机器翻译技术丛书

机器翻译系统

杨沐昀 李志升 于浩 编著

哈尔滨工业大学出版社

机器翻译技术丛书之二

机 器 翻 译 系 统

杨沐昀　李志升　于浩　编著

哈尔滨工业大学出版社
2000·哈尔滨

内 容 提 要

本书比较全面地回顾了50年来机器翻译系统的发展历程，探讨了各种机器翻译系统的特点和组成原理，并着重介绍了目前我国用户比较关心的英汉和汉英机器翻译系统，力图使读者对于目前的机器翻译系统有一个比较全面的了解。

本书不仅介绍了当前各种机器翻译系统的特点和一般使用方法，而且本书的作者结合10余年来机器翻译系统的开发经验，比较详细地阐述了英汉、汉英、翻译工作站等机译系统的总体框架和实现方法。因此，本书不但适于初学者初步了解机器翻译系统，而且对于专业的研究人员和系统开发者也具有参考价值。

机器翻译技术丛书之二

机 器 翻 译 系 统

Jiqi Fanyi Xitong

杨沐昀 李志升 于浩 编著

*

哈尔滨工业大学出版社出版发行

哈尔滨市南岗区教化街21号

邮编150006 电话0451－6414749

哈工大出版社电脑排版中心排版

地矿部黑龙江测绘印制中心印刷厂印刷

*

开本850×1168 1/32 印张5.75 字数160千字

2000年6月第1版 2000年6月第1次印刷

印数1～3 000

ISBN 7-5603-1469-4/TP·140 定价12.00元

前　言

机器翻译系统(Machine Translation System,简称 MTS)是人类利用计算机对自然语言进行翻译加工的工具,一般就是指机器翻译软件。机器翻译系统的实现,要涉及语言学、数学和计算机科学,横跨文科、理科和工科三大知识领域,所以,一个完美的机器翻译系统至今仍是研究者和开发者梦寐以求的理想。但是完善的机器翻译系统一旦成功,将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,成为人类文明史上的又一座里程碑。

本书是《机器翻译原理》的姊妹篇,重点介绍了从 1954 年第一个机器翻译系统诞生后 50 年来机器翻译系统的发展历程、各种机器翻译系统类型、典型的机器翻译系统及其应用,并论述了我国市场上最主要的两种机器翻译系统——英汉和汉英系统的组成及实现。全书共分 7 章。第 1 章讲述了机器翻译系统的发展概况和各种类型系统的特点。第 2 章和第 3 章分别介绍了英汉和汉英机器翻译系统的组成及实现,并简要介绍了它们的一般使用方法。第 4 章简要介绍了与汉语有关的其他机器翻译系统的研制情况,并重点讲述了多语机器翻译系统的概况。第 5 章介绍了近年来刚刚出现的一种新型机器翻译系统——口语机器翻译系统的概况及实现原理。第 6 章讲述的是机器辅助翻译系统的原理及使用,此类系统也在实际中有大量应用。第 7 章介绍了机器翻译系统的应用领域、怎样选择机器翻译系统以及机器翻译系统的未来。

本书的对象是有志于开展机器翻译应用和其他对机器翻译系统感兴趣的读者。虽然机器翻译系统要涉及计算机、数理统计和

语言学等多个专业，但是本书并不需要读者精通这些专业知识，对于无法避免的各种术语力求通俗易懂，因此本书可以看作是了解机器翻译的一本入门书籍。特别值得指出的是，由于本书的编写者大部分是哈尔滨工业大学机器翻译研究实验室的一线科研人员，所以他们将该实验室10余年来开发汉英系统、英汉系统、汉英双向系统、翻译工作站以及口语系统等不同的机器翻译系统的体会加以总结，并将其中主要的经验呈现在读者面前，希望能够对为未来从事和机器翻译系统有关工作的人员有所裨益，对于致力于计算语言学专业的学生和科研人员也具有一定参考价值。

参加本书编写工作的有杨沐昀、李志升(第1、第7章)、张晶(第2章)、姚建民(第3章)、于浩(第4章)、吕雅娟(第5章)、赵铁军教授(第6章)。河北师范大学的董立柱、康振国、孟俊茂、卢振林也参与了本书的编写工作。本书在写作过程中，得到了河北师范大学外语学院有关领导的大力支持和帮助，在此向他们表示衷心的感谢。由于编者的水平所限，书中难免出现疏漏甚至错误，欢迎同行专家和读者予以指正。

作　者

2000年3月于哈尔滨工业大学

第1章

机器翻译系统概况

圣经《创世纪》中说，在人类讲着同一种语言的时候，他们曾经想建造一座高达天庭的通天塔，叫做巴比塔（Babel Tower）。不料这一举动震惊了上帝，上帝便施展神通，让人们说不同的语言，使人们难以交流思想，无法协调工作，借此来惩罚异想天开的人类。结果，不仅巴比塔没有建成，而语言的不同，却又成为了人们相互交往的极大障碍。传说当然不足以解释世界上出现纷繁复杂的语言的缘由，但是语言的障碍却从古至今一直困扰着人们。

为了克服语言障碍，人们曾想通过翻译（自古有之）、通过设计国际辅助语（自莱勃尼茨的万国通用文字算起已经有300多年了，提出的方案已有150种以上）等途径来解决它，然而效果仍然难以令人满意。由于科学技术的日新月异，运输工具的日益发达，各民族间文化交流越来越频繁，语言障碍问题相对说来也就越来越严重。看来完全从人身上打主意恐怕是无法彻底消除语言壁垒，克服语言障碍的有效措施还是求助于机器。电子计算机的出现，为解决这个问题带来了新的曙光。还在计算机刚刚诞生的1946年，人们便开始讨论用它作翻译的问题了。由此而产生了一种新的人工智能系统——机器翻译系统（machine translation system，简称MTS）。

机器翻译系统，简单地说就是能够完成机器翻译功能的软件系统。目前，由于机器翻译研究对社会经济发展有着潜在的价值，各国对机器翻译十分重视，纷纷投资。欧洲共同体为了把EURO-

TRA 多语言机器翻译系统实用化，五年内投资 2 800 万美元。法国制定了一个 ESOPE 机器翻译计划，用于 ARIANE 系统的实用化，投资 5 600 万法郎。日本对机器翻译的专项投资为 140 亿日元(约相当于 1.27 亿美元)，其中，仅 CICC 计划的投资就达 62 亿日元，京都大学的 μ 计划投资 1 亿日元。而日本新一代计算技术研究所的第五代计算机系统，计划 10 年之内投资 5 亿美元，其中包括研制一个实用的日英机助翻译系统，编制 10 万词的日英机器词典，机助翻译正确率要求达到 90%。我国对机器翻译也进行了巨额的投资。“七五”计划期间，投资人民币 260 万元。可以说，经过这几十年的研究，工程师们研制并完成了一个又一个机器翻译系统，一座现代的“通天塔”的雏形已经耸立在人们的面前。

1.1 国外机器翻译系统的发展

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，人们曾经设计各种方案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，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。

20 世纪 30 年代初，法国科学家阿尔楚尼(G. B. Artsouni)提出了用机器来进行语言翻译的想法。1933 年，苏联发明家特洛扬斯基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。但是，由于 30 年代的技术水平还很低，特洛扬斯基的翻译机没有制成。

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，它惊人的运算速度，启发了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)和洛克

菲勒基金会的美国工程师韦弗(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。随后,Weaver 先生曾于 1947 年的各类交谈和信件中多次谈到机器翻译的设想,并于 1949 年 7 月发表了其有着广泛和深远影响的备忘录——《翻译》,正式提出了机器翻译问题[Hutchins,1993]。

在此后的 50 年间,伴随着机器翻译研究的曲折历程,机器翻译系统也经历了一个类似的起伏跌宕的发展过程(关于机器翻译研究的历史发展,请参阅本书的姊妹篇《机器翻译原理》的第一章有关章节)。本节将分以下几个阶段讲述机器翻译系统的发展:第一个 MTS 的诞生;以直接翻译方法为基础的第一代 MTS 的活跃期和 ALPAC 报告之后的低潮;20 世纪 70 年代中期的第二代 MTS 的研制;20 世纪 80 年代末第三代 MTS 的出现。MTS 的发展历程可由图 1.1 表示。

图 1.1　机器翻译的发展过程

1.1.1　第一个机器翻译系统——Georgetown-IBM 系统(1947 ~ 1954)

Weaver 的备忘录引起了人们对机器翻译的广泛注意。在美国麻省理工学院(Massachusetts Institute of Technology,简称 MIT),第一个全职的 MT 研究者 Bar - Hillel 到任,他对 MT 的早期发展有很大贡献。由于美英研究者的促进和洛克菲勒基金会赞助,MIT 于 1952 年 6 月举办了第一次 MT 会议,此后美国的 MT 研究开始加

速,真正在计算机上实现翻译系统。

在 1954 年 1 月,由 Georgetown 大学和 IBM 公司合作,实现了世界上第一个超出了词-词对译概念的真正 MT 系统,并于在 IBM 纽约总部进行了俄译英公开演示。他们用 IBM-701 计算机,把几个简单的俄语句子翻译成英语。尽管这个系统很小,只有 250 条俄语词汇、6 条语法规则以及精心挑选的翻译例句,但是第一次向公众和科学界展示了 MT 的可行性,并且激发了机器翻译系统研制的第一次热潮。

1.1.2 第一代机器翻译系统从繁荣到低潮(1954 ~ 1975)

从 50 年代中期开始,在美国掀起了 MT 研究的高潮。这一时期的机译系统普遍采用直接翻译(Direct Translation)方法,一般都没有进行很好的源语言句法结构分析,而是主要以词典为驱动,利用词典中的语法和语义特征来实现翻译。直接翻译方法的特点是在源语言分析阶段和目标语言综合(即生成)之间没有明确的区分,这样的系统也被称为第一代机器翻译系统。

1964 年美国国家科学院(NAS)成立了自动语言处理咨询委员会(Automatic Language Processing Advisory Committee,简称 ALPAC)。然而 1966 年的 ALPAC 报告却建议不要再对 MT 进行更多投资了,从此机器翻译系统的研制进入了 10 年的低潮时期。值得注意的是,机器翻译系统的研制在此期间却没有真正停止过,很多著名的机器翻译系统正是在低潮中开始孕育的。遗憾的是,这期间没有形成什么真正商品化的机器翻译系统,比较著名的只有该时期诞生的 Systran 系统(该系统在稍后进行了彻底的改进并得到很好的应用,参见本书 1.1.3)。可以说第一代机器翻译系统大部分都是小规模的实验系统,研制概况见表 1.1。

表 1.1　第一代机器翻译系统研制概况

单位名称	开始时间	负责人	机器翻译系统研制概况
美国			
华盛顿大学远东和斯拉夫语言系	1949	E. Beifler	研究俄英机译系统和德语复合名词等问题;1960 年夏开始汉英机译研究计划,翻译了 23 个学科的 39 篇文章,并对其中的词典和结构信息进行了研究
华盛顿大学社会学和人类学系	1961	A. H. Smith	研究日英机译系统
Rand 公司	1950	D. G. Hays	研究俄英机译问题
麻省理工学院	1951	V. H. Yngve	研究德英机译方面的问题,并开始研究阿拉伯语 - 英语翻译
乔治敦大学	1952	L. E. Dostert	研究俄英、法英机译,1960 年秋开始研究汉英机译系统
哈佛大学	1953	A. G. Oettinger	研究俄英 MTS,词典包括 300 000 个词条,约代表 15 000 个俄语词
密歇根大学	1955	Kou-tsoudas	着重理论研究
Ramo-Woold-ridge 公司	1955		制定高质量的俄英机译系统,着重语义和句法研究
国家度量衡局	1958	I. Rhodes	制定了一个为俄英机译实际应用而努力的长远规划
韦恩州大学	1958	A. H. Josselson	研究俄英机译系统
加里弗尼亚大学	1958	S. M. Lamb	着重研究俄英、汉英机译问题

续表

单位名称	开始时间	负责人	机器翻译系统研制概况
得克萨斯大学	1958	W.P.Lehmann	根据数学模型研究机器翻译的通用系统，已用德文和英文检验这个系统
Arthur D. Little 公司	1959	V. E. Giuliano	研究俄英翻译的程序自动化和机器运算综合过程
Planning Research 公司	1959	H. P. Edmundson	考虑机器装置问题
Autonetics	1961	H. J. Wolf	研究俄英机译系统
机译公司（华盛顿）		A. L. Loewenthal	研究俄英、俄德机译问题
其他研制单位：原子能委员会（橡树岭）、美国空军（Dayton）、国家航空和宇宙航行局（休斯顿）、Brigham Young 大学、耶鲁大学、Latsec 公司、世界翻译中心（拉霍亚）、Logos 开发有限公司（新汉普顿）、Smart 通信公司（纽约）			
加拿大研制单位：Lakehead 大学、蒙特利尔大学			
墨西哥			
Autonoma 大学电子计算中心		S. F. Beltràn	根据加利弗尼亚大学机译组编制的俄语词典来编制俄语－西班牙语机译词典。另外，还对西班牙语和俄语结构进行研究
苏联			
苏联科学院精密仪器与计算技术研究所	1954		在 ГЭСМ 机器上，以数学文献为材料，词典包括 952 个词，语法加工也比较复杂。后来还研究了英俄、德俄、汉俄、日俄机译系统
斯切克洛夫数学研究所	1955		在 Стрела 机上，试验了几十句的法俄翻译，还研究了英俄机译系统和程序设计自动化；1959 年进行了成段文章的翻译试验

续表

单位名称	开始时间	负责人	机器翻译系统研制概况
语言研究所应用语言小组	1956		研究媒介语和法俄/匈俄机译系统。
电模拟实验室	1956		为机译目的，对俄语进行分析，重点在句段分析
莫斯科第一外国语师范学院（机器翻译联合会）	1957		组织机译研究工作，研究机器的语义学
列宁格勒大学机译实验室	1958		进行媒介语探索研究，计划将 28 种语言（俄/英/法/德/意/西/捷/波/罗/印 地/瑞典/印 度/越/缅/日/汉/土/蒙/阿 拉 伯/斯瓦希利/豪萨/朝/匈/吉尔吉斯/孟加拉/列特/爱沙尼亚/亚美尼亚）译成媒介语，并从媒介语译成德/印尼语。
高尔基大学	1958	B. A. Arpaeb	研究从外语译成俄语和从俄语译成外语的机器翻译规则系统
无线电物理学院			
亚美尼亚科学院计算中心	1958		研究俄语－亚美尼亚语和亚美尼亚语－俄语翻译规则系统
格鲁吉亚电子、自动化和遥控力学研究所	1958		研究俄－格机译系统，并进行格鲁吉亚语的分析研究
基辅大学计算中心	1959		

续表

单位名称	开始时间	负责人	机器翻译系统研制概况
立陶宛维尔纽斯大学	1959		研究俄－立陶宛语机译，统计立陶宛语各方面的数据
里加电子学和计算研究所	1960		利用独创的电子计算机，采用穿孔纸带输入，研究俄语－拉脱维亚语机器翻译系统
其他研制单位：中央专利情报研究所（莫斯科）、原子能情报中心（莫斯科）、电力工程学院文献中心（莫斯科）、苏联科学技术文献资料翻译中心（莫斯科）			
法国			
国家科学研究中心机器翻译研究中心	1958（1960）	J. Peres	分两部分（巴黎和格勒诺布尔）进行工作，主要研究俄法机译问题，也涉及德法和日法机译工作
自动翻译和应用语言学研究和发展联合会	1959	E. Delavenay	研究英法机译问题，进行各种情报工作，出版会刊
Nancy大学自动语言学研究组	1961	M. Pottier, M. Bourguin	研究英法、法西机译问题
其他研制单位：格勒诺布尔（Grenoble）大学；法国纺织研究所（巴黎）			
英国			
柏克培克学院	1955	A. D. Booth	在APEXC机上，词典包括250个词，每小时可译1 000个词，通过电视进行了表演；研究法英、德英机器翻译系统。

续表

单位名称	开始时间	负责人	机器翻译系统研制概况
剑桥语言研究组	1955	M. Masterman	研究 Thesaurus 词典类型的媒介语
其他研制单位：伦敦大学、国家物理实验室、埃塞克斯大学(科尔切斯特)、珍珠保险股份有限公司(伦敦)			
意大利			
米兰大学	1958	S. Ceccato	研究俄英翻译程序
国立物理实验室	1959		研究俄英机译和专用机方面的问题
欧洲原子能委员会科学情报提供中心	1959		研究丹、法、德、意四种语言的翻译问题：首先要实验德法和法意两种，准备进一步研究有关高质量机器的问题
比利时			
自动翻译和情报咨询研究小组(统计研究所)		P. Gillis 等	研究俄法机译问题
其他研制单位：布鲁塞尔大学、安特卫普大学			
捷克斯洛伐克			
查理大学捷语系机器翻译小组	1958	P. Sgall	在继电器数字计算机上，研究英捷机译系统，对捷语进行分析研究，探讨媒介语问题
罗马尼亚			

续表

单位名称	开始时间	负责人	机器翻译系统研制概况
罗马尼亚科学院语言研究所数理语言学委员会		Em. Ptrovici	研究英罗汉机译系统和建立罗马尼亚语机器语法
前南斯拉夫			
实验语言学研究所	1959		编制微型词典并研制实验性的翻译机
荷兰研制单位:荷兰外交部、N.V.菲利普辉光灯工厂(埃因霍温)			
保加利亚研制单位:保加利亚科学院			
德意志联邦共和国研制单位:联邦语言局、海德尔堡大学、康斯坦茨大学、科隆大学、鲁尔大学、萨尔大学、纺织文献和情报中心、Demag 公司、西门子公司			
德意志民主共和国研制单位:东德科学院、德累斯顿工艺大学			
欧洲共同体			
日本			
日本东京电工实验室	1958	H. Wada	在 Yamato 翻译专用机上,根据初一、二年级课本编制了规则系统,词典包括 2 000 个词。研究英日机译系统
九州大学	1958	Y. Oono 等	研究英/德/日语之间的互译问题,着重句法加工,创制了一种实验性的语言翻译机(KT-1),能译 200 个词,一句话用 30 秒钟
国家防御厅		M. I. Sakai	探索用一种逻辑代数来翻译多种语言
其他研制单位:京都大学			
马来西亚研制单位:槟榔屿科学大学			
黎巴嫩研制单位:国际语言中心			

1.1.3　第二代机译系统的活跃发展（1975～1989）

由于加拿大的双语政策和欧洲共同体内部多语文本都对 MT 产生了强烈要求，因此 ALPAC 报告之后，西欧和加拿大开始出现以追求可读性和忠实性为目标的第二代 MT 系统。这些系统以基于转换的方法为代表，普遍采用以句法分析为主、辅以语义的基于规则方法，采用有抽象的转换表示的分层次实现策略。这一代 MTS 综合了多种技术：知识与算法分离，模块化设计，多种句法分析策略以及语义分析等等，并且大多引入了人工智能技术。其中许多方法和技术相对比较成熟，直到今天仍被沿用。

机器翻译系统的复苏和发展可以从 1975～1976 年欧洲原子能机构（EURATOM）安装 Systran 系统和加拿大蒙特利尔大学 TAUM 组完成 METEO 系统算起，这期间比较著名的机器翻译系统包括：

SYSTRAN 系统

托玛（P. Toma）在 Georgetown-IBM 系统的基础上，进一步开发了大型的机器翻译系统 SYSTRAN，而且随后该系统不断地得到了改进和扩展。

SYSTRAN 系统从 1970 年至今，一直为美国空军进行俄英机器翻译（词典有 16.8 万个词和 13.6 万个词组，可进行俄英机器翻译，每小时可翻译 15 个万词；），其目的仅在于浏览情报，因而对译文的质量要求不高；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 30～35 万词；设在卢森堡的欧洲共同体总部于 1976 年引入 SYSTRAN 系统进行英法和法英机器翻译；联邦德国用它进行原子能文献的德英机器翻译；法国国家科研中心用它进行文献数据库情报的机器翻译；加拿大用它进行英法机器翻译；美国 Xerox 公司自 1978 年以来，用 SYSTRAN 系统把本公司的手册从英语译为法语、

德语、西班牙语、意大利语、葡萄牙语等五种语言；王安公司等跨国企业也用它进行英法、英德等机器翻译；日本的技术服务公司从1984年开始，用它进行日英机器翻译，主要翻译与计算机有关的技术手册。

在该系统中，语言信息与分析程序不作明确区分，而系统的效率较高，能满足实用的需要。据报道，日本使用SYSTRAN系统进行英日机器翻译，1小时CPU时间（主机时间）可译200万词，相当于8 000页A4号纸的英文文章。美国XEROX公司采用该系统后，翻译效率比人工翻译提高了5倍多。

SYSTRAN系统的翻译正确率也较高。美国空军使用的SYSTRAN俄英机器翻译系统，翻译科技文章的正确率达85%以上。日本使用的SYSTRAN英日、日英机器翻译系统，翻译正确率达85%～90%。

SYSTRAN系统要求在IBM的OS/VS1，或OS/VS2，或MVS等操作系统下运行，主存在700KB（700千字节）以上，可在IBM/360/370/3033/308X4341等计算机上运行，也可在HITACM序列机或SIEMENS等计算机上运行，因此，它便于广大用户使用。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。目前，SYSTRAN系统翻译的语言如下所示（括号内表示开发中的系统）：

英语→法语、德语、意大利语、葡萄牙语、西班牙语、俄语、日语、（阿拉伯语、波斯语）

法语、德语、俄语、日语、（汉语）→英语

德语→西班牙语、（法语）

法语→德语

由此可以看出，SYSTRAN是一个多语言机器翻译系统。由于它的各个用户的词典的使用方法和翻译方法各有不同，因此没有互换性。