自然语言处理

默莎莎

联系方式

- **邮　件：**moshasha@buaa.edu.cn
- **地　址：**新主楼G座1038房间

课程内容

◆ 背景知识

1. 概率论、信息论、建模方法基础
2. 基本的语言学知识
3. 算法分析基础、编程能力

◆ 目的

1. 掌握自然语言理解的基本概念、理论、方法
2. 掌握正确分析问题、解决问题的思维方式

◆ 作业

课程内容

◆ 考核方式

1. 笔试: 60%
2. 作业: 20%
3. 平时考勤: 20%

课程内容

◆ 参考文献

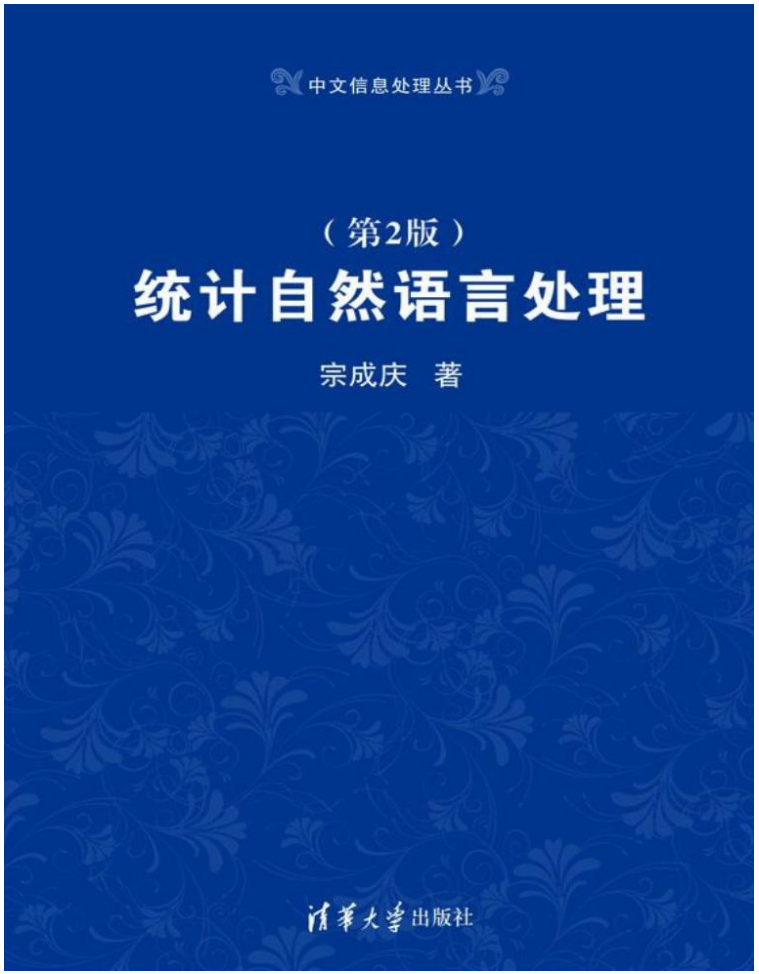
本课程教材：

书　名: 统计自然语言处理

作　者: 宗成庆

出版社: 清华大学出版社

时　间: 2013年8月



课程内容

◆ 参考文献

其它专著：

1. 冯志伟，孙乐(译)(D. Jurafsky, J. H. Martin 著)，自然语言处理综论，电子工业出版社, 2005.
2. 冯志伟，自然语言处理的形式模型，中国科学技术大学出版社， 2010.
3. 翁富良, 王野翊, 计算语言学导论, 社科出版社, 1998.
4. C. D. Manning, Hinrich Schute, Foundations of Statistical Natural Language Processing. The MIT Press. 1999.
5. James Allen, Natural Language Understanding. The Benjamin/ Cummings Publishing Company, Inc. 1995.

课程内容

◆ 参考文献

期刊：

1. Computational Linguistics
2. Natural Language Engineering
3. ACM Trans. on ALIP
4. Machine Translation
5. IEEE Trans. on Audio, Speech, and Language Processing
6. 中文信息学报/ 计算机学报/ 软件学报/ 计算机研究与发展

课程内容

◆ 参考文献

会议论文集：

1. Proceedings of ACL (Annual Meeting of the Association for Computational Linguistics)
2. Proceedings of NAACL, EMNLP
3. Proceedings of COLING (International Conference on Computational Linguistics)
4. Proceedings of IJCNLP (International Joint Conference on Natural Language Processing)
5. 国内相关会议论文集

课程内容

第1章　绪论
第2章　数学基础
第3章　形式语言与自动机
第4章　语料库与语言知识库
第5章　语言模型
第6章　概率图模型
第7章　词法分析与词性标注
第8章　句法分析
第9章　语义分析
第10章 篇章分析
第11章 应用系统介绍——机器翻译等

课程内容

机器翻译等应用	第11章
语义与篇章分析	第9，10章
语法理解与句法分析	第8章
词法分析	第7章
工具: 资源与模型	第3，4，5，6章
基础: 概念与数学	第1，2章

第1章 绪论

目 录

1.1 问题的提出

1.1 问题的提出

- 自然语言指人类日常使用的语言，如汉语、英语、法语，德语等；
- 语言是思维的载体，是人类交流思想和表达情感最自然、最直接、最方便的工具。

1.1 问题的提出

- 人类历史上以语言文字形式记载和流传的知识占知识总量的80%以上；
- 2008年1月中国互联网络信息中心（CNNIC）发布的《第21次中国互联网络发展状况统计报告》表明，中国互联网上87.8%的网页内容是由文本表示的。

1.1 问题的提出

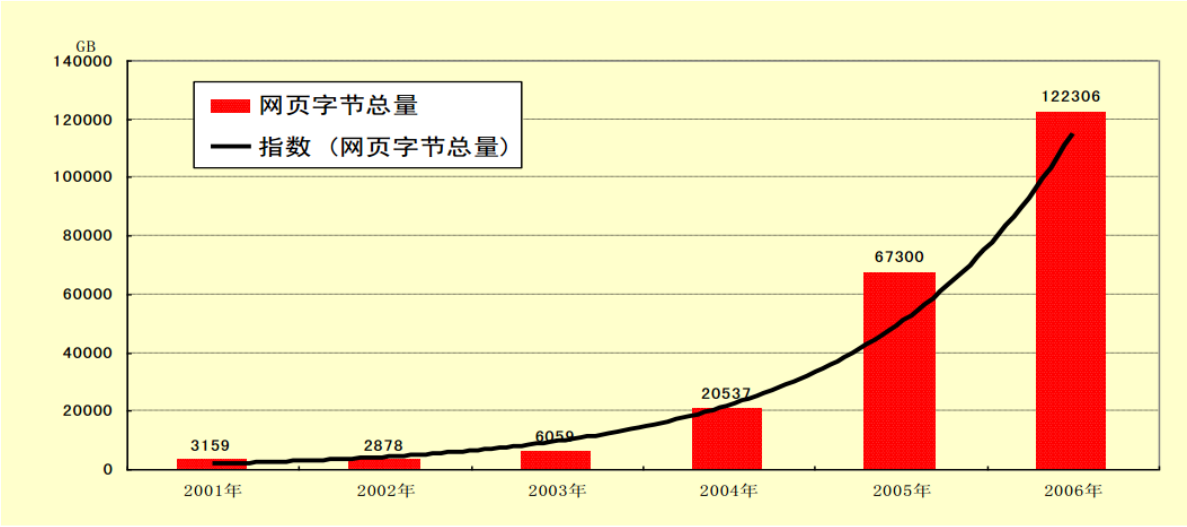
无处不在的网络、无处不在的通讯和堆积如山的文档，构成了当今社会信息爆炸的基本特征。当现代化的信息传播手段给人们的生活和工作带来极大便利的同时，也使人们面临许多难以克服的困难和障碍。有关专家指出，语言障碍是21世纪社会全球化所面临的主要困难之一。

1.1 问题的提出

信息检索市场前景广阔

- 全世界网页数量正以指数速率增长，思科公司估计，到2019年全球互联网流量将达到每年2泽字节。

1泽字节=1000艾字节，1艾字节=100亿亿字节



- 中文网页检索的最高准确率不足 40%

1.1 问题的提出

全世界正在使用的语言有1900多种

45个国家官方语言是英语，75%的电视节目是英语，80%以上科技信息用英文表达。

英语作为第一交流语言的人有3.8亿；第二交流语言的人有3.8亿；学习英语的人有7.5亿。

- 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，将超过英语成为世界上使用人数最多的语言。

1.1 问题的提出

跨语言通信与信息获取

机器翻译市场需求大

◆文化　◆商贸　◆旅游　◆体育 ……

1.1 问题的提出

舆情监测

- 利用网络组织犯罪，已成为恐怖活动的新特点；
- 信息安全问题已经成为国际社会共同关注的焦点。

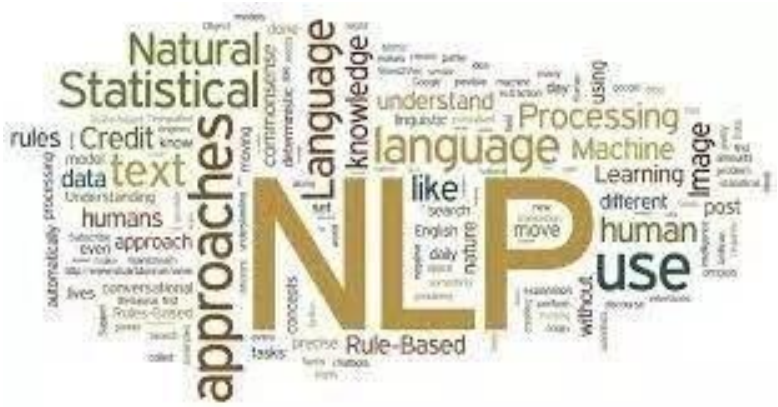
恐怖爆炸

1.1 问题的提出

- 如何让计算机实现自动的或人机互助的语言处理功能?
- 如何让计算机实现海量语言信息的自动处理、知识挖掘和有效利用?

自然语言处理

Natural Language Processing, NLP



1.2 基本概念

1.2 基本概念

- **语言 vs. 自然语言**
- **语言学 vs. 语音学**
- **自然语言理解 vs. 自然语言处理**
  vs. 计算语言学
  vs. 中文信息处理

1.2 基本概念

◆ 定义1-1：语言(Language)

人类所特有的用来表达意思、交流思想的工具，是一种特殊的社会现象，由语音、词汇和语法构成一定的系统。“语言”一般包括它的书面形式，但在与“文字”并举时只指口语。

——商务印书馆，《现代汉语词典》，1996

自然语言：指人类社会发展过程中自然产生的语言，而不是人为编造的语言，如程序语言等。

1.2 基本概念

◆ 定义1-2：语言学(linguistics)

语言学是指对语言的科学研究。

——戴维•克里斯特尔，《现代语言学词典》，1997

研究语言的本质、结构和发展规律的科学。

——商务印书馆，《现代汉语词典》，1996

语音和文字是语言的两个基本属性。

1.2 基本概念

- 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：

  (1)历时语言学 (diachronic linguistics)

  或称历史语言学(historical linguistics)

  (2)共时语言学 (synchronic linguistics)

  (3)描述语言学 (descriptive linguistics)

  (4)对比语言学(contrastive linguistics)

  (5)结构语言学(structural linguistics) ……

1.2 基本概念

◆ 定义1-3：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。

——戴维•克里斯特尔，《现代语言学词典》，1997

1.2 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。

——是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。

1.2 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)

1.2 基本概念

◆ 定义1-4：自然语言理解 (Natural Language Understanding, NLU)

从微观上讲，语言理解是指从自然语言到机器(计算机系统)内部之间的一种映射。从宏观上讲，语言理解是指机器能够执行人类所期望的某些语言功能。这些功能包括回答有关提问、提取材料摘要、不同词语叙述、不同语言翻译。

——蔡自兴、徐光佑，《人工智能及其应用》
清华大学出版社，2004

1.2 基本概念

关于“理解”的标准

- 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?

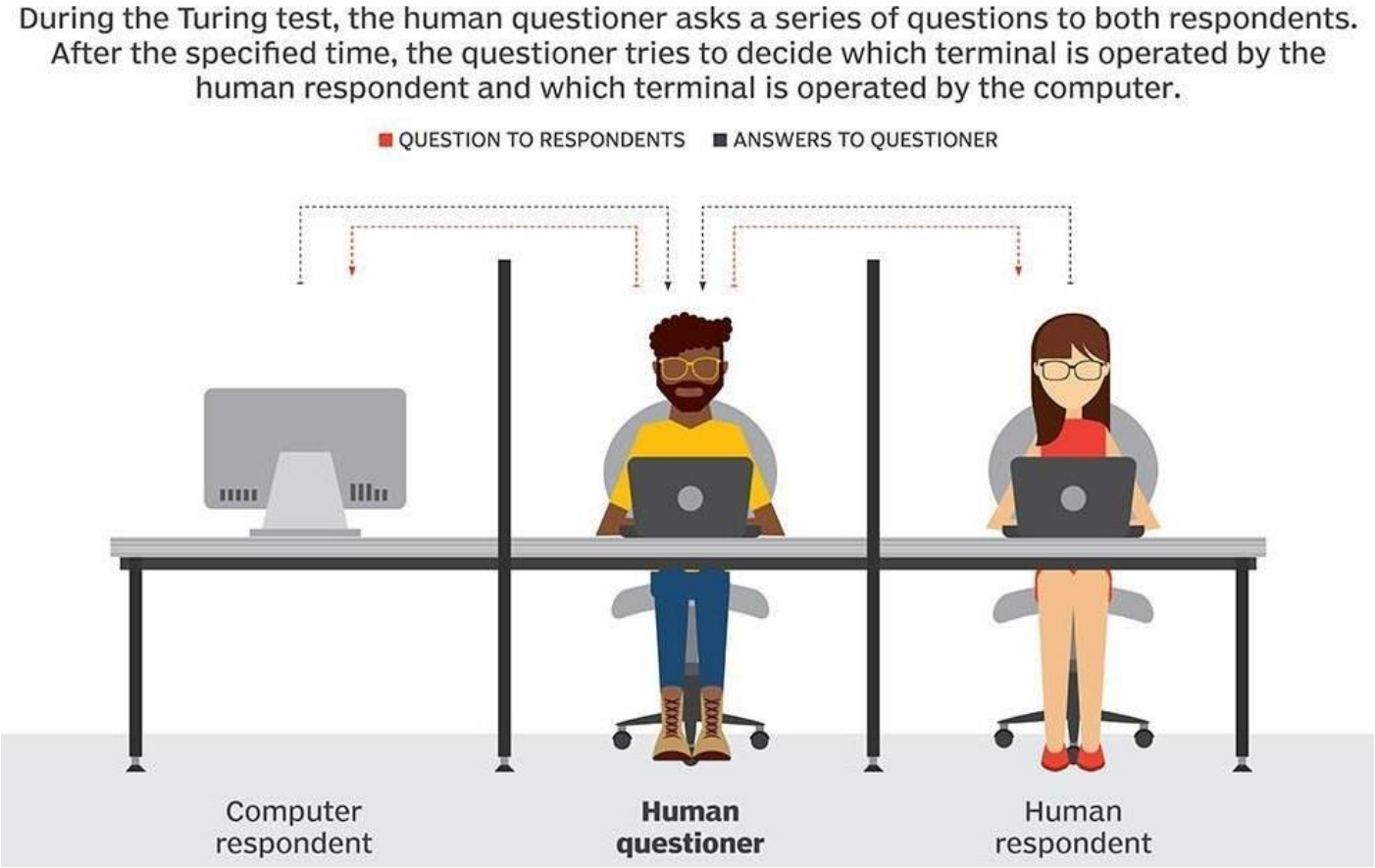
相互作用(interact)如何?

与有意识的个体（人）比较如何?

图灵设计的“模仿游戏” —— 图灵实验(Turing test)

1.2 基本概念

图灵测试(Turing Test)：



TIPS：关于图灵测试仍有争议

1.2 基本概念

图灵测试(Turing Test):

人们在自然语言处理研究领域中的任何一个应用系统都可以拿来做图灵测试，包括对答系统、文摘生成系统和机器翻译系统等。按照人的标准对这些系统的输出结果进行评价，从而判断计算机系统是否达到了“理解”的效果。

自然语言理解研究的任务：研究和探索针对具体应用目的的新方法和新技术，使实现系统的性能表现尽量符合人类理解的标准和要求

1.2 基本概念

◆ 定义1-5：自然语言处理 (Natural Language Processing, NLP)

研究在人与人交际中以及在人与计算机交际中的语言问题的一门学科。自然语言处理要研制表示语言能力(linguistic competence)和语言应用(linguistic performance)的模型，建立计算框架来实现这样的语言模型，提出相应的方法来不断地完善这样的语言模型，根据这样的语言模型设计各种实用系统，并探讨这些实用系统的评测技术 。

—Bill Manaris《从人－机交互的角度看自然语言处理》

1.2 基本概念

◆ **定义1-5：自然语言处理**
(Natural Language Processing, NLP)

自然语言处理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》
上海外语教育出版社，1996

1.2 基本概念

◆ 定义1-6：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。虽然许多其他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》

最早出现于1966年美国科学院的 ALPAC 报告。

1.2 基本概念

◆ **定义1-6：计算语言学**
(Computational Linguistics)

计算语言学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**，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）。

——戴维•克里斯特尔，《现代语言学词典》，1997

1.2 基本概念

◆三个不同的语系

1. 屈折语(fusional language/inflectional language): 用词的形态变化表示语法关系，如英语、法语等。
2. 黏着语(agglutinative language): 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、韩语、土耳其语等。
3. 孤立语/分析语(isolating / analytic language): 形态变化少，语法关系靠词序和虚词表示，如汉语。

1.2 基本概念

汉语： 汉族的语言，是我国的主要语言。

中文： 中国的语言文字，特指汉族的语言文字。

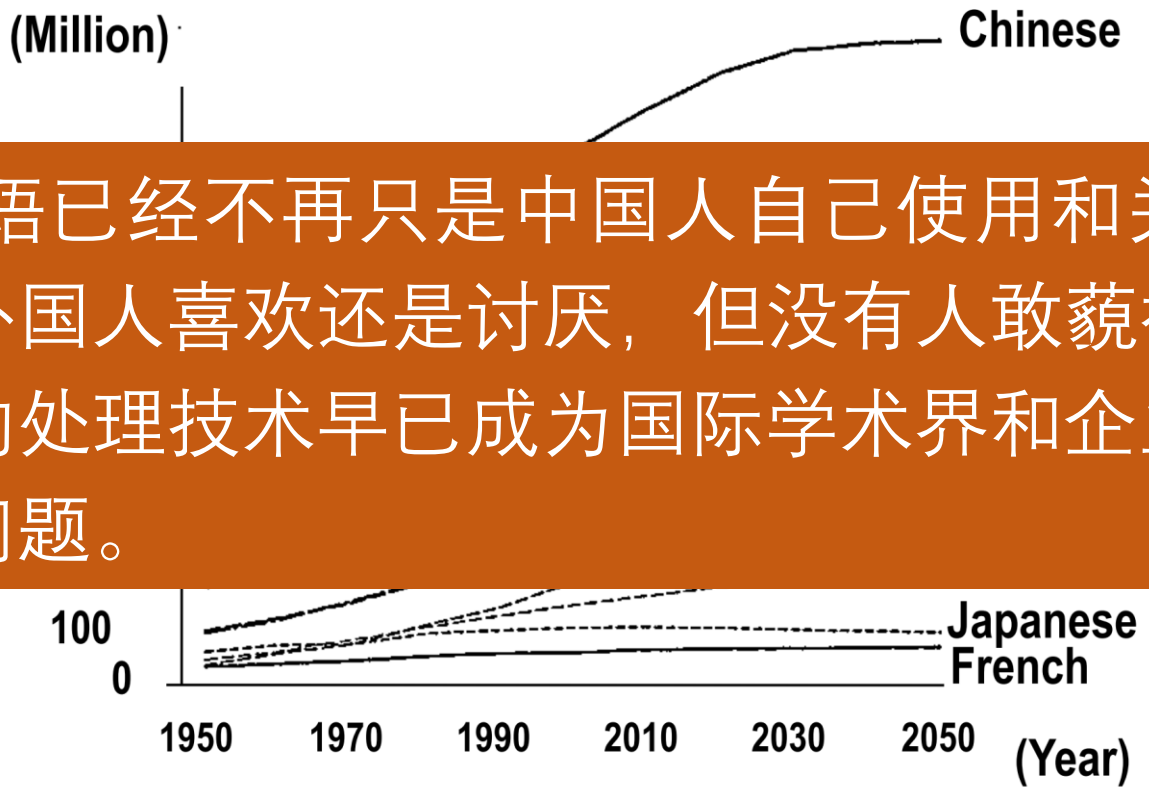
——《现代汉语词典》，1996

◆ **定义1-7：中文信息处理**
(Chinese Information Processing)

针对中文的自然语言处理技术。

1.2 基本概念

- 各语言使用人数发展趋势



汉语已经不再只是中国人自己使用和关注的语言，不管外国人喜欢还是讨厌，但没有人敢藐视她！针对汉语的处理技术早已成为国际学术界和企业界共同关注的问题。

1.2 基本概念

- 2010年十大网站使用的语言情况统计

语言种类	用户数量	用户比例	发展速度 (2000-2010)	全球分布比	人口数量
英语	536564837	42.0 %	281.2 %	27.3 %	1277528133
汉语	444948013	32.6 %	1,277.4 %	22.6 %	1365524982
西班牙语	153309074	36.5 %	743.2 %	7.8 %	420469703
日本	99143700	78.2 %	110.6 %	5.0 %	126804433
葡萄牙	82548200	33.0 %	989.6 %	4.2 %	250372925
德语	75158584	78.6 %	173.1 %	3.8 %	95637049
阿拉伯语	65365400	18.8 %	2,501.2 %	3.3 %	347002991
法语	59779525	17.2 %	398.2 %	3.0 %	347932305
俄语	59700000	42.8 %	1,825.8 %	3.0 %	139390205
韩语	39440000	55.2 %	107.1 %	2.0 %	71393343
热门10语言	1615957333	36.4 %	421.2 %	82.2 %	4442056069
其余的语言	350557483	14.6 %	588.5 %	17.8 %	2403553891
世界总计	1966514816	28.7 %	444.8 %	100.0 %	6845609960

1.2 基本概念

近几年来，自然语言处理技术迅速发展成为一门相对独立的学科，倍受关注，而且该技术不断与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。甚至有专著[刘颖，2002]干脆直接解释为：

计算语言学也称自然语言处理或自然语言理解。

1.3 NLP的产生与发展

1.3 NLP的产生与发展

源自机器翻译 (Machine Translation, MT)

- 1946年UPenn的J. P. Eckert和 J. W. Mauchly设计了世界上第一台电子计算机ENIAC；
- 英国工程师Andrew Donald Booth和美国洛克菲勒基金会(Rockefeller Foundation)副总裁W. Weaver提出机器翻译的概念。
- 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

1.3 NLP的产生与发展

- A. D. 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月至9月，曾在普林斯顿大学参与 John von Neumann 研究组，后来曾在伦敦大学工作。

- W. Weaver，信息论先驱
- 1920至1932年曾在Wisconsin 大学教授数学；
- 1932至1955年担任洛克菲勒基金会自然科学部主任。

1.3 NLP的产生与发展

- 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。
- 在随后10多年里，MT研究在国际上出现热潮，一批自然语言人机接口系统和对话系统相继出现。

随着机器翻译研究的进展，各种自然语言处理技术应运而生，并逐渐发展壮大，形成了这一语言学与计算机技术相结合的新兴学科。

1.3 NLP的产生与发展

曲折的发展历程：

- 1960S 中期之前：萌芽期
- 1960S 中期到1970S 中后期：步履维艰
- 1966年美国科学院发表 ALPAC报告
- 1970S 中后期到1980S 后期：复苏
- 1980S 后期至今：蓬勃发展

1.4 研究内容

1.4 研究内容

◆按照应用目标划分，广义上包括：

1. 机器翻译 (Machine translation, MT)
2. 信息检索 (Information retrieval)
3. 自动文摘 (Automatic summarization/Automatic abstracting)
4. 问答系统 (Question-answering system)
5. 信息过滤 (Information filtering)
6. 信息抽取 (Information extraction)
7. 文档分类 (Document categorization)

1.4 研究内容

◆按照应用目标划分，广义上包括：

8. 情感分类(Sentimental classification)
9. 文字编辑和自动校对(Automatic proofreading)
10. 语言教学 (Language teaching)
11. 文字识别 (Character recognition)
12. 语音识别 (automatic speech recognition, ASR)
13. 文语转换/ 语音合成 (text-to-speech synthesis)
14. 说话人识别/认同/验证 (speaker recognition/ identification/ verification)

1.4 研究内容

◆ 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、网页辅助浏览等。

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。

百度： http://fanyi.baidu.com/ (200多种)

有道： http://fanyi.youdao.com/ (13种语言↔汉语)

1.4 研究内容

◆机器翻译 (Machine translation, MT):

- 例:

The spirit is willing, but the flesh is weak.

心有余，而力不足

Google翻译：Have more than enough power

Google翻译：拥有足够的力量

1.4 研究内容

◆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。

- 目前至少有300多亿个网页，每天数以万计地增加，只有1%的信息被有效地利用。
- 代表系统：

  Google：http://www.google.com

  百度：http://www.baidu.com.cn

1.4 研究内容

- **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。

  观点挖掘 (Opinion mining)。

- 应用： 电子图书管理、情报获取等。

1.4 研究内容

◆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)。

社区问答(Community Question Answering, CQA)

1.4 研究内容

◆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。

◆信息抽取 (Information extraction)：

从指定文档中或者海量文本中抽取出用户感兴趣的信息。

实体关系抽取 (entity relation extraction)。

社会网络 (social network)

1.4 研究内容

◆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。

◆情感分类(Sentimental classification)：

- 应用： 图书管理、情报获取、网络内容监控、用户倾向性分析等。

1.4 研究内容

- 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。

- 语言教学 (Language teaching)
- 文字识别 (Character recognition)
- ……

1.4 研究内容

◆语音识别 (automatic speech recognition, ASR)

将输入语音信号自动转换成书面文字。

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

1.4 研究内容

◆ 文语转换/语音合成:
(text-to-speech synthesis)

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

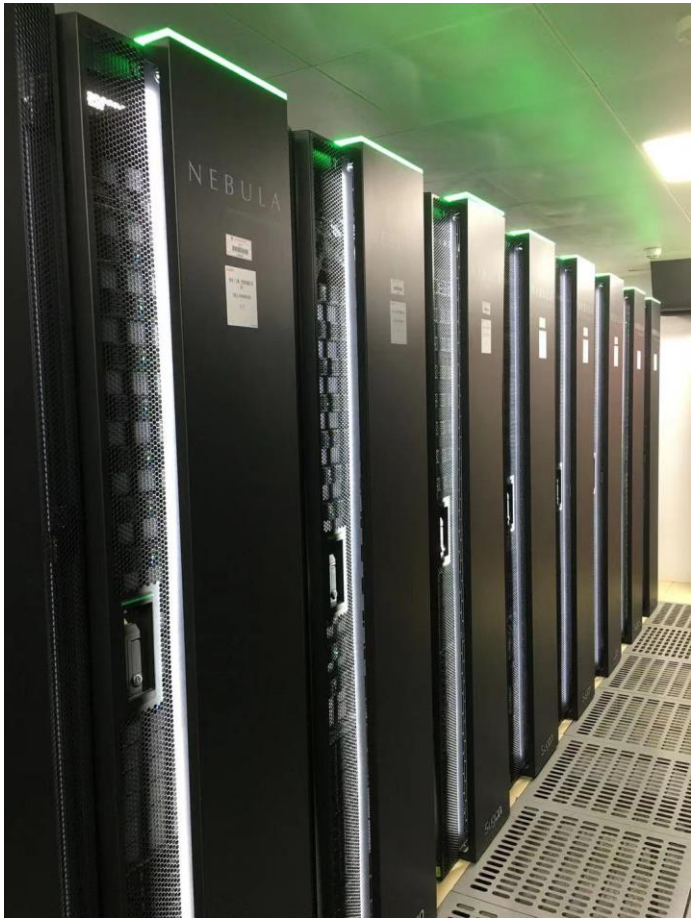
◆ 说话人识别/认同/验证情感分类:
(speaker recognition/identification/verification)

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。

1.4 研究内容

- http://www.ldmcnlp.com



1.4 研究内容

说明:

由于不同的研究方向所关注的侧重点不同，因此，一般将语音识别、语音合成和说话人识别等以语音信号为主要研究对象的语音技术独立出来，而其他以文本(词汇/句子/篇章等)为主要处理对象的研究内容作为自然语言处理的主体。

文字识别更多地涉及图像识别与理解的问题。信息检索与自然语言处理之间既有密切关联，又各自相对独立，我们暂且回避它们之间关系的争论。

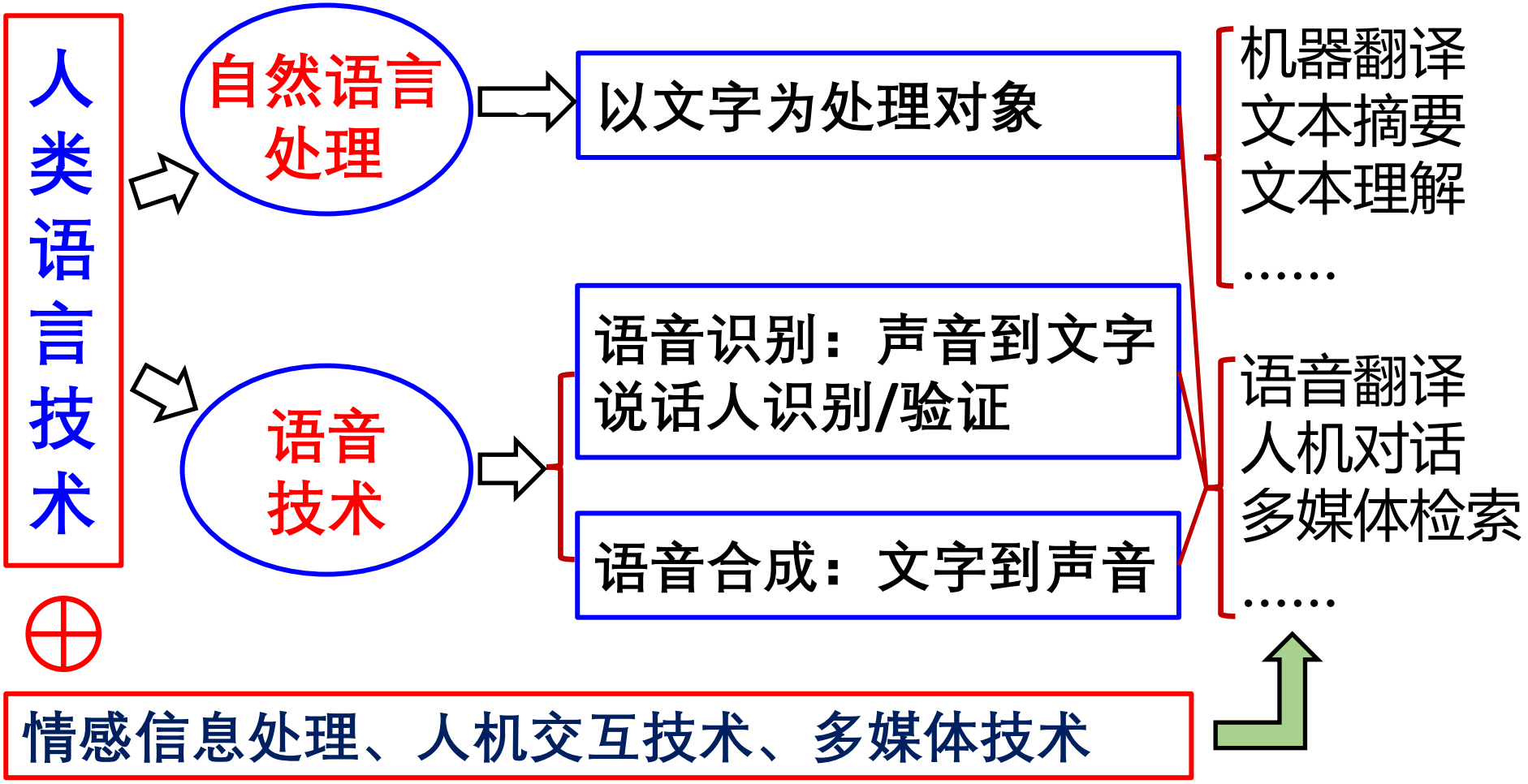
1.4 研究内容

◆很多研究方向密切相关：

- 机器翻译
- 信息检索
- 自动文摘
- 信息抽取
- 文本分类
- 舆情检测
- 人机对话
- 语音翻译
- ……

1.4 研究内容

◆ 自然语言处理用途：



1.5 基本问题和主要困难

1.5 基本问题和主要困难

◆基本问题一：形态学(Morphology)问题

又称词法，研究词(word)由有意义的基本单位—词素(morphemes)的构成问题，包括屈折变化和构词法两部分。

单词的识别 / 汉语的分词问题。

词素：词根、前缀、后缀、词尾

例如：人，蜈蚣；

老虎 ← 老 + 虎，图书馆 ← 图 + 书 + 馆

re + ex + port → reexport

1.5 基本问题和主要困难

◆ 基本问题二：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?

如何建立快速有效的句子结构分析方法?

苹果，我吃了。

我吃了苹果。

≠ 苹果吃了我。

1.5 基本问题和主要困难

◆基本问题三：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
(4) 火烧圆明园/驴肉火烧

1.5 基本问题和主要困难

◆基本问题四：语用学(Pragmatics)问题

研究在不同上下文中语句的应用，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。

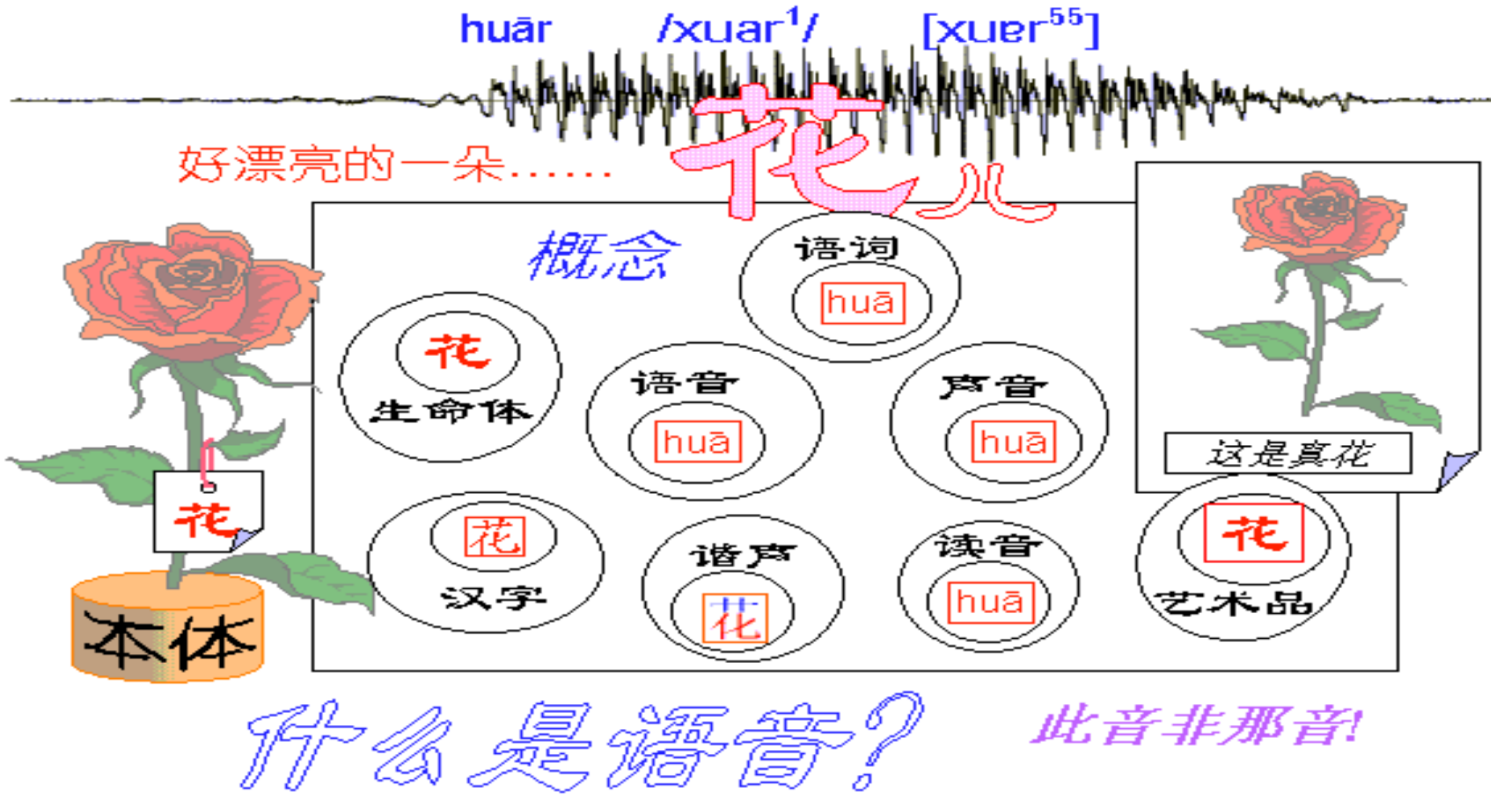
为什么要说这句话?

(1) 火，火!
(2) 看看鱼怎么样了?

1.5 基本问题和主要困难

◆基本问题五：语音学(Phonetics)问题

研究语音特性、语音描述、分类及转写方法等。



1.5 基本问题和主要困难

◆ 困难一：大量歧义(ambiguity)现象

➢ 词法歧义

例如：(1) I’ ll see Prof. Zhang home.

(2) 门把手弄坏了。

门/ 把/ 手/ 弄/ 坏/ 了/ 。

门把手/ 弄/ 坏/ 了/ 。

(3) 研究生命的起源。

研究/生命/的/起源。

研究生/命/的/起源。

1.5 基本问题和主要困难

◆ 困难一：大量歧义(ambiguity)现象

- 文章标题中的歧义比比皆是：

上大学子烛光追思钱伟长

(新浪网：http://www.sina.com.cn/, 2010.8.8)

教育部长跑活动负责人与商家总经理被曝系师生

(科学网：http://news.sciencenet.cn/,2010.11.14)

1.5 基本问题和主要困难

◆ 困难一： 大量歧义(ambiguity)现象

➢ 词性歧义

① 介词：像，好似；② 动词：喜欢

例如：(1) Time flies like an arrow.

① 动词：飞，飞翔，飞驰
② 名词：苍蝇，飞虫

时间像箭一样飞驰（光阴似箭）。

时间苍蝇喜欢箭（有一种苍蝇叫“时间”）。

(2) “动物保护警察”明年上岗

——（《环球时报》 2010年9月25日，第10版）

1.5 基本问题和主要困难

◆ 困难一： 大量歧义(ambiguity)现象

➢ 结构歧义

(1) 喜欢乡下的孩子。
(2) 关于鲁迅的文章。

(3) 今天中午吃馒头。 (4) 今天中午吃食堂。
(5) 今天中午吃大碗。 (6) 今天中午吃了闭门羹。

(7) 写文章/写毛笔/写黑板

1.5 基本问题和主要困难

◆ 困难一： 大量歧义(ambiguity)现象

➢ 结构歧义

例如：(8) Who has seen John?

(9) Who has John seen?

(10) 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.

1.5 基本问题和主要困难

◆ 困难一： 大量歧义(ambiguity)现象

➢ 结构歧义

歧义结构分析结果的数量是随着介词短语数目的增加呈指数上升，其歧义组合的复杂程度对着介词短语个数的增加而不断加深，这个结构的组合数开塔兰数(Catalan Numbers)，记作 C_n：

$$C_n = \binom{2n}{n}\frac{1}{n+1}$$

其中 $\binom{2n}{n} = \frac{2n!}{n!\times n!}$，$n$ 为句子中介词短语的个数。

1.5 基本问题和主要困难

◆ 困难一： 大量歧义(ambiguity)现象

➢ 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 基本问题和主要困难

◆ 困难一： 大量歧义(ambiguity)现象

➢ 语义歧义

人们的语言表达中大量使用缩略语和隐喻的表达方式，如：

要把权力装进制度的笼子

老虎苍蝇一起打

破四旧，除四害

消灭一切牛鬼蛇神

1.5 基本问题和主要困难

◆ 困难一： 大量歧义(ambiguity)现象

➢ 语音歧义

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。”

1.5 基本问题和主要困难

赵元任(1892-1982)：字宣仲，江苏五进人，1892年11月3日生于天津。1910年赴康奈尔大学学习数学，1914年获理学士学位。1918年获哈佛大学哲学博士学位。1919年任康奈尔大学物理学讲师。1920年回国任清华学校心理学及物理学教授。1921年再入哈佛大学研习语音学， 继而任哈佛大学哲学系讲师、 中文系教授。

1925年6月应聘到清华国学院任导师，指导范围包括“现代方言学”、“中国音韵学”、“普通语言学”等。 1929年6月底国学研究院结束后，被中央研究院聘为历史语言研究所研究员兼语言组主任，同时兼任清华大学中国文学系讲师，授“音韵学”等课程。他与梁启超、王国维、陈寅恪一起被称为清华“四大导师”。1938-1941年先后执教于夏威夷大学、耶鲁大学，之后任教于哈佛大学。1947-1962，任教于伯克利加州大学，讲授中国语文和语言学。

1.5 基本问题和主要困难

◆ 困难一： 大量歧义(ambiguity)现象

➢ 多音字及韵律等歧义

——语音合成面临的诸多问题

(1) 一字多音

例如：尾巴、亲家、削铅笔、一行

(2) 韵律、声调、语气、重音

例如：药材好药才好。

他的钱包被偷了。

今日说法

1.5 基本问题和主要困难

◆ 困难二： 大量未知语言现象

- 新词、人名、地名、术语等

例如：裸退、蜗居、非典、甲流、夏天、高山、温馨、时光、吉林、不来梅…… 布莱尔

- 新含义

例如：窗口、奔腾、同志、小姐、楼歪歪……

- 新用法和新句型等

在口语或部分网络语言中，不断出现一些“非规范的”新的语句结构。如：被长工资，很中国，百度一下……

1.5 基本问题和主要困难

◆ 归纳NLU所面临的挑战：

- **普遍存在的不确定性**：词法、句法、语义、语用和语音各个层面
- **未知语言现象的不可预测性**：新的词汇、新的术语、新的语义和语法无处不在
- **始终面临的数据不充分性**：有限的语言集合永远无法涵盖开放的语言现象
- **语言知识表达的复杂性**：语义知识的模糊性和错综复杂的关联性难以用常规方法有效地描述，为语义计算带来了极大的困难

1.5 基本问题和主要困难

- 归纳NLU所面临的挑战：
  - **机器翻译中映射单元的不对等性**：词法表达不相同、句法结构不一致、语义概念不对等

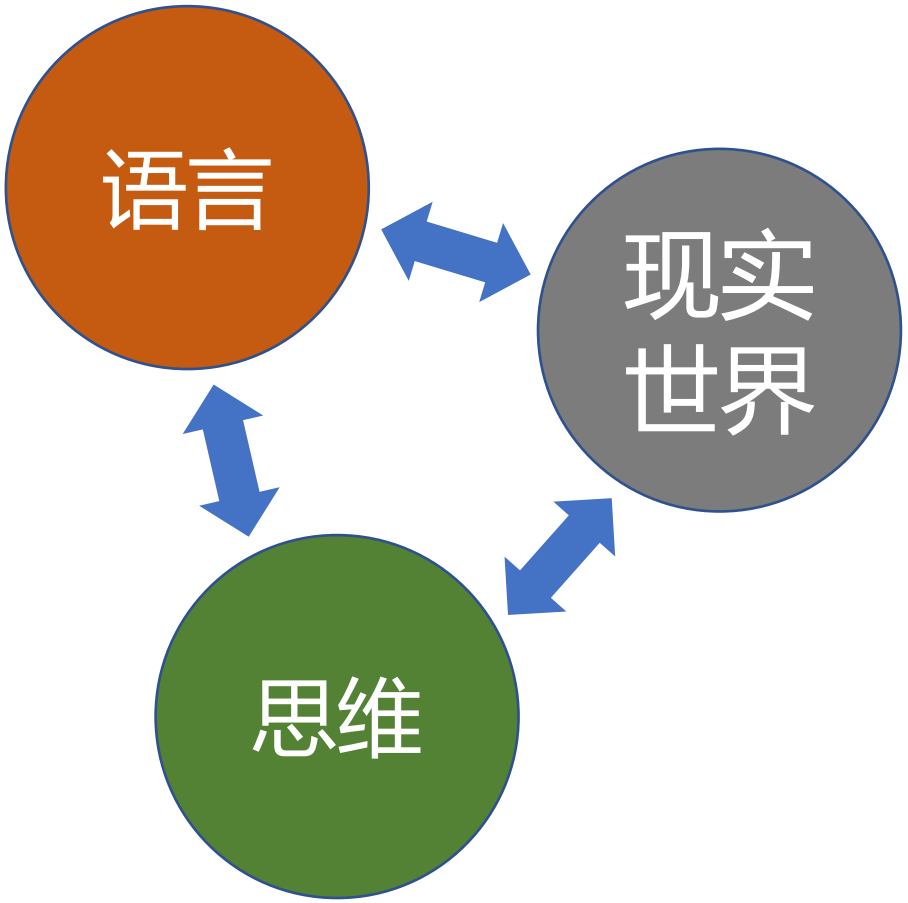
从大量复杂多样的不确定性中
寻找确定性结论

1.5 基本问题和主要困难

◆ 人脑理解语言是一个复杂的思维过程

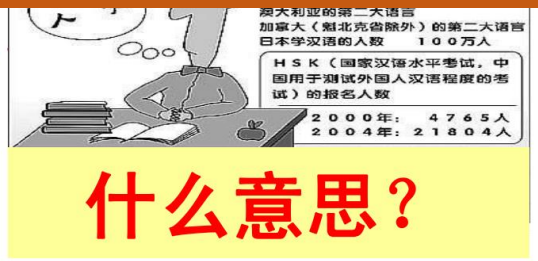
— 语言学、心理学
— 逻辑学、认知科学
— 计算机科学
— 统计学
— 背景知识、常识等

……



1.5 基本问题和主要困难

人脑的语言认知过程到底怎样？



什么意思？

1.6 基本研究方法

1.6 基本研究方法

◆ 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**：人很大一部分语言知识是与生俱来的，由遗传决定的。

诺姆•乔姆斯基（Noam Chomsky）的内在语言官能(innate language faculty) 理论被广泛接受。

1960s-1980s中期

1.6 基本研究方法

◆ 诺姆·乔姆斯基

—1928年12月生于美国费城

—1944年(16岁) 进入UPenn 学习哲学、语言学和数学

— 1949年获学士学位、1951年获硕士学位

— 1952 起在哈佛认知研究中心研究员，后来获UPenn博士学位；1957年(29岁) MIT副教授，32岁现代语言学教授、 47岁终生教授。

1.6 基本研究方法

◆ 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

- **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(泛化，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-

1.6 基本研究方法

◆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**：直接研究实际的语言数据。

1.6 基本研究方法

◆ 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**：通常基于香农(Shannon)的信息论。

1.6 基本研究方法

◆ 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。

1.6 基本研究方法

- 理性主义的问题求解方法：

  基于规则的分析方法，建立符号处理系统

  - 词典标注：#工作， N(uc)； V；
  - 规则库开发：N + N→NP
  - 推导算法设计：归约？推导？歧义消解方法？……

知识库 + 推理系统 → NLP 系统

理论基础：Chomsky 的文法理论

1.6 基本研究方法

- 经验主义的问题求解方法：

  基于大规模真实语料(语言数据)的计算方法

  - 大规模真实数据的收集、标注：真实性、代表性、标注信息 ……
  - 统计模型建立：模型的复杂性、有效性、参数训练方法……

语料库 + 统计模型 → NLP 系统

理论基础：统计学、信息论、机器学习

1.6 基本研究方法

- 以机器翻译为例：

  给定英语句子，将其翻译成汉语：

  There is a book on the desk.

➢ **基于规则的方法**

1. 对英语句子进行词法分析

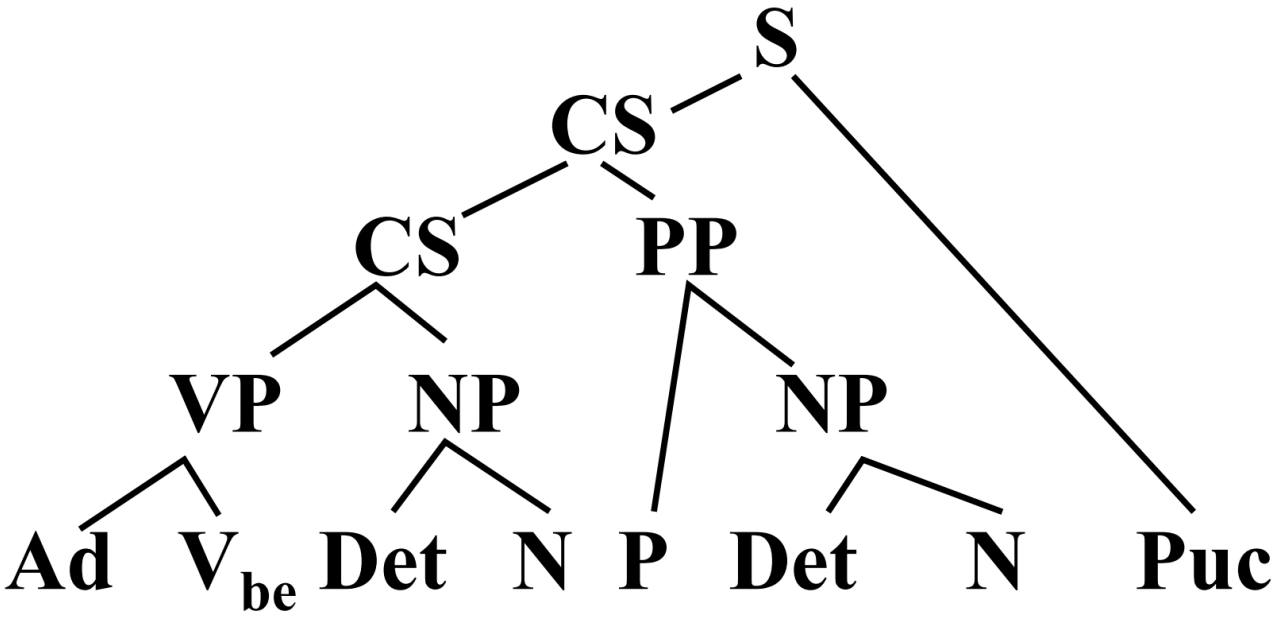
   There/Ad is/V be a/Det book/N

   on/P the/Det desk/N./Puc

1.6 基本研究方法

- **基于规则的方法**

2. 对英语句子进行句法结构分析

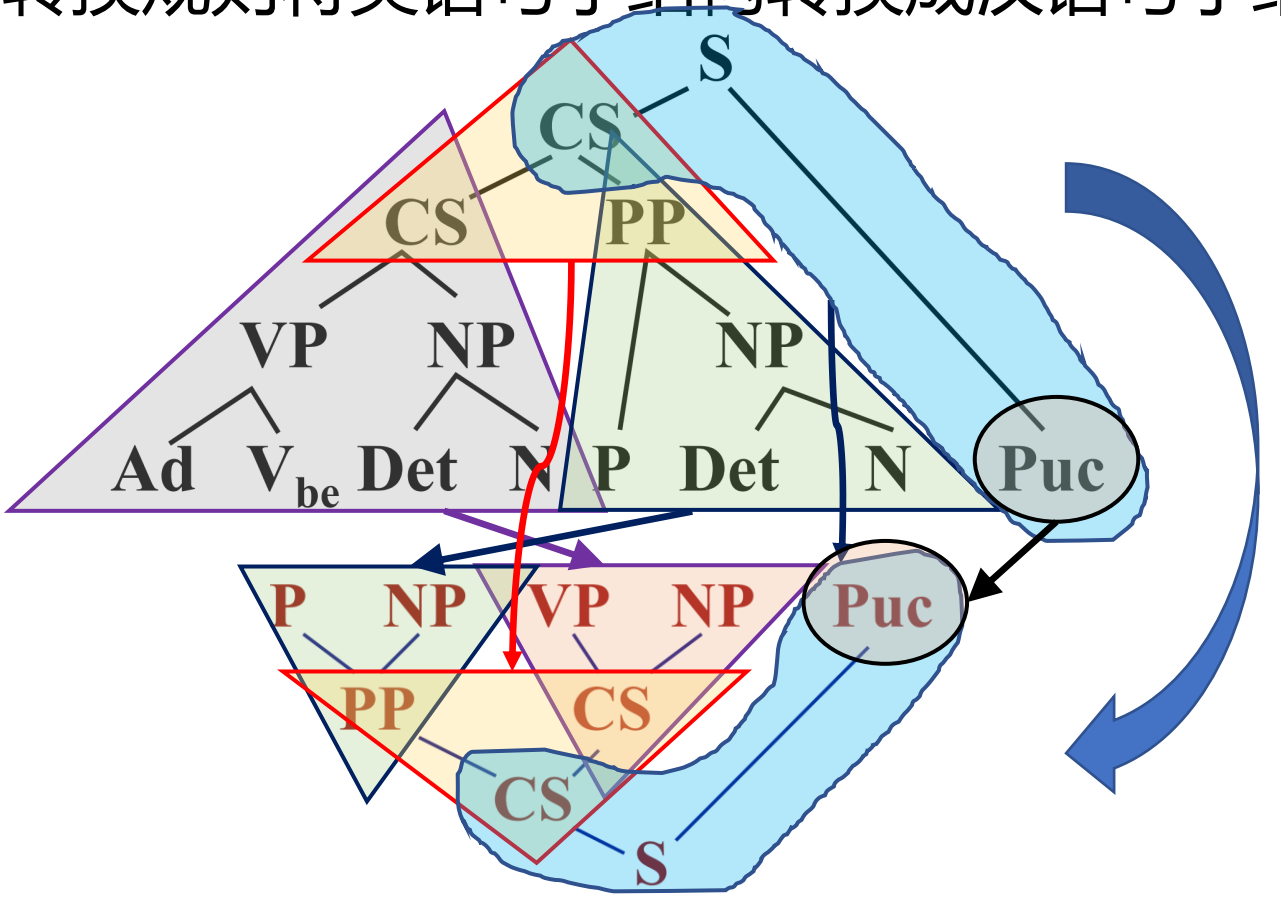


There is a book on the desk .

1.6 基本研究方法

➢ 基于规则的方法

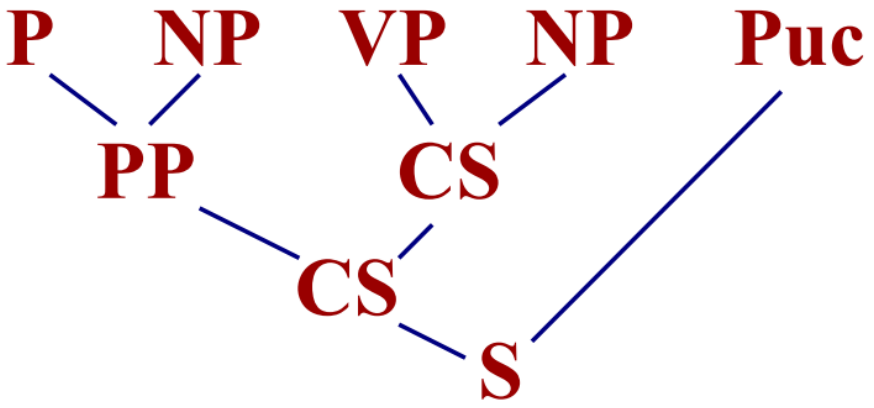
3. 利用转换规则将英语句子结构转换成汉语句子结构



1.6 基本研究方法

➢ 基于规则的方法

4. 根据转换后的句子结构，利用词典和生成规则生成翻译的结果句子。



#a, Det, —

#book, N, 书; V, 预订

#desk, N, 桌子

#on, P, 在 X 上

#There be, V, 有

输出译文：在桌子上有一本书。

1.6 基本研究方法

- **基于统计的方法**

$$E = e_1^m = e_1\ e_2 \cdots e_m \qquad C = c_1^l = c_1\ c_2 \cdots c_l$$

贝叶斯公式：$P(C|E) = \frac{P(C)P(E|C)}{P(E)}$

$$\hat{C} = arg\max_c P(C)P(E|C)$$

求解使 P 值最大的C

语言模型
(Language model, LM)

翻译模型
(Translation model, TM)

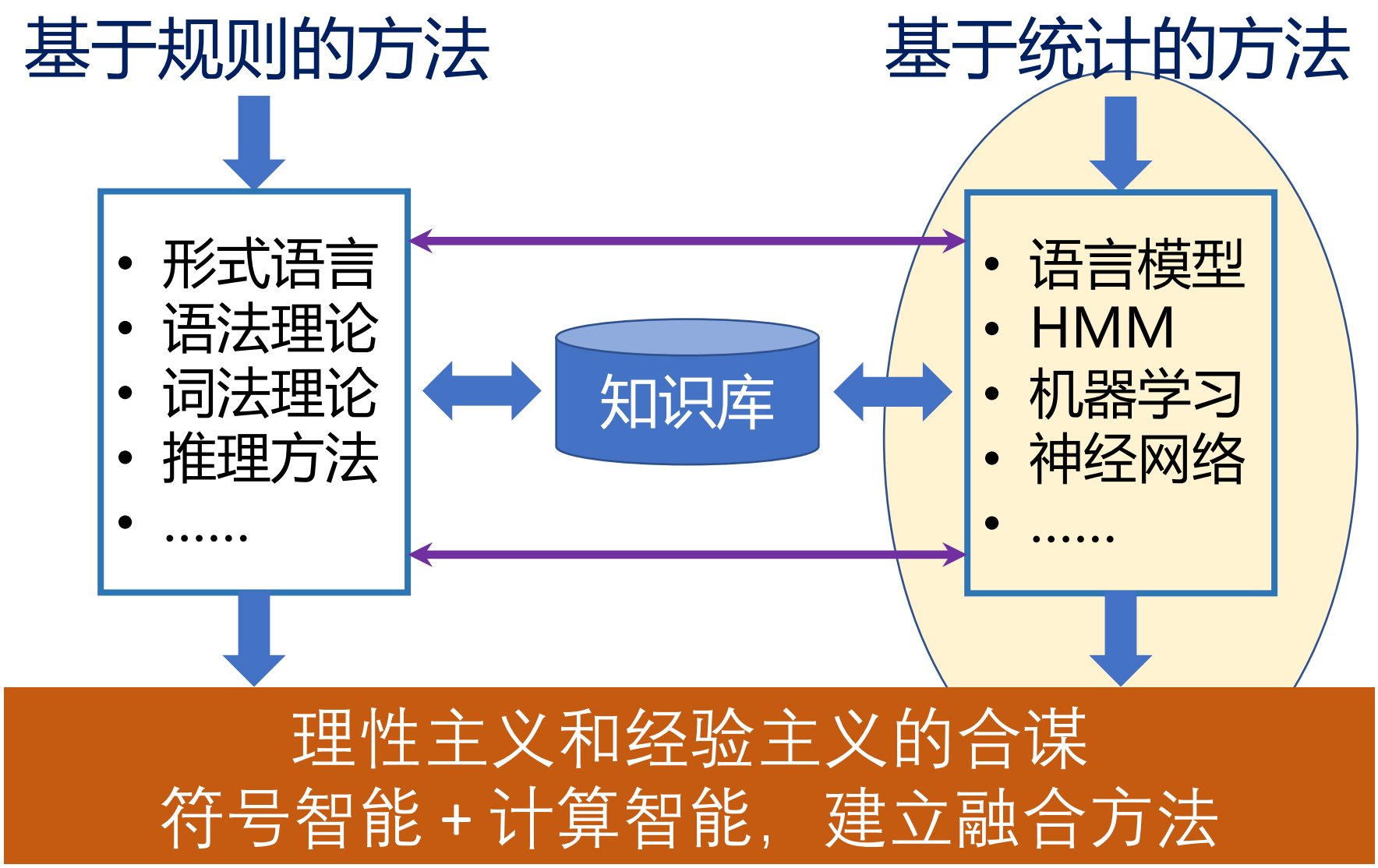
1.6 基本研究方法

- **基于统计的方法**

  主要工作：

  - 收集大规模双语句子对、目标语言句子
  - 参数训练与模型优化

1.6 基本研究方法


基于规则的方法
基于统计的方法
形式语言
语法理论
词法理论
推理方法
……
知识库
语言模型
HMM
机器学习
神经网络
……
理性主义和经验主义的合谋
符号智能 + 计算智能，建立融合方法

1.7 研究现状

1.7 研究现状

- 各种理论问题：

  从词法(汉语分词)到语义 ……

- 各种应用系统：

  从机器翻译到信息抽取 ……

哪个问题已经解决了？

哪个问题都没彻底解决！

1.7 研究现状

◆ 基本现状：

- 部分问题得到了解决，可以为人们提供辅助性帮助，如：专业领域文档翻译，电子词典，搜索引擎，文字录入等；
- 基础问题研究仍任重而道远，如：语义表示和计算、高质量的自动翻译、高精度的信息抽取等；
- 社会需求日益迫切：信息服务、通讯、网络内容管理、情报处理、国家安全等等。

1.7 研究现状

◆ 基本现状：

- 许多技术离真正实用的目标还有相当的距离，若干理论问题有待更深入的研究。

  —现有模型和方法的改进

  —在不成熟技术的基础上开发实用系统

  —期待更有效的理论体系

1.7 研究现状

1.8 国内外研究机构

1.8 国内外研究机构

◆ 国外

- Standford University, MIT, CMU, JHU, ISI, UPenn …
- IBM / Microsoft / Google / Yahoo / …
- Aachen University (RWTH), DFKI, Germany …
- ITC-irst, Italy …
- UPC, Spanish …
- 东京大学, 京都大学, 北海道大学，德岛大学， NICT 、富士通、东芝 …
- 新加坡国立大学， I^2R……

1.8 国内外研究机构

◆国内

- 一大批大学的计算机系、语言学系、研究中心等
- 中科院、社科院、教育部、科技部等若干研究所
- 华建、中软、阿里巴巴、腾讯、百度等公司
- 台湾、香港、澳门的一些大学和研究所

本章小结

- 基本概念：NLU, NLP, 计算语言学等
- 产生与发展
- 研究内容：机器翻译、信息检索...
- 基本问题：从词法到语用、语音
- 困难与挑战：歧义、未知现象 ...
- 研究方法：经验主义方法与理性主义方法
- 研究现状、国内外研究机构

思考题

1-1. 请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. 试比较汉英句子中地点状语位置的差异。

1-3. 下列语言中哪些为自然语言?

世界语、C语言、鸟语、甲骨文

谢谢!