汉语部分分析研究*

周　强

清华大学计算机系
智能技术与系统国家重点实验室
北京 100084
zhouq@s1000e.cs.tsinghua.edu.cn

摘要：本文概要介绍了近年来我们在汉语部分分析方面的研究工作，包括设计部分分析和标注体系、构建大规模的部分信息标注语料库、探索不同层次的部分分析方法等，并提出了一些应用设想。

关键词：部分分析，语料库标注，词汇知识获取

Research on Chinese Partial Parsing

ZHOU Qiang

State Key Laboratory of Intelligent Technology and Systems
Dept. of Computer Science, Tsinghua.University, Beijing 100084
zhouq@s1000e.cs.tsinghua.edu.cn

ABSTRACT: In this paper, we briefly introduce our current research on Chinese partial parsing, including hierarchical partial parsing schemes, the large scale corpus annotated with different partial parsing tags, and different partial parsing algorithms. We also propose some application tentatives and future research based on the current partial parsing schemes.

KEYWORDS: Partial Parsing, Corpus Annotation, Lexical Knowledge Acquisition

1　引言

从近年来的重要国际会议以及一些最新的研究资料中发现，目前国外对部分句法分析的研究越来越重视，已召开了数次专门讨论部分分析问题的国际研讨会。目前在英语方面比较成熟的部分句法信息描述体系是 Abney(1991)的语块（chunk）描述体系。他把语块定义为句子中一组相邻的属于同一个 s-投射（s-projection）的词语的集合，建立了语块与管辖约束（GB）理论的 X-bar 系统的内在联系，从而奠定了这个语块描述体系比较坚实的理论基础。Tjong & Buchholz (2000)据此开发了一个约 30 万词英语语块库，作为一个统一的英语语块自动分析的训练和测试平台。

近年来，我们对汉语部分分析问题也进行了比较深入的探索。本文将对有关内容进行一次简要综述。其中，第 2 节介绍了不同层次的部分分析描述体系。第 3 节介绍了部分分

*本项研究得到国家自然科学基金（项目号：69903007, 60173008）、国家 973 基金（项目号：G1998030507）、国家高技术研究发展 863 计划（项目号：2001AA114040）资助。

析信息标注语料库的构建工作。第 4 节简要描述了一些部分分析方法探索。第 5 节介绍了部分分析技术在汉语词汇搭配自动获取中的应用。最后的第 6 节进行了简要小结。

2 汉语部分分析描述体系

从汉语的具体语言事实出发，针对不同的应用需求，通过适当的任务分解，我们逐步建立了以下分层次的汉语部分句法信息描述体系：

1） 词界定信息：

主要描述句子中每个词语（w_i）在完整的句法分析树中所处的成分边界位置（b_i）信息，其中 b_i 可取值 0,1,2，分别表示该词语处于成分中间位置、左边界和右边界。它们反映了最基本的层次划分信息[19]。

2） 基本短语：

主要描述句子中相邻的、不嵌套的、内部不包含其他基本短语的、以名词、动词、形容词、数词、量词、副词等为中心词而组成的词语序列。目前我们定义了 9 种基本短语，包括名词短语（np），空间短语（sp），时间短语（tp），数量短语（mp），动词短语（vp），形容词短语（ap），副词短语（dp）等。它们描述了句子的基本组成单元。有关详细内容可参阅[11]。

3） 最长名词短语：

主要描述句子中不被其他任何名词短语所包含的名词短语信息，它们通常处于句子中主语、宾语或介词宾语位置上，用于描述句子中的各个实体概念。有关内容可参阅[20]。

4） 功能语块：

主要描述句子层面上的主语（S）、述语（P）、宾语（O）、状语（D）和补语（C）等功能成分的组合关系。它们形成了句子的基本结构骨架。有关详细内容可参阅[10]。

下面是一个具体的分析标注实例：

- 输入句子为正确的切分和词性标注结果：我/rN 哥哥/n 送/v 给/v 我/rN 一/m 本/qN 很/d 漂亮/a 的/u 书/n 。/w[1]
- 词界定输出：[我/rN 哥哥/n] [送/v 给/v] 我/rN] [一/m 本/qN] [很/d 漂亮/a] 的/u 书/n] 。/w] （对于每个词语，左括号表示它处于成分左边界，右括号表示它处于成分右边界，空标记表示它处于成分中间位置。）
- 基本短语分析：[np 我/rN 哥哥/n] [vp 送/v 给/v] [np 我/rN] [mp 一/m 本/qN] [ap 很/d 漂亮/a] 的/u [np 书/n] 。/w
- 最长名词短语分析：[np 我/rN 哥哥/n] 送/v 给/v [np 我/rN] [np 一/m 本/qN 很/d 漂亮/a 的/u 书/n] （句子中包含三个最长名词短语）
- 功能语块分析： [S 我/rN 哥哥/n] [P 送/v 给/v] [O 我/rN] [O 一/m 本/qN 很/d 漂亮/a 的/u 书/n] 。/w

从信息容量上看，目前定义的几种描述形式，存在以下关系：词界定 < 基本短语 < 最长名词短语 <= 功能语块，显示出较好的层次组织结构。其中，基本短语和功能语块层次最为重要。在基本短语层面上，对句子的各个基本信息单元，包括：实体、时间、空间、数量、性状、动作等进行了细致描述，体现了一种自底向上的句法语义组合过程。其描述信息大体上与 Abney 定义的 chunk 层次相当；在功能语块层面上，对句子的主、谓、宾、

[1] 有关的词类标记简要说明如下：rN—名代词，n—名词，v—动词，m—数词，qN—名量词，d—副词， a—形容词， u—助词， w—标点符号。

状、补等各个功能成分进行了深入描述，体现了一种自顶向下的句子整体功能成分分布的描述过程。两者可以形成很好的信息互补。以此为基础，可以将汉语分析问题很自然地分解成以下三个阶段：

1) “字→词→基本短语”的分析

主要研究汉语词语的句法语义组合模板的发现、提取和应用问题。

2) “基本短语→最长名词短语→功能语块”的分析

主要研究虚词和语序在汉语复杂结构建构中的重要作用。

3) “功能语块→小句→句子→段落”的分析

主要研究汉语句子内部的事件结构分析和句子之间的事件逻辑关系分析等问题。

它们将共同形成汉语“字→词→块→句”的一体化分析框架中的几个主要环节。

3 汉语部分信息标注语料库

以上面的部分句法信息描述体系为基础，我们对大规模的汉语真实文本进行了不同层次的信息标注，形成了高质量的汉语功能语块库、基本短语库和词界定库，为进一步进行各种部分分析方法探索和知识获取研究打下了很好的基础。其中：

功能语块库的标注规模为 200 万汉字。基础语料为清华大学开发的 200 万汉字规模的平衡语料库，主要选自 90 年代的现代汉语书面语以及准口语（包括剧本、谈话录、演讲录等）的真实文本，按文体分为文学、新闻、学术、应用四类。通过人工标注和机器检查，为其中的每个句子都标注了正确的功能语块信息[18]。

基本短语库的标注规模为 50 万汉字左右，基础语料选自清华大学开发的 200 万汉字规模的平衡语料库。通过自动标注和人工校对相结合方法完成。

词界定库的标注规模为 42 万汉字左右[15]，基础语料选自两个小规模的汉语测试树库和人民日报文本，通过树库信息自动转换和人工标注得到。

目前开发完成的 200 万汉字的功能语块库，无论在标注规模和信息容量上，都处于国际领先水平。在笔者主持的分阶段汉语句法树库构建项目中，功能语块库的使用价值也得到了充分体现[13]。

4 汉语部分分析方法探索

4.1 词语边界预测

对于词界定问题，我们主要采用了基于局部语境模板（LCT）的自动识别策略：利用从大规模标注语料库中自动训练得到的各个词类标记在局部语境下（主要依据了词类信息）的不同界定标记的频度分布数据（$BPFL_i$），选择确定合适的界定标记。目前，主要提取了以下几类 LCTs:

1) Unigram LCTs: t_i, $BPFL_i$
2) Bigram LCTs:
   - 左限制: $t_{i-1}\ t_i$, $BPFL_i$
   - 右限制: $t_i\ t_{i+1}$, $BPFL_i$
3) Trigram LCTs: $t_{i-1}\ t_i\ t_{i+1}$, $BPFL_i$

具体的识别算法则利用了 Backing-Off 控制机制，从 trigram → bigram → unigram→ 缺省设置（b_i=0），在各个不同的 LCTs 中，选择分布最显著的界定标记。

以约 26 万字的树库标注语料为训练集，对从人民日报文本中抽取的约 16 万字语料进行开放测试，词界定正确率为 91%左右，取得了较好的实验效果。有关识别算法的详细内容，可参阅论文[15]。

4.2 基本短语识别

对于基本短语识别问题，我们采用了以下两种不同的分析技术：

1） 基于有限状态自动机（DFA）的分析：

通过对标注语料库的统计分析，提取各个基本短语的常见结构组合模式。例如，对基本名词短语，有“ a | b | m | n | s | t | z | rN + {n | vN }* + n | vN ”，“ n + {n}+ ”等组合模式。据此构造不同的有限状态机，充分利用局部语境下的词类和部分词语限制信息，识别句子中所有可能的基本短语组合。

2） 基于实例学习（MBL）的分析技术：

首先利用 MBL(Memory-Based Learning)方法预测语料中的每个词所属基本短语的种类以及在基本短语中的位置，然后利用各种不同类型基本短语的内部构成信息进行左右边界配对，对有歧义的短语左右边界进行选择，对边界进行删除和添加。最终得到完整的基本短语识别信息。利用从基本短语库中抽取出的约 30 万字语料进行 Held-out（9∶1）自动识别实验，开放测试结果显示：常见基本短语识别的 F-measure 达到了 92%左右，其中，基本名词短语约为 93%，基本动词短语约为 94%，取得了较好的识别效果[21]。

在英语方面，参加 CoNLL2000 语块自动识别测试的共有 10 个系统，分别采用了有限状态分析、最大熵模型、基于实例学习、支持向量机（SVM）等分析方法，其中采用多种模型的集成分析系统识别效果最好，F-measure 达到 93%左右[9]。

4.3 功能语块识别

对于功能语块识别问题，利用功能语块标注的穷尽性（任何一个词都必须无遗漏地进入某个语块）和线形（句子中的全部语块形成一个线性序列，即没有嵌套）特点，可以简化为识别句子中的每个词语是否处于某个功能语块的左边界问题。

我们主要探索了利用比较常用的机器学习算法：C4.5 判定树模型来构建不同的统计识别模型。在基本模型中，选定局部语境信息（每个词语左右各两个词语的词类信息）和前一个识别出来的功能语块标记信息作为特征向量。在扩展模型中，进一步使用了从北大语法信息词典中提取出来的大量语法特征信息。由于词典中的特征信息描述非常丰富，因此又化了很大精力，通过人机互助的方法进行了特征选择处理，把从左右各一个词语的词典信息描述中选出的最有效的语法特征加入到判定树模型中。利用从功能语块库中抽取出的约 14 万词语料进行 Held-out（9∶1）自动识别实验，开放测试结果为：基本模型的最佳 F-measure 为 76%左右，扩展模型的最佳 F-measure 为 79%左右。有关的详细内容，可参阅（[3], [4]）。

对实验结果的分析发现，出错较多的是一些复杂的长语块，识别程序往往倾向于把一个长语块识别为多个小语块。根据我们对 200 万字语块库的统计分析，主语块和宾语块的平均长度分别为 2.53 和 4.13 个词,其中词语长度超过 5 的语块分别占 15%和 30%左右[18]。在这种情况下，目前识别算法所依据的局部语境信息（左右个 2 个词）明显不够用，边界识别错误也就在所难免了。后来，我们还尝试利用最大熵模型来进行功能语块识别[5]，处理效果与判定树模型差不多。这些基于机器学习方法的功能语块自动识别实验，也从另一个侧面反映了内部结构分析对复杂语块识别的重要性，促使我们开始探索进行基于基本短

语分析的功能语块识别的可行性。

5 词汇搭配知识自动获取

词汇搭配是描述词语间组合能力的一种重要的词汇知识，它具有句法性、任意性、可重现性等特点。传统的词汇搭配研究强调区分约束组合和自由组合的重要性。但我们在研究中发现，一些常见的高频自由组合实例可以在汉语句法结构排歧中发挥重要作用。同时，它们也是进行更深入的词汇语义关系抽取的重要素材。因此，我们把词汇搭配获取任务进行了适当扩展，通过设置不同的搭配强度阈值，从大规模的真实文本语料中自动获取约束组合和大量有用的自由组合实例。

在汉语中，针对动词的词汇搭配模式主要有“ 动+名 ”、“ 名+动 ”、“ 动+形 ”，“ 形+动 ”等。对此，我们主要采用了以下获取策略：1）对获取文本进行部分句法分析，尽可能多地引进句法限制信息；2）总结适当的启发式规则，控制选取所有可能的搭配组合实例；3）设计合理的搭配强度计算公式并选择合适的统计阈值，提取所有合理的搭配组合，排除可能的统计噪声。目前主要进行了以下两方面的探索：

乔中元(2000)主要采用了词界定和组块分析[19]预处理结果，控制选取满足下列条件的搭配组合：在观察窗口[-4,4]中的词语组合“ [v ... n] ”,“ [n … v] ”,“ [v … a] ”,“ n] …v] ”等等。对搭配强度的估计，则主要采用了 Smadja(1993)定义的三个基本统计量：强度、离散度和尖峰。实验结果表明，通过引入组块分析信息，与基于切分和词性标注信息的类似获取实验（[7],[8]）相比，较大地提高了词汇搭配获取的正确率和处理效率。

党政法(2002)则主要采用了基本短语分析预处理结果，利用以下启发式规则选取可能的搭配实例：

1） 基本动词短语内部结构分析：提取可能的“ v- p|a|v ”和“ a-v ”搭配实例
2） 基本名词短语内部结构分析：提取“ n-v ”和“ v-n ”搭配组合 (动词标注为 vN)
3） 基本短语间的搭配分析：考虑每个基本动词短语左相邻的 2 个和右相邻的 3 个基本名词短语，分别提取它们的中心词组合“ n-v ”和“ v-n ”作为可能的搭配实例。

搭配强度估计则采用了以下统计量： χ^2 分布，搭配频度，离散度分布。利用这套算法对 6 个月的人民日报切分和词性标注语料库进行了词汇搭配自动获取实验，共提取出 52 万多个搭配对，从中选取了与动词“ 发展 ”有关的 2101 个搭配对进行人工检查，计算得到搭配获取正确率约为 80%，取得了较好的处理效果。目前正在利用大规模的树库标注信息对有关算法进行改进和完善，争取达到更好的自动提取效果。

6 结语

本文概要介绍了近年来我们在汉语部分分析方面进行的研究工作，包括设计分层次的汉语部分分析描述体系、构建大规模的汉语部分信息标注语料库、探索不同层次的汉语部分分析方法等。并通过目前进行的几个汉语词汇搭配自动获取实验，初步证明了部分分析技术在语言知识自动获取方面的重要应用价值。

另外，从目前的部分分析体系和标注语料库出发，我们还进行了以下研究探索：

1）充分利用功能语块标注信息，对现有的汉语概率分析器[14]进行功能模块重组，将句子分析流程简化为以下三大步：A) 语块内部成分分析；B) 语块间的结构关系分析；C) 多个语块组成的子句逻辑语义关系分析，形成一个新的基于语块的句法分析器，初步实现了

“功能语块→小句→句子”的分析过程。它在分阶段构建大规模汉语树库[13]的语言工程实践中，发挥了重要作用。

2）通过不同语言资源的知识融合实验，初步形成比较完整的汉语动词语法搭配模板描述知识库。并以此为基础，辅之以可以从大规模真实文本中自动获取的汉语概率型上下文无关语法[16]和结构优先关系[17]知识，开发完成一个高效灵活的汉语部分分析器，初步实现了“基本短语→最长名词短语→功能语块”的分析过程。

3）充分利用有限状态分析技术，将现有的比较成熟的汉语自动切词、词性标注和基本短语分析技术进行有机融合，开发一体化的汉语“字→词→基本短语”的自动分析器。

以上研究将为我们初步设想的构建汉语“字→词→块→句”的句法语义一体化分析理解模型的长期研究目标打下坚实的基础。

参考文献

[1] Steven Abney(1991). “Parsing by Chunks”, In *Robert Berwick, Steven Abney and Carol Tenny (eds.) Principle-Based Parsing, Kluwer Academic Publishers*.

[2] 党政法（2002），“基于基本短语分析的汉语词汇搭配的自动获取”，清华大学计算机系本科毕业设计论文，2002 年 6 月.

[3] Elliott France Drabek and Qiang Zhou (2001). “Experiments in Learning Models for Functional Chunking of Chinese Text”, In *Proceedings of IEEE International Workshop on Natural Language processing and Knowledge Engineering (NLPKE’ 2001), Tucson, USA*, 895-864.

[4] Elliott France Drabek and Qiang Zhou (2001). “Use of a Lexical Feature Database for Partial Parsing of Chinese”, In *Proceedings of 6th Natural Language Processing Pacific Rim Symposium (NLPRS’ 2001), Tokyo, Japan*, 663-668.

[5] 刘畅（2001），“基于最大熵模型的汉语语块自动分析”，清华大学计算机系本科毕业设计论文，2001 年 6 月.

[6] 乔中元（2000），“基于组块分析的汉语词语搭配信息的自动获取”，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，2000 年 6 月

[7] 孙茂松，黄昌宁，方捷（1997）“汉语搭配定量分析初探”，《中国语言》，256(1)，29-38

[8] 孙宏林（1998）“词语搭配在文本中的分布特征”，*Proceedings of 1998 International conference on Chinese Information Processing, Beijing, China*, 230-236.

[9] Erik F. Tjong Kim Sang and Sabine Buchholz. (2000). “Introduction to CoNLL-200 Shared Task: Chunking”. *Proceedings of CoNLL-2000 and LLL-2000*. Lisbon, Portugal. 127-132.

[10] “汉语句子的功能语块标注规范”，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，2000 年 6 月。

[11] “汉语基本短语标注规范”，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。

[12] Smadja, Frank (1993) “Retrieving Collocations from Text: Xtract”, Computational Linguistics, 19(1), 143-177.

[13] 周强,任海波,孙茂松(2002)“分阶段构建汉语树库”, *Proc. of The Second China-Japan Natural Language Processing Joint Research Promotion Conference*, 189-197.

[14] Qiang Zhou. (1997) “A Statistics-Based Chinese Parser”, In *Proc. of the Fifth Workshop on Very Large Corpora*, 4-15.

[15] Qiang Zhou (2000). “Local context templates for Chinese constituent boundary prediction”, In *Proceedings of The 18th International Conference on Computational Linguistics (COLING’00), Germany*. 981-987

[16] 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

[17] 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

[18] 周强，任海波，詹卫东 (2001).“构建大规模汉语语块库”，黄昌宁，张普主编《自然语言理解与机器翻译》，清华大学出版社，102-107.

[19] 周强，孙茂松, 黄昌宁 (1999).“汉语句子的组块分析体系”，《计算机学报》，22(11)，1158-1165.

[20] 周强，孙茂松, 黄昌宁 (2000).“汉语最长名词短语的自动识别”，《软件学报》 11(2), 195-201.

[21] 张昱琪，周强 (2002).“汉语基本短语的自动识别”，《中文信息学报》，16(6)，1-8