新的句法标注模型探索

李良炎
(重庆大学 语言认知及信息处理研究所,重庆 400044)

摘要:由于自然语言的语义存在不确定性,形式化很困难,因此语义处理成为自然语言处理的瓶颈所在。基于大规模标注语料库的语义处理已经成为发展趋势,语料标注本质上就是语言知识(包括语义)形式化。现有句法标注模型主要包括基于短语结构语法(PSG)和基于依存语法(DG)的句法标注模型,还存在一些局限性。文章在现有句法标注模型的基础上结合认知语法(CG)的有关理论提出改进思路,以探索新的句法标注模型。

关键词:语料库语言学;语义处理;句法标注模型
中图分类号:H043 **文献标志码**:A **文章编号**:1008-5831(2007)03-0131-04

人类社会发展的基本轨迹是:原始社会→农业社会→工业社会→信息社会。人工智能的目标是用计算机模拟人的智能,以最大限度地解放和延伸人的智能,无疑是信息社会的制高点。语言是人思维的物质外壳,人不可能离开语言而具备真正属于人的高级智能。因此,模拟人类语言智能的自然语言处理无疑是人工智能的重要研究方向。然而,迄今为止的研究表明,在可以预见的将来,语义处理将是自然语言处理的瓶颈所在。原因是语义十分复杂,而基于现有计算机软硬件的自然语言处理要求语义形式化。解决这一问题的根本之道是:探索新的句法标注模型,进行大规模的语义标注,基于语料库进行语义知识获取和自然语言处理。

一、句法标注模型

语言的复杂性在于语言与认识的关系。语言具有意义,而意义是人对主客观世界的认识结果。主客观世界的复杂性决定了意义的复杂性,进一步决定了语言的复杂性。语言本身又可以视为人的主客观世界中的一部分,因此语言研究是一种特殊的认识活动,是人对语言的认识。由此可见,语言离不开认识。人对主客观世界的认识可以如此描述:认识主体借助认识工具按照认识方法处理认识对象获得认识结果。认识是由多种认识因素(主体、工具、方法、对象)共同作用的活动,认识结果是这一活动的产物,被多种认识因素共同决定,任何一种认识因素的改变必然导致认识结果出现或大或小的差异。显然,认识结果与认识对象不能等同,是认识主体对认识对象的选择性反映,认识具有主观能动性。从这个意义上讲,认识不可能也不应该去被动地还原认识对象,而是从符合主体目的性出发,力求简单有效地描述和预测认识对象。借用模型的概念,认识结果就是认识对象的模型(model),认识就是建立认识对象的模型,简称建模(modeling)。这是一种实用主义认识观。

模型一般分为心理模型(psychological model)、数学模型(mathematical model)

收稿日期:2007-04-10
作者简介:李良炎(1974-),男,重庆人,重庆大学外国语学院讲师,博士,主要从事语料库语言学和词汇语义学研究。

和物理模型(physical model)。心理模型是认识对象在人认识中的定性关系,是数学模型的基础;数学模型是认识对象在人认识中的定量关系,是物理模型的基础;物理模型是人借助特定材料和工具按照认识对象的数学模型实现的物质结构。传统意义上的建模主要指建立数学模型和物理模型,一般意义上的建模还包括建立心理模型。人的认识能力是有限的,表现在:人不能建立任意认识对象的心理模型,也不能建立任意心理模型的数学模型,也不能建立任意数学模型的物理模型。由于具有明确的实用主义特点,建模在理工科领域大行其道,在文科领域也逐渐受到青睐。人类将二进制数学模型成功实现为晶体管物理模型,并开发出越来越复杂和先进的计算机软件和硬件,从而进入信息时代。20世纪以来一些主要或次要的语言理论都或多或少应用了数学模型,特别是一些面向语言计算的语言理论。随着计算机技术的飞速发展,人们对计算机自动或辅助处理语言信息的需求越来越大。但计算机的根本缺陷在于,凡是不能建立数学模型的信息都无法处理。传统语言理论往往只在心理模型层面定性研究,无法满足这一需要。因此有必要引入数学模型研究语言,称为语言数学模型,简称语言模型(language model)。统计语言模型(statistical language model)就是一个成功的例子。但统计语言模型的性能取决于训练语料的规模和质量。目前,由于语料的不断积累和计算机技术的不断进步,语料规模已不成问题,语料中包含语言知识的数量和质量才是关键。

计算机的语言知识主要来源于人。将语料中包含的语言知识标注出来,有助于计算机获得更丰富、更有价值的语言知识,从而提高语言处理水平,这就是语料标注(corpus tagging)。一般认为主要包括词汇标注(lexical tagging,分词、词结构标注、词性标注、词义标注等)、句法标注(syntax tagging,语法树标注、语义树标注等)、语篇标注(discourse tagging,语体标注、领域标注等)等内容。经过标注的语料还可以用于语言学研究、语言教学、语言测试、词典编撰等诸多理论研究和实践应用领域,越来越受到人们重视,并形成一门新兴学科——语料库语言学(corpus linguistics)。目前,相对句法标注,词汇标注有更成熟的规范、准确率更高的技术和更大的标注规模。句法标注的主要困难在于,没有一个真正成熟的语法或语义标注模型。句法结构尤其是语义结构很难统一描述,现有的句法理论还不完善,难以制定统一规范,标注主观性很大,自动标注准确率比较低。因此,句法标注成了语料标注的瓶颈问题。由于句法知识在语言知识中的重要地位,有理由相信:如果有了大规模、高质量的句法标注语料库,围绕语料库的各种研究和应用有可能在现有基础上产生质的飞跃。因此,研究句法标注模型应是当务之急。语料库语言学属于交叉学科,句法标注模型是语料库语言学的基础理论,又与语言学的句法理论密切相关。一方面可以借鉴现有句法理论,另一方面,也可以从语料库语言学的角度研究句法,提出新的句法标注模型。

二、现有句法标注模型

句法标注(Syntax Tagging, ST)以句子的语法知识和语义知识为标注对象,是语料标注的重点、难点所在,要以一定的语法理论为基础。根据语法理论制定的句法标注规则、过程和结果,称为句法标注模型(Syntax Tagging Model, STM)。短语结构语法(Phrase Structure Grammar, PSG)和依存语法(Dependency Grammar, DG)是现有句法标注的两种基础语法理论,彼此却有很大的不同。基于PSG的句法标注模型称为短语结构句法标注模型(PSG - based Tagging Model, PSGTM),基于DG的句法标注模型称为依存句法标注模型(DG - based Tagging Model, DGTM)。根据现有语料标注的实践结果来看,PSGTM与DGTM都存在一定缺陷。

美国语言学家乔姆斯基(Noam Chomsky)于1957年出版专著《句法结构》,从而奠定了短语结构语法(PSG)的理论基础。其后发展起来的许多语法理论可以直接或间接归到这一流派,如中心词驱动的短语结构语法(HPSG)、广义短语结构语法(GPSG)等。到目前为止,PSG仍然是最重要的句法标注基础理论,为世界上众多语料库项目所采用和发展。法国语言学家特思尼耶尔(Lucien Tesnière)于1959年出版专著《结构句法基础》,从而奠定了依存语法(DG)的理论基础。其后发展起来的许多语法理论可以直接或间接归到这一流派,如词汇依存语法(WD)、概念依存理论(CD)、核心依存理论(KD)等。相对PSG而言,DG偏重于语义,在CD、KD上表现得十分明显。另外,DG更简洁、直观、经济,适应性更强,因此反而有后来居上之势,目前已经成为世界上较为通用的句法标注基础理论。不过,在具体的句法标注实践中DGTM还是暴露出一些问题,"对一些没有明确依存关系的成分,标注起来则有些力不从心"[1],存在"依存失败"[2]现象,最突出的是难以标注缺省结构。缺省结构一直是句法标注中经常出现而且很难解决的问题。

人类的自然语言符合经济性原则,而缺省结构恰恰体现了这一原则。借助句子的前后上下文省略一些成分,人们仍然能够理解,但对计算机来说却是一种挑战。句法标注的根本目的是让计算机能够正确提取句子的语法和语义知识。缺省结构在真实语料

中大量出现，常常使得原本正常的句法结构变得异常，难以按已有规则进行标注。这是任何句法标模型都必须面对的问题，目前 PSGTM 和 DGTM 都还没能够很好地解决。以 DGTM 为例，在很多情况下，DGTM 不但不能正确标注缺省结构，反而在一些语言规则的强制限定下给出违背真实语法或语义结构的标注结果，形成干扰信息。请看以下 4 个句子：

句 1：我看一下书

句 2：（真是好书啊？）我看一下

句 3：我看一本书

句 4：（好多书啊！）我看一本

句 2 是句 1 的宾语省略句，句 4 是句 3 的宾语省略句。各句的 DGTM 见图 1（为简便起见，把“一下”、“一本”作为一个词处理）。

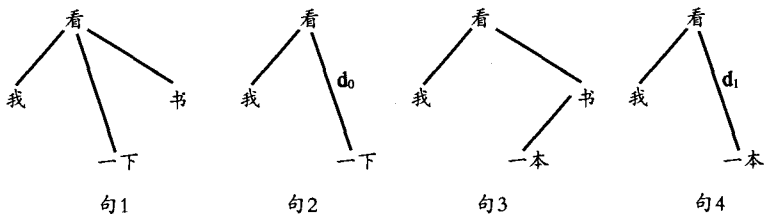

图 1　句 1－4 的 DGTM

问题出在句 4。句 1 和句 3 的依存结构是不同的，然而句 2 和句 4 却有了相同的依存结构。因为句 4 省略了“书”，根据 DG 理论，“一本”必须依存于独立谓语成分“看”。于是“看—本”和“看—下”依存结构相同，实际上违反了句 3 的正确结构。当然，我们可以采取补救措施，为 d1 标注一个特殊的依存关系属性 Cerror（即依存失败），但这不是好办法。

三、改进 DGTM

美国认知语言学家兰盖克（Ronald W. Langacher）分别于 1987 年、1991 年出版专著《认知语法基础》一、二卷，开创了认知语法（CG）理论，关于语法结构有如下观点[3]：如果一个构件 A 使另一构件 B 的一部分抽象变为具体，那么构件 A 就叫做概念自主（conceptually autonomos）的构件，构件 B 就叫做概念依存（conceptually dependent）的构件。

举例来说：独立地看，“一本”隐含一个抽象的、可数的、可用“本”量化的事物，可表示为“一本（x）”。“书”使“x”变得具体，因此“书”是概念自主的，“一本”是概念依存的。从信息表达的角度来看，“书”表达了相对完整而具体的信息，因此是概念自主的；“一本”表达了不完整不具体的信息，因此是概念依存的。从数学表达式的角度来看，“一本”类似函数，“书”类似参数，函数的地位显然是第一位的，决定了对参数的处理过程和返回参数。例如，“旧书”与“一本书”的区别不在“书”，而在“旧”和“一本”。再从阅读认知过程来看，当人们读到“一本”时，实际上已经在期待“一本”后面那个具体事物跟着出现。为什么我们觉得“我看一本”是缺省句？因为“看”和“一本”相对“书”都是概念依存的，因此人们会判定，“我看一本”的缺省成分可能是“书”。而读到“我看书”时，人们不会认为这是一个省略句，因为“书”表达的信息已经自足了。

由此有足够的理由认为：在句法结构中，“一本”应是“书”的父结点，而不是按传统的补足中心原则，中心成分总是限定成分的父结点（图 1 中句 3 所示）。依存成分是自主成分的父结点，这一原则可以称为依存中心原则（Dependency Head Principle，DHP）。采取这种原则的 DGTM 必然会有不同的标注结果。

深入研究发现，仅仅采用 DHP 是不够的，DGTM 的其他参数也需要改变。例如，“看（x）”和“一本（x）”这两个表达式在与其他词语组合时是有区别的。“看（x）”与“我”组合时由“看”与“我”产生联系。“看”与“一本（x）”组合时却是“x”（书）与“看”发生联系。代表表达式与其他词语组合的成分称为返回参数，不同表达式的返回参数是不同的。例如，“一本（x）”返回参数为“x”，“看（x）”返回参数为“看”。正因为如此，表达式“看（一本（书））”成立，“一本（看（书））”不成立。另外，表达式“（x）一下”的返回参数为“x”，即“看”；表达式“（x）看”的返回参数为“看”。根据这些定义，句 1、2、3、4 的改进 DGTM 见图 2。

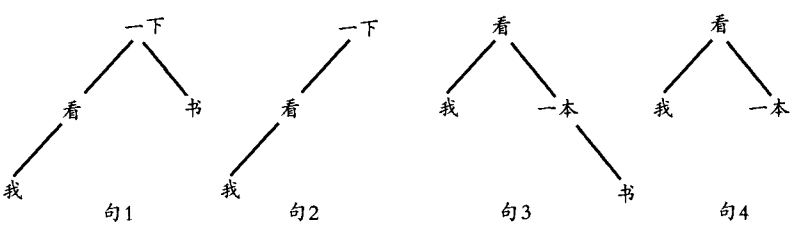

图 2　句 1－4 的改进 DGTM

根据函数、输入参数、返回参数的关系,各句结构的逆构造过程如下:

句1:我看一下书:(((我)看(x))一下)(书) = ((看(x))一下)(书) = 看(x)(书) = 看(x = 书)

句2:我看一下:((我)看(x))一下 = (看(x))一下 = 看(x)

句3:我看一本书:((我)看(x))(一本(书)) = 看(x)(书) = 看(x = 书)

句4:我看一本:(我)看(一本(x)) = 看(x)

句1和句3的x有明确取值,为完整句。句2和句4则是缺省句。基于看(x)和一本(x)的知识,可以预测并判定缺省结构及其成分。

直观看来,改进DGTM与原DGTM的标注结果有了很大的差异(对比图1和图2)。由于不采用补足中心原则,因此改进DGTM标注结果并不符合在补足中心原则影响下人们长期以来形成的语感。但更符合人们阅读认知经验,而且可以按函数标准给出形式化地解释,其解释结果符合句子本身的语法和语义结构,没有错误和干扰信息。因此,改进DGTM更适合计算机处理,更符合句法标注的本来目的。

四、结语

PSGTM的语法理论基础是PSG,DGTM的语法理论基础是DG,改进DGTM的DHP受CG的启发,其语法理论基础应该是CG。但CG只是从理论上提出了"概念自主"和"概念依存"的概念,并没有严格定义和证明依存成分与自主成分之间的主从关系。在CG的实际应用中,存在有时自主成分为短语中心语,有时依存成分为短语中心语的情况。

根据CG理论,"above"是"above the table"的中心语,"lamp"是"lamp above the table"的中心语。然而,根据CG对概念自主和概念依存的界定,相对"table"和"lamp","above"是概念依存的,具有两个抽象部分"(x)above(y)","lamp"使"x"具体化,"table"使"y"具体化。如果严格执行DHP,"above the table"和"lamp above the table"的中心语都应该是"above"。但这样一来,怎样解释"move the lamp above the table"中"move"直接依存"lamp"的关系?根据改进DGTM,可以定义"(x)above(y)"的返回参数是"x"以解决这一问题,但CG不会这样处理,而是将"lamp"限定为"lamp above the table"的中心语,从而与"move"直接联系,这样就不符合DHP的要求。

因此,改进DGTM的语法理论基础不可能是CG,必须构建一种新的语言模型。目前我们正融合哲学二元论与本体论、心理学、信息科学、网络通信模型、离散数学、语言学(依存语法、认知语法、范畴语法)、艺术学等理论的相关概念和原理,结合人的一般认知经验,建立一种新的句法标注模型,并初步用于经典汉语句式的表征,取得了较好效果。

参考文献:

[1]周强.汉语句法树库标注体系[J].中文信息学报,2004,18(4):1-8.

[2]尤昉,李涓子,王作英.基于语义依存关系的汉语语料库的构建[J].中文信息学报,2003,17(1):46-53.

[3]齐振海,张辉.导读[M]//LANGACKER RONALD W.认知语法基础(理论前提).北京:北京大学出版社,2004:13.

A New Syntax Tagging Model Exploration

LI Liang-yan

(*Institute for Languages, Cognitive and Language Processing, Chongqing University, Chongqing 400044, China*)

Abstract: Due to the uncertainty of the natural language meaning and difficulties of the semantic formalization, semantic processing becomes the key problem of Natural Language Processing. Semantic processing based on large scale tagging corpus has become the current tendency. Virtually, corpus tagging is the formalizing of the linguistic knowledge including the meanings. The current syntax tagging models are customarily based on PSG(Phrase Structure Grammar) or DG(Dependency Grammar), but all have some shortcomings. Based on CG(Cognitive Grammar), this paper presents an approach for improving the current syntax tagging models in order to explore a new syntax tagging model.

Key words: corpus linguistics; semantic processing; syntax tagging model