计算语言学刍议*

邵泽国[1,2]
(1.上海师范大学 语言研究所,上海 200234;
2.上海电子信息职业技术学院,上海 201411)

摘 要:解析了计算语言学的定义,介绍了计算语言学的发展史,讨论了计算语言学的研究方法和研究意义,着重分析了针对汉语言的计算语言学的研究状况及存在的问题。

关键词:计算语言学;汉语;自然语言处理
中图分类号:H08 **文献标识码**:A **文章编号**:1674-7720(2013)06-0068-04

Discussion on computational linguistics

Shao Zeguo[1,2]
(1.Institute of Chinese Linguistics, Shanghai Normal University, Shanghai 200234, China;
2.Shanghai Technical Institute of Electronics & Information, Shanghai 201411, China)

Abstract: This paper parses the definition of computational linguistics, introduces the development history of computational linguistics, and discusses the research methods and significance of computational linguistics, and focus on the analysis of the Chinese language in the computational linguistics research present situation and facing problems.

Key words: computational linguistics; Chinese(language); natural language processing

自20世纪40年代人类开始研究机器翻译以来,“使计算机具有人的语言能力”就成为了人们一直追逐的美好愿望。这一愿望驱使了语言科学与计算机科学的结合,从而诞生了计算语言学。经过几十年的发展,这一交叉学科涉及的学科领域不断扩大,同时在不同的学科视角下对该学科有着不同的定义和命名。本文从语言科学的视角来观察这一新学科,提出了一些不成熟的看法。

1 计算语言学综述

俞士汶的研究中这样定义计算语言学[1]:“计算语言学(Computational Linguistics)指的是这样一门学科,它通过建立形式化的数学模型来分析、处理自然语言,并在计算机上用程序来实现分析和处理的过程,从而达到以机器模拟人的全部或者部分语言能力的目的。”从上述定义中可以解析出:(1)计算语言学将人文科学(语言学)与自然科学(数学、计算机科学)紧密地结合在一起,是人文科学与自然科学的一个桥梁;(2)这门学科的研究本体是人类语言(自然语言),其目标是“使计算机具有人的语言能力”;(3)计算语言学是一个典型的边缘交叉学科,涉及到语言科学、计算机科学和数学。

随着人们对该学科的逐步认识和不断研究,如今计算语言学已开始涉及更多的学科领域,例如认知学、逻辑学、心理学、社会学、人类学等。而同时人们把用计算机处理自然语言的过程在不同时期或不同侧重点时又称为自然语言处理(Natural Language Processing)、自然语言理解(Natural Language Understanding)、人类语言技术(Human Language Technology)、计量语言学(Quantitative Linguistics)、数理语言学(Mathematical Linguistics)等。

机器翻译是人类最早用计算机来处理非数值运算的应用,它首次将自然语言与计算机联系在一起。上世纪60年代,机器翻译的研究步入了一个低谷,从而引发人们重新审视语言的计算机处理。很快人们发现语言的计算机处理过程不同于一般的数值计算,它不是一个简单的机械过程,应该注意对自然语言的理解。随后人们开始尝试用计算机来理解语言的含义。通过对语言的分析和计算让机器能够解释语言,这样的一个研究范畴被人们称为自然语言理解。随着人们对“理解”的日益加深——计算机对语言的理解离不开或者说根基于计算

*基金项目:教育部哲学社会科学研究重大课题攻关项目(09JZDH007);上海师范大学研究生优秀成果(学位论文)培育项目(B-6001-11-003105)

机对语言数据的处理，随之焦点从“理解”变成了“处理”，即而有了自然语言处理。

随着更多学科的渗入，人们开始细化、区分计算机对自然语言处理的过程。若在这个过程中侧重于从计算的角度来看待语言学的性质，或以自然语言为对象来研究算法，则称之为计算语言学，它是用计算机来模拟人去分析、处理自然语言。如果是专注于对自然语言进行各种类型的信息处理和加工技术的研究，且强调计算机实现，则称之为自然语言处理。若是以计算机作为工具手段，用数理统计方法来研究和描述自然语言，对自然语言进行计量研究，通过计算机的处理来获取语言中隐含的数量规律则称为计量语言学。如果是专注于以数学方法来刻画语言的各种特点，从而形成表述严密的语言理论体系，则称之为数理语言学。

自然语言是人类文明的一个结晶，它具有人的一定属性。有些学者认为计算语言学更接近于人类学的研究范畴，于是在人类学的领域里，人们开始用“人类语言技术”这一更确切的术语来命名面向人类语言的处理技术的研究。

事实上，计算语言学与自然语言处理、自然语言理解、人类语言技术、计量语言学、数理语言学相互之间没有严格的界限，一般人们会用计算语言学或自然语言处理来命名计算机处理自然语言的过程。

2 计算语言学的发展

计算语言学的发展历程按照时间节点分为 3 个阶段：萌芽期、发展期和应用期。

(1)萌芽期

计算语言学的萌芽期是指 20 世纪 50~60 年代。1954 年，美国乔治敦(Georgetown)大学与 IBM 公司合作，在 IBM-701 型计算机上进行了俄语翻译成英语的机器翻译实验，这是世界上首次将计算机应用在非数值计算的信息处理领域。这一实验标志着计算语言学的诞生。

该阶段的计算机语言学仅局限于机器翻译的研究，并且人们只是将语言作为一种特殊数据类型的数据交给计算机来计算，并没有将这种数据赋予语言的特性。

(2)发展期

计算语言学的发展期是指 20 世纪 60~80 年代。在此期间，人们除了继续机器翻译研究，还出现了对语言信息检索的研究。这个时期最重要的标志是人们开始注重计算机对语言的“理解”。比较有代表性的研究成果有：①20 世纪 60 年代，出现了一批基于诺姆·乔姆斯基(CHOMSKY N，美国语言学家)的转换—生成语法的语言处理系统。如麻省理工学院拉法勒(RAPHAEL B)的信息检索系统 SIR、韦森鲍姆的 ELIZA。这些系统采用的主要技术是模式识别中的句法匹配，但没有成熟的句法分析；②1972 年伍兹(Woods)在他的自然语言信息检索系统(LUNAR)中提出了著名的扩充转移网络 ATN(Augmented Transition Network)。同年，威诺甘德(WINOGAND T)的自然语言理解系统(SHRDLU)嵌入了一个句法分析程序、一个语义分析程序、一个问题求解器，是一个句法、语义和推理的组合系统。1975 年，香克(SCHANK R)设计了基于本人概念从属理论的 MARGIE(Meaning Analysis，Response Generation，and Inference on English) 系统，系统由概念分析器、推理器和篇章生产器 3 部分组成。

(3)应用期

自 20 世纪 80 年代至今统称为计算语言学的应用期，这时人们开始将计算语言学更多地称为自然语言处理。这个阶段有两个重要的变化：一是一些计算语言学(自然语言处理)系统开始走出实验室，逐渐成为被社会接受的实用系统；二是基于统计的计算语言学处理方法开始出现并逐渐成熟。

在此期间计算语言学反过来促进语言科学的发展表现得尤为明显，涌现出了各种新的语法体系，如 Gazder 的广义短语结构语法(Generalied Phrase Structure Grammar)、Bresnan 与 Kaplan 的词汇功能语法(Lexical Functional Grammar)、KAY M 的功能合一语法(Functional Unification Grammar)等。到了 20 世纪 90 年代，随着计算机技术的发展，特别是关系型数据库技术的成熟，语料库语言学(Corpus Linguistice)的研究蔚然成风，许多国家和学术机构相继推出了不同语种的超大型语料库或知识库。

这些成果大大提高了计算语言学(自然语言处理)系统的能力，涌现出了诸如美国的 METAL 和 LOGOS、日本的 PIVOT 和 HICAT、法国的 ARIANE 以及德国的 SUSY 等著名的实用性系统。

3 计算语言学的研究方法

计算语言学的研究方法一般分为基于规则的方法、基于统计的方法以及规则与统计相结合的方法。有学者从方法论上又将基于规则的方法称为理性主义方法，将基于统计的方法称为经验主义方法[2-3]。

(1)基于规则的方法

基于规则的方法(简称规则法)通常是先由语言学家撰写“规则库”(例如“词典”)，再由计算机科学家编写算法程序，对“规则库”进行解释和执行，如图 1 所示。具体地说，就是由句法分析器按照设定的自然语言语法把输入句分析为句法结构，再根据语义规则把语法符号结构映射到语义符号结构。

(2)基于统计的方法

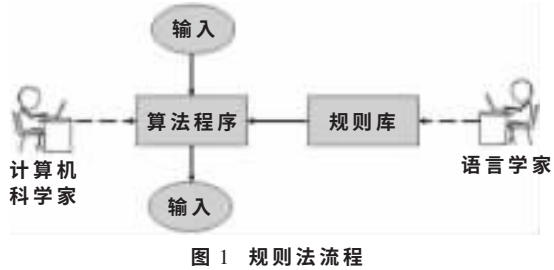

图 1 规则法流程

基于统计的方法(简称统计法)是通过对语料库中的训练数据来估计统计模型中的参数,从而建立统计性的语言处理模式。这里"语料库"由语言学家建立,计算机科学家负责建立统计模型、利用语料库训练模型参数以及编写算法解决问题,如图2所示。

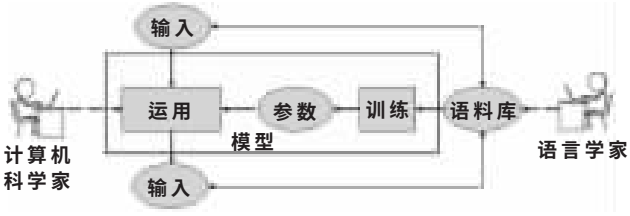

图2 统计方法的流程

(3)规则统计相结合的方法

规则统计相结合的方法(简称规则统计法)是规则法与统计法的融合,充分吸收两者的优点。规则方法易于表达复杂的语言知识且语言知识的表达较直观、灵活;但语言知识的覆盖率低,缺乏统一的语言知识冲突解决机制。而统计方法的统计模型提供了统一的冲突解决机制,且大规模数据保证了语言知识的大覆盖率;但它不善于表示复杂的、深层次的语言知识,对于数据稀缺的语言没有好的解决方案。

统计方法在发展过程中不断改进,逐渐吸收规则方法的优点来弥补自身的缺陷,统计模型趋于复杂,甚至一些模型直接建立在规则表示的基础上,从而能够表达很复杂的语言知识。两者的巧妙融合形成了规则统计法。目前来看该方法将成为计算语言学的主流方法。

4 汉语计算语言学研究的状况及存在的问题

4.1 研究的状况

属于汉藏语系的中国在计算机语言学方面的研究与应用目前还落后于印欧语系的欧美等国家,这是由汉语自身的特点以及国内计算机技术相对落后造成的。但近年来,我国的计算语言学的研究在理论以及应用方面都取得了可喜的成绩。最典型的理论方面的研究成果是黄曾阳(中国科学院声学研究所)的概念层次网络理论(Hierarchical Net Work of Concept);应用方面的代表有北京大学的《现代汉语语法信息词典》、清华大学的《汉语语素数据库》、董振东的《知网》(How Net)、潘悟云的汉语方言地理信息系统平台、中国社会科学院文学研究所的千万词级汉语语料库、台湾中央研究院的千万级古代、近代、现代汉语语料库及清华大学的《ZW大型通用汉语语料库》等。

4.2 存在的问题

虽然国外的计算语言学(自然语言处理)技术比国内先进,但国外的很多理论和方法很难照搬应用在汉语言处理上。原因在于表意体系的汉语与表音体系的印欧语在自身特征上有很大的差异。印欧语在词汇、语法、语用、语境诸层面上有明显的特征区分,相互间又有对应关系。但汉语在各层面上很难划分,特别是句法和语法间的界限相当模糊。另外,汉语没有严格意义上的形态变化(形态标记),对词没有一致认可的定义,没有明显的分词的自然形态界限。总的来说汉语的计算机处理要难于印欧语的处理,突出的几个问题是:

(1)汉语的歧义

歧义是自然语言的普遍现象,当语言形式不能完全决定语言内容(语义)时即称为歧义。在语言的语音、词汇、句法、语境上都存在歧义现象。汉语言文字是字形、字音分离的文字(不考虑有音无字的民族语),所以一字多音、一音多字现象特别多。再加上汉语词汇较难定义,句法、语法界限模糊,使得汉语的排歧相当困难。目前多是综合利用语法和语义知识,结合字典、语法规则库及上下文信息来进行排歧,但效果并不理想,特别是无法解决语境歧义。

(2)汉语语法兼类

语法兼类即词的同形异类,同一形式的词具有两种或两种以上的语法功能类别。如"连"这个词兼有副词、介词、动词、名词和量词5种词性。兼类词虽然数量不多,但出现的频率较高,且越是常用词,其兼类现象越严重。

(3)分词

多数中文句子是一长串连续的汉字(而不是以空格或其他分隔标记分开的单词),并且词汇缺少明显的形态变化[4]。

(4)词性标注

建立句法结构树的首要任务是词性标注,即明确文本中所有语法兼类词在具体语境下所属的词性。在语法平面内现有的词性标注法有:基于规则的方法、基于统计的方法、基于神经网络的方法、规则与统计混合法。

(5)电子词典

电子词典包含了语料加工处理所需的有关词的各种语言学知识,包括分词、词性标注、短语分析等。电子词典的规模和质量决定了计算机处理语言的成败,目前高质量、大规模的汉语电子词典还在建设中。

(6)规则库

语言是有规则的,规则是可以描述和处理的。规则库就是把语言学知识归纳成一套文法规则,用于判断匹配成的句子是否合法。

最典型的语言学知识表示方法有依存语法(Dependency Grammar)、格语法(Case Grammar)、语法树方法(Syntax Tree)、转换生成语法(Transformational Generative Grammar)、扩充转移网络法(Augmented Transition Network)、语义网络(Semantic Network)理论、蒙塔鸠语法(Montague Grammar)、系统语法(System Grammar)、概念依存理论(Conceptual Dependency Theory)和现代语法理论。

20世纪80年代后,国外又推出了一些新的语法理论和方法,较有影响力的有广义短语结构语法(Generalized Phrase Structure Grammar)、头驱动的短语结构语法

(Head-driven Phrase Structure Grammar)、词汇功能语法(Lexical Functional Grammar)、功能合一语法(Functional Unification Grammar)、链语法(Link Grammar)、范畴语法(Categorial Grammar)、依存语法(Dependency Grammar)和树嫁接语法(Tree Adjoining Grammar)。

而以上这些文法规则多是国外学者基于印欧语言对象的研究成果,要么完全不适用于汉语处理,要么需要系统改造后才能适用于汉语处理。

(7)统计信息库

统计信息库包含了对语料库信息的各种统计结果,如带词性标注的词频统计、邻接词同现概率统计和短语结构分布信息等,它为基于统计的语料库处理技术提供了客观的语言分布数据。这些数据可以认为是计算机从大规模语料中获得的语言学知识,不仅有助于计算机信息处理,更对语言学研究起到推动作用。同电子词典一样,汉语的统计信息库还处于发展建设阶段。

5 计算语言学的研究意义

1950 年,图灵(Alan Mathison Turing)提出了被后人称之为人工智能直接起源之一的著名的"图灵测试"。而这个测试正是机器理解人类语言的典型例子,所以有的学者把计算语言学(自然语言处理)看作是人工智能的一个分支。语言是人类智能与智慧的高度表现,因而对计算机语言学的研究也有助于人们揭开人类智能的奥秘、认识自己,为智能科学的发展和突破贡献力量。

作为一个边缘交叉学科,自然语言处理的发展受益于相关学科的发展,同时也会促进相关学科,特别是信息科学、语言学、认知学、心理学的进步。计算语言学立足于实验、理论和计算来实现计算机对语言文字信息的自动分析和理解,是实用性很强、应用范围很广的学科,它为国民经济的发展和社会的进步带来了动力。随着信息化时代的到来,特别是近几年来网络信息的大爆炸,计算语言学被认为是处理信息网络世界中语言载体的核心技术。如今,计算语言学已在机器翻译、信息检索、人机交互、语音识别、语音合成、文本分类、自动文摘、问答系统等应用领域里发挥了重要作用,这正是该学科的研究意义和实用价值所在。

语言是人类智慧的最重要特征,可以说人类的语言和大脑是世界上最复杂的两样东西,而计算机是研究它们的最有效的辅助工具。因此对计算语言学的研究及其成果的应用是人类社会发展必须且必将迈过的一道坎。另外值得一提的是,近期越来越多的学者指出汉语是世界上方言语种最多、文献资料最丰富、唯一保持历史延续性的语言,对汉语的计算语言学研究不仅对重树我国文明大国的地位有着积极的推动作用,更重要的是,未来国际计算语言学研究的突破极有可能发生在中国。

参考文献

[1] 俞士汶.计算语言学概论[M].北京:商务印书馆,2007.
[2] 冯志伟.自然语言处理的形式模型[M].合肥:中国科学技术大学出版社,2010.
[3] 江铭虎.自然语言处理[M].北京:高等教育出版社,2006.
[4] 俞士汶,黄居仁.计算语言学前瞻[M].北京:商务印书馆,2005.

(收稿日期:2012-11-26)

作者简介:
邵泽国,男,1978 年生,讲师,博士,主要研究方向:计算语言学(自然语言处理)。