自然语言理解

第一章 绪 论

宗成庆

中科院自动化研究所

模式识别国家重点实验室

cqzong@nlpr.ia.ac.cn

http://www.nlpr.ia.ac.cn/English/cip/cqzong.htm

No.95, Zhongguancun East Road
Beijing 100080, China

http://www.nlpr.ia.ac.cn
Tel. No.: +86-10-6255 4263

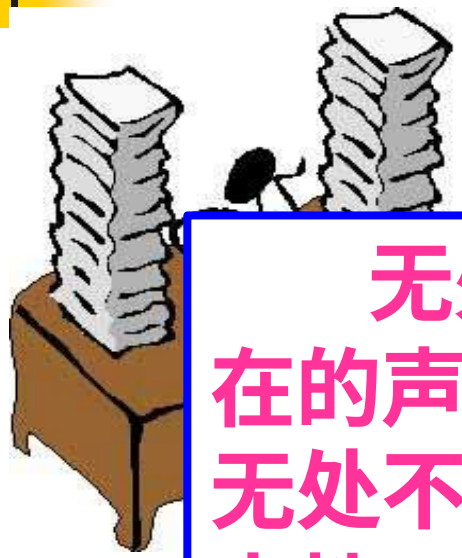
第一章 绪论

1.1 基本概念

当我们从事任何一项研究的时候，总要关注两方面的问题：一是是什么，为什么？二是做什么，怎么做？这恰恰是科学与技术紧密相关的两个方面。

自然语言处理既是一项技术，又是一门学科。

1.1 基本概念



无处不在的文字，无处不在的声音，无处不在的网络，无处不在的通讯；如此繁多的文档，如此繁多的电话，如此繁多的邮件，让我哪里能够应付的了？

1.1 基本概念

信息的主要载体－语言

语言的两种形式－文字和声音

文字和声音作为语言的两个不同形式的载体，所承载的信息占整个信息组成的70%以上（文字：70%，图象：20%；其它：10%）

- **如何让计算机实现人们希望实现的语言处理功能？**
- **如何让计算机真正实现海量的语言信息的自动处理和有效利用？**

1.1 基本概念

- **语言学** (linguistics)

是指对语言的科学研究。作为一门纯理论的学科，语言学在近期获得了快速发展，尤其从上个世纪60年代起，已经成为一门知晓度很高的广泛教授的学科。

包括：历时语言学（diachronic linguistics）**（或称历史语言学**（historical linguistics）**）和共时语言学**（synchronic linguistics）**、描述语言学**（descriptive linguistics）**、对比语言学**（contrastive linguistics）**、结构语言学**（structural linguistics）**等等。**

1.1 基本概念

- **语音学** (phonetics)

研究人类发音特点，特别是语音发音特点，并提出各种语音描述、分类和转写方法的科学。

包括: (1)**发音语音学**(articulatory phonetics)，**研究发音器官如何产生语音**；(2)**声学语音学**(acoustic phonetics)，**研究口耳之间传递语音的物理属性**；(3)**听觉语音学**(auditory phonetics)，**研究人通过耳、听觉神经和大脑对语音的知觉反应。**

1.1 基本概念

根据不同的研究方法，语音学又分为：

(a)一般语音学(general phonetics): **对语音发音、声学或知觉的一般研究。**

－与语言学的分析目的没有什么关系。

(b)**实验语音学(**experimental phonetics): **对具体语言语音特点的研究。**

－语言学研究的一部分，有人甚至认为是语言学不可或缺的基础。

1.1 基本概念

问题：

语音学究竟是一门独立的学科还是应视为语言学的一个分支呢？

➔ 复数的语言科学（linguistic sciences**）**

语言学和其它学科的交叉产生了许多语言学的新分支，包括纯理论的和应用性的，如人类语言学(anthropological linguistics)、**计算语言学**(computational linguistics)、**生物语言学**(biolinguistics)、**心理语言学**(psycholinguistics)、**教育语言学**(educational linguistics)**和社会语言学**(sociolinguistics)**等等。**

1.1 基本概念

❑**计算语言学** (Computational Linguistics)

计算语言学是利用电子数字计算机进行的语言分析。虽然许多其它类型的语言分析也可以运用计算机，计算分析最常用于处理基本的语言数据－例如建立语音、词、词元素的搭配以及统计它们的频率。

－《大不列颠百科全书》

1.1 基本概念

❑**计算语言学** (Computational Linguistics)

是语言学的一个研究分支，用计算技术和概念来阐述语言学和语音学问题。已开发的领域包括自然语言处理(natural language processing, NLP)**，言语合成，言语识别，自动翻译，编制语词索引，语法的检测，以及许多需要统计分析和领域（如文本考释）。**

—《现代语言学词典》[**戴维.克里斯特尔，**1997]

1.1 基本概念

❑自然语言处理

或称自然语言理解(natural language understanding, NLU)，**人工智能研究的重要内容之一。**

自然语言处理（natural language processing，NLP）**就是利用计算机为工具对人类特有的书面形式和口头形式的自然语言的信息进行各种类型处理和加工的技术。**

-冯志伟《自然语言的计算机处理》

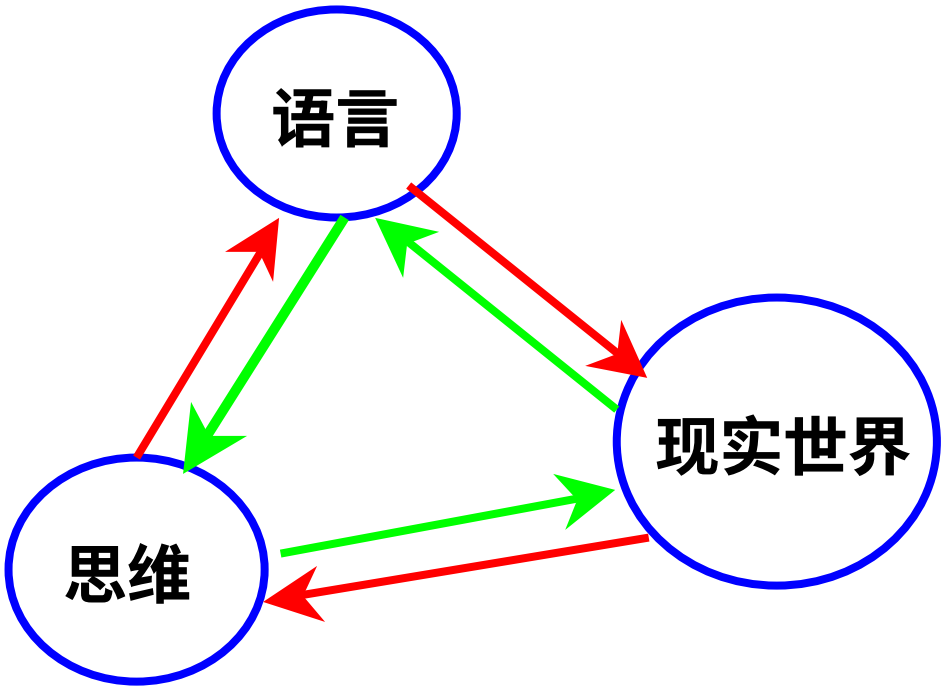
1.1 基本概念

近几年来，自然语言处理研究得到了前所未有的重视和长足的进展，并逐渐发展成为一门相对独立的学科而倍受关注，而且自然语言处理技术不断与语音识别(speech recognition)、语音合成(speech synthesis)等语音技术相互渗透和结合形成新的研究分支，因此，很多人在谈到“计算语言学”、“自然语言处理”或“自然语言理解”这些术语时，往往默认为同一个概念。甚至有些专著中干脆直接这样解释：计算语言学也称自然语言处理或自然语言理解[刘颖，2002]。

1.1 基本概念

- 人脑对语言的理解是一个复杂的思维过程
  - 语言学
  - 语言心理学
  - 逻辑学
  - 计算机科学
  - 人工智能
  - 数学与统计学

… …



1.2 关于“理解”标准

- 如何判断计算机系统的智能？

计算机系统的表现(act)如何？

反应(react)如何？

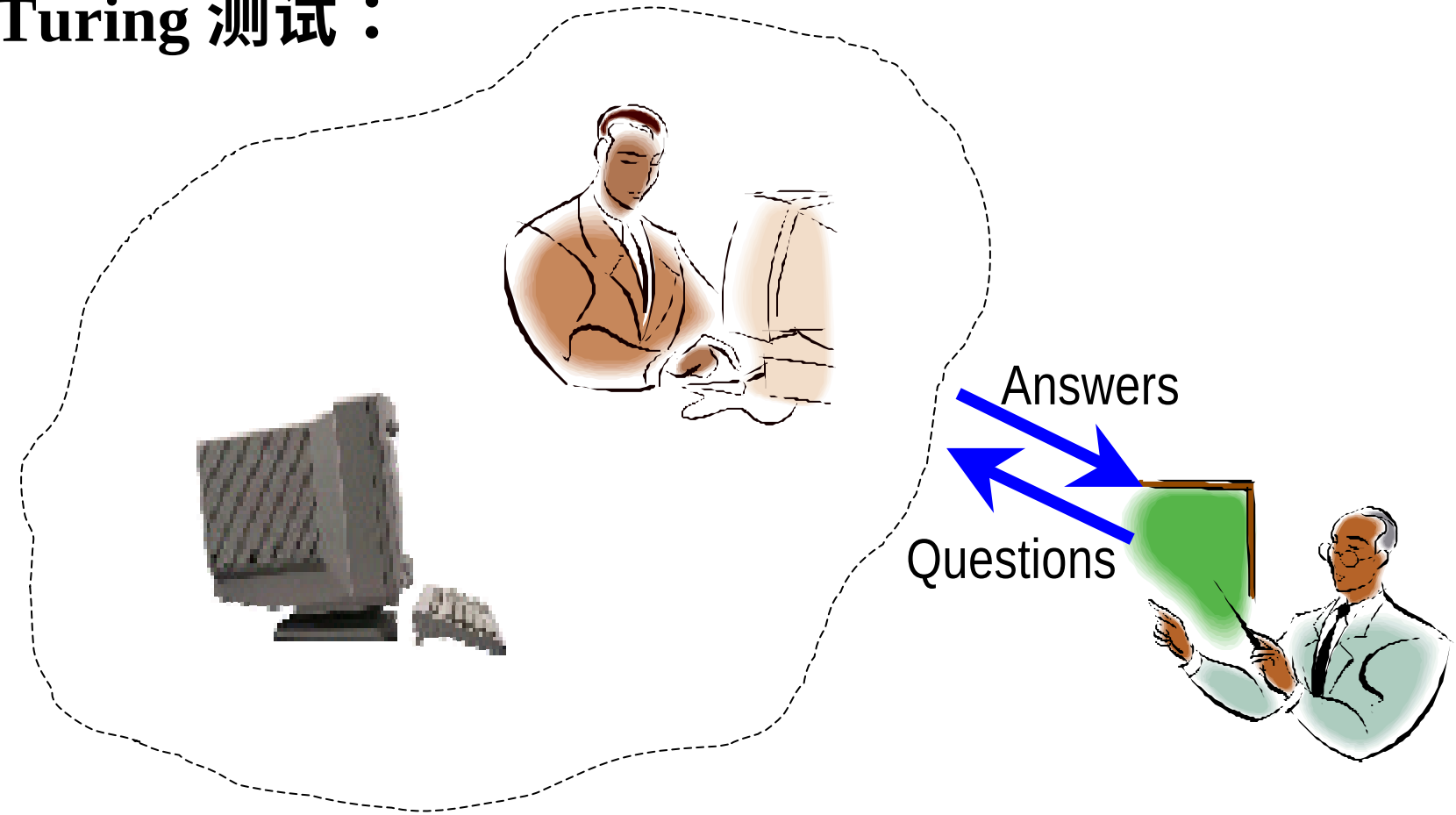
相互作用(interact)如何？

与有意识 个体（人）比较如何？

➔图灵(Turing)设计的“模仿游戏”，即图灵实验(Turing test)

1.2 关于“理解”标准

Turing 测试：



1.3 自然语言理解研究的内容

- 按照应用目标划分—
  - 机器翻译 (Machine translation, MT)：实现一种语言到另一种语言的自动翻译。
    - 应用：文献翻译、网页翻译和辅助浏览等。
    - 实用系统：Systran (http://www.systransoft.com)

      36种语言对，20个专门领域。

1.3 自然语言理解研究的内容

机器翻译现状和对机器翻译的认识

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。

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]

1.3 自然语言理解研究的内容

- **机器翻译现状和对机器翻译的认识**

用机器翻译的个别例子来批评甚至诋毁机器翻译研究是不适当的。

例1: The spirit is willing, but the flesh is weak.

（心有余，而力不足。）

精神是愿意的，但骨肉是微弱的。(Systran)

English-> Russian->English:

The wine is good, but the meat is spoiled.

(酒是好的，肉是馊的。)

1.3 自然语言理解研究的内容

例2: Out of sight, out of mind.

眼不见，心不烦。）

出于视域，在头脑外面。 (Systran)

From English to Russian:

又瞎又疯。

1.3 自然语言理解研究的内容

❖ **信息检索** (Information retrieval)：**信息检索也称情报检索，就是利用计算机系统从大量文档中找到符合用户需要的相关信息。**

面向多语言的信息检索叫做跨语言信息检索

(Cross-language / Trans-lingual information retrieval)。

- **代表系统：**Google: http://www.google.com

  百度：http://www.baidu.com.cn/

目前已有300多亿个网页，每天几百万增加，获得的信息只有1%被有效利用。

1.3 自然语言理解研究的内容

❖**自动文摘** (Automatic summarization / Automatic abstracting)：**将原文档的主要内容或某方面的信息自动提取出来，并形成原文档的摘要或缩写。**

- **应用：电子图书管理、情报获取等**

❖**文档分类** (Document categorization)：**文档分类也叫文本自动分类** (Text categorization / classification) **或信息分类** (Information categorization / classification)，**其目的就是利用计算机系统对大量的文档按照一定的分类标准（例如，根据主题或内容划分等）实现自动归类。**

- **应用：图书管理、内容管理、信息监控等**

1.3 自然语言理解研究的内容

❖**问答系统** (Question-answering system)：**通过计算机系统对人提出的问题的理解，利用自动推理等手段，在有关知识资源中自动求解答案并做出相应的回答。问答技术有时与语音技术和多模态输入/输出技术，以及人机交互技术等相结合，构成人机对话系统** (man-computer dialogue system)。

- **应用：人机对话系统、信息检索等**

1.3 自然语言理解研究的内容

❖ **信息过滤** (Information filtering)：**通过计算机系统自动识别和过滤那些满足特定条件的文档信息。**

- **应用：网络有害信息过滤、信息安全等**

❖ **语言教学** (Language teaching)：**借助计算机辅助教学工具，进行语言教学、操练和辅导等。**

- **应用：语言学习等**

1.3 自然语言理解研究的内容

❖ **文字识别** (Character recognition)：**通过计算机系统对印刷体或手写体等文字进行自动识别，将其转换成计算机可以处理的电子文本。**

- **应用：文字输入、识别等**

❖ **文字编辑和自动校对** (Automatic proofreading)：**对文字拼写、用词、甚至语法、文档格式等进行自动检查、校对和编排。**

- **应用：排版、印刷和书籍编撰等**

1.3 自然语言理解研究的内容

❖**语音识别**(speech recognition)：**将输入计算机的语音信号识别转换成书面语表示。语音识别也称自动语音识别** (automatic speech recognition, ASR)。

- **应用：文字录入、人机通讯、语音翻译等等。**
- **困难：大量存在的同音词、近音词、集外词、口音等等。**

例如：输入：美欧贸易摩擦升级

识别结果：美欧贸易摩擦生机

1.3 自然语言理解研究的内容

- 极端情况下的同音字（词）现象
  - 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。

1.3 自然语言理解研究的内容

❖ **文语转换** (text-to-speech)：**将书面文本自动转换成对应的语音表征。**

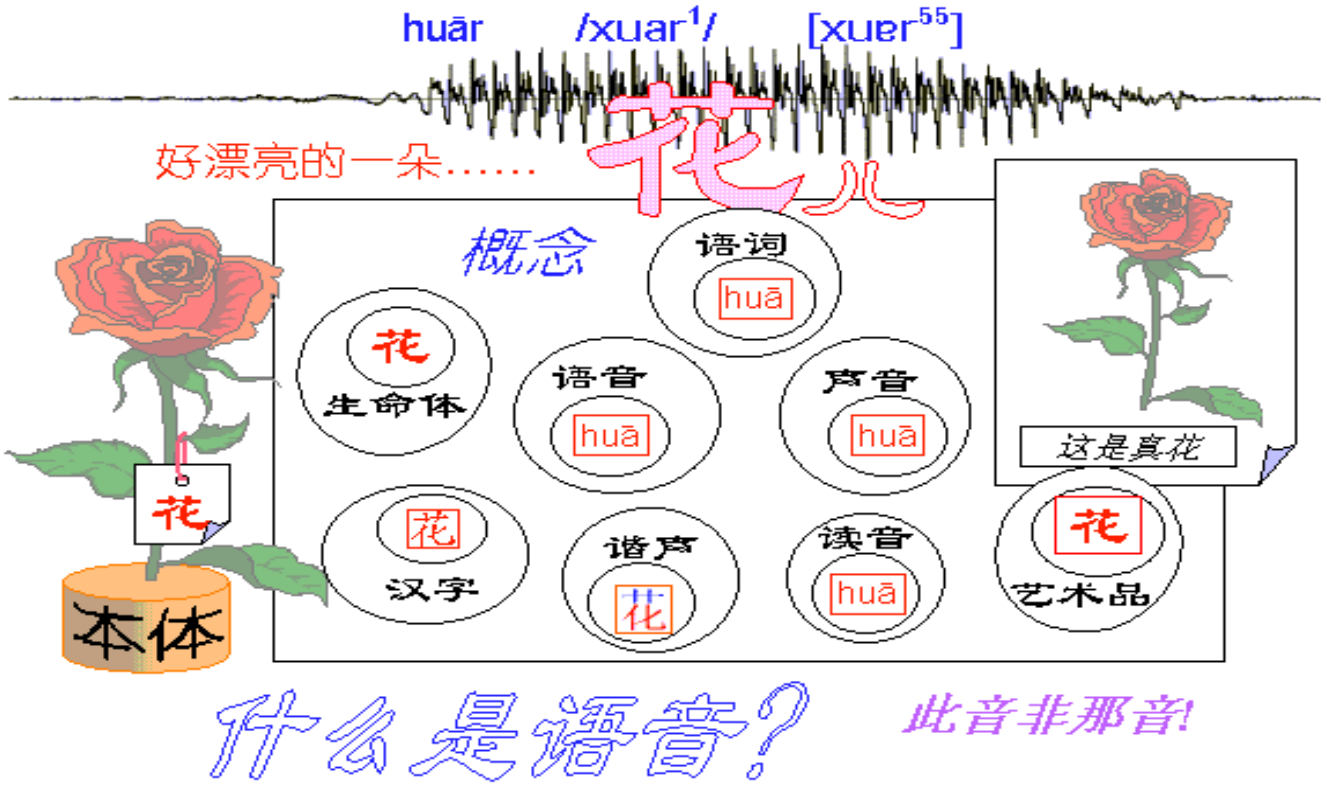
- **应用：朗读系统、人机语音接口等等。**

❖ **说话人识别/认同/验证** (speaker recognition/ identification/ verification)：**对一言语样品做声学分析，依此推断（确定或验证）说话人的身份。**

- **应用：信息安全、防伪等等。**

1.4 自然语言理解研究的基本问题

❑**语音学**(Phonetics) **问题：研究词及其语音的关联。**



1.4 自然语言理解研究的基本问题

❑**形态学** (Morphology) **问题：研究词是如何由意义的基本单位－词素（**morphemes**）构成的。**

词素（morphemes**）→ 词 （**word**） ？**

ß

词根、前缀、后缀、词尾

例： 人，蜈蚣

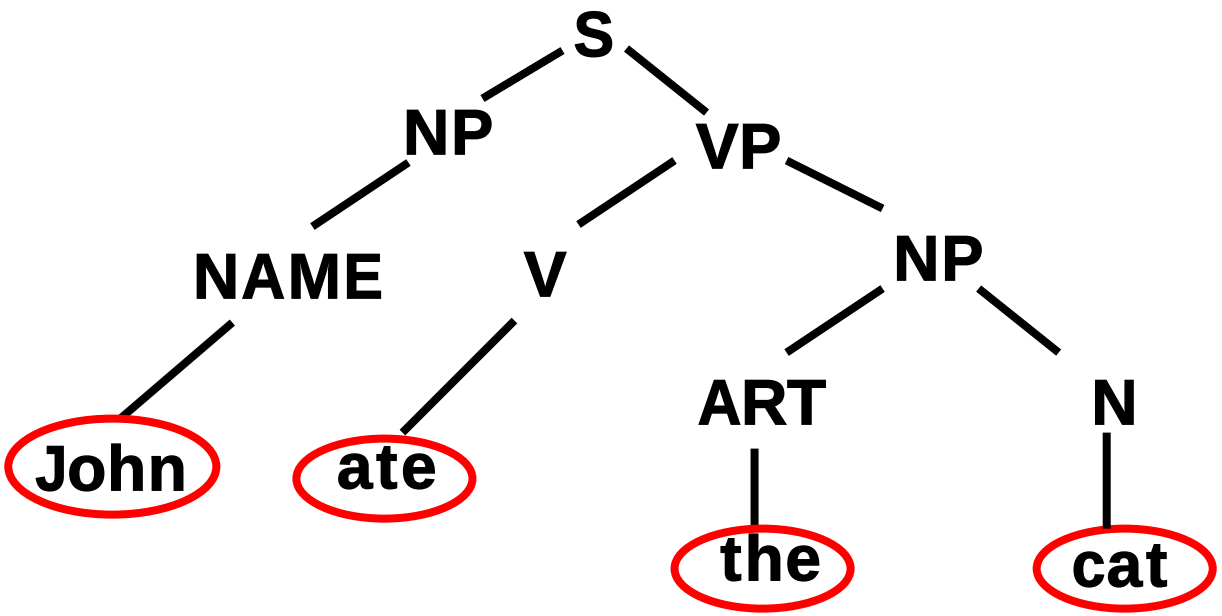
老虎 ← 老 ＋ 虎；　　图书馆 ← 图 ＋ 书 ＋ 馆

re + ex + port → reexport

1.4 自然语言理解研究的基本问题

❑语法学 (Syntax) 问题：研究句子结构成分之间的相互关系和组成句子序列的规则。

为什么一句话可以这么说也可以那么说？



1.4 自然语言理解研究的基本问题

❑语义学 (Semantics) 问题：研究如何从一个语句中词的意义，以及这些词在该语句中句法结构中的作用来推导出该语句的意义。

这句话说了什么？

(1) 苹果不吃了。

(2) 这个人真牛。

(3) 这个人眼下没些什么，那个人嘴不太好。

1.4 自然语言理解研究的基本问题

❑语用学(Pragmatics) 问题：研究在不同上下文中的语句的应用，以及上下文对语句理解所产生的影响。从狭隘的语言学观点看，语用学处理的是语言结构中有形式体现的那些语境。相反，语用学最宽泛的定义是研究语义学未能涵盖的那些意义。

为什么要说这句话？

（1）火，火！

（2）A：看看鱼怎么样了？

B：我刚才翻了一下。

1.5 自然语言理解面临的困难

❑自然语言中大量存在的歧义(ambiguity)现象

❖ 结构歧义

例如: (1) **Who** has seen John? — 主语

(2) **Who** has John seen? — 宾语

(3) 今天中午吃**馒头**。

(4) 今天中午吃**食堂**。

1.5 自然语言理解面临的困难

(5) I saw a man with a telescope.

➔ I saw [a man with a telescope].

I [saw a man] with a telescope.

➔ I saw a man with a telescope in the park. ?

歧义组合数我们称之为开塔兰数(Catalan Numbers，**记作**C_n)：

$$C_n = \binom{2n}{n}\frac{1}{n+1} \qquad \text{其中：} \quad \binom{2n}{n} = \frac{(2n)!}{n!\times n!}$$

n**为句子中介词短语的个数。**

1.5 自然语言理解面临的困难

❖ 语义歧义

他说："她这个人真有意思(funny)"。她说："他这个人怪有意思的(funny)"。于是人们以为他们有了意思(wish)，并让他向她意思意思(express)。他火了："我根本没有那个意思(thought)"！她也生气了："你们这么说是什么意思(intention)"？事后有人说："真有意思(funny)"。也有人说："真没意思(nonsense)"。

—《生活报》1994. 11. 13. 第六版

1.5 自然语言理解面临的困难

- 自然语言中存在未知的语言现象
  - 新的词汇

    例如:“非典”、专业术语、外来语、人名等
  - 新的含义

    例如：窗口、奔腾、农民等
  - 新的用法和语句结构等

    尤其在口语中或部分网络语言中，不断出现一些“非规范的”新的语句结构。

1.6 不同语言的差异

- 不同的语系

  孤立语（分析语）:形态变化少，语法关系靠词序和虚词表示，如汉语。

  曲折语：用词的形态变化表示语法关系，如英语。

  黏着语：词内有专门表示语法意义的附加成分，词根或词干与附加成分的结合不紧密，如日语。

1.6 不同语言的差异

- 不同的语言单位

  汉语：汉字（单音节、无空格）

  英语：英语（多音节、有空格）

  日语：字和词（多音节、无空格）

1.6 不同语言的差异

- 不同的语法
  - 例如:

    (1) 二楼$_1$三号房间$_2$桌子$_3$上有一本书。

    (2) There is a book on the desk$_3$ in Room 3$_2$ at the 2nd floor$_1$.

- 语义的差异

1.7 自然语言理解研究的基本方法

- **理性主义与经验主义方法的哲学分野**

  之一：对语言知识来源的不同认识

理性主义认为：人的很大一部分语言知识是与生俱来的，由遗传决定的。

Chomsky **的内在语言官能(**innate language faculty**)理论被广泛接受。**

人工编汇初始语言知识 ＋ 推理系统 ⇨ 自然语言处理系统

1960s – 1980s中期

1.7 自然语言理解研究的基本方法

经验主义认为：人的语言知识是通过感观输入，经过一些简单的联想(association)与通用化(generalization)的操作而得到的。

大量的语言数据中获得语言的知识结构。

1920s – 1950s，1980s中期-

1.7 自然语言理解研究的基本方法

- **理性主义与经验主义方法的哲学分野之二：研究对象的差异**

理性主义方法：研究人的语言知识结构（语言能力，language competence），实际的语言数据（语言行为，language performance）只提供了这种内在知识的间接证据。

经验主义方法：直接研究这些实际的语言数据。

1.7 自然语言理解研究的基本方法

- **理性主义与经验主义方法的哲学分野之三：运用不同的理论**

理性主义：通常基于 Chomsky 的语言原则（principles），通过语言所必须遵守的一系列原则来描述语言。

经验主义：通常是基于 Shannon 的信息论。

1.7 自然语言理解研究的基本方法

- 理性主义与经验主义方法的哲学分野之四：采用不同的处理方法

理性主义：通常通过一些特殊的语句或语言现象的研究来得到对人的语言能力的认识，而这些语句和语言现象在实际的应用中并不常见。

经验主义：偏重于对大规模语言数据中人们所实际使用的普通语句的统计。

1.7 自然语言理解研究的基本方法

- 理性主义方法与经验主义方法的融合

符号智能 ＋ 计算智能

理性主义研究方法 — 符号处理系统

经验主义研究方法 — 基于语言数据的计算方法

理性主义与经验主义的合谋 — 融合方法

1.8 自然语言理解的发展和研究现状

❑自然语言理解的发展

－萌芽期：1946年世界上第一台计算机出现，自然语言理解的研究起始于机器翻译。

－发展期：自1966年美国自动语言处理咨询委员会(ALPAC) 提出ALPAC报告。研究重点转写其它分支：人机接口、对话系统、信息检索等。基本方法：基于规则分析方法。

－繁荣期：自20世纪80年代末期以后，基于语料库的统计方法引入自然语言处理。

1.8 自然语言理解的发展和研究现状

❑基本研究现状

- ❖ 实用或半实用的技术已经得到广泛运用
  - —文字处理器
  - —文字输入
  - —网络搜索引擎
  - —辅助翻译、电子词典
  - —语音合成

  … …

1.8 自然语言理解的发展和研究现状

❖许多技术离真正实用的目标还有相当的距离，若干理论问题有待于进一步深入研究

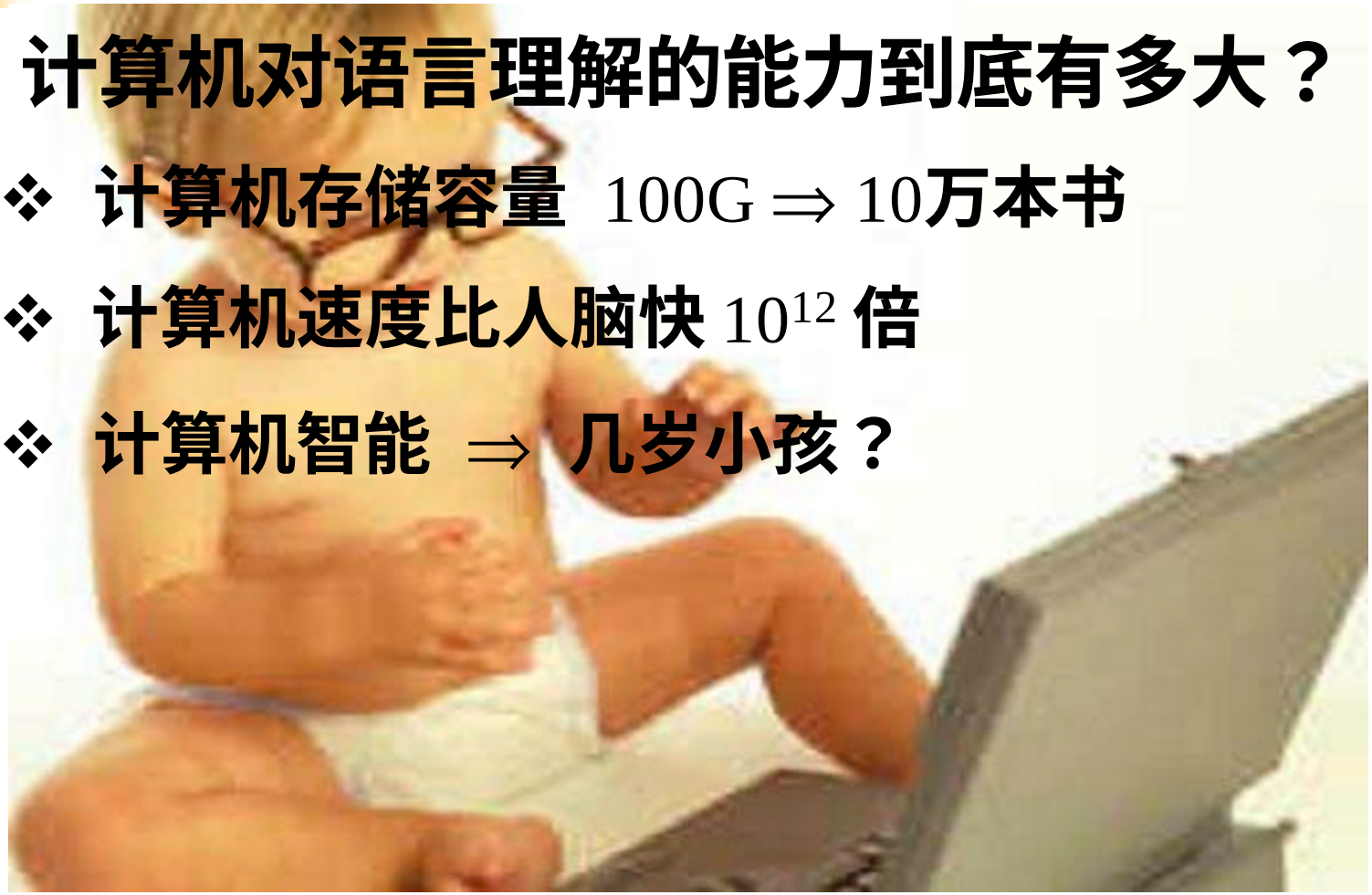
- 现有模型和方法的改进
- 期待新的理论方法

❖许多新的研究方向不断出现

- 网络内容管理、信息监控、有害信息过滤等
- 语音文摘生成

1.8 自然语言理解的发展和研究现状

- 计算机对语言理解的能力到底有多大？
  - 计算机存储容量 100G ⇒ 10万本书
  - 计算机速度比人脑快 10^{12} 倍
  - 计算机智能 ⇒ 几岁小孩？



1.9 课程安排

第一章　引论
第二章　数学基础
第三章　形式语言与自动机
第四章　语料库语言学
第五章　概率语法
第六章　词法分析技术
第七章　句法理论与句法分析
第九章　计算语义学
第十章　应用系统介绍－机器翻译、语音翻译、文本分类、信息检索、对话系统等。

1.10 参考文献

❑ 专著

[1] 瓮富良，计算语言学导论，中国社会科学出版社，1998。

[2] 冯志伟，自然语言的计算机处理，上海外语教育出版社，1996。

[3] 姚天顺，自然语言理解－一种让机器懂得人类语言的研究，清华大学、广西科技出版社，2002（第二版）。

[4] 赵铁军，机器翻译原理，哈尔滨工业大学出版社，2000。

1.10 参考文献

[5] James Allen, Natural Language Understanding. The Benjamin/Cummings Publishing Company, Inc. 1995.

[6] Christopher D. Manning, Hinrich Schute, Foundations of Statistical Natural Language Processing. The MIT Press. 1999.

[7] Rens Bod, Jennifer Hay et al. Probabilistic Linguistics. The MIT Press. 2003.

1.10 参考文献

- 期刊

1) Computational Linguistics
2) Machine Translation
3) Computer Speech and Language
4) Computational Linguistics and Chinese Language Processing
5) ACM Trans. on Asia Language Processing
6) IEEE Trans. on Speech and Audio Processing, etc.
7) 中文信息学报
8) 计算机学报、软件学报、计算机研究与发展

1.10 参考文献

❑ 会议论文集

[1] Proceedings of ACL (Annual Meeting of the Association for Computational Linguistics)

[2] Proceedings of COLING (Inter. Conf. on Computational Linguistics)

[3] Proceedings of IJC-NLP (Inter. Conf. on Natural Language Processing)

[4] **全国计算语言学联合学术会议论文集**

本章小结

- 对自然语言理解的基本认识
  - 基本概念
  - 研究内容及面临的问题
  - 研究方法
  - 课程安排
  - 参考文献

思考题

1-1. **说明如下句子有多少种不同的含义？**

(1) Time flies like an arrow.　(2) He drew one card.

(3) **咬死猎人的狗。**

1-2. **试比较汉语和英文句子中地点状语位置的差异。**

1-3. **思考一下，如果用计算机编译技术中程序设计语言的某一句法分析方法直接解析普通的英文句子，会存在什么问题？**

1-4. **思考一下，你的大脑理解一个英文句子的基本过程。**

Thanks

谢谢!