计算语言学概论

常宝宝
北京大学计算语言学研究所
chbb@pku.edu.cn

课程信息

- 名称: 计算语言学
- 授课时间: 周四 9-11节(16:50~19:40？)
- 授课地点: 一教 104
- 助教

陈亮

chenlianglucky@pku.edu.cn

成绩评定

- 平时成绩(60%)
  - 上机作业
  - 出勤
- 期末笔试(40%)

主要参考书

1. 计算语言学概论，俞士汶主编，商务印书馆，2003
2. Speech and Language Processing, Jurafsky, D. & Martin, J.H., Prentice Hall, 2000(中译本：自然语言处理综论，冯志伟等译，电子工业出版社，2005)

其它参考书(一)

1. Foundations of Statistical Natural Language Processing, Manning,C.D. & Schütze,H., The MIT press, 1999 (有中译本)
2. Statistical Language Learning. Charniak, E., The MIT Press. 1996.
3. Natural Language Understanding, Allen, J., The Benjamins/Cummins Publishing Co., 1994 (有中译本)
4. Natural Language Processing: An Introduction to Computational Linguistics, Gazdar, G. & Mellish, C., Addison-Wesley, 1989.
5. Introduction to Natural Language Processing, Harris, M.D., Reston Publishing Co. , 1985

其它参考书(二)

1. 现代汉语语法信息词典详解，俞士汶等，清华大学出版社，2003
2. 自然语言理解，姚天顺，清华大学出版社，2002
3. 自然语言处理技术基础，王晓捷、常宝宝，北京邮电大学出版社，2002
4. 计算语言学，刘颖，清华大学出版社，2002
5. 计算语言学基础，冯志伟，商务印书馆，2001
6. 计算语言学导论，翁富良、王野翊，中国社会科学出版社，1998
7. 自然语言的计算机处理，冯志伟，上海外语教育出版社，1997
8. 自然语言处理，刘开瑛、郭炳炎，科学出版社，1991

相关学术期刊和会议

1. Computational Linguistics (ACL)
2. Machine Translation
3. International Journal of Corpus Linguistics
4. 中文信息学报（中文信息学会）
5. 计算机学报、软件学报
6. 汉语语言与计算学报（新加坡）

1. Annual Meeting of the Association for Computational Linguistics (ACL年会)
2. International Conference on Computational Linguistics (COLING)
3. 全国计算语言学联合学术会议(JSCL)
4. 全国学生计算语言学研讨会(SWCL)

什么是计算语言学?

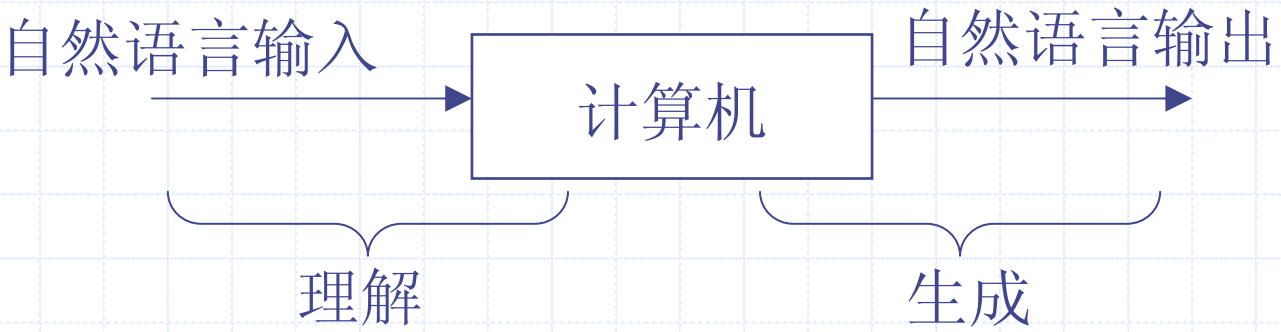
- 计算语言学是通过建立形式化的计算模型来分析、理解和处理自然语言的学科。
- 什么是自然语言?
- 其它术语
  - 自然语言处理(Natural Language Processing)
  - 自然语言理解(Natural Language Understanding)
  - 人类语言技术(Human Language Technology)

什么是计算语言学?

- 计算语言学是一门交叉学科。计算语言学研究需要多个学科的知识。
  - 语言学 （自然语言是处理对象）
  - 计算机科学（计算语言学的研究工具）
  - 数学 （自然语言的建模工具）

为什么要研究计算语言学？

- 语言障碍
  - 人－人之间的语言障碍（自动翻译）
  - 人－机之间的语言障碍（人－机接口）



计算语言学的研究目标

- 终极目标
  - 研制能理解并生成人类语言的计算机系统。
- 当前目标
  - 研制出具有一定人类语言能力的计算机文本或语音处理系统。

计算语言学的研究内容

- 建立形式化的适于计算机处理的语言模型。
- 研制分析、生成以及处理语言的各种算法。

计算语言学研究的挑战性

- 大量的词汇、大量的句子
  - OED收词40万、汉语中有多少词？
- 无法象处理人工语言那样，写出一个完备的、有限的规则系统来进行定义和描述。自然语言的规则很少没有例外。(photo、potato)
- 自然语言中有大量的歧义现象。
- 自然语言的理解不仅和语言本身的规律有关，还和语言之外的知识（例如常识）有关。因此语言处理涉及的常是海量知识，知识库的建造维护代价很高。

计算语言学研究的挑战性

- 什么是歧义？
  - 对同一个语言形式有不止一种解读。
- 歧义是自然语言的固有属性，即使对于人类自身而言，也是如此。（人工语言有歧义吗?）
- 语言单位无论大小都有歧义现象。
- 语言学家常把语言研究区分为不同的层次，例如：音韵学、形态学、句法学、语义学、语用学等，在这些层面歧义都会有所表现。

计算语言学研究的挑战性

- 歧义举例：

(1)The boy saw the girl with a telescope.

→Who has the telescope?

(2) At last, a computer that understands you like your mother

→The computer understands you as well as your mother understands you.

→The computer understands that you like your mother.

→The computer understands you as well as it understands your mother.

常见对策

- 由于歧义等因素的存在，自然语言处理的性能还不能满足一般应用的需要，为了满足某些特殊的应用需求，常采用下面的对策
  - 交互式处理
    - 人机互助进行处理
  - 受限语言
    - 限定处理文本的领域
  - 受控语言
    - 限定语言的词汇和句法，降低复杂度
- 我们在做计算语言学研究时，时刻都要避免贪大求全，应注意限定自己的研究范围。

计算语言学的研究方法

1. 规则驱动的方法
2. 数据驱动的方法
3. 二者融合的方法

计算语言学的研究方法

- 规则驱动的方法（符号主义）
  1. 研究人员（例如语言学家）对语言的规律进行总结，形成规则形式的知识库。
  2. 研制语言处理算法，利用这些规则对自然语言进行处理。
  3. 研究人员根据处理结果，调整规则，改进处理效果。

计算语言学的研究方法

- 规则方法举例

  例如:

  S → NP + VP

  NP → DET + N

  NP → NP + PP

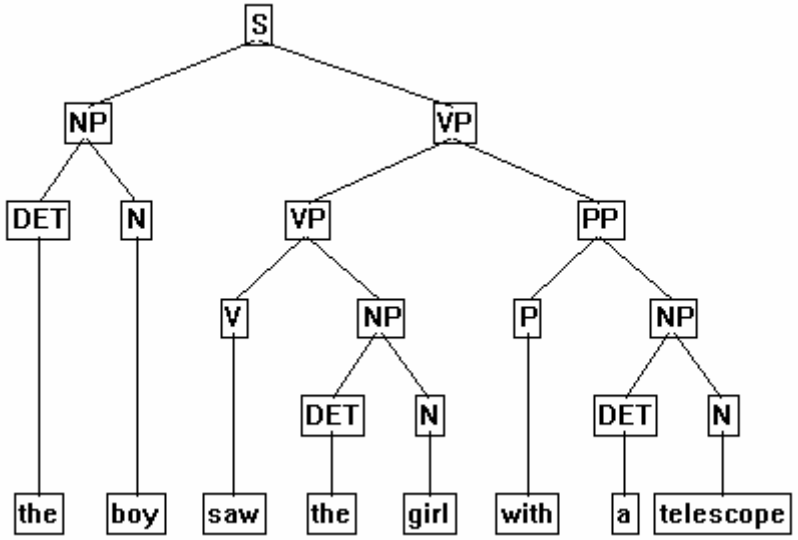
  VP → VP + PP

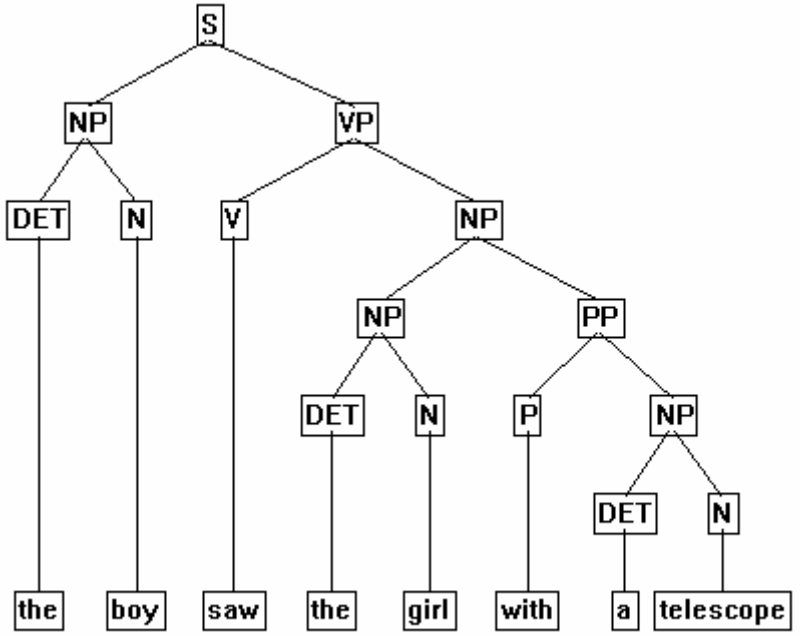
  VP → V + NP

  PP → P + NP

计算语言学的研究方法

- 用上述规则分析句子“the boy saw the girl with a telescope”





计算语言学的研究方法

- All grammar leak (Sapir 1921)

  对于自然语言而言，不大可能写出一部完备的规则集，语言规则有很强的伸缩性。

- 一般而言，很多基于规则的系统不能满足真实语言文本处理的要求，而只能处理真实语言的某个很小的子集。

  toy system?

计算语言学的研究方法

- 数据驱动的方法（统计方法）
  1. 建立可以反映语言使用情况的语料库。
  2. 研究人员对自然语言进行统计建模。
  3. 利用统计技术或机器学习技术，基于语料库训练统计语言模型。
  4. 利用得到的模型设计算法对语言进行处理。
  5. 根据处理效果改进模型，提高处理性能。

计算语言学的研究方法

- 在数据驱动的方法中，语言模型通常体现为一组参数，这些参数通常表示某个语言形式发生的概率值。例如：

  $P(w_3|w_1w_2)$

  P(公鸡|一 只) > P(供给|一 只)

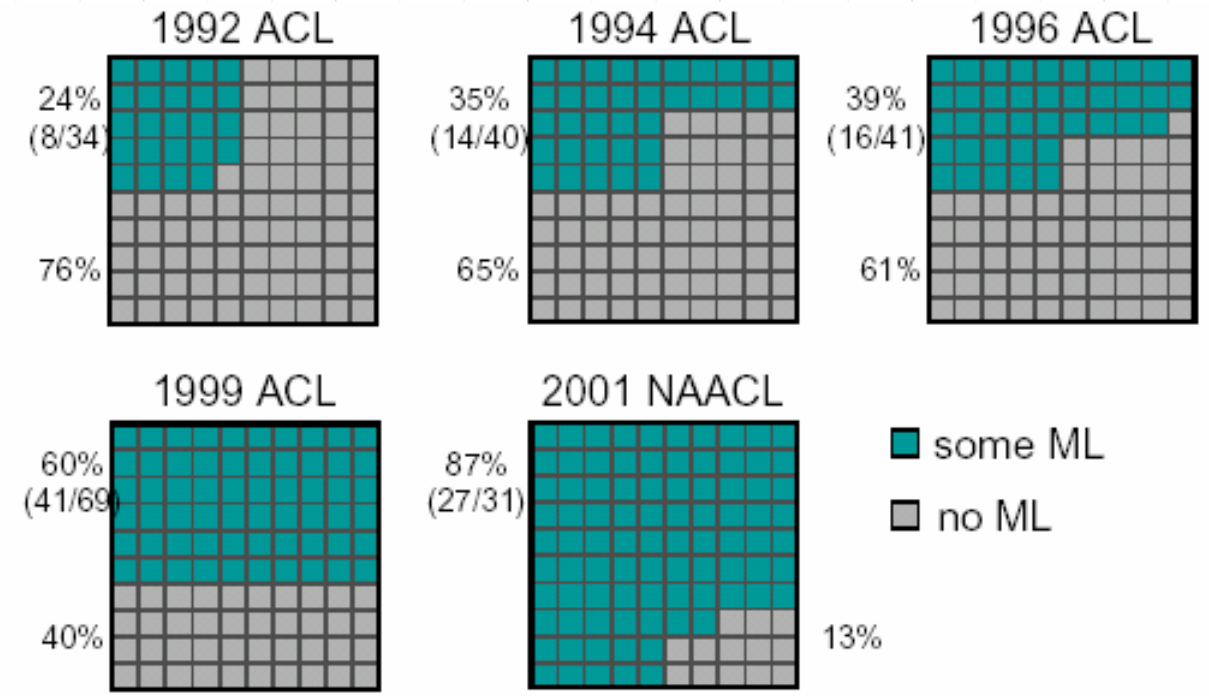
- 数据驱动的方法忽视了语言的深层结构。

计算语言学的研究方法

- 融合规则驱动和数据驱动的方法
  - 规则驱动、数据驱动的优劣不能简单评价
  - 很多研究人员（包括知名计算语言学家）建议如此
  - 已经提出了一些策略，但如何无缝结合尚须进一步探索

计算语言学的研究方法

- 从学术会议看计算语言学的研究方法
  - 机器学习以及统计技术得到了越来越多的重视



计算语言学研究中的评测问题

- 为了评价各种方法的有效性，必须进行客观公正的评测，客观公正的评测有助于引导计算语言学朝着一个健康的方向发展。
- 自然语言很复杂、关于语言处理方法和系统的评测也不容易。
- 黑箱评测 vs 白箱评测
- 能否处理大规模的非受限的语言现象值得作为一个标准。
- 规避语言学争议、制定标准测试集
- 国际、国内都在进行评测方法的研究 也在进行大规模的评测时间工作 863 973 TREC MUC SIGAHN

计算语言学的应用

- 计算语言学有着广阔的应用领域。
  1. 机器翻译
  2. 人机对话
  3. 信息检索
  4. 信息提取
  5. 自动文摘
  6. 文本分类
  7. 拼写检查
  8. 音字转换

机器翻译

- 目标是研制能把一种自然语言翻译成另外一种自然语言的计算机软件系统。
  - 例如 汉英机器翻译系统
- FAHQMT – 尚须时日
- 相关研究始于四十年代末（计算机诞生不久）。
- 机器翻译研究经历了曲折的历程，正是对机器翻译的研究导致了计算语言学的诞生。
- 目前市场上有不少翻译产品，应正确看待。

机器翻译

- 著名的例子
  - → *the spirit is willing but the flesh is weak*.
  - → *the vodaka is good but the meat is rotten.*

- 联机机器翻译网站
  - SYSTRAN http://www.systransoft.com/
  - 华建 http://www.hjtrans.com/

人机对话

- 科幻主题
  - 2001:A space odyssey(2001年太空漫游)
    1968年奥斯卡奖
  - HAL9000
    Dave: Open the pod bay doors, HAL.
    HAL: I'm sorry Dave, I'am afraid I can't do that
    Dave: What's the problem?
    HAL: I think you know what the problem is just as well as I do.

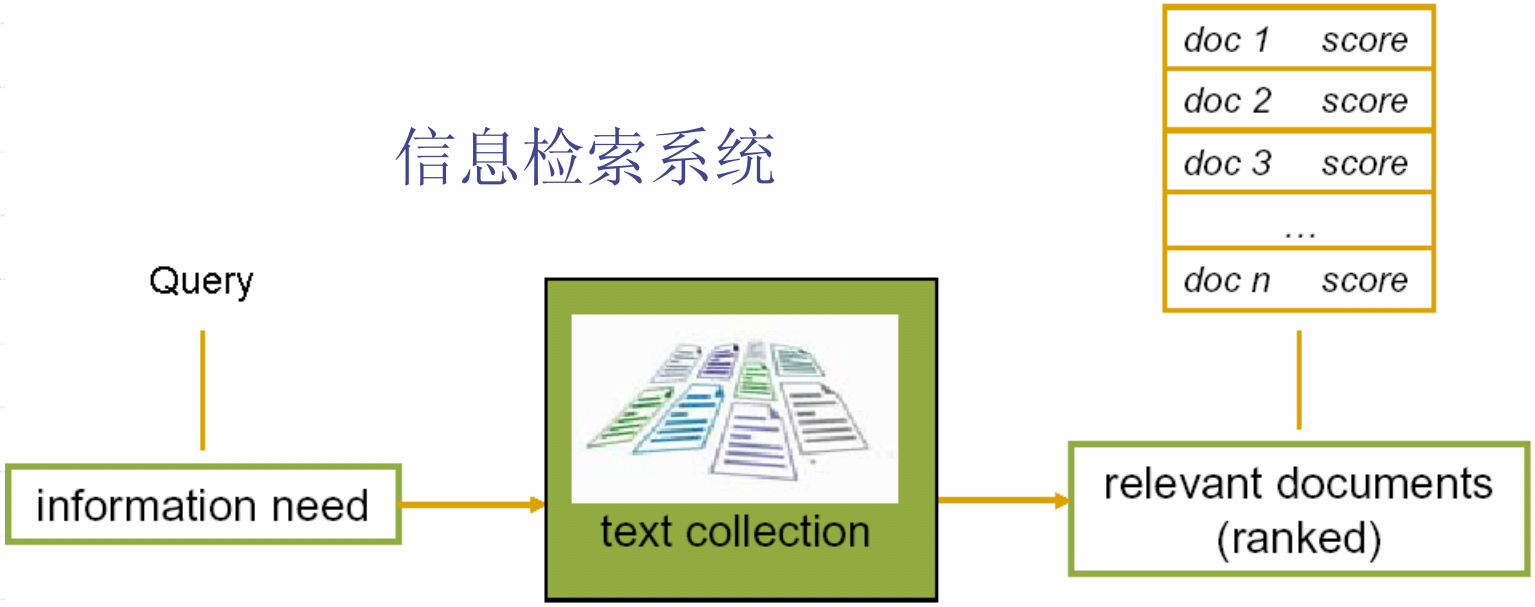
人机对话

- 自然语言接口
- Question answering system (QA系统)
- 例子：
  - **Question:**
    - Who is Ronald Reagan's wife?
  - **Possible answers:** XML TXT
    1. Nancy Davis Reagan (1923-...) is the second wife of Ronald Reagan, who served as president of the United States from 1981 to 1989.
    2. Nancy Reagan, wife of President Ronald Reagan, was born Anne Frances Robbins.
    3. ……

人机对话

- 联机QA系统
  - AnswerBus http://www.answerbus.com/
  - AskJeeves http://web.ask.com/
  - START http://start.csail.mit.edu

信息检索



- Google、百度、天网

自动文摘

About Columbia Newsblaster

Columbia Newsblaster is a system to automatically track the day's news. There are no human editors involved -- everything you see on the main page is generated automatically, drawing on the sources listed on the left side of the screen.

Every night, the system crawls a series of Web sites, downloads articles, groups them together into "clusters" about the same topic, and summarizes each cluster. The end result is a Web page that gives you a sense of what the major stories of the day are, so you don't have to visit the pages of dozens of publications.

- 访问: Columbia Newsblaster
  http://www1.cs.columbia.edu/nlp/newsblaster/
  http://newsblaster.cs.columbia.edu

信息提取

文本数据结构化

BOGOTA, 3 APR 90 (INRAVISION TELEVISION CADENA 1) – [REPORT][JORGE ALONSO] Liberal senator Federico Estrada Velez was kidnapped on 3 April at the corner of 60th and 48th streets in western Medellin, only 100 meters from a metropolitan police CAI [Immediate Attention Center]. The Antioquia department liberal party leader had left his house without any bodyguards only minutes earlier. As he waited for the traffic light to change, three heavily armed men forced him to get out of his car and get into a blue Renault.

Hours later, through anonymous telephone calls to the metropolitan police and to the media, the Extraditables claimed responsibility for the kidnapping. In the calls they announced that they will release the senator with a new message for the national government.

Last week, Federico Estrada Velez has rejected talks between the government and the drug traffickers.

信息提取

模板编号：	1
事件发生时间：	03 APR 90
事件类型：	Kidnapping
肇事人：	“Three heavily armed men”
肇事组织：	“The Extraditables”
受害人：	“Federico Estrada Velez ”
受害人数：	1
受害人类别：	Political Figure
事件发生地点：	Colombia: Medellin(city)

其它应用

- 文本分类（自动判别文本的类别）
- 音字转换（汉字整句输入法）
- 拼写检查和自动勘校系统

计算语言学简史

- 1940年代末—1960年代中期
  - Warren Weaver(49)、GeorgeTown系统(54)
  - Noam Chomsky(57)
  - 统计方法被放弃
- 1966年：ALPAC(66) 语义障碍
- 1970年代中期—1980年代
  - TAUM-METEO(76)
  - SYSTRAN(76)
  - AI繁荣
  - MT产品 如Fujitsi、Hitachi、Siemens

计算语言学简史

- 1980年代—1990年代前期
  - 欧盟 Eurotra 计划(82)
  - 日本Mu系统以及ODA计划(82)
- 1990年代—
  - 统计方法复苏、IBM统计翻译(90)
  - 规则方法、统计方法融合
  - Internet的高速发展为计算语言学发展注入了新的动力