《汉语组块分析研究综述》阅读笔记

汇报人：张贺

导师：刘茂福

《汉语组块分析研究综述》

- 作者：李业刚，黄河燕
- 期刊：中文信息学报，第27卷，第三期，2013年5月
- 链接：http://www.cnki.com.cn/Article/CJFDTOTAL-MESS201303002.htm

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。

背景

- 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率先提出了组块分析的思想（1991）后，国际会议CoNLL—2000把组块分析作为共享任务提出，组块分析逐步受到重视。

组块分析示例

从自然语言学习国际会议CoNLL一2000的共享任务中截取一段例子。作为语块的示例：

[NP He] [VP reckons] [NP the current account deficit] [VP will narrow] [PP to] [NP only # 1.8 billion] [PP in] [NP September]

其中NP、VP、PP分别表示名词短语、动词短语和介词短语。

组块的定义

- 英文组块定义：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。

- 中文组块定义：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。

中文组块定义示例

- 李素建等提出的组块定义：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。
- 周强等提出的组块定义：组块是句子中相邻的、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。
- 孙广路等提出的组块定义：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。
- 其他定义若干。

组块的类型

- 张显琪，周强等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9种基本汉语组块类型，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。这9种基本短语包括名词短语np, 动词短语vp, 形容词短语ap, 副词短语dp, 数词短语mp, 区别词短语bp, 地点短语sp, 时间短语tp, 准数词短语mbar。
- 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。
- 文中提到一句话：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。这句话该如何理解？

组块分析评测（正确率P、召回率R、F值）

- 对于某种类型的组块：

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

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

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

- 对于所有类型的组块（F值与单一类型相同）：

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

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

组块识别——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。
- 《汉语基本块规则的自动学习和扩展进化》（周强，2008）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％。

组块识别——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；
- 半指导和无指导方法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。

有指导学习方法

- 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％。

有指导学习方法

- 文献[46]将条件随机域模型应用到中文组块分析中，利用语义词典抽取语义类特征，将其加入分析模型， 在微软亚洲研究院(MSRA) 中文组块分析语料库上得到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。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，系统开放式测试的F值达到92.91％。

有指导学习方法

- 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％。

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。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％。

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%。

统计与规则相结合的方法

- 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。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

结论与展望

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。

谢谢！