计算语言学

第1讲 概论

刘群
中国科学院计算技术研究所
liuqun@ict.ac.cn

中国科学院研究生院2002~2003学年第二学期课程讲义

问题驱动的学习

要了解一门学科，首先要知道这门学科所要解决的问题。只有了解了一门学科所要解决的问题，才能真正理解一门学科的内在逻辑，才能不仅知其然，而且知其所以然。

在学习一门学科之前，不妨抛开这门学科的所有知识，直接面对这门学科所面对的最基本的问题，想一想如果要你来解决这个问题，你会用什么办法。然后在学习的过程中，不断地用你学到的知识来印证你所要解决的问题，才能深刻地理解你所学到的知识，真正做到融会贯通。

问题之一：自动翻译

- 自动翻译问题
  - 人类最古老的问题之一：《圣经》上巴比伦塔的传说
  - 计算机发明以来，几乎是人类首先想到的计算机的应用之一
  - 困难比人们想像的大得多，至今没有找到理想的解决办法
  - 是推动计算语言学研究的永恒动力
  - 是计算语言学的终极目标
  - 自动翻译研究，几乎触及到计算语言学研究的所有方方面面

问题之一：自动翻译（续）

- 一个经典的例子（早期的英俄翻译系统）：
  - 源文：The spirit is willing , but the flesh is weak.
  - 正确的译文：心有余而力不足
  - 机器的译文：酒是好的，但肉是馊的
- 又一个经典的例子（一个笑话）
  - 源文：How are you?
  - 正确的译文：你好吗？
  - 逐字翻译的译文：怎么是你？
  - 源文：How old are you?
  - 正确的译文：你多大了？
  - 逐字翻译的译文：怎么老是你？

问题之一：自动翻译（续二）

- 演示
  - Light英汉机器翻译系统
  - TransEasy汉英机器翻译系统
- 目前的水平：
  - Systran：http://www.systransoft.com
  - ReadWorld: http://www.readworld.com
  - 华建：http://www.altlan.com

问题之二：自动问答

- 图灵测试
  - 用于判断“机器是否具有智能”
  - 一个封闭的小屋
  - 屋外一个人
  - 屋内依次进入一个人和一台计算机
  - 屋外的人并不知道屋内的是人还是计算机
  - 屋外的人向屋内的人和计算机提出各种问题
  - 屋外的人根据回答来判断屋内是人还是计算机
  - 如果判断不出来，那么可以认为计算机具有了智能

问题之二：自动问答（续·例子）

User:	ELIZA:
Men are all alike.	In what way?
They’re always bugging us about something or other.	Can you think of a specific example?
Well, my boyfriend made me come here.	Your boyfriend made you come here.
He says I’m depressed much of time.	I am sorry to hear you are depressed.
It’s true, I am unhappy.	Do you think coming here will help you not be unhappy?
I need some help; that much seems certain.	What would it mean to you if you got some help?
Perhaps I could learn to get along with my mother.	Tell me more about your family.
My mother takes care of me.	Who else in your family takes care of you?
My father.	Your father.
You are like my father in some ways.	What resemblance do you see?

问题之二：自动问答（续）

- 演示
  - 机器人心理医生Frank（Eliza的变体）
  - Alice
    - 获2000 Loebner Prize
    - 基于AIML，开放源代码，变种众多
- 目前水平
  - AskJeeves
  - TREC的QA Track

其他问题

- 音字转换：语音识别、拼音输入
- 自动文摘：自动给出一篇或多篇文章的摘要
- 信息检索：在海量的信息准确找到你所需要的信息
- 信息过滤：从信息流中筛选出你所感兴趣的信息
- 信息抽取：从海量的信息中抽取出你所需要的（结构化）信息
- ……

问题驱动的学习（续）

- 本课程采用问题驱动的学习方法
- 我们将围绕“机器翻译”这一主要问题，来展开“计算语言学”这门课程的学习
- 兼顾其他问题
- 侧重汉语的处理

计算语言学定义

计算语言学是一门以**计算**为手段对**自然语言**进行**研究**和**处理**的科学。

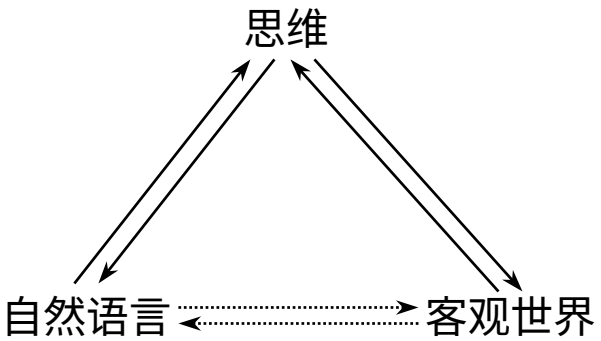
计算语言学的研究手段

- 计算语言学的研究手段是计算
- 计算的基础是冯·诺依曼结构的计算机
- 计算的表现形式是算法
- 算法：一组有穷的操作规则
  - 确定性：每一个步骤的结果都是确定的
  - 可行性：每一个步骤可在有限时间内完成
  - 输入：有输入
  - 输出：有输出
  - 有穷性：可在有限步骤内停止

计算语言学的研究对象

- 计算语言学的研究对象是自然语言
- 自然语言与形式语言的本质区别
  歧义性
- 自然语言是一种符号系统
- 语言符号的特点（索绪尔）
  - 任意性：语言符号的选择是任意的
  - 线条性：语言符号的排列是线性的

语言、思维与客观世界



语言的层面

- 语言研究的层面
  - 语音
  - 语法（包括词汇层和句法层）
    - 语法研究要回答的问题是：一句话为什么可以这么说而不能那么说？
  - 语义
    - 语义研究要回答的问题是：这句话说了什么？
  - 语用
    - 语用研究要回答的问题是：为什么要说这句话？

语言的层面（续）

- 语言各层面之间的关系
  - 语言层面的划分反映了语言在不同层次上的规律性
  - 语言的各个层面是互相交织密不可分的，语言层面的划分只是为了研究方便，对任何一个层面的研究都不能忽略其他层面所起的作用

语言在不同层面的歧义性

- 语音层面：多音字，同音词
  - 施氏食狮史（赵元任）

    石室诗士施氏，嗜狮，誓食十狮。氏时时适市视狮，十时，适十狮适市，是时，适施氏适市，施氏视是十狮，拭矢试，使是十狮逝世，适石室，石室湿，氏使侍拭石室，石室拭，始食是十狮尸，始识是十狮尸，实十石狮尸，试释是事。

语言在不同层面的歧义性（续）

- 语法层面
  - 词法歧义
    - 词性兼类：工作（动名兼类），在（动副兼类）
    - 词语切分歧义：乒乓球拍卖完了，鱼在长江中游
  - 句法歧义
    - 结构歧义：张三和李四的朋友
    - 组合关系歧义：观赏鱼

语言在不同层面的歧义性（续二）

- 语义层面
  - 一词多义：后门，人大，
    I can can the can in the can.
  - 结构语义歧义：吃饭，吃食堂，吃大碗
- 语境层面
  - 鸡蛋！
  - 他去修车了。

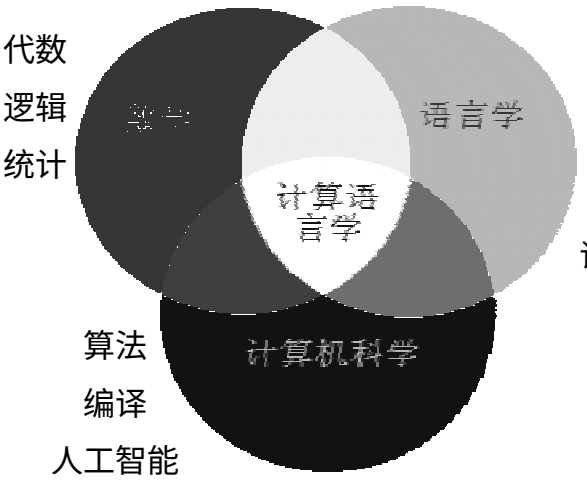
汉语的特点

- 语言的分类
  - 汉语：孤立语（分析语）
  - 英语：屈折语
  - 日语：粘着语
- 基本单位
  - 汉语：汉字（单音节，不用空格分隔）
  - 英语：词（多音节，用空格分隔）
- 词语形态变化
  - 汉语：弱（重叠、离合）
  - 英语：强（屈折）

汉语的特点（续）

- 语言的层次划分
  - 汉语：不明显：字与词、词与语、语与句、句与段，都没有明确的界限
  - 英语：明显：词、短语、子句、句子、段落之间界限分明
- 词类与句法功能的对应
  - 汉语：多对多
  - 英语：一对一

计算语言学与其他学科的关系



课程安排（1）

第一章 概论
- 第一节 计算语言学所要解决的问题（1学时）
- 第二节 计算语言学的定义及与其他学科的关系（1学时）
- 第三节 课程安排与参考文献

第二章 语言资源
- 第一节 词典（2学时）
- 第二节 语料库（2学时）

第三章 基于语言学的方法
- 第一节 词汇学与词法分析算法（4学时）
- 第二节 语法理论与句法分析算法（6学时）
- 第三节 语义学理论语义分析算法（2学时）
- 第四节 篇章分析的理论与算法（1学时）
- 第五节 语言生成的理论与算法（1学时）

课程安排（2）

第四章 基于统计学的方法
- 第一节 统计学基本常识（2学时）
- 第二节 n元语法（2学时）
- 第三节 隐马尔科夫模型（2学时）
- 第四节 概率上下文无关语法（2学时）
- 第五节 最大熵方法（2学时）
- 第五节 统计机器学习模型（2学时）

第五章 应用技术
- 第一节 音字转换（1学时）
- 第二节 自动文摘（1学时）
- 第三节 信息过滤与信息检索（2学时）
- 第四节 信息提取与问答系统（2学时）
- 第五节 机器翻译（2学时）

参考文献

刘开瑛、郭炳炎（1991）《自然语言处理》，科学出版社

冯志伟（1991）《数学与语言》，湖南教育出版社

冯志伟（1995）《自然语言机器翻译新论》，语文出版社1995年版。

姚天顺等（1995）《自然语言理解 —— 一种让机器懂得人类语言的研究》，清华大学出版社、广西科学技术出版社

冯志伟（1997）《自然语言的计算机处理》，上海外语教育出版社

翁富良、王野翊（1998）《计算语言学导论》，中国社会科学

俞士汶 等（1998）《现代汉语语法信息词典详解》，清华大学出版社、广西科学技术出版社

陈小荷（2000）《现代汉语自动分析》，北京语言文化大学出版社

刘颖（2002）《计算语言学，清华大学出版社

James Allen (1995), Natural Language Understanding, The Benjamin / Cummings Publishing Company, Inc.

Christopher D. Manning and Hinrich Schutze (1999), Foundations of Statistical Natural Language Processing, The MIT Press, Cambridge, Massachusetts

网络资源

- ACL主页：http://www.aclweb.org
- NLP新闻组：comp.ai.nat-lang
- LDC：http://www.ldc.upenn.edu
- 中文自然语言处理开放平台：
  http://www.nlp.org.cn
- 计算所自然语言处理研究组：
  http://mtgroup.ict.ac.cn
- 北京大学计算语言学研究所：
  http://www.icl.pku.edu.cn

复习思考题

- 如果让你实现一个机器翻译系统，你会如何做？
- 如果让你实现一个问答系统，你会如何做？
- 举例说明汉语和英语在不同层面上的歧义性。
- 英语句子“Time flies like an arrow”有多少种意思？
- 人机交流的语言和人类交流的语言有什么不同？
- 你觉得自然语言处理与人类的语言处理机制会有什么相同和不同之处？自然语言处理是否可能、是否需要模拟人类的语言处理机制？

致谢

- 本课程讲义（包括后续各节）直接引用了下面几位同行的课程讲义中的部分内容，在此深表感谢！
  - 詹卫东：《计算语言学概论》
  - 白　硕：《计算语言学》
  - 刘　颖：《计算语言学》