基于汉语多词块的语料库研究

钟立军，李　茹，彭洪保

（山西大学 计算机与信息技术学院，山西 太原 030006）

E-mail: zhonglijun081181@163.com

摘　要：组块分析是自然语言的处理研究领域中新近出现的一个语言处理策略，它能有效降低句法分析的难度。在汉语多词块描述体系的基础上，本文着重描述了汉语多词块库中块的结构，并对库中块的各种标记进行了深入地统计和分析。经过测试，汉语多词块库是一个较准确的参照库，在今后汉语多词块的自动识别研究中，可以很好地得到应用。

关键词：组块分析；多词块；拓朴结构；中心依存

The Study of the Corpus Based on the Chinese Multi-word Chunk

ZHONG Li-jun,LI Ru,PENG Hong-bao

(School of Computer & Information Technology , Shanxi University , Taiyuan , Shanxi 030006)

E-mail: zhonglijun081181@163.com

Abstract: The chunk parsing is a language processing strategy which recently appears in the natural language processing research area. It can effectively reduce the difficulty of the syntactic parsing.On the basis of the Chinese multi-word chunk description system, this article describes emphatically the chunk structure of the Chinese multi-word chunk bank, counts and analyzes various chunk marks in the bank. After the test, the Chinese multi-word chunk bank is a considerably accurate reference bank. This bank can be used well in the next automatic identification research of the Chinese multi-word chunk.

Key words: chunk parsing; Multi-word Chunk; topological structure; central interdependence

1　引言

句法分析是在自然语言的处理研究领域中的重点和难点。在分析大规模的真实文本中，许多研究人员发现进行完全句法分析时，面临的困难较大。所以开始尝试着把一个完全句法分析问题分解为几个易于处理的子问题，从而降低句法分析的难度，提高分析的效率。这样就引出了部分句法分析（partial parsing）。部分句法分析，又叫浅层句法分析（shallow parsing）或组块分析（chunk parsing）[1]，是近来出现的一个新的语言处理策略。它是跟完全句法分析相对的。组块分析可以看成将完全句法分析分解为两个子任务：（1）组块的识别和分析；（2）组块之间的句法关系分析。组块分析的主要任务是组块的识别和分析。这样能在一定程度上简化句法分析的任务，使句法分析技术在大规模真实文本处理系统中迅速得到应用。

目前，大部分词汇分析（包括分词和词性标注）工具对真实文本的标注正确率都较高，为在此基础上进一步进行句法分析打下了很好的基础。在英语方面,Abney (1991) 把块定义为句子中一组相邻的属于同一个s-投射的词语的集合,建立了块与管辖约束理论的X-bar 系统的内在联系,

基金项目：国家863高技术研究发展计划资助项目(2006AA01Z142)

作者简介：钟立军(1980 —)，男，硕士生，主要研究方向为自然语言处理；彭洪保(1980 —)，男，硕士生，研究方向为自然语言处理；李茹(1963 —)，女，教授，研究方向为智能信息处理。

从而奠定了这个块描述体系比较坚实的理论基础[2]。在汉语方面，国内的组块研究团队主要有清华[3]和哈工大[4]的基本短语描述体系、微软的块描述体系[5]和北大的实语块描述体系[6]等。这些体系的共同点在于它们都是从句法层面上来定义和描述块信息，主要侧重块边界确定和句法成分标注问题，不太关心各个块的内部关系分析。

汉语的多词块描述体系[7]是一套基于拓扑结构的块描述体系，它通过引入词汇关联知识库来确定基本拓扑结构，形成了较好的多词块内聚性判定标准，建立了句法形式和语义内容的有机联系桥梁。

2 汉语的多词块描述体系

多词块（Multi-word Chunk，简称MWC）[7]是一种具体的组块，它是由两个或两个以上的词语按照一定的关联关系组合形成的信息描述单位。这些词语直接相邻，互相联系比较紧密。MWC是以名词、动词、形容词等实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，如：连词、叹词、语气词、助词、标点符号等。MWC一般由1-3 个词语组成，通过不同的外部句法表现和内部词汇关联关系，形成各自特殊的概念内容描述体，成为汉语的字、词进入组块成句过程的基础和出发点。

MWC 的主要特点是块内部的各个词语按照一定的句法关系聚合到一个句法语义中心词上，可以通过这个中心词体现整个多词块的外部功能。其描述核心是以下三种基本拓朴结构[7]：左角中心结构(LCC)、右角中心结构(RCC)和链式关联结构(CHC)。左角中心结构(LCC)：块中的所有词语直接依存到左角中心词，形成一个左向中心依存结构；右角中心结构(RCC)：块中的所有词语直接依存到右角中心词，形成一个右向中心依存结构；链式关联结构(CHC)：块中的各个词语依次依存到其直接右相邻的词语，形成一个自左向右排列的多中心依存关系链。

下面的图1 显示了这三种拓朴结构的基本形状。句子中两个或多个词语能形成一个MWC的充要条件是它们的内部词汇关联能形成以上三种拓扑结构中的一种结构组合形式。根据其中不同的中心词和内部句法关系及其外部功能表现，可以把这些MWC分成三大类：体词块、谓词块和修饰块[7]。

对于符合以上条件的每个 MWC，除了使用边界标记‘[’和‘]’外，还使用三个块标记：句法标记（成分标记）、关系标记和序列标记组合。如：在多词块“[vp-AD-HI 作出/v 了/u]”[①]中，成分标记是 vp，关系标记是 AD，序列标记组合是 HI。左边界标记‘[’表示块的左边界，右边界标记‘]’表示块的右边界，vp 体现了 MWC 的外部句法功能，AD 体现了 MWC 的内部词汇关联，HI 与块内每个词建立起一一对应的关系，能更具体、详细地体现出三种重要的基本拓朴结构。下面的表 1 列出我们目前所用的主要句法标记和关系标记。其中体词块的典型关系标记为：ZX, LN, LH, AD 等。谓词块的典型关系标记为：PO, SB, AD, ZX, LH 等。修饰块的典型关系标记为：JB, AD, ZX 等。（序列标记在后面具体介绍）

为了保持体系的完整性，表中增加了‘SG’标记，以描述句子中功能类似的由单个词语直接形成的块。

---

① 注：清华大学汉语句法树库 TCT 标注句子中使用的词类（词性）标记简要说明如下：t-时间词，f-方位词，r-代词，p-介词，v-动词，n-名词，u-助词，a-形容词，d-副词，m-数词，q-量词，rN-名代词，qN-名量词，vB-趋向动词，vN-名动词。

表 1　中心聚合块的句法标记和关系标记描述集

句法标记	内容描述	关系标记	内容描述
np	名词块	ZX	右角中心结构
mp	数量块	LN	链式关联结构
sp	空间块	LH	并列关系 CHC
tp	时间块	PO	述宾关系 LCC
vp	动词块	SB	述补关系 LCC
ap	形容词块	AD	附加关系 LCC
dp	副词块	JB	介宾关系 LCC
pp	介词块	SG	单词语块
mbar	数词准块	CD	重叠关系

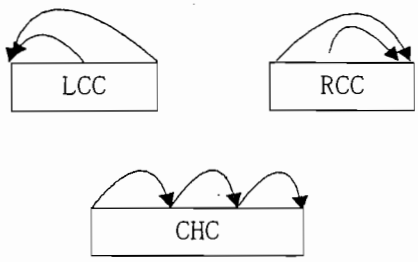

图 1　MWC 的三种典型拓朴结构

3　序列标记的设计

先介绍一下针对三种典型拓朴结构的 MWC 序列标记组合中的序列标记：

1)　左角中心结构：通过标记集合：{P, I, C, O, H}来标注不同位置词语。其中：

P：表示处于左角中心位置的谓词或介词；

I：表示中心右边的功能性附加词语，如：时态助词“了、着、过”等，对应‘AD’标记；或中间可能的结构助词‘得’，对应‘SB’标记；

C：表示中心右边的补语位置关联词，对应‘SB’标记；

O：表示中心右边的宾语位置关联词，对应‘PO’和‘JB’标记；

H：表示附加结构中心位置词语，如：“打 了”中的‘打’位置词语；

2)　右角中心结构：通过标记集合：{R, S, P, O, C, I, M,H}来标注不同位置词语。其中：

R：表示右角中心位置词语；

S：表示定语从句的主语位置词语；

P、O、C：分别表示定语从句的谓语、宾语和补语位置词语；

I：表示中间可能的分隔结构助词，如：‘的’、‘地’等；

M：表示中心左边的各个修饰位置关联词，对应‘ZX’标记；

H：表示附加结构中心位置词语，如：“红 的”中的‘红’位置词语；

3)　链式关联结构：通过标记集合：{H, K, J, I, M}来标注不同位置词语。其中：

H：表示整个结构最右边的中心词语位置，该位置上的词语将形成句法语义中心词；

K：表示结构内部的链接中心词语位置，对应‘LN’标记；

I：表示中间可能的分隔功能词，如结构助词‘的’、并列连词‘和’以及顿号等；

M：表示结构最左边的修饰位置功能词；

J：表示并列结构的各个并列成分位置，对应‘LH’标记。

另外，增加‘B’标记表示所有单独成块词语，‘X’标记表示所有不在 MWC 的其它功能词，这样形成了完整的块（包括 MWC 和单词语块）序列标记集。这些序列标记分别与块内或块外的词语形成一一对应的关系。MWC 和单词语块的标记标注不同主要是：单词语块缺省了序列标记‘B’的标注。

4 标准库中块的分布分析

对 TCT[8]标注树进行深入分析，利用其提供的丰富的句法信息，从拓扑结构出发，确定其中各个 MWC 的分布特点，设计合适的自动提取算法，从中提取出准确的 MWC 标注结果。将 TCT 中没有被这些 MWC 覆盖的实词直接上升为单词语块。这样就得到了块（包括 MWC 和单词语块）的标准标注库。经过测试，这个库中块的准确度还是很高的，只存在少数个别错误。

将由 TCT 得到的块标准标注库中 20 万词新闻语料（包括回忆录和时事报道，共有 185 篇文章）分成两个数据集：训练集和测试集。切分比例为 8：2。两者的选择方法应与其它层次的块分析器处理保持一致，以便于进行相关工具的对接和性能对比分析。

在块的标准标注库中，通过设计出的几组实验，对块的边界标记、成分标记、关系标记和序列标记组合（或序列标记 B）的相关数据分别进行统计，再结合相关的算法，将计算结果生成相应的表，具体如下：

表 2 边界标记在块标准标注库中的分布

边界标记	某标记在测试集中的个数	某标记在训练集中的个数	某标记在测试集中的个数/测试集中标记的总数	某标记在训练集中的个数/训练集中标记的总数
MWC 的边界标记	8252	38644	55.43%	54.77%
单词语块的边界标记	6635	31907	44.57%	45.23%
TOTAL	14887	70551	100%	100%

表 3 句法标记在块标准标注库中的分布

句法标记	某标记在测试集中的个数	某标记在训练集中的个数	某标记在测试集中的个数/测试集中标记的总数	某标记在训练集中的个数/训练集中标记的总数
np	6323	28980	42.47%	41.08%
vp	5213	24910	35.02%	35.31%
sp	253	1224	1.70%	1.73%
mp	483	2510	3.24%	3.56%
ap	339	1836	2.28%	2.60%
tp	641	3028	4.31%	4.29%
dp	1310	6290	8.80%	8.92%
pp	272	1404	1.83%	1.99%
mbar	53	369	0.36%	0.52%
total	14887	70551	100%	100%

表 4　关系标记在块标准标注库中的分布

关系标记	某标记在测试集中的个数	某标记在训练集中的个数	某标记在测试集中的个数/测试集中标记的总数	某标记在训练集中的个数/训练集中标记的总数
ZX	4919	22899	33.04%	32.46%
PO	900	4252	6.05%	6.03%
SG	6635	31907	44.57%	45.23%
AD	583	2941	3.92%	4.17%
LN	1170	5306	7.86%	7.52%
LH	165	674	1.11%	0.96%
SB	237	1110	1.59%	1.57%
JB	271	1400	1.82%	1.98%
CD	7	62	0.05%	0.09%
TOTAL	14887	70551	100%	100%

表 5　组合(或 B)在块标准标注库中的分布

序列标记组合(或序列标记 B)	某组合（或 B）在块标准标注库中的个数	某组合（或 B）的个数/组合和 B 的总数
B	38542	45.1111%
MR	18446	21.5899%
PO	5726	6.7019%
MKH	3700	4.3306%
MMR	3339	3.9081%
HI	2648	3.0993%
MIR	1725	2.0190%
PC	1158	1.3554%
……	……	……
TOTAL	85438	100%

从表 3 中可以看出，描述实体内容的名词块（np）和描述动作状态的动词块（vp）占了大多数，它们是研究的重点和难点。实际上，在真实文本句子中确实如此，它们富含了句子所反映的大量事件内容。具体有如下特点：(1) 一般来说，动词块的平均长度比名词块的要短，而且名词块的各种词语组合情况比动词块的要多，这说明名词块的内部结构更复杂，有更多的句法语义内容。(2) 若干名词块和动词块还可以组合成其它的更长的块，其中典型的歧义结构，如："np vp 的 np"、"vp np 的 np" 等，是句法分析中较大的处理难处。(3) 名词块、时间块（tp）、空间块（sp）之间经常容易发生混淆，主要因为汉语中的方位词（f）经常会同时表达时间方位和空间方位，因此此类混淆很容易发生，且较难区分，经常需要结合上下文和具体的语境做出准确地判断。如："[sp-ZX-MR 老百姓/n　中间/f　]"、"[tp-ZX-POR 获得/v　新生/n　后/f　]"、

“[np-LN-MKH 一/m 株/qN 银杏/n]”。

从表 4 中可以看出，单词语块（SG）、右角中心结构（ZX）、述宾关系（PO）、链式关联结构（LN）占了大多数，尤其 SG 和 ZX 占得比重较大。具体有如下特点：（1）每个单词语块中的词为实词且只含有 1 个词，其长度较短，内部结构较简单。（2）ZX 称为偏正结构，可以分为定中结构（DZ）和状中结构（ZZ），ZX 和 LN 描述的块内词汇关系较接近，区分度较小，如“[np-ZX-MMR 新疆/nS 群众/n 集会/n]”和“[np-LN-MKH 部队/n 官兵/n 会议/n]”的内部修饰关系不同：新疆/nS → 集会/n，群众/n → 集会/n；部队/n → 官兵/n, 官兵/n → 会议/n。（3）PO 包含了汉语句子所描述的事件内容的主体信息，PO 块中各种词语组合情况较多，其内部结构较复杂。

从表 5 中可以看出，序列标记 B 和序列标记组合 MR、PO、MKH、MMR 等占了大多数。具体有如下特点：（1）单词语块中 SG 和 B 是一一对应的。（2）这些序列标记组合中的序列标记分别跟其对应 MWC 中的词语形成一一对应的关系。（3）MR 对应的词性组合一般为“a+n”、“n+n”、“d+v”、“m+q”等；当 PO 对应的词之间的关系为述宾关系时，词性组合一般为“v+n”、“v+r”等；当 PO 对应的词之间的关系为介宾关系时，词性组合一般为“p+n”、“p+r”等；HI 中 I 对应的词一般为“了、着、过、地、的、等、一样、般”等；HIO 中 I 对应的词一般为“了、着、过”等；MIR 中 I 对应的词一般为“的、之”等；PC 对应的词性组合一般为“v+a”、“v+p”、“v+vB”等。

5 实验总结和块的实例

从上面的表中，可以统计得到：（1）在块标准标注库中，共有 85438 个块。其中，单词语块 38542 个，占全部块的 45.11%；多词块 46896 个，占全部块的 54.89%。（2）块中成分标记有 9 种，关系标记有 9 种，序列标记的组合有 969 种，序列标记有 12 种。（3）虽然序列标记的组合的种类很多，但出现频率较高的主要有 17 种，具体是 MR、PO、MKH、MMR、HI、MIR、PC、MKIH、MIMR、JIJ、JJ、MMMR、HIO、MKKH、MMKH、PCO、POR，这些组合和 B 一起共占全部序列标记组合和序列标记 B 的 93.17%，所占比例比较大。（4）块内词的个数范围从 1~26。含有 1~3 个词的块有 78491 个，共占全部块的 91.689%；含有 4~5 个词的块有 5637 个，共占全部块的 6.5977%；含有 1~5 个词的块有 84128 个，共占全部块的 98.4667%。（5）在由块标准标注库得到的训练集和测试集中，大部分块的边界标记、成分标记、关系标记和序列标记组合（或序列标记 B）分别占得比重差别很小。如：在训练集中，ZX 占全部关系标记的 32.46%；在测试集中，ZX 占全部关系标记的 33.04%。

经过深入分析，可以得到相应的结论：（1）在块标准标注库中，单词语块占得比重跟多词块占得比重比较接近，看来库中存在大量的单个实词，这些词不易与其它词构成组合。（2）库中块的序列标记的组合的种类较多，说明块中存在多种词组合情况。实际上，较短的块，内部结构关系较简单；而较长的块，内部结构关系较复杂。（3）这 17 种序列标记组合主要由序列标记 M、R 等组成，它们反应了真实文本句子中大量的句法语义关系。（4）库中绝大部分块的长度在 5 个词以内，尤其在 3 个词以内。看来块中含有的词一般不多，这样便于进行浅层句法分析。（5）训练集和测试集划分得比较合理，便于以后进行训练和测试。

标准标注库中块的一些实例：

（1） 如果/c 因为/c [np-SG 我们/rN] [pp-JB-PO 在/p 这里/rS] [vp-SG 办学/v] ，/，

[dp-SG 就/d] [vp-PO-PO 毁掉/v 它/rN] ,/, 对/p [np-SG 历史/n] 来说/u ,/, [np-SG 我们/rN] [dp-SG 将/d] [vp-SG 是/vC] [np-ZX-MIR 千古/t 的/u 罪人/n] 。/。

（2）[tp-ZX-POR 获得/v 新生/n 后/f] ，/, [np-ZX-MR 溥杰/nP 先生/n] [np-LN-MKIH 最/dD 大/a 的/u 愿望/n] [dp-SG 就/d] [vp-SG 是/vC] [vp-SG 能/vM] [vp-PO-PO 当好/v 普通人/n] 。/。

6 总结和展望

本文提到的块标准标注库是重要的基础资源，库中存在许多经过准确标注了的 MWC 和单词语块。块内词约占全部词的 80%，块外词约占全部词的 20%。在汉语多词块的自动识别研究，可以尝试使用几种机器学习统计模型，如条件随机场（Conditional Random Field，简称 CRF）[9]、隐马尔可夫模型（Hidden Markov Model，简称 HMM）[10]等，选择适当的特征，引入一些重要的规则，即：使用规则和统计相结合的方法，对块进行机器标注。然后参照块标准标注库，分析使用不同模型对块的标注性能。进一步理解这几种模型的算法和内在原理，结合其它的有效方法和资源，不断努力提高块的自动识别效果。

致谢：

感谢清华大学周强老师提供的语料库和相关论文资料，以及对本文提出的意见和建议。

参 考 文 献

[1] 周强, 孙茂松, 黄昌宁. 汉语句子的组块分析体系[J]. 计算机学报, 1998, 22(11): 1158-1165.

[2] Steven Abney. Parsing by Chunks [A]. In : Robert Berwick , Steven Abney and Carol Tenny (eds.) Principle-Based Parsing [C] . Kluwer Academic Publishers, 1991. 257-278.

[3] 汉语基本短语标注规范[R] . 清华大学计算机系智能技术与系统国家重点实验室, 技术资料, 2002 年2 月.

[4] Tiejun Zhao, Muyun Yang et al. Statistics Based Hybrid Approach to Chinese Base Phrase Identification [A] . In : Proc. of the Second Chinese Language Processing [C] . ACL 2000 , Hong Kong.

[5] H. Li , C. N. Huang , J . Gao , and X. Fan. Chinese Chunking with Another Type of Spec [A] . In : Proceedings of the 3rd ACL SIGHAN Workshop [C] . Barcelona , Spain , 2004. 41-48.

[6] 孙宏林. 现代汉语非受限文本的实语块分析[D]. 北京大学计算机系博士学位论文, 2001. 5.

[7] 周强. 汉语基本块描述体系[J]. 中文信息学报, 2007, 21 (3) : 21-27.

[8] 周强. 汉语句法树库标注体系[J]. 中文信息学报, 2004, 18 (4) : 1-8.

[9] J. Lafferty, F. Pereira, A. McCallum. Conditional Random Fields: probabilistic models for segmenting and labeling sequence data [C] . In : Proc. of the 18th International Conference on Machine Learning [C]. San Francisco, 2001. 282-289.

[10] Lawrence R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition [A]. In : Proceedings of the IEEE [C]. 1989, 77 (2) : 257–286.