自然语言的自动分析与生成简介[1]

詹卫东
北京大学中文系 北京大学汉语语言学研究中心
北京大学计算语言学教育部重点实验室

摘要 自然语言的自动分析（理解）和生成构成了自然语言处理研究的全部内容。但这两个直觉上对称的任务在实际中却并没有平等的地位，前者受到更多的关注和讨论，后者的研究则相对薄弱。本文对自然语言自动分析和生成各自的目标、面临的问题，所采用的基本方法等做了概要的介绍，并对如何认识二者之间的关系做了初步的讨论。

关键词 自然语言处理 句法分析 自然语言生成

A Brief Introduction to Natural Language Understanding and Generation

Weidong ZHAN
Dept. of Chinese Language & Literature, Peking University

Abstract: As a subfield of Artificial Intelligence, the aim of Natural Language Processing is to communicate between computers and human by natural language. In order to realizing it, computers should be able to understand natural language and generate natural language automatically as well. This paper gives a very brief introduction on the basic framework of NLU and NLG respectively, including the tasks, challenges and architectures of the two fields. Intuitively, the NLG can be viewed as the inverse process of NLU. The former, however, has received much less attention and research work than the latter. While there are many reasons which have been mentioned to explain why this might be so, this paper figures out a new image to illustrate the relation between them which is somewhat different from the past viewpoints.

Keywords: Natural Language Processing Syntactic Parsing Natural Language Generation

一 引言

自然语言（Natural Language）是人与人之间相互交际的最主要工具。无论是自然口语还是书面语的交际，都包括理解（understanding）和生成（generation）这两个相反的过程。在电子计算机出现之后，人们希望计算机能够模拟人的语言交际能力。对应于交际过程中的理解和生成，计算机的自然语言处理（Natural Language Processing，NLP）当然也就包含了自动分析自然语言（Natural Language Understanding, NLU）和自动生成自然语言（Natural Language Generation, NLG）这两个方面。

分析和生成语言的能力，可以说在我们的日常生活中无处不在，而且人们在学习语文课程中，这两方面的能力也都是最主要的训练内容：比如阅读理解（Reading Comprehension）就是在进行自然语言分析的训练；作文练习（Writing）则是在进行自然语言生成的训练。

[1] 本文的研究工作得到霍英东基金项目“大规模中文树库构建及其在对外汉语教学中的应用”（课题编号：111098），教育部人文社科基地重大项目“大规模中文树库建设及其应用研究”（课题编号：06JJD740001）资助，特此致谢。

一般人往往会觉得阅读理解相对容易，作文练习则相对困难一些。而有过学外语经验的人，大多数也都会有这样的体会，就是理解外语句子相对容易一些，而要自己能说出流利的外语句子或写出通顺的外语文章，则要困难不少。由此似乎可以推测，对人而言，分析自然语言与生成自然语言的关系，可能跟人走上坡路和下坡路一样，虽然方向相反，但这两个过程并不是对等的。不过，由于人类对自身所具有的语言能力的认识还很不够，特别是我们并不清楚人脑处理语言的内在过程和机制的细节到底是怎样的，因此，现在计算机对人的语言分析和生成能力的模拟，还主要是对外在功能的模拟，还无法做到在内在过程和机制层面上的模拟。跟人的经验似乎相反，许多研究人员的看法是，计算机在分析（理解）自然语言方面的难度更大，而在生成自然语言方面则相对容易一些。相应的，在整个 NLP 领域中的表现就是，关于分析的研究要明显多于关于生成的研究（Dale et al. 1998，冯志伟 2010a，b）。

下面，我们就分别对计算机自然语言分析和生成两方面的研究做概要介绍（第二、三节）。在了解了这两个从理论上说应该有密切联系，同时在实践中又几乎是各自独立的领域的基本状况后，我们有可能对现状背后的原因有更深入的认识（第四节）。

二 自然语言的自动分析

2.1 NLU 的任务

自然语言的自动分析是指计算机接收到自然语言文本后，自动将文本的“意思”表示出来的过程。纯粹从形式上看，就是将输入的自然语言文本，变换为**另一种表示形式**输出的过程。对于自动分析而言，其输入为自然语言的句子[2]，这一点是没有争议的，明确的；其输出则是对于输入的自然语言文本的一种解释。而如何来解释自然语言，则有许多不同的看法，这些不同的看法的背后实际上都联系着不同的语法理论。大致说来，当代为数众多的语法理论可以做如下粗略的划分：

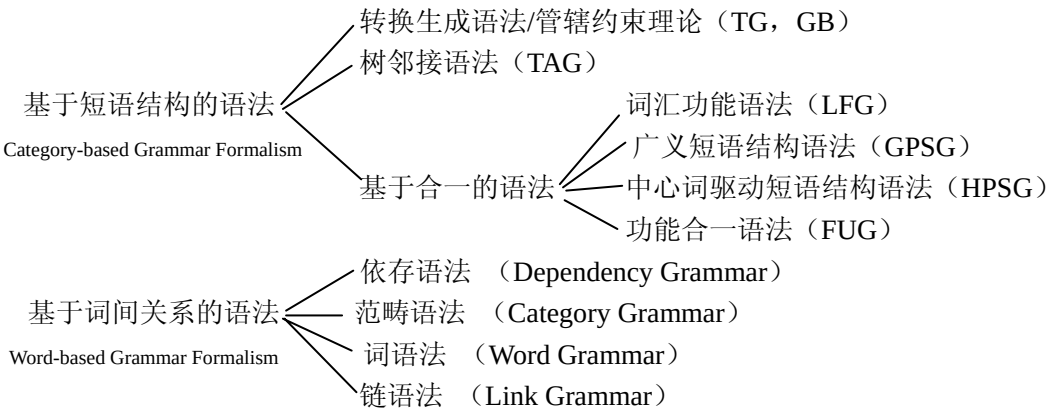


图 1：当代语法理论的划分

有关这些语法理论的详细介绍，读者可以参看冯志伟（2010a）。这里想要重点说明的是，“基于短语结构的语法”体系在语言中的“句子”和“词”两级单位之间，假设了“短语结构”（或“范畴”category）这个中间层次的概念，即一个句子可以分解为若干个短语（词组）范畴，句子的意思就是由短语的意思组合而来的；而“基于词间关系的语法”体系，则认为句子的意思可以直接由词和词之间的关系来描述，不需要假设短语（词组）范畴。

上述这些语法理论可以看作是对于同一个问题的不同回答：即什么叫做“理解”了一个

[2] 当然分析的对象也可以是自然语言中比句子小或大的单位（如词组，段落等），但目前已研究如何分析自然语言的句子最为常见。

句子的意思？这个问题我们平时不大会去关注，因为理解自然语言句子的意思，对于人来说，再平常不过了，但是，当我们要细究，到底什么叫做“理解”句子的意思，以及“理解”是一个怎么样的过程时，我们就会发现，事情并不简单，甚至可以说这是一个非常难以清楚地回答的、关于人类智能的本质为何的追问。限于篇幅，本文无法展开去讨论这个问题，下面仅通过举例的方式，来说明“基于短语结构的语法”和“基于词间关系的语法”分别是如何“理解”一个句子的意思的（以下的表述忽略了具体语法理论的诸多技术细节，只关心其中关于句子意思的表示最基本的部分）。

例 1：郭德纲打人弟子被刑拘　（新闻标题）

“基于短语结构的语法”理论认为，当人看到例 1 这个句子时，之所以能“理解”它的意思，是因为人脑能够把例 1 的“线性字符串”（linear string）形式转化为如图 2 所示的“树结构”（tree）形式[3]，并且还可以在句法树结构的基础上得到如图 3 所示的以“特征结构”（feature structure）框图形式表达的句子的深层句法关系和语义关系等。

“基于词间关系的语法”理论则认为，人之所以“理解”例 1 的意思，是因为人脑可以把例 1 中的词语组织成图 4 所示的“依存树”（dependency tree）形式。在依存树上，带箭头的连线表示两个词之间有“依存——被依存”的关系，如图 4 中“弟子”依存于“刑拘”。或者说，后者是中心词，支配着从属于它的词“弟子”。依存树上的各节点之间的关系可以进一步表示为如图 5 所示的依存关系表，即两两词语之间的语义关系描述。句中所有的词间关系被完全列举出来，就意味着“理解”了一个句子的意思。

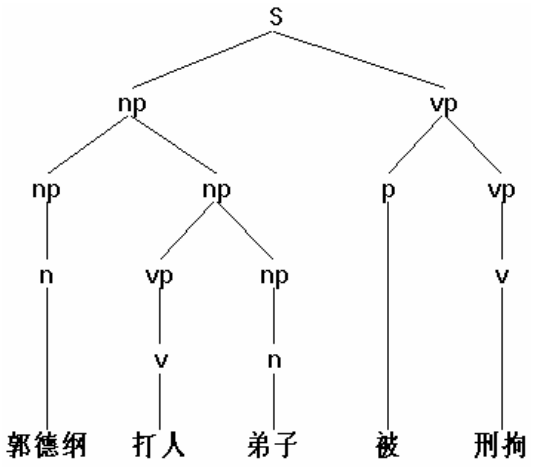


图 2：例 1 的短语结构分析树

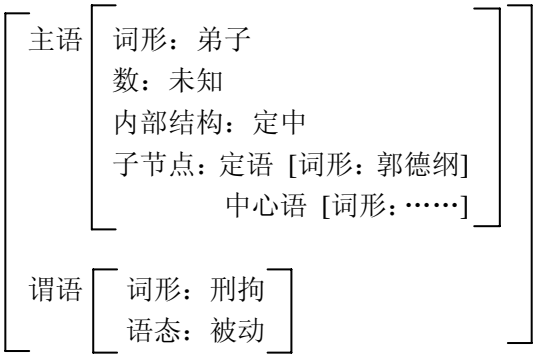


图 3：例 1 的特征结构分析框图

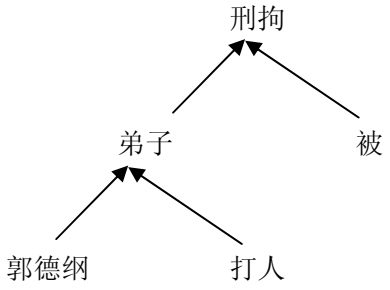


图 4：例 1 的依存分析树

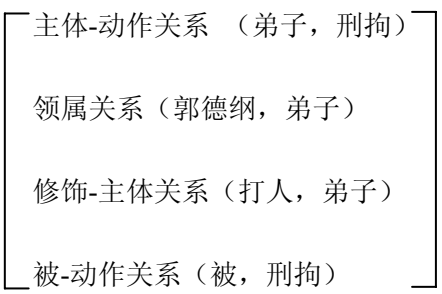


图 5：例 1 的依存关系列表

[3] 图中的符号 S，np，vp，p，……等等的含义见 2.2 小节表 1 中的说明。

通过上述关于语法理论的分类以及简单的示例，我们可以知道，计算机自动分析的输出目标（任务）在不同的语法理论背景下，形式上是有很大差异的。相比较来说，以短语结构树形式给出的分析结果所包含的信息量一般要多于依存树的分析结果。

2.2 NLU 的基本方法

上一节说明了计算机进行自然语言自动分析时输入和输出的基本情况。接下来的问题就是，中间过程是什么样的？即如何从线性的句子开始，分析得到句子的“结构”？下面以短语结构的分析为例（仍以例 1 为分析对象），来概要说明一般的做法。

概括来说，自动句法分析（Parsing）由下面两部分工作组成：

（1）构建语言模型：这部分工作的理论基础主要是上下文无关文法（context free grammar），即以文法的形式给出语言成分之间可能存在的组合模式的描述。

（2）搜索符合文法要求的语言结构树：依据给定的文法，判断一个具体句子中的语言成分（词语和词组）符合该文法的哪些组合模式。

为了分析例 1，我们需要准备如表 1 所示的语言模型：

表 1：语言模型示例

序号	规则	示例	说明
1.	S → np vp	张三 打虎	S（主谓结构型句子）由 np（名词性成分）和 vp（动词性成分）组成
2.	S → S S	学生 打架 老师 受罚	两个 S 可以并列组合为一个更大的 S
3.	np → np np	张三 那两个徒弟	两个 np 可以组合为一个更大的 np
4.	np → vp np	打虎 英雄	一个 vp 可以修饰一个 np 组合为一个更大的 np
5.	vp → vp np	伤害 他人	一个 vp 可以支配一个 vp 组合为一个更大的 vp
6.	vp → p vp	被 刑拘	p（介词）可以跟 vp 组合为一个更大的 vp
7.	vp → v	刑拘	v（动词）可以提升为 vp
8.	np → n	弟子	n（名词）可以提升为 np
9.	v → 打人 | 刑拘 | 伤害 | ……		这类规则箭头的右边是自然语言中的字符（词语），通常把这类规则记录在词典中。（“|”表示“或”）
10.	n → 人 | 弟子 | 郭德纲 | ……		
11.	p → 被 | 把 | 从 | ……		

有了语言模型后，对一个句子进行分析，就是顺序扫描该句子中的基本单位（词语），通过查规则（语言模型），去找到一棵（或多棵）句法结构树，使得这棵（些）句法结构树，能够“解释”当前的句子，即树上每个节点所对应的词语片段，都是自然语言中合乎语法（grammatical）的表达式。如果一棵树上存在一个节点，它对应的词语片段是非法（ungrammatical）的表达式，那么这棵句法树就是错误的结果（因而也不应该并且不会分析出这样的结果来）。

因为汉语词语之间没有空格，计算机在接收到输入句子后，第一步工作是要把其中的词识别出来，即把汉字字符序列转换为词序列（词间加空格分隔开），并给每个词赋词性标记：

例 1： 汉字序列： 郭德纲打人弟子被刑拘

词序列： 郭德纲 打人 弟子 被 刑拘

词性序列： 郭德纲/n 打人/v 弟子/n 被/p 刑拘/v

在词性序列的基础上，计算机就可以按照语言模型中规则的约束，从左至右，逐词扫描，

进行分析，在众多的候选树结构中，搜索得到“正确”的结构，来解释句子的意思。以下就是计算机根据表 1 的语言模型分析例 1 的过程：

① 扫描到第一个词“郭德纲”

图 6

② 扫描到第二个词“打人”

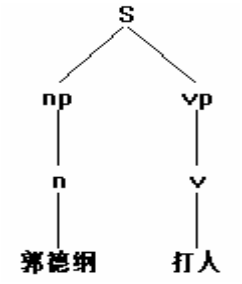


图 7

③ 扫描到第三个词“弟子”

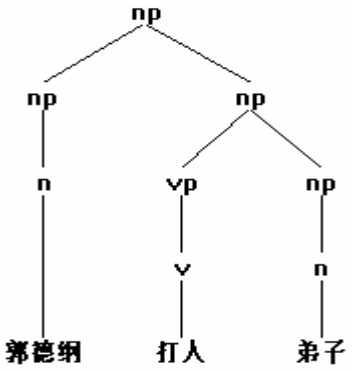


图 8

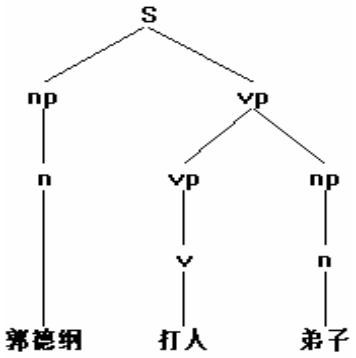


图 9

④ 扫描到第四个词“被”

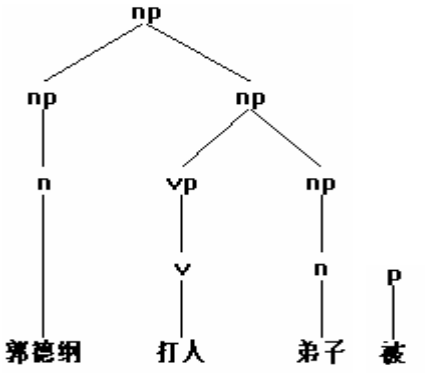


图 10

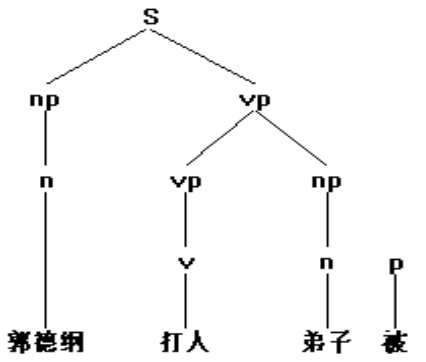


图 11

⑤ 扫描到第五个词“刑拘”

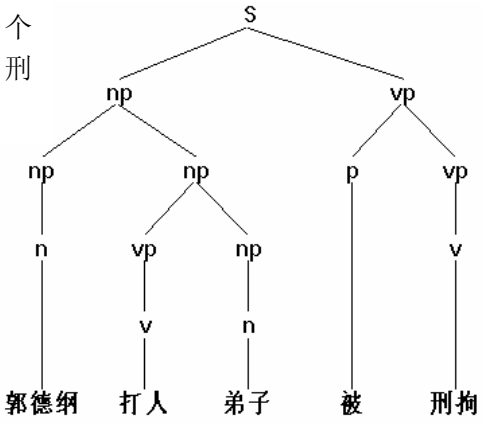


图 12

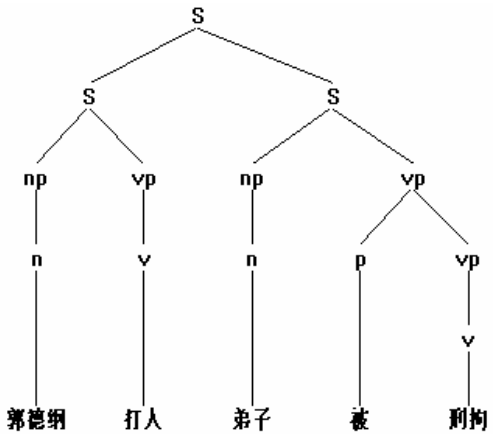


图 13

在上面的分析过程中，有的中间树结构是错误的（比如第 3 步得到的图 11），这样的结构在后续的分析中有可能根据语言模型的约束会被淘汰掉，于是在最终的句法结构树中，就不会出现这样错误的局部分析结果（图 12，13）。

2.3 NLU 面临的困难

从上述简要的介绍不难看出，自然语言的线性词序列事实上包含了大量的“隐含”信息，计算机自动分析（理解）的过程，相当于是在将这些“隐含”信息“显性化”。对计算机而言，这也是在“无中生有”，因而是一件非常困难的工作。难点可以概括为两个方面：

（一）自然语言各层级单位存在大量歧义

自然语言的一个显著特点是，各级语言单位上都存在明显的多义性，即一个形式对应多个意义。比如词汇层次上的歧义例子：

例 2　a 他真是**奇怪**极了，谁会半夜三更在这儿散步呢？

　　b 他真是**奇怪**极了，半夜三更在这儿散步。

同样一个“奇怪”，例 2a 中，“奇怪”是他对某人某事感到奇怪；例 2b 中，是别人认为“他”很奇怪。人来理解这两个“奇怪”的差异没有多大困难，但计算机却几乎无从下手。

再比如结构层次上的歧义。上面举的例 1 就是一个有结构歧义的例子，2.2 小节模拟计算机分析的结果也确实得到了两棵句法树，即对应着这句话可以有两种解释：图 12 的解释是“郭德纲的弟子打人，弟子被刑拘”；图 13 的解释是“郭德纲打人，郭德纲的弟子被刑拘”。人因为了解现实中事情的状况，因而知道例 1 应该按照图 12 的结构方式解读。但计算机不知道现实情况，只能给出图 12，图 13 两种解释[4]。

（二）自然语言的开放性

跟人为规定的语言本质上是封闭的有显著的不同，自然语言是开放的。这又表现在两方面，一是每天的生活中都可能出现不少新词新语，即在原来设定的语言模型中不存在的新单位会源源不断地冒出来[5]。比如下面这些新词：

表 2：新词示例

	新形新义（或旧形新义）
名词性	胶囊公寓 蚁族 房奴 宅青 剩女 裸官 山寨 粉丝 草根
动词性	闪婚 裸考 裸捐 拍砖 团购 人肉 灌水 晒
形容词性	雷 囧 汗 纠结 菜 酷 晕

另外，还有一些新的词语组合，超出了原先的语言模型所覆盖的范围，比如：时下流行的“被 XX”表达形式：被就业　被小康　被幸福　被涨工资　被下等人　被开心　被留学　被增长　被自愿　被出名　被失踪　被潜规则　被抄袭 ……　这些组合或者完全突破了以往“被”所修饰对象的语法限制，或者是表达了跟以往组合不同的语义，后一种情况如“被抄袭”，并不是原先的“某人被别人抄袭”的意思，而是“某人在没有抄袭的情况下，被大家说成是抄袭了别人”。

“被 XX”这种情况实际上是自然语言开放性更为本质和更为常见的表现，只是因为“新词新语”比较容易引起人的注意，给一般人的感觉是新词新语更能代表自然语言的开放性。像“被 XX”这种新的组合方式，其本质是自然语言中常常省略成分而造成的，而自然语言中大量的省略几乎无时无刻不在突破原有语言模型的限制。这里不妨再看两个普通例子：

---

[4] 如果考虑北京方言的表达习惯，例 1 还有一种解读，就是“郭德纲打别人的弟子，郭德纲被刑拘”。北京话中，“人”可以指别人、人家。

[5] 与此同时也会有原来存在的单位在“死去”，即所谓旧词的消亡。像古汉语中的文言词现在很多已经不用了。

例 3：轨道飞行器将被高温（最高可超过一千二百六十摄氏度）包围。

例 4：天桥被质疑豆腐渣 高跟鞋一踩路面就烂

例 3 中的“最高”实际上指的是“最高温度”；例 4 中“烂”的主体不清楚，到底是“高跟鞋”烂，还是“路面”烂？因为“烂”前面省略了成分，就造成意思不清楚，但人有能力根据上下文，根据现实情况中的信息来确定准确的意思，计算机就难以利用更大范围的上下文信息，更难以利用外部现实世界中的信息来进行分析了。这些都是由于人们在实际的语言使用中大量出现符号的“省略”（并进而造成符号的“转指”）所造成的分析上的困难。

以上指出的困难集中在自然语言自身（或者说是关于自然语言建模的困难），而在分析的搜索阶段，也存在着困难。这属于分析技术层面的问题，即如何提高搜索效率，比如 Earley 算法采用点规则的形式记住已经分析过的部分，从而将分析过程中的“回溯”改为“并行”，即利用动态规划（Dynamic Programming）的策略来使得分析“不走回头路”，提高分析效率，LR 算法则采用预读（Look-ahead）策略，对规则进行预分析，事先把不可能出现的组合屏蔽掉。这样就可以在分析过程中“少走弯路”，从而提高分析效率（Grune & Jacobs, 1990）。此外，通过对大规模标注语料的统计，可以对一个成分参与到不同组合模式中的分布进行概率评估，也可以在分析过程中，尽早做出“剪枝”决策，把那些不大可能的组合筛选出去。在这方面，人们已经提出过不少基于概率的句法分析模型来对句法分析技术加以改进（Charniak, 1997；Collins, 2003；宗成庆，2008）。

三 自然语言的自动生成

3.1 NLG 的分类

如果说自然语言自动分析的特点是“输入”明确，而“输出”不清楚的话，自然语言自动生成的特点就刚好相反，是“输出”明确，但“输入”是什么不清楚。

自然语言生成的目标无疑就是产生出符合人阅读（听话）习惯的，流畅的自然语言句子（篇章）。有关自然语言生成的一些有代表性的定义都阐明了这一点。比如下面两个定义就都明确指出了自然语言自动生成的输出（目标）是用自然语言表示的文本：

> 定义 1：自然语言生成是为了达到特定的交际目标而有目的地构建自然语言文本的过程。（McDonald，1987）[6]
>
> 定义 2：自然语言生成是人工智能和计算语言学的子领域，研究主旨是让计算机系统从非语言的信息表示产生出以自然语言表示的可理解的文本。（Reiter & Dale，1997）[7]

值得注意的是，定义 1 回避了 NLG 的输入是什么，而定义 2 则明确界定了 NLG 的输入是“信息的非自然语言表示形式”（non-linguistic representation of information）。事实上，这两个定义出现的时间相距十年，也基本上反映了研究人员对 NLG 研究认识上的发展。跟这种变化对应的是，NLG 系统从自然语言处理总任务中一个非独立的部分，到现在已经逐步成为一个独立的研究领域。从这个角度说，NLG 系统可以分为两大类：

一类是早期的完全依附性的，作为“机器翻译、人机对话、问答系统”等 NLP 应用中的一个后端输出模块。这类 NLG 模块的输入也是自然语言。计算机在分析了自然语言输入之后得到一个中间语言表示，再从中间语言转换生成出另一种形式的自然语言表示，就完成了“生成”。很显然，这种意义上的“生成”，在很大程度上是依赖于“分析”的，并不独立。

另一类是后期的独立的 NLG 系统。这类 NLG 系统的输入不是自然语言，而是非自然语言表示的信息，比如数据库中的数据，人的交际意图表示，甚至图像信息等。从数据库中

---

[6] 转引自 Dale et al.（1998）。

[7] 转引自 Dale et al.（1998）。

的数据生成自然语言文本是现代 NLG 系统最典型的应用模式，比如天气预报数据，股票市场价格波动数据，城市环境监测报告数据，吸烟者调查问卷数据，等等，现代 NLG 系统可以根据用户感兴趣的信息需求描述，对数据库中的数据进行筛选和整合，在事先给定的规则约束下，生成出符合自然语言表达习惯的文本（过程详见下文 3.2）。

跟上述从输入形式的不同所作出的 NLG 分类大致上对应，NLG 系统的生成机制（实现方法）也有不同的情况：

早期的 NLG 系统通常采用所谓的“罐装文本”（canned text）作为生成自然语言的基础，这可以称之为基于“模板”的生成（Template-based generation）。比如要生成报告飞机航班信息的自然语言文本，就可以用一个简单的“模板”：

[航班号] [起飞时间] 由 [出发地]起飞，预计 [到达时间] 到达 [目的地]。

模板中 [ ]中的内容由数据库中的数据填充后，就可以生成一个自然语言的句子输出。这种生成模式甚至可以追溯到 1960 年代的人机对话系统 Eliza。该系统让计算机模拟心理医生跟病人聊天，其工作原理是用正则表达式匹配用户输入的句子的模式，然后按照事先给定的模板，将模式中的某些片段替换成另外的片段，生成句子反馈给用户（Jurafsky & Martin, 2000）。据报道，这种简单地基于模板匹配的“理解——生成”系统，也可以“欺骗”一部分人，让这些人以为他们是真的在跟一个未曾谋面的人聊天，而不是在跟机器“对话”。

上世纪八、九十年代以后，NLG 系统的开发者意识到基于模板的生成机制的局限性[8]，开始探索更为系统和独立的自然语言生成方法，试图让生成系统能够适应不同领域的文本生成的需要，能更灵活地表现自然语言文本的连贯性，更有针对性地反映用户对文本生成的需求，等等（Reiter & Dale，2000）。由此发展出的 NLG 系统，可以称之为基于“规则”的生成（Rule-based generation）。跟“模板”相比，主要以树结构形式表达的“规则”[9]更具系统性，覆盖的自然语言现象更广，可以在更多的词语序列和结构中选择更符合用户需求的表达形式。

近年来，跟自然语言句法分析中越来越地依赖语料库建立概率统计模型类似，研究人员也开始尝试在基于规则的生成系统中引入统计方法，探索自然语言文本生成中句子的优化方法（比如用 N-gram 模型来优化生成句子的流利度），这类 NLG 系统，可以说是在已有的基于规则的系统基础上，又进了一步。有人称之为“可训练的生成”（Trainable generation），有兴趣的读者可以参看 Ratnaparkhi（2000），Mairesse（2008, 2010）的研究工作。

3.2 NLG 系统的构成

下面以基于规则的 NLG 系统为例，来说明目前典型的主流 NLG 系统的体系结构。在这方面，Reiter & Dale（2000）给出了被 NLG 领域称之为标准流程的所谓经典管道模型（Typical Pipeline Modal）：即一个 NLG 系统应顺序包含三个部分：（1）篇章规划（Document Planning）；（2）句子规划（Sentence planning / Microplanning）；（3）表层实现（Surface Realization）。

3.2.1 篇章规划

篇章规划是 NLG 的第一个阶段，一般由四个部分组成：

A. 知识源（The Knowledge Source）：即领域数据库和知识库，其中存放了用于生成文

[8] 一般认为基于模板的生成系统的优缺点都是很明显的。优点是概念简单，易于实现，领域确定，文本质量可保证；缺点则是不通用，生成文本的语言风格单一，难以扩展。

[9] 就表达机制而言，用于“生成”的规则跟用于“理解/分析”的规则并没有什么两样。一般都已上下文无关文法作为表达的主要工具。

本所需的信息。比如生成天气预报的知识源就是气象预报部门提供的天气数据。一个生成餐馆介绍文本的知识源就是有关餐馆的各项属性的数据库，以及有关餐饮知识的领域知识，等等。

B. 交际意图（The Communicative Goal）：即关于生成文本的用途或目的描述。这类描述的具体内容通常是人们对文本的表达功能类型的划分，比如“叙述”“比较”“说明”“论证”等等。

C. 用户模型（The User Model）：即关于用户的信息描述，包括用户所面临的任务需求，用户的背景，专业水平，倾向性（偏好）等等。通过了解用户的背景（已知）信息，NLG 系统可以给特定用户生成出更具针对性（即提供更适合的新信息）的文本。

D. 篇章历史（The Discourse History）：即关于用户和 NLG 系统先前互动（交际）的历史记录。 篇章历史有助于正确生成文本中的指称表达形式，同时也对篇章规划中如何安排新的内容有帮助。

篇章规划阶段的结果（输出）是文本结构树，树上描述了一个文本组织信息的先后顺序，以及各部分信息之间的结构关系。比如，要生成一个关于“理想的餐馆”的描述的文本，就需要在篇章规划阶段事先给出一个有关证明某个餐馆是理想的餐馆的文本的结构（如图 14 所示）。一个“证明”结构主要由“观点”和“证据”两部分组成，限于篇幅，图 14 只给出了正面的证据，没有提及反面的证据（反证）。在这个结构图中，“观点”是“核心”成分，“证据”是依附于“核心”成分的“从属”成分，“证据”内部（其下位子节点）元素之间的关系则是平等的列举关系（list relation），即我们可以通过对“厨艺、食品质量、服务、……”等的评价（描述）来支持一个餐馆是理想的就餐地点的观点。

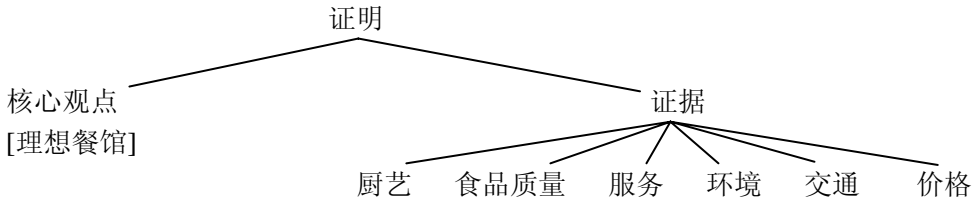


图 14：“证明”类文本的结构示例

篇章规划的结果要进一步跟自然语言文本中的表达形式关联起来，就需要引入句子层级的规划，将篇章结构跟具体句子的表达实现关联（映射）起来。在这之间，还有一些过渡性的描述，比如“证明”结构树对应的句法结构模板为：

证明（证据 X，观点 Y） → 句法结构模板：因为 X，所以 Y

这个句法结构模板，既可以视作是句子规划的内容，也可以视作是篇章规划阶段的任务，属于过渡地带的规则。

3.2.2 句子规划

句子规划（也称作“微规划”）是 NLG 的第二个阶段。这个阶段要完成的任务有三个。而且这三个任务一般是有先后顺序的：

A. 选词（Lexicalization）：即根据文本结构树中叶子节点的描述，选择合适的词语和句法结构表达式来实现信息传递。比如汉语中用于评价的词语，就可以在“好，很好，不错，差、……”等中间进行选择。

B. 整合（Aggregation）：即把选出的词和结构序列整合为自然语言的句子构造（序列）。

C. 生成实体的指称表达式（Referring Expression Generation）：即选择文本中的实体（指

称对象）的恰当的指称形式，比如是否用代词来指代，是定指（英语中用“the、this”等限定）还是不定指（英语中用“a、some”等限定），等等。

上面 B 和 C 两个任务的目标都是追求生成的自然语言文本更为自然。比如用户要生成关于“全聚德”这个餐馆的描述。从餐饮业数据库中获取的信息是：烤鸭很好，交通不方便。那么，这两个子句在整合时，NLG 系统就应该在两个子句之间加入关联成分，生成完整句子形式：“（全聚德）烤鸭很好，但交通不方便”。

3.2.3 表层实现

表层实现是 NLG 的第三个阶段。这个阶段是把第二阶段生成的文本中间表示，包括整个篇章结构，句子和词组的抽象结构进行实例化，成为符合自然语言表达习惯的真正的自然文本形式，即由空格分隔的单词序列，由标点符号分隔的句子（小句）序列，句首单词首字母采用大写形式，标题中各个单词的首字母大写，以及文本中含必要的注释[10]，等等。

值得注意的是，这个阶段的任务相对于前两个阶段，其重要性相对较低，一方面，对于不同的自然语言语种，其表层形式的呈现方式有较大差异，因而不同语言在表层实现上，需求差别较大，比如汉语缺乏形态变化，因而表层实现中就不涉及到词语变形的处理，在表层实现阶段的实际任务也就比较轻；另一方面，有的研究人员指出，那些基于语言学规则开发的表层实现模块（如 Penman，KPML，SURGE 等）尽管强调了通用性，但在实际应用中还是带来一些缺点。比如：出于语言学动机追求表层实现模块的通用性，并不能很好地适应特定领域知识本体的个性需求；对于小型的任务单一的生成系统来说，表层实现模块的复杂性不仅显得没有必要，同时还会带来效率的降低。由此，研究人员提出了跟浅层句法分析类似的想法，在自然语言生成系统中采用“浅层生成”（shallow generation）的策略，即使用相对简单的基于模板的规则，来生成受限领域的文本，比如 TEMSIS 项目中生成城市环境报告的多语文本（Busemann & Horacek, 1998），就没有将一般意义上的表层实现模块加入到 NLG 系统中，以牺牲语言学意义上的表层形式规则的覆盖率和通用性为代价，来换取快速开发 NLG 系统的好处。

以上扼要说明了 NLG 系统的不同类型和主流 NLG 系统的构成情况。跟 NLU 一样，NLG 系统也面临着自然语言固有的歧义和开放性等一系列问题。只不过问题的表现稍有不同，不是一个形式对应多个意义解释的问题，而是为表达同一个意思，如何在多个表达形式中选择一个在当前环境中最合适的形式的问题。此外，句子的形式整合，指代成分的选择，也都是 NLG 系统面临的特殊困难。以汉语述补结构的自动生成为例，汉语中可以说“我洗了衣服”“衣服变干净了”，也可以把这两个表达形式合成述补结构和“把“字句的形式来表达：“我把衣服洗干净了”。从 NLG 系统的角度来看，就是如何把两个句子整合成为一个句子，以及在何种条件下选择分析型表达方式，在何种条件下选择综合型表达方式的问题。在这方面还需要做大量的研究工作。此外，如何评估两个句子之间的连贯性，也是 NLG 系统需要面对的语言学难题。尽管研究人员已经提出了 RST 理论，CT 理论等不少评价篇章连贯性强弱的方法（Jurafsky & Martin, 2000），但离实用的要求还有很大的距离。限于篇幅，这里就不对此做展开讨论了。

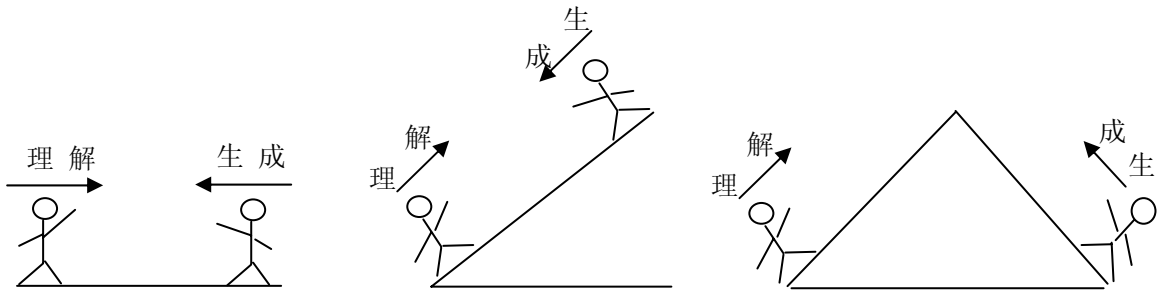
四 结语

通过上文的概要分析，不难看到，自然语言自动分析和生成并不是简单的互为逆过程的关系。二者在输入和输出，基本流程，系统实现框架等方面，还是存在较多明显差异的。

---

[10] 比如在有显示格式控制需要的文本（如 HTML 文件）中加入格式控制标记。

Dale et al.（1998）曾引用著名的机器翻译学者 Yorick Wilks 的话说，“如果自然语言理解面对的问题从某种程度上好比是从一数到无穷大，那自然语言生成面对的问题就是从无穷大数到一”[11]。对于这个比喻性的说法，可能有不同的理解，但有一点应该是可以肯定的，那就是，自然语言的自动分析，尽管很难，但至少其输入是清楚的，因而是可以着手去做的工作，而自然语言的自动生成，其输出（终点）是清楚的，但起点却很难确定，因而给人一种无从下手的感觉。或许，正是因为这样的一种情况，严格意义上的，独立的“自然语言生成”的研究在 NLP 领域一直属于少数派。本文在宏观层面较为全面地考察了自然语言自动分析和生成的概貌，我们认为，对于二者的关系，或许可以形象地对应为如下三个图景：



（I） （II） （III）

图（I）是完全基于直觉的理解，把自然语言自动分析和生成看作是以相对的方向在同一个水平面上走同一条路；图（II）则是在斜坡上相对而行，一个是上坡路，一个是下坡路，因而给人的感觉是难度有差异；图（III）则猜测理解和生成或许并不是同一条路上的相向而行，而是从不同的方向努力登上同一个高峰，需要克服的是同等难度的问题。这个高峰就是：人脑到底是如何把“意义”和“形式”对应起来的？

参考文献：

Ehud Reiter & Robert Dale, 2000, *Building Natural Language Generation System*, Cambridge University Press.

Dick Grune & Ceriel Jacobs, 1990, Parsing Techniques: A Practical Guide. ELLIS HORWOOD LIMITED.

Michael Collins, 2003, Head-Driven Statistical Models for Natural Language Parsing. In Computational Linguistics, Vol. 29, No. 4, Pages: 589 – 637.

Eugene Charniak, 1997, Statistical parsing with a context-free grammar and word statistics, In Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI 1997).

R. Dale, B. D. Eugenio & D. Scott, 1998, Introduction to the special issue on natural language generation. Computational Linguistics, Vol.24, No.3. pp 345-353.

A. Ratnaparkhi, 2000, Trainable methods for surface natural language generation, In ANLPC 6–NAACL 1.

François Mairesse & Marilyn Walker, 2008, Trainable generation of big-five personality styles through data-driven parameter estimation, In Proceedings of the 46th ACL.

François Mairesse, M. Gasic, F. Jurcicek, S. Keizer, B. Thomson, K. Yu and S. Young., 2010, Phrase-based Statistical Language Generation using Graphical Models and Active Learning,

[11] Researchers in NLG face the unique problem of deciding what to generate from: Yorick Wilks is credited with pointing out that, while the problem of natural language understanding is somewhat like counting from one to infinity, researchers in natural language generation face the problem of counting from infinity to one.

In Proceedings of the 48th ACL.

Stephan Busemann , Helmut Horacek, 1998, A Flexible Shallow Approach to Text Generation, In Proceedings of the 9th International Workshop on Natural Language Generation.

Daniel Jurafsky & James H. Martin, 2000, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Prentice-Hall Inc..

刘挺、马金山，2009，汉语自动句法分析的理论与方法，《当代语言学》2009 年第 2 期。

宗成庆，2008，《统计自然语言处理》，清华大学出版社。

冯志伟，2010a，《自然语言处理的形式模型》，中国科学技术大学出版社。

冯志伟，2010b，《自然语言生成系统的建造》导读，北京大学出版社。

詹卫东，2000，《面向中文信息处理的现代汉语短语结构规则研究》，清华大学出版社。

孙宏林、俞士汶，2000，浅层句法分析方法概述，《当代语言学》2000 年第 2 期。

张建华、陈家骏，2006，自然语言生成综述，《计算机应用研究》2006 年第 8 期。

张冬茉、李锦乾、姚天昉，1998，汉语自然语言生成的句子结构优化，《计算机工程》1998 年第 7 期。

郭忠伟、徐延勇、周献中，2003，基于 Schema 和 RST 的自然语言生成混合规划方法《计算机工程》2003 年第 6 期。