

自然语言处理

的原理及其应用

ZIRAN YUYAN CHULI DE YUANLI JIQI YINGYONG

杨宪泽 等 编著

自然语言处理的原理及其应用

杨宪泽 谈文蓉
刘 莉 肖 明 编著

西南交通大学出版社
· 成 都 ·

内 容 提 要

本书主要阐述自然语言处理的原理,内容包括机器翻译、自然语言理解、文字识别、文字码、自然语言处理的常规技术。当然,围绕着上述专题,还将论述词法分析、语法语义分析等技术。

本书的主要内容取材于编著者们的科研积累,深入浅出,内容详实。

本书可供从事自然语言处理研究方向的科技人员参考,也可作为相关专业的研究生、本科生的选修教材。

图书在版编目(CIP)数据

自然语言处理的原理及其应用 / 杨宪泽等编著. —成都:
西南交通大学出版社, 2007.3
ISBN 978-7-81104-560-4

I. ①白… II. 杨… III. 自然语言处理 IV. TP391

中国版本图书馆 CIP 数据核字 (2007) 第 032700 号

自然语言处理的原理及其应用

杨宪泽 谈文蓉
刘 莉 肖 明 编著

*

责任编辑 孟苏成
封面设计 本格设计

西南交通大学出版社出版发行
(成都二环路北一段 111 号 邮政编码: 610031 发行部电话: 028-87600564)

<http://press.swjtu.edu.cn>
四川森林印务有限责任公司印刷

*

成品尺寸: 185 mm × 260 mm 印张: 14

字数: 350 千字

2007 年 3 月第 1 版 2007 年 3 月第 1 次印刷

ISBN 978-7-81104-560-4

定价: 26.00 元

图书如有印装问题 本社负责退换
版权所有 盗版必究 举报电话: 028-87600562

前 言

自然语言处理在 20 世纪 80 年代迅速发展,主要原因是当时日本人研究“第五代智能机”的呼声如火如荼,世界各国(包括中国)纷纷响应,并称之为“迎接新技术革命挑战”。这一阶段,我国在计算机智能技术方面投入了比较大的人力、物力加紧研究。自然语言处理是人工智能研究的一个分支,涉及非常多的技术。自然语言处理就是研究计算机如何处理人类语言的学科,它是用计算机模拟人类智能的一个重要方面。

人借助于自然语言交流思想,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,创造人类的物质文明和精神文明。自然语言是人类进行通信的自然媒介,它包括口语、书面语以及动作语(如哑语和旗语)等。就计算机应用而言,统计数据表明,数学计算仅占 10%,过程控制不到 5%,其余 85%左右都用于语言文字的信息处理。在信息化社会中,语言信息处理的技术水平和每年所处理的信息总量已成为衡量一个国家现代化水平的重要标志之一。

自然语言有许多独特的性质,使得自然语言处理成为人工智能的一大难题。即使是自然语言处理中的一些重要分支,如机器翻译、自然语言理解,也是 21 世纪要解决的科技难题。

本书取材于编著者们多年的科研积累,曾先后得到国家民族事务委员会重点科研项目的资助;四川省科技攻关项目的资助;西南民族大学重点科研项目的资助;西南民族大学精品课程建设项目的资助。这些资助使得编著者们获得了一定的成果,为本书的完成奠定了良好的基础。

本书不可能涉及全部自然语言处理技术,也不可能对所涉及的技术进行全面的论述,毕竟这些技术人们还在探讨、还在不断发展,属于研究前沿。我们只能在多年科研的基础上,根据积累的资料,从软件技术的角度竭尽全力地阐述自己的、同行的研究成果。这些内容包括:机器翻译;自然语言理解;文字识别;文字码;自然语言处理的常规技术。当然,围绕着上述专题,还将论述词法分析、语法语义分析等技术。

本书第 1、2 章由杨宪泽编著;第 3、4 章由杨宪泽、谈文蓉编著;第 5 章由杨宪泽、肖明编著;第 6 章由肖明编著;第 7 章由刘莉编著;第 8 章由杨宪泽编著;全书由杨宪泽统稿。

由于自然语言处理技术的新颖性和复杂性,加之编著者的水平所限,书中错误和不妥之处难免,希望读者谅解并批评指正,以利更好地改进。

编著者

2006 年 10 月于成都

目 录

第1章 总 论	1
1.1 自然语言处理	1
1.2 机器翻译	2
1.3 自然语言理解	11
1.4 文字识别	15
1.5 文字码处理	17
1.6 自然语言处理的常用方法	17
第2章 机器翻译方法	20
2.1 机器翻译的四种实现方法	20
2.2 基于实例的机器翻译改进方法	25
2.3 基于实例的机器翻译方法实践	27
2.4 与机器翻译相似的自然语言检索接口	34
2.5 机器翻译评价	35
2.6 机器翻译的优劣 ——对机器翻译系统的评价	38
第3章 自然语言理解	44
3.1 自然语言理解概述	44
3.2 自然语言理解过程的层次	49
3.3 最简单的单词理解	53
3.4 关键字匹配法	54
3.5 复合句的理解方法	55
3.6 语言的自动生成	56
3.7 一些典型自然语言理解系统的研究与应用	57
3.8 自然语言理解技术的主要应用	62
3.9 自然语言理解的新思路 HNC	65
第4章 单词及词组的处理与分析	68
4.1 机器词典概述	68
4.2 自动分词	69

4.3 交集型和多义组合型歧义切分	71
4.4 词语的排序、检索、词库	73
4.5 词语的分类与兼类问题	73
4.6 词语处理的其他问题	77
4.7 词义消歧	83
4.8 单词的时态处理、“是”动词选择处理及调序	103
第 5 章 句法（语法）与语义理论及分析	106
5.1 句法分析的作用	106
5.2 句法分析的理论基础	107
5.3 其他语法理论介绍	116
5.4 自动句法分析方法	129
5.5 语义分析	134
第 6 章 文字识别	145
6.1 文字识别概述	145
6.2 文字识别的基础理论	148
6.3 文字识别系统的组成	151
6.4 文字识别的特征提取	155
6.5 文字的分类识别	160
6.6 文字识别后处理	167
第 7 章 文字码	169
7.1 ASCII 码和扩展 ASCII 码	169
7.2 ISO/IEC 2022	173
7.3 汉字编码的种类与中文信息处理过程中汉字编码的变换	174
7.4 常见汉字编码标准	176
7.5 INTERNET 上常用字符编码	185
第 8 章 自然语言处理使用的常规技术	192
8.1 线性表技术	192
8.2 检索技术	197
8.3 排序技术	209
参考文献	216

第 1 章 总 论

自然语言处理是人工智能研究的一个分支，涉及非常多的技术。自然语言处理就是研究计算机如何处理人类语言的学科，它是用计算机模拟人类智能的一个重要方面。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，创造人类的物质文明和精神文明。

自然语言是人类进行通信的自然媒介，它包括口语、书面语以及动作语（如哑语和旗语）等。就计算机应用而言，统计数据表明，数学计算仅占 10%，过程控制不到 5%，其余 85% 左右都用于语言文字的信息处理。在信息化社会中，语言信息处理的技术水平和每年所处理的信息总量，已成为衡量一个国家现代化水平的重要标志之一。本章将概述自然语言处理的主要研究内容，阐述相关研究方向的现状与发展。

1.1 自然语言处理

自然语言指人类语言的本族语，不同的民族具有不同的语言，如汉、英、日语等。它是相对人造语言（指世界语或计算机的各种程序设计语言等）而言的，是人类在与大自然的搏斗中，交流信息而自然形成的重要交流工具。自然语言是思维的载体，随着时代的发展和变迁，也在不断地发生着变化。在人类历史上以语言文字形式记载和流传的知识占到知识总量的 80% 以上。

自然语言是用于传递信息的表示方法、约定和规则的集合，它由语句组成，每个语句又由单词组成。组成语句和语言时，应该遵循一定的语法与语义规则。

自然语言是音、形、义结合的词汇和语法体系，是人类实现思维活动的物质表现形式。词汇和语法体系是构成自然语言的两大要素，缺一不可。词是构成自然语言的最基本单位，语法规则用来支配和控制单词，以构成有意义的、可理解的语句，进而再由语句按照一定的逻辑构成篇章。

词汇分为词和熟语。熟语就是一些词的固定组合，如汉语中的成语。词又由词素构成，词素是构成词的最小的、有意义的单位。

在计算机软件技术中，人们设计了许多人工语言，如 BASIC, PASCAL, C, LISP 等程序设计语言。这些人工语言与自然语言一样，都遵循形式语言的规律和法则，存在某些共同的性质。但人工语言是受限语言，“单词”量很少很少，与自然语言相比有以下差距：

- (1) 自然语言充满歧义，而人工语言的歧义可以控制，甚至做到没有；
- (2) 自然语言语法复杂，而人工语言语法简单得多；

(3) 自然语言的语义表达千变万化, 很难统一, 而人工语言语义由人定义, 完全可以统一;

(4) 自然语言的结构和语义之间错综复杂, 一般不存在一一对应的关系, 而人工语言可以做成一一对应关系。

自然语言这些独特的性质, 使得自然语言处理成为人工智能的一大难题。像自然语言处理中的一些重要分支, 如机器翻译、自然语言理解, 均成为 21 世纪要解决的科技难题。

本书不可能涉及全部自然语言处理技术, 也不可能对涉及的技术进行非常全面的论述, 毕竟这些技术人们还在探讨、还在不断发展, 属于研究前沿。我们只能在多年科研的基础上, 根据积累的资料, 从软件技术的角度竭尽全力地阐述自己的、同行的研究成果。这些内容包括:

- (1) 机器翻译;
- (2) 自然语言理解;
- (3) 文字识别;
- (4) 文字码;
- (5) 自然语言处理的常规技术。

当然, 围绕着上述专题, 还将论述词法分析、语法语义分析等技术。

1.2 机器翻译

1.2.1 简 叙

机器翻译是自然语言处理的一部分, 也是人工智能研究的一个分支。机器翻译是用电脑代替人做翻译工作, 这里特指笔译, 口译问题本书不讨论。那么, 什么是翻译呢? 翻译是把一种自然语言记述的表达式(单语、短语、句子、单词标记符)变换为用另一种自然语言记述的表达式。同样是翻译, 却有不同的水平。最基本的是把原文的单词与句法结构映射到译文相对应的单词与句法结构。

人做翻译工作的必要条件是懂得两种语言。一般情况下, 有一种语言是其母语, 另一种是外语, 因而工作的难点在外语上, 需要记外语单词, 学外语语法。人能记住的单词与语法规则总是有限的, 翻译时免不了要查词典或参考书。还应当注意到, 人在翻译时会自觉或不自觉地运用自己头脑中所有的关于客观世界的各种各样的知识, 而人的知识又是可以不断积累与更新的。因此, 高水平的翻译, 人是可以达到的。当然, 不是人人都能达到最高境界, 正确认识人的知识在翻译过程中的作用正是把握机器翻译能做到什么程度以及做不到什么程度的关键。

显然, 计算机系统能够翻译的必要条件也是要能“懂得”两种语言, 而且, 这两种语言对于计算机系统来说都是“外语”。不难想象, 机器翻译是很难、很难的研究课题, 通观当今世界上现有的机器翻译系统, 大多数是建立在句法学的基础上的, 当然也注意吸收语义学的研究成果, 因此一般以句子为单位进行翻译。

机器翻译不可能达到人工翻译的水平, 因此, 它的用途主要有:

- (1) 进行粗译, 经过人工修改达到目标;
- (2) 有低水平的翻译总比没有翻译好, 可在译文中发现所需要的东西。

翻译过程要尽可能多地把源语言的意思、感觉和语言艺术传递给目标语言, 但是, 如果源语言中有的词汇目标语言中找不到, 翻译就只能近似地表达了。对于这种情况, 机器翻译不可能超过人。若把自然语言文本分成:

- (1) 诗歌与文学作品;
- (2) 法律文件与合同;
- (3) 科技文献;
- (4) 文章题目和一般句子。

四种情况, 机器翻译合适的处理对象是第(3)、(4)类语言材料。

机器翻译的特点:

- (1) 多学科合作。
- (2) 可计算性。即能用计算机程序实现。
- (3) 难解性。由于人类对语言的认知过程还不清楚, 所以计算机不可能达到人类对语言的熟知程度。因此机器翻译被称为要在 21 世纪解决的科技难题之一。主要困难是自然语言在各层次上的歧义性, 也称二义性或多义性。机器翻译的根本任务就是要在处理过程中逐步消除这些歧义。

(4) 实用性。尽管机器翻译的研究存在巨大的困难, 但人们却对它抱有很高希望。正如一些专家所说: 机器翻译研究者扮演的是科学的和商业的双重角色, 可随时在语言这个无底洞和它的使用者之间作出正确的妥协。

机器翻译有着重要的实践意义和理论价值:

(1) 实践方面的意义。国际合作与交流, 语言的差异是一个非常严重的障碍。各行各业的人们面对大量他们不熟悉语言的文档资料, 如果单纯依靠人工翻译, 这些日益增加的待翻译材料将是一种非常沉重的负担。而机器翻译可大幅度减轻这种负担。

(2) 学术研究方面的意义。机器翻译对于了解人类语言和思维的基本机制, 探讨人工智能技术有着重要的意义。

(3) 商业方面的意义。机器翻译广泛的应用前景有可观的经济效益和社会效益。

1.2.2 机器翻译常用的处理技术

自然语言的机器翻译目前一般以句子为翻译单位, 句子又分为句、短语、词三个层次。分析阶段一般分为词法分析、句法分析、语法分析、语义分析、语境分析、语用分析等, 其中以词法分析和句法分析为主。

词法分析有单词切分(包括歧义切分)和单词形态分析; 还有专有名词问题, 如政府机构名称、地理名称、公司名称、人名、日期、专业头衔等, 若不能正确识别, 就不会有正确的翻译。

句法分析的任务是确定句子中每个词的词性(或称词类), 确定词与词之间的关系以便构成短语, 确定短语之间的关系以便构成更大的短语或者组成句子。

对于复杂真实句子来说, 要一次性给出句子的完整分析是相当困难的, 而给出句子的某

些基本分析或部分分析结果却相对容易,这样既可以提高整个句子分析的正确性,又对于某些应用很有用处。所以,当前句法分析的一个热点是注重真实句子的部分分析。其内容包括名词短语的确定、句子中部分或全部短语边界的划定等。

语法分析可以确定通过句法分析的句子内部成分搭配是否合理,例如:

小孩玩游戏(句法分析正确,语法分析也正确,是主谓宾结构句子);

游戏玩小孩(句法分析正确,语法分析就不正确,虽是主谓宾结构句子,但主宾颠倒。人一眼就能看出来,计算机若没有语法分析就无能为力)。

语义分析就是解决句子中意义的不同或可不可能搭配,如下面两句:

(1) The fish was bought by the cook. 鱼是厨师买的。

(2) The fish was bought by the river. 鱼是从河边买的。

以上两个句子结构上完全相同,机器翻译必须依靠语义分析来区分含义,“cook”是有生命的人,而“river”是无生命的地点。

语境分析研究句子与句子间的关系,也就是上下文关系。例如两个句子中某个代词指代那一个名词。

语用分析研究源语言与目标语言不同的文化背景,有典故的词、句(成语)的翻译,例如:

(1) How are you? 不能译成“你怎么样?”,只能译成“你好!”

(2) “指鹿为马”不能译成“call a stag a horse”,只能译成“deliberately misrepresent”。

翻译在源语言和目标语言词汇级之间对应情况有:

(1) 一对一,直接根据词典译文替换;

(2) 一对多,需要根据上下文选择词典给出的译文之一。

居于机器翻译系统核心地位的语言处理程序,包括原文的分析程序与译文的生成程序。实际翻译过程是利用从原文分析出的信息生成与原文等价的译文。

机器翻译系统的基础是词典与语法。这里所说的词典同通常的书本形式的词典不同,它是存放在计算机系统内的供语言处理程序使用的电子词典,它记述了构成句子的词所具有的各种有关词法、语法、语义的性质,其中当然包括两种语言的对译关系。这里所说的语法也不是通常的用于教学或研究的描述性的语法,而是存储在计算机系统内的高度形式化的语法,它以规则形式描述了句子的结构,语言处理程序依靠这些规则分析原文并生成译文。这种电子词典与语法规则显然是相当庞大的数据库,因而机器翻译系统应配备词典与语法的管理系统。

为了实际运行机器翻译系统,在使用者与系统之间必须有一种友好的人机界面,使得原文与译文能够方便地输入与输出。

总之,机器翻译是一种高技术,它综合运用了计算机科学、语言学、认知心理学等多学科的技术成就。机器翻译的理论与技术还在不断发展之中,机器翻译系统在社会生活中的价值将会日益显现出来。

1.2.3 机器翻译的过程

可以用下式高度抽象地描述机器翻译的原理:

$$L1 \xrightarrow{\Phi} L2$$

这里 L_1 、 L_2 分别代表源语言、目标语言中的所有表达式的集合， Φ 是一种算法；它能确定 L_1 中任一个表达式与 L_2 中的某个表达式之间的等价对应关系。显然，如果这样的 Φ 存在并且能够把它构造出来，机器翻译的问题也就迎刃而解了。可惜的是，无论在理论上还是在实践上都还没能找到这样的 Φ ，实际上，由于 L_1 、 L_2 都是无限集，在只有有限资源（计算速度、存储容量等）的计算机系统中，这样的 Φ 是不存在的。因此，从工程实现的角度，只能对 L_1 、 L_2 加以限制，并给出 Φ 的近似描述。

Φ 的最直观的描述是枚举。在计算机系统中建立一部词典即数据库文件，它的每个记录只要包含英语与对应的汉语这样两个字段就够了。例如，

beautiful—美丽的	pretty—可爱的	flower—花	very—很
garden—花园	Good morning—早晨好	girl—姑娘	

如果要翻译的东西有限，机器的存储又足够，这种办法是再恰当不过的了，会翻译得又准确又迅速。不过这种办法对于变化多端的语句的翻译是没有实用价值的。因为即使有限多单词也能组合出无穷多个不同的语句。例如，由上述词典中的 7 个单词至少可以组合出 24 种不同的短语：pretty girl, beautiful girl, very beautiful, very pretty flower, …都是这样的短语。将这么多短语全部存入机器显然太笨拙了，实际上也是不必要的。词典中只要存入基本的 7 个单词就够了，另外把英语与汉语中都有的“名词可以受形容词修饰”、“形容词可以受副词修饰”这样的语法知识以下列规则的形式存入计算机，计算机程序参照这些规则，通过查英汉词典，得到相应的汉语词，就可以合成相应的汉语短语，即“可爱的姑娘”，“美丽的姑娘”，“很美丽的”，“很可爱的花”，……

自然语言的句子是由单词连接而成的，但是句子不仅仅是单词的线性序列，句子是有着层次结构的。例如，句子是由名词短语后接动词短语组成的，而名词短语又是由名词或形容词后接名词组成的，动词短语是由不及物动词或及物动词后接名词短语组成的。显然，同一种语言内部存在不同的句子结构。不同的语言之间，表达等价意义的句子的结构可能相同，也可能不同。在翻译的时候，通常的做法就是按照句子结构的本来面貌把源语言的一个句子分解成较小的构成要素，如短语。再将较小的结构进一步分解成更小的构成要素，直到基本的构成要素。这样便得到了关于原文句子的结构。所谓基本的构成要素就是那些登录在词典中的项目，单词显然属于这样的基本构成要素。这时通过查词典就可以得到在意义上等价的目标语言基本构成要素，于是参照原文的句子结构以至语义就可以选定目标语言的对应的句子结构，用目标语言的基本构成要素逐步生成较大的结构、更大的结构直至完整的句子。

以上所述的要素合成原理是当代机器翻译的最基本的原理。虽然说这个原理是不难理解的，但要将它付诸实施，就会碰到许多棘手的难题。

首先碰到的一个问题是如何选取基本的构成要素，单词是基本构成要素，但仅仅将单词作为基本构成要素是不可能得到较好的译文的。请看下例：

Good morning.
How do you do?
Here is a cup of tea for you.

如果词典仅登录单词，并假定系统中的语法规则是完备的，处理程序是正确的，以上各句也只会分别译成以下的汉语句子：

好的早晨。

你怎样做？

这里有一杯为了你的茶。

看了会莫名其妙吧？因此，通常也要把一些固定词组作为基本构成要素。但是这个口子一开，犹如洪水决堤，即使当代计算机的“海量存储”也难以包容。因此，对超越单词的基本构成要素必须加以限制，机器翻译还是要回到要素合成原理上。

即使将“Good morning”这样的固定词组吸收到词典中，也不能解决所有的翻译问题，因为翻译需要根据具体的语言环境选择恰当的目标语言的表达方式。例如，朋友之间问好，将“Good morning”译成“你早”是恰当的。但如果是晚辈问候长辈或下级问候上级，最好译成“您早”。一种取巧的办法是不分“你”和“您”，而含混地译成“早晨好”。但英美人在上午 11 点见面仍可以说“Good morning”，如果在此场景中仍译成“早晨好”，中国人又会觉得别扭了。

由此可见，好的翻译决不是仅靠要素合成原理就能实现的。不过，这么多复杂而又微妙的问题不是眼下立刻能够解决的。现在的机器翻译系统主要用于翻译科学论文等客观记述事实的、不带感情色彩的文章，因而可以认为要素合成原理是适用的。

1.2.4 句子的剖析与翻译

句子是表达一个完整意思的语言单位。现在的机器翻译系统基本上是一句对一句地进行翻译的，由于机器翻译依据的是要素合成原理，因此就有必要研究如何把原文的句子分解成基本构成要素（简称原文分析）以及如何根据基本构成要素结合成译文句子（简称译文生成）。为了开发一个好的机器翻译系统，原文分析与译文生成都有很多的问题需要深入研究，不过，分析似乎处于更重要的地位。

分析与生成都要依据某种语言理论去进行，计算语言学为适应机器翻译等自然语言处理技术的需要发展了一系列新的语言理论，如上下文无关短语结构语法、扩充转移网络、格语法、广义短语结构语法、词汇功能语法、功能合一语法等。我们不打算系统介绍这些语法理论及其在分析生成中的应用，而只选用这些语法以及传统语法中的一些基本概念与规则，用以阐述原文分析的方法步骤。请考察下面的英语句子及一些语法公式：

I like a red apple (I)

句子→代词+动词+名词短语 (II)

名词短语→冠词+名词 (III)

名词短语→冠词+形容词+名词 (IV)

与 (I) 类似的句子的结构都可以用规则 (II) 和 (IV) 加以解释。当然为了覆盖英语的各种各样的句型，仅有规则 (II) ~ (IV) 是远远不够的，但是这样的规则毕竟有限，数量可以控制。采用这种办法实现了利用有限的规则组合无穷多的英语句子。机器翻译系统不可能在任意的英语句子与汉语句子之间建立一一对应的关系，但是可以在反映英语句子结构的语法公式与反映汉语句子结构的语法公式之间建立对应关系。就 (II)，(IV) 而言则有：

(英) 名词+动词+名词短语 对应 (汉) 名词+动词+名词短语 (V)

(英) 冠词+形容词+名词 对应 (汉) 数量词+形容词+名词 (VI)

(V)、(VI)表示了英语与汉语某些句法结构的对应关系。除了英语的冠词要替换为汉语的数量词外,两者之间几乎没有什么区别,因而两种语言之间句法结构转换的必要性表现得不明显。但这里只涉及肯定陈述句及名词短语的一部分结构。如果考虑到疑问句、否定句、被动句等,情况就不一样了。

依据英汉句法结构对应关系(V)、(VI)就可以将英语句子(I)翻译成汉语,得到的汉语句子是:

我喜欢一个红苹果。(VII)

如果考虑到汉语名词前的数量词“一个”在不强调数量的情况下是可以省略的,而在(VII)中,“红苹果”只是“喜欢”的一般对象,“一个”应该省去,于是有:

我喜欢红苹果。

从这个简单的例子可以看出剖析一个句子的结构在翻译中所起的关键作用,同时也可以看到,为了生成一个合格的译文句子,还需要给计算机配备足够多的语言知识与客观世界知识。

众所周知,英语有以下5个基本句型:

$S+V$ $S+V+C$ $S+V+O$ $S+V+O+C$ $S+V+O_1+O_2$

这里S、V、O、C分别代表主语、谓语动词、宾语、补语,O₁、O₂则分别表示间接宾语与直接宾语。只要对英语的这5个基本句型分别给出对应的汉语句型,就成为英语基本句型到汉语的转换规则。如果能够识别一个英语句子属于哪一个基本句型(显然,计算机关于其操作对象即符号或数据的异同的逻辑判断能力在这里大有用武之地,这也就是通常所说的模式匹配),根据英语汉语之间句法结构转换规则,就能把给定的英语句子翻译成汉语,如可把下句:

She told me a very interesting story.

识别为 $S+V+O_1+O_2$ 的句型(其中O₁为me, O₂为a very interesting story),则根据转换规则,不难把它翻译为

她给我讲了一个很有趣的故事。

当然,实际的英语句子要复杂得多,并不都这么容易处理。识别一个英语句子的谓语动词是关键,围绕谓语动词的是作为S、O、C的各个名词短语。而英语的名词短语又有复杂的结构:

(英语)名词短语→名词

- 人称代词
- 冠词+名词
- 冠词+形容词+名词
- 副词+形容词+名词
- 名词+介词+名词
- 名词+关系代词+句子
- ……

英语的句子中包含名词短语,而从上面的最后一个公式可以看出,名词短语中又可以包含句子,因此剖析英语句法结构的程序必须要有处理这种复杂的递归结构的能力。

通过以上介绍,可以归纳机器翻译的通常方法要点如下:首先通过剖析确立原语言句子

的语法（或语义）结构，将原语言的基本构成要素译成目标语言的基本构成要素，根据两种语言间的句子结构转换规则，可以由原语言的句子结构找到目标语言的适当的句子结构，按照这个结构就可以将目标语言的基本构成要素组合成句子。

1.2.5 机器翻译的常规步骤

下面给出包括输入原文及输出译文在内的常规机器翻译的全过程，简要地解释一下各个步骤。虽然某些步骤是纯技术性的，但它们对一个实用的机器翻译系统也是不可或缺的。

（1）原文输入。输入原文的方法多种多样，最常规的是键盘输入。不过大量输入原文也并不是一件轻而易举的事。由中国操作员录入某些语种更会感到困难。因此，要考虑文字识别等先进的输入手段。机器翻译系统与科技文献检索系统相结合，也可以收到很好的效益。

（2）译前编辑。为了加快处理速度，可将原文中一些无需翻译的图表、公式、符号标记出来。为了降低翻译的难度，可以在原文的句子中加进一些符号以排除歧义或者补上一些省略了的成分，这项工作通常由只懂原文的人根据系统的要求去做。

（3）词法分析。英语有词尾变化，而词典中不可能收录各个单词的所有形态，因此系统中应另存一套形态变化规则。根据这套规则，可将变化了的形态还原成基本形，再去查词。如果原文是汉语，还需要将构成句子的字符序列切分成单词。这也是一件颇困难的工作。

（4）句法分析。这一步已详细介绍过了。但是句法剖析并不能解决机器翻译中的所有问题。特别是碰到有歧义结构的句子时，仅仅靠句法分析便决定不了如何取舍。例如

The boy saw a girl with a telescope. (Ⅶ)

在（Ⅶ）中，介词短语 *with a telescope* 既可以作为谓语动词 *saw* 的状语，也可以作为名词 *girl* 的定语。当解释为状语时，会译成：

男孩用望远镜看见了一个女孩。

当解释为定语时，则译为：

男孩看见了一个带望远镜的女孩。

因此，仅有句法分析是解决不了两种语言间的转换问题的，还需要作深入的语法分析、语义分析和语境分析。

（5）语法分析。前面已经介绍过，通过语法分析可以确定通过句法分析的句子内部成分搭配是否合理。

（6）语义分析。虽然句法分析程序确定不了（Ⅶ）中的介词短语在句子结构中的地位，但并不等于说这类问题都是不可解决的。请看下面两个例子：

I bought a table with three dollars. (Ⅸ)

I bought a table with three legs. (Ⅹ)

这两个句子的结构与（Ⅶ）完全一样。但从语义的角度来看，在（Ⅸ）中，*dollars* 是货币，可以作为 *bought* 的工具格，*dollars* 与 *table* 并没有直接关系，因此介词短语 *with three dollars* 只能是状语。而在（Ⅹ）中，*legs* 是 *table* 的组成部分，不能用 *legs* 购物，因此 *with three Legs* 只能是修饰 *table* 的。如果在英语词典中，除了指明 *dollars*，*legs* 属于名词这个语法范畴外，还注记其语义信息，即 *dollars* 是“货币”，*legs* 是“物体的组成部分”，分析

程序就可以参照这些语义信息在两种可能的句法结构中选择一个合理的保留下来。这项工作就叫做语义分析。

当代的机器翻译系统或多或少都加进了语义分析的功能。不过,多数系统还只是把语义分析作为一种辅助的手段。即使增加了语义分析,通常分析的跨度基本上也还是局限在一个句子的范围内。因此,对于(Ⅳ)那样的句子还是无法排除其歧义的。也许词典中指出了 telescope 是工具,可以作为 saw 这个动词的工具格,据此可将 with a telescope 确定为 saw 的状语,但并不能保证这种判断一定符合原文的意思。因此,还需要进行语境分析。

(7) 语境分析。分析的范围超出一个句子,在上下文的环境中确定句子的语义,这就是语境分析。语境分析除了可以解决(Ⅳ)所表现的那一类歧义性外,还可以解决句子中的省略、指代等自然语言中习以为常而计算机处理起来却十分困难的问题。语境分析程序必须包含有力的推理功能。

(8) 语用分析。研究源语言与目标语言不同的文化背景,有典故的词、句(成语)的翻译。

以上分别介绍了语言的词法分析、句法结构分析、语法分析、语义分析、语境分析和语用分析,实际的分析过程大致上也是这样进行的。但不等于说,每一个实用的系统都必须包含直到语境分析的每一个步骤,也不是说这些步骤之间一定是分明的。实际的系统可能融合词法分析、句法分析、语义分析于一体。

设计机器翻译系统时,应该牢记工作的目标是要得到与原文等价的译文。因此,分析到哪一步为宜要根据能否由此得到恰当的译文来决定,不要以为分析得越深就越好。

(9) 内部表示的转换。由句法结构分析、语义分析及语境分析皆可得到原文的某种形式的内部表示。上下文无关语法的分析树是一种内部表示,格语法的格框架也是一种内部表示。机器翻译的决定性步骤就是将这种相对独立于原文表层表达方式的内部结构转换为译文的相对应的内部结构,这种转换根据一定的规则和算法进行。这一步完成了,剩下的工作就是根据转换了的内部结构生成译文的句子。

(10) 译词选择。原文总是要分解成基本构成要素的,这些基本构成要素的相对应的译词可以在双语词典中找到。译词应当置于转换后的结构的某个位置上,如树结构的叶结点上或格框架的格槽中。

由于一个英语词可能对应若干个不同的汉语词,例如 table 作为名词就有“桌子”、“平板”、“表格”3种不同的意思,因此需要根据专业领域、句法结构、语义以及译文的前后搭配关系加以选择。为了适应这种需要,词典很可能不仅仅是数据库,而且可能包括可执行的过程。

(11) 译文句子的生成。根据转换后的内部结构及选译的译词就可以按一定的算法将单词排列成线性的序列,译文的句子也就形成了。

(12) 词形变化。如果是由英语译成汉语,由于汉语单词的形态变化不丰富,这一步也许没有什么事情可做,只要按通常的书写印刷习惯将词与词之间的空格挤掉就行了。如果是由汉语译成英语,虽然无需挤掉单词间的空格,却应改变某些单词的词形,使主谓一致,使时态表示正确。例如:

I be a student.

在这一步就应当改为

I am a student.

(13) 译后编辑。可以在终端屏幕上将译文显示出来供人校对。熟悉机器翻译规律的校对者不看原文也可以作一些编辑修改工作。如有一句译文为

我看见了在桥上游泳的人。

编辑者不难判断,“在桥上”这个短语应是修饰动词“看见”的。译后编辑可在相当程度上提高译文的可读性。

为了提高译后编辑的效率,机器翻译系统通常会提供良好的界面,如实现两种语言对照的多窗口,灵活的编辑功能等。

(14) 译文输出。译文可用打印机输出,也可同电子印刷系统联结起来,还可以通过网络将译文送给远程终端的用户。

1.2.6 机器翻译的分类

前面介绍了机器翻译的基本原理及典型过程,实际的机器翻译系统却呈现各种各样的形态,现从不同角度对机器翻译系统进行分类,以期把握机器翻译的概貌。

(1) 从涉及语种的角度分类。

a. 一对一的系统,又可细分为单向的与双向的。

b. 多语种系统。虽然说多对一与一对多的机器翻译系统可划为多语种系统,但本质上不过是多个一对一系统的简单组合。

(2) 按自动化的程度分类。

a. 自动翻译的系统。这类系统通常采用批处理作业方式,在翻译过程中不需要人的干预。但不同的自动翻译系统又可按有无译前编辑或译后编辑分成不同的小类。

b. 翻译支援系统。这类系统按人机互助方式完成翻译作业,通常采用会话方式工作。又可细分为人助机译和机助人译两小类。如果构思精巧,这类系统是很有价值的,可以充分发挥人与机器各自的优点,从而获得最大的效益。

(3) 按处理方式分类。

a. 直接翻译。直接翻译指从原文句子的表层出发,将单词或者与单词同样看待的成分(固定的词组、短语甚至句子)直接置换成目标语言的对应成分,必要时也可能调整一些词序,这样生成译文的句子。以词汇为主的第一代机器翻译系统主要采用这种直接翻译方式。计算机型的旅游翻译器基本上也采用这种方式。

b. 间接翻译。与直接翻译相对照,间接翻译则是要分析原文的句子,得到原文句子的句法结构、语义表达等中间形式,然后将这种中间形式进行转换,得到目标语言的中间形式,再由此生成译文的句子。以句法为主的第二代机器翻译系统与以语义为主的第三代系统采用的主要是间接翻译方式。按照分析深度及中间形式与源语言、目标语言关系的密切程度又可将间接翻译方式分成不同的小类。不过,这种划分的界限不是十分清晰的。理想的中间形式是完全独立于源语言与目标语言的,如果这样的中间形式找到了,也可以将这种中间形式叫做媒介语或中间语言。

目前,多数系统的中间形式或多或少地依赖于源语言与目标语言,这样的方式就叫做转

换方式。这是当代机器翻译系统的主流。在主要采用转换方式的系统中,当然也允许用直接方式翻译某些语句。

1.2.7 机器翻译的进展

纵观机器翻译的历程,已经走过了理论研究的几十年,并走向了实用化。美、英、法、日、俄、德等国都相应进行了母语与“外语”机器翻译,这一领域的国际化研究与合作非常活跃,取得了不少的成绩。然而,机器翻译确实是一项很难的课题,包含了语法、语义和语用等基础问题的研究,需要计算机科学、语言学、认知心理学等多种学科的技术成就,因此,至今各种语言的机器翻译都尚未达到满意的效果。

必须肯定的是,机器翻译的理论与技术近十几年发展非常迅速,机器翻译系统在社会生活中的价值日益显现出来。若其效果满意,一方面,它不仅可被各国政府机构所用,而且可促进世界范围内的科学、文化交流;另一方面,由于创造和使用自然语言是人类高度智能的表现,因此,机器翻译的理论部分有助于揭开人类智能的奥秘,深化我们对语言能力和思维本质的认识。可以说:机器翻译的研究在理论方面和应用方面都有重大意义。

我国是继美国、俄罗斯、英国之后,世界上第四个开展机器翻译研究工作的国家。与国外机器翻译的发展情况相比较,我国机器翻译除了有草创期、复苏期和繁荣期之外,由于受“文化大革命”的影响,还有一个非常特别的停滞时期。而且,由于我国机器翻译在理论上和方法上以及设备上的底子都很薄,我国机器翻译的每一个时期又都比国外机器翻译的同样时期稍微落后。这是我国机器翻译发展的不足之处,也有待从事这一领域研究的科技工作者去赶超世界先进水平。

1956—1966 年是我国机器翻译草创期;1966—1975 年是停滞期;1975—1987 年是复苏期;1987 年至今是繁荣期,这个时期是以“科译1号”机器翻译系统的问世为标志的。继“科译1号”之后,一系列实用化商品化的机器翻译系统如雨后春笋般地推向市场,我国的机器翻译迈向了实用化和商品化的阶段。

1.3 自然语言理解

1.3.1 什么叫自然语言理解

自然语言理解就是研究如何让计算机理解人类自然语言的一个研究领域。从宏观上看,自然语言理解就是指计算机能够执行人类所期望的某些语言功能。

自然语言理解可以分为口语理解与文字理解两方面的工作。所谓口语理解就是让计算机“听懂”人们所说的话;所谓文字理解就是让计算机“看懂”输入到计算机中的文字资料,并能用文字作出响应。

自然语言理解作为语言信息处理技术的一个高层次重要方向,一直是人工智能界关注的核心课题之一。显然,如果计算机能够理解自然语言,那将是计算技术的一项重大突破。什么叫“理解”?现在普遍采用著名的图灵(Turing)试验判断计算机是否“理解”了某种自然语言,具体的判别标准至少有如下四条: