汉语组块的定义和获取

李素建 刘群

北京大学计算语言学研究所 100871
Email: lisujian@pku.edu.cn

摘要：组块是介于词语和句子之间的一种语言结构，目前还没有明确的定义。本文总结了当前对组块的各种研究，对汉语组块进行了定义。同时组块的获取和收集也是一项迫切的任务，由于不易直接获取到具有组块标注的语料，我们从现有树库中抽取组块。本文根据汉语特点提出了 12 种汉语组块类型，并根据这些组块类型和宾州大学中文树库短语类型的对应关系进行转化获得组块库。

关键词：组块，组块语料库，树库，语法分析

Research on Definition and Acquisition of Chunk

Li Sujian, Liu Qun

Institute of Computational Linguistics, Peking University, Peking, China, 100871
Email: lisujian@pku.edu.cn

Abstract: Chunk is a kind of linguistic structure between word and sentence, which isn't defined definitely now. This paper summarizes various current researches on chunks, and defines what is a Chinese chunk. At the same time, the acquisition and collection of chunks are a hard but urgent work. Due to the difficulty of acquiring chunked corpus, we adopt the method of converting from Treebank available. According to the characteristics of Chinese, 12 Chinese chunk categories are proposed. Then our chunked corpus is obtained by extracting from Upenn Chinese Treebank.

Keywords: Chunk, Chunked corpus, Treebank, Syntactic parsing

1 引言

当前 Internet 的发展促进了信息的交流，文本的获取和收集变得相对容易。然而构建大规模标注语料库的任务却依然严峻，这是因为标注标准的不一致性，同时需要耗费大量的人力和物力。目前经过词性标注加工的英文、中文语料库已经具有一定规模，对于更高层次上的语言加工，Upenn 英文树库是一个大规模的句法标注语料库，汉语方面的成果包括：清华大学的汉语测试树库[1]，美国宾州大学的中文树库[2]，和东北大学的中文语义树库[3]。随着部分分析技术的发展和应用范围的不断扩大，粒度处于词和句之间的组块标注语料的开发也越来越受到重视。CoNLL-2000[4]会议提供了从 Upenn 英文树库中抽取出来的英文组块库；虽然清华大学提出了一套语块标注体系，构建了 200 万字的汉语语块库。但汉语组

块的定义及组块库的构建，仍然需要做进一步的研究。

第 2 节详细介绍了相关的组块研究工作，并据此给出了本文的组块定义；第 3 节详细说明了所定义的组块类型及相应类型组块的获取；第 4 节对全文进行了小结。

2. 组块的研究

人们一直都是对词或整句进行研究。组块是较词语复杂、句子简单的成分，对它的定义，一直没有一个明确的定义。下面介绍一些典型的组块研究，并给出本文对组块的界定。

2.1 组块的研究

Abney[5]最早提出了一个完整的组块描述体系，对组块有着权威性的定义。他把组块定义为从句内的一个非递归的核心成分。这种成分包含核心成分的前置修饰成分，而不包含后置附属结构。组块不一定覆盖整个句子，例如：常有一些介词、连词等不是任何一个组块的部分。Buchholz[6] 、Veenstra[7]也分别对 NP、VP、PP 等组块类型及自动识别方法进行了比较完整的研究工作。这些研究都为 CoNLL-2000 提出的组块共享任务奠定了基础。

中文组块最初侧重对基本名词短语、最长名词短语、以及命名实体等的研究[8,9,10]。但汉语句法中除了大部分名词块外，还有很多其他结构的组块。东北大学针对机器翻译提出了扩展组块（E-Chunk）的概念[3]。清华大学对整理和加工中文组块库作了大量工作，建立了一个完整的组块划分体系，其中设计了 8 个标记的语块标记集（包括主语语块、述语语块、宾语语块、兼语语块、状语语块、补语语块、独立语块、语气块）[1]。

2.3 本文对组块的界定

虽然在汉语学习中我们对语句划分的标准经常是主语、谓语、宾语、状语等，然而这种划分属于一种从全局考虑的划分方式，如果没有对语句深入的理解，就不能正确标注出这些成分，这就违背了组块分析的原则。组块分析又称浅层分析，意在不用通过深入的理解就可以得到确定的一个片段。同时从组块的大小来看，组块粒度越大，组块概念的确定性就越强，进一步的分析也就越容易，而组块本身的正确识别却比较困难。因此组块粒度的选取是一个大问题，粒度过小时，组块分析的任务就成了词性标注的问题；粒度过大，则成了完全句法分析问题。这样，选取组块要粒度适当，同时保证组块简单性和概念确定性的均衡问题。因此我们确定建立类似 Abney 组块的汉语组块体系。

从组块分析角度来看，Abney 提出的组块是有级别的，高层次的组块由低层次组块构成。本文对所有组块都一视同仁，使它们都处于一种平等的地位。这里组块的定义借用了 Abney 组块定义的思想，但也存在着差别。为汉语组块定义如下：

定义 1：组块是一种结构，是符合一定句法功能的基本短语。每个组块都有一个核心词，并围绕核心词展开，以核心词作为组块的开始或结束。

这里，如果一个词序列可以构成某种类型的组块，那么它的内部即使有形成其他类型组块的可能性，也不会产生其他类型的组块。组块之间不存在级别问题，即所有组块都位

于同一个层次上，是平等的关系。此外，这里的组块核心词也可以作为组块的开始。同时通过引入*非组块*的组块类型，保证句中任何一个词都属于且只属于一种组块。对语句组块划分遵循以下的原则：

(1) 各种组块类型在构成上都是平等的，任一个组块都严格符合一定的语法规则，且不能由其他类型的组块构成。

(2) 组块之间不发生重叠。句中任一词只能属于一个组块，且组块之间不存在嵌套的现象，在发生歧义时遵守最长匹配原则，能够构成大组块的情况下，屏蔽小组块。

(3) 覆盖原则，我们在划分组块时，要保证句中每一个词语都能够归入一个组块内，对于一些词（如：连词、虚词），不能被归并到其他组块时，则归入到*非组块类型*的组块内。

3. 组块库的获取

组块库的获取是一项繁重和迫切的任务。由于已经存在一定规模的树库，因此利用现有资源完成组块库的构建可以减少部分劳动量。我们选用宾州大学中文树库抽取中文组块。宾州大学中文树库共 4,185 个句子，约 100,000 个词语。语料库中每一句都形成了一个以词语为叶子节点，以整句为根的树状图。

3.1 树库到组块库的转化

组块类型	组块描述
ADJC	形容词组块
ADVC	副词组块
DNC	“的”字组块
DVC	“地”字组块
LCC	方位组块
LST	列举标示组块
NC	名词组块
PC	介词组块
QC	量词组块
VCC	动词组块
NOC	非组块
O	主要用于表示标点符号

表 1：中文组块类型

本文定义了 12 种组块类型，如表 1。这里的组块类型与宾州树库的短语类型有很多对应之处。下面将具体描述一下如何利用中文树库抽取组块。

(1)ADJC：为形容词组块，和宾州树库的形容词短语 ADJP 基本保持一致。不同的是 JJ（形容词）和 VA（形容动词）都可以作为组块核心词。而 ADJP 只以 JJ 作为核心词。例如：

(ADJC (AD 不) (JJ 完全))

(2) ADVC：为副词组块，与宾州大学的副词短语 ADJP 基本保持一致。ADJP 是以副词 AD 作为短语的核心词，在副词组块中除此之外，CS（句子的连接词）或者 VA（形容动词）也可以作为它的核心词，VA 作为核心词时后面经常跟着词语“地”。例如：

(ADVC (VA 方便)) (VA 快捷))

(DVC (DEV 地))

(3) DNC：‘的’字组块是根据宾州树库中‘的’字短语（DNP）抽取出来的，在树库中 DNP 是由任何一个结构（XP）与词性标注为 DEG 的“的”词共同构成的短语，但是由于“的”字在汉语语法中的复杂性，在组块库中通常是“的”字作为单独的一个组块。即：

(DNC (DEG 的))

(4) DVC：‘地’字组块与‘的’字组块的情况相似，它是根据树库中“地”字短语 DVP 进行抽取的，这里也通常把“地”作为一个单独的组块，

(DVC (DEV 地))

(5) LCC：定位组块，由树库的定位短语 LCP 中提取出来，有些直接由 LCP 转换，例如：

(LCC (NT 去年)
(LC 初))

当树库中的 LCP 短语很长时，并由其他短语组成，我们则将其分解成多个组块。例如：

(LCP(VP (VV 改善)
(NP-OBJ (NP-PN (NR 中国))
(NP (NN 出口) (NN 商品) (NN 结构)))))
(LC 中))

从该短语中获取的组块如下：

(VCC (VV 改善))
(NC (NR 中国))
(NC (NN 出口) (NN 商品) (NN 结构))
(LCC (LC 中))

可以看出这里一个 LCP 短语分解出来了四个组块，“中”单独作为一个定位组块。

(6) LST：列举组块，和宾州大学 LST 短语的定义一致，一般是一个词作为一个组块。例如：

(LST (CD 一))
(VP (AD 不) (VV 怕))
(VP (VV 吃亏))
(O (PU ，))
(LST (CD 二))
……

在该句中的“一”、“二”分别组成列举组块。

(7) NC：名词组块，在树库中名词短语 NP 可以递归定义。名词组块实际上是基本的名词短语，内部不含有其他名词组块。因此句中最底层的名词短语为所要抽取的名词组块。例如：

(NP (NP (DT 全) (NN 国))

(NP (NN 劳动) (NN 模范)))

从树库中抽取组块时，我们可以获得两个名词组块，如下：

(NC (DT 全) (NN 国))

(NC (NN 劳动) (NN 模范))

(8) PC 表示介词组块。由于组块不能包含其他组块，因此在形成介词组块时，不需先得到名词组块后才获得。例如，在宾州树库中介词短语如下：

(PP (P 按)
(NP (NN 国家) (NN 政策)))

转化为组块构成，为：

(PC (P 按) (NN 国家) (NN 政策))

(9) QC：为数量组块，从树库数量短语 QP 中抽取出来。宾州树库中存在着一种短语类型—量词短语 CLP，多由一个或两个词构成。我们一般直接将其转化为数量组块。例如：

(QP (CD 三十多)
(CLP (M 项)))

转化为组块构成，为：

(QC (CD 三十多) (M 项))

也有基数词 CD 单独作为数量组块的情况。

(10) VCC：为动词组块。一般由动词短语 VP 转化或抽取得来。因为动词短语在树库中经常包含其他类型的短语，这种情况下，我们把动词短语中和动词相关的词剥离出来，同时在进行校对时根据一些规则抽取一个动词组块，核心词为动词。例如：

(VP (ADVP (AD 及时)) (VV 指定)
(NP (NN 法规性) (NN 文件)))

转化后，为两个组块，一个动词组块和一个名词组块。为：

(VCC (AD 及时) (VV 指定))

(NC (NN 法规性) (NN 文件))

这里，首先从大动词短语中抽取名词组块，然后一个单独的动词“指定”随之被剥离出来，再根据校对规则：副词 AD 和动词 VV 可以构成一个动词组块。如果被抽取的动词短语所含有的其他类型的短语只有一个词，则该词不再单独构成其他组块。例如：

(VP (VV 走)
(NP (NN 亲戚)))

转化为动词组块为：

(VCC (VV 走) (NN 亲戚))

在表 1 中，我们还定义了两个特殊的组块类型：非组块(NOC)用来对于一些经常不能组成组块的规则总结出来，在落单时进行识别和错误纠正。O 专门用来表示标点符号。

3.2 组块库构成

我们根据以上的对应关系从中文树库中抽取组块，获得组块库，共包含 67,734 个组块，各种类型的组块数目统计如表 2。其中名词组块和动词组块所占比例最高，分别为 38.4% 和 17.7%。平均每个组块含有 1.46 个汉语字符（汉字或标点）。如果不计算表示标点符号的

组块 O，平均每个组块含有 1.57 个汉字。

组块类型	ADJC	ADVC	DNC	DVC	LCC	LST
组块数目	875	856	2100	87	1470	9
组块类型	NC	PC	QC	VCC	O	NOC
组块数目	26002	3863	3270	11971	12802	4429

表 2：组块数目统计

4. 结论

信息检索、信息抽取、文本聚类/分类等领域的发展都迫切需要粒度较词语大的确定性成分——组块，这些反过来也促进了组块的研究工作。本文在借鉴其他研究者思路的基础上，对组块进行了定义。只有对组块和组块类型有了明确的定义，才能有效地进行下一步的工作，正确地划分和识别各种类型的组块。同时，我们还提出了利用树库抽取组块库的方法，虽然目前获得的组块语料库规模还比较小，这些语料可以作为种子库，作为构建组块分析器的试验语料，同时为今后大规模的组块获取和收集工作奠定良好的基础。

参考文献

[1] 周强，詹卫东，任海波，构建大规模的汉语语块库，清华大学出版社：自然语言理解与机器翻译，2001，pp102-107.

[2] Nianwen Xue, Fei Xia, The Bracketing Guidelines for the Penn Chinese Treebank(3.0), 2000, http://morph.ldc.upenn.edu/ctb/

[3] 姚天顺等，自然语言理解——一种让机器懂得人类语言的研究，北京：清华大学出版社，1995

[4] Erik F. Tjong Kim Sang and Sabine Buchholz, Introduction to the CoNLL-2000 Shared Task: Chunking. In: Proceedings of CoNLL-2000 and LLL-2000, Lisbon, Portugal, 2000

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

[6] Buchholz S., J. Veenstra and W. Daelemans, Cascaded grammatical relation assignment, In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, College Park, MD, 1999, pp.239--246

[7] Jorn Veenstra, Memory-Based Text Chunking, In: Nikos Fakotakis (ed), Machine learning in human language technology, workshop at ACAI 99, 1999

[8] 赵军，汉语基本名词短语识别及结构分析研究，清华大学工学博士学位论文，1998

[9] 孙宏林，现代汉语非受限文本的实语块分析，北京大学博士学位论文，2001

[10] 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