第十章 机器翻译

(1/3)

内容提要

1、机器翻译概论

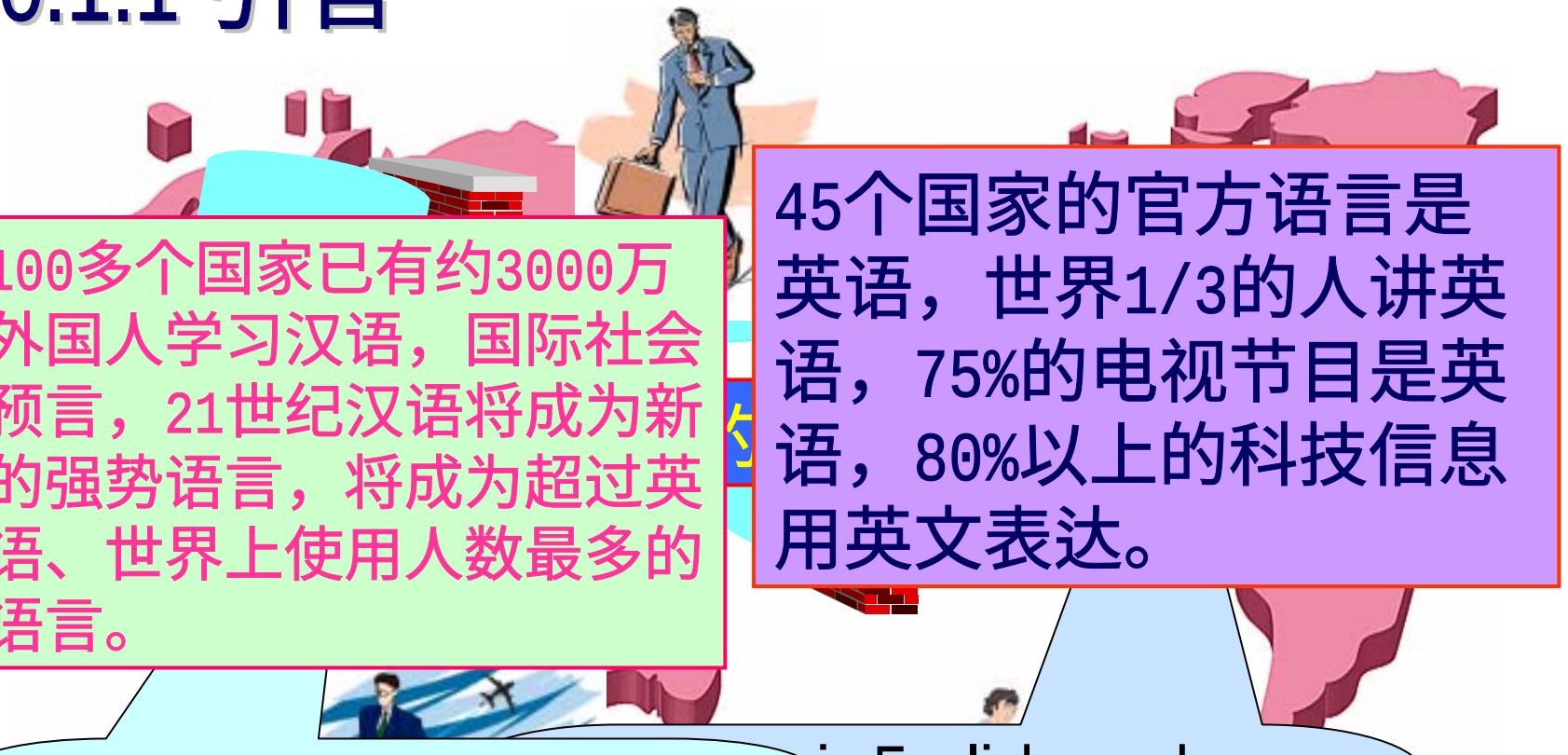
2、统计机器翻译

3、CASIA SMT系统

4、语音翻译

1、机器翻译概论

10.1.1 引言



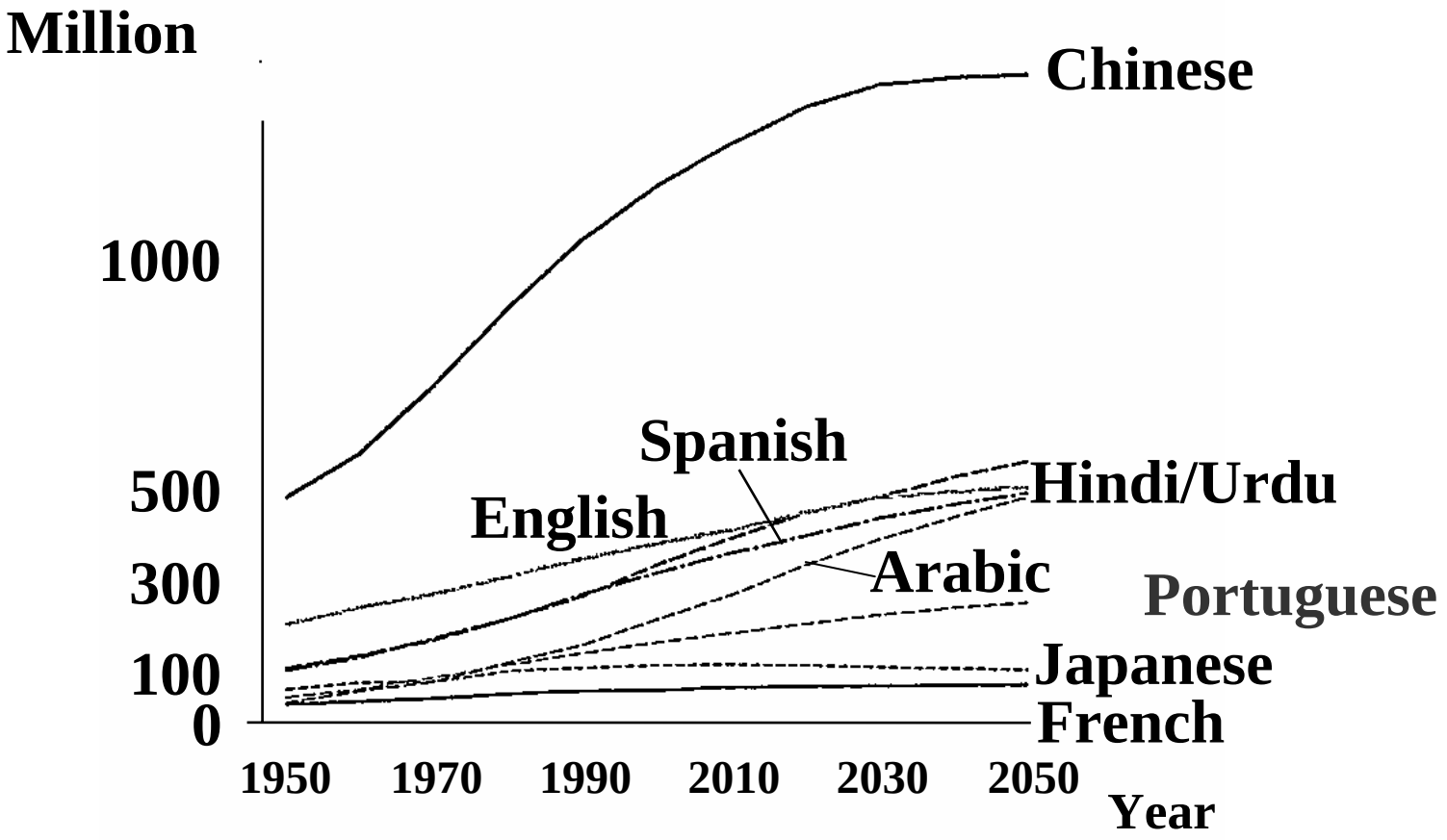
100多个国家已有约3000万外国人学习汉语，国际社会预言，21世纪汉语将成为新的强势语言，将成为超过英语、世界上使用人数最多的语言。

45个国家的官方语言是英语，世界1/3的人讲英语，75%的电视节目是英语，80%以上的科技信息用英文表达。

汉语是世界上使用人数最多的语言。

in English speakers
guage: 380 million
language: 380 million
Learners: 750 million

10.1.1 引言


Million
Chinese
1000
Spanish
500
English
Hindi/Urdu
300
Arabic
Portuguese
100
Japanese
0
French
1950
1970
1990
2010
2030
2050
Year

10.1.1 引言

- 有关专家已经指出，语言障碍是21世纪国际社会全球化面临的主要困难之一。
- 机器翻译涉及语言学、计算语言学、认知科学、通信技术等多种学科，具有重要的科学意义。
- 具有巨大的社会需求（包括奥运会在内）。
- 欧洲有380多种语言，目前欧盟20种官方语言，每年为11种语言翻译、转录文件耗费的人力费用大约 549M 欧元。
- 汉语已经不再仅仅是中国人关注的语言。

10.1.2 机器翻译的产生与发展

❑ 概念：机器翻译 (machine translation, MT) 是用计算机把一种语言 (源语言，source language) 翻译成另一种语言 (目标语言，target language) 的一门学科。

机器翻译是一门学科也是一门技术。

10.1.2 机器翻译的产生与发展

- 起始与发展
  - 古希腊
  - **17世纪**：笛卡儿(Descartes)莱布尼兹(Leibniz)试图用统一的数字代码编写词典；17世纪中页贝克(Cave Beck)等人出版类似的词典。
  - **1930s**：亚美尼亚法国工程师阿尔楚尼(G. B. Arsouni)提出了用机器来进行语言翻译的想法，并在1933年7月22日获得了一项“翻译机”的专利，叫做机器脑 (mechanical brain)

10.1.2 机器翻译的产生与发展

- 1933年，前苏联发明家特洛扬斯基设计了用机械方法把一种语言翻译成为另一种语言的机器。
- 1946年 世界上第一台电子计算机 ENIAC 诞生以后，英国工程师 A. D. Booth 和美国洛克菲勒基金会 (Rockefeller Foundation) 副总裁 W. Weaver 提出了利用计算机进行机器翻译的设想。

10.1.2 机器翻译的产生与发展

- Andrew Donald Booth 数学物理学家，曾研究利用X射线确定晶体结构，二战中参与计算机研制，在程序化计算机研究中成绩卓著；
- 1947年3月－9月，曾在普林斯顿大学参与 John von Neumann 研究组
- 曾在伦敦大学工作

- 信息论先驱
- 1920-32曾在Wisconsin大学教授数学
- 1932-55 Rockefeller Institute自然科学部主任

10.1.2 机器翻译的产生与发展

- 1947年3月 W. Weaver 给 N. Wiener 写信，讨论机器翻译问题，但遭到了 N. Wiener 的反对
- 1949年，W. Weaver 发表了以《翻译》为题目的备忘录，正式提出机器翻译问题。

10.1.2 机器翻译的产生与发展

◆ **Weaver 的两个基本观点**：

a) 翻译类似于解读密码的过程：当我阅读一篇用俄语写的文章时，我可以说这篇文章实际上是用英文写的，只不过它用另外一种奇怪的符号编了码，当我阅读时，我是在进行解码；

10.1.2 机器翻译的产生与发展

b) 原文和译文“说的是同样的事情”，因此，当把语言A翻译为语言B时，就意味着，从语言A出发，经过某一“通用语言(universal language)”或“中间语言 (interlingua)”，然后转换为语言B，这种“通用语言”或“中间语言”可以假定是全人类共同的。

10.1.2 机器翻译的产生与发展

- 美国和英国的学术界对机器翻译产生了浓厚的兴趣，并得到了实业界的支持
- **1954年** Georgetown 大学在 IBM 协助下，用IBM-701计算机实现了世界上第一个 MT 系统，实现俄译英翻译，1954年1月该系统在纽约公开演示。系统只有250条俄语词汇，6条语法规则，可以翻译简单的俄语句子。
- 在随后10 多年里，MT 研究在国际上出现热潮。

10.1.2 机器翻译的产生与发展

- **1964年**，美国科学院成立语言自动处理咨询委员会(Automatic Language Processing Advisory Committee, ALPAC)，调查机器翻译的研究情况，并于1966年11月公布了一个题为“语言与机器”的报告，简称 ALPAC 报告，宣称：

“在目前给机器翻译以大力支持还没有多少理由”

“机器翻译遇到了难以克服的语义障碍(semantic barrier)”

从此，机器翻译研究在世界范围内进入低迷状态。

10.1.2 机器翻译的产生与发展

- **1970～1976年**，法国、日本、加拿大等国，仍坚持机器翻译研究，而且，在这一阶段语法与算法分开，机器翻译研究开始复苏。代表系统：法国的 ARIANE-78。

10.1.2 机器翻译的产生与发展

- **1976～至今**，繁荣时期：
  - 加拿大蒙特利尔大学与加拿大联邦政府翻译局联合开发的实用机器翻译系统 TAUM-METEO，用于天气预报翻译。每小时可以翻译6－30万个词；每天翻译1500－2000篇天气预报资料，并通过电视、报纸等立即公布。
  - 1978年欧共体启动多语言机器翻译计划；

10.1.2 机器翻译的产生与发展

- **1982－1986**，日本在提出第五代机的同时，研究日英双向机器翻译系统Mu和亚洲多语言机器翻译（日语、汉语、印尼语、马来西亚语、泰国语）
- **1990年** IBM提出统计机器翻译模型，机器翻译研究进入了一个空前辉煌的繁荣时期。

10.1.2 机器翻译的产生与发展

- 概括为三个阶段：

  1954 ～ 1970 (ALPAC)：草创时期；

  1970 ～ 1976：复苏阶段；

  1976 ～ 现在：繁荣时期。

10.1.3 机器翻译的困难

- 自然语言中普遍存在的歧义和未知现象：
  - 句法结构歧义/词汇歧义/语用歧义 …
  - 新的词汇、术语、结构、语义 …
- 机器翻译不仅仅是字符串的转换
  - 不同语言之间文化的差异
  - 现有方法无法表示和利用世界知识和常识
- 机器翻译的解不唯一，而且始终存在的人为的标准

10.1.4 机器翻译研究现状

- 若干翻译系统已实用化或接近实用化
  - **Systran (http://www.systransoft.com)**
  - TAUM-METEO / Google
  - 华建英汉翻译系统等
- 仍面临若干问题
  - 理论模型与方法
  - 实现技术问题
  - 用户认识问题

10.1.4 机器翻译研究现状

- 应尽快消除对机器翻译的误解
  - 编造的故事

1962年8月号《哈泼杂志》(Harper's Magazine) 发表 John A. Kouwenhoven 的文章 "翻译的困扰 (The trouble with translation)"，编造了如下故事：

一个机器翻译系统包含1500个基础英语词汇和相应的俄语词汇，翻译测试：

英语：Out of sight, out of mind. （眼不见，心不烦。）

翻译成俄语句子的意思：Invisible idiot.

（看不见的白痴。）

10.1.4 机器翻译研究现状

英语：The spirit is willing, but the flesh is weak.

（心有余而力不足。）

翻译成俄语句子的意思：

The liquor is holding out all right, but the meat has spoiled.

(酒保存的很好，但肉已经腐烂。)

10.1.4 机器翻译研究现状

黛玉自在枕上感念宝钗……又听见窗外竹梢焦叶之上，雨声淅沥，消寒透幕，不觉又滴下泪来。
(《红楼梦》第45回)

As she lay there alone, Dai-Yu's thoughts turned to Bao-chai …Then she listened to the insistent rustle of the rain on the bamboos and plantains outside her window. The coldness penetrated the curtains of her bed. Almost without noticing it she bad begun to cry.

文学翻译家 David Hawkes

摘自冯志伟著《机器翻译研究》，2004

10.1.4 机器翻译研究现状

- 实际水平

(SYSTRAN: http://www.systransoft.com/index.html)

(1)输入：我可以帮助您吗？

英文：I may help you?

(2)输入：I am not feeling well.

中文：我不感觉好。

10.1.4 机器翻译研究现状

❑ 基本观点

在机器翻译研究中实现人机共生(man-machine symbiosis), 人机互助比追求完全自动的高质量的翻译(Full Automatic High Quality Translation, FAHQT)更现实、更切合实际[Hutchins, 1995]。

我们需要的是计算机帮助人类完成某些翻译工作，而不是完全替代人，人与机器翻译系统之间应该是互补的关系，而不是相互竞争[Hutchins, 2001].

10.1.4 机器翻译研究现状

机器翻译研究在过去的五十多年曲折发展经历中，无论是它给人们带来的希望还是失望我们都必须客观地看到，机器翻译作为一个科学问题在被学术界不断深入研究的同时，企业家们已经从市场上获得了相应的利润。

- 机器翻译还不成熟 (in state-of-the-art)，需要的是人与系统的配合，而不是有意为难，辅助机器翻译可以大大减轻人的负担。
- 追求“信、达、雅”是人类翻译的目标，计算机在这方面永远都不会替代人。

10.1.5 基本翻译方法

- 直接翻译方法
- 基于规则的翻译方法
- 基于中间语言的翻译方法
- 基于语料库的翻译方法
  - 基于事例的翻译方法
  - 统计翻译方法

10.1.5 基本翻译方法

❑直接翻译方法

从源语言句子的表层出发，将单词或者短语甚至句子直接置换成目标语言译文，必要时进行一些词序的调整，便可以生成译文的句子。对原文句子的分析仅仅满足于特定译文生成的需要。这类翻译系统一般真对某一个特定的语言对，将分析与生成、语言数据、文法和规则与程序等都融合在一起。

例如：I like Mary. → Me(I) gusta(like) Maria(Mary).

X like Y → Y X gusta

10.1.5 基本翻译方法

❑基于规则的方法

1957年美国学者V．Yingve在《句法翻译框架》(Framework for Syntactic Translation) 一文中提出了对源语言和目标语言均进行适当描述、把翻译机制与语法分开、用规则描述语法的实现思想，这就是基于规则的**(Rule-based)** 转换翻译方法。

10.1.5 基本翻译方法

基于规则的翻译过程分成六个步骤：

a) 对源语言句子进行词法分析

b) 对源语言句子进行句法分析/语义分析

c) 源语言句子结构到译文结构的转换

d) 译文句法结构生成

e) 源语言词汇到译文词汇的转换

f) 译文词法选择与生成

10.1.5 基本翻译方法

由于基于规则的翻译方法执行过程为：

“独立分析－独立生成－相关转换”

因此，又称基于转换的翻译方法。

其代表系统是法国格勒诺布尔机器翻译研究所(GETA)开发的ARIANE翻译系统。

1976年加拿大蒙特利尔大学与加拿大联邦翻译局联合开发的实用性机器翻译系统 TAU-METEO：天气预报信息服务。

10.1.5 基本翻译方法

基于规则转换的翻译方法的评价：

主要优点：可以较好地保持原文的结构，产生的译文结构与源文的结构关系密切，尤其对于语言现象已知的或句法结构规范的源语言语句具有较强的处理能力和较好的翻译效果。

主要不足：分析规则一般由人工编写，工作量大，规则的主观性强，一致性难以保障，不利于系统扩充，对非规范语言现象缺乏相应的处理能力 。

10.1.5 基本翻译方法

❑基于中间语言的翻译方法（**Interlingua-based**）

- 方法：输入语句→ 中间语言→ 翻译结果
- 代表系统：JANUS (CMU) 早期版本
  - ☆ 源语言解析器
  - ☆ 比较准确的中间语义描述语言(Interlingua)
  - ☆ 目标语言生成器(Target Language Generator)

10.1.5 基本翻译方法

—国际先进语音翻译研究联盟(C-STAR)定义的中间转换格式 (Interchange Format, IF)

—日本东京联合国大学 (United Nations University) 提出的通用网络语言 (Universal Networking Language, UNL)

10.1.5 基本翻译方法

对基于中间语言的翻译方法评价：

主要优点：中间语言的设计可以不考虑具体的翻译语言对，因此，该方法尤其适合多语言之间的互译。

主要弱点：如何定义和设计中间语言的表达方式，以及如何维护并不是一件容易的事情，中间语言在语义表达的准确性、完整性等很多方面，都面临若干困难。

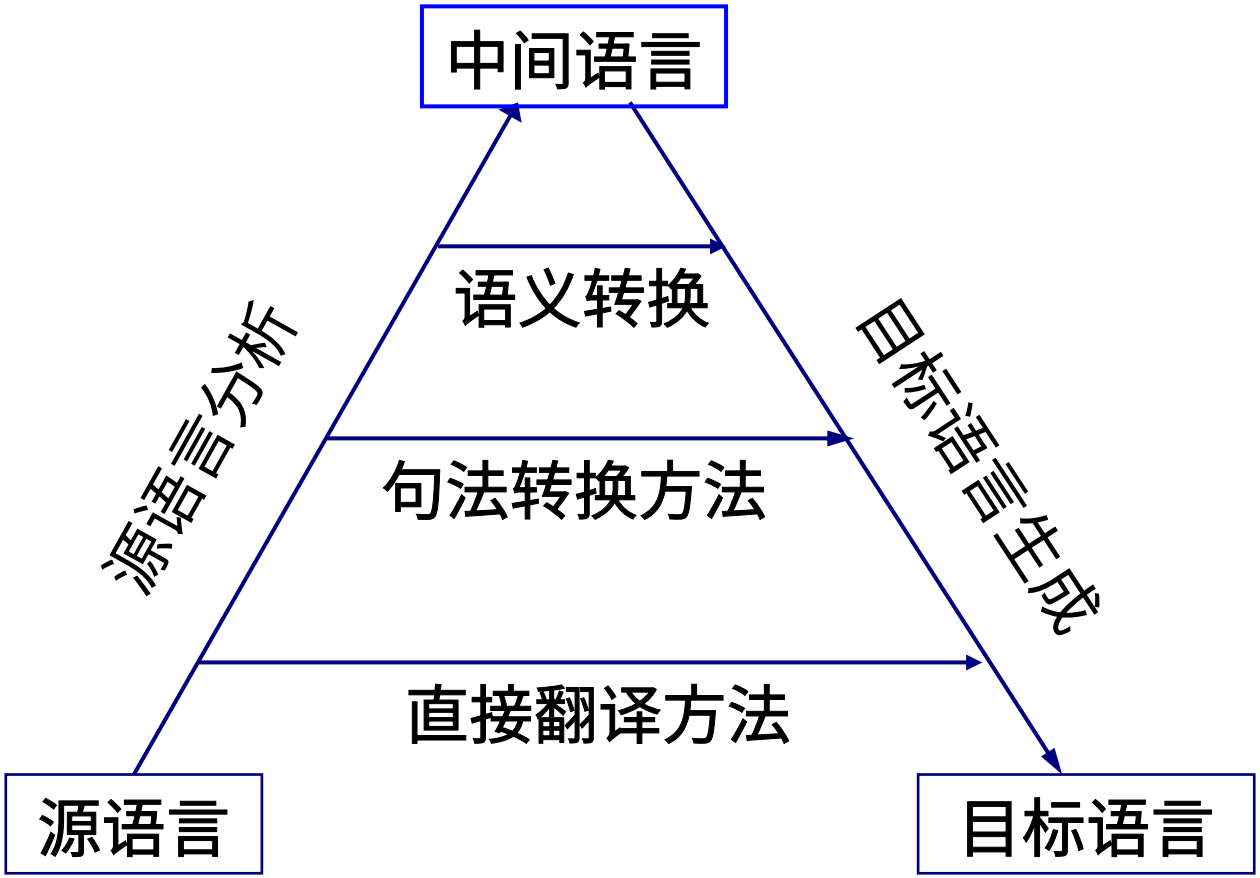
10.1.5 基本翻译方法

- 基于事例的翻译方法 **(Example-based) [1984, 长尾真]**
  - 方法：输入语句→ 与事例相似度比较→翻译结果
  - 资源：大规模事例库
  - 代表系统：ATR-MATRIX（ATR, Japan)

10.1.5 基本翻译方法

- 其它翻译方法
  - 基于记忆的翻译方法 (memory-based MT)
  - 基于神经网络的翻译方法 (neural network MT)
  - **基于统计的翻译方法（statistical method）**
  - 基于多引擎的翻译方法(multi-engine)

10.1.5 基本翻译方法



本讲小结

- 机器翻译的产生与发展
- 研究现状
- 机器翻译基本方法
  - 直接翻译(基于模板)
  - 基于规则的翻译方法
  - 基于中间语言的翻译方法
  - 基于事例的翻译方法

Thanks

谢谢!