Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2018, 7(4), 152-163
Published Online November 2018 in Hans. http://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2018.74018

Hans汉斯

Syntactic Automatic Analysis Based on Chinese Information MMT Model

Fuyi Yang

Anshan Normal University, Anshan Liaoning
Email: yangfuyi@sina.com

Received: Sep. 26th, 2018; accepted: Oct. 12th, 2018; published: Oct. 19th, 2018

Abstract

This paper describes the engineering practice of syntactic automatic analysis of sentences using the Chinese information MMT model. A syntactic analysis expert system that can run online on the Internet is developed. The theories, methods and tools used are introduced in detail. The application of granular computing and semiotics in syntactic analysis is introduced. The structure, implementation mode and system design of expert system for syntax analysis are studied. Finally, examples of sentence analysis are listed. The rationalism based rule method is adopted in the decision-making of research methods. The expert system of syntactic parser is constructed with the theory of algebraic linguistics, and the MMT model of Chinese information is used in the research. The results show that the MMT model based on Chinese is a useful tool to solve the current syntactic analysis problems, which reduced ambiguity and analysis level. The significance of the research results is that it can test grammatical rules, establish sentence grammatical model structure through deep processing of corpus, and expand grammatical knowledge base for deep understanding of natural language. It provides practical tools and models for in-depth study of syntactic and semantic meaning.

Keywords

Chinese Information MMT Model, Grammatical Symbolic Language, Phrase Structure Grammar, Syntactic Analysis, Natural Chunks

基于中文信息MMT模型的句法自动分析

杨福义

鞍山师范学院 辽宁 鞍山
Email: yangfuyi@sina.com

收稿日期：2018年9月26日；录用日期：2018年10月12日；发布日期：2018年10月19日

文章引用: 杨福义. 基于中文信息 MMT 模型的句法自动分析[J]. 人工智能与机器人研究, 2018, 7(4): 152-163.
DOI: 10.12677/airr.2018.74018

摘　要

本文叙述采用中文信息MMT模型对句子进行句法自动分析的工程实践。研制了可在互联网在线运行的句法分析专家系统。对使用的理论、方法和工具作了详细介绍。介绍了粒计算与符号学理论在句法分析中的应用。研究了句法分析专家系统的组成结构、实施方式和系统设计。最后列举了句子分析实例。在研究方法的决策中采用的是基于理性主义的规则方法。运用代数语言学的理论构建句法分析器专家系统，在研究中使用了中文信息MMT模型，研究结果表明，基于中文MMT模型是进一步解决处理当前句法分析问题的可借鉴的手段，减少了歧义和分析层次。其研究成果的意义在于可以检验语法规则，可以通过语料库的深加工而建立句子的语法模型结构，为深层次的自然语言理解，扩充语法知识库，为句法语义的深入研究提供实用的工具和模型。

关键词

中文信息MMT模型，语法符号语言，短语结构语法，句法分析，自然语块

1. 引言

自然语言处理是计算机科学、人工智能、语言学关注计算机和人类语言之间的相互作用的领域，这一领域中产生了大量的人工智能研究成果和产品，是现阶段人工智能领域的研究热点。

句法分析技术指的是依据语法规则来确定句子结构的分析方法[1]。

目前，自然语言处理的最重要的问题是，热点趋于统计方法，陷入缺少严格理论指导的困境。

宗成庆指出了中文信息处理的现状和遇到的问题如下[2]。

“在规范的汉语文本上最好的句法分析性能(短语准确率)也只有 86%左右，而日语和英语的句法分析性能已经超过 90%。”

“近几年来随着国内指标(SCI/SSCI 论文数量、引用次数、高被引论文数等)导向的各种学术评估愈演愈烈，很多研究开始一味地跟踪热点、追逐新潮，只是为了早出成果、快发论文，而最终忘记了解决中文语言理解这一问题的根本目标。这正是我们担忧的关键所在。”

“而当统计方法一统天下之后，对语言学特性和认知规律的研究在自然语言处理领域并没有得到应有的重视。”

“如何针对汉语自身的特点和规律建立专用的模型和算法，恐怕才是最终解决汉语理解问题的正确出路。”

刘安远、崔安颀指出[3]：“深度学习其实是一个没有太严格理论基础的体系”“缺少完善理论”“缺少更为宏观的框架”。

冯志伟指出：“在自然语言处理的研究中，我们不能采取像蜘蛛那样的理性主义方法，单纯依靠规则，也不能采取像蚂蚁那样的经验主义方法，单纯依靠统计，我们应当像蜜蜂那样把理性主义和经验主义‘更紧密地’‘更精纯地’结合起来，推动自然语言处理的发展[4]。”

本文所进行的研究，不仅仅是完成中文信息的句法分析，而是旨在通过语料库的深加工，提取规则，

建立基于规则和逻辑代数理论上的句法分析专家系统，进而实现语义与篇章结构分析，达到建立以汉语文本(包括现代汉语和古代汉语)语料库的深加工而探索汉语语法理论的实用模型方法与工具。

2. 中文信息 MMT 模型与粒计算理论简介

随着自然语言处理应用的日益广泛，特别是对文本类型的处理需求增加，使得句法分析的作用更加突出，它在机器翻译、信息检索与抽取、问答系统、语音识别等研究领域中都有重要的应用价值。

中国学者冯志伟针对乔姆斯基(Chomsky)短语结构语法的弱点和汉语语法的特点，在 20 世纪 80 年代初期提出了“多叉多标记树形图分析法”(Multiple Branched and Multiple Labeled Tree Analysis)，又叫作“中文信息 MMT 模型”[4]。这是对乔姆斯基短语结构语法的重大改进。

中文信息 MMT 模型的根本与核心是向自低向上逐层抽象。用多叉树代替二叉树，采用多标记方法，包括词性标记、语义标记、语用标记等各种语法标记，最终聚合形成系统的语法树。可以从叶节点向上的而逐层进行概念单元的句子组成单位抽象后进行解析。

采用中文信息 MMT 模型，可以解决自然语言理解与中文信息处理的组合爆炸问题。可以解决避开深度计算中的计算复杂度问题而采用类脑处理的计算方式。

“粒计算研究的对象所具有的结构称为粒结构，其组成元素包括粒、层次及分析结构。”“知识结构是粒化和多层次的；用来交流的自然语言也是粒化与多层次的。”“粒存在特定的层次中，他们是该层次岩觉得主体人们在粒计算的不同层次中研究不同类型的粒，这些粒之间是有联系的，同一层次粒与粒之间可以是不相交的关系，也可以是交叠的关系。层次中每一个粒表述了一个特定的粒化观点。同一层次的所有粒形成了对此层的覆盖[5]。”

通俗的讲，就是在中文句法分析中，以一个汉字或词作为一个基本粒为单位来进行计算。粒是我们对现实的抽象，它的目标是建立高效的以用户为中心的对于外界世界的视点，从而支持和帮助我们对周围物理和虚拟世界的感知。人类具有根据具体的任务特性对相关数据和知识抽象或者泛化成不同程度、不同大小的粒的分析判断逻辑运算，以及进一步根据这些粒和粒之间的关系建立数学模型进行求解的能力。

目前，联合国的工作语言以及各国语言的文本都是以字符描述。各国语言所用的文本，都具有基本粒，英文最小的粒度单位是字母，比字母大的单位是语素或词(Word)。中文的最小单位是笔画，比笔画更大的单位是汉字构件，构件组成字。本文以英文的词和中文的字作为基本粒。

把句法分析的过程用粒计算的理论与方法结合中文信息 MMT 模型来实践，可以解决自然语言处理中句法自动分析实用化而遇到的一些问题。

粒计算主要是指以下 3 个方面：

1) 研究信息分类、被分成的块是两两分离的划分还是两两可能有交的模糊分割。

2) 研究分成的信息组块粒度大小、不同大小的粒度层之间的关系。

3) 研究粒度分解与合并的方法。

目前模糊集、粗糙集和商空间理论可以看作是三种不同的粒度计算理论。这三者从思考问题的出发点和解决问题的任务各具特色。但三者有一个共同的特点是在不同的粒度层次上观察问题。

自细视大者不尽，自大视细者不明。

从细小的角度看庞大的东西不可能全面，从巨大的角度看细小的东西不可能真切。

中英文平行语料库对比句法自动分析过程则是把句子从细小的角度的对比向较大角度的过度。

粒度结构：粒度结构给出了一个系统或者一个问题的结构化描述。通过从系统思维、复杂系统理论和层次结构理论(技术)中得到的启发，我们至少需要确定一个粒度网中三个层次的结构：粒的内部结构；

粒集的集体结构；粒度网的层次结构。粒集的集体结构可以看作是全部层次结构中一个层次或者一个粒度视图中的结构。它本身可以看作是粒的内部连接网络。对于同一个系统或者同一个问题，许多解释和描述可能是同时存在的，所以粒度结构需要被模型化为多种层次结构以及在一个层次结构中不同层次。

在句法分析树中，顶层确定了树的根－句子。

而各国语言文本实际构成的句子。除汉语古文只分段而没有标点外，都含有以标点符号分隔的语言单位，杨福义(2018)对自然语块作了定义[6]。语块与语块的关系，可以抽象成句子的各种构成模式。从而可以依靠大规模语料库统计分析句子模式，使句子模型有了实证和统计分析发现规律的依据。

从粒计算的角度分析，字是汉语语言研究客观世界并且传递信息的最重要的基本单位，字的进一步细化可以分解为近似语音单位(字的音素)、字的意义单位(字的义素)、字的构形单位(字的形素)。从而构成汉字的形音义与客观事物的概念映像，建立事物之间的抽象关系。以字作为概念的基本粒，进行粒度更大的抽象集合运算，则有词。可以认为词是字与字关系的集合。

对词的聚合则生成短语。短语是词语词连接的进一步集合。

短语与短语连接以及他们之间的关系，则构成粒计算的进一步抽象。

例如：

把副词短语与动词短语链接构成含副词短语动词短语。DP—>DP + DP

把时间短语与动词短语链接构成含时间短语的动词短语。VP—>TP + VP

把介词短语与动词短语链接构成含介词短语的动词短语。VP—>PP + VP

把数量短语与名词短语链接构成含数量短语的名词短语。NP—>MP + NP

把树形图结构的某一层向上则聚合中文构成自然语言处理的句子合一运算，成为更大粒度的语法单位(暂且称这种单位为长语，或者称为大语块)。

基于中文信息 MMT 模型的一次合一运算就是一次聚合。就是一次由粒度小的语言单位向粒度大的语言单位聚合的过程。

对任意长度句子从基本粒做起，聚集成不能再次聚集的语言单位。则聚合结束。从而可以分析出汉语句子的较大粒度的构造模型。

在以上理论的指导下，使用符号学的方法，把汉语由字作为基本单位组成的文本句子转化为符号语语言，进行运算，开发研制了可在互联网上运行的小型句法分析专家系统。可供使用者实时在线输入句子，进行句法分析。

3. 句法自动分析专家系统的设计

3.1. 专家系统的组成

专家系统第一个重要组成部分是知识库，其中有专家那里得到的关于某个领域里的专门知识。专家系统的第二个组成部分是推理机，它具有推理的能力，即能够根据知识推导出结论，而不是简单的去搜索现成的答案[7]。

句法自动分析专家系统的知识库由两部分组成：一是关于中文信息处理的词汇知识，词汇及其语法语义属性，另一个是语言学专家所描述的现代汉语语法学知识，也就是汉字组词造句的结构规律的知识。现代汉语的理论还在发展中，句法自动分析系统不仅能分析文章中句子的句法结构，而且能够根据标准汉语文本语料库的句法分析提出的问题，进一步完善现代汉语语法的理论体系。

词汇知识由分词词典，词性词典和语义词典组成。

语法知识由语法规则库描述。把专家的语法知识系统化存储，构成计算机可读的语法词典。

推理机的设计，参考 LISP 语言的结构，由一整套的句法自动分析函数构成函数库。由计算机自动按语法规则完成推理运算。中文句法自动分析器的组成如图 1 所示。

3.2. 工作环境

句法分析器的工作环境是互联网人人可交互使用的形式。因此，算法设计采用可以在网络服务器下的 PHP 语言。在大量丰富的 PHP 函数库的基础上，构建自然语言处理专用函数库，实现句法分析器的网络在线运行。

系统输入：

输入是任何有意义的中文文本，可以是一句，一段或一篇文本。这样，就可以自动分析语料库的全部句子。而获得基于语料库的句子分析库、从而形成大规模的树库。

系统输出：

采用类似 LISP 语言表达式作为系统输出，也可以用短语标注或图形方式输出。

LISP 语言是最早和最重要的逻辑性编程语言之一，有以下特点[7]：

1) 主要数据结构是表(符号表达式)。而不是作为算术运算对象的数。

2) 特性表简单，便于进行表处理。

3) 最主要的控制结构是递归，适于过程描述和问题求解。

4) LISP 程序内外一致，全部数据均以表形式表示。

5) 能够产生更复杂的函数和解释程序。

6) 对大多数事物的约束发生在尽可能晚的时刻。

7) 数据和过程都可以表示成表使得程序可能构成一个过程并执行这个过程。

8) 大多数 LISP 系统可以交互方式运行，便于开发各类程序，包括交互程序。

参考 LISP 的特点，结合可在网络服务器运行的 PHP 语言。建立了 PHP 自然语言处理专用函数库。应用于句子语法分析过程的各个阶段。

对于中文句子语法自动分析中的短语结构，采用两种方式输出。一种是以 LISP 语言表达式表示的结果。另一种是以树图表示。图 2 所示是系统对句法分析语句的图形输出。

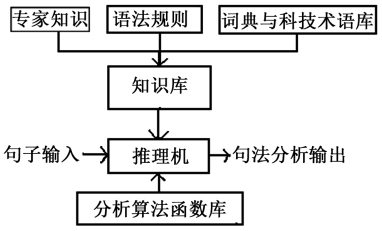

Figure 1. Composition of a Chinese syntactic automatic analyzer
图 1. 中文句法自动分析器的组成

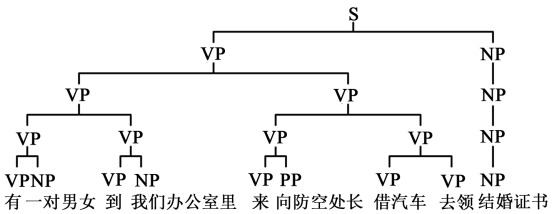

Figure 2. Automatic generation of phrase structure tree diagram by parser
图 2. 句法分析器自动生成短语结构树形图

4. 句法分析方法与工具

4.1. 概述

句法分析“是要根据给定的文法，自动地识别出句子所包含的句法单位和这些句法单位与句法单位之间的关系”。

句法分析的工具使用语法符号语言，把系统获取的汉语文本转换为符号语言表达式。构成符号句子，使用符号句字和汉语文本句子形成二分图模型。

推理机的工作包括两方面内容，一方面是确定语言的语法体系，即对语言中合法句子的语法结构给予形式化的定义；另一方面是句法分析技术，即根据给定的语法体系，自动推导出句子的语法结构，分析句子所包含的句法单位和这些句法单位的关系[2]。

基于语料库的句子分析是依据语料库提供的数据资源，对汉语或英语等进行句子分析并生成报告的过程。

句子分析包括句法功能分析和句法结构分析。有浅层句法分析和含有标点符号的句子全文分析。

语法知识库的设计采用了冯志伟的中文信息 MMT 模型。构成语法词典。

对初级标准现代汉语语料库的全部语块自动分析的结果。对使用的理论、方法、工具作了介绍。对语料库全部句子的分析统计分析。

4.2. 术语及其定义

本文作者定义的术语说明如下：

语法符号：语法词类属性标记符号。

语法符号语言：以有限的语法描述符表述的语法关系符号序列。

单字节语法符号句：用一个 ASCII 码表示的语法符号句子。例如：n 表示名词，v 表示动词，a 表示形容词，d 表示副词等。

双字节语法符号句：用两个 ASCII 码字母表示的语法符号句子。例如：Ns 地名词，Nd 处所名词 Ud 助词“的”，V0 一般动词，Vd 趋向动词，P0 一般介词，R0 介词等。

语料库的数据规模，一般以字词数计算，分为小型，中型，大型，超大型多种。

本文作者以人民教育出版社的 1~6 年级全部课文的自然语块构建了中型自然语块库。对全部记录进行了句子自动分析。

“英语的句子结构是主语部分 + 谓语部分”。汉语句子并不限于这种唯一的模式，各种类型的自由短语都可实现为句子。汉语句式结构的多样性也增加了汉语句子分析的难度[6]。

古汉语文章没有标点符号，依靠读者对汉语字词的虚实掌握句法结构，引入西方语法体系后，构建了和西方语法体系类似的汉语语法体系。在现代汉语的语法教学中，可以应用主语 + 谓语构成句子的简单模式。但对于现代汉语，由于使用标点符号，使得汉语句子结构之间的嵌套性更加突出，难以分析。有的句子内部嵌套子句，子句又进一步嵌套。因此，到目前为止，还没有成套的理论与方法解决现代汉语句子内部各语法单位之间关系的成熟理论与方法。现代汉语的语法理论还在发展中，这种情况下，构建大型语块库，用中文信息处理推动现代汉语语法研究也具有重要的意义。

因此，在大规模汉语语料库的句子分析实践中，探索与解决现代汉语复杂句子的分析理论与实践也就成为自然语言处理技术的关键。

汉语句子结构模式的确定，需要大规模语料库资源数据的验证。也需要语言学家对古代汉语向现代汉语转换发展过程中的语法现象做详细精准的分析。

郑家恒(2010)指出：汉语句子的构造原则与短语的构造原则基本一致。另一个重要特点是各种类型的短语的组成成分又可以是各种类型的短语。这表现出汉语句法成分特有的套叠现象[8]。基于规则的自底向上分析方法可以通过递归获得具有套叠结构的大型自然语块，从而构成长语。

因此，笔者根据大量统计分析的实践，提出汉语语法句子结构的模式是：

名词短语，动词短语，时间短语，介词短语中一个及多个的嵌套组合。

在嵌套与包含中，会出现涉及名词属性的形容词短语，数词短语，数量词短语。涉及名词相互关系的名词短语。

动词短语中会嵌套涉及动词属性的副词短语。

介词短语涉及到事件的地点、位置、地域和空间等。

4.3. 句子分析算法

本文采用中文信息 MMT 模型，构建了语法知识库的规则词典。使用了多叉结构描述语法规则。是基于规则的自底向上的合一运算算法。

多标记的确定，作为汉字，可以分类。安子介把全部汉字的 170 个部首分为 12 大类。根据这个分类体系，可以分析在归一运算过程中。同类汉字构词和不同类汉字构词的理论实证材料。

作为汉字的标记，可以有语法词性标记，语义标记，拼音标记，汉字分类标记等。

笔者在计算机自动句法分析中，采用**语法符号语言**。使用中文信息 MMT 模型多叉树的语法词典和带有词性标记的分词词典，对语料库的文本进行句法自动分析。对句法分析的结果，可以分为两个语块集合，1) 分析成功集合。2) 待处理分析失败语块集合。对待处理部分，进行人工检验，根据系统给出的分析，补充分词词典或补充语法规则，直至全部句子，段落或篇章句法自动分析完成。

按语法规则进行替换规约操作，通过递归最终获得句子的语法结构模式。

形而上者谓之道，形而下者谓之器。本系统把现代汉语的具体词汇，向上规约转换为语法和短语符号，实现汉语句子复杂特征集的运算。

运用符号学的理论构建语法符号句子。对语法符号句子进行递归分析的合一运算，是本文的核心与关键技术，也是根据冯志伟数理语言学的中文信息 MMT 模型理论在语言工程实践中的应用。

4.4. 分析过程与方法

在中文信息 MMT 模型双态原则推导下，冯志伟提出：汉语句子的自动分析，应包括如下步骤：

1) 对输入的汉语句子进行切词，确定单词与单词之间的界限，这就是所谓的自动切词。

2) 在词典中查出句子中各单词的静态特征。这就是所谓的自动标注。

3) 根据语法规则和语义规则检查这些静态的相容性，把静态特征相容的单词结合成词组，并求出词组类型特征。

4) 根据语法规则和语义规则由静态类型和词组类型出发，计算出句法功能特征，并进一步计算语义关系特征和逻辑关系特征[9]。

依据人工智能的知识，构建一系列句法分析的函数，形成句法分析函数库。对于句法分析，可以提供基于网络的在线句子分析。供国内研究单位和学者试用。

自动分词函数。对于汉语切词，最重要的是需要对汉语文本进行自动切词。自动切词有多种方法。目前采用最多的是基于分词词典的分词方法。有正向最大分词法和逆向最大分词法。笔者采用正向最大分词法。

自动标注函数。词性标注，主要依靠教育部语言研究所对大型语料库的分词结果，进行归并整理，

扩充，构成基本实用带有按词性标注桂发标注的分词词典。在分词词典中，对单字词也标注词性。构成基于语法规则的新词及属性的发现。

语法单位合一运算函数，求出短语类型和句子分析底层结构。

递归句法分析函数。使用语法符号语言。对语法符号句递归分析，获得最终结果。并进行自动分析。

4.5. 工具与关键技术

句法分析自动机依靠根据特定任务研制的一整套函数。实现了算法 = 数据 + 程序。基于以上方法，采用中文信息 MMT 模型构建句法分析自动机，可以大批量进行句子分析计算。系统把中文文本转换为语法符号语言，使用语法符号对符号句子进行移进—归约运算，将线性有序小粒度字符串转换为上层大粒度的字符串处理方法，把句法分析解析成层次构造。系统反应速度快，没有计算复杂度问题。可以对大型语料库的资源进行大数据环境下的句子分析计算。获得句子分析报告和句子各种结构类型的统计制图分析。从而使中文信息的句子自动分析过程进入实用化与工程化的阶段。

句法自动分析自动机的关键技术是使用中文 MMT 模型，使用 MMT 模型构建语法规则词典。目前已经实现多叉树的工程应用。对于多标记，已建立多个代码体系，采用平行构建语块分别标注，再按基本单位合一的方法运算，引入科技术语部件知识库，构建分类语义块，为文本自动分类，信息抽取提供可在线操作的使用方式。

5. 句法分析模型与句法分析专家系统

句法自动分析器是一个小型的专家系统。它由知识库和算法构成。

5.1. 句法自动分析知识库

通用汉语句子分析器的实质是一个先进先出自动机。对于任何文章，按顺序输入句法分析器，可以按序列输出句法分析结果。

可以进行多次的递归分析，也可以有多台自动机根据输入的不同阶段，采用不同的语法词典多个自动机进行分析。没有文章长度的限制。

基于规则的中文 MMT 模型句子分析器使用如下资源：

分词词典 6 万，用于进行中文文本的自动分词与词汇属性的自动标注。

语法词典依据中文 MMT 模型规定的语法规则 340 多条，通过语法规则，对句子进行规约合并，获取短语。

5.2. 算法分析程序。

分析器可以通过网络在线运行。对于缺少的词汇与需要补充的语法规则，可以再不修改算法程序的情况下，通过补充与修改知识库的方式随时进行。

句法分析在网络环境生物的运行见图 3。

软件工作环境：使用网络服务器，数据库，XML 语言，PHP 程序设计语言构建系统开发工作环境。对语料库的句子进行句子分析。

在 PHP 语言丰富函数库的基础上构建自然语言处理函数库。

主要有：文本净化与标准化函数，自动分词函数，自动标注函数，句子初级分析函数，句子自底向上递归输出函数等。

基于符号语言的抽象层符号句子处理函数。

使用函数可以对各种规模的文章，段落，句子，语块进行句子自动分析，提出分析报告。

Figure 3. Running view of syntactic parser in Network Environment
图 3. 句法自动分析器在网络环境下的运行视图

采用可视化输出方法。把不同层次的输出，分别用不同彩色的文本输出，方便人工监测自动机的运行。

本系统到目前为止，使用了 340 多条语法规则。其中：语法规则详细界定了名词短语(NP)、动词短语(VP)、形容词短语(AP)、副词短语(DP)、介词短语(PP)和时间短语(TP)。此外，在底层还使用数量词短语(MP)、叹词短语(EP)等。

6. 句子自动分析机的分析实例

6.1. 未登录新词的发现

依据中文 MMT 模型。凡是符合语法规则的短语，可能构成词组、新词和短语。

在系统分词词典和词性标注词典中，没有的词。句子自动分析可以发现：下例的黑熊与紫貂是系统依据 MMT 语法模型自动发现的。

实例如下：

原文：紫貂和黑熊不得不躲进各自的洞里(引自小学 3 年级上学期语文 22 课《美丽的小兴安岭》)。

分词：紫|貂|和|黑|熊|不得不|躲进|各自|的|洞里。

分词标注：紫/a 貂/n 和/c 黑/a 熊/n 不得不/d 躲进/v 各自/r 的/ud 洞里/nl。

语法原文：A0N0C0A0N0D0V0R0UdNl。

语法分块：A0N0|C0|A0N0|D0V0|R0Ud|Nl。

短语属性：(NP(a)(n)) + (CP(c)) + (NP(a)(n)) + (VP(d)(v)) + (AP(r)(u)) + (NP(n))。

短语分词：紫貂|和|黑熊|不得不躲进|各自的|洞里|。

短语标注：紫貂/NP，和/CP，黑熊/NP，不得不躲进/VP，各自的/AP，洞里/NP。

中文输出：S((NP(紫)(貂)) + (CP(和)) + (NP(黑)(熊)) + (VP(不得不)(躲进)) + (AP(各自)(的)) + (NP(洞里)))。

紫貂/NP	和/CP	黑熊/NP	不得不躲进/VP	各自的/AP	洞里/NP
紫貂	和	黑熊	不得不躲进	各自的	洞里
(NP(a)(n))	(CP(c))	(NP(a)(n))	(VP(d)(v))	(AP(r)(u))	(NP(n))

对句子自底向上分析则得到如下分析过程：

原文：紫貂和黑熊不得不躲进各自的洞里

整理原文：紫貂和黑熊不得不躲进各自的洞里

紫 貂 和 黑 熊 不得不 躲进 各自 **的** 洞里

|紫|貂|和|黑|熊|不得不|躲进|各自|的|洞里---1、中文文本分词处理结果。

紫/a 貂/n 和/c 黑/a 熊/n 不得不/d 躲进/v 各自/r 的/ud 洞里/nl---2、词性标注处理结果。

A0N0C0A0N0D0V0R0UdNl---3、属性提取结果(语法符号句)。

A0N0 **C0** A0N0 D0V0 R0Ud Nl

A0N0|C0|A0N0|D0V0|R0Ud|Nl---4、短语语法分词结果。

A0N0 **C0** A0N0 D0V0 R0Ud Nl

(NP(a)(n)) + (CP(c)) + (NP(a)(n)) + (VP(d)(v)) + (AP(r)(u)) + (NP(n))---5、短语属性标注结果。

紫貂|和|黑熊|不得不躲进|各自的|洞里|---6、短语字块显示。

紫貂/NP 和/CP 黑熊/NP 不得不躲进/VP 各自的/AP 洞里/NP---7、短语标注输出串。

继续向上递归处理

紫貂/n 和/c 黑熊/n 不得不躲进/v 各自的/a 洞里/n**

N0C0N0V0A0N0**

N0C0N0 **V0** A0N0 N0C0N0 **V0** A0N0

1) 输出：紫貂和黑熊/NP 不得不躲进/VP 各自的洞里/NP---3

紫貂和黑熊/n 不得不躲进/v 各自的洞里/n**

N0V0N0**

N0 V0N0 **N0** V0N0

2) 输出：紫貂和黑熊/NP 不得不躲进各自的洞里/VP---2

紫貂和黑熊/n 不得不躲进各自的洞里/v**

N0V0**

N0 V0 **N0** V0

最终结果：

紫貂和黑熊/NP 不得不躲进各自的洞里/VP

至此，单句分析结束。

6.2. 基于名著选读的实例

例句：有一对男女到我们办公室里来向防空处长借汽车去领结婚证书(引自“张爱玲文集第一卷——烬余录”)。

在这个句子中出现多个动词多个宾语。动词如下：有、到、来、借，去、领，一共有六个动词。

转换为语法符号语言如下：V0Y0N0V0R0N0NdVdP0N0V0N0V0V0N0N0W0

短语属性：

(VP(v)) + (NP(m)(n)) + (VP(v)) + (NP(r)(n)(n)) + (VP(v)) + (PP(p)(n)) + (VP(v)(n)) + (VP(v)(v)) +

(NP(n)(n)) + (BH(w))

类 LISP 语言的系统输出：S((VP(有)) + (NP(一对)(男女)) + (VP(到)) + (NP(我们)(办公室)(里)) + (VP(来)) + (PP(向)(防空处长)) + (VP(借)(汽车)) + (VP(去)(领)) + (NP(结婚证)(书)) + (BH(。)))

作为自底向上递归为：有一对男女到我们办公室里来向防空处长借汽车去领结婚证书/VP。/BH

6.3. 新闻语料库复杂句子分析(详细过程略)

原文：尼日利亚生活水平极其低下，许多人都指望通过体育成名成家，摆脱贫困，因此使用兴奋剂取得好成绩仍然成为一些尼日利亚运动员的手段之一。

语法符号语言句子：

NsN0N0D0V0W0A0N0D0V0P0N0V0V0W0V0N0W0C0V0N0V0A0N0D0V0N0NsN0UdN0N0W0

首次短语标注结果：

尼日利亚生活水平/NP 极其低下/VP，/BH 许多人/NP 都指望/VP 通过体育/PP 成名成家/VP，/BH 摆脱/VP 贫困/NP，/BH 因此/CP 使用兴奋剂/VP 取得/VP 好成绩/NP 仍然成为/VP 一些/NP 尼日利亚运动员/NP 的/UP 手段之一/NP。

最终句子分析结果：

最终结果：

尼日利亚生活水平/NP 极其低下/VP，/BH 许多人/NP 都指望通过体育成名成家/VP，/BH 摆脱贫困/VP，/BH 因此/CP 使用兴奋剂取得好成绩仍然成为一些尼日利亚运动员的手段之一/VP。/BH

7. 结语与问题讨论

7.1. 结语

通过对句子自动分析的应用实践，证明了：

1) 中文信息 MMT 模型是解决汉语句子复杂特征集的有效方法。

2) 使用语法符号语言对句子自动分析，为基于规则的分析方法提供了工程使用的手段，全部程序可以互联网在线运行。

3) 中文信息 MMT 模型，有很广阔的研究空间，模型分析的结果迫切需要现代汉语语言学语法理论的而进一步解释与完善。

7.2. 存在问题的讨论

1) 虽然句法分析器可以对任意句子进行层次分析，但有最终结果正确而中间层产生歧义问题，在分析的过程中，初始分词标注正确，最终结果也正确，但中间层次有多种解释的问题。

2) 没有着手解决多义词处理。

3) 没有解决兼类词问题。

致　谢

感谢冯志伟老师、胡凤国老师对句法分析器研制的指导和帮助。

参考文献

[1] (美)艾伦(Allen, J.)著. 自然语言理解[M]. 第 2 版. 刘群, 等, 译. 北京: 电子工业出版社, 2005.

[2] 宗成庆. 中文信息处理研究现状分析[J]. 语言战略研究, 2016, 1(6): 19-26.

[3] 刘知远, 崔安颀. 大数据智能: 互联网时代的机器学习和自然语言处理技术[M]. 北京: 电子工业出版社, 2016.
[4] 冯志伟. 自然语言计算机形式分析的理论与方法[M]. 合肥: 中国科技大学出版社, 2017: 819-820.
[5] 苗夺谦, 等, 编著. 粒计算: 过去、现在与展望[M]. 北京: 科学出版社, 2007: 6-7.
[6] 杨福义. 基于双语平行语料库的术语自动抽取[J]. 中国科技术语, 2018(2): 13.
[7] 蔡自兴, 徐光祐. 人工智能及其应用 [M]. 第 2 版. 北京: 清华大学出版社, 1996: 366.
[8] 郑志恒. 智能信息处理-汉语语料库加工技术[M]. 北京: 科学技术出版社, 2010: 168.
[9] 冯志伟, 胡凤国. 数理语言学[M]. 北京: 商务印书馆, 2012: 148.