翻译专业本科生系列教材

TRANSLATION

A Concise Course in Machine Translation

机器翻译简明教程

◎ 主编　李正栓　孟俊茂

A CONCISE COURSE IN MACHINE TRANSLATION

上海外语教育出版社
外教社 SHANGHAI FOREIGN LANGUAGE EDUCATION PRESS
www.sflep.com

翻译专业本科生系列教材

TRANSLATION

A Concise Course in Machine Translation

机器翻译简明教程

◎主　编　李正栓　孟俊茂
副主编　冯　梅　姬生雷

上海外语教育出版社
外教社 SHANGHAI FOREIGN LANGUAGE EDUCATION PRESS

图书在版编目(CIP)数据

机器翻译简明教程 / 李正栓，孟俊茂主编. —上海：
上海外语教育出版社，2009
（翻译专业本科生系列教材）
ISBN 978-7-5446-1420-7

Ⅰ. 机… Ⅱ. ①李…②孟… Ⅲ. 机器翻译-高等学校-
教材 Ⅳ. H085

中国版本图书馆 CIP 数据核字(2009)第 088281 号

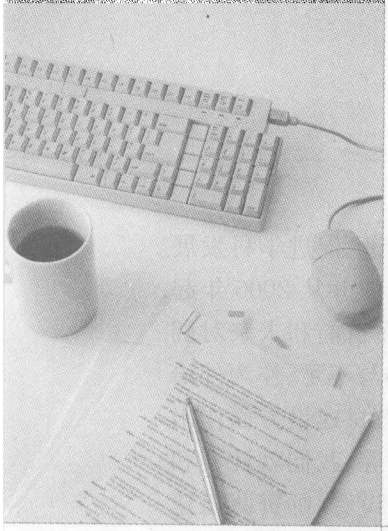
出版发行：上海外语教育出版社
（上海外国语大学内） 邮编：200083
电　　话：021-65425300（总机）
电子邮箱：bookinfo@sflep.com.cn
网　　址：http://www.sflep.com.cn http://www.sflep.com
责任编辑：李法敏

印　　刷：上海敬民实业有限公司长阳印刷厂
经　　销：新华书店上海发行所
开　　本：787×965 1/16 印张 20 字数 349千字
版　　次：2009 年 9月第 1版 2009 年 9月第 1次印刷
印　　数：3 100 册

书　　号：ISBN 978-7-5446-1420-7 / H · 0574
定　　价：32.00 元

编委会名单

（以姓氏笔画为序）



序

2006年初，国家教育部颁布了《关于公布2005年度教育部备案或批准设置的高等学校本科专业结果的通知》，“翻译”专业（专业代码：0502555，作为少数高校试点的目录外专业）获得批准：复旦大学、广东外语外贸大学、河北师范大学三所高校自2006年开始招收“翻译专业”本科生。这是迄今教育部批准设立本科“翻译专业”的首个文件，是我国翻译学科建设中的一件大事，也是我国翻译界和翻译教育界同仁数十年来勇于探索、注重积累、不懈努力、积极开拓创新的重大成果。2007年、2008年教育部又先后批准了10所院校设置翻译专业；2007年国务院学位办批准了15所院校设立翻译专业硕士点（Master of Translation and Interpretation，简称MTI），从而在办学的体制上、组织形式或行政上为翻译专业的建立、发展和完善提供了保障，形成了培养学士、硕士、博士的完整的教育体系。这必将为我国翻译学科健康、稳定、快速和持续发展，从而形成独立的、完整的专业学科体系奠定坚实的基础，亦必将为我国培养出更多更好的高素质的翻译人才，为我国的改革开放，增强与世界各国的交流和沟通，促进政治、经济、文化、教育、科技和社会各项事业的发展作出更多更大的积极贡献。

上海外语教育出版社（简称外教社）作为全国最大最权威的外语出版基地之一，自建社以来，一直将全心致力于中国外语教育事

业的发展、反映外语教学科研成果、繁荣外语学术研究、注重文化建设、促进学科发展作为义不容辞的责任。在获悉教育部批准三所院校设置本科翻译专业并从2006年起正式招生的信息后，外教社即积极开展调查研究，分析社会和市场在目前和未来对翻译人才的需求，思考翻译专业建设问题与对策、学科建设方面的优势与不足、作为外语专业出版社如何更好地服务于翻译学科的建设与发展以及如何在教材建设方面作出积极的努力和贡献。通过问卷调查、召开师生座谈会与专家咨询会等，我们就社会和市场对翻译人才的需求，我国翻译人才培养的目标、培养规格、课程设置、师资队伍建设、教学材料选择、教学方法和手段、教学测试与评估等有了初步的了解，并作了更深入的分析、思考、研究，以期在全面探索翻译专业和学科建设的基础上，承担起翻译专业教材建设的任务，为保证培养目标的实现尽一份力量。

在广泛调研和对社会和市场需求分析的基础上，外教社邀请了全国部分外语院校、综合性大学、师范院校中长期从事翻译教学与研究的近30名教授和专家，组成了“翻译专业本科生系列教材编委会”。编委会先后召开了数次工作会议，就教材的定位、体系、特点和读者对象等进行广泛而深入的讨论；尤其是对翻译作为一门课程与一个专业的异同与特点、翻译专业的定位与任务、人才培养目标与规格、教学原则与大纲、课程结构与特点、教学方法与手段、测试与评估、师资要求与培养等进行了深入的探讨和细致的分析；而后撰写了本系列教材的编写大纲，确定教材的类别，选定教材目录，讨论和审核样稿。经过两年多的努力和辛勤工作，终于迎来了“翻译专业本科生系列教材”的出版。

本系列教材由翻译理论、翻译实践与技能和特殊翻译等数个板块组成，涉及中外翻译史论、中外翻译理论、英汉—汉英互译、文学翻译、应用文翻译、科技翻译、英汉对比与翻译、计算机辅助翻译、汉语文言翻译、同声传译与交替传译、语言学与翻译、文化与翻译、作品赏析与批评等；尤其值得一提的是，在本系列教材中还针对翻译专业学生的现状和未来发展需要，专门设计和编写了中文读写教程，以丰富和提高翻译专业学生的汉语知识和应用能力。教材总数近40种，可以说比较全面地覆盖了当前我国高校翻译专业本科所开设的基本课程，可以比较好地满足和适应教学需要。

本系列教材的设计与编写，尽可能针对和贴近本科翻译专业学生的需求与特点，内容深入浅出，反映了各自领域的最新研究成果；编写和编排体例采用国家最新有关标准，力求科学、严谨、规范，满足各门课程的需要；突出以人为本，既帮助学生打下扎实的专业基本功，又着力培养学生分析问题、解决问题的能力，提高学生的人文、科学

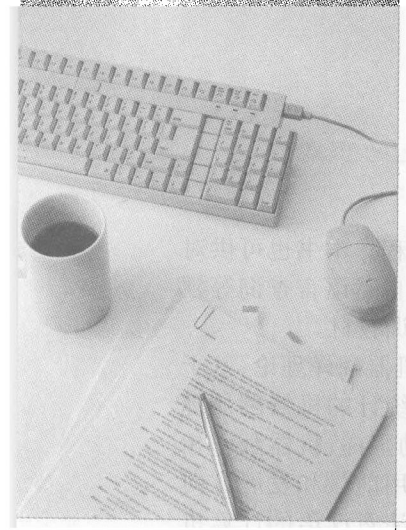
素养，培养他们奋发向上、积极健康的人生观，从而使他们全面提高综合素质，真正成为能够满足和适应我国改革开放、建设中国特色社会主义所需要的翻译专业人才。

本系列教材编委会的委员和承担各教程的主编们，大多是在我国高校长期从事翻译教学和研究的专家和学者，具有相当丰富的教学经验和科研成果，都有多年指导翻译硕士和博士研究生的经历和经验，在翻译实践和理论方面有比较深的造诣。从某种意义上说，本套教材的编写队伍和水平代表了我国当前翻译教学和研究的发展方向和水准。

鉴于本科翻译专业在我国内地是首次设立（我国台湾和香港地区早已设立本科翻译专业），教学大纲、教材建设、教学方法和手段、师资队伍建设、教学评估和管理等还有待进一步探索和实践，有待于在办学中不断提高和完善。同样，本系列教材在设计和编写中亦不可避免地存在不足和缺陷，有待广大教师和学生在使用过程中帮助我们不断完善，使其更好地服务于我国翻译专业本科生的教学学科建设及翻译人才的培养。

庄智象

2008 年 4 月



前　言

目前，有关机器翻译(Machine Translation，MT)方面的专业书籍和文章已经有很多，但是能作为教材的书还不多见。机器翻译涉及语言学、计算机科学、数学和人工智能等多个学科，是一门交叉学科，其中要求掌握的知识很广，对于一般读者而言难度很大，使他们望而却步。我们在机器翻译的教学方面已经有了一定积累和经验。因此，很多学校和出版社希望我们能把我们在机器翻译方面的教学经验和我们对机器翻译的认识整理出来，编成教材。任务很艰巨，但对机器翻译教学做一下总结还是有必要的，便硬着头皮写成此书。

机器翻译是研究计算机如何翻译人类间语言的学问。在计算机问世之初，人们就想，如果计算机能够理解和翻译人的语言，懂得人们的意图是什么，那么，我们就可以使用计算机，达到人与人之间交流的目的，那就太好了。这样计算机就能帮助人们去做枯燥、繁琐、劳动量又大的翻译工作，甚至代替人去翻译。但是在当时的条件下，这只是一种梦想。在实现这一梦想的过程中，出现过技术和认识上的危机，人们认为这样的梦想是不可能实现的。经过大批专家的研究，情况有很大的变化，计算机的功能、容量和速度都有几个数量级的提高，机器翻译的理论研究有了很大进展。因此，人们又想起了这个梦想，很多人再次为此努力奋斗，特别是新一代计算机技术和机器翻译理论的研究，使得梦想逐渐变成现实。机器翻译的研究已成为计算机科学界和语言学界的热门课题。

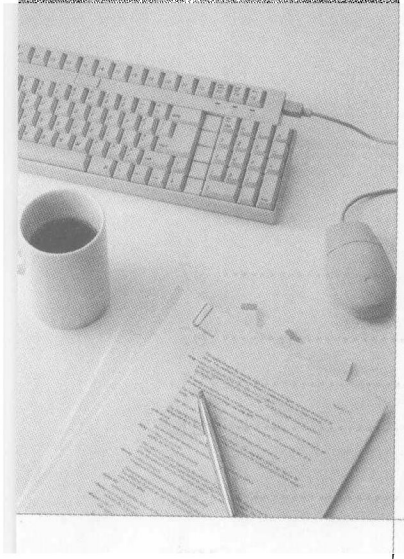
本书是机器翻译的入门导读教材，主要用作翻译专业本科生教程。本书也可供对机器翻译感兴趣，欲从事计算语言学、机器人语音对话、大型数据库自然语言查询等领域工作的研究人员参考使用，或作为使用机器翻译系统的人员的培训教材。

本书内容包括机器翻译概述、机器翻译类型介绍、基于转换的机器翻译理论、基于转换的机器翻译系统、基于中间语言的机器翻译理论和实践、基于统计的机器翻译理论和实践、基于实例的机器翻译理论和实践、译文处理、计算机辅助翻译、机器翻译相关知识、机器翻译的资源及其建设、机器翻译评价以及机器翻译应用前景和发展方向等；较为系统地介绍了机器翻译当前最主要的理论和方法，并特别注意汉语的计算机处理问题。在本书中，我们重点把汉语的处理和机器翻译的各种类型工作原理介绍给大家，其中包括机器词典、词汇语义驱动理论、中间转接语言、目标语言生成、语义关系集、规则描述语言和机译的实例等。我们的工作虽已基本完成，但还要不断地完善。请各位专家批评指正，并希望得到各位的指导。在本书编写过程中，我们参考了许多有关的论文和书籍，在此一并致谢。

所有这些内容都是我们多年来教学和研究的结果。在书稿的写作过程中，我们又重新作了整理，较多地吸收了近期机器翻译的发展成果和其他著作里的一些有意义的内容，希望尽可能反映当代机器翻译的内容体系和学术思想。相信本书对有志于从事这方面研究的读者会有所帮助。

编者

2009 年 1 月



目　录

第一单元　机器翻译概述 …… 1

第一节　什么是机器翻译 …… 1
第二节　机器翻译发展的历史 …… 6

第二单元　机器翻译类型介绍(一) …… 14

第一节　第一代机器翻译系统 …… 14
第二节　第二代机器翻译系统 …… 19
第三节　第三代机器翻译系统 …… 27

第三单元　机器翻译类型介绍(二) …… 31

第一节　实用机器翻译系统类型 …… 31
第二节　我国机器翻译系统 …… 33

第四单元　基于转换的机器翻译理论(一) …… 40

第一节　实现过程综述 …… 40
第二节　词法、词性、句法分析 …… 45

第五单元　基于转换的机器翻译理论(二) …… 97

第一节　语义分析 …… 97
第二节　译文的转换与生成 …… 121
第三节　词义消歧 …… 121

第六单元　基于转换的机器翻译系统(一) …… 133

第一节　英汉机器翻译实现过程 …… 133
第二节　英汉机器翻译系统介绍 …… 143

第七单元　基于转换的机器翻译系统(二) …… 153

第一节　汉英机器翻译实现过程 …… 153
第二节　汉英机器翻译系统介绍 …… 170

第八单元　基于中间语言的机器翻译理论和实践 …… 177

第一节　基于中间语言的机器翻译理论 …… 177
第二节　基于中间语言的机器翻译系统介绍 …… 181

第九单元　基于统计的机器翻译理论和实践 …… 187

第一节　基于统计的机器翻译理论 …… 187
第二节　基于统计的机器翻译系统介绍 …… 190

第十单元　基于实例的机器翻译理论和实践 …… 194

第一节　基于实例的机器翻译理论 …… 194
第二节　基于实例的机器翻译系统介绍 …… 198

第十一单元　译文处理 …… 203

第一节　译文转换与生成 …… 203
第二节　译文生成方法 …… 218

第十二单元 计算机辅助翻译 …… 224

第一节 计算机辅助翻译理论 …… 224
第二节 计算机辅助翻译系统 …… 233

第十三单元 机器翻译相关知识 …… 243

第一节 语法 …… 243
第二节 汉语语法概说 …… 246
第三节 英语语法概说 …… 252

第十四单元 机器翻译的资源及其建设 …… 263

第一节 词典 …… 263
第二节 语料库 …… 271

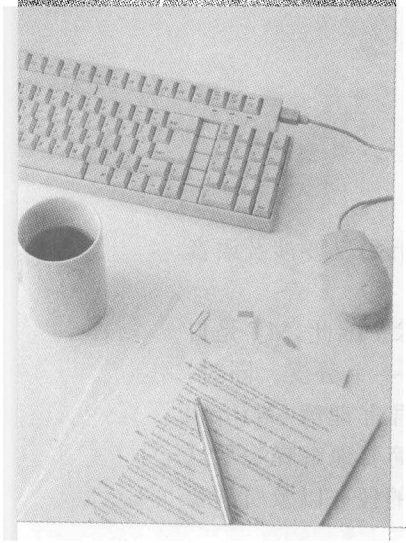
第十五单元 机器翻译评价 …… 275

第一节 机器翻译评价和其必要性 …… 275
第二节 机器翻译的评价历史 …… 278
第三节 机器翻译系统评价的内容和方法 …… 283

第十六单元 机器翻译应用前景和发展方向 …… 293

第一节 机器翻译的应用领域 …… 293
第二节 机器翻译发展方向 …… 300

参考文献 …… 304



第一单元

机器翻译概述

翻译是从一种语言到另外一种或几种语言的转换过程，是为了人与人之间的交流和沟通。世界上不同国家或民族的人们使用不同的语言，在大多数情况下需要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？能不能找到一种使用计算机来翻译的方式来满足人们的需要？这便是机器翻译要研究的内容。在这一单元里，我们将介绍什么是机器翻译，以及机器翻译是如何发展的。

第一节　什么是机器翻译

机器翻译(Machine Translation，简称 MT)就是利用计算机实现从一种自然语言文本到另一种或多种自然语言文本的翻译。它涉及语言学、计算机科学、数学等多个学科，是一门交叉学科。

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。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。计算语言学的目标从某种意义上说是试图捕捉人类的语言能力。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，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。

一、机器翻译的用途和处理对象

机器可以做翻译工作，但是 MT 不能像人一样进行各种各样的翻译，至少在将来相当长的一段时间内是无法实现的。原因如下：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。

我们应该对机器翻译系统（Machine Translation System，简称 MTS）和工具提什么样的翻译要求呢？按照英国学者 Hutchins 的分析，MT 的应用可以分为以下四类：

1）用于发行　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。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。

2）用于浏览　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。

3）用于交流　随着国际交流的日益广泛，特别是 Internet 的普及，产生即时翻译的大量需求。MT 应在这种需求当中找到其应用的角色。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，将给人们带来极大的便利。

4）用于信息获取　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。

而按照法国学者 Boitet 的分类，MT 可以分为如下四种类型：

1）用于浏览者（For The Watcher），称之为 MT-W，目的是提供粗糙的译文，以便获取某些信息；

2）用于修订者（For The Revisor），称之为 MT-R，目的是得到类似手工译文初稿的翻译；

3）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；

4）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，都说明不能笼统地谈论 MT 应用，而应该指明 MT 的具体用途。

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，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：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”？只靠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。因为根据常识或世界知识（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），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“怎么是你?”，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，例如：“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，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 系统肯定不会超过人。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 的翻译结果绝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：

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。

则对于第一类艺术作品，MT 是不能问津的，即便是能简单处理，其匹配度与可读性也是较低而不可信赖的；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。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，而不管这些句子是不是 MT 合适的处理对象。因此，说明 MT 的有限性是 MT 研究者的一个责任。

二、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：

1）学科交叉性（Cross-disciplinary）　MT 涉及计算机科学与语言学。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。

2）可计算性（Computable）　既然是 MT 而不是人类翻译，那么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。

3）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 21 世纪亟待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。

4）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。

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：

1）实践上的意义　在当今信息社会，国际交流与合作日益广泛和深入。在交流过程中，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他们所