学习机器翻译

Cyril Goutte Nicola Cancedda
Marc Dymetman George Foster 著

曹海龙 赵铁军
朱聪慧 杨沐昀 译

Learning Machine Translation

科学出版社

学习机器翻译

Cyril Goutte　Nicola Cancedda
Marc Dymetman　George Foster　著

曹海龙　赵铁军
朱聪慧　杨沐昀　译

科学出版社
北京

图字：01-2013-5035

内 容 简 介

本书首先讨论若干使能技术，即解决那些不是机器翻译本身但却是与机器翻译系统开发相关的技术，其中包括从可比语料中获取双语句子对齐数据、多语名称词典的自动构造、词对齐技术等。随后介绍若干新的、改进的统计机器翻译技术，包括利用句法信息的判别式训练框架、半监督学习方法和基于核的学习方法的应用以及多机器翻译译文的组合以改进整个翻译系统的质量。

本书适合于从事机器翻译的研究者和研究生阅读，读者应具备统计机器翻译的入门知识。

图书在版编目(CIP)数据

学习机器翻译/(加)古特(Goutte, C.)等著；曹海龙等译. —北京：科学出版社，2014. 10

书名原文：Learning Machine Translation

ISBN 978-7-03-042297-2

Ⅰ. ①学… Ⅱ. ①古…②曹… Ⅲ. ①机器翻译-翻译机-基本知识 Ⅳ. ①H085②TP391.2

中国版本图书馆 CIP 数据核字(2014)第 250852 号

责任编辑：张艳芬 范 勃 / 责任校对：张小霞
责任印制：肖 兴 / 封面设计：陈 敬

科学出版社出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

北京佳信达欣艺术印刷有限公司印刷
科学出版社发行 各地新华书店经销
*
2014 年 10 月第 一 版 开本：720×1000 1/16
2014 年 10 月第一次印刷 印张：19
字数：362 000

定价：98.00 元

（如有印装质量问题，我社负责调换）

译 者 序

当今互联网时代对于机器翻译的需求越来越迫切，因为人们随时都可能浏览自己根本不懂的语言的网页。机器翻译这个极具商业和社会价值的研究领域受到了全球同行们的极大关注。改进技术使之更好地服务于人类需求，一直是研究者们孜孜追求的目标。在国际计算语言学学会的年度盛会（ACL）上，机器翻译始终是近年来发表文章最多的主题。

当前机器翻译的主流技术是统计机器翻译技术，其采用了大量机器学习特别是统计学习方法。统计机器翻译以大规模的双语数据为驱动，通过统计学习的手段挖掘出其中的翻译规律。这就是本书题名为“学习机器翻译”的缘故：计算机以“学习”为手段完成“机器翻译”。

本书是几年前一个研讨会论文的扩充，其中所涉及的问题虽有待继续解决，但所介绍的方法仍具有参考价值。本书除了第一章关于统计机器翻译的概述外，其余章节分为两部分：第一部分称为“使能”技术，这些技术虽然不直接服务于机器翻译系统的开发，却与统计机器翻译系统所依赖的基础资源相关，其中包括双语对齐语料的构建、多语名称词典的构建、命名实体识别、词对齐和语言模型构建技术等；第二部分介绍统计机器翻译在训练、调参、解码各个实现阶段的技术，包括判别式训练、翻译短语选择、基于核的翻译模型、句子重构翻译模型、半监督学习方法应用、翻译结果重排序和多机器翻译系统的译文融合等，多方实验表明这些技术对于改进整个统计翻译系统的输出质量有很大帮助。

本书的翻译分工如下：赵铁军负责第 1、4、12 章，曹海龙负责第 2、6、7、9、10 章，朱聪慧博士负责第 3、5、11 章，杨沐昀负责第 8、13 章。在本书翻译过程中，哈尔滨工业大学计算机学院机器智能与翻译实验室的多位研究生参与了初稿的翻译，他们是：刘淋（第 2 章）、崔一鸣（第 3 章）、李婷婷（第 4 章）、朱晓宁和崔一鸣（第 5 章）、赵弈欧和李晓倩（第 7 章）、张宇（第 8 章）、张文文（第 9 章）、张捷鑫（第 10 章）、史华兴（第 12 章）、朱俊国（第 13 章）等。

本书的出版得益于多方面的支持。感谢译者团队的成员侯亚楠同学，她为本书的校对、排版等付出了辛勤劳动。感谢哈尔滨工业大学教育部-微软语言语音重点实验室为本书出版提供了资助。在翻译过程中，译者承担了机器翻译方向的多项国家科研课题，包括国家 863 计划重大项目“互联网语言翻译系统研制”（2011AA01A207）、国家自然科学基金项目（61173073、61100093、61272384），这些研究工作无疑为翻译工作提供了坚实的知识基础。

限于译者水平，书中难免存在疏漏和不足之处，恳请各位读者不吝指教。译者邮箱：hailong@mtlab. hit. edu. cn。

原书前言

外国语总是围绕在我们身边。现代通信技术使得人们能够获取自身并不完全理解的语言信息。数以亿计的互联网用户任何时间都可以在线，问题是语言障碍使得人们不能相互沟通。自动翻译的梦想激发着人们对自动或半自动翻译方法的持续兴趣。尽管自动翻译在有限领域和应用上取得了成功，千百万网页每天都被自动翻译，但是机器翻译却经常遭受那些并不欣赏产生全自动翻译这种挑战的人们的怀疑。但是，至少在现阶段，这是一个统计方法占据主流的快速发展和异常丰富的领域。

本书是2006年12月“神经信息处理系统会议”组织的用于多语言信息获取的机器学习研讨会的一个续篇。本书的几位作者在2006年的会议上就其工作作了介绍，但涉及的内容并不全面，而本书中约有半数内容是未公开发表的。

与以往的研讨会比较，本书也是坚持把统计机器翻译作为重点。本书分为两部分。第一部分涉及使能技术，即那些解决并不是机器翻译本身但与开发机器翻译系统紧密相连的问题的技术。例如，第2章涉及从可比语料库中获取双语句子对齐数据，这对于那些没有平行语料的领域或语言对是一个至关重要的任务。第3章和第4章探讨各种命名实体的多语言等价体的识别问题。该项技术的一个应用就是改进跨语言命名实体的翻译。第5章涉及词对齐，对于大多数统计机器翻译系统来说是一个基本的使能技术。它展示如何利用多预处理机制来改进对齐的质量。第6章展示词序核如何能被用于不同类型的语言学信息，并且指出这种方法可以改进判别式语言模型。

本书第二部分介绍新的统计机器翻译技术以及依靠统计方法或机器学习在现有技术基础上的改进。第7章探讨在翻译模型中纳入句法信息的问题，在一个判别式训练框架下进行参数估计。第8章提出一种在极大规模语料库上训练的统计机器翻译系统的假设输出重排序新方法。第9章介绍一种机器翻译新方法，它避开传统的特征函数对数线性组合，代之以基于核的方法（据作者所知，这是机器翻译领域涉及该方向的第一个研究）。第10章和第11章重点放在翻译模型的词或短语选择的改进上。在第10章中，基于源语言的全局信息，一个判别式过程决定目标语言词汇表中一个词是否出现在目标语言句子中，同时使用语言模型合作的加权转换机制为所选择的语言词汇进行排序。在第11章中，一个判别式短语选择模型被集成在一个基于短语的统计机器翻译系统中。这一章同时给出对大量的自动机器翻译评价标准的有趣分析和比较。第12章探索半监督学习的使

用，通过利用源语言中大量未翻译的材料来改进机器翻译输出。第 13 章展示多机器翻译系统的输出如何能够被组合起来，以便改进整体翻译质量。这种方法允许若干合作者贡献不同的机器翻译系统进行合作。系统融合目前在国际评测中产生了较好的结果。

作者期待本书对机器学习和统计机器翻译两个领域的研究者有所帮助。希望机器学习研究者将对统计机器翻译各个前沿方向有一个很好的整体了解，对于学习方法不同形式的直接影响力有所了解，并对一个具有挑战性的重要问题进行探索。也希望统计机器翻译研究者能够对本书产生兴趣，特别是对导论级统计机器翻译教材中可能没有覆盖的某些先进题目的介绍，对某些机器学习所激发的、具有产生本领域新方向潜力的新方法的介绍给予关注。

特别感谢 Susan Buckley 和 Robert Prior 在书稿撰写方面给予的支持。特别感谢为提高本书质量而提出意见和建议的以下人员：Caroline Brun，Marine Carpuat，Mauro Cettolo，Hal Daumé III，Hervé Déjean，Andreas Eisele，Alex Fraser，Patrick Haffner，Xiaodong He，Pierre Isabelle，Roland Kuhn，Dragos Munteanu，Miles Osborne，Jean-Michel Renders，Antti-Veikko Rosti，Craig Saunders，Libin Shen，Michel Simard，Sandor Szedmak 和 Dan Tufis。

Cyril Goutte
Nicola Cancedda
Marc Dymetman
George Foster

目　录

第一部分:使能技术

第 1 章　统计机器翻译初步

Nicola Cancedda, Marc Dymetman, George Foster, Cyril Goutte

第 1 章主要是关于统计机器翻译(statistical machine translation, SMT)方面的一个简短介绍。特别是,覆盖了机器翻译(machine translation, MT)输出的自动评价、语言建模、基于词和基于短语的翻译模型以及 MT 中句法的应用。也对一些较新的方向做了一个快速综述,这些方向在未来具有重要意义。将 SMT 置于机器学习研究的一般情况下,并将重点放在它与标准机器学习问题和实践的相似和差异之上。

1.1　背　景

MT 带着雄心勃勃的目标和未完成的承诺开始了其长期发展的历史。早期的工作一如当时的认识,以自动的或者"机械式"的翻译面貌出现,至少可以追溯到 20 世纪 40 年代。尽管其发展过程中存在少数磕绊如美国的 ALPAC 报告,但是 MT 依然以各种方式伴随着计算机科学和人工智能的进步而迅速发展(Hutchins, 2003)。

越来越强大的计算能力的提供使得 MT 的获取和使用更加直接。MT 也通过诸如搜索引擎服务这样的专门服务给公众以更广泛的展示。许多互联网用户熟悉至少一种如巴别鱼[①]、谷歌语言工具[②]、视窗在线翻译[③]一类的工具。这些服务中的大部分由基于规则的系统所支持,如 SYSTRAN 所开发的系统[④],可是其中的某些系统(如谷歌和微软)至少部分地使用了统计方法。

在本章和本书的其他部分,翻译被定义为这样一个任务:将一个现存的源语言书写的文本转换为一个等价的不同语言即目标语言文本。传统的 MT(在本章论中指"统计之前的翻译")依赖于源语言端不同层次的语言学分析和目标语言端的语言生成(图 1.1)。

---

① http://babelfish.yahoo.com/ or http://babelfish.altavista.com/。

② http://www.google.com/language_tools。

③ http://translator.live.com/。

④ http://www.systransoft.com/。

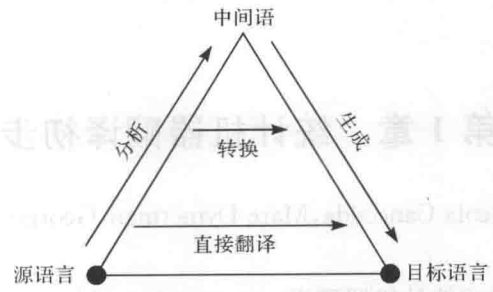

图 1.1 MT 金字塔

各种方法根据对分析和生成的依赖程度而不同：**中间语言**方法全部依赖分析和生成，而**直接翻译**方法最少依赖分析和生成，**转换**方法介于二者之间

在 20 世纪 80 年代后期，来自 IBM 的一个研究组开拓了第一个 MT 统计方法（Brown et al，1990）。事实上这可被看做是计算语言学总体潮流转移的一部分：在大约十年内，统计方法成为了这个领域中压倒性的支配方法。例如，在计算语言学协会（Association for Computational Linguistics，ACL）的年度会议论文集里的情况就是如此。

SMT 一般是指如何从源语言和目标语言的等价句对构成的大语料库中学习翻译。这是一个典型的机器学习框架：我们有一个输入（源语言句子），一个输出（目标语言句子），和一个试图为每一个给定输入产生正确输出的模型。

这里包含一系列关键问题，然而，其中一些关乎于 MT 应用。一个至关重要的问题是翻译质量的评价。机器学习技术典型地依赖于某些代价优化，以便学习输入和输出数据之间的关系。不过，自动地评价一个译文的质量或与某个给定的 MT 输出所相关的代价，是一个非常困难的问题。它可被归入于语言理解得更广泛问题，并且在很大程度上将停留在尚待解决的阶段。关于 MT 评价代价的定义与自动计算的困难，将在本章 1.2 节详述。

由 IBM 研究组倡导的 SMT 早期方法是信源通道方法。这基本上是两个模型组合的一个框架：一个基于词的翻译模型（1.3 节）和一个语言模型（language model，LM）（1.4 节）。翻译模型保证了系统生成源语言句子对应的目标语言假设，而 LM 保证了输出的译文是尽可能符合语法和流畅的。

基于词的翻译模型取得了一些进展。然而，一个显著的突破却是采用对数线性模型和基于短语的翻译所带来的。这些内容将在 1.5 节中有更详细的描述。

虽然早期的 SMT 模型基本上忽略了语言学方面，但已有一系列努力尝试在翻译模型或者 LM 方面重新引入语言学思考。这些内容由 1.6 节和后续的某些章节所包含。此外，我们在 1.7 节给出了一个关于 SMT 目前趋势的概述，其中的一些内容在后面的章节中也会介绍。