

<<统计机器翻译>>

图书基本信息

书名：<<统计机器翻译>>

13位ISBN编号：9787121175923

10位ISBN编号：7121175924

出版时间：2012-9

出版时间：电子工业出版社

作者：菲利普·科恩

页数：301

字数：525000

版权说明：本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问：<http://www.tushu007.com>

## <<统计机器翻译>>

### 内容概要

Philipp

Koehn所著的《统计机器翻译》是介绍统计机器翻译理论和方法的教材。

全书分三部分(共11章), 分别讨论基础知识、核心方法和前沿研究。

全书首先简要介绍语言学和概率论基础知识, 然后全面讨论各种经典统计机器翻译模型和系统实现方法, 最后深入探讨统计翻译领域的最新进展和研究热点。

对核心方法的论述按照统计机器翻译模型发展的过程逐步展开: 基于词的模型、基于短语的模型和基于句法树的模型。

从技术实现的角度, 本书还介绍了统计翻译模型的参数训练方法、语言模型和参数平滑方法、解码算法和译文自动评测方法及系统整合方法等。

《统计机器翻译》是统计机器翻译和自然语言处理课程的理想教材, 适合研究生和本科生教学使用, 也是所有对机器翻译技术和系统有兴趣的研究者、开发者和使用者的指南性读物。

同时, 本书还可作为人工智能、语言学等相关专业的辅助读物。

## &lt;&lt;统计机器翻译&gt;&gt;

## 作者简介

作者：（德国）菲利普·科恩（Philipp Koehn）译者：宗成庆 张霄军 菲利普·科恩，英国爱丁堡大学信息学院讲师（lecturer）。

欧洲EuroMatrix项目的科学协调员，同时参与了美国DARPA资助的研究项目。

与机器翻译领域的知名公司如Systran和AsiaOnline等都建立了合作。

实现了广为使用的解码器Pharaoh，同时领导着开源机器翻译工具Moses的开发。

宗成庆，1998年3月毕业于中国科学院计算技术研究所，获博士学位。

1998年5月至2000年4月在中国科学院自动化研究所从事博士后研究，博士后出站后留自动化所工作至今，现为模式识别国家重点实验室研究员、博士生导师。

曾于1999年和2001年两次在日本国际电气通信基础技术研究所（ATR）做客座研究员，2004年在法国Grenoble信息与应用数学研究院机器翻译研究组（GETA—CLIPS，IMAG）做短期高访。

主要研究方向为自然语言处理基础、机器翻译、文本分类和自动文摘等相关技术。

作为项目负责人承担国家自然科学基金项目、国家“863”项目、国家支撑计划项目和国际合作研究项目等10余项，在国内外重要学术期刊和会议上发表论文100余篇，其中在ComputationalLinguistics

、Information Sciences、IEEE TASLP、ACM TALIP、Machine Translation及ACL、COLING、EMNLP等本领域权威期刊和会议上发表论文20多篇，出版学术专著1部，获8项国家发明专利。

目前担任国际计算语言学联合会（ACL）汉语特别兴趣组（SIGHAN）候任主席（ChairElect）和亚洲自然语言处理联合会（AFNLP）执行理事，并担任国际学术期刊IEEE IntelligentSystems副主编

（Associate Editor）、ACM TALIP副主编、UCPOL副主编、Machine Translation编委、JCST编委、《自动化学报》编委，以及中国中文信息学会常务理事、中国人工智能学会理事和中国计算机学会中文信息技术专委会副主任等职务。

2008年获中国科学院研究生院集中教学突出贡献奖。

2009年获亚太地区语言、信息与计算国际会议（PACLIC）最佳论文奖，2010年获中国科学院“朱李月华优秀教师”奖。

张霄军，2008年6月毕业于南京师范大学，获博士学位。

现为陕西师范大学外国语学院副教授，硕士生导师。

2010年至2011年在英国曼彻斯特大学访学，研究方向为现代翻译技术。

目前承担国家社科基金项目1项，参与国家自然科学基金项目1项及国家社科基金项目1项。

在国际学术期刊Computational Linguistics、Information Retrieval、Language Learning&Technology

和Applied Linguistics等发表学术论文4篇，在《当代语言学》和《计算机应用研究》等国内期刊发表学术论文50余篇。

出版学术专著《语义组合与机器翻译》（科学出版社，2010），主（参）编教材多部。

## &lt;&lt;统计机器翻译&gt;&gt;

## 书籍目录

## 第1章 绪论

## 1.1 概述

- 1.1.1 第1章：绪论
- 1.1.2 第2章：词、句子和语料
- 1.1.3 第3章：概率论
- 1.1.4 第4章：基于词的翻译模型
- 1.1.5 第5章：基于短语的翻译模型
- 1.1.6 第6章：解码
- 1.1.7 第7章：语言模型
- 1.1.8 第8章：评测
- 1.1.9 第9章：判别式训练
- 1.1.10 第10章：整合语言学信息
- 1.1.11 第11章：基于树的翻译模型

## 1.2 机器翻译简史

- 1.2.1 肇始
- 1.2.2 ALPAC报告及其后果
- 1.2.3 首批商用系统
- 1.2.4 基于中间语系统的研究
- 1.2.5 数据驱动方法
- 1.2.6 目前的开发商
- 1.2.7 技术现状

## 1.3 应用

- 1.3.1 全自动高质量机器翻译
- 1.3.2 要旨翻译
- 1.3.3 集成语音技术
- 1.3.4 手持设备中的翻译
- 1.3.5 后编辑
- 1.3.6 译者的工具

## 1.4 可用资源

- 1.4.1 工具
- 1.4.2 语料
- 1.4.3 评测竞赛

## 1.5 小结

- 1.5.1 核心概念
- 1.5.2 延伸阅读

## 1.6 习题

## 第2章 词、句子和语料

## 2.1 词

- 2.1.1 词例化
- 2.1.2 词的分布
- 2.1.3 词性
- 2.1.4 形态学
- 2.1.5 词汇语义学

## 2.2 句子

- 2.2.1 句子结构

## &lt;&lt;统计机器翻译&gt;&gt;

- 2.2.2 语法理论
- 2.2.3 句子结构的翻译
- 2.2.4 语篇

## 2.3 语料

- 2.3.1 文本的类型
- 2.3.2 获取平行语料
- 2.3.3 句子对齐

## 2.4 小结

- 2.4.1 核心概念
- 2.4.2 延伸阅读
- 2.4.3 习题

## 第3章 概率论

## 3.1 概率分布估计

- 3.1.1 估计分析
- 3.1.2 常见概率分布
- 3.1.3 基于统计的概率估计

## 3.2 概率分布计算

- 3.2.1 形式定义
- 3.2.2 联合概率分布
- 3.2.3 条件概率分布
- 3.2.4 贝叶斯法则
- 3.2.5 插值

## 3.3 概率分布的特性

- 3.3.1 均值和方差
- 3.3.2 期望和方差
- 3.3.3 熵
- 3.3.4 互信息

## 3.4 小结

- 3.4.1 核心概念
- 3.4.2 延伸阅读
- 3.4.3 习题

## 第二部分 核心方法

## 第4章 基于词的翻译模型

## 4.1 基于词的机器翻译

- 4.1.1 词汇翻译
- 4.1.2 数据统计
- 4.1.3 估计概率分布
- 4.1.4 对齐
- 4.1.5 IBM模型1

## 4.2 学习词汇翻译模型

- 4.2.1 语料不完备问题
- 4.2.2 期望最大化算法
- 4.2.3 IBM模型1中的期望最大化算法
- 4.2.4 困惑度

## 4.3 确保流畅的输出

- 4.3.1 流利译文的经验证据
- 4.3.2 语言模型

## &lt;&lt;统计机器翻译&gt;&gt;

## 4.3.3 噪声信道模型

## 4.4 更高级的IBM模型

## 4.4.1 IBM模型2

## 4.4.2 IBM模型3

## 4.4.3 训练模型3：采样对齐空间

## 4.4.4 IBM模型4

## 4.4.5 IBM模型5

## 4.5 词对齐

## 4.5.1 词对齐任务

## 4.5.2 词对齐质量评估

## 4.5.3 基于IBM模型的词对齐

## 4.6 小结

## 4.6.1 核心概念

## 4.6.2 延伸阅读

## 4.6.3 习题

## 第5章 基于短语的翻译模型

## 5.1 标准模型

## 5.1.1 基于短语的翻译模型提出的动因

## 5.1.2 数学定义

## 5.2 学习短语翻译表

## 5.2.1 从词对齐中抽取短语

## 5.2.2 一致性定义

## 5.2.3 短语抽取算法

## 5.2.4 应用实例

## 5.2.5 短语翻译概率估计

## 5.3 翻译模型的扩展

## 5.3.1 对数线性模型

## 5.3.2 双向翻译概率

## 5.3.3 词汇化加权

## 5.3.4 词语惩罚

## 5.3.5 短语惩罚

## 5.3.6 作为分类问题的短语翻译

## 5.4 调序模型的扩展

## 5.4.1 调序限制

## 5.4.2 词汇化调序

## 5.5 基于短语模型的期望最大化训练

## 5.5.1 短语对齐的联合模型

## 5.5.2 对齐空间的复杂度

## 5.5.3 模型训练

## 5.6 小结

## 5.6.1 核心概念

## 5.6.2 延伸阅读

## 5.6.3 习题

## 第6章 解码

## 6.1 翻译过程

## 6.1.1 翻译一个句子

## 6.1.2 计算句子的翻译概率

## &lt;&lt;统计机器翻译&gt;&gt;

## 6.2 柱搜索

## 6.2.1 翻译选项

## 6.2.2 通过假设扩展的解码过程

## 6.2.3 计算复杂度

## 6.2.4 翻译假设重组

## 6.2.5 栈解码

## 6.2.6 直方图剪枝和阈值剪枝

## 6.2.7 调序限制

## 6.3 未来代价估计

## 6.3.1 不同的翻译困难

## 6.3.2 翻译选项的未来代价估计

## 6.3.3 任意输入跨度的未来代价估计

## 6.3.4 在搜索中使用未来代价

## 6.4 其他解码算法

## 6.4.1 基于覆盖栈的柱搜索算法

## 6.4.2 A\*搜索算法

## 6.4.3 贪婪爬山解码

## 6.4.4 有限状态转换机解码

## 6.5 小结

## 6.5.1 核心概念

## 6.5.2 延伸阅读

## 6.5.3 习题

## 第7章 语言模型

## 7.1 n元文法语言模型

## 7.1.1 马尔可夫链

## 7.1.2 估计

## 7.1.3 困惑度

## 7.2 计数平滑

## 7.2.1 加1平滑法

## 7.2.2 删除估计平滑法

## 7.2.3 古德图灵平滑法

## 7.2.4 评估

## 7.3 插值和后备

## 7.3.1 插值

## 7.3.2 递归插值

## 7.3.3 后备

## 7.3.4 预测词的差异性

## 7.3.5 历史的差异性

## 7.3.6 修正的Kneser-Ney平滑算法

## 7.3.7 评估

## 7.4 控制语言模型的大小

## 7.4.1 不同的n元文法的数目

## 7.4.2 在磁盘上进行估计

## 7.4.3 高效的数据结构

## 7.4.4 减小词汇表规模

## 7.4.5 抽取相关的n元文法

## 7.4.6 根据需要加载n元文法

## &lt;&lt;统计机器翻译&gt;&gt;

## 7.5 小结

## 7.5.1 核心概念

## 7.5.2 延伸阅读

## 7.5.3 习题

## 第8章 评测

## 8.1 人工评测

## 8.1.1 流利度和忠实度

## 8.1.2 评测目的

## 8.1.3 其他评测标准

## 8.2 自动评测

## 8.2.1 准确率和召回率

## 8.2.2 词错误率

## 8.2.3 BLEU：一个双语评测的替代指标

## 8.2.4 METEOR

## 8.2.5 关于评测的争论

## 8.2.6 评测指标的评测

## 8.2.7 自动评测不足的证据

## 8.3 假设检验

## 8.3.1 计算置信区间

## 8.3.2 成对比较

## 8.3.3 自举重采样

## 8.4 面向任务的评测

## 8.4.1 后编辑的代价

## 8.4.2 内容理解测试

## 8.5 小结

## 8.5.1 核心概念

## 8.5.2 延伸阅读

## 8.5.3 习题

## 第三部分 前沿研究

## 第9章 判别式训练

## 9.1 寻找候选译文

## 9.1.1 搜索图

## 9.1.2 词格

## 9.1.3 n-best列表

## 9.2 判别式方法的原理

## 9.2.1 译文的特征表示

## 9.2.2 标注译文的正确性

## 9.2.3 监督学习

## 9.2.4 最大熵

## 9.3 参数调节

## 9.3.1 实验设置

## 9.3.2 Powell搜索方法

## 9.3.3 单纯型算法

## 9.4 大规模判别式训练

## 9.4.1 训练问题

## 9.4.2 目标函数

## 9.4.3 梯度下降



## &lt;&lt;统计机器翻译&gt;&gt;

9.4.4 感知机

9.4.5 正则化

9.5 后验方法与系统融合

9.5.1 最小贝叶斯风险

9.5.2 置信度估计

9.5.3 系统融合

9.6 小结

9.6.1 核心概念

9.6.2 延伸阅读

9.6.3 习题

第10章 整合语言学信息

10.1 直译

10.1.1 数字和名字

10.1.2 名字翻译

10.1.3 直译的有限状态方法

10.1.4 资源

10.1.5 反向直译与翻译

10.2 形态学

10.2.1 词素

10.2.2 简化丰富的形态变化

10.2.3 翻译形态丰富的语言

10.2.4 单词拆分

10.3 句法重构

10.3.1 基于输入语言句法的调序

10.3.2 学习调序规则

10.3.3 基于词性标记的调序

10.3.4 基于句法树的调序

10.3.5 预留选择

10.4 句法特征

10.4.1 方法论

10.4.2 数的一致性

10.4.3 一致性

10.4.4 句法分析概率

10.5 因子化翻译模型

10.5.1 因子化翻译的分解

10.5.2 因子化模型训练

10.5.3 模块的融合

10.5.4 高效解码

10.6 小结

10.6.1 核心概念

10.6.2 延伸阅读

10.6.3 习题

第11章 基于树的翻译模型

11.1 同步文法

11.1.1 短语结构语法

11.1.2 同步短语结构语法

11.1.3 同步树替换文法

## <<统计机器翻译>>

### 11.2 同步文法的学习

#### 11.2.1 层次短语模型的学习

#### 11.2.2 句法翻译规则的学习

#### 11.2.3 规则的简化

#### 11.2.4 文法规则的打分

### 11.3 基于句法分析算法的解码

#### 11.3.1 线图分析

#### 11.3.2 核心算法

#### 11.3.3 线图的组织

#### 11.3.4 假设重组

#### 11.3.5 栈剪枝

#### 11.3.6 文法规则的使用

#### 11.3.7 立方剪枝

#### 11.3.8 文法二叉化

#### 11.3.9 外向代价估计

### 11.4 小结

#### 11.4.1 核心概念

#### 11.4.2 延伸阅读

#### 11.4.3 习题

### 参考文献

### 索引

## 章节摘录

版权页：插图：6.4其他解码算法 前面已经详细介绍了基于短语的翻译模型中常用的解码算法：柱搜索栈解码器。

该解码器也可以用在基于词的翻译模型中。

现在再介绍相关文献中出现的其他几种解码算法，作为本章的总结。

6.4.1 基于覆盖栈的柱搜索算法 根据翻译过的外语输入单词的数量在栈里组织翻译假设，给未来代价估计带来了额外的复杂度。

但是，如果为每个覆盖了相同外语输入单词的跨度都设立一个栈，就不会有额外的复杂度。

如果仅仅比较翻译相同跨度的外语单词的假设，根据定义，它们的未来代价是一样的，就可以忽略未来代价。

需要注意的是，这样做仍然有可能造成搜索错误：当某个翻译假设看起来可能比搜索图中的另一点的候选翻译假设更好时，它可能以一个导致在下一步的语言模型得分上更差的英语单词结束，最终不是最优路径中的一部分。

这种覆盖栈（coverage stacks）的问题是指数级的复杂度，从而导致算法不可行。

但是，回顾一下6.2.7节为使用调序限制所做的阐述。

调序限制可将可能的外语单词覆盖向量的数量减少到与句子长度成线性关系（尽管与调序限制仍然呈指数关系）。

因此，使用调序限制的覆盖栈解码方法是可行的。

6.4.2 A\*搜索算法 这里介绍的柱搜索算法与很多人工智能教科书上介绍的A\*搜索（A\*search）算法非常类似。

A\*搜索算法允许零风险地对搜索空间剪枝，换句话说，防止了搜索错误。

A\*搜索算法对在未来代价估计中使用的启发式方法进行了限制。

A\*搜索算法使用一种可接纳的启发式方法（admissible heuristic），该方法要求估计代价不能过高。

注意这种方法是如何安全地用于对翻译假设进行剪枝的：如果某翻译假设的局部得分加上已估计的未来代价，仍然小于最小的完整翻译假设路径的代价，就可以安全地将该翻译假设删除。

可接纳的启发式机器翻译解码方法 6.3节介绍的未来代价的启发式方法并不是一种可接纳的方法：它可能会过高地或过低地估计真实翻译代价。

那么，如何才能适应这种启发式方法呢？

如果忽略调序代价，仅仅使用翻译表中的真实短语翻译代价，就不会冒过高地估计模块代价的风险。

但是，估计语言模型的代价是非常粗略的，它忽略了前面的上下文信息，因此有可能过高或过低地估计真实的翻译代价。

也可以考虑有用的历史信息建立优化的语言模型估计。

例如，对于短语中第一个单词的概率，可以找到在给定任意历史条件下的最高概率。

搜索算法 为了使A\*搜索更有效，必须快速地找到一个真实的、完整的、代价最低的早期候选。

为此，使用图6.12所示的深度优先方法。

## <<统计机器翻译>>

### 编辑推荐

近年来，机器翻译领域因统计技术的出现而充满了活力，从而使人类语言自动翻译的梦想与实现更加接近。

这本由该领域一位活跃的研究者撰写、经过课堂检验的教科书，向读者简要、通俗地介绍了该领域的最新研究方法，使读者能够通过《国外计算机科学教材系列:统计机器翻译》的学习为任何语言对构建机器翻译系统。

版权说明

本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问:<http://www.tushu007.com>