机器翻译技术丛书之一

机 器 翻 译 原 理

赵铁军等 编著

哈尔滨工业大学出版社
2000·哈尔滨

内 容 提 要

本书是国内第一本全面、系统论述机器翻译实现原理和技术的著作。

本书以作者的实际研究与开发经验为基础，全面介绍了当前国内外机器翻译研究的最新进展，取材丰富，内容深入。每章后面均配有思考题，便于教学。书后列出全部参考文献，便于读者查找相关资料作进一步研究。本书可作为高等院校高年级本科生和研究生的教材，也可作为机器翻译与计算语言学研究者的参考书。

机器翻译技术丛书之一

机 器 翻 译 原 理

Jiqi Fanyi Yuanli

赵铁军等 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 13.625 字数 360 千字

2000 年 6 月第 1 版 2000 年 6 月第 1 次印刷

印数 1～3 000

ISBN 7-5603-1468-6/TP·139 定价 25.00 元

前　言

●缘起

两年多以前，河北师范大学外语学院（原河北师范学院英语系）建立了我国第一个机器翻译专业，并开始招生。创业之初，一切都要从头做起。困难之一便是没有机器翻译方面的教材。于是，当时的系主任一行专程来到哈尔滨工业大学机器翻译研究室，请我们为其编写有关教材。盛情难却，我们只能拿出初生牛犊的精神，勉力为之。我和机器翻译研究室的博士生们在 1997 年暑假结束之前匆忙赶就了两本内部讲义《机器翻译原理》和《机器翻译系统》，以满足新学期教学的需要。书稿写就之时，就颇感惶恐，现在再来翻看，更觉汗颜。

去年五月，河北师范大学外语学院和哈尔滨工业大学出版社协商后决定正式出版机器翻译技术丛书。值此机会，我将原书大纲重新审理了一番，开始了重新编写。经过近 3 个月紧张、不懈的努力，终于使《机器翻译原理》一书以新面貌呈现在读者面前。实际上，3 个月的时间还是太短了，因此本书中还存在着种种不能令人满意的地方。今后如有机会，仍需加以修正。

●内容

众所周知，机器翻译是一项难度很大的研究课题，许多方法和技术尚处在探索之中。因此，本书在力图阐述机器翻译基本原理和技术的同时，必然要涉及不少当前计算语言学领域的最新研究成果。本书力图兼顾这两方面，既能为机器翻译的初学者提供入门的引导，又能成为机器翻译和计算语言学研究者的参考书。所以本书就内容上来说是一本专著，从体例上来说则努力向教材方面靠拢。

机器翻译所涉及的技术与计算语言学或自然语言处理的诸多

研究分支都有联系。如何将有关内容串接在一起，以利于教学？最直观的方式就是按照机器翻译系统实现的处理流程去讲解各个步骤和模块所包含的方法与技术。这很像分立元件时代的收音机或电视机工作原理的介绍。因此，本书就按照一般机器翻译系统应该具备的各个处理步骤的顺序讲述机器翻译的实现原理，即按照词法分析、词性标注、句法分析及语法理论、语义分析及词义消歧、转换与生成的次序组成本书的主要脉络。由于近年来自然语言处理已经成为计算机应用研究的热点，新方法、新技术层出不穷，许多内容都可以应用到机器翻译研究与系统实现当中来，所以本书不可能一一收集、面面俱到，挂一漏万在所难免。但是通过本书，总体上还是可以掌握机器翻译基本原理的。

本书的读者应该具有一定的计算机科学方面的基础知识，如离散数学、数据结构、形式语言、人工智能等，以便更好地理解本书的内容。当然，读者如果对于某些部分感到困难的话，完全可以跳过，而着重基本内容的理解。

●编著者

本书的编著得到了哈尔滨工业大学机器翻译研究室的其他教师和研究生的分工协助。参与本书写作的人员有：

杨沐昀(第 7 章、第 11 章)

刘　芳(第 5 章大部、2.6 节)

钱丽萍(第 3 章前 3 节)

孟　遥(第 6 章前 3 节、2.7 节)

陈丹琪(第 4 章大部、3.4 节、5.3 节)

张　晶(第 9 章)

于　浩(2.8 节、3.5 节、5.8 节)

姚建民(6.4 节)

根据他们写就的初稿，我作了仔细的校改和润色，以使全书的内容前后连贯、风格相对统一。

●致谢

尽管本书重新进行了构思和编写，但原内部讲义作为本书写

作的基础仍然功不可没。值本书正式出版之际，我向当年曾参与编写讲义的张民博士、刘小虎博士、王海峰博士、董涛硕士表示感谢，向曾提供帮助的荀恩东博士、蔡萌硕士表示感谢。

一本书的背后往往都有许多人的支持，本书也是如此。在此，感谢哈尔滨工业大学李生教授在本书写作过程中的大力支持；感谢前河北师院英语系系主任袁宪军教授、现任有关领导张艳军老师、冯梅老师等的积极联络和鼓励；感谢在本研究室工作的研究生和进修教师方高林、刘继武、孟俊茂、康振国等在各方面的帮助。

感谢本研究室博士生吕雅娟为本书编制了英汉术语对照表。

本书的写作得益于近年来本研究室在机器翻译技术和系统实现方面的广泛而深入的研究工作，这些研究都是在有关部门的资助下开展的。在此，我代表研究室全体师生向多年来一直资助我们进行机器翻译研究的国家 863 计划智能计算机主题专家组、国家自然科学基金委、原航天工业总公司科技部表示我们最真诚的谢意！本书的出版就作为我们一点小小的回报吧。

●欢迎批评指正

由于本人水平和时间的限制，本书难免存在不少疏漏和不足之处。恳请各位读者不吝指教，指出书中不足，以帮助作者不断提高。欢迎与我们联系，通信地址是：

150001 哈尔滨工业大学计算机系 321 信箱

e - mail: tjzhao@mtlab.hit.edu.cn

赵铁军

2000 年 3 月于哈尔滨工业大学

第 1 章　机器翻译概述

翻译是从一种语言到另外一种或多种语言的变换过程。世界上不同国家或民族的人们使用不同的语言，在大多数情况下必须通过翻译才能进行交流。如何克服由语言不同而带来的不便？能不能找到一种自动翻译的方式来满足人们的需要？在计算机技术飞速发展的今天，大家很自然地会想到使用计算机来帮助我们，这便是机器翻译要研究的内容了。在这一章里，我们将介绍什么是机器翻译，计算机是怎么进行翻译的，如何看待计算机的翻译结果，以及机器翻译是如何发展的。

1.1　机器翻译的任务和意义

1.1.1　什么是机器翻译

机器翻译(Machine Translation)的英语缩写为 MT，在本书中常常使用这个缩写来代表机器翻译。首先我们要说明什么是机器翻译？简单地说，机器翻译就是应用计算机实现从一种自然语言文本到另一种自然语言文本的翻译[Hutchins, 1986]。显然，这里的机器专指计算机。由于某种习惯的原因，人们继续使用机器翻译这一名称，而没有改为计算机翻译之类的名字。MT 的处理对象是自然语言(Natural Language，简称 NL)，以区别于任何人工语言如计算机程序设计语言。同时注意，这里专指对文本的翻译，未涉及话语的翻译。因为话语翻译(或者称为口语翻译)又要涉及语音识别与合成，而这些是相对独立的研究领域。

MT 要实现对自然语言的翻译，必然涉及对自然语言的处理技术。因此，MT 是自然语言处理(Natural Language Processing，简称 NLP)研究领域的一个分支。同时，MT 和计算语言学(Computational

Linguistics,简称 CL)、自然语言理解(Natural Language Understanding,简称 NLU)都有密不可分的联系。下面简要地解释一下这几个术语所包含的内容及其相互联系,从而更好地理解 MT 所担负的任务以及所需要的方法和技术。

计算语言学是对理解和生成自然语言的计算机系统的研究[Grishman,1986]。这里之所以强调计算机系统,就是因为只有当一种语言学理论或方法能够被计算机所处理时,才能称得上是计算语言学。计算语言学和自然语言处理研究的内容应该是一致的,二者的着重点有所不同。从理论和方法的角度称为计算语言学,从技术和应用的角度称为自然语言处理。总之,这是一个相当广泛的研究领域,一般来说,凡是和自然语言相关的计算机理论、方法、技术、系统,都可以纳入自然语言处理的研究范围。从某种意义上说,计算语言学的目标是试图捕捉人类的语言能力[Grishman,1986]。相比之下,自然语言理解研究的范围就小一些,它研究的是自然语言的词汇已被识别以后所要进行的处理[Allen,1995],它的研究从词汇开始。自然语言理解是计算语言学的核心内容,也是 MT 的基础,因为 MT 就是从处理词汇开始的。由于自然语言理解是人工智能(Artificial Intelligence,简称 AI)的一个研究分支,所以 MT 也是 AI 的应用。MT 作为计算语言学的应用和分支,既广泛应用了计算语言学的方法和技术,又有自己的专门技术,如涉及双语的计算技术、中间语言表示等。

1.1.2 机器翻译的用途和处理对象

机器可以作翻译,但是 MT 能像人一样进行各种各样的翻译吗?答案是否定的,至少在未来相当长的一段时间内是这样的。首先,人类的翻译能力是经过长期学习和训练而培养出来的。要想翻译好,必须请专门的翻译人员才行。其次,计算机的智能远远无法和人相比。所以,我们怎么能期望机器做一般人都做不到的事情呢?

那么,应该对机器翻译系统和工具提什么样的翻译要求?按照英国学者 Hutchins 的分析,MT 的应用可以分为以下 4 类

[Hutchins,1999]:

·用于发行(dissemination)　期待 MT 的翻译结果达到人工翻译的水平,可直接分发给阅读者。这是一种最传统的要求,但是 MT 系统的输出必然总是要经过人工修改才能达到其目标。或者把待翻译的文本及其语言格式限制在一个非常狭窄的范围内,以便于 MT 系统处理。

·用于浏览(assimilation)　虽然 MT 的译文不能达到直接发行的质量,但是有一些低水平的翻译总比没有翻译要好。有些用户在 MT 输出的未经编辑的译文里发现了他们所需要的东西,因此第二种应用在某种意义上是第一种应用的副产品。

·用于交流(interchange)　随着国际交流的日益广泛,特别是 Internet 的普遍开通,产生即时翻译(immediate translation)的大量需求。MT 应在这种需求当中找到其自然的角色。更进一步地,与语音识别和语音合成结合起来,构成语音翻译系统如电话翻译系统等,将给人们带来极大的便利。尽管语音翻译实现起来还非常困难,但是已经得到了人们的注意,并开始了这方面的探索。

·用于信息获取(information access)　MT 可作为各类信息获取系统的一部分,构成多语言环境下信息检索、信息抽取、文摘、数据库查询等应用中不可缺少的部件。在欧洲,这一研究领域已经成为欧盟一些项目的关注点。

而按照法国学者 Boitet 的分类,MT 可以分为如下 4 种类型[Boitet,1995]:

·用于浏览者(for the watcher),称之为 MT - W,目的是提供粗糙的译文,便于获取某些信息。

·用于修订者(for the revisor),称之为 MT - R,目的是得到类似于手工译文初稿的翻译。

·用于翻译者(for the translator),称之为 MT - T,目的是帮助人类翻译者进行翻译,如提供在线词典、同义词词典等。

·用于作者(for the author),称之为 MT - A,目的是通过人机共同工作,输出比较满意的译文。

上述两种分类方法有相似之处,但都说明不能笼统地谈论 MT 应用,而应该指明 MT 的具体用途。这对于认识和评价 MT 的能力是十分必要的。

MT 适合翻译什么样的文本? 下面再分析一下 MT 的处理对象。按照美国语言学家 Nida 的观点[Nagao,1989],翻译可以分为三个层次:第一个层次是源语言(Source Language,简称 SL)与目标语言(Target Language,简称 TL)之间的词汇和语法结构的映射;第二个层次是根据交际(communication)原则来生成目标语言;第三个层次是基于特定文化背景的翻译。这三个翻译层次大致对应于语言学研究的三个层次,即句法(syntax)、语义(semantics)和语用(pragmatics)。

有时只使用句法知识对于翻译来说是不够的。例如有英语句子[桂诗春,1988]:

Mary and John saw the mountains while they were flying to California.

句子中的"they"代表"Mary and John"还是"mountains"? 只有语法规则可能做不出正确判断,因为有的语法书说"代词代替最接近它的先行词"。而根据常识也就是根据一种交际原则,我们只理解为前者。所以翻译结果是"当玛丽和约翰飞往加利福尼亚时,他们看见了山"。要达到这种理解,需要进行语义分析。因为根据常识,这样的世界知识(world knowledge),"mountain"和"fly"不能搭配,而"人"可以和"fly"搭配。这种搭配关系可以通过词汇所属的语义类来判断。

至于涉及到语用方面的翻译,由于要研究句子本身意义之外的意义(语用学的研究内容)[何自然,1988],就必须依靠源语言和目标语言所处的不同文化背景才能得到正确的译文。例如,有下列两方面的翻译问题:一类是社交方面的文化差异造成的翻译,一类是含有典故的词、句(成语)的翻译。显然,这两类翻译问题都不能按照一般的句法分析和转换的方式去处理,否则就会产生笑话或者让人感到莫名其妙。如"How are you?"不能译为"你怎么样?"而是"你好!"同样,成语也不能按照其字面意思去翻译。汉语成语

的翻译就是最明显的例子，如“指鹿为马”不能译为“call a stag a horse”，而是译为“deliberately misrepresent”。这些成语只能整个作为一个词来进行翻译，一般不能再拆开分析了。

翻译过程要尽可能多地把源语言的意思、感觉和语言艺术(artistic value)传递给目标语言[Nagao,1989]，但是如果源语言中有的词汇和概念在目标语言中找不到对应物的话，翻译也就只能近似地传达了。对于这样的情况，MT 系统决不可能超过人。从目前计算机技术的发展来看，它还不能像人一样来理解自然语言，即使限制一个极其狭隘的范围也是不能完全理解的，同样难以应用世界知识。人工翻译是以他或她的全部知识积累作为翻译支持的，而 MT 只会利用人教给它的有限知识，也许今后机器学习的发展会改变目前的状况。所以，这里再次强调指出：MT 的翻译结果决不可能和人工翻译相比。MT 所适合的翻译材料只能是自然语言中表述客观事实的部分。因此，如果把自然语言文本分为下述 4 种不同类型，即

·诗歌与文学作品

·法律文件与合同

·科技文献

·文章题目和一般句子

则对于第一类艺术作品，MT 是不能问津的；第二类由于对翻译质量要求非常严格，MT 只能起到辅助手段的作用。这样，MT 合适的处理对象是第三、四类语言材料[Nagao,1989]。实际上，MT 仅仅把第三类语言材料中的事实从一种语言传达到另一种语言，对事实的理解还依赖于本领域的专家。

上述分类只是一种分类方法，我们完全可以构造自己的分类，然后分别考察其人工翻译的难易程度，从而认识 MT 处理对象的范围。MT 的实践表明，许多用户希望 MT 系统能够为自己翻译各种句子，而不管这些句子到底是不是 MT 合适的处理对象。因此，说明 MT 的有限能力是 MT 研究者的一个责任。

1.1.3 机器翻译研究的特点和意义

总体上说,MT 研究具有以下特点。

·学科交叉性(Cross - disciplinary) MT 涉及计算机科学与语言学的交叉。显然,如果不研究语言学规律、汇集语言使用的知识,MT 系统只能是无源之水;反之,如果只有语言学研究成果而不能用计算机加以实现,MT 就是一句空话。因此,需要计算机工程师与语言学者密切合作,才能推动 MT 不断发展。

·可计算性(Computable) 既然是 MT 而不是人工翻译,所以,有关翻译的方法和知识都必须具有可计算的性质,即能够用计算机程序实现,才能应用于 MT。

·难解性(Intractable) 因为 MT 的处理对象是自然语言,而人类对于语言认知的过程仍然不清楚,所以计算机不可能达到人类对语言的驾驭程度,因而要实现全自动、高质量的 MT 至少在目前是极其困难的。因此 MT 被称为是要在 21 世纪解决的科技难题之一。主要困难就是自然语言在各个层次上的歧义性(ambiguity),也称为二义性或多义性。MT 的根本任务是要在处理过程中逐步消除这些歧义,从而正确地理解并翻译一个句子或篇章。

·实用性(Practical) 此条似乎与上一条有矛盾,但现实往往就是如此奇怪。尽管 MT 研究存在着极大困难,还是面临着人们对它抱有过高期望的巨大压力。各种各样的 MT 技术研究的最终目标就是要建造一个实用的 MT 系统。倘若 MT 研究不是朝着部分替代人类翻译的目标前进,那么它也就失去了存在的价值。可以说这是 MT 最重要的特点。诚如国外专家所说,MT 研究者不得不扮演科学的和商业的双重角色,以便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[Bourbeau,1993]。

机器翻译的研究与实用系统的最终实现有着重要的实践意义和理论价值,可归纳为如下几方面:

·实践上的意义 在当今信息社会,国际间的交流与合作日益广泛和深入,"地球村"的概念正在为越来越多的人所接受。在这种过程中,语言的差异是一个非常严重的障碍。各行各业的人们

每天都要面对大量用他们所不熟悉的语言写成的文档资料，要与和他们持不同语言的人进行交流。如果单纯依靠人工翻译，这些日益增加的待翻译材料无疑是一种非常沉重的负担，而机译系统的成功运行必将大幅度地减轻这种压力。

·学术研究上的意义　MT 研究对于了解人类语言和思维的基本机制，探索计算机及人工智能的潜力和极限都有着重要意义。

·商业上的意义　MT 产品的广阔应用前景和很高的技术含量决定了它必将为 MT 系统的开发与销售带来可观的经济利益以及社会效益。

1.2　机器翻译的实现过程

1.2.1　源语言到目标语言的转换

MT 的实现过程是 MT 系统(MT System，简称 MTS)完成由源语言向目标语言的转换，MTS 的输入是源语言句子，输出是目标语言句子。它的工作过程和一个人进行翻译的过程有许多相似之处。下面举一个英译汉的例子。

The book which was given to him was a workbook of English grammar.

我们翻译这个句子时，首先判断“which was given to him”是“book”的定语，然后判断整个“The book which was given to him”是主语，译成汉语时把定语位置提前。对宾语部分“a workbook of English grammar”也作类似处理。“was”是谓语，整个句子的“主语 + 谓语 + 宾语”结构在翻译时不改变。就具体词的翻译来说，我们要把上面句子中“which”去掉不译，而“a”译成汉语时，我们要添加一个量词“本”，译为“一本”才合乎汉语表达的要求。

实际上在人的翻译过程中，把一个源语言句子译为目标语言句子，涉及到三个基本操作：调序、删词、增词。在 MTS 的操作过程中也不例外。这就是 MT 的从源语言到目标语言的转换过程，称之为转换阶段(transfer phase)，而在这之前判断主谓宾的过程称

之为分析阶段(analysis phase)。MTS 要生成合乎目标语言语法的句子,通常还要经过一个生成阶段(generation phase)。这就是标准的 MT 实现过程,采用这样一种实现过程的 MT 方法称为基于转换(transfer - based)的方法,是 MT 系统实现的主要方法。其他方法可能采取与之不同的过程,本书在后面还要介绍。下面主要通过基于转换方法来讨论 MT 的一般实现过程,并简要介绍 MT 系统的一个实例。

1.2.2 基于转换的机器翻译过程

基于转换的 MT 系统要经过三个处理阶段:分析、转换、生成。这种方法被认为是模拟人类翻译活动的最恰当的机制[Modina & Shalyapina,1994]。从源语言到目标语言的转换是 MT 的核心操作。即使基于统计(Statistics - Based)的 MT 方法(简称 SBMT)号称不要任何语言学知识,也必须实现源语言词汇到目标语言词汇的映射。基于实例(Example - Based)的 MT 方法(简称 EBMT)通过结构化的翻译例子直接把源语言的短语和句子与目标语言的短语和句子对应起来。方法的不同使得处理步骤或多或少,但都必须实现从源语言到目标语言的转换,其映射关系或者是词到词,或者是短语或句子片断到与之相应的等价物,或者是一棵句法树到另一棵句法树。

基于转换的方法采取了一系列的分析和转换生成层次,使一个源语言句子经过不同的中间表达形式,最终到达目标语言句子的表示。其目的是尽可能地加深对源语言的理解,生成尽可能恰当的目标语言形式。这种方法比较全面地体现了语言学知识在 MT 中的应用,是了解 MT 实现过程的非常合适的一个样本。本书正是以这个过程中各个步骤应用的技术为主线来安排各章节的。

图 1.1 给出了一个比较完整的处理流程,其中列出了从源语言译为目标语言的处理步骤和所需的资源。每个步骤可以看作是计算机程序的一个相对独立的模块(module)。当然,不同的 MTS 不一定与之完全一样,可能根据需要而减少或增加某些步骤与资源,其执行的先后次序也不一定如此,可以适当地重新组合各个步

骤。处理过程中只要保证一点，即后面步骤所需的信息已从前面步骤的结果中得出。随着 MT 研究的不断深入，处理步骤也在不断变化，例如原来没有一个单独的词性标注模块，现在则增加了。

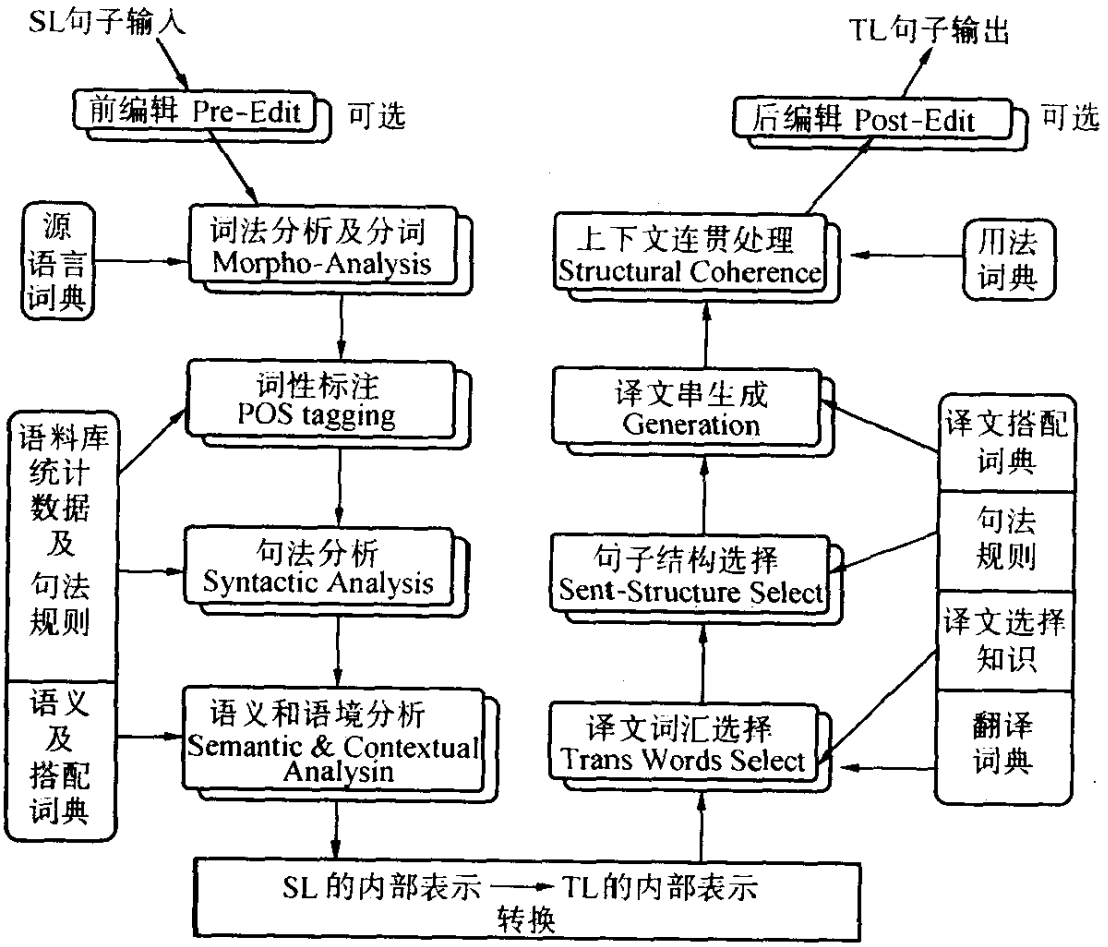

图 1.1　基于转换方法的 MT 处理过程

图 1.1 中前、后编辑的处理步骤是可选的。为了使 MTS 能够得到更好的翻译结果，有的系统加上前编辑处理模块，将被译句子作适当修改，使之更加适合本系统的翻译机制。例如，一个 MTS 附加上一个受控语言(Controlled Language)的写作环境，让写作环境来提示作者写出满足机译系统要求的句子，从而限制句子的复杂度和句型，保证系统处理的正确率[Mitamura & Nyberg, 1995]。就目前的 MT 技术来说，机译系统的译文对于参照原文进行理解一般还可以做到，但是还难以达到直接输出译文稿而无需人工润色的质量，所以后编辑处理在许多情况下必须使用。只要人工后

编辑的工作量远远小于直接从原文翻译的工作量，那么 MT 的优势就体现出来了。实际上，一些机译研究的评价就是说明改 MT 的译文句子比直接翻译原文少写了多少字符[Brown, et al.,1990; Nirenburg,1995]。

自然语言的机器翻译目前一般都是以句子为翻译单位的，句子又分为句、短语、词三个层次。翻译一个句子首先要理解源语言，即将源语言句子分析到一定程度，表示为计算机的某种内部形式。这就是图 1.1 中的左半部分，属于分析阶段。然后从内部形式经过一系列变换，形成目标语言的词串，构成目标语言句子。这就是图 1.1 中的右半部分，属于转换和生成阶段。整个机器翻译过程的输入是源语言句子，即源语言词汇串，输出是目标语言句子，即目标语言词汇串。

分析阶段一般分为词法分析、句法分析、语义分析、语境分析等几个步骤，其中以词法分析和句法分析为主。

词法分析需要实现的功能可能是单词切分(如汉语、日语)，也可能是单词形态分析(如英语、德语、法语、俄语等)，此外还有某些词组的切分等。

句法分析有一个专门的术语，叫做 Parse，通常使用其动名词形式 Parsing。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，是整个分析过程的主要部分。确定词性(或者称为词类，Part - of - Speech，简称 POS)的阶段称为词性标注(POS Tagging)，基于语料库的统计方法已成为当前词性标注的主要方法，详见第 4 章。句法分析一直是 MT 也是计算语言学研究的重点，产生了不少算法。随着研究的深入，研究者把目光转向了真实文本的分析，从而发现：对于复杂的真实句子来说，要一次性地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既可以提高整个句子分析的正确率，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，称为 Shallow Parsing。其内容包括基本

名词短语(BaseNP)的确定、句子中部分或全部短语边界的划定等。因此本书在句法分析一章里对这一部分单独进行了论述。

语义分析相对来说比较困难,涉及的知识较多,既缺乏统一表示,也缺乏有效的处理机制。但是语义对于提高译文质量是非常重要的,因而也受到研究者的普遍重视。MT 中的语义分析在很大程度上是和句法分析相联系的,因为 MT 不是强调理解一个句子,而是要在理解的基础上再转换成另一种语言的句子,而且在句法形式上要相似。

语境分析研究的是句子与句子之间的联系,也就是上下文(context)关系,此时分析已经从句子扩展到段落或语篇。上下文关系可能对句子内部的理解和翻译有影响,例如对某个代词指代哪一个名词的分析。上下文中指代关系的分析是语境分析的一个重要研究内容。

句子经过分析阶段以后,就得到了源语言句子的一种计算机内部表示,其形式一般是树型结构或者是网络结构,树型结构称为句法树(Syntax Tree)。分析 1.2.1 中例句的句法树如图 1.2 所示。

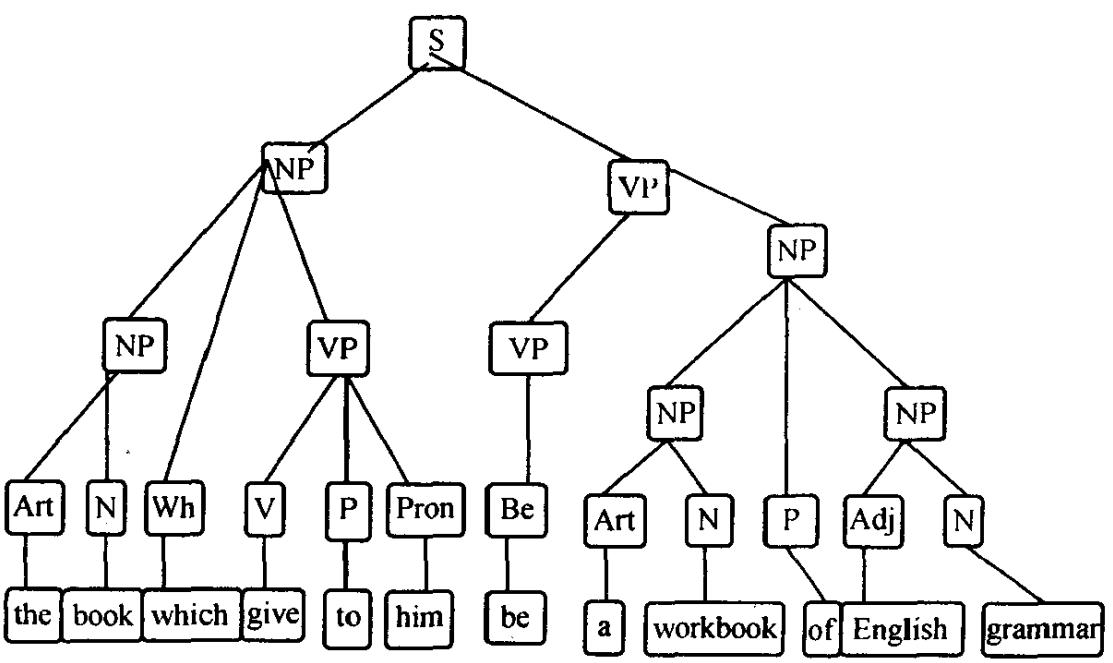

图 1.2　一个英语句子的句法树

图 1.2 中的符号说明如下：

Art——冠词

N——名词

V——动词，原文中的“was given”已还原

P——介词

Pron——代词

Be——动词 be，原文中的“was”已还原

Adj——形容词

Wh——疑问代词

NP——名词性短语

VP——动词性短语

S——句子

句子中每个单词都至少有一个词类，词类的不同组合构成了短语，短语再构成句子。按照上图的表示，一个句子内部的语法结构已经分析完毕。

句子经过分析阶段之后，进入转换和生成阶段。这两者一般联系得比较紧密，有时无法严格区分，因为转换到目标语言也就是要生成目标语言的词汇、短语、句子。在图 1.2 中是把句子结构的转换也就是源语言的句法树转换成合适的目标语言对应形式（依然是句法树结构）当作转换，而后续的许多处理都当作生成阶段的处理步骤来看待。

自然语言翻译的一个根本特征就是词汇对应关系是一对多的，那么确定一个词在一定的上下文中选择哪个译文，就是 MT 所要完成的重要任务。词汇译文选择的基础是词汇语义消歧（Word Sense Disambiguation），这方面的研究已经成为计算语言学研究的一个热点，本书在第 9 章专门介绍。句子结构的选择只有当目标语言内部表示与待生成的目标语言句子结构也存在一对多的情况下才需要，通常因为源语言到目标语言内部表示的转换就是句子结构的转换，所以经过转换后，目标语言的句子结构就已经确定下来。译文串生成一般指词汇的形态生成，根据不同的人称、时态等