文章编号：1003-0077(2018)12-0031-10

基于中文 AMR 语料库的非投影结构研究

闻 媛[1]，宋 丽[1]，吴泰中[2]，李 斌[1]，周俊生[2]，曲维光[2,3]

(1. 南京师范大学 文学院，江苏 南京 210097；
2. 南京师范大学 计算机科学与技术学院，江苏 南京 210023；
3. 闽江学院 福建省信息处理与智能控制重点实验室，福建 福州 350121)

摘 要：非投影结构是指依存树上的词语节点与原句中的词语序列出现错位的现象，对于句法分析器的影响较大，在语言理论上也有较大研究价值。在世界多种语言的依存树或图库上，都发现了含有非投影结构的句子，并对比展开了相关研究。而汉语的非投影结构尚未得到重视，语料库构建过程中也因遵循了投影性原则而缺乏对非投影结构的标注。该文基于概念对齐版的中文 AMR 语料库，在 10 149 句语料上统计出带有非投影结构的句子比例为 31.62%，其三种主要类型为模态词提升、话题化和成分分离，并提出了相应的自动分析方案，以提高中文 AMR 自动分析效果。

关键词：抽象语义表示；概念对齐；非投影；语义分析；中文信息处理

中图分类号：TP391 **文献标识码**：A

Research on Non-projective Structure Based on the Chinese Abstract Meaning Representation Corpus

WEN Yuan[1], SONG Li[1], WU Taizhong[2], LI Bin[1], ZHOU Junsheng[2], QU Weiguang[2,3]

(1. School of Chinese Language and Literature, Nanjing Normal University, Nanjing, Jiangsu 210097, China;
2. School of Computer Science and Technology, Nanjing Normal University, Nanjing, Jiangsu 210023, China;
3. Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University, Fuzhou, Fujian 350121, China)

Abstract: The non-projective structure refers to the phenomenon that the word nodes on the dependency tree are misplaced with different word sequence in the original sentence. It has not been discussed in Chinese, following only the projection principle in the construction of Chinese dependency corpus. In this paper, we construct a Chinese abstract meaning representation (AMR) corpus of 10 149 sentences, in which 31.62% sentences have non-projective structures. Then we distinguish the three main types of the non-projective structures, modal words, topicalization and the component separation. Finally, we provide the solutions for the structures in the AMR parsing.

Keywords: abstract meaning representation; concept-to-word alignment; non-projective; semantic parsing; Chinese information processing

0 引言

近年来，随着依存语法的研究在自然语言处理中的比重逐步增大，句法依存树库[1]、语义依存图库的建设也开始覆盖越来越多的语言[2]。在诸多语言的依存语料库上都发现了一定数量的非投影结构(non-projective structure)。非投影结构是指依存树上的词语节点与原句中的词语序列出现的错位结构(图 1)。在国际上，非投影结构引发了语言学领域的分析和讨论[3]，也有研究对自动分析算法进行了改进[4]。目前国内对依存语法研究较少，在语料库构建时也大都遵循了投影性原则，这使得非投影结构在汉语里是否存在、有哪些类型，成为难以解答

收稿日期：2018-09-29 **定稿日期**：2018-10-29

基金项目：国家社会科学基金(18BYY127)

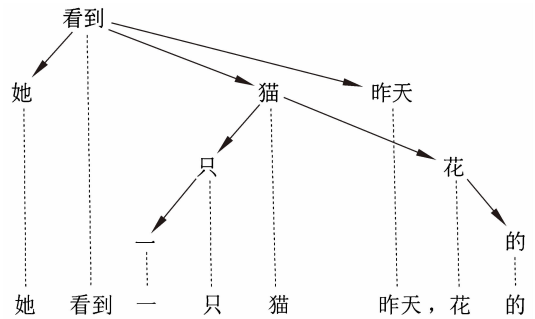
的问题，对非投影结构的自动分析更是无从谈起。

本文针对非投影现象展开了系统的讨论，回顾其在语言学理论和句法语义资源建设过程中从被排斥到认可的过程，分析了在依存树中的非投影现象和转换生成语法理论的关系。为了寻找和分析汉语中的非投影结构，我们使用了新的语义表示方法——抽象语义表示(abstract meaning representation，AMR)。这种方法脱胎于依存语法，引入了超越树结构的图结构来表示句子语义，增加了概念增删修改机制，语义表示能力强[5]。但由于缺少和原句词语对齐的信息，无法直接使用抽象语义表示发现非投影结构。我们使用“概念—词语”对齐的中文AMR语料库[6]，统计出非投影的具体类型和比例，并为中文句法语义自动分析提出相应对策。

全文结构如下：第1节回顾和梳理非投影结构的研究历史和现状；第2节介绍对齐版中文抽象语义库的基本情况；第3节展示我们基于该语料库得到的汉语非投影结构占比情况，并进行分类分析和理论探讨；第4节是结论和未来工作。

1 非投影结构的研究历史和现状

非投影结构是依存语法中存在的一种特殊现象，特指依存树上的节点垂直投影到句子上出现的交叉现象。如图1中的句子“她看到一只猫昨天，花的”在依存树上的节点向原句中的词语做投影时，就会出现“昨天”和“猫—花”的交叉。这种包含非投影结构的句子在传统的语言学理论中往往被作为有问题的句子，或者被解释为生成语法理论的移位(movement)现象，没有引起足够的重视。但是后来发现这种句子在捷克语等形态丰富、语序自由的语言中出现较多，这引发了理论探讨、资源建设，乃至对分析算法的讨论。



图中带箭头的实线表示依存关系，虚线表示投影关系。

图1 含非投影结构的依存树示例

非投影结构是根据依存语法的树结构发现的，早期的依存语法对非投影结构是持忽视和排斥态度的，但是在后来的语料分析中，发现这一结构是真实存在的，并且在越来越多的语言材料中得到验证。于是非投影现象才逐渐得到关注，进而出现了对不同语言非投影结构的专门研究。国际上对非投影现象的研究大致可以分为忽视期、发现期和发展期三个阶段。

1.1 忽视期

法国的Tesnière提出依存语法理论时[7]，采用普通的多叉树来描述句子的结构，没有论及非投影问题。之后Ihm和Lecerf提出了投影结构[8]。美国的Hays进一步指出，图2中实线部分表示的是依存关系，低位置的节点依存于高位置的节点[9]。与依存树上的节点用虚线连接的，是最小句法单位(minimal syntactic unit)，且这些最小句法单位是有序的。当句子的依存树被准确地分析出来后，依存树上的依存关系一般不会交叉，这种特性就是“投影性(projective)”，它与直接成分理论(immediate-constituent theory)中的成分的非断续性(non-discontinuity)很相似。此后，罗马尼亚的Marcus对投影性的原则又进行了详细规定，正式提出了投影原则[10]，为树结构对应到句子词语的线性序列提供了理论基础。

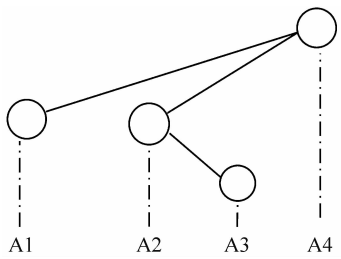


图2 Hays对投影性结构的定义

Robinson更为系统地提出依存语法中关于依存关系的四条公理：①一个句子只有一个独立的成分；②句子的其他成分都从属于某一成分；③任何一个成分都不能依存于两个或两个以上的成分；④如果成分A直接从属于成分B，而成分C在句子中位于A和B之间，那么，成分C或者从属于A，或者从属于B，或者从属于A和B之间的某一成分[11]。现在看来，这四条公理相当于将依存树的形式约束为单根(single rooted)、连通(connective)、无环(acyclic)和投影(projective)，从而保证句子的依存分析结果是一棵单根投影树。

在依存树库建设的早期，遵循了投影性原则，忽

视和回避了非投影结构。将句子的结构限制在一棵投影树上，有助于计算机的自动分析和处理，却不够尊重语言事实。随着依存树库的建设，在标注形态复杂、语序自由的语言时，非投影结构占有相当比例，无法再被忽视了。

1.2　发现期

从 20 世纪 70 年代开始，非投影现象在捷克语中逐步受到关注[12-15]，这些研究多为传统的语言学分析，主要以捷克语为单一研究对象。在布拉格树库[16]的建设过程中就没有遵循投影性原则，而允许非投影结构的出现。发现 Hajičová 等统计了布拉格树库的7 308 句捷克语，共有 23.2%的句子含有非投影结构，产生原因包括模态词提升、量化、名词配价及控制结构等，还为计算机处理部分类型的非投影结构提出了相应的解决方案，包括把话题化成分移动到原始位置，对特定谓词进行标注等[17]。

1.3　发展期

随着更多语言的依存树库的建设，非投影结构在多种语言中的普遍存在逐渐得到认可。Mannem 和 Ambati 均发现印地语中非投影结构占有一定比例，并归纳出成对连接词、小句补语和关系子句三种类别[18-19]。此外，许多语言的依存树库中都存在非投影结构[1,3,20]，但从语言结构的角度进行详细分析的研究则相对缺乏。表 1 总结了 Zeman[20]的数据，给出了 29 种语言的依存语料库中含有非投影弧的比例，即造成非投影的那些弧（词语关系）占到所有弧的比例。

表 1　Zeman 给出的 29 种语言的非投影弧比例

语言	非投影弧的比例/%	语言	非投影弧的比例/%	语言	非投影弧的比例/%
阿拉伯语	0.37	芬兰语	0.51	葡萄牙语	1.31
巴斯克语	1.27	德语	2.33	罗马尼亚语	0.00
孟加拉语	1.08	希腊语	1.17	俄语	0.83
保加利亚语	0.38	古希腊语	19.58	斯洛文尼亚语	1.92
加泰罗尼亚语	0.00	印地语	1.12	西班牙语	0.00
捷克语	1.91	匈牙利语	2.90	瑞典语	0.98
丹麦语	0.99	意大利语	0.46	泰米尔语	0.16
荷兰语	5.41	日语	1.10	泰卢固语	0.23
英语	0.33	拉丁语	7.61	土耳其语	5.33
爱沙尼亚语	0.07	波斯语	1.77		

可以看到，这 29 种语言中大都存在非投影现象，特别是语序自由的古希腊语，其比例接近 20%。只有西班牙语、罗马尼亚语、加泰罗尼亚语三种语言没有统计到非投影结构，主要是由于这三个依存树库的构建遵循了投影性原则。而根据 Havelka[3]对于 12 种采用非投影原则标注的依存树库的统计结果，西班牙语中的含有非投影结构的句子比例为 1.72%（表 2）。

表 2　Havelka 给出的 12 种语言的非投影句子比例

语言	非投影弧的比例/%	语言	非投影弧的比例/%	语言	非投影弧的比例/%
西班牙语	1.72	阿拉伯语	11.16	斯洛文尼亚语	22.16
日语	5.29	土耳其语	11.6	捷克语	23.15
保加利亚语	5.38	丹麦语	15.63	德语	27.75
瑞典语	9.77	葡萄牙语	18.94	荷兰语	36.44

这些数据表明，非投影结构在多种语言的树库中都普遍存在。传统的句子依存自动分析算法，也都是基于投影树的，自然无法处理这种结构。McDonald 则抛开投影原则，引入了针对有向图的最小生成树算法来分析含有非投影的句子[4]。而随着学界对于非投影和论元共享现象的承认，以图结构取代了树结构，发展出句法依存图和语义依存图[2]，以及包含了概念增删机制的抽象语义表示[5]。虽然图结构包含了非投影树结构，但是图结构主要还是由论元共享、指代问题造成的。把非投影结构表示为树结构，能够体现出语言中的错序现象，仍然是学界的重要研究对象。

对于汉语依存树库来说，目前已有的资源，都有意或无意地遵循了投影原则，如 CoNLL 评测中使用的汉语依存树库，是按照投影原则从短语结构树库转换而来的，无法从中统计出非投影结构。郑丽娟等[21]基于哈尔滨工业大学的依存图库报告了汉语中的非投射现象，但讨论的是超越投影树结构的图结构。李斌等[6]在中文抽象语义库的 7 000 句语料上，初步介绍了非投影结构的比例和类型，但没有介绍非投影结构的研究历史、语言学意义和对自动分析的影响。

本文基于更大规模的 10 149 句中文抽象语义库，探究汉语非投影结构的存在情况，并对汉语非投影结构进行分类，探索汉语非投影结构的特点，并为自动分析处理非投影结构提供一些对策。

2 对齐版中文 AMR 语料库

抽象语义表示(abstract meaning representation)是一种将句子语义抽象为一个单根有向无环图的整句句子语义表示方法,拥有增删修改概念和语义关系的较强表示能力[5],是目前最充分的句子语义表示方法。其主要思想是将句子中的实词(如名词、动词、形容词等)作为概念节点,用 45 种语义关系[如 arg0(原型施事)、arg1(原型受事)、quant(数量)等]作为弧,从而形成表示句子语义的图结构。

不过,AMR 忽视概念和词语的对齐信息,即忽略图 1 和图 3 中虚线表示的对应关系,使得人们无法在 AMR 语料库上提取非投影结构。李斌等提出了将概念和词语对齐的方法,构建了中文 AMR 语料库[6],使得我们能够考察汉语中的非投影现象。图 3 给出了具体实例,左侧是利用词语的下标来锁定词与概念的关系,如 x2 表示第 2 个词对应的概念"看到-01"①;右侧则是将其绘制为依存树结构的可视化结果,能清楚地显示出"昨天"的虚线和"猫—花"的关系存在交叉,是非投影结构。

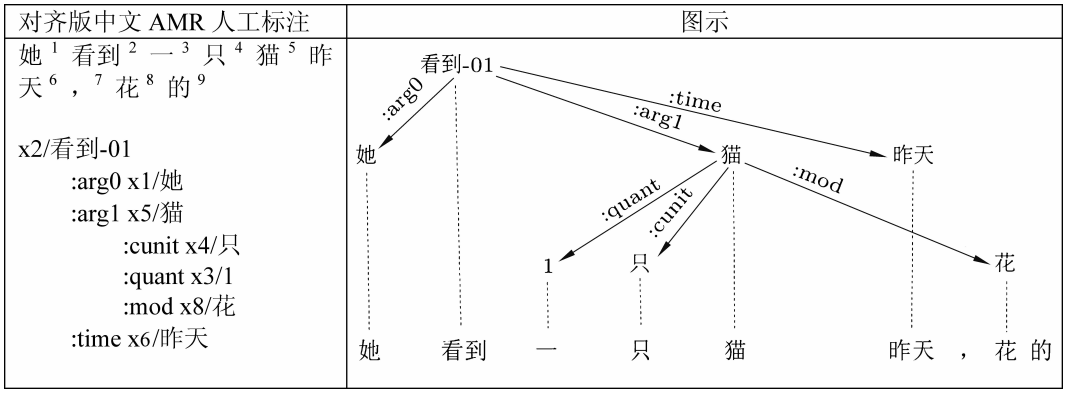


图 3 概念对齐的抽象语义表示实例

本文选取了宾州中文树库 CTB 8.0 语料(以下简称 CTB)中的网络媒体语料,共 10 149 句②,按照概念对齐的方式,标注形成中文 AMR 语料库。在随机抽样的 500 句语料上,双人标注一致性达到 0.83 的 Smatch 值[22],与英文 AMR 的标注一致率基本相当。谓词义项及角色框架参考的是中文命题库(CPB)的谓词框架词典[23]。该词典是从 CPB 标注语料中抽取出来的,含有每个谓词在不同义项下的语义角色框架,共收录了 24 510 个中文谓词(包括动词、形容词等)的 26 650 个义项的不同语义角色框架。这部词典较好地覆盖了 CTB 语料。少量未覆盖到的谓词的语义角色则根据标注规范从 AMR 规定的语义关系中补充。

3 汉语非投影结构类型及比例统计

语料标注完成后,我们根据非投影规则自动提取出所有的非投影结构。在中文 AMR 语料库的 10 149 个句子中,有 3 208 个句子含有非投影结构(非投影树),比例为 31.62%。从弧的比例来看,一共有 193 955 条弧,造成非投影的弧有 3 358 条,占 1.73%,说明非投影结构在汉语中也是较为常见的。其次,和其他语言一样,汉语的非投影结构也是由许多具体的语言现象导致的,如模态词提升、话题化、成分分离等。此外,复句中两个小句成分的分离也可能导致非投影结构。表 3 给出了非投影结构的详细分类和比例,比例之和超过 1,是因为分子按弧、分母按句子计数,每个句子可能含有多处非投影现象。这样统计方便观察出有多少句子出现了非投影结构。

可以看到,在所有的非投影结构类型中,模态词的提升占比最高(52.37%),超过一半;其次是成分分离(28.49%)、话题化(13.34%)以及一般移位(5.14%)。下面我们来逐一说明。

3.1 模态词提升

模态词(modal word)提升是中文 AMR 语义结构中非投影比例最高的一种类型,这种非投影类型

① 01 表示"看到"的第一个义项。

② 选取的原始语料共 10 325 句,其中 176 句存在断句错误、句子意义错乱或句子格式错误,未予标注。

表 3　对齐版中文 AMR 语料中非投影结构类别

序号	非投影类别			数量	比例(%)	
1	模态词提升			1 680	52.37	52.37
2	话题化	连谓结构下的成分前置	小句宾语前置	212	6.61	13.34
3			小句主语前置	68	2.12	
4			小句同位语前置	9	0.28	
5			小句定语前置	8	0.25	
6			小句其他成分前置	64	2.00	
7		从属关系分离		38	1.18	
8		数量结构后置		13	0.41	
9		整体/部分关系分离		16	0.50	
10	成分分离	复句中的小句分离	小句前件/后件内部分离	3	0.09	28.49
11			主体感受插入	7	0.22	
12		成对结构分离	一般分离	902	28.12	
13			动词拷贝结构	2	0.06	
14	一般移位	副词(状语)前置		67	2.09	5.14
15		定语后置		17	0.53	
16		同位语移位		38	1.18	
18		数量结构中的副词移位		43	1.34	
18	其他			171	5.33	5.33
合计				3 358	104.67	

也存在于捷克语[17]等其他语言的依存语料库中。在中文 AMR 中产生此现象的原因是我们将模态词进行了提升处理，即将模态词作为谓词的上层节点。

这类模态词包括“可能”“也许”“似乎”“可以”等。下面以“大多数人可以做到”这个句子为例进行分析，如图 4 所示。

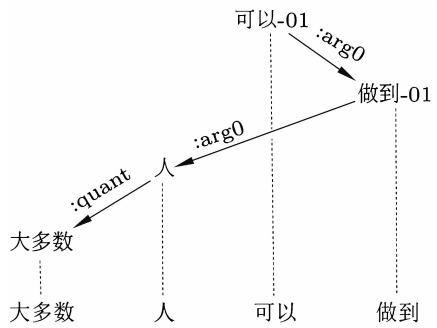


图 4　模态词提升的非投影结构示例

在这个句子中，“可以”作为句子的最上层节点。根据谓词库，“可以”的第一个义项是“可以-01”。这个义项中有一个论元 arg0，表示被允许的事件内容，“做到”作为“可以”的 arg0。“人”是“做到”的 arg0，表示施事主体。“大多数”则表示“人”的数量成分，用 quant 表示“人”和“大多数”的关系。

从图 4 可以看出，由于“可以”位于上层，所以“可以”的投影弧与“人”和“做到”之间的弧有交叉，形成了非投影结构。而传统的句法语义分析是将模态词依附于谓词的，所以不会产生这种非投影结构。

3.2　话题化

话题化指的是将句子中某些成分提前，语用上起到将该成分作为句子关注焦点的作用，在生成语法中研究较多。捷克语中也存在话题化导致的非投影结构[17]，印地语的非投影结构中也存在 15.3%的话题化[19]，说明话题化导致非投影是跨语言的共性。

话题化导致的非投影结构又分为连谓结构下的成分前置、从属关系分离、数量结构后置及整体/部分关系分离四种子类。篇幅限制，下面仅就连谓结构下的成分前置、整体/部分关系分离进行较为详细

的举例分析。

(1) 连谓结构下的成分前置

成分前置的情形是较为典型的“话题化”(topicalization)现象。通过分析,我们发现简单句的成分前置一般是不会造成非投影结构的,而谓词较多的嵌套句中的成分前置才更容易造成非投影结构。当一个小句中有多个谓词(广义的连谓结构)时,这些谓词各自有一套论元。这些论元在一个句子中的排列就容易出现错序情况。当某个小句中处于语序较后位置的谓词的论元发生了前时,就容易形成非投影结构。例如“必然导致对此案做出不公正判决”(图 5)。

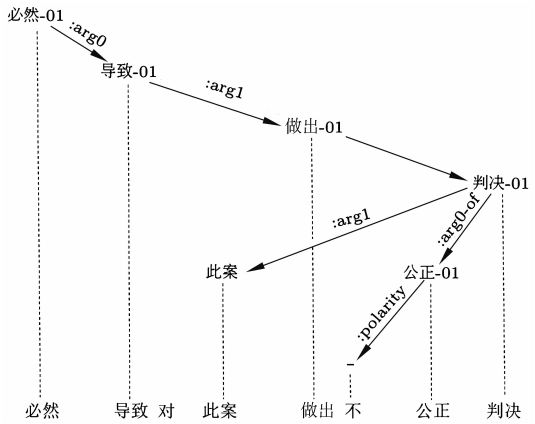


图 5 连谓结构下论元前置的非投影结构示例

在这个句子中,“必然”是最上层节点,“导致对此案做出不公正判决”是“必然”下面的子事件,“导致”及其下层所有节点充当“必然”的 arg0。“对此案做出不公正判决”是“导致”的 arg1。而“对此案不公正判决”则是“做出”的 arg1,表示“做出”的行为事件。“此案”是“判决”的 arg1,“判决”是“公正”的 arg0。这里用了一个反关系“arg0-of”,目的是为了保证有向图的单根性,polarity(极性)为-,表示否定。

从图 5 的非投影结构的可视化表示中可以看到,“判决”与“此案”之间的 arg1 关系与“做出”的投影线有了交叉。这种交叉正是由于判决的 arg1,即句法层面上谓词“判决”的论元“此案”前置所导致的,这种前置由介词“对”引导。

(2) 整体/部分关系分离

整体/部分(part-of)关系往往由两个概念构成,如果这两个概念在句子中被谓词分开了,可能会造成非投影结构,例如“活熊取胆残忍无比”(图 6)。

可以看到,这个句子中的“熊”与“胆”之间有整

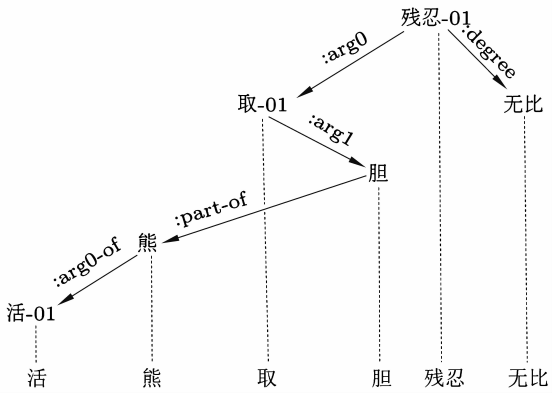


图 6 整体/部分关系分离的非投影结构示例

体/部分(part-of)关系,但是由于强调这个行为的残忍性,所以在表面词序上将“活熊”提到了整个句子的最前面,最终导致了“胆”和“熊”之间的整体/部分关系的分离。从可视化结果可以看到,“熊”和“胆”之间的整体/部分关系(part-of)与“取”的投影线发生了交叉,导致了非投影结构。

类似的话题化现象还有从属关系(poss)的分离,如“给儿子补身体”,“身体”从属于“儿子”;数量结构后置,如“苹果有五个”。

3.3 成分分离

成分分离又分为由复句关系的小句拆分导致的非投影结构和一般成对结构的分离。

(1) 复句关系的小句拆分

复句关系的小句拆分又分为前件和后件的分离,以及主体感受插入两类。例如“如果国家不及时采取措施,我觉得会给国家带来经济危机。”(图 7)

图 7 中圈出来的部分表示的是中文 AMR 中对复句结构(discourse relation)处理时添加的“condition”概念节点。在这个句子中,“觉得”是整个句子的最上层节点,“我”是“觉得”的 arg0,即感受主体,条件复句“如果国家不采取措施,会给国家带来经济危机”是“觉得”的 arg1,即内容。“国家不采取措施”和“会给国家带来经济危机”分别是条件复句的前件和后件。从图 7 不难看到,“觉得”的插入使得条件句的前件和后件被割断开来,形成了交叉。

(2) 成对结构的分离

成对结构的分离往往导致树结构上有一个节点对应表面词序中的多个词的情况,这种情况没有造成投影边的交叉,但是破坏了正常的投影结构。如“法官以事实为依据”(图 8)。

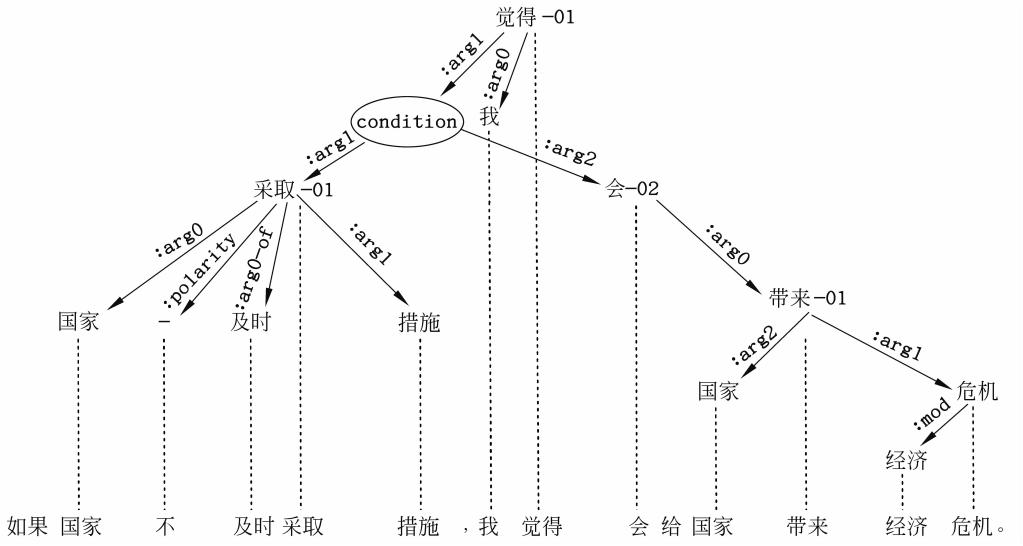


图 7　复句关系中间插入主体感受的非投影结构示例

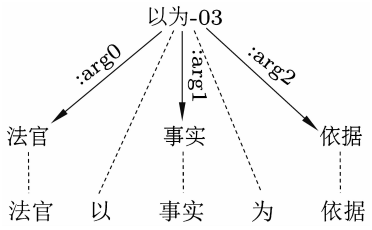


图 8　一般的成对结构分离的非投影结构示例

在这个句子里，“以……为”按照 AMR 的要求被合并为一个概念“以为-03”，是句子的核心，处于最上层结构。“法官”是“以为”的 arg0，表示感受主体；“事实”是“以为”的 arg1，表示“以为”的对象；“依据”是“以为”的 arg2，表示“以为”的结果。从可视化的中文 AMR 语义结构可以看到，由于“以为”在表面词序上的分离，导致了其被“事实”隔断，不是节点与词语一一对应的投影结构。当然，这种类型不一定算作是非投影树结构，也可以直接作为图结构的一种类型。

3.4　一般移位

除此以外，一些普通的移位（movement），也会导致非投影结构的产生，主要包括状语、定语、同位语及其他介词结构的移位。下面以同位语的移位为例，如“我们在这儿等你，地下车库”。

这个句子中，状语“地下车库”发生了移位，其 AMR 语义结构表达的一般语序是“我们在地下车库这儿等你”，“地下车库”的后置导致了非投影结构（图 9）。

3.5　非投影结构的理论探讨与处理对策

从上面四种非投影结构的示例，我们可以看出基于概念对齐的抽象语义表示能够清晰地刻画出汉语中的非投影结构。在传统的基于投影原则的依存树上，是无法找到这些非投影结构的。即使是基于图结构的依存图，如果不从语义的角度来描写，也很难找出这么多真实的用例。对齐版 AMR 更真实地刻画了句子的语义结构，能够表示出“活熊取胆”等非投影结构。

（1）理论探讨

在非投影结构中，模态词提升占的比例较高，主要源于 AMR 标注体系的处理方式。在传统的句法依存标注中，模态词一般都依附于谓词。在图 5 的例子中，如果“必然”依附于“导致”，就不会形成非投影结构了。但是在比较新的生成语法和依存语法的研究中，模态词的位置一般认为处于更高层。因为，“必然”是说话人对整个命题的判断，而非命题的附属。在其他语言的依存语料库中，模态词提升也占据了一定比例[19]。AMR 遵从了语言学的理论分析，而非强行约定。

话题化和一般的移位，在生成语法中有较多研究[24]，但在依存语法中却存在较大局限。依存语法没有像生成语法那样，区分移位前的深层结构和移位后的表层结构，依存语法更多的是直接描写移位后的句子结构，所以在体系上不如生成语法严密。另一方面，生成语法虽然可以用转换（transformation）操作来描写移位，但往往需要在句法树上增加很多层次和空位，但在标注真实语料时，又做了很多

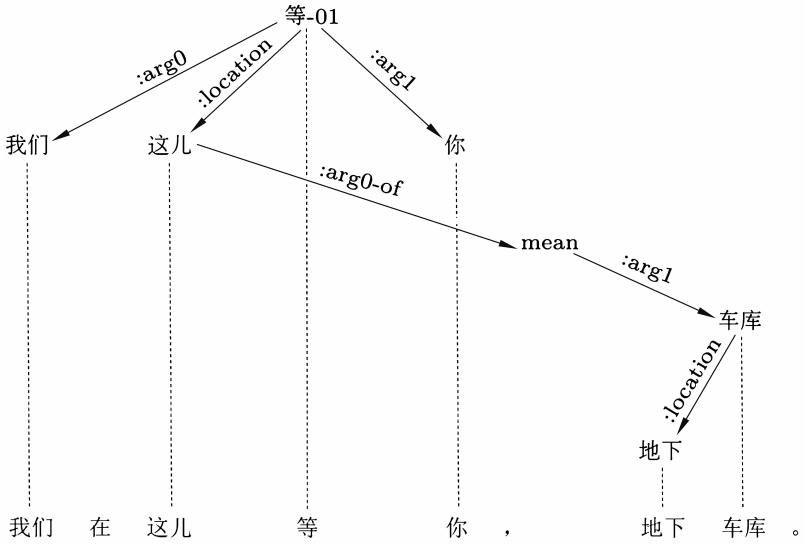


图 9 同位语后置的非投影结构示例

简化,使得移位标注并不那么完整。而对于从属关系分离、复句关系中插入主体感受、成对结构的分析,生成语法和依存语法也尽量回避。

对于自然语言处理来说,句子的语义结构需要更为清晰的描写和表示方法。如果按照简约的句法表示,虽然自动分析的 F 值很高,但不能完整而正确地表示句子的语义结构,对后续的处理会产生负面影响。例如,将"活熊取胆"简化为"施事—谓词—受事"结构,显然是不妥的。AMR 则在语义依存图的基础上增加了概念和关系的灵活处理机制,能够更好地刻画句子的语义结构。而"概念—词语"对齐机制的加入和非投影结构的研究,能够让我们进一步看清汉语中真实存在的移位和特殊的语序现象,从而为语言学理论提供更多的讨论素材,提供相应的处理对策,为汉语的语义自动分析奠定基础。非投影结构的正确分析也能够提升汉语句子的句法语义分析效果,为文本摘要、舆情分析等应用提供更准确的结果。

(2) 自动处理对策

目前,英文 AMR 自动分析的 F 值最高为 74%[25],汉语仅有 58%左右[26]。非投影结构是汉语处理的一大难点。通过上面对非投影结构的分类和具体分析可以看到,非投影结构产生的原因虽然情况复杂、种类较多,但也具有一定的规律性。其中由模态词提升导致的非投影结构占据了超过 50%的比例,一般的成分分离占据了将近 30%的比例。这样对模态词和成分可以分离的词语建立相应的词典,对这两种类型的句子进行预处理或做特殊标记进行机器学习,80%左右的非投影结构就有望得以解决。剩下 20%稍显零散的非投影结构,则需要进一步深入探究,或可考虑对词语移位进行建模计算。我们也期待着基于图结构的一体化句子语义分析方法能有算法上的突破,将本文的分析结果更好地融合到分析算法中。

4 结论及未来工作

近年来,随着句法依存和语义依存在理论和资源建设上的进展,非投影结构在越来越多的语言中被发现和研究,但汉语中的非投影结构一直没有得到较好的理论与实证研究。本文系统地梳理了国际上对于非投影结构的研究历程,并且基于 AMR 的新体系,在增加概念对齐的机制后的 10 149 句中文 AMR 语料库上,通过程序自动提取和人工统计分析得出,带有非投影结构的句子比例为 31.62%。总结出非投影的产生原因主要是模态词提升、话题化、成分分离和一般移位,其中模态词提升和成分分离的情况最为普遍。进而提出利用这两种情况与特定动词之间的较强联系,为其构建相应的词库,对其进行特殊处理,以提升中文 AMR 的自动分析效果。

在未来的工作中,我们将继续分析抽象语义库中超越单纯的投影树结构的语言现象,包括非投影结构和图结构。同时,我们会借助宾州树库等语料标注的移位信息,更为系统地对比分析和研究汉语中的语序问题,从而为语言学研究提供更多理论探讨的空间。最后,我们希望基于中文 AMR 语料库进行非投影结构的自动分析,可提高 AMR 分析器的效果。

参考文献

[1] Nivre J, et al. The CoNLL 2007 shared task on dependency parsing[C]//Proceedings of Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2007, 117(1):53-55.

[2] Oepen S, et al. SemEval 2014 task 8: Broad-coverage semantic dependency parsing[C]//Proceedings of International Workshop on Semantic Evaluation, 2015: 63-72.

[3] Havelka J. Beyond projectivity: Multilingual evaluation of constraints and measures on non-projective structures[C]//Proceedings of 45th Annual Meeting of the Association of Computational Linguistics, 2007: 608-615.

[4] McDonald R, et al. Non-projective dependency parsing using spanning tree algorithms[C]//Proceedings of Conference on Human Language Technology and Empirical Methods in Natural Language Processing. 2005:523-530.

[5] Banarescu L, et al. Abstract meaning representation for sembanking[C]//Proceedings of Linguistic Annotation Workshop and Interoperability with Discourse. 2013:178-186.

[6] 李斌，等. 融合概念对齐信息的中文 AMR 语料库的构建[J],中文信息学报,2017, 31(6):93-102.

[7] Tesnière L. Eléments de Syntaxe Structurale[M]. Librairie C. Klincksieck, 1959.

[8] Ihm P, Lecerf Y. Éléments Pour une Grammaire Générale des Langues Projectives[M]. Bruxelles: Presses Académiques Européennes, 1963.

[9] Hays D G. Dependency theory: A formalism and some observations[J]. Language, 1964, 40(4):511-525.

[10] Marcus S. Sur la Notion de Projectivité[J]. Mathematical Logic Quarterly, 1965, 11(2):181-192.

[11] Robinson J J. Dependency structures and transformational rules[J]. Language, 1970, 46(2):36.

[12] Uhlírová L. On the non-projective constructions in czech[J]. Prague Studies in Mathematical Linguistics, 1972, (3): 171-181.

[13] Štícha F. K řízení vět v češtině[J]. Naše Řeč, 1996 (79):26-31.

[14] Oliva K. Některé aspekty komplexity českého slovního nepořádku[J]. Čeština-univerzália a specifika, 2001, (3):163-172.

[15] Petkevič V. Neprojektivní Konstrukce v Češtině z Hlediska Automatické Morfologické Disambiguace Českých Textů[J]. Čeština-univerzália a Specifika. Brno: Masarykova univerzita, 2001:197-205.

[16] Hajič J, et al. The Prague dependency treebank: A three-level annotation scenario[C]//Proceedings of the Treebanks: Building and using parsed corpora, amsterdam. Kluwer, 2000:103-127.

[17] Hajičová E, et al. Issues of projectivity in the prague dependency treebank[J]. Prague Bulletin of Mathematical Linguistics, 2004, (81):5-22.

[18] Mannem P, Chaudhry H, Bharati A. Insights into non-projectivity in Hindi[C]//Proceedings of 4th International Joint Conference on Natural Language Processing, 2009: 10-17.

[19] Ambati B R, Deoskar T, Steedman M. Hindi CCG Bank: A CCG treebank from the Hindi dependency treebank[J]. Language Resources and Evaluation, 2018, 52(1):67-100.

[20] Zeman D, et al. HamleDT: Harmonized multi-language dependency treebank[J]. Language Resources and Evaluation, 2014, 48(4): 601-637.

[21] 郑丽娟，邵艳秋，杨尔弘. 中文非投射语义依存现象分析研究[J]. 中文信息学报，2014，28(6):41-47.

[22] Cai S, Knight K. Smatch: An evaluation metric for semantic feature structures[C]//Proceedings of Meeting of the Association for Computational Linguistics, 2013:748-752.

[23] Xue N, et al. The Penn Chinese TreeBank: Phrase structure annotation of a large corpus[J]. Natural Language Engineering, 2005, 11(2): 207-238.

[24] Carnie A. Syntax: A generative introduction[M]. Wiley-Blackwell, 2013.

[25] Lyu C, Titov I. AMR parsing as graph prediction with latent alignment[C]//Proceedings of 56th Annual Meeting of the Association for Computational Linguistics, 2018: 397-407.

[26] Wang C, Li B, Xue N. Transition-Based Chinese AMR parsing[C]//Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018, 2: 247-252.

闻媛(1992—),硕士研究生,主要研究领域为计算语言学。
E-mail: wenyuan.njnu@gmail.com

宋丽(1993—),硕士研究生,主要研究领域为计算语言学。
E-mail: songli1105@sina.com

吴泰中(1993—),硕士研究生,主要研究领域为自然语言处理。
E-mail:wtz_njnu@163.com

欢迎订阅《中文信息学报》

《中文信息学报》(Journal of Chinese Information Processing)是全国一级学会—社团法人中国中文信息学会和中国科学院软件研究所联合主办的学术性刊物,创刊于 1986 年 10 月,现为单月刊。由商务印书馆出版,为商务印书馆期刊方阵中的期刊之一,清华大学印刷厂印刷。

《中文信息学报》是我国计算机、计算技术类中文核心期刊。主要刊登中文信息处理基础理论与应用技术方面的高水平学术论文,内容涵盖计算语言学(包括语音与音位、词法、句法、语义、语用等各个层面上的计算),语言资源建设(包括计算词汇学、术语学、电子词典、语料库、知识本体等),机器翻译或机器辅助翻译,汉语和少数民族语言文字输入输出及其智能处理,中文手写和印刷体识别,中文语音识别及文语转换,信息检索,信息抽取与过滤,文本分类、中文搜索引擎,以自然语言为枢纽的多模态检索,与语言处理相关的数据挖掘、机器学习、知识获取、知识工程、人工智能研究,与语言计算相关的语言学研究等。也刊登相关综述、研究报告、成果简介、书刊评论、专题讨论、国内外学术动态等稿件。

读者对象主要是从事中文信息处理的研究人员、工程技术人员和大专院校师生等。《中文信息学报》(国内统一刊号:CN11-2325/N;国际统一刊号:ISSN 1003-0077)国内外公开发行,国内定价每期 30 元,全年 360 元;海外 US $ 50/年(平邮)。

国内发行处:《中文信息学报》编辑部

国外发行处:中国图书进出口总公司 100020 北京 88-E 信箱

1. 支付宝转账:(请注明期刊征订)

账号:cips_pay@163.com

姓名:中国中文信息学会

2. 银行转账

开户银行:工商行北京市分行海淀西区支行

户 名:中国中文信息学会

账 号:0200004509014415619

《中文信息学报》编辑部

地 址:北京海淀区中关村南四街 4 号 7 号楼 201 房间

电 话:010-62562916

电子信箱:jcip@iscas.ac.cn