成分边界辨识问题从完整的句法分析任务中分离出来, 形成具有不同层次的成分边界限制信息的组块描述体系。作为一种基本上独立于各种句法描述形式的句子拓朴结构, 在此基础上可以方便地进行更深层次的句法知识自动获取和自动句法分析方法研究。

* 本研究得到国家自然科学基金资助, 项目号为: 69705005.

本文主要侧重于介绍这种组块描述体系的自动分析算法，即给定一句经过正确切分和词性标注处理的汉语句子，如何构造一个有效的自动分析算法，快速分析出句子的组块描述形式。

二、组块分析体系介绍

给定一句经过正确切分和词性标注的汉语句子，我们的组块描述包括以下内容：

1) 词界块 WB (Word Boundary Block)：描述了句子中每个词语所处的成分边界位置信息，简记为 $wb_i = \langle w_i/t_i, b_i \rangle$ ， $i \in [1, n]$ ，其中 w_i 为句子中的第 i 个词， t_i 为它的词性标记， b_i 可取值 0, 1 或 2，分别表示此词处于某个句法成分的中间位置、左边界或右边界。

2) 成分组 CG (Constituent Group)：描述了句子中具有如下分布特点的一些特殊分区域：I) 区域中的词界块只能与区域中的其他词界块发生句法作用，II) 整个区域作为一个整体与句子中的其他成分发生句法作用。简记为 $cg_j = \langle lp_j, rp_j, ctag_j \rangle$ ， $j \in [1, k]$ ，表示句子中总共有 k 个成分组， lp_j 和 rp_j 分别表示其中第 j 个成分组在句子中的左边界和右边界词位置， $ctag_j$ 则为其标记。目前我们总结的成分组主要有以下几种：

- 并列结构和并列成分 ($ctag \in \{CS, CC\}$)，如：{CS 哥哥 和 弟弟}
- 固定搭配组合及内部结构 ($ctag \in \{LP, MD\}$)，如：当 {MD} 时
- 标点分隔结构 ($ctag \in \{CR, SR, PR\}$)，如：, {CR}，

有关这两部分内容的精确定义和详细描述，可参阅论文[1]。

下面给出了一个具体的组块描述实例：

- 输入句子：商业/n 是/v 在/p 商品生产/n 和/c 商品交换/n 出现/v 之后/f , /w 经过/p 第/m 三/m 次/q 社会/n 大/a 分工/n , /w 在/p 简单/a 商品流通/n 的/u 基础/n 上/f 产生/v 的/u 。 /w
- 词界块描述：<商业/n, 1> <是/v, 1> <在/p, 1> <商品生产/n, 1> <和/c, 0> <商品交换/n, 2> <出现/v, 2> <之后/f, 2> <, /w, 0> <经过/p, 1> <第/m, 1> <三/m, 2> <次/q, 2> <社会/n, 1> <大/a, 1> <分工/n, 2> <, /w, 0> <在/p, 1> <简单/a, 1> <商品流通/n, 1> <的/u, 0> <基础/n, 2> <上/f, 2> <产生/v, 2> <的/u, 2> <。 /w, 2>
- 成分组描述：{<1, 26, MD>, <1, 25, PR>, <1, 8, CR>, <3, 24, MD>, <4, 6, CS>, <10, 16, CR>, <18, 24, CR>, <19, 22, MD>}
- 简化表示¹：{MD {PR {CR 商业/n [是/v {MD 在/p {CS 商品生产/n 和/c 商品交换/n} 出现/v] 之后/f} , /w {CR 经过/p [第/m 三/m] 次/q} [社会/n [大/a 分工/n] , /w {CR 在/p {MD 简单/a 商品流通/n] 的/u 基础/n} 上/f} 产生/v} 的/u }} 。 /w }

三、组块分析算法概述

近年来，国外在句法分析方法研究上的一个重要动向，是有限状态 (Finite-State) 分析方法重新得到重视和发展。文献[2]提出了一种用有限状态模型来近似描述上下文无关语法的方法，并在限定领域的语音识别实验中取得了较好的分析效果。文献[3]提出可以

¹ 为简单起见，我们用中括号表示词界标记，用大括号对表示成分组边界位置，并标上特殊标记。当两者位置重叠时，则省去中括号。

利用一组有限状态转换器来有效地分析自然语言句子的句法层次树。事实上，有限状态模型描述能力是与 Chomsky 句法层次中的 3 型文法，即正规文法等价的。因此在对一些简单句法形式的自动分析中，更能显示出它描述简单、分析效率高的处理优势。文献[4]讨论了它在词法分析中的应用。文献[5]将有限状态转换器思想应用与 Eric Brill 的基于规则的自动词类标注工具[6]中，大大提高了处理效率。S. Abney(1996)开发了瀑布式 (Cascade) 的有限状态成分组块(chunk)自动识别工具[7]。J. Senellart(1998)则利用类似的方法来准确定位大规模真实文本中的专有名词短语[8]。以上这些研究对我们进行汉语组块自动分析都有一定的借鉴意义。

针对汉语组块描述的特点，我们提出了这样的组块自动分析思想：

1) 将基于规则的有限状态成分组分析和基于统计的词界块界定预测相结合，充分发挥两种不同分析方法的处理优势，提高整体分析效率。

首先利用一组有限状态分析器，根据各个成分组的结构组合特征，自动分析出成分组边界位置。它们可以将一个句子划分为具有如下特征的若干个词段 (span) ws_j : ① $ws_j = wb_{i1} wb_{i2} \dots wb_{ik}$ ② $b_{i1} \in \{0, 1, 2\}$, $b_{ik} \in \{0, 1, 2\}$, 即词段边界上的词界块信息已确定。③ $b_{ij} = -1$, $j \in (1, k)$, 即词段中间的词界块信息没有确定。

然后利用统计模型，进行基于各个词段的词界块界定预测 (可参阅论文[9])，最终得到句子的完整的词界块和成分组信息描述。

2) 建立“瀑布式”的多层次有限状态转换器系统，自底向上地识别具有不同结构分布特点的成分组，充分利用不同成分组信息的相关性，形成整体处理优势。

首先，利用一组有限状态转换器，自动识别出句子中的一些简单成分组，包括数词串，如“一/m 九/m 八/m 五/m”，动词、形容词重叠结构，如“高兴/a 高兴/a”，“看/v 了/u 看/v”等，并设置其相应的词界块信息。它们可以作为一个整体出现在并列成分中。

接着，利用一组固定搭配词表，包括标号对，如括号对，单引号对，双引号对等，自动识别出句子中的具有不同层次的固定搭配结构 (允许有嵌套)，其中的某些固定搭配结构，如标号对结构等，也可以作为一个整体出现在复杂的并列成分中。

然后，利用并列成分的内容相似性特征，进行并列结构的边界自动识别，有关算法的详细情况将在下一节介绍。

最后，利用汉语点号描述的层次性

表 1 汉语标点符号的层次性

点号	分隔成分作用	层次标记
、	分隔并列结构中的并列成分	1
,	分隔短语片段和小句	2
； :	分隔比较大的分句和引句	3
。 ? !	分隔完整的句子	4

特点 (见表 1)，自底向上地识别出具有不同层次的点号成分组。在这一过程中，需对两种不同的点号，即受限点号 (在并列结构或固定搭配结构中出现的点号) 和自由点号进行不同方式的处理。

综合以上的处理思路，我们形成了这样的汉语组块分析算法：

输入：一个经过正确切分和词性标注处理的汉语句子。

输出：词界块描述： $\{ \langle w_i / t_i, b_i \rangle \}$ ，成分组描述： $\{ \langle l_j, r_j, ctag_j \rangle \}$ ，两者组合形成简化表示形式。

背景知识：1). 汉语简单成分组组合规则集。2). 汉语固定搭配词语表。3). 词语界定信息分布数据。

分析流程：

- 简单成分组的自动识别；

- 固定搭配结构的自动识别;
 - 并列结构的自动识别;
 - 标点分隔结构的自动识别;
 - 基于统计的词界块自动界定预测;
-

四、并列结构的自动识别

一个并列结构(Conjunctive Structure,CS) 短语是由两个或两个以上的并列成分(Conjuncts, CCs)平等相联在一起而构成的。按照并列短语是否包含形式标记,我们把它分为两大类:1). 使用形式标记的并列短语,主要有以下几种标记:a). 并列连词,包括“和、与、并、或……”等;b). 关联副词,如,“能说也能写”;c). 标点符号,主要是顿号、逗号和分号。2). 没有形式标记的并列短语,如:“繁荣富强”,“伟大的光荣的正确的”等。

两种不同类型的并列短语,为自动识别处理提出了不同的要求:

首先,通过检测特殊的形式标记,可以很容易地发现句子中是否存在属于类型1)的并列结构。不过,需要注意的是,这里的几个形式标记对并列结构的标识作用是有强弱差别的:一般来说,并列连词和顿号是很强的并列标记,据此可以比较明确地确定句子中存在着并列结构;而关联副词、逗号和分号则是比较弱的并列标记,往往还需根据左右成分的初步相似性判断(如:判断词类是否相等)来确定是否存在可能的并列结构。

而对于属于类型2)的短语,则只能根据两个成分之间的相似性判断来确定句子中是否存在这样的并列结构。但由于此类并列结构的并列成分之间一般存在着较强的结构对应关系,因此通过对词类信息的相等性判断一般可以发现大多数可能的此类并列结构。

发现了并列结构的可能存在,下一步的处理是要准确地确定其中不同的并列成分,进而确定并列结构的边界位置。这是并列结构识别中的难点所在,也是我们所要研究的主要问题。文献[10]提出了一种日语并列结构边界自动识别算法。它通过构造一种日语词节的相似度计算方法,利用动态规划方法计算路径相似度评分,从中选择具有最高评分的路径来设置并列结构的边界,正确地识别出了句子中的绝大部分并列结构短语。考虑到汉语和日语在并列结构描述中的一些相似性,我们吸收了他们的基本处理思想,形成了效率较高的汉语并列结构自动识别算法。

我们的处理依据了汉语中这样一个基本假设,即认为“词性相同、结构相同、语义类相同、音节数相同的项并列是最理想、最严格的并列”。它主要包括以下几个步骤:

1). 通过形式标记检索和成分模糊匹配处理,发现句子中所有可能的并列成分中界位置(对于类型1)的并列短语,为中间的形式标记所处的位置;对于类型2,则为左边的并列成分的最后一个词语的位置)及其最大边界,形成一组可能的并列短语描述向量: $\langle \text{MidPos}, \text{MaxLen} \rangle$ 。

2). 对每个可能的并列短语描述向量,设置成分相似度计算矩阵,通过基于动态规划的最佳路径选择和路径评分阈值的限制判断,首先确定该描述向量是否可能形成一个并列结构,然后再设定并列成分的准确边界,以形成一个部分并列结构描述向量 $\langle \text{LP}, \text{MP}, \text{RP} \rangle$ 。

3). 考虑到句子中的多个并列结构之间可能存在的成分嵌套和边界重叠现象,还需对这些并列结构描述向量进行边界调整和成分合并操作,以形成完整的并列结构,据此可以设定并列短语和它的并列成分的边界位置。

因论文篇幅限制,更详细的内容将另文介绍。

五、实验结果分析

表 2 实验语料的基本统计数据

我们的实验语料主要分为两大部分（表 2 列出了它们的一些基本统计数据）：

1. 封闭测试语料：选自作者在北大开发的一个汉语树库[11]，内

容涉及两大类：① 汉英翻译研究的测试题库语料，② 新加坡小学语文课本语料。其特点是句子较短，句型多样，覆盖的语言现象较全面。从中我们提取了组块分析算法所需的各种知识，包括成分组结构规则和词语界定统计分布数据等。

2. 开放测试语料：选自真实文本，内容涉及经济、军事、新闻等领域。其特点是每个句子中都至少包含一个并列结构标志词（顿号或并列连词）。因此句子较长，句法结构也比较复杂。对这部分语料进行了自动组块分析和人工校对，形成了正确的组块标注结果，以作为分析比较的依据。

对这两部分语料进行组块自动识别实验，并对实验结果进行了以下性能指标的分析，得到了表 3 和表 4 的结果。在 PII—233，内存 32M 的 PC 机上，分析速度约为 25 句/秒（开放测试）。

1) 成分组边界正确率(CGP) = 具有正确的边界位置的成分组总数 (Cort_CG) / 识别出的成分组总数 (ECG)

2) 成分组边界交叉率(CGC) = 与正确成分交叉的成分组总数 (Crossed_CG) / 识别出的成分组总数 (ECG)

3) 词界块界定预测正确率(WBP) = 具有正确边界预测的词界块总数(Cort_BP) / 语料中的词项总数(WBSum)

从表中数据可以看出，大多数成分组的自动识别准确率都在 90%以上，反映了目前的有限状态模型对成分组边界识别的有效性。但其中的并列结构识别效果较差，主要原因是在真实文本中，并列结构的组合情况比较复杂，与传统意义上的并列结构假设有较大差距。目前我们正通过搜集大量真实文本中出现的并列结构实例，不断改进和完善现有的并列结构自动识别算法，争取逐步提高其分析精度。

表 4 组块分析实验的词界块边界预测结果

	WBSum	Cort_BP	WBP
封闭测试	64426	62333	96.75%
开放测试	32389	29891	92.29%

一些基本名词短语自动识别技术，提高界定预测算法的处理精度，将是我们今后研究的一个重点。

	句子总数	词总数	汉字总数	平均句长 (词/句)
封闭测试	5573	64426	89492	11.56
开放测试	1071	32389	51171	30.24

表 3 组块分析实验的成分组边界识别结果

成分组 标记	封闭测试		开放测试	
	CGP(%)	CGC(%)	CGP(%)	CGC(%)
CS	66.25	31.09	64.46	26.16
CC	91.39	7.09	72.95	7.65
LR	100.00	0.00	100.00	0.00
MD	98.97	0.60	98.82	0.07
CR	97.00	2.78	92.90	2.87
SR	88.57	9.14	92.86	2.04
PR	99.79	0.15	99.48	0.21

在词界块边界预测方面，开放测试语料的分析准确度下降了 4%。对其错误实例进行分析，发现它们主要集中在一些与双音节动词有关的基本名词短语 (baseNP) 组合上，如：“n v”和“v n”结构等。由于这些组合在真实文本中出现较频繁，因此产生了较多的界定预测错误。如何利用

六、结语

本文提出了一种高效的汉语组块分析算法。它通过采用基于规则的有限状态成分组分析和基于统计的词界块界定预测相结合的处理策略和多个有限状态成分组转换器相互配合的处理机制,在对真实文本的汉语句子的组块自动识别实验中取得了较好的处理效果。

对目前的组块分析算法的不断改进和完善,将有助于我们逐步形成一个描述能力介于线性的词语、词类标记序列和完整的句法树之间的汉语浅层句法知识描述平台:组块分析体系。在此基础上,可以方便地进行各种句法分析和知识获取实验,包括开发基于统计的汉语句法分析器[12],自动识别汉语最长名词短语[13],自动获取汉语概率性上下文无关语法知识[14]和结构优先关系知识[15],以及自动获取汉语动词的搭配知识等,从而为中文信息处理的许多应用领域的研究,包括信息检索、机器翻译、信息抽取等,提供有力的支持。

参考文献

- [1] 周强,孙茂松,黄昌宁(1998).“汉语句子的组块分析体系”,已被《计算机学报》录用。
- [2] Fernando C. N. Pereira and R. N. Wright (1997). “Finite-State Approximation of Phrase-Structure Grammars”, In *Emmanuel Roche and Yves Schabes (eds.) Finite-State Language Processing, The MIT press, USA, 149-174.*
- [3] Emmanuel Roche(1997). “Parsing with Finite-State Transducers”. In *Emmanuel Roche and Yves Schabes (eds.) Finite-State Language Processing, The MIT press, USA, 241-282.*
- [4] David Clemenceau(1997). “Finite-State Morphology: inflections and Derivations in a Single Framework Using Dictionaries and Rules”, In *Emmanuel Roche and Yves Schabes (eds.) Finite-State Language Processing, The MIT press, USA, 67-98.*
- [5] Emmanuel Roche & Yves Schabes(1995). “Deterministic Part-of-Speech Tagging with Finite-State Transducers”. *Computational Linguistics*, 21(2), 227-253.
- [6] Eric Brill (1992). “A Simple Rule-Based Part Of Speech Tagger”. In *Proceedings, Third Conference on Applied Natural Language Processing*. Trento, Italy, 152-155.
- [7] Steven Abney(1996). “Partial Parsing via Finite-State Cascades”. In *Proc. of the ESSLLI'96 Workshop*.
- [8] Jean Senellart (1998). “Locating noun phrase with finite state transducers”, In *Proceedings of COLING-ACL '98*. Menthol, Canada, 1212-1218.
- [9] 周强(1996). “一个汉语短语自动界定模型”,《软件学报》第7卷,增刊,315-322.
- [10] S. Kurohashi & M. Nagao. (1994). “A syntactic Analysis Method of Long Japanese Sentences Based on the Detection of Conjunctive Structures”, *Computational Linguistics*, 20(4), 507-534.
- [11] 周强 (1996). “汉语语料库的短语自动划分和标注研究”,博士学位论文,北京大学计算机系,1996.6.
- [12] Qiang Zhou. (1997) “A Statistics-Based Chinese Parser”, In *Proc. of the Fifth Workshop on Very Large Corpora*, 4-15.
- [13] 周强,孙茂松,黄昌宁(1998).“汉语最长名词短语的自动识别”,已被《软件学报》录用。
- [14] 周强,黄昌宁(1998).“汉语概率型上下文无关语法的自动推导”,《计算机学报》,21(5),385-392.
- [15] 周强,黄昌宁(1999).“汉语结构优先关系的自动获取”,《软件学报》,10(2),149-154.