中文信息处理丛书

汉英机器翻译若干关键技术研究

刘群 著

清华大学出版社

中文信息处理丛书

汉英机器翻译若干关键技术研究

刘群 著

清华大学出版社

北京

内 容 简 介

本书是作者所在的课题组近年来在汉英机器翻译研究方面所取得进展的一个阶段性总结。内容涉及汉英机器翻译的各个主要方面及关键技术，包括对目前国际上机器翻译研究进展的综述，汉语词法分析技术、汉语句法分析技术、汉语词汇语义相似度计算、汉英双语语料库的词语对齐、语料库的结构对齐、基于结构对齐语料库的翻译模板抽取、多引擎机器翻译方法等多方面的研究成果。

本书可供从事计算语言学、自然语言处理、中文信息处理、机器翻译等领域研究工作的人士参考，也可以作为大学相关专业高年级本科生和研究生课程的参考书。

图书在版编目（CIP）数据

汉英机器翻译若干关键技术研究/刘群著. —北京：清华大学出版社，2008.10
（中文信息处理丛书）
ISBN 978-7-302-18358-7

Ⅰ.汉…　Ⅱ.刘…　Ⅲ.英语－机器翻译－研究　Ⅳ.H315.9

中国版本图书馆 CIP 数据核字（2008）第 120702 号

责任编辑：赵彤伟　赵从棉
责任校对：赵丽敏
责任印制：王秀菊

出版发行：清华大学出版社　　**地　　址**：北京清华大学学研大厦 A 座
http://www.tup.com.cn　　**邮　　编**：100084
社　总　机：010-62770175　　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质　量　反　馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：三河市春园印刷有限公司
经　　销：全国新华书店
开　　本：185×260　　**印　　张**：10.75　　**字　　数**：254 千字
版　　次：2008 年 10 月第 1 版　　**印　　次**：2008 年 10 月第 1 次印刷
印　　数：1～3000
定　　价：32.00 元

中文信息处理丛书

编　委　会

序言

自然语言处理技术的产生可以追溯到20世纪50年代，它是一门集语言学、数学、计算机科学和认知科学等于一体的综合性交叉学科。近几年来，随着计算机网络技术和通信技术的迅速发展和普及，自然语言处理技术的应用需求急剧增加，人们迫切需要实用的自然语言处理技术来帮助人们打破语言屏障，为人际之间、人机之间的信息交流提供便捷、自然、有效的人性化服务。但是，自然语言处理中的若干科学问题和技术难题尚未得到解决，有待于来自不同领域的学者深入研究和探索。

中文信息处理作为自然语言处理中的一个分支，近几年来备受关注。一方面，随着中国经济的迅速发展和中国国力的不断增强，汉语正在成为一种新的强势语言而被世人瞩目，汉语理解所涉及的科学问题让国际计算语言学界无法回避；而另一方面，汉语使用者所拥有的巨大市场潜力令国际企业界不敢轻视。因此，中文信息处理成为全球自然语言处理研究者们共同关注的问题已经是不争的事实。目前国际上每年举行的颇具影响的几种技术评测，包括机器翻译评测、信息抽取评测和句法分析评测等，无不与汉语密切相关。因此，作为炎黄子孙，我们没有理由不在这一领域的研究中做出应有的贡献。

中文信息处理所面临的困难既有其他任何一种自然语言处理都会遇到的共性问题，如生词识别问题、歧义消解问题等，也有中文处理本身所具有的个性问题，如汉语自动分词问题、词性定义规范问题等。因此，从某种意义上讲，中文信息处理更具挑战性。值得欣慰的是，中文信息处理在引起国际学术界和企业界关注的同时，得到了中国政府的重视和大力支持，它已经被列入国务院批准的“国家中长期科学技术发展规划纲要”。因此，中文信息处理面临着前所未有的大好机遇。

近几年来，我国的中文信息处理技术得到了快速发展，无论是在基础理论研究方面，还是在技术开发和产业化发展方面，都取得了显著成绩，一大批青年学者投身到这一领域中。为了使这一领域的广大学者，尤其是青年学生，全面了解中文信息处理的技术现状，进一步推动中文信息处理及其相关学科的快速发展，我们组织编写并出版了这套中文信息处理丛书，力求将这一领域的

最新技术和理论方法全面、系统地介绍给广大读者。随着丛书的陆续出版，我们相信这套丛书必将在促进中文信息处理技术的发展和培养后继人才队伍方面发挥它应有的作用。

感谢清华大学出版社给予的支持。

倪光南

中国工程院院士

中国中文信息学会理事长

2007 年 12 月 20 日

前言

自从1949年Warren Weaver发表《翻译》备忘录，正式提出机器翻译的思想以来，到现在已经经过了半个多世纪。虽然机器翻译的现状离人们的期望和市场的需求都还有相当大的距离，远远不能满足人们的要求，不过人们对机器翻译研究的热情依然很高。这一方面是因为机器翻译的巨大需求和应用前景在不断激励着人们从事这方面的研究工作；另一方面，仅从学术角度看，机器翻译也是一个非常有意义的研究课题，其复杂性、挑战性和高难度的特点对研究者而言充满了魅力。机器翻译的研究，大大加深了人们对于语言、知识、智能等问题的了解，促进了相关学科的发展。作者认为，对全自动高质量机器翻译的不懈追求，正是计算语言学研究的终极目标之一和不竭动力的源泉。

最早的机器翻译是建立在简单的单词对译、词频统计和词序变化的基础上。当人们认识到这种方法的局限性后，开始加强了对自然语言理解的研究。伴随着人工智能研究的发展和乔姆斯基语言学的大行其道，规则方法成为了机器翻译研究的主流。研究者发现，在一些小规模应用或演示环境中表现出色的规则方法，在真正的大规模应用中却表现得非常糟糕。于是，从1990年初开始，统计方法又被重新引入到自然语言处理研究中，在机器翻译方面，IBM公司提出了著名的基于信源信道模型的统计机器翻译方法。在这以后的一段时间内，尽管统计方法在自然语言处理的很多领域都获得了成果，但对于机器翻译来说，统计方法并没有马上建立起优势地位。由于机器翻译问题本身的复杂性和计算机运行能力的限制，在很长一段时间内，很少有人能够重复IBM的统计机器翻译工作，以至于很多人对统计方法在机器翻译中的效果产生了怀疑。不过近年来，在一批研究者的不断努力下，也得益于计算能力的普遍提高，统计机器翻译终于开始表现出明显的优势并受到了普遍的重视。在最近的一些机器翻译评测中，基于统计方法的机器翻译系统取得了很好的成绩。

统计机器翻译方法近年来发展迅猛。从早期的基于词的信源信道模型的方法，到目前比较成熟的基于短语的对数线性模型的方法，再到目前热门的基于句法的统计翻译模型研究，统计机器翻译也经历了一个转换层次由浅入深的过程。机器翻译的水平比传统的基于规则的方法有了较大幅度的提高。

不过，总体上，目前机器翻译的水平依然不高。机器翻译还没有达到让一般人基本可读的水平，尤其是汉英机器翻译。

汉语是我们的母语，是数千年中华文化的主要载体，同时又是一种非常独特的语言。目前，汉字的输入、输出等方面的问题已基本解决，而汉语更深层次的处理，如词法、句法、语义分析、机器翻译等，和世界上其他一些主要语种的处理技术相比，还有一定的差距。这可能有语言学上的原因。通常人们认为，汉语是孤立语，由于缺乏形态上的标记，汉语的自动分析和处理会比其他语言更加困难。目前，自然语言处理研究的实践也证明了这一点。在句法分析方面，同样是以美国宾州大学开发的树库作为训练语料，同样采用词汇化概率上下文无关语法训练出来的汉语句法分析器和英语句法分析器，汉语句法分析器的标记正确率和召回率比英语大约低 10 个百分点。在美国国家标准技术局（NIST）举办的 2005 年机器翻译评测中，在所提供的训练语料类型和规模都大致相当的情况下，最好的阿拉伯语到英语的机器翻译系统的 BLEU[①] 评分是 0.5131，而最好的汉语到英语的机器翻译系统的 BLEU 评分只有 0.3531。这些事实表明，关于汉语的自然语言处理研究，困难确实比其他语言要更大一些。应该说，加强这方面的研究工作对中国的自然语言处理研究者来说是责无旁贷的。

本书共有以下 10 章，主要围绕汉英机器翻译中的一些关键技术展开讨论。

第 1 章是综述，介绍了机器翻译最近一段时间的研究进展，以及我们对机器翻译的一些认识。机器翻译技术方法各异，种类繁多，非常复杂。我们主要从范式和分类这两个方面对现有的机器翻译技术作了介绍。范式方面，主要介绍了传统的基于规则的转换方法以外的几种范式，包括基于平行语法的机器翻译方法、基于实例的机器翻译方法、基于信源信道模型的统计机器翻译方法、基于对数线性模型的机器翻译方法、多引擎机器翻译方法等。分类方面，主要从机器翻译的转换层面和机器翻译的知识表示形式这两个角度对现有的机器翻译方法进行了分类。

第 2 章提出了一种基于层叠的隐马尔可夫模型汉语词法分析算法。这个算法由多个层叠的隐马尔可夫模型构成，粗切分采用基于 N 最短路径的算法，简单未定义词和复合未定义词采用基于角色的隐马尔可夫模型识别新词，并采用基于角色的词语生成模型估计未定义词的概率；细切分采用词汇化的隐马尔可夫模型；词性标注采用基于词性的隐马尔可夫模型；多种模型紧密结合，下层模型不仅提供多个最好的分析结果供高层模型使用，而且也给出了这些结果的概率。模型之间环环相扣，互为补充，最终达到整体结果的最优化，同时保持算法的高效率（线性时间复杂度）。

第 3 章介绍了一种融合语义知识和词汇化上下文概率语法的汉语句法分析方法。现在主流的句法分析研究都是基于词汇化概率上下文无关语法来进行的，这种研究主要的知识来源都是事先由人工制作的树库（treebank）。由于树库的制作需要耗费大量人力，因此，树库的规模都不可能太大，因而存在比较严重的数据稀疏问题。我们通过引入两部同义词词典（“同义词词林”和“知网”），在一定程度上缓解了数据稀疏问题，提高了汉语句法分析的准确率和召回率。

第 4 章介绍了一种汉语句法分析和词法分析的融合策略。由于句法分析器训练时所采用的树库规模较小，存在比较严重的数据稀疏问题，而词法分析器则可以采用一些公开的大规模汉语切词和词性标注语料库。但由于切词和词性标注的标准不一致，用这种大规模语料库训练出来的词法分析器还不能直接用在句法分析器中。为此我们分别采用错误驱动的

① 一种机器翻译质量的自动评分指标。

方法和条件随机场算法，将词法分析器分析得到的结果通过切词和词性标注两方面的转换，适应了树库的切词和词性标注标准，成功地将不同标准的词法分析器和句法分析器进行了融合，显著提高了句法分析的性能。

第5章提出了一种基于“知网”的词汇语义相似度计算模型。这种方法充分利用了“知网”中所包含的丰富的人类语言学知识，直接计算两个词语的语义相似度，而无需通过大规模语料库的训练，方法简单有效。这种方法可广泛用于词义排歧、基于实例的机器翻译等多个领域。

第6章提出了一种对数线性模型的词语对齐方法。这一方法首次将判别训练的思想引入词语对齐研究中，使得我们可以利用各种形式的特征来改进词语对齐的性能，大大拓宽了词语对齐研究的思路，也显著降低了词语对齐的平均错误率。

第7章提出了一种高效的双语短语对齐搜索算法。这种算法的主要优点是可以尽可能避免词语对齐错误给短语对齐带来的干扰，使得短语对齐的正确率和召回率比词语对齐的相应指标都要高出很多，效果很好。算法采用柱形搜索策略，时间消耗随着句子长度线性增长，效率也非常高。

第8章定义了一种可以刻画两种语言深层句法结构对应关系的短语结构转换模板，并给出了从双语短语对齐的语料库中抽取这种模板的算法。对实验结果的初步分析表明，从一个八千句子对的短语对齐语料库中抽取出来的模板，已经可以覆盖各种常见的汉英句法结构的转换模式。

第9章提出了一种微引擎流水线机器翻译系统结构。在这种结构下，整个机器翻译过程被分解成若干个串行的阶段，每个阶段可以有若干个功能相似的部件（微引擎）同时工作。通过添加和删除微引擎以及调整流水线的结构很容易实现各种机器翻译构件的协调工作，而无需修改系统的总体翻译算法和数据结构，有利于提高机器翻译系统的开发效率以及尝试新的机器翻译方法。其中介绍了一个基于这种结构实现的面向新闻领域的汉英机器翻译系统，并给出了实验结果。

第10章对本书进行了全面总结，介绍了下一步的工作计划。

本书涵盖的研究工作反映了作者所在研究团队在汉英机器翻译研究方面进行的一系列努力，本书也是对我们现有研究工作的一个阶段性总结。我们的这些工作受到国家重点基础研究项目（“973”计划）子课题“面向大规模真实文本的汉语计算理论、方法和工具”（课题编号G1998030507-4）、国家自然科学基金重点项目“融合语言知识与统计模型的机器翻译方法研究”（项目批准号60736014）和中国高技术研究发展计划（“863”计划）重点项目课题“面向跨语言搜索的机器翻译关键技术研究”（课题编号2006AA010108）资助，特此表示感谢！

机器翻译研究涉及的领域非常广，技术门类也非常多。本书每一章都是对汉英机器翻译中某一项具体关键技术的研究。虽然在本书中我们还没有将这些技术集成到一个完整的机器翻译系统中去，不过所有这些技术对于构造一个完整的机器翻译系统都是非常重要的。我们希望本书对真正希望从事汉英机器翻译研究的人士有所帮助，对相关领域的研究人员和学生也能够具有参考价值。

刘　群

2008.5

目录

第1章

机器翻译方法综述

经过50多年的发展，产生了很多不同的机器翻译方法。比如，人们常常提到基于规则的方法、基于统计的方法、同基于规则相结合的方法、基于实例的方法、中间语言方法、转换方法、基于知识的方法，等等。这些方法种类繁多，都有各自的优缺点。但这些方法往往是从不同角度、不同层面出发的，互相之间并不一定具备可比性。人们在初次接触机器翻译时，往往会被如此众多的方法所迷惑，如坠雾中，不容易理解这些方法之间内在的区别与联系。

本章将从范式(paradigm)和分类这两个角度对机器翻译方法进行初步梳理，不仅要对各种机器翻译方法作一个概要介绍，而且试图刻画出它们之间的联系与区别。

范式，指的是对某些具体的机器翻译实现方法的一种抽象和归纳。范式往往要对机器翻译方法的某些方面作出明确规定，而对另外一些方面可以没有明确要求。但由于范式往往都有一些典型的实现方法或具体系统，所以即使对那些没有明确要求的方面，人们往往也都有一些默认的理解。比如说，基于转换的方法，作为一种范式，本身并没有规定采用规则方法还是统计方法，但人们谈到这种方法时，往往把它理解成一种基于规则的方法。这是由于该方法出现时，还没有出现现在意义上的统计机器翻译方法。而且一些典型的基于转换的系统，也都是采用规则方法实现的。另外，不同的范式往往对机器翻译方法的不同方面和不同层次作出规定，所以范式之间往往不具有可比性。可以这么说，通过范式对机器翻译方法进行总结，就是人们常说的抓典型的方法，或解剖麻雀的方法。通过对范式的研究，可以起到解剖麻雀的作用，有助于对机器翻译的实现技术进行比较全面和深入了解。

但范式并不等同于分类。科学的分类往往要求执行统一的分类原则，分类的结果要求满足完整性和互斥性。而范式的定义并没有统一的标准和原则，因而也不满足完整性和互斥性。也就是说，可以有一些具体的机器翻译实现方法，同时满足两种或两种以上的范式，而有些机器翻译的实现方法可能不符合现有的任何一种范式。而分类则不然，一旦分类原则确定下来，任何一种具体的机器翻译实现方法必定落入一种且只能落入一个类别当中。根据分类原则的不同，可以有各种不同的分类结果。通过对各种机器翻译方法进行分类研究，将各种机器翻译方法进行多角度多层次的比较，可以对机器翻译的各种方法之间的区别和联系有比较清楚的了解。

本章先介绍一些机器翻译中常见的范式，然后试图用分类的方法对这些方法进行总结和归纳。

1.1 机器翻译的范式

机器翻译的范式很多，常见的一些如下：

(1) 直接翻译方法。早期的不经过句法分析直接进行词语翻译和词序调整的方法。

(2) 基于转换的方法。基于源语言和目标语言的深层表示形式进行转换的方法，典型的转换方法要求独立分析，独立生成。注意，这里的深层表示既可以是句法表示，也可以是语义表示。

(3) 基于中间语言的方法。利用独立于具体语言的某种中间表示形式(称为中间语言)实现两种语言之间翻译的方法。

(4) 基于语言学的方法。以语言学知识为推理基础，在对源语言的句法语义进行深入分析和理解的基础上进行翻译的方法。

(5) 基于知识的方法。利用人工智能中知识表示、知识推理技术进行机器翻译的方法。

(6) 基于平行语法的方法。通过为源语言和目标语言构造一套平行的语法体系，在分析的同时完成翻译的方法。

(7) 基于实例的方法。通过将源语言句子和实例库中已有的句子进行类比得到译文的方法。严格地说，基于实例的方法不是一种机器翻译的典型范式，因为其中的不确定性因素太多，各种具体的基于实例的机器翻译方法实施方案相差较大。

(8) 基于信源信道模型的统计机器翻译方法。将翻译理解为信息传输的过程，通过对语言模型和翻译模型的估计来求解最佳译文的方法。

(9) 基于对数线性模型的统计机器翻译方法。直接将翻译的概率分解为一组特征函数的乘积，通过对数线性模型进行参数估计的方法。

一些传统的机器翻译方法早已广为人知，如直接翻译方法、基于转换的方法、基于中间语言的方法等，这里不再详细介绍。本书主要介绍一些比较新的或者与本书关系较为密切的方法，包括基于平行语法的方法、基于实例的方法和基于统计的方法等。

1.2 基于平行语法的机器翻译方法

这种方法的基本思想是，用一套双语平行的语法模型，即两组相互对应的规则，同时生成两种语言的句子，在对源语言句子进行理解的同时，就可以得到对应的目标语言句子的生成过程。

这种方法的基本特点是：有明确的规则形式；源语言规则和目标语言规则一一对应；如果采用概率形式，那么源语言与目标语言服从相同的概率分布，即对应的规则在两种语言中出现的概率相同；对于两种语言的转换过程不使用概率模型进行描述。

以下分别介绍这种方法中几种具体形式。

1.2.1 Alshawi 的基于加权中心词转录机的统计机器翻译方法

有限状态转录机(finite-state transducer，FST)和有限状态识别器(finite-state recognizer，FSR)是有限状态自动机(finite-state automata，FSA)的两种基本形式。其主要区别在于有限状态转录机在识别的过程中同时可以产生一个输出，其每一条边上面同时有

输入符号和输出符号两个标记，而有限状态识别器只能识别，不能输出，其每一条边上只有一个输入符号标记。

中心词转录机(head transducer，HT)是对有限状态转录机的一种改进。对于中心词转录机，识别的过程不是自左向右进行，而是从中心词开始向两边执行。所以在每条边上，除了输入输出信息外，还有语序调整信息，用两个整数表示。图 1.1 是一个能够将任意 a、b 组成的串逆向输出的中心词转录机示意图。

基于加权中心词转录机(weighted head transducer，WHT)的统计机器翻译方法是由 AT&T 实验室的 Alshawi 等提出的，用于 AT&T 的语音机器翻译系统。该系统由语音识别、机器翻译、语音合成三部分组成。其中机器翻译系统的总体工作流程如图 1.2 所示。

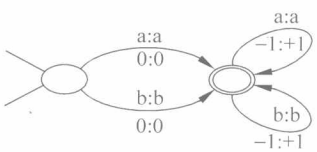

图 1.1　中心词转录机示意图

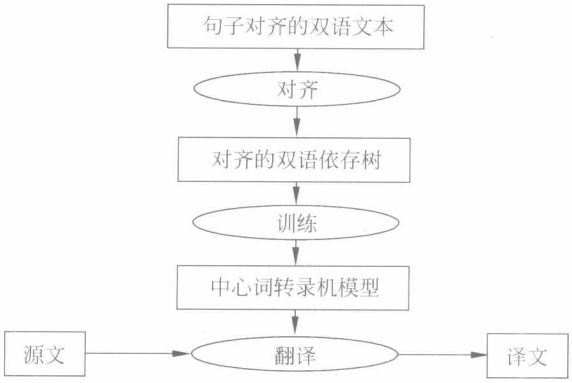

图 1.2　基于中心词转录机的机器翻译系统工作流程

在加权中心词转录机模型中，中心词转录机是唯一的知识表示方法，所有的机器翻译知识，包括词典，都表示为一个带概率的 HT 的集合。知识获取过程是全自动的，从语料库中训练得到，但获取的结果(就是中心词转录机)很直观，可以由人进行调整。中心词转录机的表示是完全基于词的，不采用任何词法、句法或语义标记。

整个知识获取的过程实际上就是一个双语语料库结构对齐的过程。句子的结构用依存树表示(但依存关系不作任何标记)。他们经过一番公式推导，把一个完整的双语语料库的分析树构造并对齐的过程转化成了一个数学问题的求解过程。这个过程可用一个算法高效实现。得到对齐的依存树后，很容易就训练出一组带概率的中心词转录机，也就得到了一个机器翻译系统。不过要说明的是，通过这种纯统计方法得到的依存树，与语言学意义上的依存树并不符合，而且相差甚远。

这种方法的主要特点是：①训练可以全自动进行，效率很高，由一个双语句子对齐的语料库可以很快训练出一个机器翻译系统；②不使用任何人为定义的语言学标记(如词性、短语类、语义类等)，无需任何语言学知识；③训练得到的参数包含了句子的深层结构信息，这一点比 IBM 的统计语言模型更好。

这种方法比较适合于语音翻译这种领域比较受限、词汇集较小的场合。

1.2.2　吴德凯的反向转录语法

反向转录语法(inversion transduction grammar，ITG)是香港科技大学吴德凯(Dekai Wu)提出的一种供机器翻译使用的语法形式[Wu 1997]。

这种语法的特点是，源语言和目标语言共用一套规则系统。