基于规则的多层次组块合并研究

余小鹏[1] 黄雨菲[1] 徐健儿[1] 姚小桐[2]
YU Xiaopeng HUANG Yufei XU Jianer YAO Xiaotong

摘 要 针对句法分析组块研究中存在的组块识别粒度难衡量问题，为解决组块识别后语法易丢失等问题，提出一种基于规则的多层次组块合并模型。先定义词和组块的二元或三元合并规则，然后考虑组块合并粒度和规则的优先顺序，相比于目前的一次合并，引入多层层级关系，设置组块合并的层次顺序，通过在不同层级中设置对应的合并规则形成优先顺序来合并组块。该模型通过层级关系和合并规则对组块合并进行约束，解决了组块粒度难以衡量和语法易丢失等问题，可以帮助计算机更好地抽取文本信息、理解语义，提高自然语言理解的准确性和效率。

关键词 组块；句法分析；规则；层次结构

doi：10.3969/j.issn.1672-9528.2025.01.019

0 引言

在自然语言处理中，组块分析作为一种重要的技术手段，正逐渐受到广泛关注。组块分析旨在将文本中的单词组合成有意义的、句法相关的单元，即组块。这种分析方法为语言理解和处理带来了诸多益处，能够简化文本的句法结构，使复杂的句子变得更易于理解和处理。通过识别组块，可以清晰地把握句子的核心成分和语义关系，从而为信息检索、机器翻译、问答系统等应用提供更准确的基础。例如，在信息检索中，组块分析可以帮助其提取关键信息，提高检索的准确性和效率；在机器翻译中，有助于更好地理解源语言的句法结构，从而生成更符合目标语言习惯的译文。组块分析无疑为自然语言处理的进一步发展和应用开辟了新的途径，具有重要的研究价值和实践意义。

组块的研究中主要有基于规则的方法、基于统计学习的方法、基于深度学习的方法以及混合的方法。目前使用较多的是基于规则和统计学习混合的方法，杨陈菊等人[1]提出的一种结合条件随机场和多元规则的层次化句法分析模型，可以更好地进行句法分析。旦正吉等人[2]提出了基于藏文音节向量和 BiLSTM-CRF 混合模型相结合的藏语语义组块识别方法。该方法在组块分析中取得了一定效果，但建立的规则难以跨领域移植，且统计学习依赖大量的训练文本。为了避免这些不足，相关学者尝试使用深度学习的方法，谷波等人[3]提出了一种基于 RNN 的中文二分结构句法分析，但忽略了中文部分语句不满足二分结构而满足三元结构的语句状况，导致其句法分析正确率偏低。Kitaev 等人[4]提出了一种改进的神经网络模型进行词性识别和依存句法分析，在实验语料中效果较好。但基于深度学习方法的可解释性较差，数据资源和计算力成本高，且仍不能得到较好的组块分析结果。

本文提出了一个基于规则的多层次组块合并模型，在预先定义好的规则中，设置规则的优先级去定义层级，通过规则和层次顺序合并出粒度合适的组块。

1 相关研究

Abney（1991）[5]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，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[6]。

目前组块识别有基于规则的方法、基于机器学习的方法、规则和统计相结合的方法、基于深度学习的方法和规则。下面将从这 4 种方法进行阐述。

（1）基于规则即根据人工书写的或（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，主要困难在于语法规则的获取以及语法规则之间的优

1. 武汉工程大学管理学院 湖北武汉 430000
2. 武汉工程大学计算机学院 湖北武汉 430000
[基金项目] 湖北省社科基金后期资助项目“代数应用题表征辅导系统”（HBSKJJ20233292）

先顺序排列。主流方法为基于有限自动机、基于转换的错误驱动和模式匹配的方法。钱小飞[7]利用有限状态自动机，对法语和英语双语语料进行了NP短语的自动抽取。Ramshaw等人[8]用基于转换的错误驱动的方法来识别基本名词短语（BaseNP）和NP、VP块。

（2）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需要花费大量人力物力，而无指导的缺点则在于一般迭代算法的复杂度高，运算效率差，且不能很好地保证最终训练结果的语法可靠性。李珩等人[9]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%。

（3）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处理方法。于鸿霞[10]将支持向量机方法和基于转换学习的方法相结合。

（4）随着深度学习的发展，神经网络逐渐应用于图像（XUE B 2019）、文本等方面，对句法的研究也不再停留于浅层分析。皮乾东等人[11]基于汉语语序算式化融合设计了句法分析器；贾继康等人[12]通过规则合成的方法进行了层次化语句识别。

目前对组块分析的研究中，现有的组块识别环节通常只能识别出一种结果，根据需要只识别出细粒度组块或者只识别粗粒度组块。而在浅层句法分析中，需要识别出组块之间的依存关系，细粒度组块会使得关系错综复杂，信息不足，存在组块语法识别错误的情况；粗粒度组块会使得关系过于简单，无法捕捉到细微但重要的语义差异。因此，需要训练不同的模型去识别不同粒度的组块。

2 提出问题

目前基于规则、统计、深度学习的方法在组块研究上已取得较好成果，但当前对组块分析上还存在一些不足。

（1）对组块的粒度难以衡量。现有的句法分析只能识别出细粒度或粗粒度组块，而过大的组块粒度可能会丢失细粒度信息，无法捕捉到细微但重要的语义差异，同时该粒度组块包含的信息过多，使得组块的具体信息模糊；过小的粒度组块会导致组块包含信息量不足，并且小组块增加了模型计算的复杂性，也会更容易受到噪音的影响。

例如，数学应用题中“一辆汽车走高速路的速度是80千米/时”，细粒度分析是先进行词性标注，使用哈工大LTP分析“一 /m 辆 /q 汽车 /n 走 /v 高速路 /n 的 /u 速度 /n 是 /v 80/m 千米 / 时 /q”，然后按照一定的语法将标注后的字词进行合并“[NP 一辆汽车] [VP 走高速路的] [NP 速度] [VP 是] [QP 80 千米 / 时]”（请落实）；粗粒度分析是直接将句子分解为较大的语法成分，上句进行分析就是“一辆汽车”名词短语作主语。只能按照需求识别出细粒度组块或是粗粒度组块。

（2）组块合并丢失语法规则。中文语法词序较为固定，通常采取“主谓宾”结构，通过对不同成分的修饰，句子中还存在着其他成分，如定语、状语、补语等。句子成分之间的关系主要有主谓关系、动宾关系、定中关系、状中结构和动补结构等。目前组块合并的方法中，基本都是进行一次合并，识别出句子中不同的短语结构，如名词短语、动词短语等。而句子中的某些名词或者动词可能作修饰成分和中心词之间的关系为定中或状中关系，并非简单的名词或动词短语。在进行组块合并时丢失了语法规则。

例如，数学应用题中“这个粮仓存放的稻谷约有多少千克？”对这句话使用LSTM+CRF进行组块合并，结果是“[NP 这个粮仓] [VP 存放的] [NP 稻谷] [VP 约有] [QP 多少千克？]”。使用中文语法分析，其中被识别为动词短语“存放的”和名词短语“稻谷”实际上是一个动词“存放”加上“的”字修饰的词组，表示一种修饰关系。这里的“存放的”是一个定语，修饰名词“稻谷”，在句子中是定中结构，整体上“存放的稻谷”表示一个名词词组。

清华大学教授黄昌宁在1992年开始研究中文句法，其团队前后提出了多种依存关系集，包括32种、44种以及更为详细的106种依存关系[13-14]。他认为要想把汉语复杂的依存关系描述清楚，需要建立许多依存关系。例如，按照谓语成分，可以把主语分为动词性主语、从句性主语等，甚至可以更加细致划分。因此在对中文句子进行组块识别时，被识别出的组块之间应满足对应的依存关系，即语法规则。

针对以上基于组块研究的不足及分析，本文根据句子的合并规则及优先级，以数学应用题语义理解为研究对象，提出基于规则的多层次组块合并模型。

3 基于规则的多层次组块合并模型研究

3.1 构建词和组块合并规则

为构建基于规则的多层次组块合并模型，首先应定义组块合并的规则，本文以数学应用题文本为例，结合哈工大LTP[15]对于已词性标注的句子进行词合并组块，通过黄昌宁授团队定义的44种依存关系并且结合数学应用题文本提炼出3种词的合并规则如表1，其中涉及到的组块标准参考文献[16]。

（1）合并为名词性短语：名词中有修饰成分，修饰成分和名词之间的依存关系为定语，最常见的有“的”字结构、

形容词、名词。例如“存放的稻谷”为“的”字结构名词性短语；“普通公路”为形容词构成的名词性短语；“小学校园”为两个名词构成的名词性短语。

（2）合并为动词性短语：动词与修饰成分之间的依存关系为状语，一般位于动词前面，二者可以合并为动词性短语。例如：“约有”“大约是”。在数学应用题中也会有两个动词连续出现的时候，如“宣传教育”，因此两个动词连续出现也可以合并为一个动词性短语。

（3）合并为其他词性短语：可以将数字和量词进行合并，如“200 平方米”“480 千米”等。复数关系也可以直接合并为名词性短语，如“同学们”“工人们”等。还有名词的“和”字关系，可以合并为名词性短语，如“小明和小黄”。代词和名词一起出现时，通常可以合并为名词性短语，如“这个粮仓”。以及代词和量词可以合并为其他词性短语，如“多少米”。

表 1 词和组块合并规则

词性短语	合并规则
名词性短语	a+n=np; r+n=np; n(p)+n(p)=np; n/v(p)+ 的 +n(p)=np
动词性短语	v+v=vp; d+v=vp
其他词性短语	m+q=qp; 这 +q=qp; n+ 们 =np; (n)+ 在 /p+n=np; n(p)+ 和 +n(p)=np; r+q=qp

3.2 组块合并的层次顺序

在定义词合并规则之后，要考虑词合并规则的优先顺序，以及将词或组块合并的粒度，组块粒度过小会导致包含的信息不足，过大会导致组块信息过于模糊，丢失细粒度信息，因此在一个句子中将组块合并到合适粒度至关重要。文献 [1] 在识别出的组块中层层引入不同优先级的二元、三元规则，实现了同时进行细粒度和粗粒度组块的识别。

本文在以数学应用题为例，在对数学应用题文本进行组块分析时，发现引入三层层级关系可以将该文本组块划分为粒度大小合适的组块，并在每个层级关系中设置对应的词合并规则形成优先顺序，通过层级关系和词合并规则对组块进行合并，最终合并成粒度合适的组块，如表 2。

表 2 层次关系

层级	合并规则

层次一	a+n=np; n+n=np; r+n=np; d+v=vp; v+v=vp;
	m+q=qp; r+q=qp;
	这 +q=qp; n+ 们 =np; (n)+ 在 /p+n=n; n+ 和 +n=np

层次二	n(p)+ 的 +n(p)=n; v(p)+ 的 +n(n)=np; n(p)+ 和 +n(p)=np
层次三	n(p)+n(p)=np

3.3 模型构建步骤

在确定词合并规则和层级关系以后，开始构建基于规则的多层次组块合并模型，模型构建步骤如下：

（1）第一层：首先对句子进行预处理，使用哈工大 LTP 对句子进行词性标注，输入经过预处理的语句块，在第一层中使用预定的规则进行词合并组块工作。

（2）第二层：先判断，语句块是否有符合组块合并规则的词或组块，没有则直接输出组块合并后的句子，有则按照第二层词和组块合并规则进行合并。

（3）第三层：同样是先进行判断，然后按照预定规则进行合并。如果语句块可以进行到这一层，表示语句中合并出了粒度较大的组块，同时也包含了粒度小的关键动词组块，解决了组块合并中对组块粒度难以衡量的问题。

以数学应用题“这个粮仓存放的稻谷约有多少千克？”为例，说明模型的构建方法。（1）预处理：使用哈工大 ltp 对该句子进行分词和词性标注处理，结果为“这个 /r 粮屯 / n 存放 /v 的 / 稻谷 /n 约 /d 有 /v 多少 /r 千克 /q ？ /wp”。（2）第一层合并，根据第一层级的合并规则，该语句块满足的合并规则有 r+n=np，即“这个 /r 粮屯 /n”合并为“这个粮仓 /np”和 d+v=vp，即“约 /d 有 /v”合并为“约有 /vp”以及 r+q=qp，即“多少 /r 千克 /q”合并为“多少千克 /qp”。第一层词合并结束后组块合并结果为“这个粮仓 /np 存放 / v 的 / 稻谷 /n 约有 /vp 多少千克 /qp”。（3）第二层合并，判断有符合第二层规则的语句块，符合的规则有 v(p)+ 的 +n(n)=np，即“存放 /v 的 / 稻谷 /n”合并为“存放的稻谷 / np”。第二层词合并结束后组块合并结果为“这个粮仓 /np 存放的稻谷 /np 约有 /vp 多少千克 /qp”。（4）第三层合并，判断可以合并，符合该层合并规则有 n(p)+n(p)=np，即“这个粮仓 /np 存放的稻谷 /np”合并为“这个粮仓存放的稻谷 /np”，最后词合并结束后组块合并结果为“这个粮仓存放的稻谷 /n 约有 /v 多少千克 /qp ？ /wp”见表 3。

表 3 基于规则的多层次组块合并分析图

输入： 这个粮仓存放的稻谷约有多少千克？
ltp 预处理：这个 /r(代词) 粮屯 /n 存放 /v 的 / 稻谷 /n 约 /d 有 /v 多少 /r 千克 /q ？ /wp
第一层：r+n=np 这个粮仓 /np d+v=vp 约有 /vp r+q=qp 多少千克 /qp
第二层： v+ 的 +n=np 存放的稻谷 /np
第三层： n(p)+n(p)=np 这个粮仓存放的稻谷 /np
组块： 这个粮仓存放的稻谷 /np 约有 /vp 多少千克 /qp ？ /wp

由构建步骤可知，基于规则的多层次组块合并模型是根据预先定义的词合并规则和组块合并层次顺序逐步生成的，

词和组块合并规则是根据依存句法中依存关系定义，组块合并层次用来设置合并规则的优先级如图 1，因此在数学应用题领域中，该模型可以将句子合并为粒度大小合适的组块，并且符合中文语法规则。

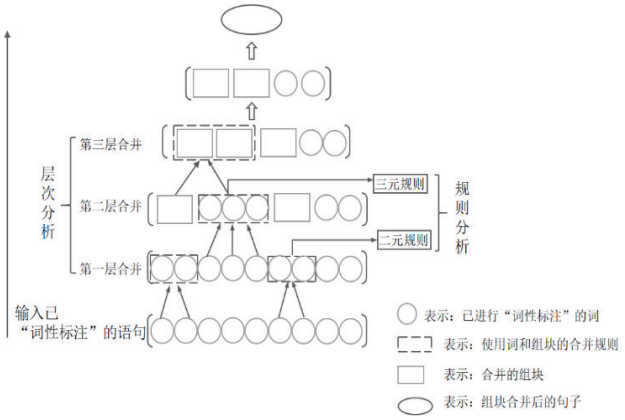

图 1 基于规则的多层次组块合并模型示意图

4 结语

组块分析可以帮助计算机更好地抽取文本信息、理解语义，提高自然语言理解的准确性和效率。但目前对组块的研究存在组块粒度难以衡量、易丢失语法规则等问题。本文从数学应用题领域入手，深入分析了句子间的相互依存关系，基于黄昌宁教授的 44 种依存关系构建了词和组块的合并规则，并加入层次关系，设计出了基于规则的多层次组块合并模型。在数学应用题领域中对组块进行分析，使用该模型可以有效的解决上述问题，提高处理文本的效率。此外，该模型可以通过定义不同的词和组块合并规则和层次运用到其他领域，提高自然语言处理的准确性和效率。

参考文献：

[1] 杨陈菊 , 孙俊 , 皮乾东 , 等 . 基于 CRF 和多元规则的层次化句法分析 [J]. 吉林大学学报 (理学版), 2020,58(6):1452-1460.

[2] 旦正吉 , 华却才让 , 完么措 , 等 . 基于藏文音节结合 BiLSTM-CRF 的藏语语义组块分类标注 [J]. 高原科学研究 , 2024, 8(2):118-125.

[3] 谷波 , 王瑞波 , 李济洪 , 等 . 基于 RNN 的中文二分结构句法分析 [J]. 中文信息学报 ,2019,33(1):35-45.

[4]KITAEV N, KLEIN D. Constituency parsing with a self-attentive encoder[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Brussels:ACL,2018:2676-2686.

[5]ABNEY S P. Parsing by chunks[J]. Principle-based parsing, 1991, 44: 257-278.

[6] 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+124.

[7] 钱小飞 . 组块分析研究综述 [J]. 现代语文 , 2018(6):166-170.

[8]RAMSHAW L A, MARCUS M P. Text chunking using transformation-based learning[J].Natural language processing using very large corpora. 1995,11:157-176.

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

[10] 于鸿霞 . 统计与规则相结合的中英文组块分析 [D]. 哈尔滨 : 哈尔滨工业大学 , 2006.

[11] 皮乾东 , 邵玉斌 , 龙华 , 等 . 汉语语句算式化融合句法分析 [J]. 电子测量技术 , 2020,43(6):123-127.

[12] 贾继康 , 邵玉斌 , 龙华 , 等 . 一种基于规则与句法合成的层次化语句分析识别算法 [J]. 吉林大学学报 (理学版), 2020, 58(4): 885-892.

[13] 周明 , 黄昌宁 . 面向语料库标注的汉语依存体系的探讨 [J]. 中文信息学报 ,1994,8(3):35-52.

[14] 周明 , 黄昌宁 , 张敏 , 等 . 统计与规则并举的汉语句法分析模型 [J]. 计算机研究与发展 , 1994, 31(2):40-49.

[15]CHE W X, FENG Y L, QIN L B, et al. N-LTP: an open-source neural language technology platform for Chinese[DB/OL].(2021-09-23)[2024-04-11].https://doi.org/10.48550/arXiv.2009.11616.

[16] 仵永栩 , 吕学强 , 周强 , 等 . 汉语概念复合块的自动分析 [J]. 中文信息学报 ,2016,30(2):1-11.

【作者简介】

余小鹏（1974—），男，湖北咸宁人，博士，教授，研究方向：信息系统与电子商务、数据挖掘、教育信息技术。

黄雨菲（2001—），女，湖北孝感人，硕士研究生，研究方向：信息系统与电子商务、数据挖掘，email: 2814544609@qq.com。

徐健儿（2000—），女，广东广州人，硕士研究生，研究方向：信息系统与电子商务、数据挖掘。

姚小桐（1999—），女，江苏南京人，硕士研究生，研究方向：自然语言处理、数据挖掘。

（收稿日期：2024-10-10）