语文现代化

自然语言的计算机处理

冯志伟

自然语言处理(Natural Language Processing, 简称NLP)就是利用电子计算机为工具对人类特有的书面形式和口头形式的自然语言的信息进行各种类型处理和加工的技术。这种技术现在已经形式一门专门的边缘性交叉性学科,它涉及语言学、数学和计算机科学,横跨文科、理科和工科三大知识领域。自然语言处理的目的,在于建立各种自然语言处理系统,如机器翻译系统、自然语言理解系统、情报自动检索系统、电子词典和术语数据库系统、计算机自动检索系统、语音自动识别系统、语音自动合成系统、文字自动识别系统等。由于自然语言处理离不开电子计算机,因此,自然语言处理又可以叫做"自然语言的计算机处理"(Natual Language Processing by Computers),以强调电子计算机对自然语言处理的作用。

自然语言处理又是语言文字应用的一个新课题,从语言学的观点来看,我们可以把它作为应用语言学的一个分支。

自然语言处理又是人工智能(Artificial Intelligence, 简称AI)的一个主要内容,它是电子计算机模拟人类智能的一个重要方面。因此,自然语言处理还是研制智能化电子计算机的一项基础性工作。目前,科学技术的发展突飞猛进,信息的数量与日俱增,电子计算机技术得到越来越广泛的运用,正在联成世界性的网络,并向更高的层次迈进,向智能化方向发展。智能化的电子计算机已经不是十分遥远或虚无缥缈的幻想。而是近在眼前、指目可待的现实。当前,美国、英国、日本等发达国家,都投入大量的人力、物力和财力,把智能化电子计算机的研制放在十分突出的地位。预计智能化计算机将会在本世纪末问世,并对人类社会产生不可估量的影响。它同过去人类历史上语言的出现、文字的创造、造纸技术的发明以及印刷技术的发明一样,将成为人类文明史上的又一件大事。

自然语言是人类区别于其它动物的重要标志之一。人借助于自然语言交流思想,达到互相了解,组成人类社会生活;人还借助于自然语言进行思维活动,认识事物的本质和规律,创造了人类的物质文明和精神文明。

自然语言是人脑的高级功能之一。心理学研究表明:人脑的语言功能具有一侧化的性质。它主要定位在大脑左半球,由大脑左半球控制。因此,自然语言是人类特有的一种最重要的智能。智能化电子计算机的研究离不开自然语言处理,自然语言处理的研究水平,在智能化计算机的研制中,起着举足轻重的作用。我们中国的自然语言处理工作者,应该站在电子计算机智能化的高度,以战略的眼光来看待自然语言处理技术的研究,把我国的自然语言处理提高到一个新的水平。

在电子计算机软件中,早已设计了许多人工语言,如BASIC, PASCAL, COBOL, PROLOG, LISP等程序设计语言。这些人工语言与自然语言一样,都遵循着形式语言的规律和法则。美国语言学家乔姆斯基(N. Chomsky)的形式语言理论,既适用于人工语言,也适用于自然语言。这有力地说明,自然语言与人工语言之间,在形式描述方面,确实存在着某些共同的性质。

但是,自然语言毕竟是人类历史长期发展而约定俗成的产物。它带着几千年人类历史的痕迹,比人工语言要复杂得多,因而用计算机处理起来也就困难得多。

自然语言起码在下面4个方面与人工语言

大相径庭：

1. 自然语言中充满着歧义，而人工语言中的歧义则是可以控制的；

2. 自然语言的结构复杂多样，而人工语言的结构则相对简单；

3. 自然语言的语义表达千变万化，迄今还没有一种简单而通用的途径来描述它，而人工语言的语义则可以由人来直接定义；

4. 自然语言的结构和语义之间有着千丝万缕的、错综复杂的联系，一般不存在一一对应的同构关系；而人工语言则常常可以把结构和语义分别进行处理，人工语言的结构和语义之间有着整齐的一一对应的同构关系。

由于自然语言的这些独特性质，使得自然语言处理成为人工智能的一大难题。自然语言处理的种种难题，常常使研究陷入束手无策、一筹莫展的困境中。然而，恰恰因为自然语言处理的这些困难，却吸引了一大批敢于迎着困难上、毫无畏惧的探索者。他们以战胜困难为乐，以克服困难为荣，就象科学战线上的侦察兵。正是这种对未来的坚强信念，从本世纪50年代以来，国内外学者在这个新的学科领域进行了不屈不挠的探索，历时40余年，现在已经取得了可喜的成绩。

自然语言处理有时也叫做"计算语言学"(Computational Linguistics)。本书着重论述自然语言处理的方法，当涉及到自然语言处理的基本理论的时候，我们才使用计算语言学这个术语。也就是说，自然语言处理这个术语主要用于说明方法；计算语言学这个术语主要用于说明理论。两个术语各有分工，以体现它们各自的特点。

本书共分八章。第一章至第七章讲述基本知识，第八章讲述应用系统。各章内容简述如下：

第一章讲述自然语言处理与理论语言学的关系，说明自然语言处理对语言学各个方面的深刻影响。

第二章讲述自动词法分析，以有限状态转移网络为工具，说明黏着型语言和分析型语言的自动词法分析方法，并介绍了书面汉语的自动切词与文本自动标注的方法。

第三、第四、第五章讲述自动句法分析。以递归转移网络和扩充转移网络为工具，说明基本的剖析技术，提出了"潜在歧义论"，分析了科技术语和日常语言中的潜在歧义，并介绍了良构子串表和线图分析法。

第六章讲述复杂特征理论以及合一运算方法(这是当前自然语言处理中最有影响的方法)，并介绍了中文信息处理中的多叉多标记树模型。

第七章讲述自动语义分析，介绍了义素分析法、语义场、语义网络等语义分析方法。

第八章讲述各种自然语言处理系统，使读者进一步了解到自然语言处理研究的实用价值。

本书的重点是自然语言处理的方法，而不是理论。对于自然语言处理的许多理论(如广义短语结构语法、词汇功能语法、功能合一语法等)，仅在说明方法时加以简要的介绍，不作详尽的叙述，以便提高本书的通俗性和实用性。有中等文化程度的广大读者，理解本书的内容将不会有很大的困难。

本书特别注意介绍自然语言处理中的新方法，尽可能深入地、具体地描述每一种方法的操作过程。还尽量地考虑到不同学科读者的需要，使语言学工作者可以从中了解到计算机处理自然语言的有关技术，使计算机工作者可以从中了解到现代语言学的有关知识。希望本书的出版，对于语言学工作者和计算机工作者在自然语言处理这个学科中的进一步合作，能够有所裨益。

[编者附记]《自然语言的计算机处理》一书已由上海外语教育出版社出版。此文是本书作者写的前言。计算语言学专家刘海涛高工推荐"本书内容丰富、观点新颖，对从事语言学、计算机科学、数学、逻辑学、人工智能、自然语言等领域的研究者都有重要的参考价值"。