现代化知识文库

自动翻译

知识出版社

自动翻译
知识出版社

现代化知识文库

倪海曙 主编

自动翻译

冯志伟 杨 平 编著

知识出版社

上 海

现代化知识文库

自 动 翻 译

Zidong Fanyi

冯志伟 杨 平 编著

知识出版社出版发行
(上海古北路650号)

(沪 版)

新华书店上海发行所经销 上海东方印刷厂印刷

开本 850×1035 毫米 1/32 印张 11 字数 255,000
1987 年 11 月第 1 版 1987 年 11 月第 1 次印刷
印数：1—2,500

ISBN 7-5015-5307-6 / TP·9

定价：2.30 元

内 容 提 要

自动翻译是建立在语言学、数学和计算技术这三门学科基础上的一门边缘学科，从第一次自动翻译试验30年来，这门学科已逐渐成熟。本书共分九章，力图全面地、系统地介绍自动翻译的最新成果，一至六章介绍自动翻译的基本知识，主要包括语法分析、语义描述、媒介语、数学模型、程序技术等，七至九章介绍自动翻译的分析、转换、生成的全过程。全书内容丰富，插图实例较多，可供具有高中、大专文化程度的广大读者阅读，也可以作为自动翻译初学者的一本入门书。

总　序

社会主义现代化建设需要知识，需要在不断更新中的现代化知识。

人类的知识是不断发展、不断更新的。现代的社会，文化科学突飞猛进，人类知识的更新速度空前加快；假定19世纪的知识更新周期是80～90年，现在已缩短为15年，而某些领先学科更缩短为5～10年。知识体系不断更新，人的知识结构也必须不断更新，进学校求得适用一辈子的知识的“一次教育”已经成为陈旧的观念。这样，不断地进行更新知识的再学习，也就成为现代人生活和工作的需要。“活到老，学到老”这句格言有了新的含义。现在，好些国家已经在研究和推行“终身教育”，又称为“知识更新教育”，它的主要方法是提供对最新知识的深入浅出的介绍，以便自学。现代化的人才要由实行全面的终身教育来造就。

人类认识日新月异，各门科学的新分支层出不穷，边缘性、交叉性学科随着发展，形成了人类知识结构的综合化和整体化的新趋向。因此，现代化社会不仅需要“专才”，而更需要“通才”，也就是具有新的知识结构的科学人才。现在许多成就卓著的科学家，极少是只限于一门专业的，他们往往在边缘性、交叉性学科领域中以博识多才取胜。当然，一个人不可能通晓一切知识的细节；但是，如果知识深广，视野开

阔，就可以具有融会贯通、触类旁通的创造能力。我国的现代化事业正需要成千上万这样的通才。

《现代化知识文库》就是为了提供知识更新的学习材料而出版的。它将系统地、全面地、通俗地介绍从自然科学到社会科学各个部门的最新成就，特别是边缘性、交叉性学科的新进展以及它的难题和解决的方向。《文库》的有些内容在国内还是第一次作系统介绍，希望它的出版对正在探索科学文化新境界的读者有所帮助。

这套文库将不断补充新的选题，分辑出版，每辑10本。编著者大多是中年科研人员，由老一辈的著名科学家担任编审。从内容到文体都将按照客观情况的发展不断更新。

知识就是力量，我们的工作希望得到大家的支持和帮助。

《现代化知识文库》编辑部

1982年5月

前　言

本书力图全面地、系统地介绍自动翻译这门综合性边缘学科的最新成果，着重介绍自动翻译中的语法分析、语义描述、数学模型、程序技术以及媒介语等问题，可供具有高中、大专文化程度的广大读者阅读，对于有志于从事自动翻译的青年同志来说，本书也可以作为一本入门书。

本书由两人合编，第一、二、三、四、五、六章由冯志伟执笔，第七、八、九章由杨平执笔，全书最后由冯志伟定稿。

在本书写作过程中，曾参考过国内外学者的许多著作和论文，我们在每章末都列出了主要的参考文献。如果没有学者们做的大量的研究工作，本书是写不出来的，我们仅是对他们的研究成果作了一点系统化的总结而已，在本书出版之际，我们谨对他们深致谢意。我们还要感谢我们的老师和同事们，他们给了我们很大的支持，并对本书的写作提出了许多有益的建议。

自动翻译涉及的知识面很广，我们学识有限，不妥之处，恳请广大读者批评指正。

编　者

1984 年 1 月 1 日

第 1 辑 10 种

软科学
语言文字的信息处理
电子显微术
老年学与老年病
现代语言学
二十世纪数学史话
怎样使用电脑
——程序设计的方法和技巧
国民经济结构学浅说
现代作战模拟
科学方法和科学动力学
——现代科学哲学概述

（以上书均已出版）

目　录

第一章　自动翻译概述

1．什么是自动翻译

语言是人类最重要的交际工具。在科学技术的领域内，使用不同语言的人们为了交流科技情报，达到互相了解，就需要进行科技文献的翻译。

近年来，科学技术的发展日新月异，国际科学技术交流日趋频繁，出现了世界性的“情报爆炸”现象。据统计，目前世界上一年出版的文献量约500万篇，每年按10～12%的速度持续增长，每年约增加2亿个情报单元(例如，一个数据，一个公式等等)。全世界出版的科技期刊大约7万种左右，文摘杂志2 000种左右，这些期刊每年以60～70种文字发表75万多位作者的300万篇论文，期刊总数每15年到20年翻一番，科技图书在世界各类书籍总数中占20～30%，70年代以来，全世界每年出版图书50万种以上，其中科技图书12万种，从1950～1970年20年间，图书品种增长一倍，印数增长两倍，出书品种和印数每年分别以4～6%的速度递增，全世界每一分钟就有一本书问世。

面对着这与日俱增、浩如烟海的科技文献，科技工作者为了了解各国科技发展的动态，不得不做难以数计的翻译工作，而传统的、手工式的情报工作方法和翻译方式，远远满足不了科技文献工作的需要，这就必须在翻译技术上来一个革新。恩格斯指出：“社会一旦有技术上的需要，则这

种需要就会比十所大学更能把科学推向前进。"① 正是这种技术上的迫切需要，促使着人们去寻找新的翻译手段。1946年电子计算机问世后，为翻译技术的革新提供了可能性，在实践的要求下，就出现了一门新兴的研究科目——自动翻译（automatic translation），又叫机器翻译（machine translation）。

自动翻译是建立在语言学、数学和计算技术这三门学科基础上的一门边缘学科。为了建立一个自动翻译系统，需要解决语言学、数学和计算技术三方面的问题。

（1）语言学方面的问题　在选定的学科领域内，编制机器词典，确定应该收进词典的各种语法信息；选择语法类型，决定语法分析的策略并建立机器语法；进行语义的形式化工作。

（2）数学方面的问题　制定自动翻译算法的总体结构，制定翻译过程中各个阶段的算法，制定记录语言数据以及记录算法的公式。

（3）计算技术方面的问题　建立自动翻译程序系统，编制实现算法的程序，建立各种类型的服务程序以及人机联作程序。

要解决这些问题，需要语言学家、数学家和计算技术专家共同努力，相互协作。事实上，很难要求一个自动翻译的研究人员对语言学、数学和计算技术三个方面都很在行，但是，自动翻译工作要求从事某一方面的研究人员对自己的本专业很内行，对本专业之外的其他两方面不外行。这样，自动翻译的研究人员就有必要不断地更新自己的知识结构，努力地开拓自己的科学视野，才能适应自动翻译这一门边缘学科发展的要求。

---

① 《恩格斯致符·博尔吉乌斯》，见《马克思恩格斯选集》，第4卷，第505页。

2. 自动翻译研究的兴起

关于用机器把一种语言翻译为另一种语言的想法，比计算机本身的历史还要早。远在本世纪30年代，苏联的П. П. Смирнов-Троянский(斯米尔诺夫-特洛扬斯基)和法国的G. B. Artsouni(阿尔楚尼)就提出了这样的思想。П. П. Смирнов-Троянский甚至还亲自设计了把一种语言翻译成另一种语言的机器，并在1933年9月5日向当时的苏联政府登记了他的发明。但是，他们的工作并没有引起人们的重视，他们关于用机器来进行翻译的卓越思想被埋没了十多年。

1946年，美国宾夕法尼亚州摩尔工学院的J. P. Eckert(埃克特)和J. W. Mauchly(莫奇莱)设计并制造出了世界上第一台电子计算机ENIAC。美国在第二次世界大战时，曾计算过炮弹从发射到命中目标的弹道上40个点的位置，以研究炮弹的运动规律，这种计算过去由人来做需要7个小时，而当时用ENIAC电子计算机来计算，只要3秒钟就行了。电子计算机这种惊人的计算速度，启示着人们考虑翻译技术的革新问题。与此同时，Claude Shannon(申农)建立了信息的数学理论，Norbert Wiener(维纳)提出了控制论(Cybernetics)，Pitts(皮茨)和McCullough(麦卡洛)提出了关于神经网络和大脑功能的新思想，在第二次世界大战中发展起来的解读密码的技术也更为成熟，这些都为自动翻译思想的提出准备了有利的条件。

不过，在自动翻译研究工作才开始的时候，在程序技术方面，人们还是采用手编程序，计算机语言还是使用机器语言，数组和子程序等概念还没有出现，至于诸如后进先出栈、编译程序、递归过程等等新的概念，则根本还没有。在语言研究中，也没有任何一个人听到过诸如上下文自由文法

(context-free grammar)、上下文敏感文法(context-sensitive grammar)、转换语法(transformational grammar)、扩展转移网络(augmented transition networks)等新名词。这些，又使得早期的自动翻译研究工作难以达到较高的水平，缺乏坚实的理论基础。

1946年，英国工程师A. Donald Booth(布斯)和美国工程师Warren Weaver(韦弗)在讨论电子计算机的应用范围时，第一次提出了用电子计算机进行语言自动翻译的想法。由于他们两个人都很熟悉如何根据字母频率或词的频率用计算机来解读密码的技术，因此，他们认为，似乎也可以用类似的方法来搞自动翻译，所不同者，只不过在自动翻译中需要编制一部完善的双语言词典。当然，当时他们也认识到了词典并不能解决所有的问题，这是因为：(a)许多词由于上下文的不同而有若干个不同的译法；(b) 两种语言的词序是不尽相同的；(c)成语不能逐词翻译，而必须把它们作为一个整体来进行翻译。

自动翻译的想法刚一提出，立即引起人们极大的兴趣，许多学者进行了初步的研究。1947年，Booth和D. H. V. Britten(布里顿)编出了词典的找查程序，他们编的词典是一种原形词典，每个词的全部变化形式(如love, loves, loving)都被当作单独的条目编入词典中。1948年，R. H. Richens(瑞琴斯)提出，在程序中应增加关于词的屈折变化的规则，词典中只存词干，这样可以大大地减少词典的条目。

1949年，Weaver发表了以《翻译》为题的备忘录。在这个备忘录中，他除了提出各种语言都有许多共同的特征这一论点之外，还有两点值得我们注意：第一，他认为翻译类似于解读密码的过程。他说："当我阅读一篇用俄文写的文章的时候，我可以说，这篇文章实际上是用英语写的，只不过它是用另外一种奇怪的符号编了码而已，当我在阅读

时，我是在进行解码。"第二，他认为原文与译文"说的是同样的事情"，因此，当把语言A翻译为语言B时就意味着，从语言A出发，经过某一中介的"通用语言"(universal language)或"中间语言"(interlingua)，然后转换为语言B，这种"通用语言"或"中间语言"，可以假定是全人类共同的。

在Weaver的备忘录发表之后，美国有好几个单位进行了自动翻译的研究工作。Erwin Reifler（雷夫勒）提出了在自动翻译中采用译前编辑(pre-editor)和译后编辑(post-editor)的思想。译前编辑可以不懂译语，译后编辑可以不懂原语。译前编辑的任务是准备输入文句，使其尽可能地没有歧义，同时，处理一些原语中机器不好处理的一些困难问题，译后编辑的任务是把机器译文改写成合乎语法的、可理解的译文。

1952年，由洛克菲勒基金会主持，一些英美学者在美国麻省理工学院召开了第一次机器翻译会议，会上听取和讨论了15篇报告。与会者一致认为，今后自动翻译研究可分两个阶段：首先在大范围内研究词的频率及词汇对应问题，其后再开始句法分析问题的研究。同时，会议根据Weaver提出的关于自动翻译的中间语言的想法，认为有必要开展通用语言的研究，并把这种语言取名为Machinese。会后，英美学者根据会议的精神分头进行研究。V. A. Oswald(奥斯瓦德)和R. H. Lawson(劳森)对于神经外科学的专业文献进行了词的频率分析，A. D. Booth提出了如何在自动翻译中节省查词典时间的方法．K. E. Happer（哈珀)和A. G. Oettinger（埃丁格尔）从自动翻译的角度出发，对俄语进行了多方面的研究。通过这些研究，人们对于英语和俄语的语言结构有了更为精确的了解。这样，直接在电子计算机上进行自动翻译试验的时机便成熟了。

1954年初，美国乔治敦大学在国际商用机器公司的协同下，用IBM-701电子计算机进行了世界上第一次自动翻

译试验，比较顺利地把若干个俄文句子译成了英文句子。这次试验所用的机器词典包括 250 个词，机器语法规则只有 6 条。这次试验从实践上证明了自动翻译是可能的。与此同时，世界上第一份自动翻译杂志创刊，取名为《机器翻译》，简称 MT(Mechanical Translation)。

苏联于 1954 年也开始了自动翻译研究工作。1955 年，苏联科学院精密机械与计算技术研究所利用 БЭСМ 大型通用电子计算机，进行了英俄自动翻译试验。这次试验以数学文献为材料，机器词典包括 952 个词，并制定了一套独立的语法加工系统。

此后，自动翻译工作陆续在各国开展起来。

1955 年，英国柏克培克学院利用 APEXC 计算机，进行了法英自动翻译试验，这次试验的机器词典包括 250 个词，每小时可译 1 000 个词，并通过电视进行了表演。

1956～1958 年，苏联科学院数学研究所利用 Стрела 电子计算机进行了法俄自动翻译试验，翻译了十几句话。

1959 年，日本东京电工实验室设计了世界上第一台翻译专用机 YAMATO，并利用它进行了英日自动翻译试验。这次试验的自动翻译规则系统是根据初中一、二年级教科书编制的，机器词典包括 2 000 个词，利用穿孔纸带输入，翻译了一些简单句。

这一年，苏联科学院数学研究所利用 Стрела 电子计算机，又进行了一次法俄自动翻译试验，翻译成段的文章。

1960 年，捷克斯洛伐克利用继电器数字计算机，进行了英捷自动翻译试验，只翻译了一句话。

同年，苏联里加电子学和计算技术研究所进行了俄语—拉脱维亚语的自动翻译试验，用穿孔纸带输入。

我国早在 1957 年就开始了自动翻译的研究。1959 年 9 月，中国科学院语言研究所和计算技术研究所利用我国自己制造的 104 大型通用快速电子计算机成功地进行了俄

汉自动翻译试验。这次试验的机器词典包括 2 030 个词条，机器语法规则系统由 29 个线路图表组成，试验了 9 句不同类型的较复杂的句子，解决了某些困难的词序问题，利用穿孔纸带输入，输出的是代码不是汉字。1961 年，我国学者在原有成就的基础上，完成了 61 型俄汉自动翻译规则系统的编制工作，这是我国俄汉自动翻译的第二个方案。1961 年的新方案与 1959 年的老方案比较起来，无论在机器词典和机器语法方面都有了很大的改进，而且与同时期的国外自动翻译规则系统相比，也达到了较高的水平。另外，语言研究所和北京外国语学院协作，于 1961 年初还编制了一个英汉自动翻译方案。

自动翻译试验的初步成功，使自动翻译很快地进入了高潮，各种有关的研究机构象雨后春笋般地在美国、苏联、日本和欧洲建立了起来。

这个时期，在美国建立的自动翻译机构有：

(1) 华盛顿大学西雅图机器翻译组　该组早期的研究重点是德英自动翻译，后期的研究重点是俄英自动翻译，此外，还编写过汉英自动翻译程序。该组在自动解决语法多义性问题方面作过较多的工作，同时还研究过自动翻译的译文质量的评价问题，提出过评价程序 (evaluation program)。在句法分析上，该组采用核心分析法 (kernel analysis)，根据句子中核心成分的语法特征和词序类型把它们加以分类，然后对这些核心成分进行机械定位。

(2) 麻省理工学院(简称 MIT)　他们主要研究德英及俄英自动翻译，也研究过法英自动翻译和阿拉伯语—英语的自动翻译问题。他们的句法分析主要是根据短语结构语法，研究范围较广，从程序设计技术到理论语言学，从形态学、句法学到语义学，几乎各个领域都有所涉及。为了满足程序设计自动化的需要，该组主要成员 V. Yngve(英格维)设计了专门用于自动翻译的 COMIT 语言。