计算语言学简介

俞士汶

1. 计算语言学的研究内容

社会的需求和技术的进步推动历史悠久的语言学和新兴的计算机科学相结合，产生了一门交叉学科---计算语言学(computational linguistics)。计算语言学为利用计算机处理语言信息（包括语言中信息成分的发现和提取，语言数据的存储、加工和传输，语言的翻译和理解）提供理论模型、计算方法和实现技术，因此考虑问题都是站在计算机的立场上的，这同过去以人为对象的语言研究有着明显的不同。稍微学过一点英语的中国人都不难把第一次见到的下面两句英语正确地翻译成汉语。

I bought a table with three legs.（我买了一张有三条腿的桌子。）

I bought a table with three dollars.（我花三美元买了一张桌子。）

计算机要翻译这两句话，却碰到了不易处理的歧义结构问题。尽管当代计算机的存储容量极大，但也不可能一一记住所有英语句子的汉语译文。计算机不难记住数量有限的英语句子结构及其对应的汉语句子结构。一般地说，计算机系统里有了这样的知识，就可以通过句法分析和查词典实现自动翻译了。这一步算是句法理论(syntax)的成就。

上述两个句子的结构都是

名词短语+动词+名词短语+介词短语

最后的介词短语既可以修饰句中的动词，也可以修饰后一个名词短语。要计算机针对具体的句子决定取舍，可就犯难了。因此要以某种形式给计算机灌输诸如“桌子有腿，用美元可以购物”之类的知识，并要教会计算机如何运用这些知识，这属于语义学(semantics)和语义分析的研究范围。在很多情况下，计算机死记住一些静态的知识还不能消解这种结构的歧义，还要学会分析上下文和谈话的环境，从语境中获取并活用动态的知识，这又要靠语用学(pragmatics)和语境分析(context analysis) 发挥作用了。除了要分析语言，文章生成(text generation)也是一门学问。

以上理论都是建立在基于规则的语言模型上的。基于规则的理论模型用于指导语言信息处理实践历史虽久，却常常捉襟见肘。与此同时，计算机技术飞跃进步，这又推动了基于统计模型的语料库语言学(corpus linguistics)的发展。同样，基于统计的理论模型也有其自身的局限性。 有机结合两种模型，不断地实践，并吸收相关学科（如脑科学，认知科学等）的成果，人类理解语言的奥秘总会被逐步揭开，模拟这个过程的自然语言处理系统也会逐渐接近真正理解的目标。语言信息处理是数字计算机在非数值领域的最早应用，50年来，虽历经坎坷，终究取得了长足的进步，并在社会生活中发挥作用。计算语言学从定名起，也有了30多年的历史，已成为一个稳定而且活跃的学科。

2. 研究计算语言学的意义

计算机和网络正走进千家万户，社会日益信息化，而语言是信息载体，因此语言信息处理越来越受到人们的重视。计算语言学研究积累下来的技术和资源有的已经形成产品（最有影响的可能是机器翻译产品），有的正在被集成到新的信息处理系统中。

语言科学是人文科学与自然科学之间的桥梁，而计算语言学又是其最活跃的一个分支，开展语言信息处理研究，可以带动多种学科和技术的发展。我国学者可以在汉语信息处理的这一具有天然优势的领域大有作为。

智能的本质是当代科学难题之一。在计算机上建立自然语言处理系统可以为人类了解自身的语言活动提供一个可以观察的“窗口”。自然语言理解的研究可以为智能科学的突破贡献力量。

3. 计算语言学的研究方法

3.1 重视基础设施的建设

目前计算机处理自然语言的能力还很差，很重要的一个原因在于计算机缺乏知识。建立大规模的综合型语言知识库是必不可少的基础工程。这个知识库既包括词法、句法知识，也包括语义乃至语用知识；这个知识库中的基本语言单位既有词，也有语素和短语；这个知识库既包含原始的语料库，也包含经过多级加工的语料库，知识含量高、存储格式规范的词典数据库更是必不可少的组成部分。为了实现机器翻译，这个知识库不仅包含汉语知识，还要包含汉语和其他语言的对译知识。北大计算语言学研究所积12年之努力开发出来的“现代汉语语法信息词典”可以成为综合语言知识库的组成部分。

3.2 重视适合汉语的的理论体系和计算模型的探索

这方面的研究既要提倡与国际接轨，又要重视对汉语实际情况的调查分析。北大计算语言学研究所正在研究基于词组本位语法的面向信息处理的现代汉语语法体系，希望为这方面的探索能作出一些实际的贡献。

3.3 重视应用研究

开发实用产品获得的收益可以支持理论研究和基础设施建设，使理论、基础、应用之间形成良性循环，这样的技术路线从总体上看无疑是可取的。不过，具体到一个小单位，常常会顾此失彼，这也是语言信息处理学界的苦恼。

3.4 重视人才培养

为了增强我国在语言信息处理这一高新技术领域的竞争力，大力培养计算语言学的人才，特别是青年人才是十分重要的。我国现在只能在其他一些学科（计算机科学或语言学）内培养计算语言学研究方向的博士生与硕士生。笔者希望能在一些有条件的大学试验建立计算语言学的博士点与硕士点，加速语言信息处理领域高级人才的培养。

(摘要发表于《中华读书报》1998年3月4日第6版)