文章编号：1003-0077(2013)03-0001-08

汉语组块分析研究综述

李业刚[1,2]，黄河燕[1]

(1. 北京理工大学 计算机学院，北京 100081；
2. 山东理工大学 计算机科学与技术学院，山东 淄博 255049)

摘 要：组块分析作为浅层句法分析的代表，既可以满足很多语言信息处理系统对于句法功能的需求，又可以作为子任务，在词法分析和完全句法分析以及语义分析中间架起一座桥梁，为句子进行进一步深入分析提供有力的支持，因此众多的研究将注意力集中于组块分析上。该文主要对组块的定义和分类、组块识别方法、组块的标注和评测以及组块内部关系分析等几方面的研究进展进行详细的综述。最后，探讨了组块分析存在的问题并对未来的发展方向进行了展望。

关键词：中文信息处理；浅层句法分析；组块分析；组块识别

中图分类号：TP391　　**文献标识码**：A

A Survey on Chinese Chunk Parsing

LI Yegang[1,2], HUANG Heyan[1]

(1. School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China;
2. Department of Computer Science and Technology, Shandong University of Technology, Zibo, Shandong 255049, China)

Abstract: Chunking, as a typical shallow parsing, serves for many language information processing system for their demands on syntactic information, as well as a bridge between the lexical analysis, syntactic parsing and semantic parsing. This paper surveys the rich researches on chunking in several aspects: the definition and classification of chunks, the chunks identification, the chunks annotation and evaluation, and the internal relationship in chunks. Finally, this paper draws conclusions and discusses the future work.

Key words: Chinese information processing; shallow parsing; chunk parsing; chunk identification

1 引言

句法分析是自然语言处理中的重点和难点，虽然经过几十年的研究和发展，仍是自然语言处理的一个瓶颈问题。采用“分而治之”的方法，进行浅层的句法分析可以降低完全句法分析的难度。组块分析作为浅层句法分析的代表致力于识别句子中的某些结构相对简单、功能和意义相对重要的成分，只限于把句子解析成较小的单元，而不揭示这些单元之间的句法关系。

继 Abney[1] 率先提出了组块分析的思想后，国际会议 CoNLL-2000 把组块分析作为共享任务[2]提出，组块分析逐步受到重视。人们对于基本名词短语、介词短语，继而扩大到所有类型短语的识别等问题做了大量的研究。目前的组块分析技术由于受到相关语言处理研究及信息处理技术应用两个方面的驱动，已成为自然语言领域中一个重要研究内容而受到广泛关注。

2 组块分析的任务

Abney[1] 将句法分析问题分为三个阶段：

(1) 块识别：利用基于有限状态分析机制的块

收稿日期：2011-12-24　**定稿日期**：2012-10-11

基金项目：国家自然科学基金资助项目(61132009,61201352)

作者简介：李业刚(1975—)，男，博士研究生，副教授，主要研究方向为自然语言处理；黄河燕(1963—)，女，博士生导师，研究员，主要研究方向为自然语言处理与机器翻译。

识别器识别出句子中所有的块。

(2) 块内结构分析：对每个块内部的成分赋予合适的句法结构。

(3) 块间关系分析：利用块连接器将各个不同的块组合成完整的句法结构树。

浅层句法分析的结果并不是一棵完整的句法树，各个组块是完整句法树的一个子图，只要加上组块之间的依附关系，就可以构成完整的句法树，对语块的识别是组块分析的主要任务[3]。

3 组块的定义和类型

Abney[1]最早提出了一个完整的组块描述体系，他把组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，建立了组块与管辖约束理论的X-bar系统的内在联系，从而奠定了语块描述体系的比较坚实的理论基础。在自然语言学国际会议(CoNLL-2000)的共享任务组块分析中，在Abney描述的组块定义框架的基础上，重新分解和细化了组块的定义，对英文组块的定义达成了共识：句子是由一些短语构成，而每一个短语内是由句法相关的词构成，这些短语彼此不重叠、无交集，不含嵌套关系。

3.1 中文组块的定义

受限于中文句法分析的定义问题，最初中文组块并不是覆盖整个句子的，而是侧重对基本名词短语、介词短语以及短语自动界定的研究。文献[4-5]首次提出了中文的基本名词短语的形式化定义，提出了用词语潜在依存关系分析baseNP结构的模型，将依存语法知识融入概率模型中，使得baseNP结构分析在依存语法知识的指导下进行，开放测试精确率82%，召回率91.5%。文献[6]设计了一种基于转换的基本名词短语识别模型，该模型可同时结合表示基本名词短语句法组成的基本结构模板(静态知识)与表示基本名词短语出现的上下文环境特征的转换规则(动态知识)识别基本名词短语，开放测试精确率89.3%，召回率92.8%。文献[7]使用了基于最大熵的方法识别中文基本名词短语。在开放语料Chinese TreeBank上，只使用词性标注，达到了88.09%的准确率。文献[8]对汉语中最常用的介词“在”进行了实验，开放测试的准确率93%。

但是除名词组块和介词组块外，中文句子中还有很多其他结构的组块，如动词组块，形容词组块等。汉语的句法体系至今还没有一个像英文那样统一的完全公开的训练语料库[9]为各种汉语组块分析方法提供统一的评测平台。从公开的研究成果可以看出，研究者们根据自己的研究目的提出了各自不同的块描述体系。

文献[10]在Abney定义的基础上，对汉语组块定义为：组块是一种语法结构，是符合一定语法功能的非递归短语，每个组块都有一个中心词，并围绕该中心词展开，以中心词作为组块的开始或结束。任何一种类型的组块内部不包含其他类型的组块。并提出了非递归、不重叠、覆盖三个组块划分原则。文献[11]与CoNLL-2000类似，将基本短语定义为句子中相邻的、不嵌套的(允许在黏合式定中结构中出现一级嵌套)、内部不包含其他基本短语、主要由实词(名词、动词、形容词、数词、量词、副词等)组成的词语序列。文献[12]提出了基于拓扑结构的基本块描述体系，通过引入词汇关联信息确定基本拓扑结构，形成了基本块内聚性判定准则，确定不同基本块的内部关系标记，将紧密结合的述宾结构关系纳入基本块描述体系中。文献[13]通过引入词汇关联信息确定基本拓扑结构，形成了很好的基本块内聚性判定准则，建立了句法形式与语义内容的有机联系桥梁。这套描述体系大大简化了从现有的句法树库TCT中自动提取基本块标注语料库和相关词汇关联知识库的处理过程，为进一步进行汉语基本块自动分析和词汇关联知识获取研究打下了很好的基础。文献[14]定义组块是一种具有一定句法功能的非递归、不重叠、不嵌套的短语。包含一个中心成分以及中心成分的前置修饰成分，而不包含后置附属结构。它对组块的基本划分原则为：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；组块是严格按照句法定义的，不能破坏句子的句法结构，也不体现句子的语义和功能；组块的划分只依据局部的表层信息，例如词信息、词性信息等，而不能考虑远距离约束以及句子的整体句法结构。

Abney定义的组块强调对局部的句法进行相关描述，侧重于从底向上把句子分割成不同的组块，文献[10,14]与Abney的定义类似；清华大学的组块体系[11-13]强调对句子整体功能的描述，侧重于自顶向下地描述句子的基本骨架。CoNLL的组块一般比较简单，平均每个块只包含1～2个词语，而清华大学的组块比较复杂，有的组块甚至包含10～20

个词语。组块粒度越大，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。

3.2 中文组块的类型

文献[11]根据宾州大学中文句法分析树库的语料和句法标记类型，并结合汉语特点从中抽取出了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了转化得到组块库。其定义的组块长度较短，平均每个组块只含有1.57个汉字。文献[15]针对机器翻译提出了扩展组块(E-Chunk)的概念及其体系。更多研究者根据自己的研究目的提出了各自不同的组块类型[16-23]。从组块包含词的个数来看，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。组块粒度过大，组块分析任务就成了完全句法分析问题；而粒度过小，则成了词性标注的问题。因此组块粒度的选取是一个重要问题，要同时保证组块简单性和概念确定性。另外，中国香港理工大学计算机系的陆勤教授[24]和中国台湾"中央研究院"的许闻廉教授[25]在中文简体和繁体组块分析语料库的建设方面做出了卓有成效的工作。

4 组块分析结果的评测

通常用正确率(P)，召回率(R)和F值作为组块分析结果的评测指标。对于某种类型的组块，其正确率、召回率和F值分别为：

$$P=\frac{\text{正确标注类型 X 的组块个数}}{\text{标注类型 X 的组块个数}}\times 100\%$$

$$R=\frac{\text{正确标注类型 X 的组块个数}}{\text{类型 X 的组块总数}}\times 100\%$$

$$F=\frac{2PR}{(P+R)}\times 100\%$$

对于所有类型的组块，识别的正确率和召回率分别为：

$$P=\frac{\text{正确标注的组块个数}}{\text{标注的组块个数}}\times 100\%$$

$$R=\frac{\text{正确标注的组块个数}}{\text{组块总数}}\times 100\%$$

F值的计算方法跟单一类型相同。

对于所有类型的组块，在计算正确标注组块的个数时，不仅要考虑组块的前后界划分要正确，而且组块的类型标注也要正确。如果被识别出来的组块，其类型标记错误，那么这个组块也不是被正确标注的组块。比如把数量词组块标注成名词组块，即使是组块的边界划分正确，也不是被正确标注的组块。

继CoNLL-2000设计了英文组块分析共享分析任务，文献[26]针对汉语的描述特点，提出了三项汉语组块分析评测任务：基本组块分析、功能组块分析和事件描述小句识别。

5 组块的标注形式

组块的标注形式主要包括两类：第一类是Inside/Outside表示方法；第二类是Start/End表示方法。Inside/Outside的表示方法首先由Ramshaw和Marcus[27]提出，采用了组块标记集合{I,O,B}，在识别多种类型的组块时，组块标记的含义为：B-X表示X类型组块的开始并且其前面的词属于另一个组块；I-X表示X类型组块的内部，可以是组块的开始；O表示不属于任何组块。文献[28-29]把上述表示方法称为IOB1，并在此基础上提出了IOB2，IOE1和IOE2表示方法。在IOB2中，B-X表示X类型组块的开始；I-X表示X类型组块的内部，但不是组块的开始；O表示不属于任何组块。在IOE1中，E-X表示X类型组块的结尾，并且其后面的词属于另一个组块；I-X表示X类型组块的内部，可以是组块的结尾；O表示不属于任何组块. 在IOE2中，E-X表示X类型组块的结尾；I-X表示X类型组块的内部，但不是组块的结尾；O表示不属于任何组块。

Start/End表示方法是曾用于日语实体名词识别的IOBES方法[30]。B-X表示X类型组块的开始，该组块至少包含两个词；E-X表示X类型组块的结尾，该组块至少包含两个词；I-X表示X类型组块的内部，该组块至少包含三个词；O表示不属于任何组块；S-X表示该X类型的组块由一个词组成。

6 组块识别

利用机器学习方法来解决组块识别问题主要有两种基本思路：基于统计的方法和基于规则的方法，当然也可以采用规则和统计相结合的方法。

英文的组块分析已经建立了统一的标准和数据集，很多学者尝试了大量的机器学习算法[31-40]来解决组块分析问题。Church[31]将英语的基本名词短语定义为简单非嵌套名词短语，并将文本中的基本

名词短语识别问题看作是给每个词加标记的过程，利用基于词性标记的N元同现的概率统计方法和Viterbi方法来解决。文献[32]在Church的研究基础上，采用了基于转换的错误驱动学习方法来解决基本名词短语识别问题，并得到了召回率88%的实验结果，这也是机器学习方法首次被应用到短语识别问题中。文献[33]提出了基于word-only思想的组块分析模型。模型只利用了词特征和词缀特征，对CoNLL-2000英文组块分析训练语料库的规模进行扩充，在训练语料库达到50 000万句的情况下，性能曲线超过了利用词和词形特征的模型的性能曲线。文献[34]应用了Winnow的方法，并引入了训练语料之外的英文槽语法来解决组块分析问题，取得了94.17%的分析性能。文献[35]应用了基于存储的机器学习方法，结合手写规则的方式解决组块分析问题，在韩语组块分析语料库上取得了94.21%的性能。文献[36]应用了半指导学习的方法解决组块分析问题，一方面使用了人工标注好的CoNLL-2000数据，另一方面使用了大量的未标注数据来训练分析模型，取得了94.39%的分析性能。文献[40]采用了多个支持向量机模型融合，结合动态规划技术的机制进行组块分析，取得了2000年的CoNLL-2000会议评测中最佳的分析性能93.48%。汉语组块识别借鉴英语组块识别的方法也有大量的尝试。由于中文和英文在书写方法上存在着根本的不同，中文词与词之间没有显式的分隔标记，词的定义也比较模糊。在组块分析之前的语言处理任务除了跟英文相同的词性标注和未登录词识别外还有分词。这也就意味着中文组块识别的难度比英文要更大一些。

6.1 基于规则的方法

规则方法就是根据人工书写的或(半)自动获取的语法规则标注出短语的边界和短语的类型。在基于规则的方法中，主要的困难在于语法规则的获取以及语法规则之间的优先顺序排列。现在一般都采用机器学习的方法来自动获取规则。

Abney提出组块的概念后，针对英语，在文献[41]中提出把句法分析的过程分成很多个层次，每个层次都只输出一个结果，而在每个层次内部只使用简单的有限状态自动机进行分析。汉语方面，文献[42]在词汇知识库支持下，从标注语料库中自动获取所有基于词类的基本块规则，通过设置规则置信度自动排除大量低可靠和无效规则。针对其中的高频低可靠规则，不断引入更多的内部词汇约束和外部语境限制知识，使之逐步进化为描述能力更强的结构化规则。数量组块、时间组块和形容组词块的F值达到了93%左右，多词语动词组块、名词组块和空间组块的F值分别为87%、84%和83%。

6.2 基于统计的方法

机器学习方法可以分为有指导学习方法、无指导学习方法和半指导学习方法。有指导方法难点在于构造一个大规模的标注语料库是要花费大量的人力物力的，而无指导的缺点则在于一般的迭代算法的复杂度都很高，运算效率较差，并且不能很好地保证最终训练结果的语法可靠性。

6.2.1 有指导学习方法

有指导学习方法是通过学习已知数据的特征以及对应的结果度量，建立起预测模型来预测并度量未知数据的特征和结果。虽然无指导和半指导的学习方法取得了一定的成果，但是大规模语料库支撑下的有指导学习仍旧是中文语言处理的主流方法。

文献[43]采用了一种基于增益的隐马尔可夫模型的方法来进行汉语组块的研究。在哈尔滨工业大学树库语料测试的F值为82.38%。文献[44]将中文组块识别问题看成分类问题，并利用SVM加以解决，在哈尔滨工业大学树库语料测试的F值是88.67%。文献[45]在SVMs模型的基础上，提出基于大间隔方法的汉语组块分析方法，给出判别式的序列化标注函数的优化目标，并应用割平面算法实现对特征参数的近似优化训练。通过在宾州中文树库CTB4数据集上的实验数据显示，各种类型组块识别的总的F值为91.61%。文献[46-47]将条件随机域模型应用到中文组块分析中，其中文献[46]利用语义词典抽取语义类特征，将其加入分析模型，得到92.77%的F值。文献[22]将有向图语言模型应用于汉语组块分析，将候选组块标记映射为有向图节点，根据候选组块标记之间的接续关系确定节点之间是否存在有向边。利用词、词性和组块标记的统计信息为有向边赋值。组块分析的F值为84.99%。文献[50]提出了一种基于CRFs的分布式策略及错误驱动的方法识别汉语组块，首先将11种类型的汉语组块进行分组，结合CRFs构建不同的组块识别模型来识别组块；之后利用基于CRFs的错误驱动技术自动对分组组块进行二次识别；最后依据各分组F值大小顺序处理类型冲突。系统开放式测试的F值达到92.91%。

文献[21]提出基于 Stacking 算法的多分类器组合方法，通过构造一个两层的叠加式框架结构，将4种分类器(fnTBL、SNoW、SVM、MBL)进行了组合，并融合各种可能的上下文信息作为各层分类器的输入特征向量，组合后的分类器在哈尔滨工业大学树库语料的测试中 F 值达到 93.64。文献[51]给出了双规则(DR-AdaBoost)分类算法。算法在每次迭代中将双规则(最优弱分类规则和次优弱分类规则)的线性组合作为迭代的评价标准，应用在汉语组块分析中 F 值为 89.92%。

6.2.2 半指导和无指导学习方法

半指导学习是使用大量的未标注数据和一部分标注的数据来构建分类器或者模型，对未标注的数据进行标注和判断。无指导学习是利用从总体给出的样本信息来做出推断和描述数据的组织和聚类。

文献[52-53]提出了一种基于信息熵的层次词聚类算法，并将该算法产生的词簇作为特征应用到中文组块分析模型中。利用中文组块语料库中的词及其组块标记作为基本信息，采用二元层次聚类的方法形成具有一定句法功能的词簇。用词簇特征代替传统的词性特征应用到组块分析模型中，并引入命名实体和仿词识别模块，F 值为 82.69%。文献[54]采用 co-training 实现中文组块识别。选取增益的隐马尔可夫模型和基于转换规则的分类器(fnTBL)组合成一个分类体系，对 co-training 算法中两种不同的策略进行了比较，一种是选择缓存器中的所有实例的方法，一种是保证两个分类器在未带标数据的一致性方法，在小规模标注的汉语树库语料和大规模未标注汉语语料上进行中文组块识别，F 值分别达到了 85.34%和 83.41%。

6.2.3 混合学习方法

使用有指导的统计方法和无指导的聚类方法结合，可以提高无指导聚类的准确率，避免有指导方法因汉语组块语料库规模较小而导致的数据稀疏现象。文献[55]提出了改进 K-均值聚类方法。分为3个过程：首先根据从语料库中统计的数据，采用基于中心词扩展的策略把句子中的单词先分到不同的类中；然后运用聚类算法调整中心，进行聚类；最后根据单词在句子中的位置确定短语的边界。应用改进 K-均值聚类方法对7种汉语组块进行识别，F 值达到了 92.94%。优于基于中心词扩展的方法 89.90%，也优于 K-均值聚类算法 87.12%。

6.3 统计和规则结合的识别方法

规则和统计相结合的方法出发点是充分发挥基于统计方法和基于规则方法各自的优势，为组块分析寻找一种较好的处理方法。

文献[10]由语言学知识得到初步的组块划分语料，通过校正和学习不断对规则进行调整，完善规则模型，并在不断增大的标注语料基础上对统计模型进行训练，得到组块划分的统计模型。从封闭测试和开放测试的试验结果来看，两种方法结合进行标注的正确率分别达到了 96.2%和 94.6%。文献[48]采用基于实例的学习方法，对汉语基本短语的边界及类别进行识别，并利用短语内部构成结构和词汇信息对预测中出现的边界歧义和短语类型歧义进行了排歧处理。实验结果中对基本组块的识别正确率达到 95.2%，召回率达到 93.7%。文献[49]给出了一种错误驱动学习机制与 SVM 相结合的汉语组块识别方法。该方法在 SVM 组块识别的基础上，对 SVM 识别结果中的错误词语序列的词性、组块标注信息等进行分析，获得候选校正规则集；之后按照阈值条件对候选集进行筛选，得到最终的校正规则集；最后应用该规则集对 SVM 的组块识别结果进行校正。实验结果表明，与单独采用 SVM 模型的组块识别相比，加入错误驱动学习方法后，精确率、召回率、F 值都有了不同程度的提高。文献[56]实现了一种针对并行语料库进行双语组块自动识别的方法。首先根据规则库，分别对源语言句子和目标语言句子中所有符合规则的子块进行标记，然后利用统计模型，对所有可能的源语子块在可能的目标语子块集合中搜索其最佳的对应，最终形成双语句对的可能的双语组块划分。在一个6万句的旅馆预定领域口语语料库中的实验中，正确率可达到80%左右。

7 组块内部结构分析

相比于组块识别，对于中文的组块内部结构研究还比较少。文献[13]的汉语基本块标注体系中，提出了基本块的关系标记描述集，包括了右角中心结构、链式关联结构、并列关系 CHC、述宾关系 LCC、述补关系 LCC、附加关系 LCC 和单词语基本块。基于基本块标注体系，文献[57]设计了一套关系标记集。其设计思路是针对4种关系：修饰关系(ZX,LN)、述宾关系(PO)、述补关系(SB)和并列关系(LH)，对句子中的每个词所处的功能位置进行描述，如表1所示。利用条件随机场模型对句子中的每个词进行序列关系标注，然后通过有限自动机

规则自动获取句子的基本块标注结果。其句法标记识别性能与使用经典的边界标记(IOB)相比略有下降。文献[57]进一步提炼出了三种典型的拓扑结构：左角中心结构(LCC)、右角中心结构(RCC)和链式关联结构(CHC)，它们覆盖了基本块内部修饰关系、并列关系、述宾、述补和附加关系。

表1 关系标记集

标记	功能说明	标记	功能说明
M	修饰关系中心词左侧的词	I	述补关系中的功能词如“了”
R	修饰关系的中心词	J	并列关系
P	述宾和述补关系中的谓语	H	该位置为整个结构的句法语义中心词
O	述宾关系中的宾语	B	单词
C	述补关系中的补语	X	块外其他成分

8 结论和展望

英语方面已有在组块分析基础上进行完全句法分析的研究。其中文献[58]把句法分析分解为一系列的组块识别任务，并用 CRFs 模型实现。虽然正确率略低，但是时间和空间复杂度却低了很多，在对实时性要求较高的系统中有很好的应用前景。相比之下中文组块分析技术，由于缺乏一个明确、公开的定义方法和训练语料库，在语料库建设角度上还有很多工作没有进行。清华大学在整理和加工中文组块库方面做了大量工作，同时建立了一个完整的组块划分体系：基本组块、功能组块、事件描述小句识别。从其已经公开发布的成果来看，除了基本组块外，对功能组块也有一些较为成熟的研究[59-60]，但是第三层次事件描述小句识别的研究则很少见，距离完全句法分析尚有一定的距离。文献[61]实现了基于组块的日英统计机器翻译模型，这也对中文组块的应用提出了一种可尝试的研究方向，不以完全句法分析为目标，用组块代替词或者短语实现具体的应用，当然因此带来的数据稀疏问题也是必须要面对的。为了更好地研究和解决组块分析及其应用，笔者认为还应该在以下几个方面进行进一步的研究和探索。

(1) 由于目前的中文组块分析定义还没有一个统一的标准，一方面，对组块定义的统一和规范的制定，是研究者们共同的发展方向；另一方面，对于利用剪枝从句法树库中抽取组块的定义方式，如何根据应用领域的不同，实现可定制的剪枝和抽取策略，自动地构建符合需求的组块语料库，也是一项有意义的研究。

(2) 对于组块识别，一方面要进一步提升模型的性能，在模型中加入其他类型的上下文信息，如搭配信息、语义信息和共现信息等，并辅之以规则的方法以进一步提高组块识别的性能；另一方面在已有组块研究成果的基础上，适当地增大组块粒度，以便能更好地实现完全句法分析或者应用到其他语言处理任务中。

(3) 在组块识别的基础上，块内结构分析和块间关系分析也值得做更多的进一步的研究。

(4) 在很多实际的信息处理技术应用中，组块分析也起到了很重要的作用。伴随着中文组块分析的发展，组块在机器翻译、问答系统、信息抽取、信息检索、文本分类等领域的进一步应用也是值得期待的研究。

参考文献

[1] Abney S. Parsing by Chunks[C]//Berwiek R, Abney S, Carol T, eds. Principle-Based Parsing. Dordrecht: Kluwer Academic Publishers, 1991: 257-278.

[2] Erik F, Tjong Kim Sang, Buchholz S. Introduction to the CoNLL-2000 Shared Task: Chunking[C]//Proceedings of CoNLL-2000 and LLL-2000, Lisbon, Portugal, 2000: 127-132.

[3] 孙宏林，俞士汶. 浅层句法分析方法概述[J]. 当代语言学，2000，2(2)：74-83.

[4] 赵军，黄昌宁. 结合句法组成模板识别汉语基本名词短语的概率模型[J]. 计算机研究与发展，1999，36(11)：1384-1390.

[5] 赵军，黄昌宁. 基于转换的汉语基本名词短语识别模型[J]. 中文信息学报，1999，13(2)：1-7，39.

[6] 赵军，黄昌宁. 汉语基本名词短语结构分析模型[J]. 计算机学报，1999，22(2)：141-146.

[7] 周雅倩，郭以昆，黄萱菁，等. 基于最大熵方法的中英文基本名词短语识别[J]. 计算机研究与发展，2003，

40(3)：440-446.

[8] 王立霞，孙宏林. 现代汉语介词短语边界识别研究[J]. 中文信息学报，2005，19(3)：80-86.

[9] Y Tan，T Yao，Q Chen，et al. Applying Conditional Random Fields to Chinese Shallow Parsing[C]//David：Computational Linguistics and Intelligent Text Processing 6th International Conference，Mexico City，Mexico，2005：527-536.

[10] 李素建，刘群，白硕. 统计和规则相结合的汉语组块分析[J]. 计算机研究与发展，2002，39(4)：385-391.

[11] 张昱琪，周强. 汉语基本短语的自动识别[J]. 中文信息学报，2002，16(6)：1-8.

[12] 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，1999，22(11)：1158-1165.

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

[14] 孙广路. 基于统计学习的中文组块分析技术研究[D]. 哈尔滨：哈尔滨工业大学，2008.

[15] 李沐，吕学强，姚天顺. 一种基于 E-Chunk 的机器翻译模型[J]. 软件学报，2002，13(4)：669-676.

[16] Zhou M. A block-based robust dependency parser for unrestricted Chinese text[C]//Cardie C，Daelemans Nedelle C，Tjong Kim Sang E F：Proceedings of the 2nd Chinese Language Processing Workshop Attached to ACL. HongKong：Association for Computational Linguistics，2000：78-84.

[17] Chen WL，Zhang YJ，Hitoshi I. An empirical study of Chinese chunking[C]//Morristown，Proc. of the COLING/ACL 2006 Main Conf. Poster Sessions. Sydney，Australia：Association for Computational Linguistics，2006：97-104.

[18] 谭咏梅，王小捷，周延泉，等. 使用 SVMs 进行汉语浅层分析[J]. 北京邮电大学学报，2008，31(1).

[19] 刘芳，赵铁军，于浩，等. 基于统计的汉语组块分析[J]. 中文信息学报，2000，14(6)：28-33.

[20] Z Tiejun，Y Muyun，L Fang，et al. Statistics Based Hybrid Approach to Chinese Base Phrase Identification[C]//Cardie C，Daelemans Nedelle C，Tjong Kim Sang E F：Proceeding CLPW '00 Proceedings of the 2 Workshop on Chinese Language Processing. Hong Kong：Association for Computational Linguistics，2000：73-77.

[21] 李荫，朱靖波，姚天顺. 基于 Stacking 算法的组合分类器及其应用于中文组块分析[J]. 计算机研究与发展，2005，42(5)：844-848.

[22] H Gao，DG Huang，YS Yang. Chinese Chunking Using ESVM-KNN[C]//YM Cheng，YP Wang，HL Liu：Proceedings of the 2006 International Conference on Computational Intelligence and Security，Guangzhou：IEEE，2006：721-734.

[23] Li H，C N Huang，J Gao，et al. Chinese Chunking with Another Type of Spec[C]//Oliver Streiter，Qin Lu：Proceedings of the 3rd ACL SIGHAN Workshop. Barcelona，Spain：Association for Computational Linguistics，2004：41-48.

[24] B Li，Q Lu，Y Li. Building a Chinese Shallow Parsed Treebank for Collocation Extraction[C]//Proceedings of 4th International Conference on Intelligent Text Processing and Computational Linguistics，Mexico City，Mexico，2003：402-405.

[25] S H Wu，C W Shih，C W Wu，et al. Applying Maximum Entropy to Robust Chinese Shallow Parsing[C]//Proceedings of ROCLING-2005，Taiwan，China，2005：23-30.

[26] 周强，李玉梅. 汉语块分析评测任务设计[J]. 中文信息学报，2010，24(1)：123-128.

[27] Ramshaw L A，M M P. Text chunking using transformation-based learning[C]//Yarowsky D，Church K，eds. Proceedings of the 3rd ACL Workshop on Very Large Corpora. Massachusetts：Association for Computational Linguistics，1995：82-94.

[28] Tjong Kim Sang E F，Veenstra J. Representing text chunks[C]//Osborne M，Tjong Kim Sang E F，eds，Proceedings of EACL'99. Bergen：Assoeiation for Computational Linguistics，1995：173-179.

[29] Erik F，Tjong Kim Sang，Sabine Buchholz. Introduction to CoNLL-2000 Shared Task：Chunking[C]//Proceedings of CoNLL-2000. Lisbon，Portugal，2000：127-132.

[30] K Uehimoto，Q Ma，M Murata，et al. Named entity extraction based on a maximum entropy model and transformation rules[C]//Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics，2000：326-335.

[31] K Church. A Stochastic Parts Program and Noun Phrases Parser for Unrestricted Text[C]//Proceedings of the 2nd Conference on Applied Natural Language Processing，New Jersey，USA，1988：136-143.

[32] Ramshaw L，Marcus M. Text Chunking Using Transformation-Based Learning[C]//Proceedings of 3rd Workshop on Very Large Corpora. Massachusetts：Association for Computational Linguistics，1995：82-94.

[33] A V D. Bosch，S Buchholz. Shallow Parsing on the Basis of Words Only：A Case Study[C]//Eisner：Proceedings of 40th Annual Meeting of the Association for Computational Linguistics，Philadelphia，PA，USA：Association for Computational Linguistics，2002：433-440.

[34] T Zhang，F Damerau，D Johnson. Text Chunking Based on a Generalization of Winnow. Journal of Ma-

chine Learning Research[J]. 2002,(2): 615-637.

[35] S B Park, B T Zhang. Text Chunking by Combining Hand-Crafted Rules and Memory-Based Learning [C]//Erhard W, Dan Roth: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, Japan: Association for Computational Linguistics, 2003: 497-504.

[36] R K Ando, T Zhang. A High-Performance Semi-Supervised Learning Method for Text Chunking[C]// Kevin Knight, Hwee Tou Ng, Kemal Oflazer: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, Ann Arbor, Michigan: Association for Computational Linguistics, 2005: 1-9.

[37] Erik F. Tjong Kim Sang. Memory-Based Shallow Parsing[J]. The Journal of Machine Learning Research. 2002: 559-594.

[38] F Pla, A Molina, N Prieto. Improving chunking by means of lexical-contextual information in statistical language models[C]//Alan: Proceedings of the 2nd Workshop on Learning Language in Logic and the 4th Conference on Computational Natural Language Learning. Stroudsburg, PA, USA: Association for Computational Linguistics, 2000. 148-150.

[39] Koeling Rob. Chunking with maximum entropy models[C]//Alan: Proceedings of the 2nd Workshop on Learning Language in Logic and the 4th Conference on Computational Natural Language Learning. Stroudsburg, PA, USA: Association for Computational Linguistics, 2000: 139-141.

[40] Kudoh Taku, Matsumoto Yuji. Use of support vector learning for chunk identification [C]//Alan: Proceedings of the 2nd Workshop on Learning Language in Logic and the 4th Conference on Computational Natural Language Learning. Stroudsburg, PA, USA: Association for Computational Linguistics, 2000: 142-144.

[41] Abney S. Part of speech tagging and partial parsing [C]//Church K, Young S, Bloothooft G, eds, Proc. of the Corpus-Based Methods in Language and Speech, An ELSNET Volume. Dordrecht: Kluwer Academic Publishers, 1996: 119-136.

[42] 周强. 汉语基本块规则的自动学习和扩展进化[J]. 清华大学学报(自然科学版), 2008,4(1): 88-91.

[43] 李珩,谭咏梅,朱靖波,等. 汉语组块识别[J]. 东北大学学报(自然科学版),2004,25(2): 114-117.

[44] 李珩,朱靖波,姚天顺. 基于 SVM 的中文组块分析[J]. 中文信息学报,2004,18(2): 1-7.

[45] 周俊生,戴新宇,陈家骏,等. 基于大间隔方法的汉语组块分析[J]. 软件学报,2009,20(4): 870-877.

[46] 孙广路,郎非,薛一波. 基于条件随机域和语义类的中文组块分析方法[J]. 哈尔滨工业大学学报,2011,43(7): 135-139.

[47] Tan YM, Yao TS, Chen Q, et al. Applying conditional random fields to Chinese shallow parsing[C]// David: Computational Linguistics and Intelligent Text Processing 6th International Conference. Mexico City, Mexico: COCLing 2005. 2005: 167-176.

[48] 张昱琪,周强. 汉语基本短语的自动识别[J]. 中文信息学报,2002,16(6): 1-8.

[49] 黄德根,王莹莹. 基于 SVM 的组块识别及其错误驱动学习方法[J]. 中文信息学报,2006,20(6): 17-24.

[50] 黄德根,于静. 分布式策略与 CRFs 相结合识别汉语组块[J]. 中文信息学报,2009,23(1): 16-22.

[51] Gao Hong, Huang Degen, Liu Wei, et al. Double Rule Learning in Boosting[J]. International Journal of Innovative Computing, Information & Control. 2008,4(6): 1411-1420.

[52] G Sun, C Huang, X Wang, et al. Chinese Chunking Based on Maximum Entropy Markov Models[J]. International Journal of Computational Linguistics and Chinese Language Processing. 2006, 11(2): 115-136.

[53] G Sun, Y Guan, X Wang. A Maximum Entropy Chunking Model with N-Fold Template Correction [J]. Journal of Electronics. 2007,24(5): 690-695.

[54] 刘世岳,李珩,张俐,等. Co-training 机器学习方法在中文组块识别中的应用[J]. 中文信息学报,2005,19(3): 73-79.

[55] 梁颖红,赵铁军,于浩,等. 基于改进 K-均值聚类的汉语语块识别[J]. 哈尔滨工业大学学报, 2007,39(7): 1106-1109.

[56] 程葳,赵军,刘非凡,等. 面向口语翻译的双语语块自动识别[J]. 计算机学报,2004,27(8): 1016-1020.

[57] 宇航,周强. 汉语基本块的内部关系分析[J]. 清华大学学报(自然科学版),2009,49(10): 136-140.

[58] Yoshimasa Tsuruoka, Jun' ichi Tsujii, Sophia Ananiadou. Fast Full Parsing by Linear-Chain Conditional Random Fields[C]//Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), Association for Computational Linguistics, Athens, Greece,2009: 790-798.

[59] 周强,赵颖泽. 汉语功能块自动分析[J]. 中文信息学报,2007,21(5): 18-24.

[60] 陈亿,周强,宇航. 分层次的汉语功能块描述库构建分析[J]. 中文信息学报,2008,22(3): 24-31,43.

[61] Taro Watanabe, Eiichiro Sumita, Hiroshi G Okuno. Chunk-based Statistical Translation[C]//Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Sapporo, Japan,2003: 303-310.

汉语组块分析研究综述

作者： 李业刚， 黄河燕， LI Yegang， HUANG Heyan
作者单位： 李业刚,LI Yegang(北京理工大学计算机学院,北京100081;山东理工大学计算机科学与技术学院,山东淄博255049)， 黄河燕,HUANG Heyan(北京理工大学计算机学院,北京,100081)

刊名： 中文信息学报 ISTIC PKU
英文刊名： Journal of Chinese Information Processing
年，卷(期)： 2013,27(3)