百种语文小丛书

曹先擢 主编

机器翻译今昔谈

JIQI FANYI JINXI TAN

冯志伟 著

语 文 出 版 社
http :// www. ywcbs. com

百种语文小丛书

JIQI FANYI JINXI TAN

机器翻译今昔谈

冯志伟 著

YUWEN CHUBANSHE
语文出版社

机器翻译今昔谈

冯志伟

圣经《创世纪》中说，古代人类说的原是一种统一的语言，交流思想非常方便，劳动效率也很高，他们曾经想建立一座高达天庭的通天塔，叫做“巴比塔”，来显示他们的丰功伟绩。建造巴比塔的壮举震惊了上帝，上帝便施伎俩，让不同的人说不同的语言，使人们难于交流思想，无法协调工作，以此来惩罚异想天开的巴比塔建造者。结果，巴比塔没有建成，而语言的不同，却成为人们相互交往的极大障碍。这样的传说当然是不可信的，但是，语言的障碍却时时刻刻在困扰着人们。

现在我们已经进入了信息化的时代，语言是信息的最主要的负荷者，如何有效地使用现代化手段来突破人们之间的语言障碍，成为了全人类面临的共同问题。机器翻译采用电子计算机来进行不同语言之间的自动翻译，是解决这个问题的有力手段之一。机器翻译有可能成为消除人们语言障碍的真正的通往理想境界的巴比塔。由于自然语言的极端复杂性，机器翻译是当代科学技术的十大难题之一。

可以毫不夸张地说，在进入21世纪之后，几乎每一个生活在信息网络时代的现代人，都要直接或间接地与机器翻译打交道，因此，就有必要了解机器翻译的基本知识，本书不可能讲述机器翻译的各种复杂技术，只想从语言应用和多语言交际的角度，尽量用不带技术色彩的语言，以最为浅显的方式向读者介绍一下机器翻译发展的曲折道路，使读者对于机器翻译有一个初步的认识，从而产生兴趣，并投身到这个语言学领域中唯一被列入当代科学技术十大难题的研究中去，贡献出自己的聪明和智慧。

机器翻译就是用计算机来进行不同自然语言之间的翻译，它是自然语言计算机处理的一个历史悠久的部门，是横跨语言学、数学、计算机科学的综合性学科，也是信息时代语言应用的一个重要领域。随着计算机网络的迅速普及和推广，随着信息高速公路的发展，网络上不同语言之间交际越来越普遍，语言的障碍也显得越来越严重，机器翻译是克服信息时代的语言障碍的不可缺少的手段，它在现代信息社会中的巨大作用将会越来越明显。

关于用机器来进行语言翻译的想法，远在古希腊时代就有人提出过了。当时，人们曾经试图设计出一种理想化的语言来代替种类繁多形式各异的自然语言，以利于在不同民族的人们之间进行思想交

流，曾提出过不少方案，其中一些方案就已经考虑到了如何用机械手段来分析语言的问题。

在17世纪，一些有识之士提出了采用机器词典来克服语言障碍的想法。笛卡儿（Descartes）和莱布尼兹（Leibniz）都试图在统一的数字代码的基础上来编写词典。在17世纪中叶、贝克（Cave Beck）、基尔施（Athanasius Kircher）和贝希尔（Johann Joachim Becher）等人都出版过这类的词典。由此开展了关于“普遍语言”的运动，一些人试图在逻辑原则和图形符号的基础上，创造出一种无歧义的语言，这样一来，人们就不必再由于误解而产生交际方面的困惑了。维尔金斯（John Wilkins）在《关于真实符号和哲学语言的论文》（An Essay towards a Real Character and Philosophical Language，1668）中提出的中介语（Interlingua）是这方面最著名的成果，这种中介语的设计试图将世界上所有的概念和实体都加以分类和编码，有规则地列出并描述所有的概念和实体，并根据它们各自的特点和性质，给予不同的记号和名称。

1903年，古图拉特（Couturat）和洛（Leau）在《通用语言的历史》一书中指出，德国学者里格（W. Rieger）曾经提出过一种数字语法（Zifferngrammatik），这种语法加上词典的辅助，可以利用机械将一种语言翻译成其他多种语言，首次

使用了“机器翻译”（德文是 ein mechanisches Uebersetzen）这个术语。

本世纪三十年代之初，亚美尼亚裔的法国工程师阿尔楚尼（G. B. Artsouni）提出了用机器来进行语言翻译的想法，并在 1933 年 7 月 22 日获得了一项“翻译机”的专利，叫做“机械脑”（mechanical brain）。这种机械脑的存储装置可以容纳数千个字元，通过键盘后面的宽纸带，进行资料的检索。阿尔楚尼认为它可以应用来记录火车时刻表和银行的帐户，尤其适合于作机器词典。在宽纸带上面，每一行记录了源语言的一个词项以及这个词项在多种目标语言中的对应词项，在另外一条纸带上对应的每个词项处，记录着相应的代码，这些代码以打孔来表示。要查询的词项也利用键盘打孔来表示，检索一个词项的时间大约是 10 到 15 秒。阿尔楚尼的原型机于 1937 年正式展出，引起了法国邮政、电信部门的兴趣。但是，由于不久爆发了第二次世界大战，阿尔楚尼的机械脑无法安装使用。

1933 年，苏联发明家特洛扬斯基（П. П. ТРОЯНСКИЙ）设计了用机械方法把一种语言翻译为另一种语言的机器，并在同年 9 月 5 日登记了他的发明。特洛扬斯基认为翻译可以分为三个阶段，第一个阶段由只懂源语言的编辑，将输入的原

文分析成特定的逻辑形式，将带有屈折词尾的变形词还原成原形词，并分析出各个单词的句法功能，为此，他创造了一套逻辑分析符号。第二阶段是利用他的翻译机，把源语言的原形词和逻辑符号转换成目标语言的原形词和符号。第三阶段由只懂目标语言的编辑，把目标语言的原形词和符号转换成目标语言。特洛扬斯基认为，他的翻译机只能在第二阶段作为自动词典来使用。不过他相信，只要能够建造出一部专门处理逻辑分析过程的机器，总有一天，上述的整个翻译程序都能够用机器来实现。特洛扬斯基这种认识，已经超越了“机器词典”的简单想法，比阿尔楚尼又进了一步。1939 年，特洛扬斯基在他的翻译机上增加了一个用“光元素”操作的存储装置；1941 年 5 月，这部实验性的翻译机已经可以运作；1948 年，他计划在此基础上研制一部“电子机械机”（electro – mchanical machine）。但是，由于当时苏联的科学家和语言学家对此反映十分冷淡，特洛扬斯基的翻译机没有得到支持，最后以失败告终了。

1946 年，美国宾夕法尼亚大学的埃克特（J. P. Eckert）和莫希莱（J. W. Mauchly）设计并制造出了世界上第一台电子计算机 ENIAC，电子计算机惊人的运算速度，启示着人们考虑翻译技术的革新问题。因此，在电子计算机问世的同一年，

英国工程师布斯（A. D. Booth）和美国洛克菲勒基金会副总裁韦弗（W. Weaver）在讨论电子计算机的应用范围时，就提出了利用计算机进行语言自动翻译的想法。1947 年 3 月 6 日，布斯与韦弗在纽约的洛克菲勒中心会面，韦弗提出，“如果将计算机用在非数值计算方面，是比较有希望的”。在韦弗与布斯会面之前，韦弗在 1947 年 3 月 4 日给控制论学者维纳（N. Wiener）写信，讨论了机器翻译的问题，韦弗说：“我怀疑是否真的建造不出一部能够作翻译的计算机？即使只能翻译科学性的文章（在语义上问题较少），或是翻译出来的结果不怎么优雅（但能够理解），对我而言都值得一试。”可是，维纳给韦弗泼了一瓢冷水，他在 4 月 30 日给韦弗的回信中写道：“老实说，恐怕每一种语言的词汇，范围都相当模糊；而其中表示的感情和言外之意，要以类似机器翻译的方法来处理，恐怕不是很乐观的。”不过韦弗仍然坚持自己的意见。1949 年，韦弗发表了一份以《翻译》为题的备忘录，正式提出了机器翻译问题。在这份备忘录中，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：“当我阅读一篇用俄语写的文章的时候，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。”备忘录中记载了一个有趣的故事，布朗大学数学系的吉尔曼（R. E. Gilmam）曾经解读了一篇长约100个词的土耳其文密码，而他既不懂土耳其文，也不知道这篇密码是用土耳其文写的。韦弗认为，吉尔曼的成功足以证明解读密码的技巧和能力不受语言的影响，因而可以用解读密码的办法来进行机器翻译。

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“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。

可以看出，韦弗把机器翻译仅仅看成一种机械的解读密码的过程，他远远没有看到机器翻译在词法分析、句法分析以及语义分析等方面的复杂性。

由于学者的热心倡导，实业界的大力支持，美国的机器翻译研究一时兴盛起来。1954年，美国乔治敦大学在国际商用机器公司（IBM公司）的协同下，用IBM—701计算机，进行了世界上第一次机器翻译试验，把几个简单的俄语句子翻译成英语，接着，苏联、英国、日本也进行了机器翻译试验，机器翻译出现热潮。

早期机器翻译系统的研制受到韦弗的上述思想的很大影响，许多机器翻译研究者都把机器翻译的过程与解读密码的过程相类比，试图通过查询词典的方法来实现词对词的机器翻译，因而译文的可读性很差，难于付诸实用。

1964 年，美国科学院成立语言自动处理咨询委员会（Automatic Language Processing Advisory Committee，简称 ALPAC 委员会），调查机器翻译的研究情况，并于 1966 年 11 月公布了一个题为《语言与机器》的报告，简称 ALPAC 报告，对机器翻译采取否定的态度，报告宣称："在目前给机器翻译以大力支持还没有多少理由"；报告还指出，机器翻译研究遇到了难以克服的"语义障碍"（semantic barrier）。

在 ALPAC 报告的影响下，许多国家的机器翻译研究进入低潮，许多已经建立起来的机器翻译研究单位遇到了行政上和经费上的困难，在世界范围内，机器翻译的热潮突然消失了，出现了空前萧条的局面。

不过，尽管在萧条时期，法国、日本、加拿大等国，仍然坚持着机器翻译研究，于是，在 70 年代初期，机器翻译又出现了复苏的局面。

如果我们把从 1954 年第一次机器翻译试验到 ALPAC 报告发表后出现的萧条看成是机器翻译的

草创期（1954 年—1970 年），那么，从 70 年代初期开始，机器翻译便进入了它的复苏期（1970 年—1976 年）。

在这个复苏期，研究者们普遍认识到，原语和译语两种语言的差异，不仅只表现在词汇的不同上，而且，还表现在句法结构的不同上，为了得到可读性强的译文，必须在自动句法分析上多下功夫。

早在 1957 年，美国学者英格维（V. Yingve）在《句法翻译的框架》（Framework for syntactic translation）一文中就指出，一个好的机器翻译系统，应该分别地对原语和译语都作出恰如其分的描写，这样的描写应该互不影响，相对独立。英格维主张，机器翻译可以分为三个阶段来进行。

第一阶段：用代码化的结构标志来表示原语文句的结构；

第二阶段：把原语的结构标志转换为译语的结构标志；

第三阶段：构成译语的输出文句。

第一阶段只涉及原语，不受译语的影响，第三阶段只涉及译语，不受原语的影响，只是在第二阶段才设计到原语和译语二者。在第一阶段，除了作原语的词法分析之外，还要进行原语的句法分析，才能把原语文句的结构表示为代码化的结构标志。

在第二阶段，除了进行原语和译语的词汇转换之外，还要进行原语和译语的结构转换，才能把原语的结构标志变成译语的结构标志。在第三阶段，除了作译语的词法生成之外，还要作译语的句法生成，才能正确地输出译文的文句。

英格维的这些主张，在这个时期广为传播，并被机器翻译系统的开发人员普遍接受，因此，这个时期的机器翻译系统几乎都把句法分析放在第一位，并且在句法分析方面取得了很大的成绩。

这个时期机器翻译的另一个特点是语法（grammar）与算法（algorithm）分开。

早在1957年，英格维就提出了把语法与“机制”（mechanism）分开的思想。英格维所说的“机制”实质上就是算法。所谓语法与算法分开，就是要把语言分析和程序设计分开，程序设计工作者提出规则描述的方法，而语言学工作者使用这种方法来描述语言的规则。语法和算法分开，是机器翻译技术的一大进步，它非常有利于程序设计工作者与语言工作者的分工合作。

这个复苏期的机器翻译系统的典型代表是法国格勒诺布尔理科医科大学应用数学研究所（IMAG）自动翻译中心（CETA）的机器翻译系统。这个自动翻译中心的主任沃古瓦（B. Vauquois）教授明确地提出，一个完整的机器翻译过

程可以分为如下六个步骤：

（1）原语词法分析

（2）原语句法分析

（3）原语译语词汇转换

（4）原语译语结构转换

（5）译语句法生成

（6）译语词法生成

其中，第一、第二步只与原语有关，第五、第六步只与译语有关，只有第三、第四步牵涉到原语和译语二者。这就是机器翻译中的“独立分析—独立生成—相关转换”的方法。他们用这种研制的俄法机器翻译系统，已经接近实用水平。

他们还根据语法与算法分开的思想，设计了一套机器翻译软件 ARIANE—78，这个软件分为 ATEF，ROBRA，TRANSF 和 SYGMOR 四个部分。语言工作者可以利用这个软件来描述自然语言的各种规则。其中，ATEF 是一个非确定性的有限状态转换器，用于原语词法分析，它的程序接收原语文句作为输入，并提供出该文句中每个词的形态解释作为输出；ROBRA 是一个树形图转换器，它的程序接收词法分析的结果作为输入，借助语法规则对此进行运算，输出能表示文句结构的树形图；ROBRA 还可以按同样的方式实现结构转换和句法生成；TRANSF 可借助与双语词典实现词汇转换；

SYGMOR 是一个确定性的树—链转换器，它接收译语句法生成的结果作为输入，并以字符链的形式提供出译文。

通过大量的科学实验的实践，机器翻译的研究者们认识到，机器翻译中必须保持原语和译语在语义上的一致，也就是说，一个好的机器翻译系统应该把原语的语义准确无误地在译语中表现出来。这样，语义分析在机器翻译中越来越受到重视。

美国斯坦福大学威尔克斯（Y. A. Wilks）提出了“优选语义学”（preference semantics），并在此基础上设计了英法机器翻译系统，这个系统特别强调在原语和译语生成阶段，都要把语义问题放在第一位，英语的输入文句首先被转换成某种一般化的通用的语义表示，然后再由这种语义表示生成法语译文输出。由于这个系统的语义表示方法比较细致，能够解决仅用句法分析方法难于解决的歧义、代词所指等困难问题，译文质量较高。

本世纪 70 年代末，机器翻译进入了它的第三个时期——繁荣期（1976 年——现在）。繁荣期的最重要的特点，是机器翻译研究走向了实用化，出现了一大批实用化的机器翻译系统，机器翻译产品开始进入市场，变成了商品，由机器翻译系统的实用化引起了机器翻译系统的商品化。

机器翻译的繁荣期是以 1976 年加拿大蒙特利

尔大学与加拿大联邦政府翻译局联合开发的实用性机器翻译系统 TAUM—METEO 正式提供天气预报服务为标志的。这个机器翻译系统投入实用之后，每小时可以翻译 6 万—30 万个词，每天可以翻译 1500—2000 篇天气预报的资料，并能够通过电视、报纸立即公布。TAUM—METEO 系统是机器翻译发展史上的一个里程碑，它标志着机器翻译由复苏走向了繁荣。

日本富士通公司开发的 ATLAS—Ⅰ（Automatic Translation System—Ⅰ）系统是一个建立在大型计算机上的英日机器翻译系统，该系统以句法分析为中心，可进行科学技术文章的翻译，在 FACOM M380 计算机上，每小时可翻译 60000 词。

日本富士通公司开发的 ATLAS—Ⅱ机器翻译系统也建立在大型计算机上，但其翻译方式与 ATLAS—Ⅰ不同。ATLAS—Ⅰ以句法分析为中心，而 ATLAS—Ⅱ则以语义分析为中心。该系统建立了用于表示概念之间关系和客观世界知识的“世界模型”，在译文生成时，特别注意单词之间的搭配关系和邻接关系，在机器翻译过程中，采用一种叫做“概念构造”的中间语言来作为原语和译语的共同表达。该系统目前用于日英机器翻译。

此外，日本的实用化机器翻译系统还有：日立公司开发的 HICATS（Hitachi Computer Aided

Translation System）英日、日英机器翻译系统，日本电气公司开发的 PIVOT 英日、日英机器翻译系统，三菱电机公司开发的 MELTRAN 日英机器翻译系统，冲电气公司开发的 PENSEE 日英机器翻译系统，理光公司开发的 RMT 英日机器翻译系统，三洋电气公司开发的 SWP—7800 日英机器翻译系统，东芝公司开发的 TAURAS 英日机器翻译系统，日本布拉维斯公司（BRAVICE INTERNATIONAL）研制的 BRAVICE PAK 11/73 日英机器翻译系统等。

欧美除 TAUM—METEO 机器翻译系统之外，还陆续推出了一批实用化的机器翻译系统。

法国纺织研究所的 TITUS—Ⅳ系统，可以进行英、德、法、西班牙等四种语言的互译，每种语言都有一部 14000 个词的机器词典，每秒钟可译 240 个词，主要用于翻译纺织技术方面的文献。

美国在乔治敦大学机器翻译系统的基础上，进一步开发了大型的机器翻译系统 SYSTRAN，已提供试用。例如，提供给美国空军的 SYSTRAN 系统，词典有 168000 个词干形式和 136000 个词组，可进行俄英机器翻译，每小时可翻译 150000 词；提供给美国拉特塞克（Latsec）公司的 SYSTRAN 系统，可进行俄英、英俄、德英、汉法、汉英机器翻译，每小时可译 30 万—35 万个词。SYSTRAN 是目前应用最为广泛、所开发的语种最为丰富的一

个实用化机器翻译系统。

美国罗各斯（LOGOS）公司开发的 LOGOS—Ⅲ机器翻译系统，可进行英语—越南语机器翻译和英俄机器翻译，词典有 10 万个词。

美国国家航空和航天的 NASA 系统，可进行俄英和英俄机器翻译。

美国魏德纳（WEIDNER）通讯公司 WCC 的 WEIDNER 机器翻译系统，可进行英语与法语、英语与德语、英语与西班牙语、英语与葡葡牙语之间的双向机器翻译，并可进行英语—阿拉伯语的单向机器翻译。

设在华盛顿的泛美卫生组织研制成的 PAHO 系统，可进行西班牙语—英语的机器翻译。从 1980 年以来，已经翻译了 100 多万词的资料。近来，他们又推出了 ENGSPAN 和 SPANAM 两个实用化系统。

德国西门子（SIMENS）公司与美国德克萨斯大学（Texas University）合作，研制成 METAL 系统，可进行德英机器翻译，词典包含 1 万个词条。

德国萨尔大学（Universitaet des Saarlandes）研制成 SUSY（Saarbruecken Automatic Translation System）系统，以德语为中介，可以进行俄语、英语、法语、世界语的机器翻译。比如，由英语译成法语，首先要由英语译成德语，再由德语译成法