依存语法和机器翻译

刘 海 涛

提要 回顾了依存语法的发展过程,介绍了依存语法的基本原则和构建方法,比较了短语结构语法与依存语法的异同以及它们在机器翻译中的作用,说明了依存语法与配价语法的关系。认为依存语法在自然语言的计算机处理中有着重要的作用。

一 Tesnière 和依存语法理论的产生

虽然语法依存和动词中心论的概念古已有之,但一般认为现代依存语法理论的创立者是法国语言学家 Lucien Tesnière(特思尼耶尔,1893—1954)。他的主要思想反映在 1959 年出版的《结构句法基础》一书中,事实上,早在 1934 年,特思尼耶尔就发表了阐述依存语法基本观点的论文。特思尼耶尔的《结构句法基础》和其他著作对于语言学的特殊价值在于,它们是建立于对人类语言进行广泛对比研究的基础之上的,这与目前某些语言学者仅仅通过一种语言的研究而得出语言普遍现象的方法迥然不同。特思尼耶尔本身掌握多种古典与现代语言,这使得他的许多发现和理论相似于今天的语言类型和语言共性研究的成果。特思尼耶尔的本意是建立一门跨越各国语言界限、客观揭示人类语言内在规律的句法理论。特思尼耶尔所用的术语"结构句法"与今天我们用的"依存语法"指的是一回事。

《结构句法基础》可分为三个主要部分:La connexion(联系或依存),La jonction(组合),La translation(转移或词类转换)。作者说:"联系、组合和转移是概括一切结构句法现象的三大核心。"遗憾的是特思尼耶尔对什么是结构句法(依存语法)未作正面的定义,我们对于这一概念的理解只能从他对其他问题的论述得出。周国光将配价(依存)语法定义为:"一种结构语法。它主要研究以谓词为中心而构句时由深层语义结构映现为表层句法结构的状况及条件,谓词与体词之间的同现关系,并据此划分谓词的词类。"(见沈阳,1995)这一定义基本反映了依存语法的实质和核心内容。

依存语法在机器翻译领域的应用,开始于 60 年代初的美国和苏联。我国学者冯志伟教授在 80 年代初,利用依存语法作为机器翻译中语言分析和生成的语法模型,取得了不少有意义的成果,他是国内最早研究和应用依存语法的学者之一。当然这些研究与应用不一定与特思尼耶尔的著作有直接的关系,但至少其基本原理是相似的。

二 依存语法的基本原则和建构方法

特思尼耶尔认为结构句法的目的在于句子的研究,句子是一个"有组织的整体",其组织性体现于构成句子的词和词与相邻词之间的"联系",所有这些联系构成了句子的框架。虽然这些联系不是显见的,但这些联系是客观存在的,否则句子便是不可理解的。所谓造句就是在一堆不定形的词之间建立起成为一个整体的"各种联系",反之,语句的理解,就是要找出联结句子中各个不同的词的各种联系。在特思尼耶尔的眼中,"联

系”的概念是如此重要,是它赋予了句子的“有机性”和“生命力”,它是句子的“根本成分”。特思尼耶尔的出发点使得他的句法理论具有良好的可操作性,同时他这些对于句子生成和理解过程的看法,注定了依存语法理论对于计算语言学应用会有极大的价值,因为计算语言学研究和应用的实质就是通过计算机等智能机器去仿真人类的语言处理机制和能力。机器翻译作为计算语言学中最为复杂、涉及面最广的一个应用领域,不但涉及语言的理解,而且也要顾及语言的生成,为此依存语法和机器翻译研究之间具有密切的关系,可能是必然的。

结构联系建立起词与词之间的“依存”关系。每一项联系原则上将一个上项和一个下项联结起来,上项叫支配词,下项叫从属词。一个词可以同时是某个上项词的从属词和另一个下项词的支配词,这样句子里的所有词便构成一个真正的“分层次的体系”。动词是一个句子的中心,支配句中的其他成分。

1970 年美国计算语言学家 J. Robinson 在一篇题为《依存结构和转换规则》的论文中,提出了依存关系的四条公理,这为依存语法的形式化描述及在计算语言学中的应用奠定了基础,这四条公理是:1. 一个句子只有一个成分是独立的;2. 其他成分直接依存于某一成分;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 A 成分直接依存于 B 成分,而 C 成分在句子中位于 A 和 B 之间,那么 C 或者直接依存于 A,或者直接依存于 B,或者直接依存于 A 和 B 之间的某一成分。由于 Robinson 是在转换语法的框架内研究依存概念的,他的这几条公理只是对于依存语法在计算语言学应用中的基本要求,对于完整的、目标为语言自动处理的依存语法而言,仅仅这样是不够的。据笔者所知,目前在机器翻译,乃至整个计算语言学界,对于依存语法理论研究和实践最为详尽的当数荷兰 BSO 公司在研制多语翻译机器系统 DLT 时所做的有关工作。K. Schubert(1987)在其研究专门用于机器翻译的依存句法的著作中认为,一个面向句法形式的、用于计算语言学应用的依存句法应符合以下基本原则:1. 句法只与语言符号的形式有关;2. 句法研究从词素到语篇各层次的形式特征;3. 词在句中通过依存关系相互关联;4. 依存关系是一种有向的同现关系;5. 词的句法形式通过词法、构词法和词序体现;6. 词对于其他词的句法功能通过依存关系来描述;7. 词组是一种作为一个整体与其他词和词组产生聚合关系的单位,它的成分间存在着句法关系,形成语言组合体;8. 一个语言组合体只有一个内部支配者,该支配者代表本语言组合体与句中其他成分发生联系;9. 除句子的主支配词外,句中的每一个词只有一个支配者;10. 每一个词只在依存结构中出现一次;11. 依存结构是一种真正的树结构;12. 应在依存结构中避免出现空节点。Schubert 对于依存语法的定义和原则具有这样的特点:1. 吸取了 Robinson 公理的优点;2. 只将句法限于语言符号的形式方面;3. 充分考虑到句法模型的可操作性和可计算性;4. 将原限于句子的依存语法扩展到词素和语篇层;5. 建构模型时考虑了其普适性,即对多种语言的有效性。

虽然目前在理论和计算语言学中有各式各样的语法体系和理论。但就实质来说,只有两种:其一是以转换生成语法(TGG)为代表的各类模型;另一种便是依存语法(DG),也称从属关系语法。两种语法模型的不同之处大致有以下几点:1. 在 DG 模型中,不存在短语节点。每一节点都与句子中的一个确定的词对应。词的语法范畴、词性等特征可作为附加信息标于同一节点之上,它是一种多标记的语言模型。2. DG 表示比 TGG 所用的节点少。另一方面,DG 语法不能直接反映句子中词的线性次序。相反,在 TGG 中具有明显的词序信息,事实上,词序是一种表示词间关系的手段。利用一种基于 DG 树中结构关系的线性化规则,从 DG 语法树到一般句子的转换不难办到。3. 在 DG 树中,每个节点都至少与一个其他节点之间存在着结构关系。一个父节点支配子节点,或者说是其子节点的支配者。一个子节点依从于父节点,或者说是其父节点的从属者。我们一般将这种依存关系标于树枝之上。4. TGG 注重的是句子的表形特征,也就是说,在 TGG 中句法关系是隐含的,它比较适合描述词序固定的语言。与 TGG 相反,DG 更注重于句法功能或关系。5. 在依存语法树中,树是用(a,b,r)三元组来表示的,这里 a、b 为词汇单元,r 是 a 和 b 之间的有向弧。这种不对称性正体现了自然语言的实际情况,它能有效地表示语言的结构,但 TGG 却难以表述这种成分之间的不对称性。6. TGG 也无法表达句子结构中的中心词及其作用,而 DG 突出了中心词在句法上的功用,这有助于深入的分析。

DG语法模型可能更适用于语言结构的描述。因为语法的主要目的在于描述与揭示构成语言的元素及元素之间的关系,而DG重视的正是语言系统中的关系。DG的这种优点使得它能比较容易地扩展到比词更高或更低一级的语言单位,如:词素与篇章。而利用TGG类的模型很难做到这一点。这可能就是理论语言学界目前正逐渐向DG类语法模型或思想靠近的原因之一。另外,对计算语言学应用而言,TGG的生成能力过强,分析能力却有所不足。所有这些使得TGG本身在计算语言学界很少得到应用。当然我们也不能否认TGG对现代语言学的发展所起的巨大作用,特别是在语言的形式化与精确化描述方面。虽然目前也有一些派生于TGG的文法,如管约论、扩充短语结构语法、词汇功能语法等,但无一例外,这些新的理论和文法都改善并注重了以下方面:依存关系,中心词之类的概念;句法功能的表述;由词汇限定的句法,即在词汇描述中包含了更多的句法信息;基于复杂特征的合一算法等。而这些方面原本就是依存语法的特征。

依存语法中除了依存关系外的另一个重要概念便是"价","价"主要体现了一种词汇的组合能力,它是针对词汇而言的。"依存"与"价"不同之处在于,依存是一种句法概念。依存语法对于词汇的重视使得它格外关注词汇"价"信息的研究,这可能就是依存语法也被称之为"配价语法"的原因。按照特思尼耶尔的原著,"价"只是动词所特有的,即动词所能"支配的人物语的数目"。现代语言学家所理解的"价"的含义一般更为广泛,它被认为是一种"特定次类的支配能力(Schubert,1987)",这样就不单动词有价,诸如名词、形容词等词类也可用价来描述。在依存语法中,词汇成了体现语言符号系统中各相关成分的关系建构的核心,以词汇为基点的词典与依存关系的汇合是依存语法的特质,依存语法离不开一部含有"价"信息的词典。而这种词汇"中心论"是现代句法和计算语言学研究的方向之一。

依存语法是一种建立于"同现"基础之上的语法理论,建立一种自然语言的依存语法一般可分为两步:首先通过同现来确定语言所具有的词类,这基本类似于其他语法中确定词类的方法;在确定词类的过程中,我们不仅仅研究了词的问题,也发现了词类的同现关系,通过将这些同现关系整理、比较和分类,我们可以得出词类与词类之间的同现关系,然后加以命名,便得到了这种语言的依存关系。经过这两步的工作,研究者便可以编写出语言的依存语法。像任何语法描述一样,依存语法的建立同样存在"任意性"的问题。这里所谓的"任意性"指的是在语法的建立过程中,语法的编写者在符合基本原则的前提下,具有一定的主观可选择性。这也是同一种语言的同一类型语法,因编者的不同而略有差异的原因。

依存语法中依存关系的数目不宜过多或过少,过少的依存关系使得语言描写的深度和精度不够,而太多的依存关系又会使语言分析和处理的过程太为繁杂,代价太高。为此考察以下几种语言的依存关系数目,对于这一问题的理解可能是有益的。各种语言依存语法模型均选自Maxwell(1989)一书:德语26种,丹麦语15种,波兰语18种,孟加拉语20种,芬兰语21种,匈牙利语21种,日语20种,世界语18种,法语21种,汉语36种。这些数据说明了两个问题,一是依存语法理论确实是一种跨越各国语言界限、客观揭示人类语言内在规律的句法理论;二是依存语法中依存关系的数目不能太多,清华大学黄昌宁教授等人的研究证明了这一点,在他们的系统中汉语依存关系的数目从原有的106种依存关系减为目前的44种。

三　依存语法和语义

特思尼耶尔认为句法有别于语义,他将二者用"结构平面"和"语义平面"加以区分,结构平面属于语法的研究范畴,而语义平面则属于心理学和逻辑学。"句法是独立自主的"这一前提使得我们有可能集中精力研究语言系统的形式方面,这也就是说"依存语法"的本质只是一种与句法结构有关的语言理论。

为了便于研究,从理论上可以认为句法和语义二者是独立的,而事实上这两个平面是并行的,这体现了语言符号的二重性,即符号的形式与内容。句法是形式,语义是内容,研究形式的目的在于更好地表述和理解内容。"这种并行性体现于各种联系之中,实际上语义联系是覆盖在结构联系之上的",按照我们对这一句话的

理解,句子的语义是可以由句子的依存结构推断或变换而来的。依存结构表达语义的方式为:"从属词的语义附加在它所依附的支配词的语义上",特思尼耶尔说:"词语在层次体系中的重要性(结构平面)和语义的重要性成反比。一个词在结构层次上越低,就越有可能对句子的意义有举足轻重的作用"。换言之,语义联系可以用从属成分限定支配成分来说明。有理由认为,在特思尼耶尔的语言理论中句法关系与语义关系间存在着一种天然的同形关系,这有利于对依存语法结构表述作进一步的语义处理。

目前在理论语言学和机器翻译界处理语义问题的理论中,应用最为广泛的是由美国语言学家菲尔墨创立的"格语法"理论。格语法认为核心句的基本结构是由一个动词和几个名词组构成的,各名词组通过某种格关系与动词发生联系。与依存语法相似,格语法中的"中心语"也是动词,句中的其他成分通过某种关系与中心语发生联系。虽然在格语法中,这种联系被称之为"格关系",在依存语法中这种联系是依存关系,但二者在本质上存在一种共性。这一点就连"格语法"的创立者菲尔墨自己也承认他的理论与特思尼耶尔的DG语法有异曲同工之处。研究从依存语法表示到格语法表示的转换过程,无疑对于机器翻译应用中的语义处理及句法结构到语义结构的映射有极其重要的意义。格语法和依存语法这种结构上的相似性,可能说明了在重视语义处理的今天,"格语法"和"依存语法"在机器翻译界和计算语言学界为什么正日益流行的原因。

由于"格语法"或"概念依存理论"等方法十分强调语义作为其基础,在这些理论中是由语义模式主导、句法为辅的。我们知道语义较之句法而言,是难于客观处理的。对于语义的研究和处理,适宜的方法可能是通过句法这条有形的线,进而深入到语义,达到理解之目的。根据特思尼耶尔依存语法和相关概念发展起来的自然语言语义处理技术可以从以下两方面来看,一是以研究词汇配价信息为主的"词汇中心论",如:"词汇—语法",二是目前针对大规模真实文本处理的、基于语料库的语言处理技术。

根据自然语言语义的本质和属性,我们曾提出过一个词的意义就是语境关系的总和的观点。而语境关系是一个词在各种语境中所遇到的全部关系。具体而言,这些关系有句法关系和语义关系,如:词的搭配关系、同现关系、支配关系等。语义是一种隐含的、难于量化的东西。如果采用"用法"的观点来处理语义,我们首先遇到的问题就是如何将一个词的语境关系清楚地告诉计算机。经过结构化或标注处理的语料库,可以提供每一个词的搭配关系、垂直同现约束和水平同现约束关系。换言之,语料库可以作为基于词汇"用法"语义处理机制的知识库使用,使基于维特根斯坦语言哲学中的语义机制能在计算语言学应用中得以实现。所有这些都与依存语法对于自然语言描述的适宜性分不开。

四 依存语法和机器翻译

如果把一篇文章看作为一个巨大的符号,那么翻译就是在保持内容不变的前提下,用另一种符号替代原符号的过程,换言之,翻译是一种改变语言形式,而保留其内容的活动,这一过程可用下式来描述:

$F_s(C_s) => F_t(C_t)$; $C_s \approx C_t$

式中:F_s 指源语的形式,C_s 为源语形式中所含的内容

F_t 指目标语的形式,C_t 为目标语形式中所含的内容

$C_s \approx C_t$ 是从源语影射到目标语的前提与条件,没有这种限制,这里所说的转换就称不上是翻译。从理论上讲,$C_s = C_t$ 是最理想的状况,但事实上,由于语言的模糊性、信息的多样性和语义网的粒散性,C_s 与 C_t 之间很难有绝对的等值存在。

翻译过程是一种分层次的转换过程,一篇文章的层次结构一般为:篇章→句段→句子→子句→短语→词→词素,层次越低,翻译难度就越小。但语言的不同性决定了不能总在最低的层次上进行翻译,而语言的模糊性和歧义性又迫使我们将低一层次的成分放到更高一级的层次去考虑。就翻译的理想状况来说,当然最好是能在源语语篇与目标语语篇之间达到等值。所谓机器翻译就是利用具有智能处理能力的机器(目前主要是计算机)来完成这一语言转换的过程,它是人类翻译过程的仿真。

根据以上讨论,我们可以认为,自然语言的依存语法描述,对于用计算机来处理语言的句法和语义问题而

言,较之其他语法理论,有一定的优越之处。如果撇开翻译过程中源语的分析表述和目标语的生成问题不谈,翻译过程中最为重要的一环就是语际转换。由于客观存在的可译性,将一种语言形式转化为另一种语言形式并保持内容不变的过程是可行的。也就是说,源语到目标语的转换是可以行得通的。

《结构句法基础》中有专门讨论翻译过程中结构转换的章节,特思尼耶尔用术语 metataxe 来表示在翻译过程中由于语言结构的不同,而导致的"结构变换"这一概念。翻译理论家 P. Newmark 认为特思尼耶尔著作中的这一章节是"40 页宝贵的翻译理论"。特思尼耶尔"结构变换"的思想告诉我们:翻译不仅仅是一个机械的替换字词的过程,必须将句子作为一个整体来考虑。对于机器翻译而言,我们的语际转换算法应该具有在更高层次上转换的能力。在依存语法的奠基性著作中有关于翻译的章节,说明特思尼耶尔在考虑依存语法的基本框架时注意了翻译问题。机器翻译和依存语法之间的渊源,正起自于特思尼耶尔本人的思想。

在采用依存语法体系的机器翻译系统中,语际转换显然不应只包括词类的转换,也应包括结构关系的变换,在这里就是依存关系与词类的转换。为了进行这种转换,必须有转换规则,按规则的常用度可分为"无标记"规则与"有标记"规则。"有标记"规则是一种特殊的规则,"无标记"规则可算作一种缺省的规则,如英语中的宾语一般可转换为汉语中的宾语就是一条"无标记"规则。这种结构转换的步骤一般为:1. 遍历应转换的树,从上到下,从左到右;2. 寻找下一个应转换的符号(标签或词);3. 寻找适宜于该符号的转换规则,如果规则与符号本身及它的相关词或标签吻合,并满足规则的使用条件,则可用;4. 从可用规则中,选取优先级最高的;5. 采纳此条规则,利用双语词典进行词的翻译;6. 返回第一步,直到句末停止。

从形式化的观点看,如果每个节点与标签都能以上下文无关的方式转换,则是最理想的,但由于对比句法不是如此简单,故我们必须采取其他办法,原则是应尽可能缩小转换规则的范围,范围愈小,规则就愈有一般性,所需的规则也就愈少。但这本质上是一个小词典或小规则集之间的折中。为了保证转换的准确性,规则库的规则分层次排列,如:"有标记"规则优于"无标记"规则,包含多符号的规则优于少符号的规则等等。系统选择规则时按优先级选用。通过这种方法,可实现源语到目标语的映射,语义在结构转换过程中保持隐含。通过语际转换,目前就可得到目标语的从属结构树。

参考文献

[1] Engel, U, 1992: *Deutsche Grammatik*. 北京语言学院出版社/ Groos.

[2] Fillmore, C.J, 1968: (*The Case for Case*)《"格"辩》,载《语言学译丛》第二辑,1980。

[3] Maxwell, D/ Schubert, K, 1989: *Metataxis in Practice*. Dordrecht: Foris.

[4] Schubert, K, 1987: *Metataxis: contrastive dependency syntax for MT*. Dordrecht: Foris.

[5] 冯志伟《现代语言学流派》,陕西人民出版社,1987。

[6] 冯志伟《自然语言机器翻译新论》,语文出版社,1995。

[7] 胡明扬(主编)《西方语言学名著选读》,中国人民大学出版社,1988。

[8] 刘海涛《维特根斯坦语言哲学对计算语义学的影响》,载《计算语言学研究与应用》,北京语言学院出版社,1993。

[9] 刘涌泉、乔毅《应用语言学》,上海外语教育出版社,1991。

[10] 沈阳、郑定欧(主编):《现代汉语配价语法研究》,北京大学出版社,1995。

[11] 周明、黄昌宁《面向语料库标注的汉语依存体系探讨》,《中文信息学报》1994 年第 3 期。