汉语句子分块研究

王厚峰
(华中师范大学计算机系，武汉大学计算机学院)
戴大为
(武汉大学计算机学院)

摘要：句法分析的复杂性很大程度上依赖于句子的长度。为了降低其复杂度，本文根据汉语的特点，提出了一种进行句子分块的分析方法。从而，为句法分析提供较短的构件。

关键词：句法分析，词类，分块

Research on Chunking Chinese Sentence

Wang Hou_feng
(Hua Zhong Normal Univ.,430070, Wu Han Univ.,430072)
Dai Da_wei
(Wu Han Univ. 430072)

Abstract: The complexity of syntactic analysis depends on length of sentences to some extent. In this paper, we present a method of chunking Chinese sentence, By which sentence will be divided into short constitutes for syntactic analysis.

Keyword : Syntactic analysis, part of speech, Chunking

一、引言

在自然语言处理中，句法分析是关键的一个步骤。而这一处理是非常困难的。一个重要的原因是句法构件(词或短语)之间的组合非常复杂。开始，人们试图找出其构件之间的结合规律，从而总结出句法规则。但一方面，由于语言现象复杂，要获得完整规则几乎不可能；另一方面，即使得到了相应规则，在进行句法分析时，又可能导致歧义性。

例 1　叫李兵来打球

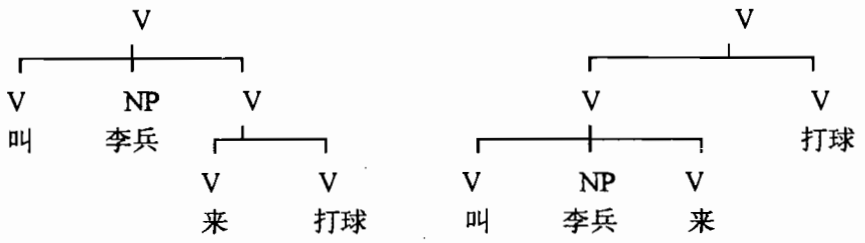
这可以用如下二条规则表示:

V [兼] +NP+V → V

V+V → V

可表示为如下二种结构树。

那么，我们如何从下面的两种结构中选择一种呢?



基于统计的方法可以回避上述的问题。但其方法却又面临着新的困难。首先，基于统计的方法对句法规律进行统计，这需要庞大的语料（树库）；而且，即使得到了统计数据，当分析句法结构时，其复杂性亦会随句子长度而迅速增加。为了使算法趋于实用，可以引入一些明显的规则，并结合统计法，先将句子分块，类似于算法设计中的分而治之法，有望降低分析算法的复杂性。先看一下简单句子的句法结构:

例 2　他叫李明

如果对于任意一个句子，无论它多么复杂，我们都能将之表示成类似例 2 的结构，那么，其分析也可以象简单句子那样进行简化。

例 3　那个小姑娘的言语和行动深深地感动了全村的老少。

可以分块成

A　　B　　C

[那个小姑娘的言语和行动] [深深地感动了] [全村的老少]

当以块为单位后，其结构同例 2 完全一样。

其中，A，B 及 C 由原来的词变成了块。

在块的抽取方面，[LAR95] 曾讨论了基于变换的名词块(NP)和动词块(VP)的分界方法。[LWJ95] 也针对中文作了中文最长名词短语的抽取研究。对于汉语，其句法中除了大部分向心结构(如名词块，动词块)之外，还存在着很多离心结构， 如主谓结构和介宾结构。因此，我们在进行系统设计时，将汉语的块划分为四大类: 名词(体词)块，动词(谓词)块，介词块和主谓块，用虚词驱动及块分界消歧策略进行分块分析。

汉语句子结构的表示一般采用具有层次的树结构，在这种结构中， 叶结点就是词，而内部点便是短语或块。块的划分则简化了中间的许多结合。但其结构仍然具有层次。

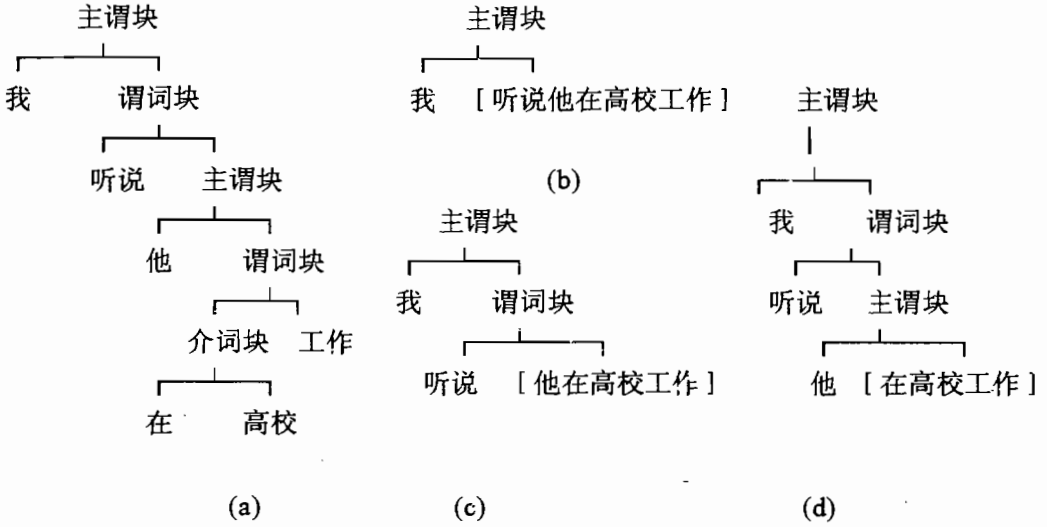
例 4　我听说他在高校工作

其结构可表示为(a)

若用块表示，则可以非常简单，也可以比较细致（如 b,c d)。

于是，块划分的层次直接关系到系统的复杂性。划分得越细， 则越接近于最终完整的句法分析，但系统也越复杂。为了控制复杂性， 我们采用的策略是适可而止，在没有确切

界止的情况下，不再考虑细分。



二 、虚词驱动分块

在汉语中，虚词有着特殊的句法属性:

· 大部分虚词没有明显的词汇意义，只起句法作用。

· 绝大部分虚词在句法结构中的位置是固定的，是粘着的，这有利于句法结构的识别。为了降低句法分析的复杂性，我们假定书面语中的句子是规范的， 每一句子都限定在主谓结构上，主、谓部分可以有修饰部分，通常情况下， 虚词在修饰部分与中心部分起连接作用。我们主要考虑了以下结构块。

(1) 介词及介词块

介词本身并不构成句子成分，它与后面的部分形成介宾结构——介词块，主要充任状语，补语和主语。

主要结构模式为:

介词+体词块——→介词块

但对个别情况，也有例外。

例 5　他为了解决这一问题一直忙到深夜。

为此，在作词类划分时， 我们将介词分为两类: 接体词的介词和接谓词的介词。

除了上述两种情况之外，有些介词也有自身的固定模式:

从+宾语+起

从(自从)+宾语+以来

对于+宾词+来说

在+宾语+看来

对于这些情况，系统需要进行特殊处理。

(2) 结构助词“的”

在汉语中，结构助词“的”与前导部分形成“的”字结构，具有体词功能;

例 6 这支笔是他的

在更多情况下，“的”用来连接前导结构与后继结构，形成体词性的偏正词组。具有如下模型:

X+的+名词块

X 可以是体词块，谓词块。主谓块及介词块。

例 7 (a) 部队的纪律很严

(b) 他是新来的学生

(c) 他写的海报贴在宣传栏上

(d) 他们还没有听过关于美人鱼的故事

由于 X 的取值太灵活，导致了修饰关系很难确定。

例 8 他写字的姿式不正。

由于 X 可以是体词块，谓词块和主谓块，那么，究竟是由

[字] [写字] [他写字]

中的哪个修饰“姿式”产生了歧义。

另外，“的”字结构本身可以作为体词块(例 6)，当“的”字结构接谓词块时，又可以作为修饰部分:

例 9 (a) 这一问题的解决完全消除了他俩之间的隔阂

(b) 不及格的留下来，其余的可以走了。

其中“的”字结构是作为修饰语，还是作为独立的体词块，也产生了歧义。

(3)结构助词“地”

在汉语中，结构助词“地”用于连接前面的修饰语和后面的动词短语。因此，可以表示为:

X+“地”+动词短语

动词短语是谓词块的一种。

X 取值通常可以是体词块，谓词块和主谓块。

例 10 (a) 我们要历史地看待这一问题

(b) 他不停地唠叨那件事

(c) 张教授精神焕发地谈论他的学术论点

这同结构助词“的”一样，X 的定界同样会产生歧义。

(4)结构助词“得”

“得”用于连接谓词块和修饰块，其结构为:

谓词块+“得”+X

其中，x 取谓词块和主谓块

例 11 (a) 雷声把他吓得出不来气

(b) 雷声吓得他出不来气

对于这种结构，歧义性较少。

(5) 前助词“所”带谓词块

这种结构与前面讨论的“的”字结构(例 6)一样，构成非向心的体词块，其结构为:

“所”+谓词块

例 12 所看的书是一般的小说

对于这种结构，大部分与“的”字联用。

(6) 副词与谓词块连接

在汉语中，副词用于修饰形容词和动词，因此，绝大部分副词都修饰谓词块。

例 13 他亲自动手做实验

但有些情况下，副词可以修饰体词块和主谓块。

例 14 (a) 仅张三就有九十公斤

(b) 那天恰巧他在北京

那么，当出现(b)时，“恰巧”是修饰“他”，还是修饰“他在北京”出现了歧义。

(7) 连词连接两个并列结构

在一个句子内，连词连接的两个结构一般是平行的(相同的)。因此，具有模式:

X+连词+X

连接的结果块与 x 块具相同属性，即当 x 为体词块，则上述模式连接的结果亦为体词块。

(8) 其它

除上述情况之外，还有许多其它虚词，如“被”、“给”、“把”，后助“似的”、“一般”等。

三、块分界歧义的消除策略

块可以看成是若干构件的结合体。如何确定哪些构件可以优先结合成块，系统基于如下的步骤进行处理:

(1) 适当的词类标注;

(2) 虚词驱动块结合;

(3) 规则与统计并举的方法。

词类的划分是多方面的，我们在系统设计中，着重考虑到的是句法特征。词类主要取

自于清华大学提出的标注集，根据块分界的句法特征作了适当修改。如果标注系统能对句中的各个词进行准确标注，则在作块构造时，便有了重要的句法导引。例如，在副词中，有少数可以修饰体词；而绝大部分同于修饰谓词块或主谓块，因此，有必要将副词分为二类：可接体词的和一般的。同样，对于非谓形容词可以分为修饰体词的和修饰谓词的。在作词性标注时，不能受某些词常见用法的影响，而是要根据具体应用而定，对于例 9(a) 中的“解决”，在系统中，我们便看成为名词而不是动词，于是，对于“的”字后面的中心语必为体词块。

在使用虚词驱动方法分块时，副词具有最优级，其次是连词，再其次是助词。但考虑到句子结构的灵活性，按上述方法分块并不一定完全正确，因此，在分块时，并不总是严格地作修饰与被修饰之间的划分。

例 15 非常形象而有趣的表演和十分幽默的语言深深吸引住了所有观众。

其中体词块的修饰与被修饰部分比较复杂，我们只将之划分为：

[非常形象而有趣的表演和十分幽默的语言] 体词块

[深深吸引住了所有观众] 谓词块

并不是要弄清“非常”修饰的对象究竟是什么。在作块划划分时，我们既使用了规则方法，又使用了基于统计的方法。其中对于十分明确的结合规律，我们用规则进行描述，在并不完全能够确定的地方，我们用统计方法——相关信息计算；但规则优先。在我们的系统中，除了虚词驱动的规则外，还有关于谓词配对的规则等，在此，不详细讨论。

此外，为了降低后续处理的复杂性，系统总是尽可能使块不要太大，对于例 15 的划分，主语部分就显得过长，可以继续划分成：

[[非常形象而有趣的表演] 体词块和 [十分幽默的语言] 体词块] 体词块

[深深吸引住了所有观众] 谓词块

四、总结

本文系国家自然科学基金支助的课题，该课题旨在获取汉语短语结构规则，正如前言所述，要手工构造其规则几乎是不可能的。因此，我们试图通过语料库来训练。但这又需要庞大的树库，树库的构造也需要相当的工作量。本文所讨论的问题主要是通过分块，自动得到树库，而后，再训练短语规则。由于这一工作正在进行之中，尚未得到最终结果。

参考文献

[LAR95] L.A.Ramshaw & M.P.Marcus Text chunking using Transformation-based learning: Proceeding of the third workshop on very large corpora,1995

[LWJ95] 李文捷，周明等，基于语料库的中文最长名词短语的自动抽取，《计算语言学进展与应用》，陈力为，袁琦主编,1995

[LZY] 李子云，汉语句法规则，安徽教育出版社 1991