基于规则的汉语基本块自动分析器

周 强

清华信息科学与技术国家实验室（筹）
清华大学信息技术研究院语音与语言技术中心，北京 100084
zq-lxd@mail.tsinghua.edu.cn

摘要：本文提出了一种规则驱动的汉语基本块自动分析方法，它的主要分析资源是从大规模标注语料库和词汇关联知识库的交互作用中自动习得的融合内部词汇关联和外部语境限制约束知识的分层次、多粒度的基本块规则库。利用其中各条规则的置信度信息，可以有效地驱动汉语真实文本句子的多词语基本块的自动识别过程，同时完成歧义结构自动排歧。初步的实验结果表明，现有分析器可以在 95%以上的开放测试语料上达到 90%左右的 F-measure 值，同时又保留了约 5%的在现有知识库条件下很难判断的复杂歧义结果供后续分析器选择使用，显示出较好的处理灵活性和有效性。

关键词：基本块，部分分析，规则驱动排歧

A rule-based Chinese base chunk parser

ZHOU Qiang

Tsinghua National Laboratory for Information Science and Technology（TNList）
Center for Speech and Language Technology, Research Institute of Information Technology
Tsinghua.University, Beijing 100084

ABSTRACT: This paper proposed a rule-driven Chinese chunking algorithm, whose main parsing resource is a hierarchical rule base automatically learned from the interaction of a large-scale annotated corpus and a lexical knowledge base. Each rule in it can obtain a confident score to evaluate the construction reliability of a special multiword chunk under some refined knowledge about its internal lexical relationship and external contextual restriction. These confident values give us alternative opportunity for efficient multiword chunk recognition and disambiguation. Some primitive experimental results indicate that our current parser can achieve the chunking performance of about 90% overall F-measure under 95% open testing texts, and still keep 5% ambiguous regions due to the deficiency of reliable enough knowledge. Therefore, the more complex parser can be developed to select suitable chunks among them based on some new knowledge and larger contexts.

KEYWORDS: Base Chunk, Partial Parsing, Rule-driven Disambiguation

1．引言

块（chunk）分析作为一种重要的部分分析技术，可以通过对完整分析问题的合理任务分解，大大降低自动分析的处理难度，在自然语言处理领域的信息抽取、问答系统、文本挖掘等应用系统研究都可以发挥重要作用。

Abney(1991)最先提出了块分析概念，Ramshaw & Marcus (1995)通过‘ BIO ’模型把块分析转化为序列标记确定问题，为各种机器学习方法的应用打下了很好的基础[TB00]。本文则从另一个角度定义了块分析问题：针对一个输入句子，首先通过词语聚合性和周围语境约束限制分析，确定哪些

词语组合可以形成一个多词语基本块；然后把剩余词语中的实义词直接上升为单词语基本块，这样就可以形成一个包含多词语基本块、单词语基本块和其他功能词的完整的块描述序列。与‘ BIO ’序列标记识别模型不同的是，这种处理思路更强调不同基本块的内部词汇聚合分析，从而可以比较方便地建立起语料库的块描述实例与词汇语义知识库之间的内在联系。

从这个思路出发，本文提出了一种基于规则的汉语基本块分析方法，并据此开发完成了一个面向汉语真实文本的基本块自动分析器。在下面的几节中，第 2 节介绍了分析算法的基本处理策略；第 3 节给出了详细的实验结果分析数据；第 4 节介绍了相关研究工作；最后的第 5 节是结语。

2．基本块分析策略

基本块分析器的设计目标，是在基本块规则库和词汇知识库的支持下，对经过词语切分和词性标注处理的汉语真实文本句子进行自动分析，识别出其中的各个基本块的边界位置，确定其成分和关系标记以及分析置信度，得到句子的基本块分析标注结果。有关汉语基本块的详细定义方法可参阅文献[ZQ06]。

为了充分发挥自动习得的分层次描述规则的处理能力，提高匹配效率，我们设计了以下基本块规则保存结构：1）基本规则表：保存所有词类标记串描述规则。如：v+n → {vp-PO, 3140; np-DZ, 48}, 3671 [1]；2）扩展规则表：保存所有经过进化学习得到的扩展规则。如：v(词语=WC-L)+n(词语=WC-R)__v → { vp-PO, 308}, 23[2]。通过在基本规则表中记录相关的扩展规则的索引信息（在扩展规则表中的起止位置）建立起两者之间的内在联系。其中每条规则都可以计算出一个置信度值 θ，表示在该规则的约束条件下形成一个多词语基本块的可靠程度。这个信息将在后续的块组合分析和排歧中发挥重要作用。

在具体匹配分析过程中，首先获取句子中待分析位置的词类标记串，检索基本规则表，如果发现某个能匹配的基本规则，则进一步检查是否存在扩展规则。如果存在，则对句子中的相关位置进行分层次的信息扩展，检查扩展组合是否在扩展规则库中出现。如果发现了扩展规则，则从所有的规则中选择置信度最高的规则作为匹配规则输出。否则，使用基本规则作为缺省的匹配规则。为了保证得到比较可靠的多词语块，我们设置了一个排除阈值（DelTh=0.5）。只有当匹配规则的置信度大于 DelTh 时，才选择作为一个有效的多词语块加入分析结果表中。

针对分析过程中碰到的各种歧义情况，我们采用了基于规则置信度的动态排歧策略：1）对于交集型歧义，即分析得到的基本块与其他基本块发生边界交叉的情况，计算这两条基本块规则的置信度差值。如果这个差值超过某个阈值 Th1（目前选择 0.2），则删除其中置信度较低的基本块；2）对于组合型歧义，即某个词语组合在某些语境下可以组合为一个完整的基本块，在其他语境下又可拆分成多个基本块的情况，计算‘ 分 ’、‘ 合 ’情况下的基本块规则的置信度，如果它们的差值超过某个阈值 Th2，则删除其中置信度较低的组合情况。

另外，对于分析过程中发现的一类特殊的交集型歧义情况，即多个交集型组合基本块可以形成一个链式关联结构，如：“ 中国 医学 宝典 ”，可能分析为两个交叉基本块：“ 中国 医学 ”和“ 医学 宝典 ”，我们通过分析其内部聚合性和外部语境限制性确定是否需合并为一个更大的链式结构基本块，从而完成排歧任务。

经过以上排歧处理，句子中还会剩下一部分歧义情况不能排除。它们或者是由于目前规则描述能力所限，对某些歧义组合不能产生明显的置信度差异；或者是由于某些歧义情况的排除需要考虑更大的语境信息，超过了目前规则的处理能力。对于这些歧义情况，我们将保留在分析结果表中，

---

[1] 表示在训练语料中，有 3140 个相邻的“ v+n ”组合可以归约为“ vp-PO ”块，48 个可归约为“ np-DZ ”块，有 3671 个不能成块（即标注反例），因此其置信度为 0.46。

[2] 表示在训练语料中，共有 308 满足条件“ 形成动宾关联对，同时在句子中右相邻词类为 v ”的“ v+n ”组合可以归约为“ vp-PO ”块，23 个不能成块，因此其置信度为 0.93。

留待我们获取更多的词汇语义知识后或送到更高层次的句法分析器进行处理。

3．实验结果分析

为了准确测试目前开发完成的汉语基本块自动分析器的处理性能，我们从目前的 TCT 标注语料[ZQ04]中选择了所有的新闻类文本，总规模约 20 万词，包括 8137 个句子，平均句长为 25 个词。将它分成两部分：80%作为训练语料，主要用于规则学习；20%作为测试语料，主要用于分析器的性能评价。

在训练语料上，通过规则学习和扩展进化处理，我们得到了以下分层次、多粒度的基本块规则库：1) 在基本规则层面，包含 61 条高可靠和 150 条中低可靠的词类标记描述规则；2）在扩展规则层面，包括 2503 条高可靠规则和 2469 条中度可靠规则，规则总数为 4972 条，它们与上面的 150 条中低可靠的词类标记描述规则相对应，通过引入更多的内部词汇关联和外部语境限制约束知识，大大提高了相应的基本规则的处理置信度。

同时，与上面扩展规则的具体应用相配合，我们还使用了以下词汇知识库：1）动宾关系词汇关联库，基本规模是，动词词条 5346 个，词汇关联对 52390 个；2）特征动词表，基本规模为，体宾动词 5712 个，谓宾动词 1065 个；3）名词语义信息表，包含常用名词的 11 个语义大类信息，基本规模是名词词条 26905 个。

3.1 识别性能分析

考虑到目前分析器的具体处理情况，我们首先把分析结果按照是否包含歧义分成以下三大类：1）无歧义区间；2）组合型歧义区间；3）交集型歧义区间。针对目前的处理语料，在开放测试情况下，上述 3 个区间覆盖的词语占处理词语总数的比率分别为：0.955、0.026、0.020。表明目前的块分析器对大部分处理语料，可以很好地完成分析和排歧工作。

在两个歧义区间内，我们通过以下指标分析歧义结果对正确分析结果的覆盖能力：1）正确结果招回率，描述所有歧义分析结果中包含正确结果的比率，计算公式为：歧义结果中包含的正确结果总数/歧义区间涉及的正确结果总数*100%；2）歧义分布率，描述针对一个正确结果可能形成的歧义分析平均数目，计算公式为：歧义结果总数/歧义区间涉及的正确结果总数。表 1 显示了目前的处理结果。从中可以看出，剩余的组合型歧义结果保存了绝大部分的正确分析结果，对它们的分合选择需要参考更大的语境信息。复杂的交集型歧义是目前的处理难点，其自动排歧需要引入更多有效的词汇语义信息描述。

表 1　歧义区间处理结果分析数据

	组合型歧义		交集型歧义	
	正确结果招回率	歧义分布率	正确结果招回率	歧义分布率
封闭测试	96.25	1.88	76.60	2.72
开放测试	97.75	1.75	67.79	2.58

在无歧义区间内，我们通过以下指标分析基本块的识别能力：1）基本块识别正确率（P），计算公式为：分析正确的基本块总数/自动识别出的基本块总数*100%；2）基本块识别招回率（R），计算公式为：分析正确的基本块总数/正确的基本块总数*100%；3）正确率和招回率的几何平均值 F-Measure，计算公式为：2*P*R/(P+R)；

针对不同块，确定不同的正确性判定标准：对多词语基本块，考虑以下两个层次：1）块边界、成分标记和关系标记完全相同(B+C+R)；2）块边界和成分标记相同，关系标记可以不同(B+C)；对单词语基本块，主要判断块边界和成分标记的相同性(B+C)；对所有块（多词语块+单词语块），主要

判断块边界和成分标记的相同性(B+C)；

表 2 显示了所有块的整体处理结果，表 3 进一步显示了开放测试中不同类型基本块的处理结果。从中可以看出，多词语数量块（mp）、时间块（tp）和形容词块（ap）达到了很高的处理 F-M 值，表明针对这三个块的自动习得规则已经达到了很好的描述能力，基本上覆盖了这些块的各种分布情况。多词语动词块（vp）、名词块（np）和空间块（sp）还有很大的提升空间。其中空间块的识别效果较差，主要原因是我们目前的基本块标注体系[ZQ06]中引入了单个名词与方位词形成的方位结构组合，这是汉语中一种很难分析的结构，在许多情况下需要考虑更大的语境信息才能确定合适的方位结构左边界。而目前的规则只考虑了左右各一个词语的语境信息，因此在许多情况下很难给出准确的分析判断。类似的方位组合也影响了时间块的识别精度。

vp 和 np 块在真实文本语料中占了 76%以上，对它们的准确识别是我们研究的重点。从目前的处理结果看，vp 块的 F-M 值高出 np 块 3-4 个百分点，这种性能差异情况在考虑关系标记的情况下更为明显。这充分显示了动宾词汇关联信息在提高基本块的边界识别和内部关系分析性能方面的重要作用。而 vp 块和 np 块开放测试 F-M 值比封闭测试有所下降，表明目前针对它们的规则描述还很不充分，许多测试语料中出现的新分布情况没有被训练语料所覆盖。特别是在应用词汇语义信息的条件下，目前的十几万词规模的训练语料是明显不够的。需要我们在今后进一步探索相关的词汇关联知识自动获取和规则学习应用方法。

表 2 所有块（多词语块+单词语块）的整体实验结果（B+C）

	封闭测试			开放测试		
标记	正确率	召回率	F-M	正确率	召回率	F-M
np	90.32%	88.74%	89.53%	88.33%	86.62%	87.46%
vp	91.13%	92.36%	91.74%	89.44%	90.58%	90.00%
mp	95.00%	96.50%	95.75%	92.18%	94.72%	93.43%
ap	93.55%	93.45%	93.50%	94.28%	95.65%	94.96%
tp	91.49%	91.84%	91.67%	89.97%	89.82%	89.89%
sp	81.88%	88.08%	84.87%	80.12%	88.97%	84.31%
合计	91.04%	91.39%	91.21%	89.33%	89.61%	89.47%

表 3 开放测试实验结果

	多词语块 1：B+C+R			多词语块 2：B+C			单词语块：B+C		
标记	正确率	召回率	F-M	正确率	召回率	F-M	正确率	召回率	F-M
np	75.25%	75.76%	75.50%	83.68%	84.25%	83.97%	91.74%	88.28%	89.97%
vp	83.23%	81.46%	82.34%	87.35%	85.49%	86.41%	90.65%	93.69%	92.15%
mp	94.89%	95.26%	95.08%	94.89%	95.26%	95.08%	54.55%	83.33%	65.93%
ap	93.99%	97.33%	95.63%	93.99%	97.33%	95.63%	94.42%	94.83%	94.62%
tp	92.75%	88.18%	90.40%	93.52%	88.92%	91.16%	83.78%	91.63%	87.53%
sp	78.76%	86.41%	82.41%	79.65%	87.38%	83.33%	81.25%	92.86%	86.67%
合计	81.76%	81.44%	81.60%	87.01%	86.67%	86.84%	89.33%	89.61%	89.47%

3.2 错误原因分析

我们把所有自动分析错误按照不同错误原因分成以下几类：

1) 分析不足情况：自动分析器将一个完整的多词语块拆分为几个单词语块。如：… 是/vC

[np-ZX 我们/rN 自己/rN] 的/u … → … 是/vC [np-SG 我们/rN] [np-SG 自己/rN] 的/u …（‘ → ’左边为正确结果，右边为自动分析的错误结果，下同）

2) 分析过度情况：自动分析器将相邻的几个单词语块合并为一个错误的多词语块。如：…处/n [np-ZX 古/b 建筑群/n] [np-SG 飞檐斗拱/iN] … → … 处/n [np-LN 古/b 建筑群/n 飞檐斗拱/iN] …

3) 并列结构问题：由于某个块处于并列成分位置（顿号或并列连词的左右侧）而产生的边界识别错误。如：… [vp-SG 清洗/v] [np-SG 雕像/n] 、/、 [np-SG 壁画/n] 等/u …→ … [vp-PO 清洗/v 雕像/n] 、/、 [np-SG 壁画/n] 等/u …

4) 成分标记错误：对于某个基本块，自动分析器确定为错误的成分标记。如：… [tp-ZX 二战/nR 期间/f] ，/，… → … [sp-ZX 二战/nR 期间/f] ，/，…

5) 其他情况：主要包括自动分析结果与正确结果产生的边界交叉情况，如：… [vp-ZX 喜/aD 送/v] [np-SG 贵子/n] … → … [dp-SG 喜/aD] [vp-PO 送/v 贵子/n] …

其中的第 1 类错误是由于相关规则缺失引起的，主要原因是目前的训练语料规模还不够大，对一些特殊组合不能很好地习得比较可靠的基本块描述规则。第 2、4、5 类错误是由于相关规则的约束条件学习得不够充分，致使在分析和排歧中不能很好地选择到合适的基本块组合。第 3 类错误则涉及到更大语境的成分关系确定问题，在目前的规则描述体系下还很难解决。表 4 列出了这些分析错误的具体分布数据，从中可以看出，由于规则缺失和规则信息描述不准确引起的分析错误占了绝大多数，这为我们下一步改进系统指明了方向，即引入更多有效的词汇关联知识描述资源和标注语料库，并与自学习模型相配合，不断提高现有规则库的信息容量和知识描述准确度。另外，在几类最常见的基本块，包括 np、vp、sp、ap 等的分析实例中，由于并列结构问题引起分析错误也占了很大比例，需要我们针对这种特殊结构的自动分析提出新的解决方案，在更大的处理语境下，把基本块和复合块的边界识别任务有机结合起来。

表 4 不同基本块分析错误的分布百分比

	封闭测试（%）				开放测试（%）			
标记	Case 1	Case 2	Case 3	Case 4	Case 1	Case 2	Case 3	Case 4
np	49.71	27.76	14.64	2.08	54.14	24.94	11.70	1.54
vp	64.97	20.46	6.15	2.05	60.49	21.49	8.67	1.39
mp	83.41	8.53	0.95	4.27	85.71	11.11	0.00	3.17
ap	71.37	19.66	6.84	0.43	72.50	25.00	0.00	2.50
tp	72.53	21.03	1.29	1.29	77.59	15.52	0.00	0.00
sp	40.82	47.62	2.04	8.16	33.33	46.67	3.33	10.00

4．相关研究工作

近年来，汉语块分析方面的研究工作主要采用了基于统计的方法，典型工作包括：1）张昱琪等（2002）使用基于实例学习(MBL)方法，在约 20 万词的汉语 TCT 语料上进行了基本块自动识别实验，采用与本文类似的基本块整体评价方法，最好的 F-measure 为：np 块 93%、vp 块 94%，整体效果 93%。2）孙广路等（2006）使用最大熵马尔可夫模型（MEMM），在 10 万词规模的汉语宾州树库语料（CPTB）和 47 万词规模的人民日报标注语料库(MSRA)上进行了基本块识别实验，采用 CoNLL-2000 的基于词标记的评价方法，最好的 F-measure 分别是：在 CPTB 上，np 块 85%、vp 块 95%，整体效果 93%；在 MSRA 上，np 块 88%、vp 块 96%，整体效果 91%。

由于采用的基本块标注体系、训练测试语料以及最终结果评价方法的不同，这些结果与本文的实验结果并不具有绝对的可比性。其中对分析性能影响最大的是标注体系的选择问题。我们目前的

基本块标注体系与以上研究工作所用的标注体系最大的不同点是引入了粘合式述宾结构的处理。这个改变对基本块自动识别方法处理难度的影响是巨大的。根据我们目前的统计，在 20 万词的新闻语料库中，“ vp-PO ”结构基本块占了所有多词语 vp 块总数的 39%。如果不处理这些“ vp-PO ”块，把它们转化为两个单词语块进行分析，不仅降低了自动分析的处理难度，而且可以大大减少汉语中一些典型歧义结构，如：“ v v n ”、“ v n 的 n ”、“ v n n ”等对基本块自动分析算法的影响，从而将改变最终处理结果中基本 vp 和 np 块的分布格局和相应的分析性能评价结果。

基于以上的分析，我们可以得出这样的初步结论：在分析难度加大的条件下，目前的规则驱动的汉语基本块分析器，在可以得出明确分析结果的无歧义区间内的分析性能（开放测试 F-measure 为 90%左右）已基本达到了目前最好的统计分析器的处理水平。同时，利用规则描述信息，现有分析器还可以确定不同其本块的内部关系标记，开放测试的 F-measure 已达到 82%左右，为进一步进行基本块语义内容的深入分析打下了很好的基础。另外，我们目前的分析器在规则知识描述能力不足的条件下，还保留了约 5%的歧义区间，将目前分析器可能给出的所有分析结果，包括绝大部分的正确分析结果都保留下来。从而为进一步进行更高层次的分析和词汇知识自学习处理提供了灵活有效的基础数据，使后续系统可以根据所掌握的丰富知识和更大的语境信息从中选择最佳的分析结果。这种功能是目前所有的基于统计的分析器所不具备的。

5．结语

本文提出了一种规则驱动的汉语基本块自动分析方法，它具有以下基本特征：1）利用自动习得的融合内部词汇关联和外部语境限制约束知识的分层次、多粒度的基本块规则库驱动基本块分析流程，可以对汉语句子中的各个多词语基本块进行有效识别；2）利用规则置信度信息进行歧义结构自动排歧，可以保留一部分在现有知识库条件下很难判断的复杂歧义现象供后续分析器选择使用，从而提高了分析结果的灵活性和有效性；

目前的初步实验结果基本上达到了我们的预期目标。在今后的研究中，我们将在以下方面进一步改进目前分析器的处理性能：1）融合现有语言资源，开发更大规模的汉语词汇关联库，其中将包含汉语中主要的述宾、述补、定中、状中、并列等词汇关联关系描述对，将它们应用于基本块规则学习和分析应用中，可望提高目前分析器对基本 np 和 vp 块的识别精度；2）在现有基本块分析器基础上，启动大规模语料库的基本块分布知识和词汇关联知识自动学习过程，不断提高目前基本块规则库的信息容量和描述能力。

致谢：本项研究得到了国家自然科学基金资助（项目号：60573185，60520130299）。

参考文献

[Abn91] Steven Abney(1991). “Parsing by Chunks”, In *Robert Berwick, Steven Abney and Carol Tenny (eds.) Principle-Based Parsing, Kluwer Academic Publishers*.

[RM95] Lance A.Ramshaw and Mitchell P.Marcus (1995). “Text Chunking Using Transformation-Based Learning”. In Proceedings of the Third ACL Workshop on Very Large Corpora,Cambridge MA,USA.

[SH06] Guang-Lu Sun, Chang-Ning Huang (2006) “Chinese Chunking Based on Maximum Entropy Markov Models”, *Computational Linguistics and Chinese Language Processing*, 11(2).

[TB00] 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.

[ZQ04] 周强 (2004) 汉语句法树库标注体系. 中文信息学报, 2004, 18(4): 1-8.

[ZQ06] 周强 （2006）“ 汉语基本块描述体系 ”，清华大学计算机系信息科学与技术国家实验室，技术报告。

[ZZ02] Zhang, Y., and Q. Zhou. (2002) “Automatic Identification of Chinese Base Phrases,” *Journal of Chinese Information Processing*, 16(6), pp. 1-8.