中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

机器翻译原理与方法

第一讲 概论

刘群
中国科学院计算技术研究所
liuqun@ict.ac.cn

中国科学院计算技术研究所 2011 年秋季课程

内容提要

- **机器翻译定义**
- 课程目的与特点
- 课程概况
- 机器翻译历史
- 机器翻译现状
- 机器翻译应用系统类型

什么是机器翻译

English

中文

什么是机器翻译

- 机器翻译 (machine translation) 是使用电子计算机把一种自然语言 (源语言 ,source language) 翻译成另外一种自然语言 (目标语言 ,target language) 的一门学科
- 这门学科同时也是一种技术 . 它涉及到语言学、计算机科学、数学等许多部门 , 是非常典型的多边缘的交叉学科
  - 在语言学中 , 机器翻译是计算语言学的一个研究领域
  - 在计算机科学中 , 机器翻译是人工智能的一个研究领域
  - 在数学中 , 机器翻译是数理逻辑和形式化方法的一个研究领域 .

以上定义引自冯志伟《**澄清对机器翻译的一些误解（论文提要）**》，现代语文（语言研究），2005.1，做了个别修改

内容提要

- 机器翻译定义
- **课程目的与特点**
- 课程概况
- 机器翻译历史
- 机器翻译现状
- 机器翻译应用系统类型

课程目的

- 讨论：为什么要学习机器翻译？
- 直接目的：学会机器翻译
  - 了解机器翻译基本原理
  - 掌握机器翻译实践技能：
    能够编写一个简单的机器翻译系统
- 间接目的：学会做研究
  - 学会分析问题和解决问题
  - 学会发现问题和提出问题：
    问题是嵌套的，尽量探究最原始、最本质的问题！

课程特点

- 目标驱动
  - 目标：解决机器翻译问题
  - 目标驱动：
    - 不是为学习而学习，为方法而方法
    - 学习任何问题都要跟目的联系起来，多想想为什么
- 注重实践：知行合一（王阳明）
  - 学不懂的时候，去编程实现！
  - 编程效果不好的时候，去看书、看论文！

内容提要

- 机器翻译定义
- 课程目的与特点
- **课程概况**
- 机器翻译历史
- 机器翻译现状
- 机器翻译应用系统类型

课程概况

- 教师介绍
- 学生情况
- 时间安排
- 课程内容
- 作业安排
- 网络资源
- 学术会议
- 学术刊物
- 参考文献

教师介绍

- 主讲教师：刘群
  - 中国科学院计算技术研究所 研究员
  - 办公电话： 010-62600552
  - 办公地点：计算所 552 室
  - 电子邮件： liuqun at ict.ac.cn
  - 个人主页： http://mtgroup.ict.ac.cn/~liuqun

    课件可在个人主页下载（教学工作⇨机器翻译）
- 课代表：腾志扬
  - 手机： 15110082354
  - 邮件： tengzhiyang@ict.ac.cn

课程邮件列表

- 邮件列表地址：

  mt-course-at-ict-2011@googlegroups.com
- 邮件列表网址：

  http://groups.google.com/group/mt-course-at-ict-2011
- 选课或旁听的同学同学都可以访问邮件列表网站并申请加入邮件列表，申请时请注明你的姓名和单位
- 申请邮件列表需要先注册一个 Google Account，最好是用 Gmail 信箱注册，也可以用非 Gmail 信箱注册，但有可能会丢失邮件
- 所有课程通知、作业相关资料都通过邮件列表发放

时间安排 (1)

周	月	日	一	二	三	四	五	六
1	十月	9	10	11	12	13	14	15
2		16	17	18	19	20	21	22
3		23	24	25	26	27	28	29
4	十月 / 十一月	30	31	1	2	3	4	5
5	十一月	6	7	8	9	10	11	12
6		13	14	15	16	17	18	19
7		20	21	22	23	24	25	26
8	十一月 / 十二月	27	28	29	30	1	2	3
9	十二月	4	5	6	7	8	9	10
10		11	12	13	14	15	16	17
11		18	19	20	21	22	23	24
12		25	26	27	28	29	30	31

授课或讨论　　可能需调课　　考试

时间安排 (2)

- 上课时间（12 次共 48 学时）：
  - 每周四晚上： 18:30-21:30
  - 课堂讲授： 7 次
  - 项目报告 + 专题讲座： 4 次
- 考试时间（1 次 2 小时）：
  - 12 月 16 日晚上： 18:30-20:30

课程内容

讲课 28 学时（每一次课 4 学时）

- 第一讲：机器翻译概述
- 第二讲：词法分析技术
- 第三讲：句法分析技术
- 第四讲：基于规则和基于实例的机器翻译方法
- 第五讲：基于词的统计机器翻译方法
- 第六讲：基于短语的统计机器翻译方法
- 第七讲：基于句法的统计机器翻译方法

课堂讨论 16 学时。

答疑 2 学时，考试 2 学时。

作业安排——项目 (1)

- 目标：
  - 利用开源的统计机器翻译工具 Moses （摩西），实现一个英汉机器翻译系统
  - 自己实现一个统计机器翻译解码器
- 资源：
  - 开源的统计机器翻译系统 Moses （摩西）
  - 开源的汉语词法分析系统 ICTCLAS
  - 开源的语言模型工具 SRILM
  - 英汉双语语料库

作业安排——项目 (2)

- 第一阶段：
  - 能够掌握 Moses，并在给定的数据上跑完完整的统计机器翻译训练和解码流程
- 第二阶段：
  - 在 Moses 的基础上，自己实现一个解码器，要求达到跟 Moses 接近的 BLEU 值
- 完成方式：每人独立完成
- 考核：
  - 在线提交测试结果并得到 BLEU 值
  - 提交完整的实验报告
  - 在课堂上做演讲和展示

开源统计机器翻译系统“摩西”简介

- http://www.statmt.org/moses
- 目前最有影响的开源统计机器翻译系统
- 代码经过大量优化，性能很高，已成为这一领域研究最主要的 Baseline 系统
- 开发单位：英国爱丁堡大学、德国亚琛工业大学、意大利 ITC-IRST 研究所、美国卡内基梅隆大学、美国麻省理工学院、捷克查尔斯大学在美国和欧盟的一些资助下完成
- 基本特点：
  - 基于短语的统计机器翻译方法（ Phrase-based Approach ）和基于层次短语的统计机器翻译方法（ Hierarchical Phrase-based ）
  - 基于混合网络的解码（ Confusion Network Decoding ）
  - 基于要素的翻译模型（ Factored Translation Model ）
  - 柱搜索算法（ Beam Search Algorithm ）

网络资源

- ACL主页（ACL Anthology）
- Machine Translation Archive
- LDC（Language Data Consortium）
- ChineseLDC
- 中文自然语言处理开放平台
- 中科院计算所自然语言处理研究组
- 北京大学计算语言学研究所

国际会议

- ACL (NAACL, EACL, AFNLP)
- EMNLP
- COLING
- MT Summit (AMTA, EAMT)
- JSCL （全国计算语言学联合学术会议）
- CWMT （统计机器翻译研讨会）
- 相关领域会议：计算机、人工智能、互联网、语音

学术刊物

- Computational Linguistics
- ACM Translations on Asian Language & Information Processing (ACM TALIP)
- Machine Translation
- 中文信息学报
- 相关领域刊物：
  - 计算机
  - 人工智能
  - 语音
  - 互联网

参考书目

冯志伟（1995）《自然语言机器翻译新论》，语文出版社 1995 年版

翁富良、王野翊（1998）《计算语言学导论》，中国社会科学

陈小荷（2000）《现代汉语自动分析》，北京语言文化大学出版社

赵铁军（2000）《机器翻译原理》，哈尔滨工业大学出版社

杨沐昀（2000）《机器翻译系统》，哈尔滨工业大学出版社

姚天顺等（2002）《自然语言理解 —— 一种让机器懂得人类语言的研究（第二版）》，清华大学出版社、广西科学技术出版社

俞士汶 主编（2003）《计算语言学概论》，商务印书馆

冯志伟（2005）《机器翻译研究》，中国对外翻译出版公司

宗成庆（2008）《统计自然语言处理》，清华大学出版社

刘群（2008）《汉英机器翻译若干关键技术研究》，清华大学出版社

参考书目

James Allen (1995), Natural Language Understanding (Second Edition), The Benjamin / Cummings Publishing Company, Inc. , 中译本：刘群等译，自然语言理解（第二版），电子工业出版社， 2005

Christopher D. Manning and Hinrich Schutze (1999), Foundations of Statistical Natural Language Processing, The MIT Press, Cambridge, Massachusetts , 中译本：苑春法等译，统计自然语言处理基础，电子工业出版社， 2005

Daniel Jurafsky, James H. Martin, Speech and Language Processing: an Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Prentice Hall, US Ed edition, January 26, 2000, 中译本：冯志伟，孙乐译，自然语言处理综论，电子工业出版社， 2005

Philipp Koehn, Statistical Machine Translation, textbook, Cambridge University Press, August 2009

参考文献

- Top 10 SMT papers (2007 By Yang Liu)
- Bibliography for SMT (2007 By Yang Liu)
- 其他
  - A highly selective MT bibliography (1996 by Adam Berger)
  - Bibliography for Machine Translation Evaluation (2003 by Florence Reeder et al.)
  - Bibliography for Statistical Alignment and Machine Translation (2003 by Adrià de Gispert & Patrik Lambert)
  - Bibliography for Statistical Machine Translation (2003 by Kevin Knight)

内容提要

- 机器翻译定义
- 课程目的与特点
- 课程概况
- **机器翻译历史**
- 机器翻译现状
- 机器翻译应用系统类型

机器翻译的历史

- W. J. Hutchens, latest Development in MT Technology: Beginning a New Era in MT Research. In : Proceedings of Machine Translation Summit-IV, Kobe, Japan, 1993
- 冯志伟，自动翻译，上海知识出版社， 1987 年
- 冯志伟，自然语言机器翻译新论，语文出版社，1994 年
- 冯志伟，自然语言的计算机处理，上海外语教育出版社， 1996 年

以下有关机器翻译历史的资料大部分取材于冯志伟先生的相关著作，特此向冯志伟先生表示感谢！

机器翻译的萌芽期 (1)

- 关于用机器来进行语言翻译的想法，远在古希腊时代就有人提出过了。
- 在 17 世纪，一些有识之士提出了采用机器词典来克服语言障碍的想法。
- 笛卡儿（Descartes）和莱布尼兹（Leibniz）都试图在统一的数字代码的基础上来编写词典。
  在 17 世纪中叶，贝克（Cave Beck）、基尔施（Athanasius Kircher）和贝希尔（Johann Joachim Becher）等人都出版过这类的词典。由此开展了关于“普遍语言”的运动。
- 维尔金斯（John Wilkins） 在《关于真实符号和哲学语言的论文》(An Essay towards a Real Character and Philosophical Language, 1668) 中提出的中介语（Interlingua）是这方面最著名的成果，这种中介语的设计试图将世界上所有的概念和实体都加以分类和编码，有规则地列出并描述所有的概念和实体，并根据它们各自的特点和性质，给予不同的记号和名称。

机器翻译的萌芽期 (2)

- 1930 年代之初，亚美尼亚裔的法国工程师阿尔楚尼（ G.B. Artsouni ）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“机械脑”（ mechanical brain ）。
- 这种机械脑的存储装置可以容纳数千个字元，通过键盘后面的宽纸带，进行资料的检索。阿尔楚尼认为它可以应用来记录火车时刻表和银行的帐户，尤其适合于作机器词典。在宽纸带上面，每一行记录了源语言的一个词项以及这个词项在多种目标语言中的对应词项，在另外一条纸带上对应的每个词项处，记录着相应的代码，这些代码以打孔来表示。机械脑于 1937 年正式展出，引起了法国邮政、电信部门的兴趣。但是，由于不久爆发了第二次世界大战，阿尔楚尼的机械脑无法安装使用。

机器翻译的萌芽期 (3)

- 1903 年，古图拉特 (Couturat) 和洛 (Leau) 在《通用语言的历史》一书中指出，德国学者里格 (W. Rieger) 曾经提出过一种数字语法 (Zifferngrammatik)，这种语法加上词典的辅助，可以利用机械将一种语言翻译成其他多种语言，首次使用了“机器翻译” （德文是 ein mechanisches Uebersetzen) 这个术语。
- 1933 年，苏联发明家特洛扬斯基（П．П．ТРОЯНСКИЙ）设计了用机械方法把一种语言翻译为另一种语言的机器，并在同年 9 月 5 日登记了他的发明。1939 年，特洛扬斯基在他的翻译机上增加了一个用“光元素”操作的存储装置； 1941 年 5 月，这部实验性的翻译机已经可以运作； 1948 年，他计划在此基础上研制一部“电子机械机”(electro-mechanical machine)。但是，由于当时苏联的科学家和语言学家对此反映十分冷淡，特洛扬斯基的翻译机没有得到支持，最后以失败告终了。

机器翻译的草创期 (1)

- 1946 年，美国宾夕法尼亚大学的埃克特（J. P. Eckert）和莫希莱（J.W.Mauchly）设计并制造出了世界上第一台电子计算机 ENIAC，在电子计算机问世的同一年，英国工程师布斯（A. D. Booth）和美国洛克菲勒基金会副总裁韦弗（W. Weaver）在讨论电子计算机的应用范围时，就提出了利用计算机进行语言自动翻译的想法。
- 1947 年 3 月 6 日，布斯与韦弗在纽约的洛克菲勒中心会面，韦弗提出，“如果将计算机用在非数值计算方面，是比较有希望的”。
- 在韦弗与布斯会面之前，韦弗在 1947 年 3 月 4 日给控制论学者维纳(N. Wiener）写信，讨论了机器翻译的问题，韦弗说：“我怀疑是否真的建造不出一部能够作翻译的计算机？即使只能翻译科学性的文章（在语义上问题较少），或是翻译出来的结果不怎么优雅（但能够理解），对我而言都值得一试。”可是，维纳在 4 月 30 日给韦弗的回信中写道：“老实说，恐怕每一种语言的词汇，范围都相当模糊；而其中表示的感情和言外之意，要以类似机器翻译的方法来处理，恐怕不是很乐观的。”

机器翻译的草创期 (2)

- 1949 年，韦弗发表了一份以《翻译》为题的备忘录，正式提出了机器翻译问题。在这份备忘录中，他除了提出各种语言都有许多共同的特征这一论点之外，还有两点值得我们注意：
  第一，他认为翻译类似于解读密码的过程。他说：“当我阅读一篇用俄语写的文章的时候，我可以说，这篇文章实际上是用英语写的，只不过它是用另外一种奇怪的符号编了码而已，当我在阅读时，我是在进行解码。”
  第二，他认为原文与译文“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。
- 由于学者的热心倡导，实业界的大力支持，美国的机器翻译研究一时兴盛起来。1954 年，美国乔治敦大学在国际商用机器公司（IBM 公司）的协同下，用 IBM-701 计算机，进行了世界上第一次机器翻译试验，把几个简单的俄语句子翻译成英语，接着，苏联、英国、日本也进行了机器翻译试验，机器翻译出现热潮。

机器翻译的萧条期

- 1964 年，美国科学院成立语言自动处理谘询委员会（ Automatic Language Processing Advisory Committee，简称 ALPAC 委员会），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，对机器翻译采取否定的态度，报告宣称：“在目前给机器翻译以大力支持还没有多少理由”；报告还指出，机器翻译研究遇到了难以克服的“语义障碍”（ semantic barrier ）。
- 在 ALPAC 报告的影响下，许多国家的机器翻译研究低潮，许多已经建立起来的机器翻译研究单位遇到了行政上和经费上的困难，在世界范围内，机器翻译的热潮突然消失了，出现了空前萧条的局面。

机器翻译的复苏期 (1)

- 尽管在萧条时期，法国、日本加拿大等国，仍然坚持着机器翻译研究，于是，在七十年代初期，机器翻译又出现了复苏的局面。
- 在这个复苏期，研究者们普遍认识到，源语和译语两种语言的差异，不仅只表现在词汇的不同上，而且，还表现在句法结构的不同上，为了得到可读性强的译文，必须在自动句法分析上多下功夫。

机器翻译的复苏期 (2)

- 这一时期的机器翻译系统开始采用的分析、转换、生成三个阶段的做法
- 思想提出：早在 1957 年，美国学者英格维（ V. Yingve ）在《句法翻译的框架》（ Framework for syntactic translation ）一文中就指出，一个好的机器翻译系统，应该分别地对源语和译语都作出恰如其分的描写，这样的描写应该互不影响，相对独立。英格维主张，机器翻译可以分为三个阶段来进行。
  - 第一阶段：用代码化的结构标志来表示源语文句的结构；
  - 第二阶段：把源语的结构标志转换为译语的结构标志；
  - 第三阶段：构成译语的输出文句。

机器翻译的复苏期 (3)

- 这个时期机器翻译的另一个特点是语法（grammar）与算法（algorithm）分开。
- 思想提出：早在 1957 年，英格维就提出了把语法与“机制”（mechanism）分开的思想。英格维所说的“机制”，实质上就是算法。所谓语法与算法分开，就是要把语言分析和程序设计分开，程序设计工作者提出规则描述的方法，而语言学工作者使用这种方法来描述语言的规则。语法和算法分开，是机器翻译技术的一大进步，它非常有利于程序设计工作者与语言工作者的分工合作。

机器翻译的复苏期 (4)

- 这个复苏期的机器翻译系统的典型代表是法国格勒诺布尔理科医科大学应用数学研究所（IMAG）自动翻译中心（CETA）的机器翻译系统。这个自动翻译中心的主任沃古瓦（B. Vauquois）教授明确地提出，一个完整的机器翻译过程可以分为如下六个步骤：

  （1）源语词法分析
  （2）源语句法分析
  （3）源语译语词汇转换
  （4）源语译语结构转换
  （5）译语句法生成
  （6）译语词法生成

  其中，第一、第二步只与源语有关，第五、第六步只与译语有关，只有第三、第四步牵涉到源语和译语二者。

- 这就是机器翻译中的“独立分析－独立生成－相关转换”的方法。他们用这种研制的俄法机器翻译系统，已经接近实用水平。

机器翻译的复苏期 (5)

- 美国斯坦福大学威尔克斯（Y.A. Wilks）提出了“优选语义学”（preference semantics），并在此基础上设计了英法机器翻译系统。
- 这个系统特别强调在源语和译语生成阶段，都要把语义问题放在第一位，英语的输入文句首先被转换成某种一般化的通用的语义表示，然后再由这种语义表示生成法语译文输出。
- 由于这个系统的语义表示方法比较细致，能够解决仅用句法分析方法难于解决的歧义、代词所指等困难问题，译文质量较高。

机器翻译的繁荣期

- 1970 年代末，机器翻译进入了它的第三个时期 -- 繁荣期（ 1976 年— 1980 年代末）。
- 繁荣期的最重要的特点，是机器翻译研究走向了实用化，出现了一大批实用化的机器翻译系统，机器翻译产品开始进入市场，变成了商品，由机器翻译系统的实用化引起了机器翻译系统的商品化。

机器翻译的平台期 (1)

- 整个 1990 年代，机器翻译进入了一个平台期
- 基于规则的机器翻译方法理论上无法突破
- 在应用上，机器翻译遇到问题，在面对大规模真实文本时翻译质量难以满足用户需求，反而是基于翻译记忆思想的计算机辅助翻译获得了巨大进展

机器翻译的平台期 (2)

- 就在机器翻译进入平台期的时候，一些新的因素也在萌芽
  - 基于实例的机器翻译思想
  - 基于统计的机器翻译思想
  - 互联网的出现大大促进了机器翻译的需求

统计机器翻译的新热潮 (1)

- 1999 年开始，出现了一个机器翻译的新热潮，其最主要的特征是统计机器翻译方法开始占据主导地位，机器翻译的质量出现了一个跨越式的提高

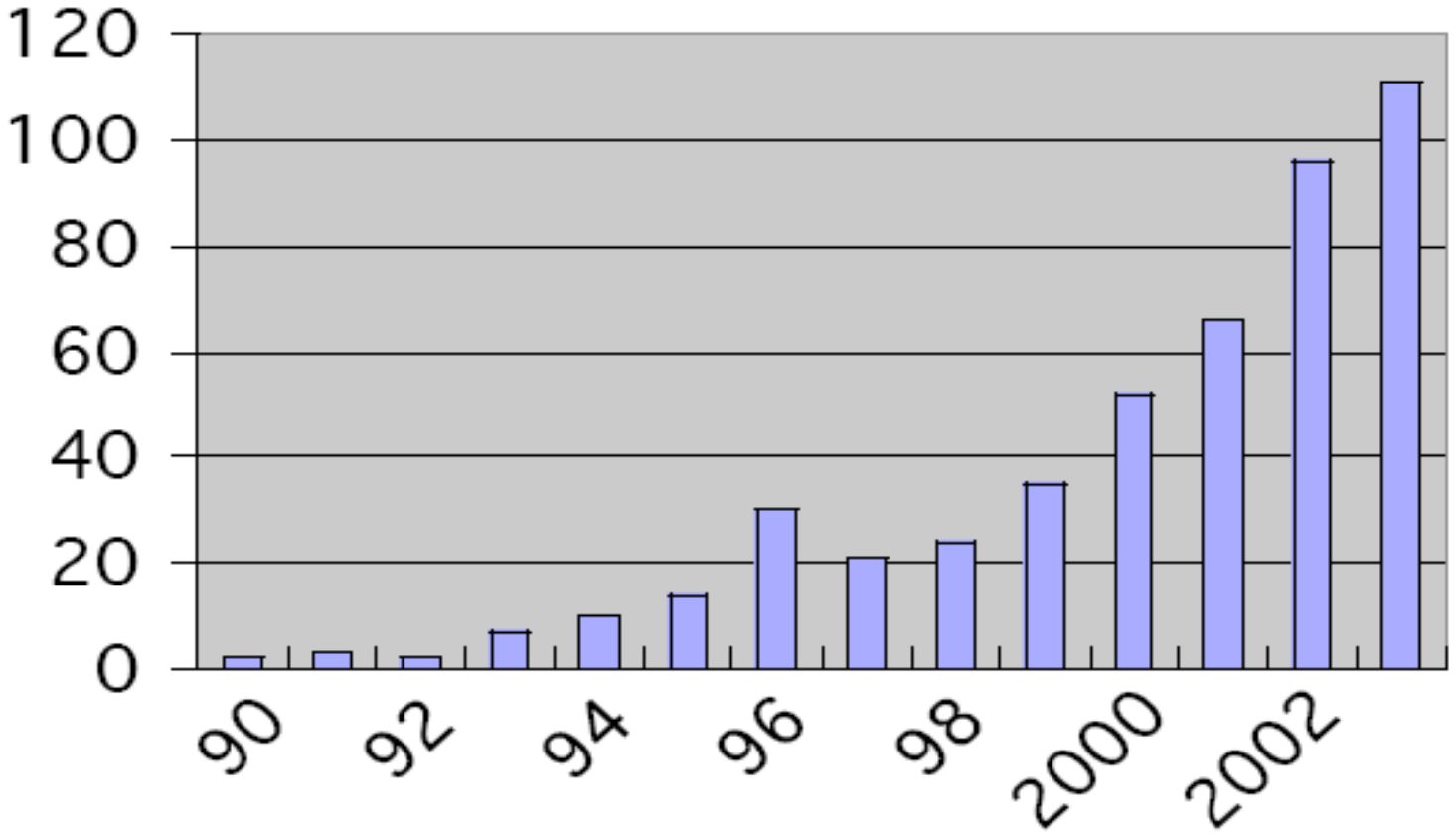
统计机器翻译的新热潮 (2)

- 1980 年代末 IBM 首次开展统计机器翻译研究
- 1992 年 IBM 首次提出统计机器翻译的信源信道模型
- 1993 年 IBM 提出五种基于词的统计翻译模型 IBM Model 1-5
- 1994 年 IBM 发表论文给出了 Candide 系统与 Systran 系统在 ARPA 评测中的对比测试报告
- 1999 年 JHU 夏季研讨班重复了 IBM 的工作并推出了开放源代码的工具
- 2001 年 IBM 提出了机器翻译自动评测方法 BLEU
- 2002 年 NIST 开始举行每年一度的机器翻译评测
- 2002 年第一个采用统计机器翻译方法的商业公司 Language Weaver 成立

统计机器翻译的新热潮 (3)

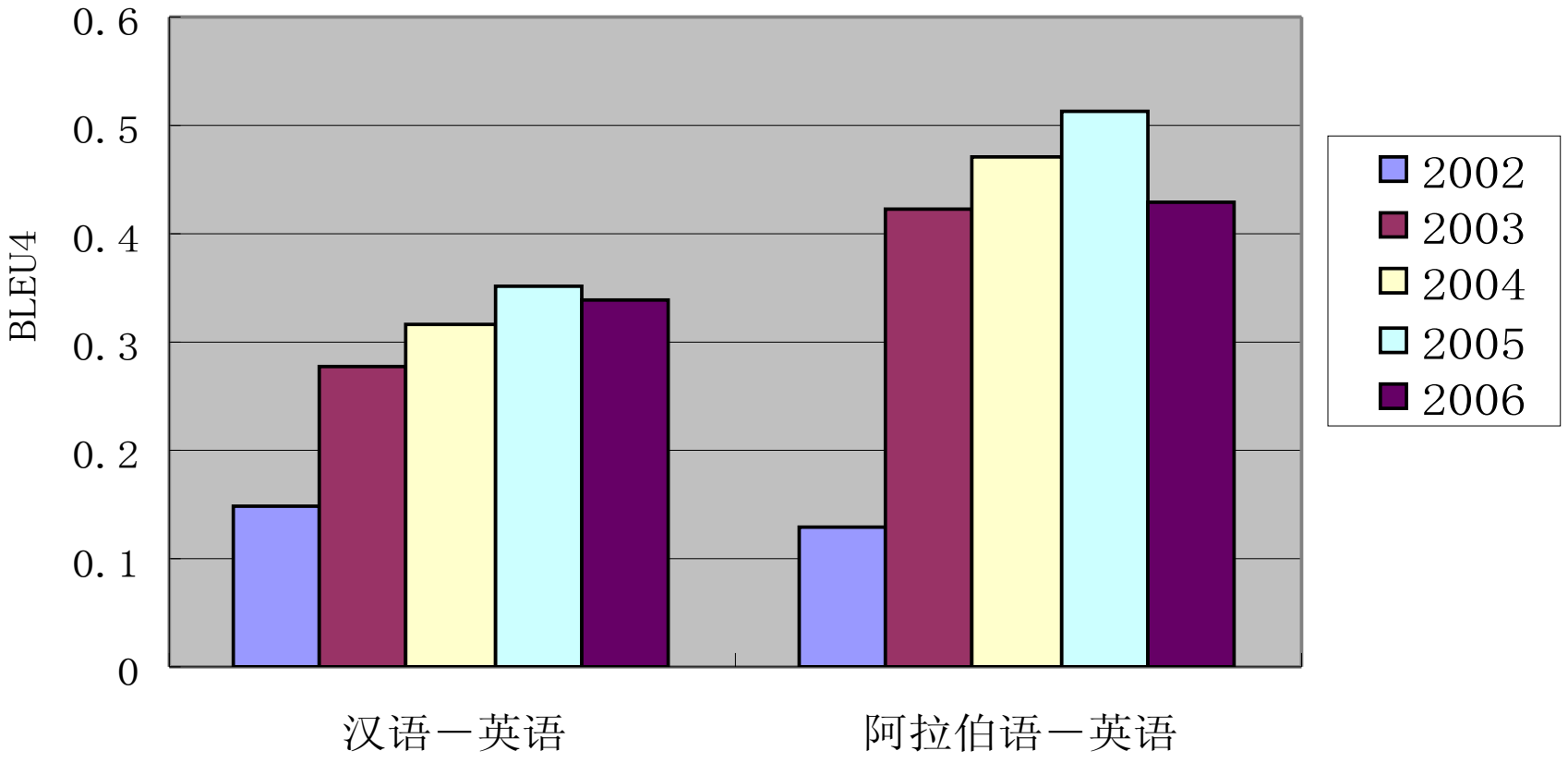
- 2002 年 Franz Josef Och 提出统计机器翻译的对数线性模型
- 2003 年 Franz Josef Och 提出对数线性模型的最小错误率训练方法
- 2004 年 Philipp Koehn 推出 Pharaoh （法老）标志着基于短语的统计翻译方法趋于成熟
- 2005 年 David Chiang 提出层次短语模型并代表 UMD 在 NIST 评测中取得好成绩
- 2005 年 Google 在 NIST 评测中大获全胜，随后 Google 推出基于统计方法的在线翻译工具，其阿拉伯语 - 英语的翻译达到了用户完全可接受的水平，目前已经可以支持 40 多种语言的互译
- 2006 年 NIST 评测中 USC-ISI 的串到树句法模型第一次超过 Google （仅在汉英受限翻译项目中）

近年来统计机器翻译论文发表数量



引自 Franz Josef Och, Statistical Machine Translation: Foundations and Recent Advances, Tutorials on MT Summit X, September 13-15, 2005, Phuket, Thailand

近年来国际 NIST 评测最好成绩



中国机器翻译的历史 (1)

- 黄果，再建巴比塔——冯志伟研究员谈我国机器翻译发展历程，计算机世界报， 1999 年 9 月 27 日
- 董振东，中国机器翻译的世纪回顾，《中国计算机世界》 2000 第一期， 2000/01/03

（这一部分讲义部分引用上述两篇文章内容，特此感谢！）

中国机器翻译的历史 (2)

- 第 41 项：计算技术的建立

  本任务必电子计算机的设计制造与运用为主要内容。一、二年内，首先着重于快速通用数字电子计算机的设计与制造，从中掌握各种电子计算机的基本技术与运用方法，以建立计算技术的基础。二、三年内，开始掌握专用电子计算机的设计与制造，进而根据需要研究制造各种专用计算机。**关于利用电子计算机进行自动翻译的工作，首先由语言学家与数学家协同研究翻译中字汇范围和文句结构，并编制运算程序，然后进行实际操作的研究。**此外，有关计算机技术的数学问题，如程序设计与近似计算方法等，也包括在本任务之内（关于模拟计算机以及穿孔式及检式计算机的制造问题，已列入第 54 项任务内）。

以上摘自《一九五六——一九六七年科学技术发展远景规划纲要》（又称十二年科技规划纲要）

中国机器翻译的历史 (3)

- 1959 年建国十周年之际， 1959 年，我国第一个机器翻译系统诞生了，这就是中国科学院语言研究所与计算技术研究所合作的《俄译汉的翻译系统》，标志着我国继美国，苏联，英国和日本之后，成为世界上第五个机器翻译实验成功的国家。
- 这个系统还很不成熟，当时只做了 9 个不同类型的句子翻译的实验，所采用的机器是我国早期自行研制的 104 机，当时还没有解决计算机的汉字信息处理问题，外围设备也差，无法在屏幕上显示翻译好的汉字，只能输出穿孔纸带，外行人根本看不懂，所以谈不上什么实用性；但是，这个成果证明了用机器帮助人把外文翻译成中文是可行的，因此，当时的《科学通报》发表了这个有重要学术价值的成果。

中国机器翻译的历史 (4)

- 文革后期的 748 工程，对机器翻译重新给予重视。1975 年，成立了由情报所、语言所、计算所、冶金部、林业部、核工业部、化工研究院、中国医学科学院等单位参加的全国机器翻译协作研究组，以冶金题录 5000 条为试验材料，制定英汉机器翻译方案并上机试验。 1976 年粉碎“四人帮”后，机器翻译研究全面复苏。 1978 年，在计算所 111 机上进行《英汉冶金题录机器翻译系统》的抽样试验，抽样 20 条，达到了预期的效果。当时用机器翻译系统翻译整篇文章还比较困难，因此这个机器翻译系统主要用于翻译外文资料的题目。

中国机器翻译的历史 (5)

- 军事科学院曾经开发出“科译一号 (KY-1)”实用型全文与题录兼容的英汉机器翻译系统，在富士通中型机上运行，由 COBOL 语言编写，原型系统运行效果良好，译文质量较高。它获得了国家科技进步二等奖。（董振东）
- 1986 年，中软公司发现了这个机器翻译系统，就把它买下来，用 C 语言改编成在 PC 上运行的机器翻译系统，并且把它商品化，取名叫做“译星”。译星的诞生有着特殊的历史意义，首先，它是我国第一个运行在 PC 机上的机器翻译软件；其次，它也是我国第一个商品化的机器翻译系统。（董振东）

中国机器翻译的历史 (6)

- 邮电科研院研制的 "MT-IR-EC"，这是一个非常实用的通讯题录系统， 人们利用它翻译出版通讯题录刊物，从而使刊物的发行效率得到很大的提高，它因此成为了第一个荣获国家科技进步奖的机译系统。
- 1980 年代中后期，中国参加了由 日本发起的亚洲五国机器翻译研发的合作项目。国内近 10 个单位参加了这一长达 7 年的国际项目。这次的大协作对于培养人才、传播技术、积累资源 (如词典等)， 以及使中国的机译研究走向世界，都有着深远的影响。另外，这个时期又正值“七五”，它给了更多的单位和研究人员参与机器翻译研究的机会。（董振东、黄昌宁、俞士汶、姚天顺、袁琦……）

中国机器翻译的历史 (7)

- 1990 年代初，高立公司与社科院语言研究所联合开发的高立机器翻译系统，这是在 50 年代的工作基础上，按照比较严格的语言学规则进行开发。（刘倬）
- 1980 年代后期到 1990 年代中期，中科院计算所在 863 项目支持下研制了“智能型机器翻译系统”，该系统授权给香港“权智”公司并合作开发了“快译通 863”系列产品，在市场上取得了巨大的成功，带来了十分可观的效益。该系统获得了国家科技进步一等奖。后成立华建集团应用推广该项技术。（陈肇雄、黄河燕）

中国机器翻译的历史 (8)

- 桑夏公司在“ 863”计划支持下研究的开发出“光翻译系统”（ Light ）。以该系统为引擎建立的“看世界 (ReadWorld)”网站是最早的可以提供网上全文翻译的网站。（史晓东）
- 工信部（原电子部、信产部）所属的北京赛迪翻译技术有限公司开发的赛迪机器翻译系统从 1980 年代起就一直开展机器翻译研究和应用推广，也取得了较好的成绩。（袁琦、孙广泛）

中国机器翻译的历史 (9)

- 国家 863 计划专家组在 1990 年代年举行多次全国性的中文信息处理技术评测，其中包括 1994 、 1995 、 1998 三次机器翻译评测，大大推动了我国机器翻译研究的进展。（钱跃良，俞士汶）
- 863 计划专家组于 2003-2005 年委托中科院计算所恢复并连续举办了三个年度了中文信息处理技术评测，包括机器翻译评测，评测中开始采用国际上通用的一些自动评测指标。（钱跃良，刘群）

中国机器翻译的历史 (10)

- 2004 年以后，中科院计算所、自动化所、厦门大学等单位开始从事统计机器翻译研究工作，并于 2005 年在厦门大学联合举办了第一次“全国统计机器翻译研讨会”，以后该研讨会每年举办一次，并改名“全国机器翻译研讨会”。（徐波，史晓东，刘群，宗成庆）
- 2006 年第二届统计机器翻译研讨会上，中科院计算所、自动化所、软件所、厦门大学、哈尔滨工业大学等五个单位联合推出了开放源代码的统计机器翻译系统“丝路”。（刘群，宗成庆，史晓东，赵铁军，孙乐）

中国机器翻译的历史 (11)

- 2007 年第三届统计机器翻译研讨会开始，每次研讨会前都举办公开的机器翻译评测，并在会上就评测中的技术进行交流。（刘群，赵红梅）
- 中国的研究机构在国际机器翻译评测中表现不俗。中科院计算所在竞争最激烈的 NIST 机器翻译评测中获得过第 3 名（ 2009 年汉英项目总成绩），中科院自动化所、东芝中国研究中心、中科院计算所在国际口语机器翻译评测 IWSLT 多次获得第 1 名。（刘群、王海峰、宗成庆）

中国机器翻译的历史 (12)

- 中科院计算所提出的基于句法的树到串系列统计机器翻译模型，连续多年在国际自然语言处理最重要的 ACL 、 EMNLP 、 COLING 等学术会议上发表了一系列论文，引起广泛的关注和引用。（刘群、吕雅娟、刘洋、熊德意、何中军、米海涛）
- 东芝中国研究开发中心、微软亚洲研究院、中科院自动化所在 ACL 、 EMNLP 、 COLING 等重要学术会议上也发表了很多统计机器翻译相关的研究论文，涉及领域包括机器翻译领域自适应、基于桥接语言的机器翻译、基于句法的机器翻译、机器翻译的系统融合等（王海峰、吴华、周明、李沐、李志灏、张冬冬、宗成庆）

内容提要

- 机器翻译定义
- 课程目的与特点
- 课程概况
- 机器翻译历史
- **机器翻译现状**
- 机器翻译应用系统类型

机器翻译现状

- Babelfish (Powered by Systran) (点击)
- Google Translate (点击)
- Bing Translate (点击)
- 百度翻译(点击)
- 有道翻译(点击)
- 华建翻译中心 (点击)
- 中科院计算所机器翻译在线演示（点击）

从当日新闻中（新浪和 CNN ）选择中英文句子各两个，送到上述网站进行翻译，并对翻译的结果进行分析比较

内容提要

- 机器翻译定义
- 课程目的与特点
- 课程概况
- 机器翻译历史
- 机器翻译现状
- **机器翻译应用系统类型**

机器翻译应用系统类型 (1)

- 理想的机器翻译
  - 全自动高质量， FAHQ MT
    Full Automatic High Quality Machine Translation
- 按人机关系划分
  - 全自动机器翻译， FAMT
    Full Automatic Machine Translation
  - 人助机译， HAMT
    Human Assisted Machine Translation
  - 机助人译， CAT
    Compute-Aided Translation

机器翻译应用系统类型 (2)

- 按应用方式划分
  - 信息分发型 MT for dissemination
    - 要求高质量，不要求实时
    - 采用人机互助，或者受限领域、受限语言等方式提高翻译质量
  - 信息吸收型 MT for assimilation
    - 不要求高质量，要求方便、实时
    - 翻译浏览器、便携式翻译设备、……

机器翻译应用系统类型 (3)

- 按应用方式划分（续）
  - 信息交流型 MT for interchange
    - 不要求高质量，通常要求实时，语言随意性较大
    - 语音翻译、网络聊天翻译、电子邮件翻译
  - 信息存取型 MT for information access
    - 将机器翻译嵌入到其他应用系统中
    - 跨语言检索、跨语言信息抽取、跨语言文摘、跨语言非文本数据库的检索……

思考题

- 形式语言的翻译（如编译）和自然语言的翻译有何异同？
- 在机器翻译中，对源语言的分析是否越深越好？为什么？
- 你认为英汉机器翻译和汉英机器翻译的难点有何不同之处？
- 如何自动评价一个机器翻译系统译文的质量？