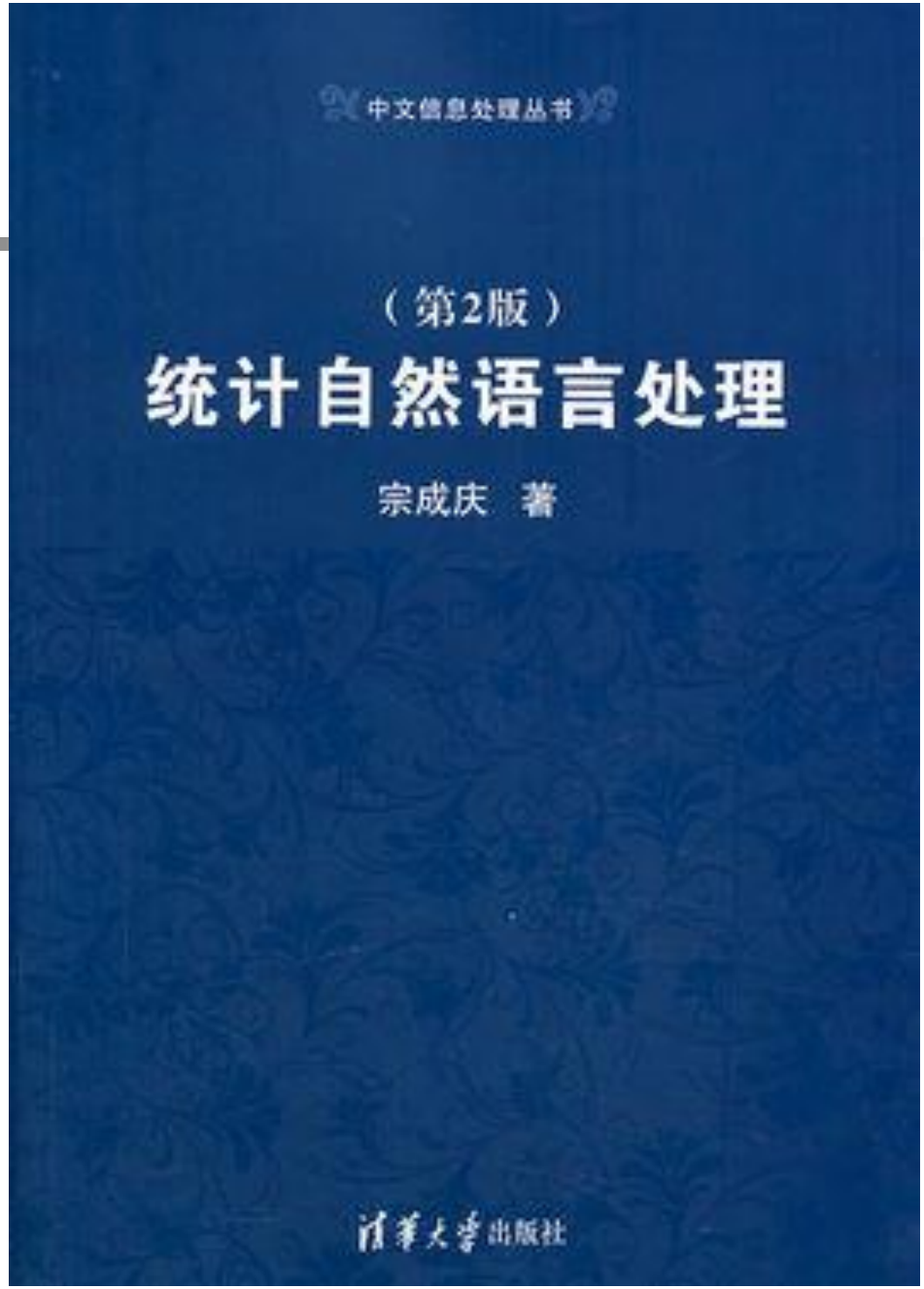
自然语言处理

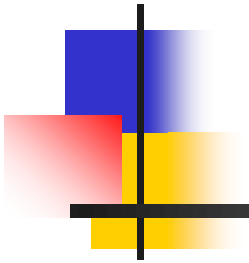
自然语言-NLP

话说普京与拜登赛跑，普京第一，拜登第二

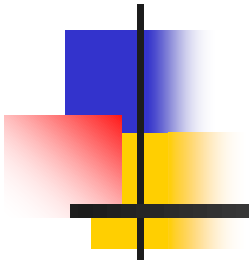
- 俄罗斯媒体
  - 在国际领导人赛跑中，普京勇夺冠军，拜登倒数第一！
- 美国媒体
  - 在国际领导人赛跑中，拜登勇夺亚军，普京倒数第二！

教材





第1章 绪论



1.1 问题的提出

1.1 问题的提出


Skype


微信


短信


邮件


WIFi
免费无线上网
FREE HOTSPOT ACCESS


WIFi
Free Hotspot
Access...
无线上网
自由自在


QQ


论坛


微博

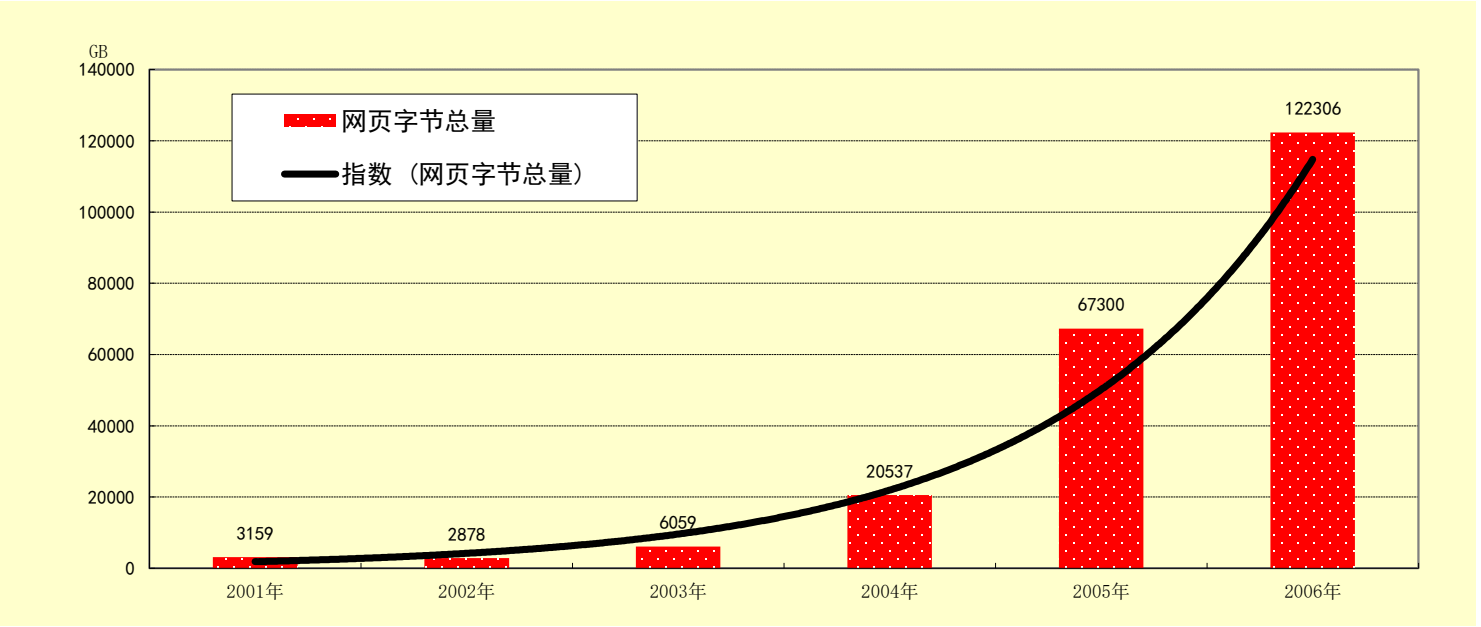
1.1 问题的提出

- 任意时间、任意地点、任意语言的自由通讯无时无刻不在改变着人们的思维方式和生活方式
- 语言是思维的载体，是人类交流思想、表达情感最自然、最直接、最方便的工具
- 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，中国互联网上有**87.8%**的网页内容是文本表示的
- 面对文本**大数据**，我们面临怎样的机遇和挑战？

1.1 问题的提出

网络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



◆中文网页检索的最高准确率不足40%

1.1 问题的提出

随着社会全球化时代的到来，机器翻译市场潜力巨大：

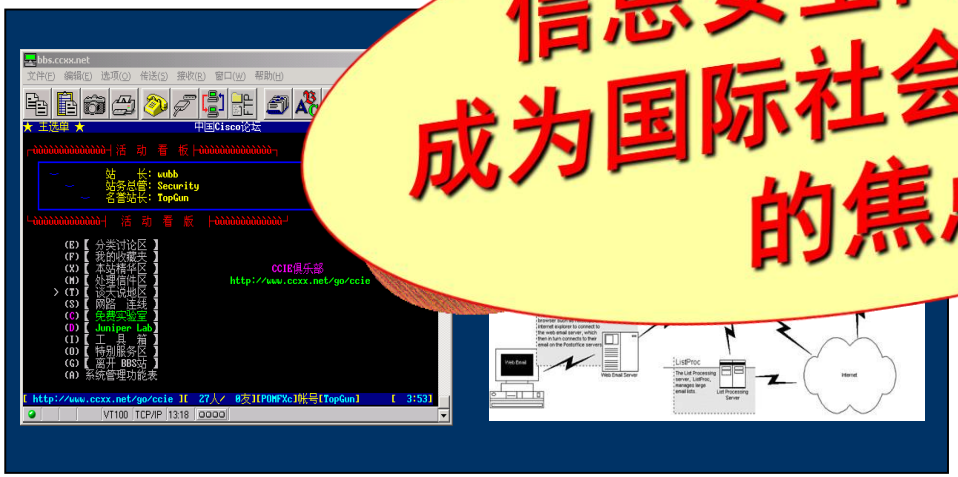
- 文化
- 商贸
- 旅游
- 体育
- ……

跨语言通讯和信息获取技术具有重要的用途

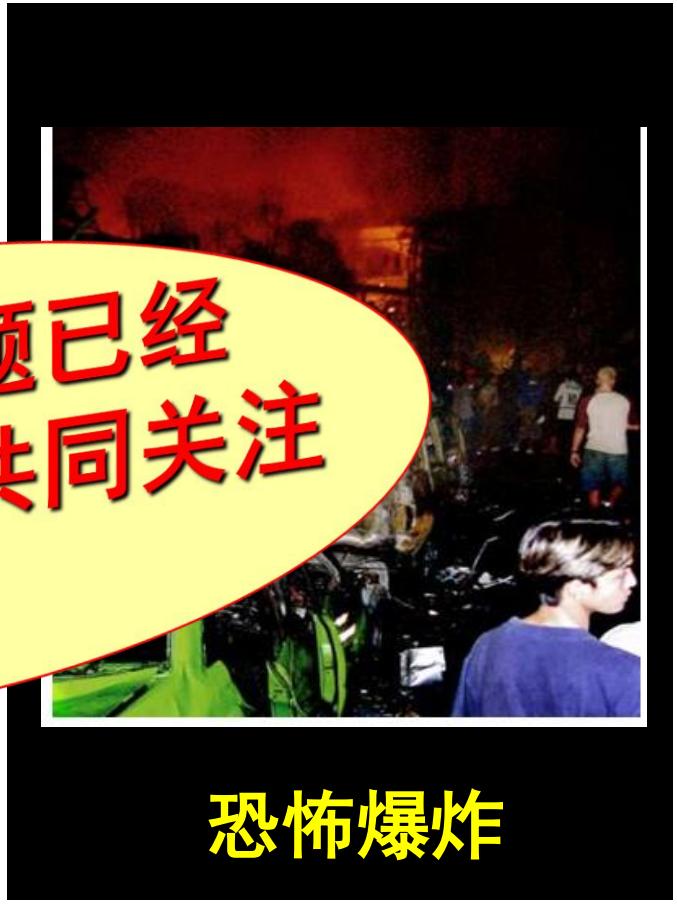
全世界正在使用的语言有1900多种.

1.1 问题的提出

利用网络组织犯罪，已成为恐怖活动的新特点

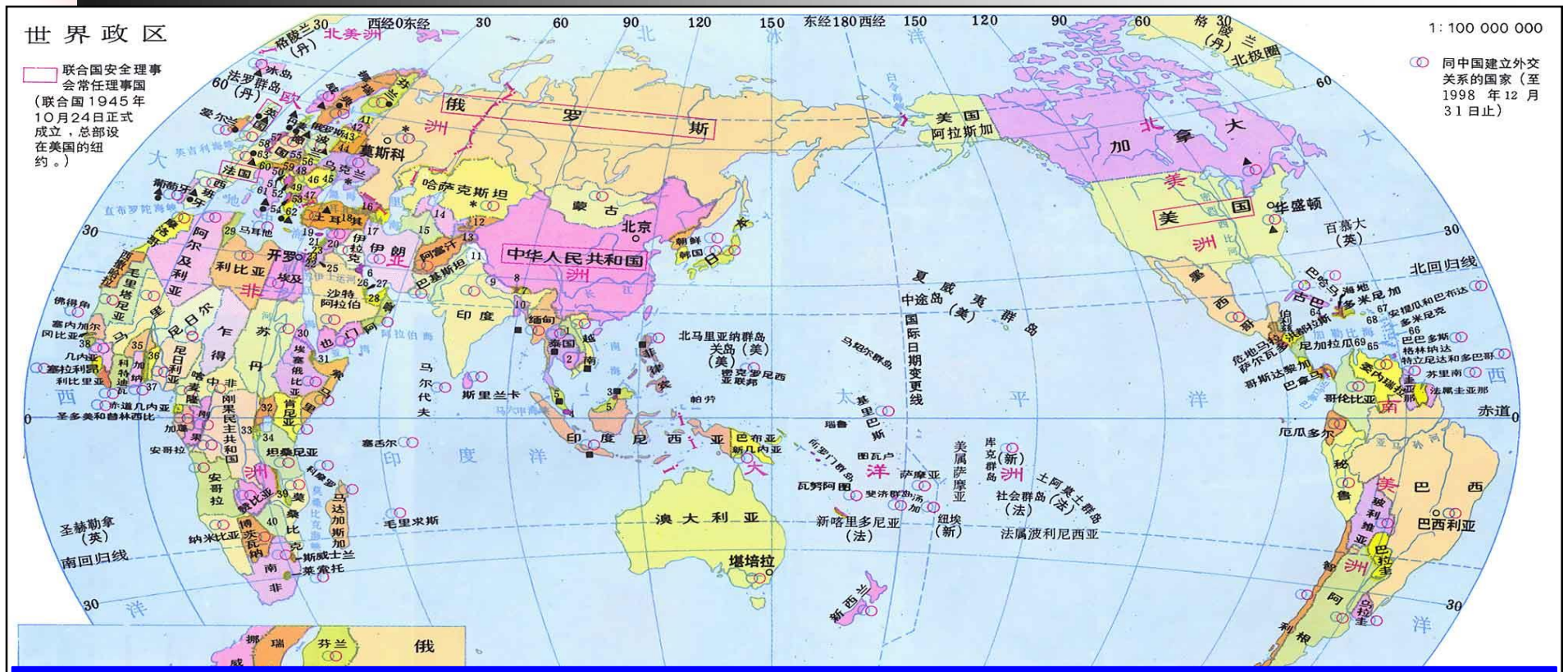


信息安全问题已经成为国际社会共同关注的焦点



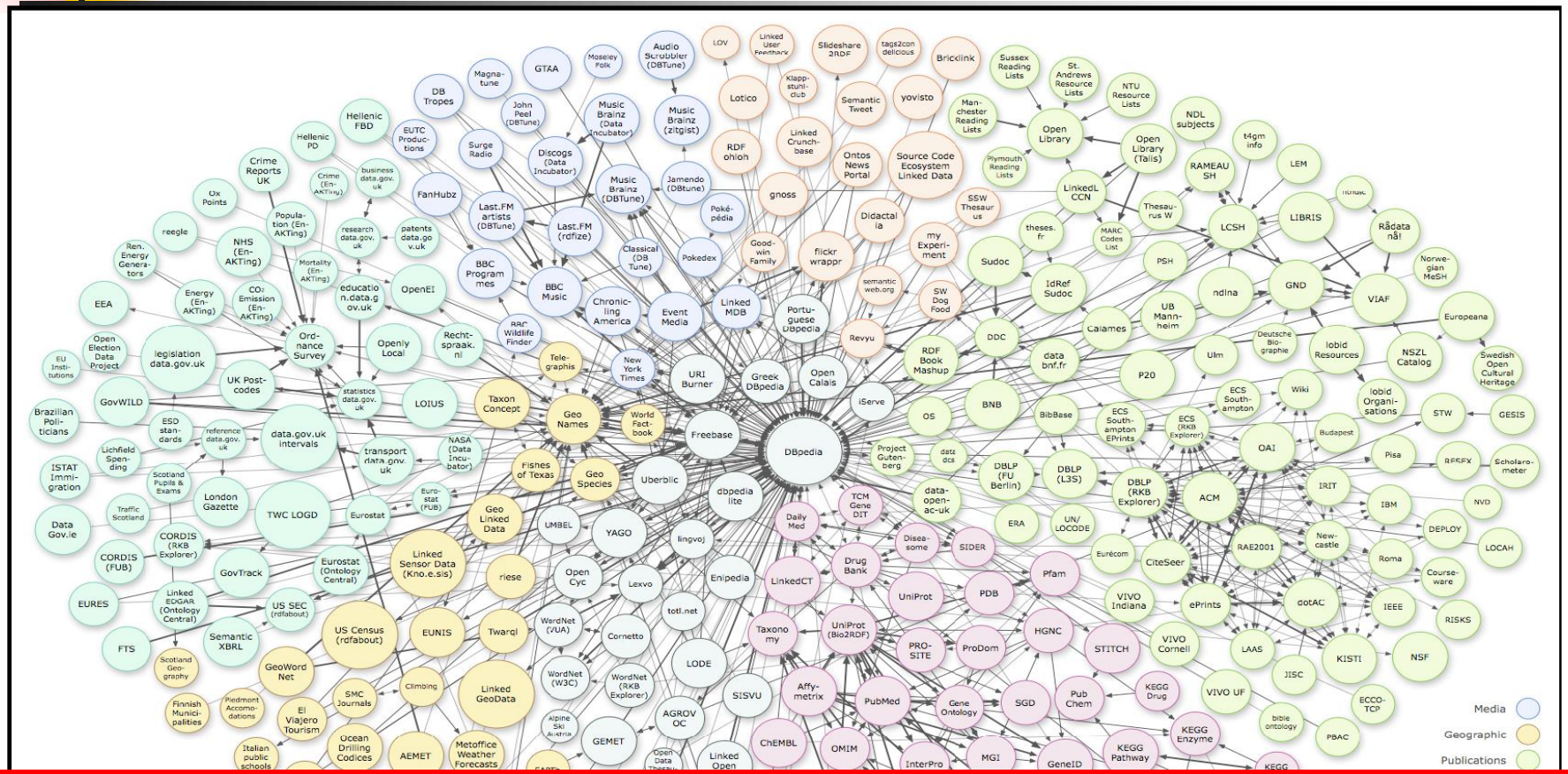
恐怖爆炸

1.1 问题的提出



人们处在不同的国家，使用不同的语言，在不同的地方发表过不同的言论(专著、论文、博客、网页等)，千丝万缕的关系将他们联系在一起，构成一个特定的社会网络。如何发现或挖掘这种网络？如何确定不同的实体、事件和知识之间的关联？

1.1 问题的提出



人工难以应对的数据“灾难”！

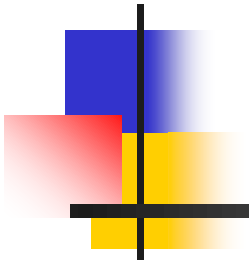
1.1 问题的提出


图片
文本
Flash
扫描文档
视频
肇事者撞人后自称扶摔倒老人被诬 监控证实说谎_新闻中心_新浪网 - 360安全浏览器 3.18 正式版
http://news.sina.com.cn/s/2012-03-08/233624083315.shtml
中新网
Chinanews.com
3月1日事发现场。左2为一再喊冤的广西“许云鹤”张都。 林增崇 摄
中新网南宁3月8日电 （林增崇 孙洁）被炒得沸沸扬扬的广西版“许云鹤”事件目前水落
称“学雷锋”扶倒地老人而被冤枉成肇事者的广西玉林男子，在交警部门展示其撞
终于不再四处喊冤，承认自己就是肇事者。
3月2日，天涯社区、广西红豆社区等论坛出现了《“许云鹤案”“再现广西？？》的贴子，作者自称“张都”，是广西玉林市博白县人，他以“做好事被交警和当事人家属冤枉为肇事者”为由发帖，寻求网上舆论支持：
“3月1日上午11点40分左右，我开着自家的面包车从玉林大北路西城停车场出来，在大门口我看见一个老人站在马路中间的隔离栏边
热门博客
“红泥人”的奇异婚俗
地狱般的印度陨石矿场
买个秦始皇的村官要花多少钱（图） Qing
毛泽东何时下决心打倒刘少奇（图）
彭德怀如何交代百团大战“罪行”
林彪要求对叶群是处女的两个证据
男人有外遇跟老婆不好没有关系
微博惊现卖萌考勤机 发微博曝光迟到者
太阳风暴袭击地球 GPS与飞机航行受影响
The new iPad发布 多项改进你会买吗
湖南卫视《锋尚之王》票选微公益
智投导购
探秘领导一个月
搞定英语
出国不用翻译
查看详情
教育
解酒保肝一不醉有秘诀
赚时尚男女财富的绝招
45岁前停经不正常！
做别人没想到的生意！
完成
开始
收件箱 - Outlook...
NSFCProposal2012
Microsoft PowerP...
360安全网址导航_...
新浪首页 - 360安...
肇事者撞人后自称...
驱动人生
20:27

1.1 问题的提出

- 如何让计算机能够自动或半自动地理解自然语言文本，懂得人的意图和心声？
- 如何让计算机实现海量语言文本的自动处理、挖掘和有效利用，满足不同用户的各种需求，实现个性化信息服务？

自然语言处理
Natural Language Processing, NLP



1.2 基本概念

1.2 基本概念

- 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：历时语言学 (diachronic linguistics)或称历史语言学 (historical linguistics)、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。

1.2 基本概念

◆ 定义1-2：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

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1-3：计算语言学 (Computational Linguistics)

通过建立形式化的计算模型来分析、理解和生成自然语言的学科，是人工智能和语言学的分支学科。计算语言学是典型的交叉学科，其研究常常涉及计算机科学、语言学、数学等多个学科的知识。与内容接近的学科**自然语言处理**相比较，计算语言学更加侧重基础理论和方法的研究。

《计算机科学技术百科全书》（常宝宝）

1.2 基本概念

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

自然语言理解是探索人类自身语言能力和语言思维活动的本质，研究模仿人类语言认知过程的自然语言处理方法和实现技术的一门学科。它是人工智能早期研究的领域之一，是一门在语言学、计算机科学、认知科学、信息论和数学等多学科基础上形成的交叉学科。

《计算机科学技术百科全书》（宗成庆）

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？

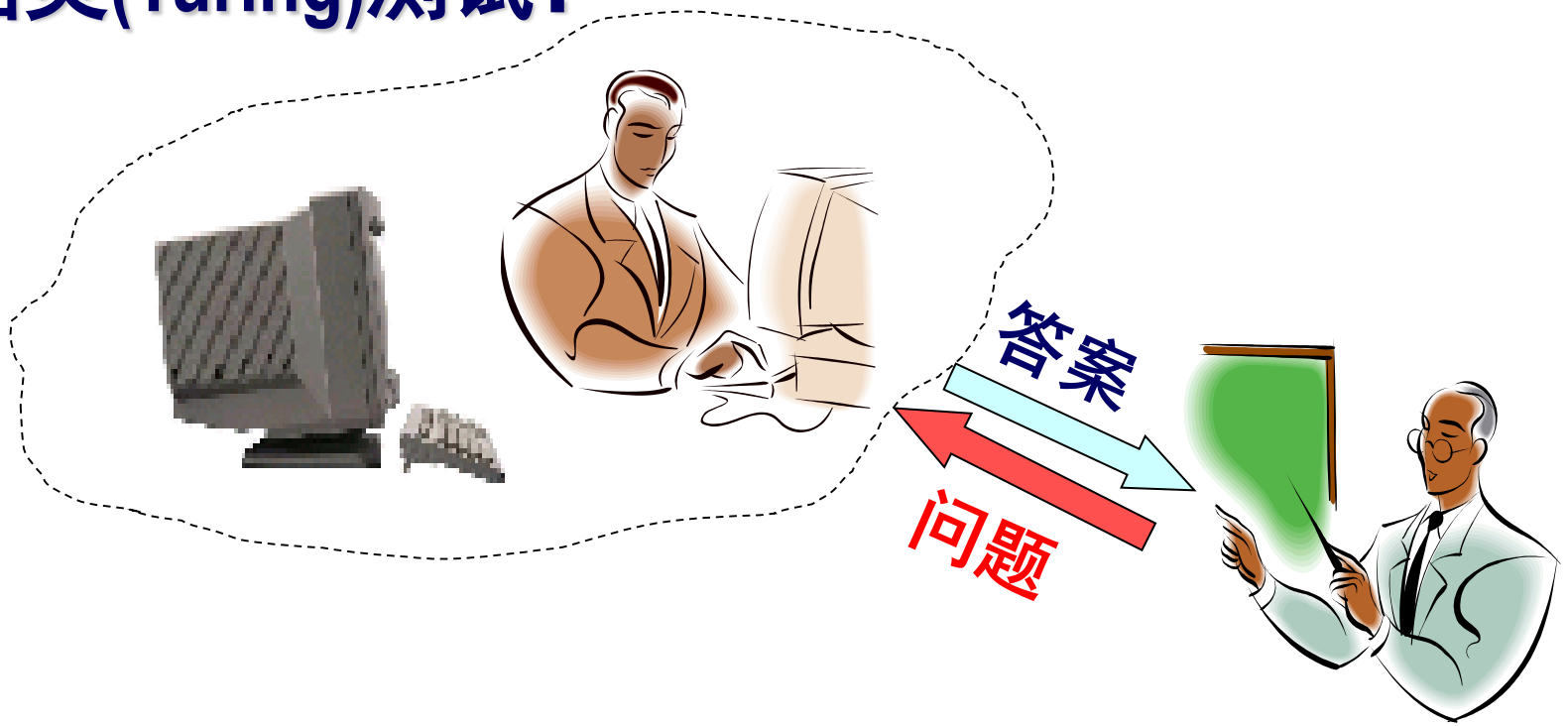
相互作用(interact)如何？

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(Turing)测试：



悄悄地提醒：关于图灵测试仍有争议。

1.2 基本概念

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

自然语言处理是研究如何利用计算机技术对语言文本（句子、篇章或话语等）进行处理和加工的一门学科，研究内容包括对词法、句法、语义和语用等信息的识别、分类、提取、转换和生成等各种处理方法和实现技术。

《计算机科学技术百科全书》（宗成庆）

1.2 基本概念

◆ 三个不同的语系

- **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。
- **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。
- **孤立语**(analytic language)(**分析语**，isolating 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。

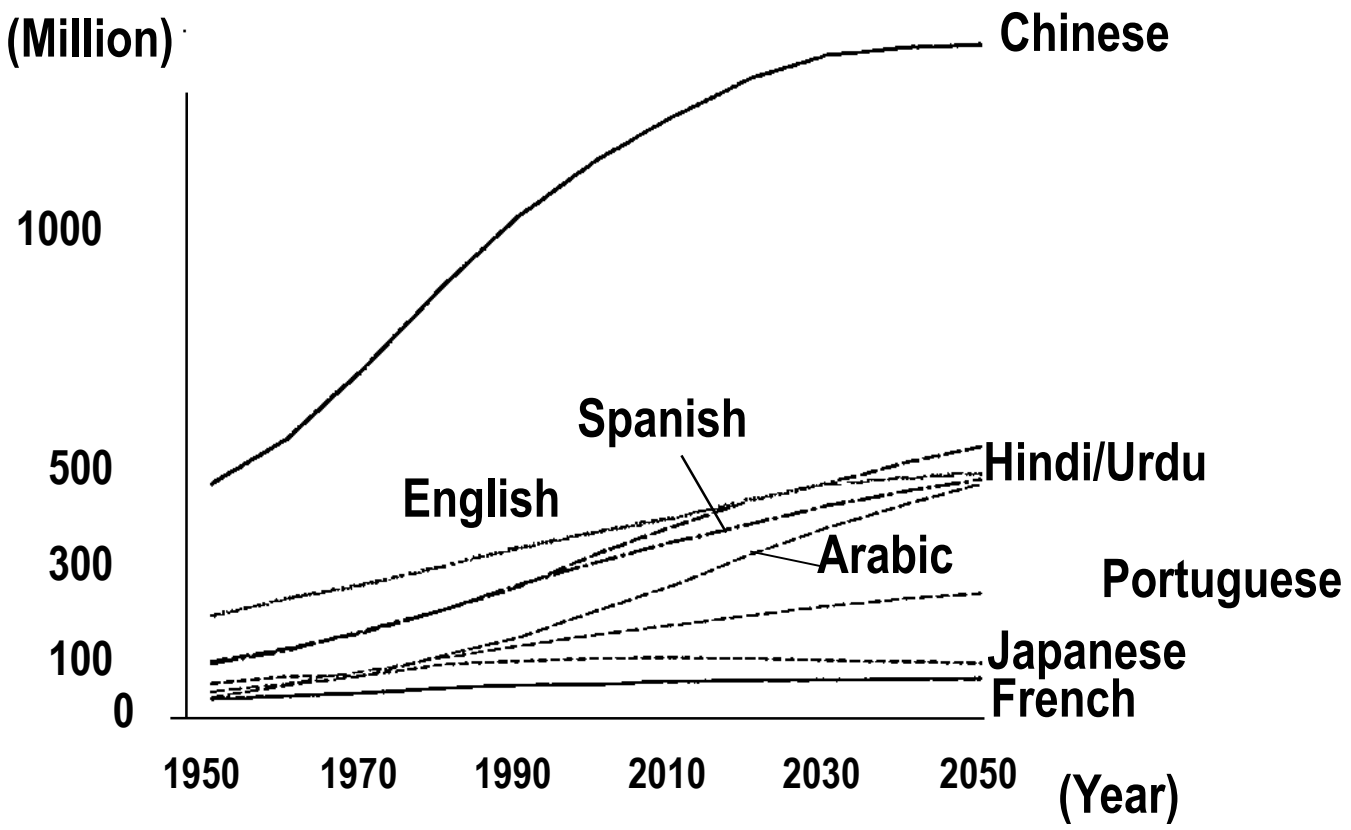
中文：中国的语言文字，特指汉族的语言文字。

-《现代汉语词典》，1996

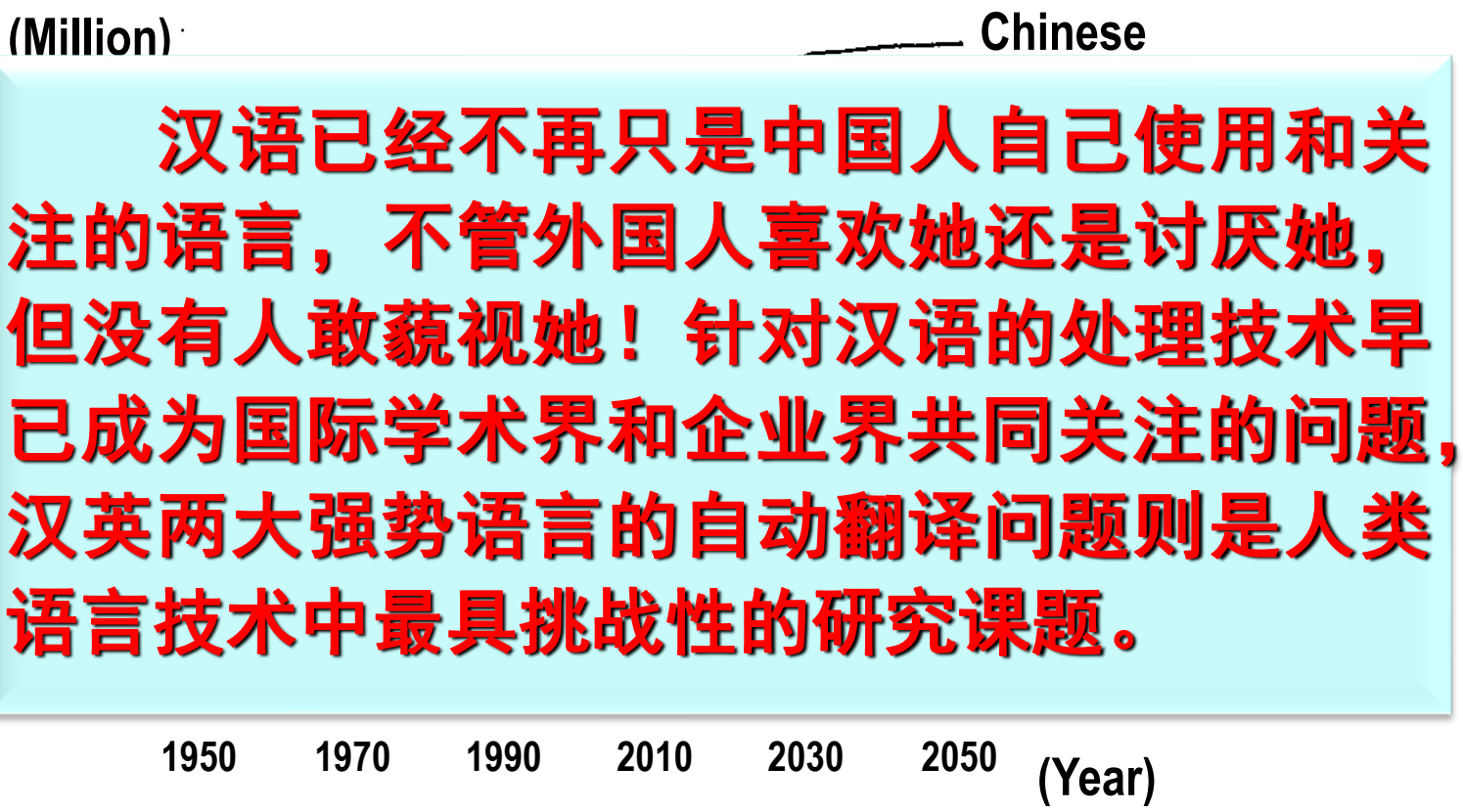
◆ **定义1-6：中文信息处理**
(Chinese Information Processing)

针对中文的自然语言处理技术。

1.2 基本概念


(Million)
Chinese
1000
Spanish
500
English
Hindi/Urdu
Arabic
300
Portuguese
100
Japanese
0
French
1950
1970
1990
2010
2030
2050
(Year)

1.2 基本概念



汉语已经不再只是中国人自己使用和关注的语言，不管外国人喜欢她还是讨厌她，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，汉英两大强势语言的自动翻译问题则是人类语言技术中最具挑战性的研究课题。

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。

人类语言技术

(Human Language Technology, HLT)

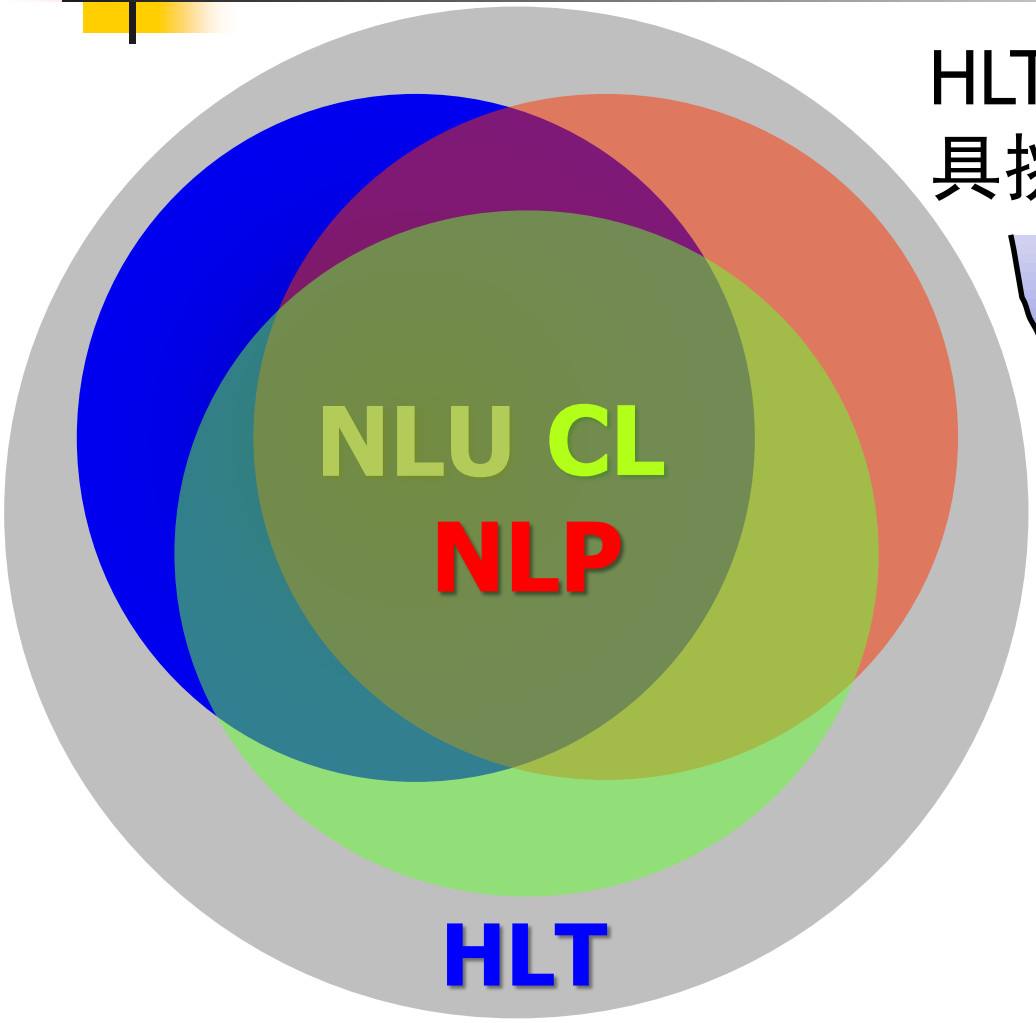
1.2 基本概念

- **自然语言理解**(natural language understanding, NLU)是人工智能最重要的研究方向之一，是当今“**人工智能皇冠上的明珠**”。
- **计算语言学** (Computational Linguistics, CL)

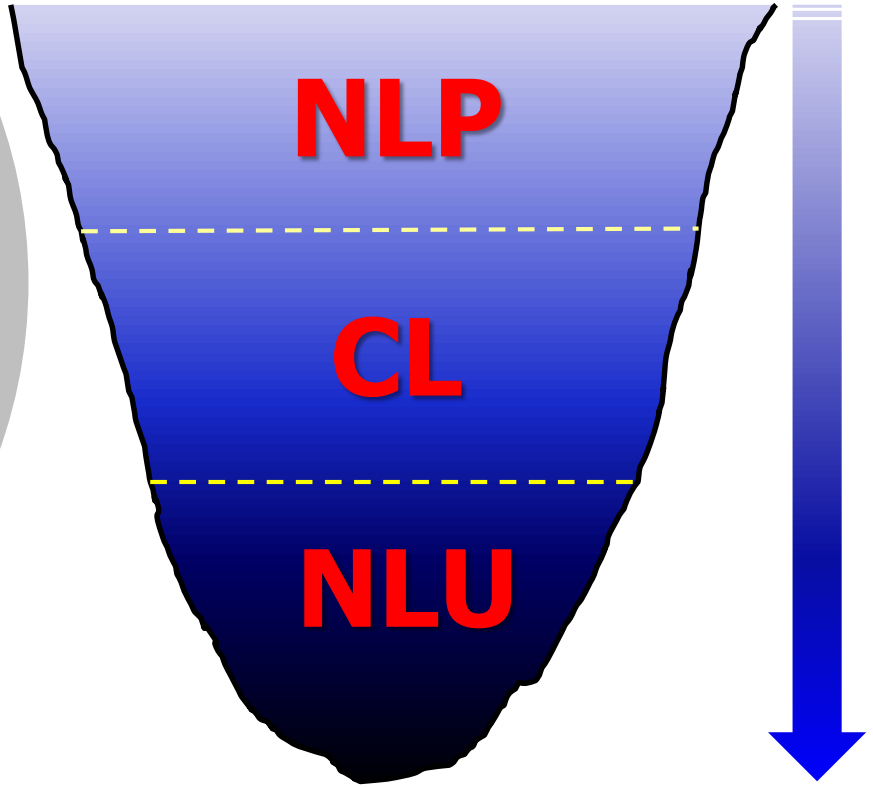
  1960S，形成相对独立的学科。1962年国际计算语言学学会(ACL)成立，1965年国际计算语言学委员会(ICCL)成立，1966年“计算语言学”首次出现在美国国家科学院ALPAC报告里。
- **自然语言处理** (Natural Language Processing, NLP)

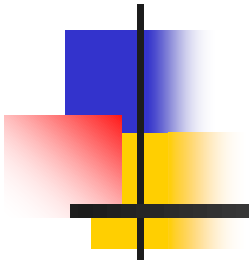
  1980S，面向计算机网络和移动通信，从系统实现和语言工程的角度开展语言信息处理方法的研究。专门针对中文的语言信息技术研究称为中文信息处理。

1.2 基本概念



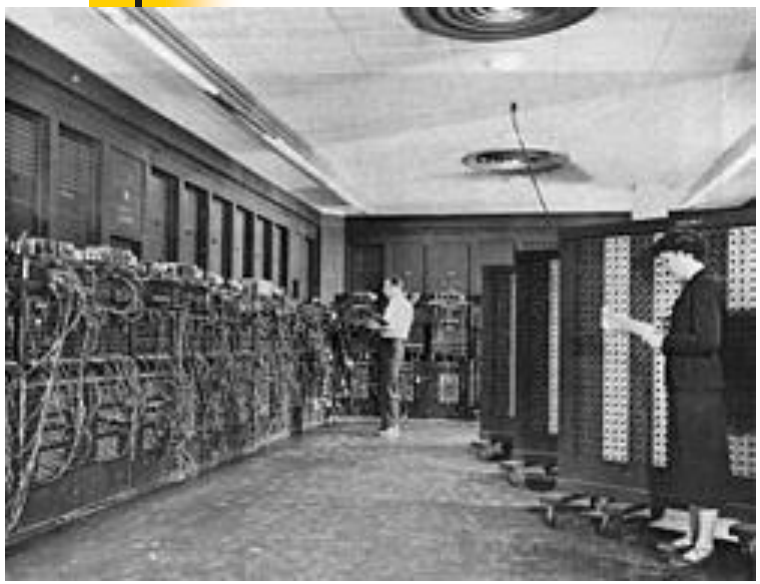
HLT 是当前人工智能领域最具挑战性的研究方向之一。





1.3 HLT的产生与发展

1.3 HLT的产生与发展



1946年，世界上第一台计算机ENIAC诞生

Warren Weaver (July 17, 1894 – Nov. 24, 1978)

- ✧ 信息论先驱
- ✧ 1920至1932年Wisconsin大学数学教授
- ✧ 1932至1955年担任Rockefeller Institute自然科学部主任

- ◆ A. D. Booth 数学物理学家，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。

1.3 HLT的产生与发展

诺伯特·维纳 (Norbert Wiener) (1894年11月26日 ~ 1964年3月18日)

[Reproduced by permission of the Rockefeller Foundation Archives]

March 4, 1947

Dear Norbert:

I was terribly sorry, when in Cambridge recently, that I got unavoidably held up by several unexpected jobs, and did not get a chance to see you.

One thing I wanted to ask you about is this. A most serious problem, for UNESCO and for the constructive and peaceful future of the planet, is the problem of translation, as it unavoidably affects the communication between peoples. Huxley has recently told me that they are appalled by the magnitude and the importance of the translation job.

I wondered if it were unthinkable to design a computer which would translate

Also knowing nothing official about, but having guessed and inferred considerable about, powerful new mechanized methods in cryptography - methods which I believe succeed even when one does not know what language has been coded - one naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say "This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode."

Have you ever thought about this? As P. linguist and expert on computers, do you think it is worth thinking about?

Cordially,

Warren Weaver.

Professor Norbert Wiener
Massachusetts Institute of Technology
Cambridge 39, Massachusetts

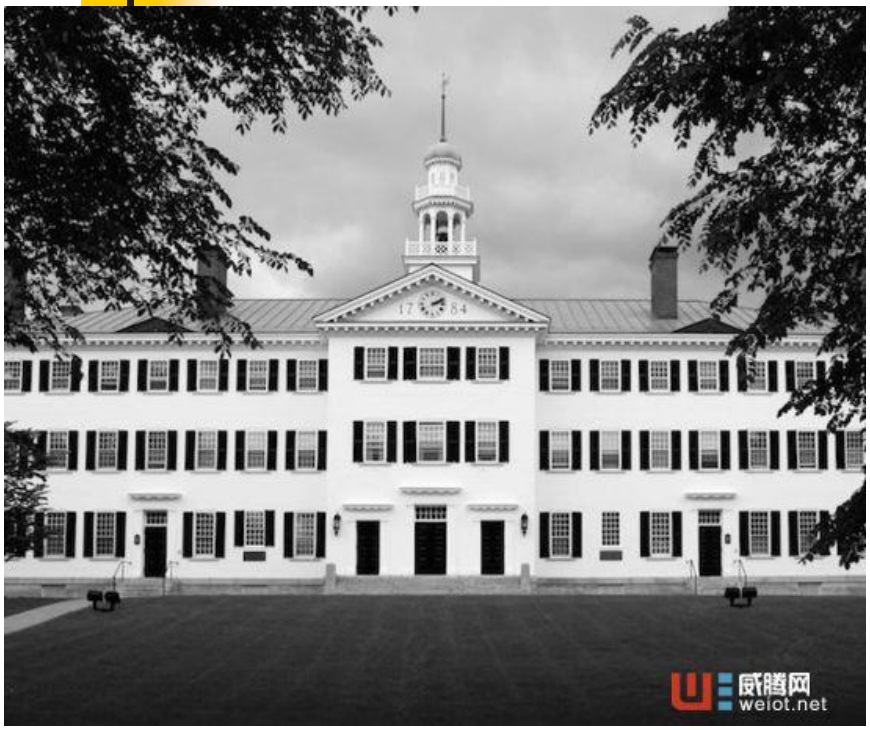
WW:AEB

1.3 HLT的产生与发展

- 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
- 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 HLT的产生与发展



达特茅斯学院 (Dartmouth College)
(成立于1769年)

左起：摩尔、麦卡锡、明斯基、赛弗里奇(Oliver Selfridge)、所罗门诺夫

人工智能夏季研讨会（大茅斯会议, 1956）

Summer Research Project on Artificial Intelligence (Dartmouth Conference)

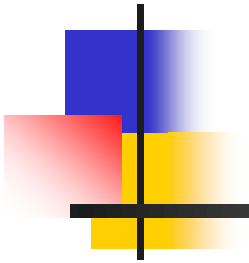
1.3 HLT的产生与发展

- 1962年美国成立“机器翻译和计算语言学协会(Association for Machine Translation and Computational Linguistics)”并组织召开了第一届国际计算语言学学术年会(ACL)
- 1965年杂志 *Machine Translation* 改名为 *Machine Translation and Computational Linguistics*
- 1965年成立国际计算语言学委员会 (The International Committee on Computational Linguistics, ICCL)，并组织召开了第一届国际计算语言学大会 (The International Conference on Computational Linguistics, COLING)
- 1966年术语 Computational Linguistics 正式出现在ALPAC (Automatic Language Processing Advisory Committee)中

1.3 HLT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至2010左右：快速发展
- 2010至今：繁荣时期



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：
  - 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。
  - 代表系统：
    - Google：http://translate.google.cn (110+ 种语言)
    - 百度： http://fanyi.baidu.com/ (200+种语言，包括文言文和简繁转换)
    - Systran：http://www.systransoft.com
    - 有道： http://fanyi.youdao.com/

1.4 研究内容

机器翻译研究现状和对机器翻译的认识

机器翻译研究在过去五十多年的曲折发展经历中，无论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)和人机互助，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]

1.4 研究内容

全球数万亿网页，
80%非汉语文字

出境游人数破亿，前
20出境游目的地
有12种语言

64个国家和地区
50多种语言
44亿人口

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。

▶代表系统： Google: http://www.google.com

百度： http://www.baidu.com.cn/

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。

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)

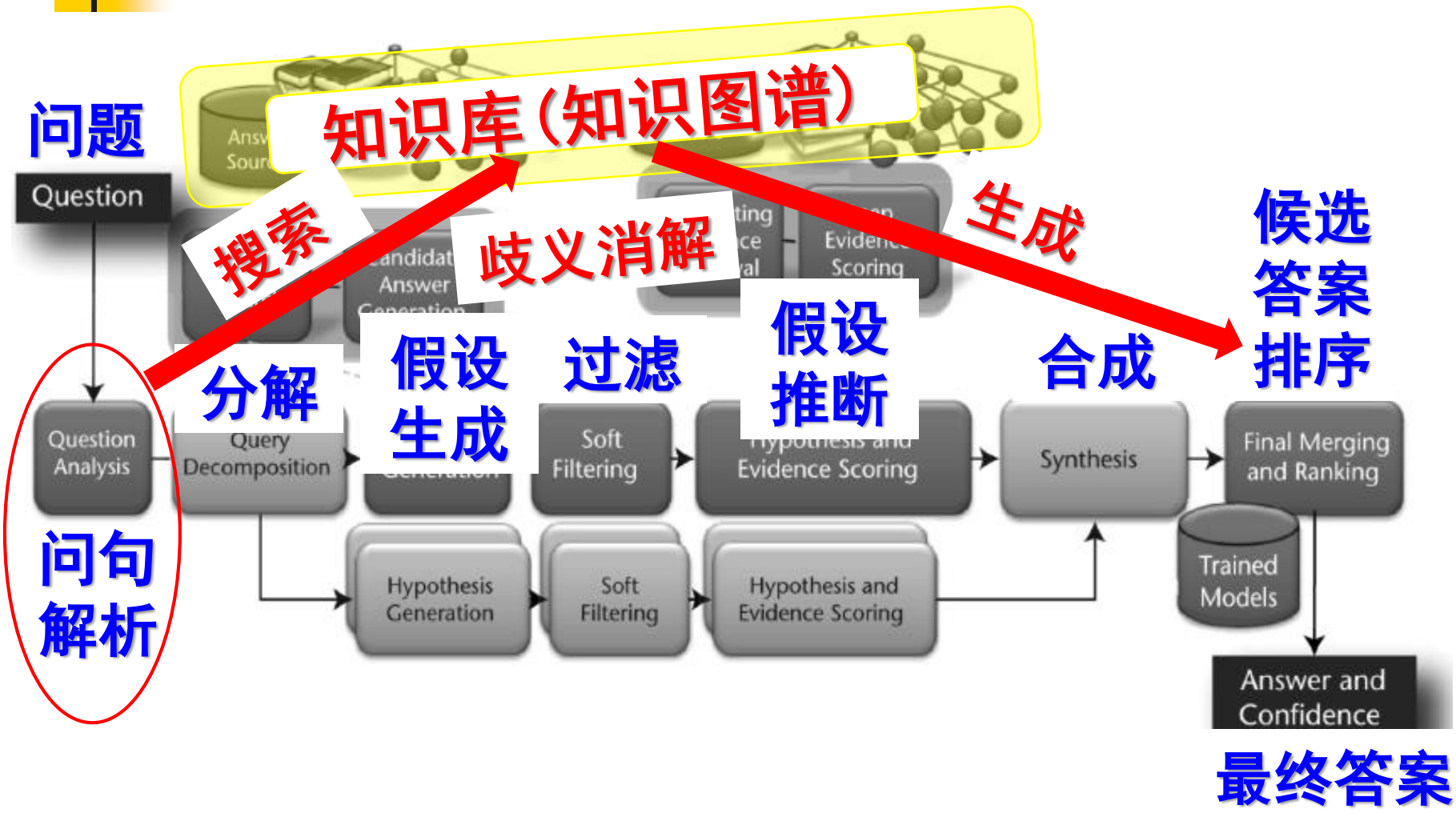
- **百度知道：**用户群体智慧
- **IBM Watson 自动问答系统**

1.4 研究内容

“沃森”(Watson)在2011年2月美国热门的电视智力问答节目“危险边缘”(Jeopardy!)中战胜了两位人类冠军选手。

D. Ferrucci *et al*., Building Watson: An Overview of the DeepQA Project. AI Magazine, Fall 2010, pp.59-79

1.4 研究内容


问题
Question
知识库(知识图谱)
搜索
歧义消解
生成
候选答案排序
分解
假设生成
过滤
假设推断
合成
Question Analysis
Query Decomposition
Soft Filtering
Evidence Scoring
Synthesis
Final Merging and Ranking
Hypothesis Generation
Soft Filtering
Hypothesis and Evidence Scoring
Trained Models
Answer and Confidence
问句解析
最终答案

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、语音翻译等等。
  - 困难：大量存在的同音词、近音词、集外词、口音等等。

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** (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 研究内容

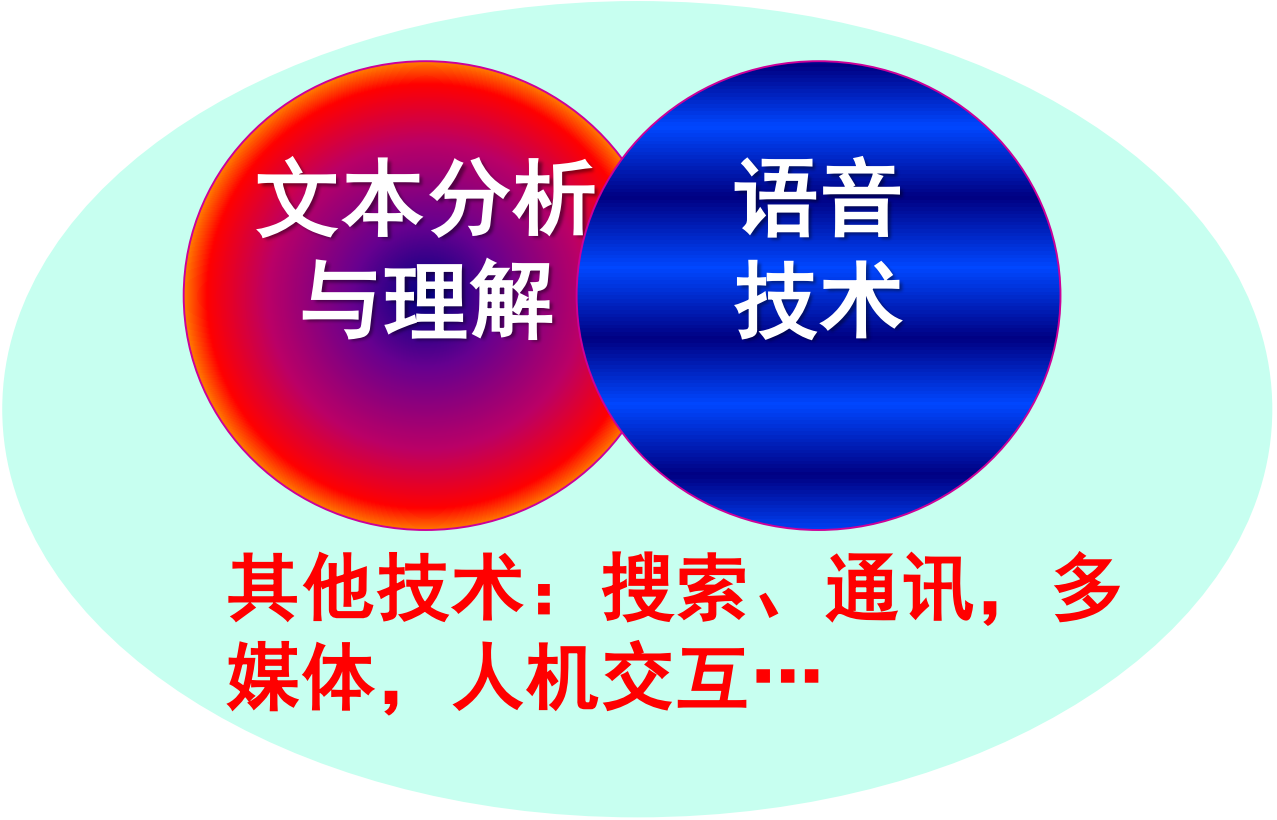
◆说明

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

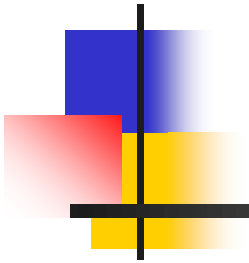
◆很多研究方向密切相关



人机对话

语音翻译

……



1.5 基本问题和主要困难

1.5 基本问题和主要困难

◆ 基本问题之一：形态学 (Morphology) 问题

研究词(word) 由有意义的基本单位 - 词素(morphemes)的构成问题。

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

◆ 基本问题之二：句法 (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

◆ 基本问题之三：语义 (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/火烧驴肉

1.5 基本问题和主要困难

◆ 基本问题之四：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。

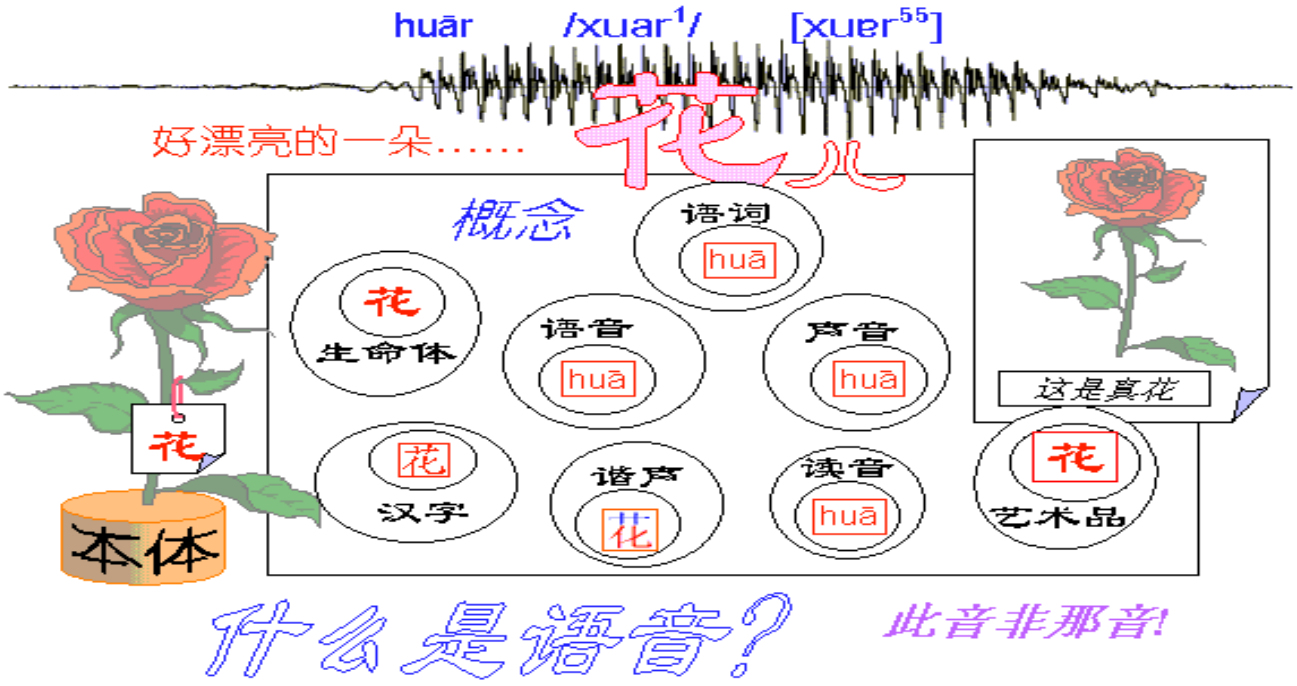
为什么要说这句话?

（1）火，火！
（2）看看鱼怎么样了？

1.5 基本问题和主要困难

◆ 基本问题之五：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

◆ 困难之一：大量歧义(ambiguity)现象

❖ 词法歧义

例如: (1) I'll see Prof. Zhang home.

(2)自动化研究所取得的成就

自动化/研究所/取得/的/成就

自动化/研究/所/取得/的/成就

(3)门把手弄坏了

门/ 把/ 手/ 弄/ 坏/ 了

门把手/ 弄/ 坏/ 了

1.5 基本问题和主要困难

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

❖ 词性歧义

①介词：像，好似；②动词：喜欢

(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

❖ 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

(8) 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. ?

英语句子歧义组合的开塔兰数(Catalan Numbers) C_n:

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

***n* 为句子中介词短语的个数。**

1.5 基本问题和主要困难

❖ 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6版

人们的语言表达中大量地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

❖ 语音歧义：大量同音现象

施氏食狮史

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。

赵元任(1892-1982)

1892年11月3日生于天津。1914年获康奈尔大学数学学士学位。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，任哈佛大学哲学系讲师、中文系教授。与梁启超、王国维、陈寅恪并称为清华“四大导师”。1938~41年先后执教于夏威夷大学、耶鲁大学，之后任教于哈佛大学。1947~62，任教于伯克利加州大学，讲授中国语文和语言学。1982年2月24日逝世，享年90岁。

1.5 基本问题和主要困难

- **多音字及韵律等歧义**

一 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/小心地滑/聊吧/说吧

1.5 基本问题和主要困难

◆ 困难之二：大量未知语言现象

❖ 新词、人名、地名、术语等，如：裸退、非典、夏天、高山、温馨、时光、吉林、失联、冰墩墩、葱桶

❖ 新含义

如：苹果、奔腾、同志、小姐、老虎、苍蝇等

❖ 新用法和新句型等，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。如：被长工资，很中国，百度一下

1.5 基本问题和主要困难

◆ 归纳起来，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

- **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

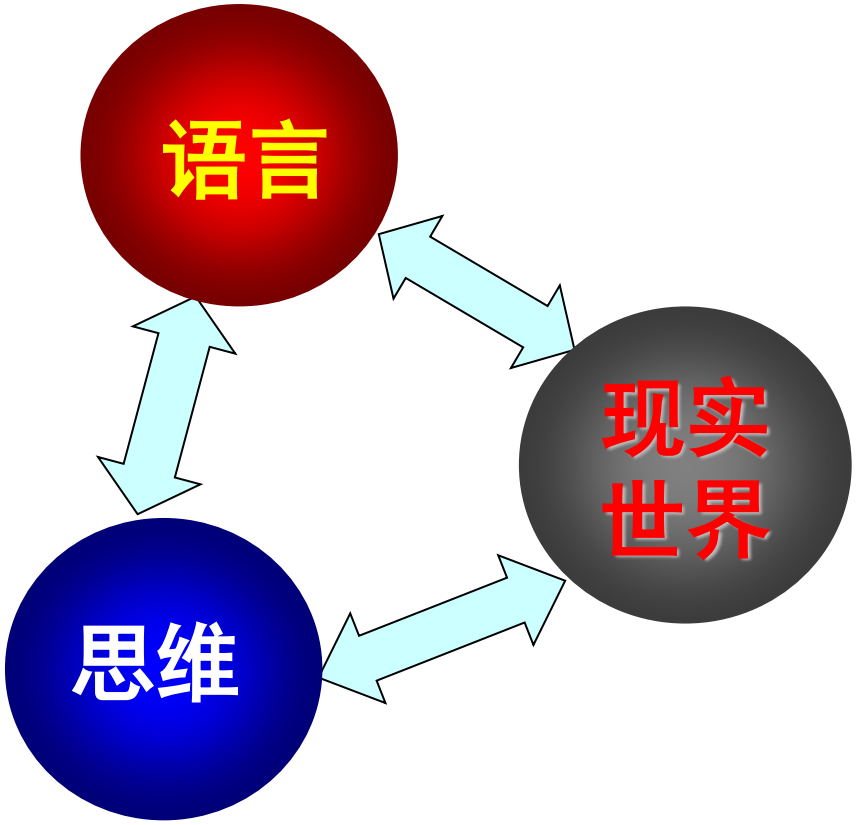
从大量复杂多样的不确定性中寻找确定性结论

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

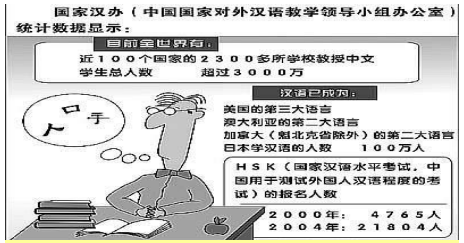
- 语言学、心理学
- 逻辑学、认知科学
- 计算机科学
- 统计学、信息论
- 背景知识、常识等

……



1.5 基本问题和主要困难

爸爸在说什么？

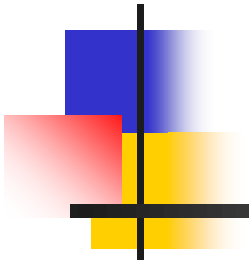


什么意思？

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：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。

- 问题求解的基本思路：基于规则的分析方法建立符号处理系统
  - 规则库开发：N + N → NP
  - 词典标注：　#工作，N(uc)；V；
  - 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：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。

- 求解问题的思路：基于大规模真实语料(语言数据)建立计算方法
  - 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：

There is a book on the desk.

将其翻译成汉语。

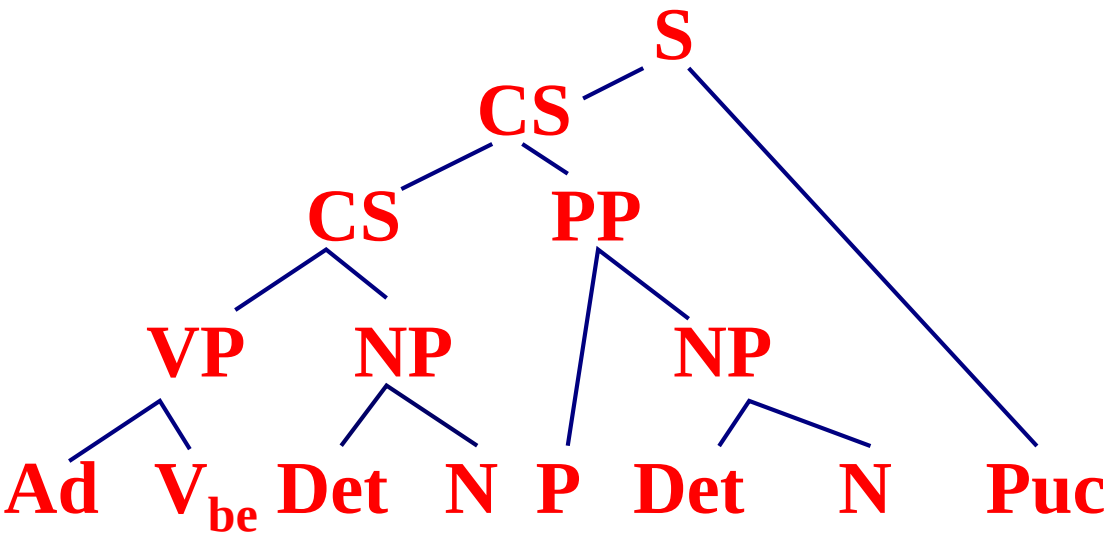
1.6 基本研究方法

➢ 基于规则的方法

■词法分析：

There/Ad is/V_{be} a/Det book/N on/P the/Det desk/N ./Puc

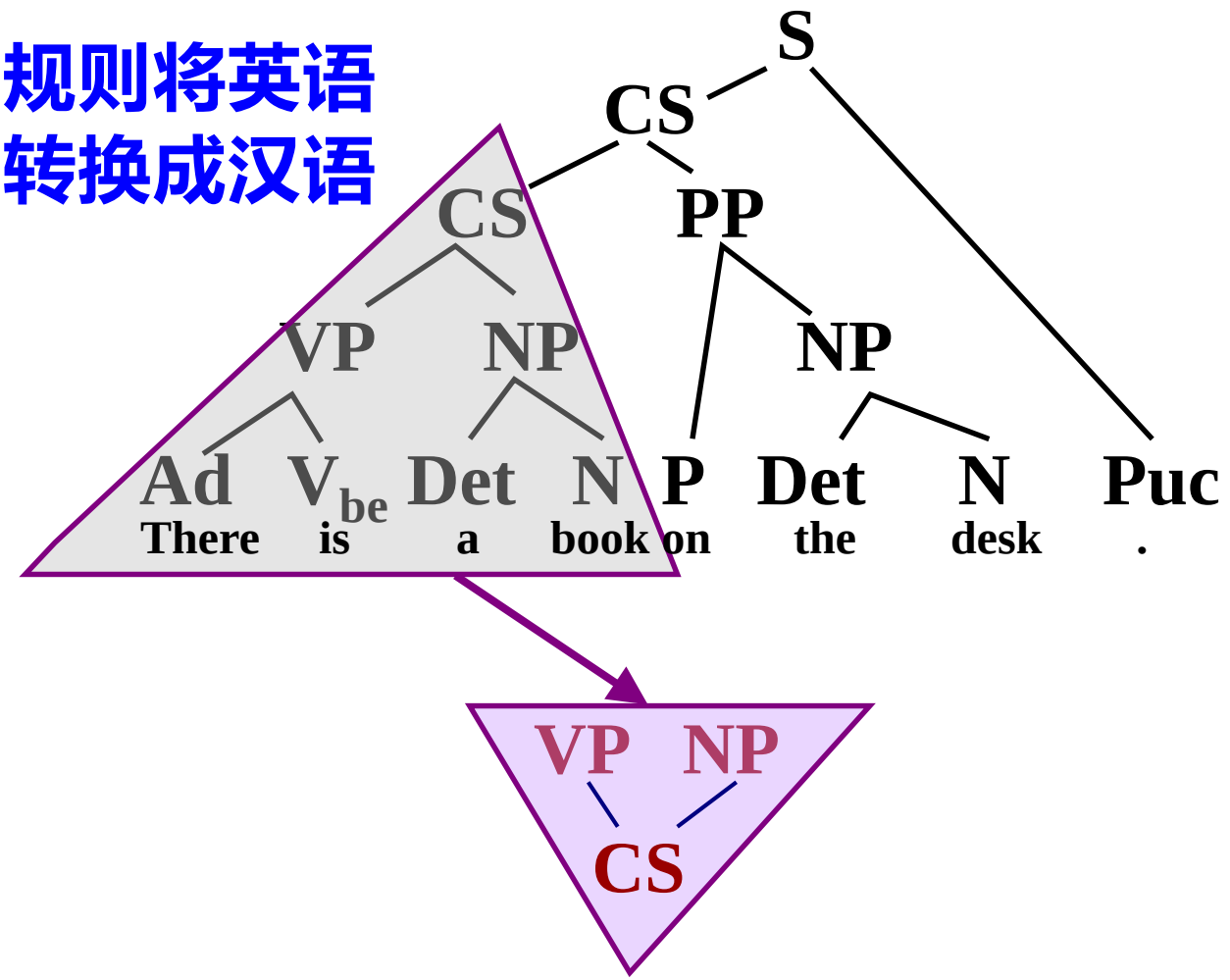
■利用句法规则进行句法结构分析：



动词短语（verb phrase, VP）
名词短语（noun phrase, NP）
介词短语（preposition, PP）
2类连词（conjunction, 分别记作： CC, CS）

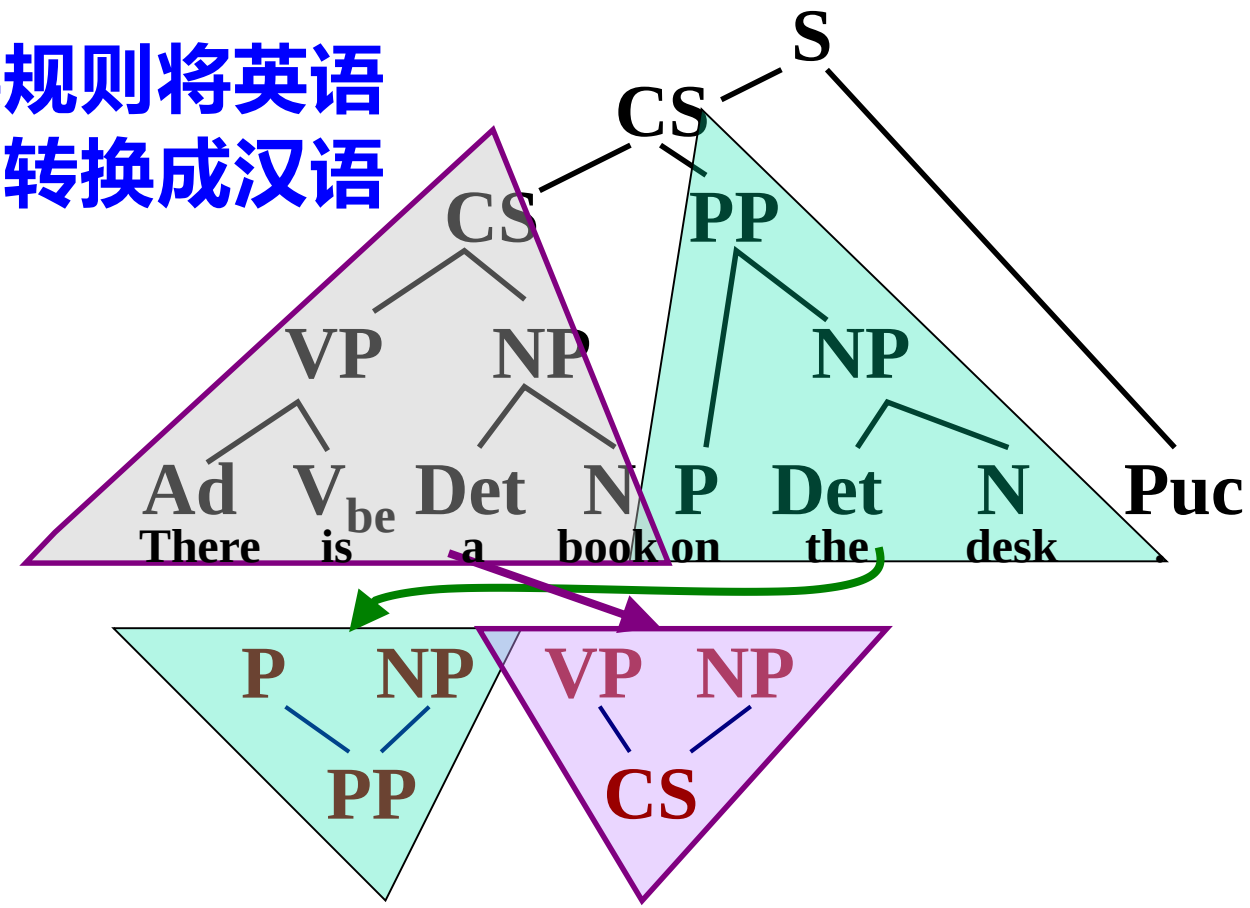
1.6 基本研究方法

- 利用转换规则将英语句子结构转换成汉语句子结构



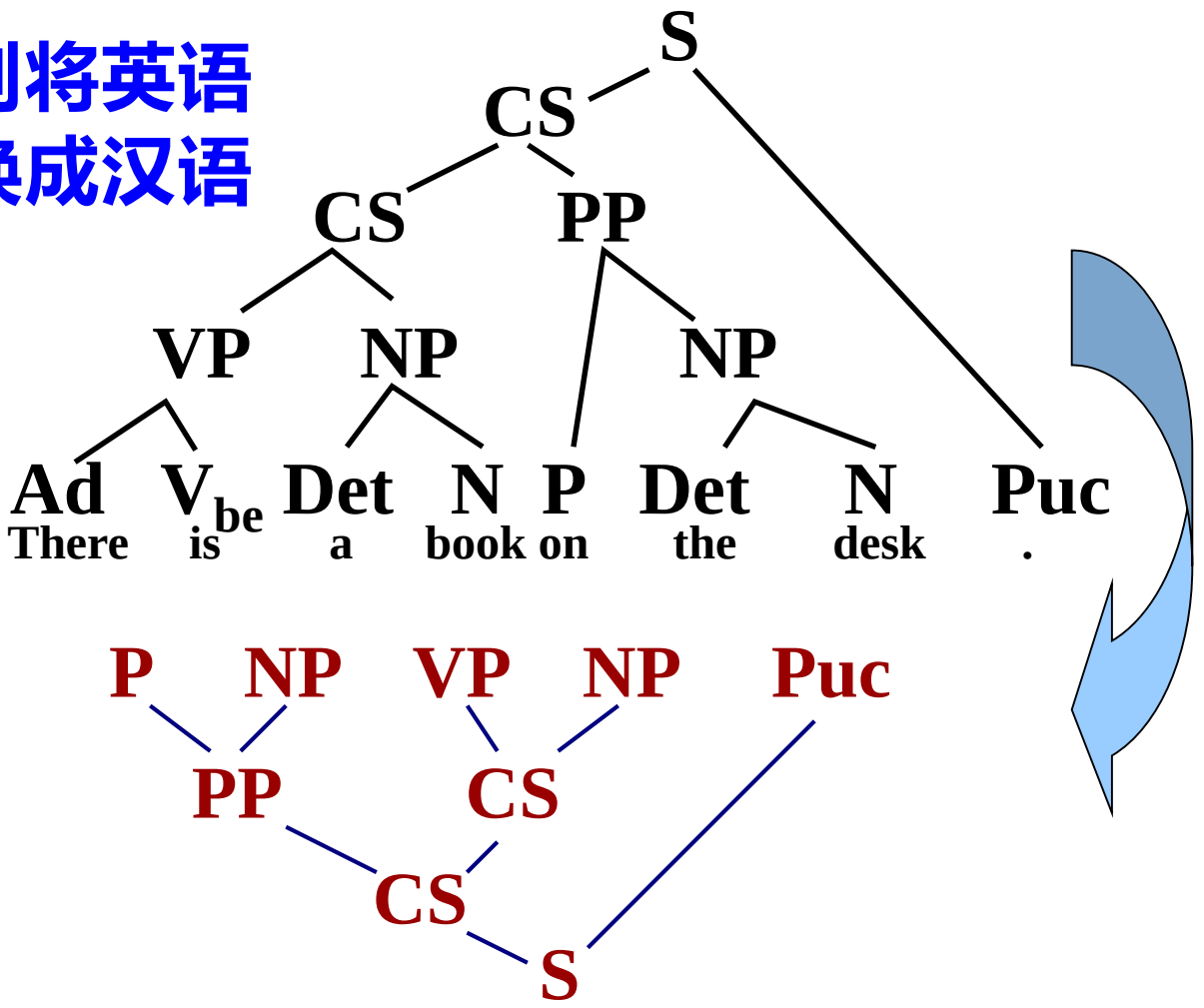
1.6 基本研究方法

- 利用转换规则将英语句子结构转换成汉语句子结构



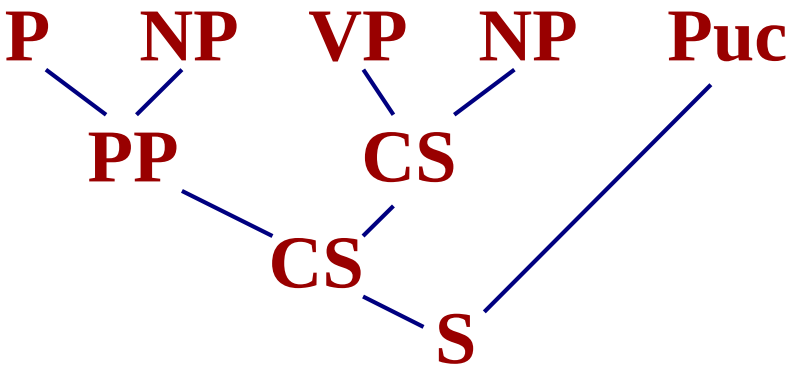
1.6 基本研究方法

- 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

- 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, 一

#book, N, 书; V, 预订

#desk, N, 桌子

#on, P, 在 X 上

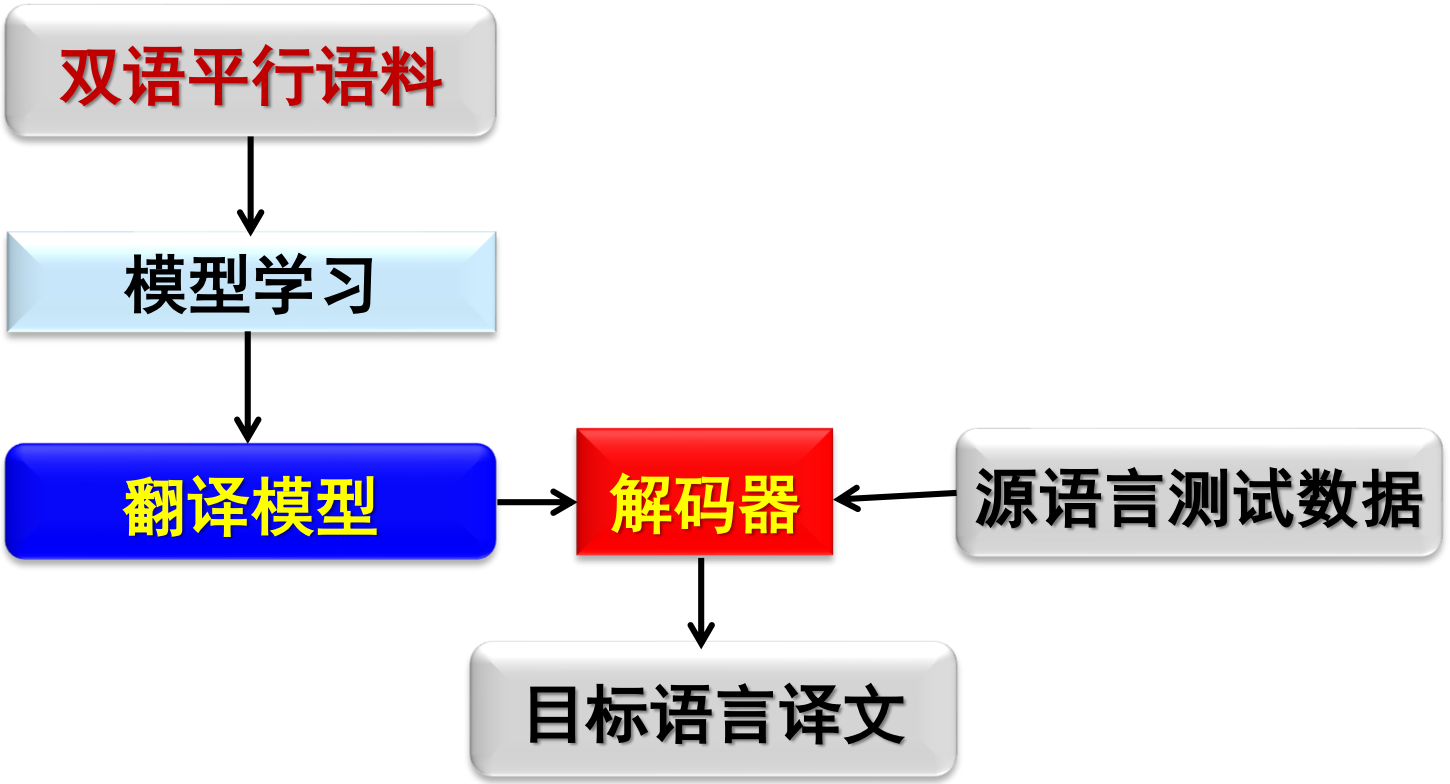
#There be, V, 有

输出译文:

在桌子上有一本书。

1.6 基本研究方法

- 数据驱动的翻译方法（如SMT和 NMT）



1.6 基本研究方法

merkezdiki dölet apparatliri bilen jaylardiki dölet apparatlirining xizmet hoquqi merkezning bir tutash rehberlikide jaylarning teshebbuskarliqi we aktipliqini toluq jari qildurush prinsipi boyiche ayrilidu.
中央和地方的国家机构职权的划分，遵循在中央的统一领导下，充分发挥地方的主动性、积极性的原则。

madda jungxua xelq jumhuriyitide hemme millet bapbarawer.
中华人民共和国各民族一律平等。

herqandaq milletni kemsitish we ëzishni men'i qilidu, milletler ittipaqliqini buzidighan we milliy bölgünchilik qilidighan qilmishlarni men'i qilidu.
禁止对任何民族的歧视和压迫，禁止破坏民族团结和制造民族分裂的行为。

……

1.6 基本研究方法

- 基于统计的方法

给定源语言句子：$E = e_1^m \equiv e_1 e_2 \cdots e_m$

将其翻译成目标语言句子：$C = c_1^l \equiv c_1 c_2 \cdots c_l$

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

求解使P值最大的C

$$\hat{C} = \arg\max_{C} P(C)P(E \mid C)$$

语言模型
(Language model, LM)

翻译模型
(Translation model, TM)

1.6 基本研究方法

构建解码器(decoder)，快速搜索最优翻译候选：

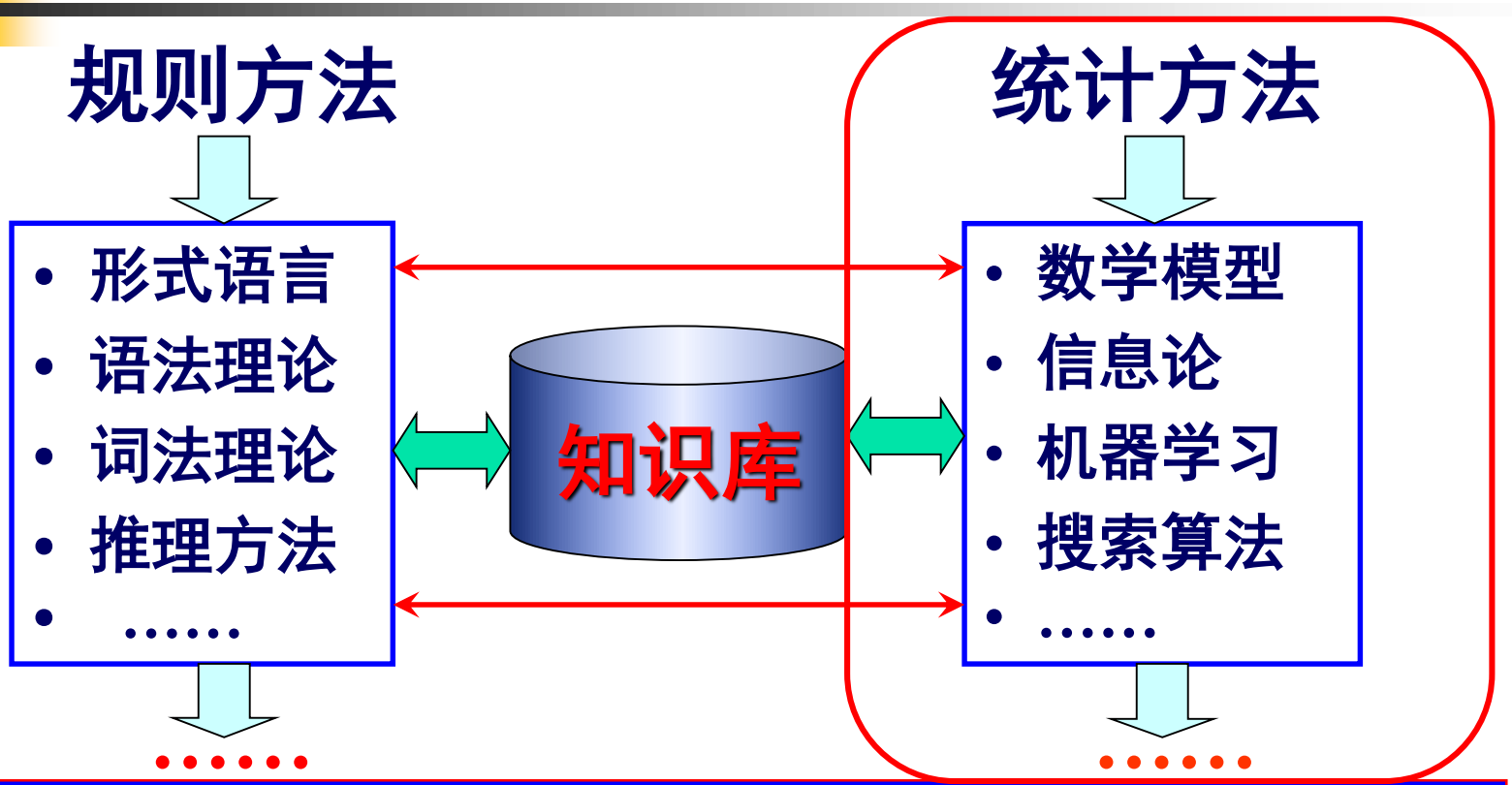
◆三个关键问题：

- 估计语言模型概率 $P(C)$；
- 估计翻译模型概率 $P(E|C)$；
- 快速有效地搜索候选译文C，使$P(C)\times P(E|C)$最大。

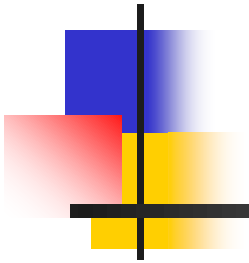
◆主要任务：

- 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



理性主义与经验主义的合谋 —
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？

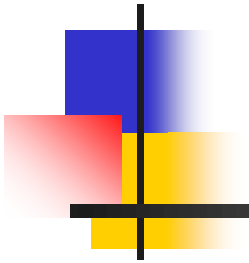
哪个问题都没彻底解决！

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；
- 许多技术离真正实用的目标还有相当的距离，尚未建立起有效、完善的理论体系。

1.7 研究现状



1.8 课程内容

1.8 课程内容

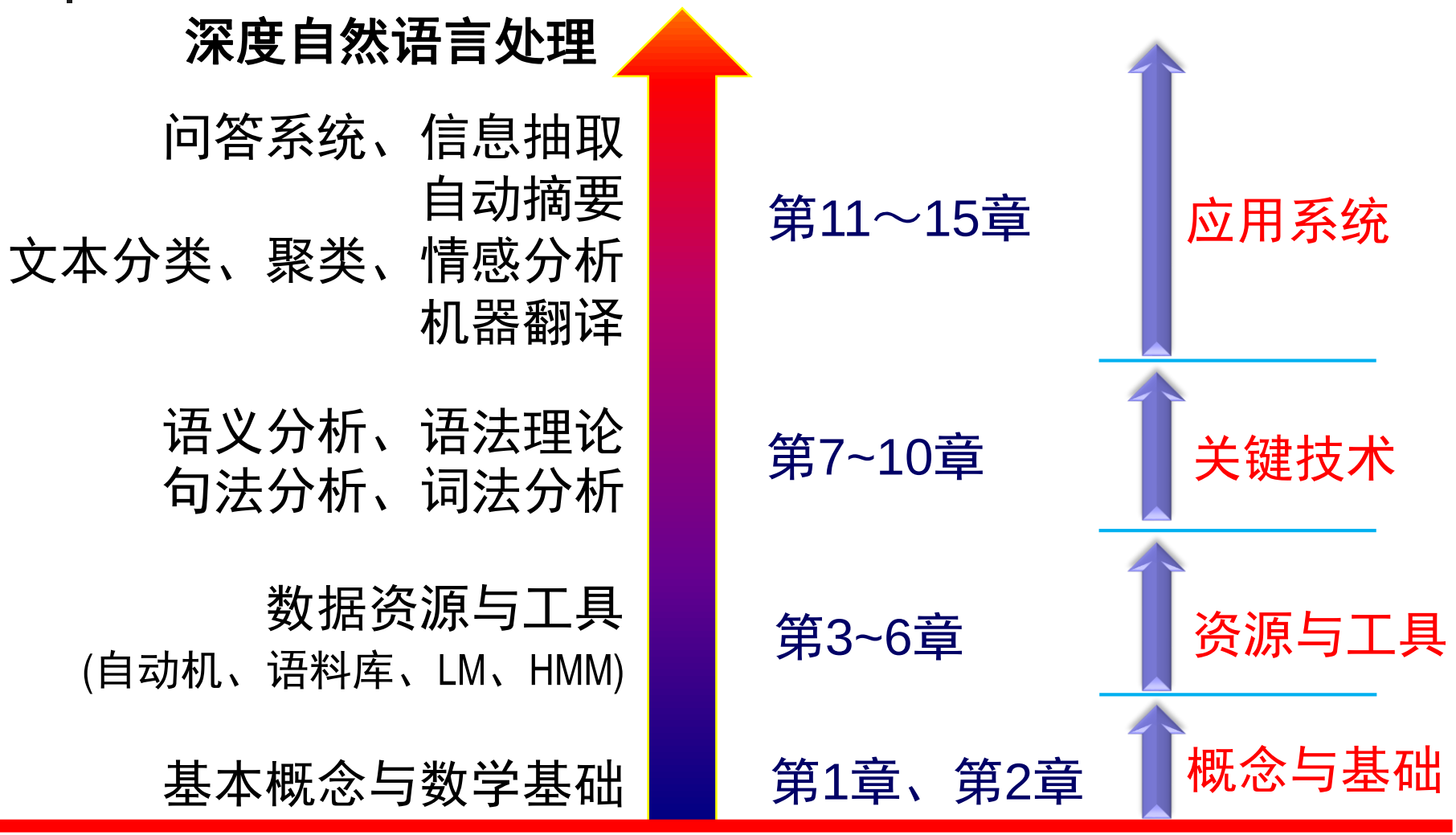
- ◆背景知识
  - 概率论、信息论、建模方法基础
  - 基本的语言学知识
  - 算法分析基础、编程能力
- ◆目的
  - 掌握自然语言理解的基本概念、理论、方法
  - 掌握正确分析问题、解决问题的思维方式
- ◆实验

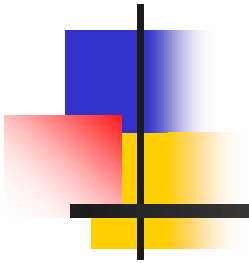
1.8 课程内容

其余各章：

第2章 数学基础
第3章 形式语言与自动机
第4章 语料库语言学
第5章 语言模型
第6章 HMM与CRFs
第7章 词法分析技术
第8章 句法理论
第9章 句法分析
第10章 语义分析
第11章 机器翻译
第12章 文本分类、聚类
第13章 文本自动摘要
第14章 信息抽取
第15章 问答系统
补充： 深度自然语言处理

1.8 课程内容


深度自然语言处理
问答系统、信息抽取
自动摘要
文本分类、聚类、情感分析
机器翻译
语义分析、语法理论
句法分析、词法分析
数据资源与工具
(自动机、语料库、LM、HMM)
基本概念与数学基础
第11～15章
第7~10章
第3~6章
第1章、第2章
应用系统
关键技术
资源与工具
概念与基础

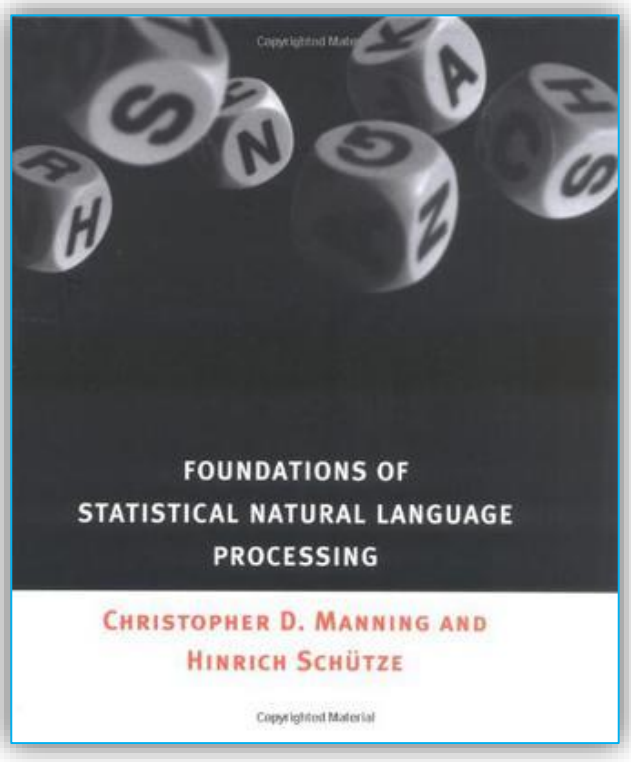


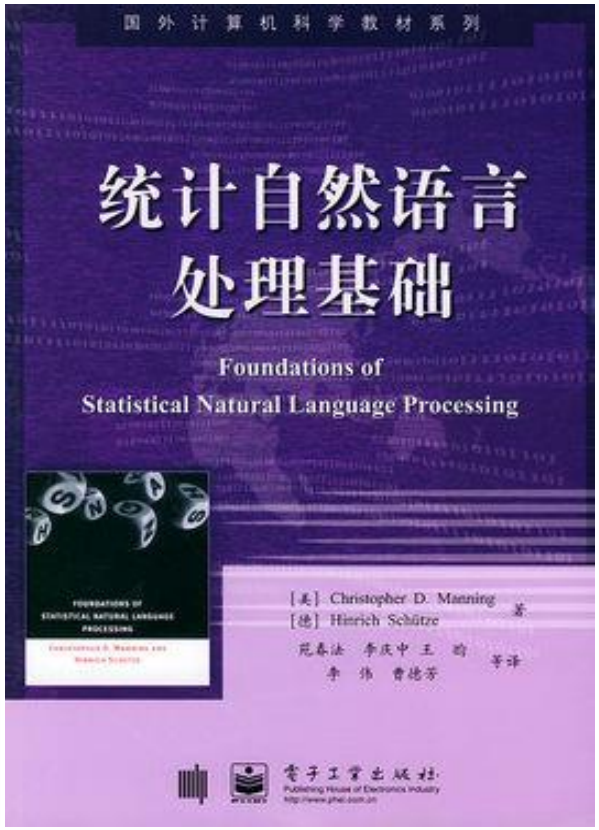
1.9 参考文献

1.9 参考文献

◆专著：

1. C. D. Manning, H. Schüte. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press





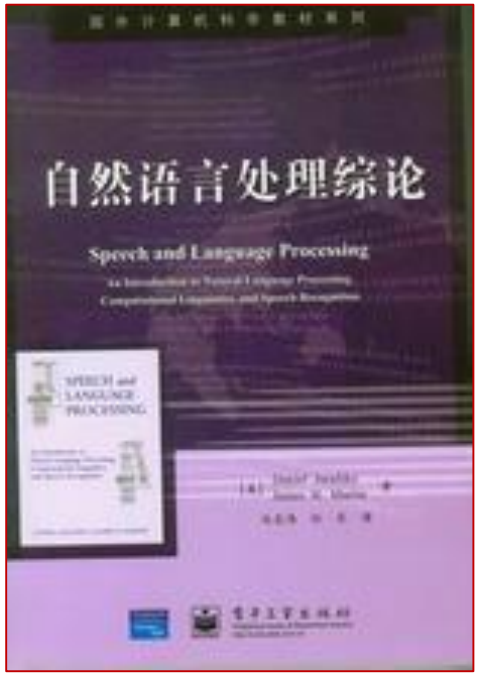
苑春法等译，统计自然语言处理基础，电子工业出版社，2005。

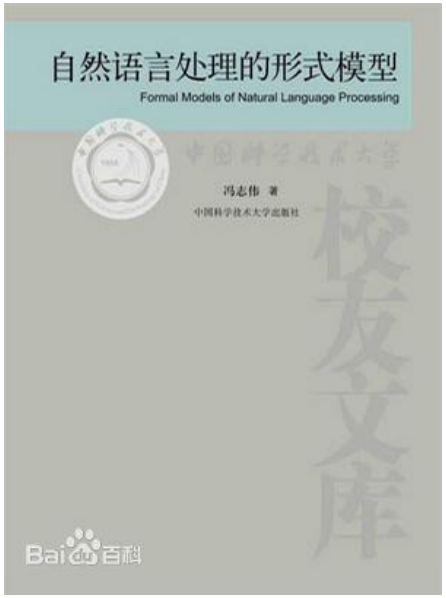
1.9 参考文献

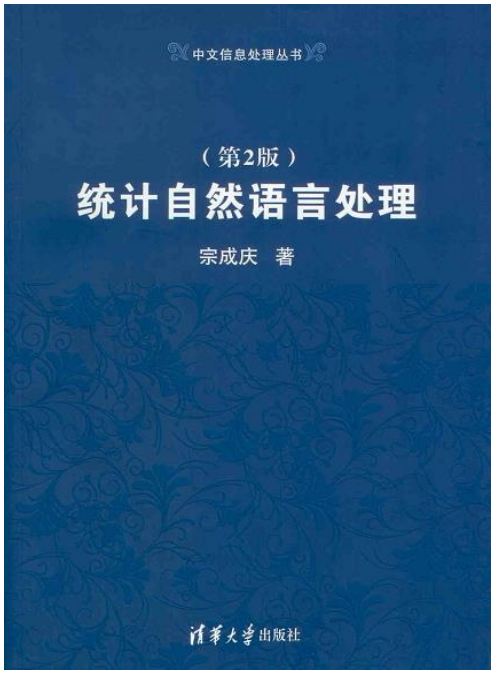
2. D. Jurafsky, and J. H. Martin. 2000. *Speech and Language Processing*, Prentice Hall, 2000 **（冯志伟，孙乐 译，自然语言处理综论，电子工业出版社）**

3. 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

4. 宗成庆，统计自然语言处理(第2版)，清华大学出版社

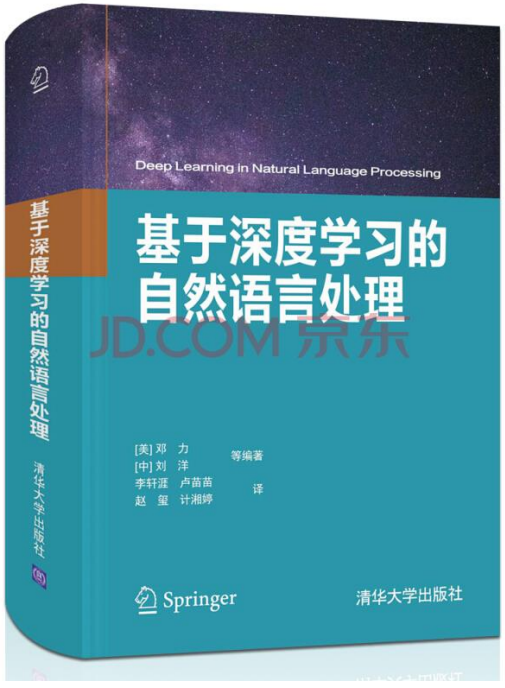


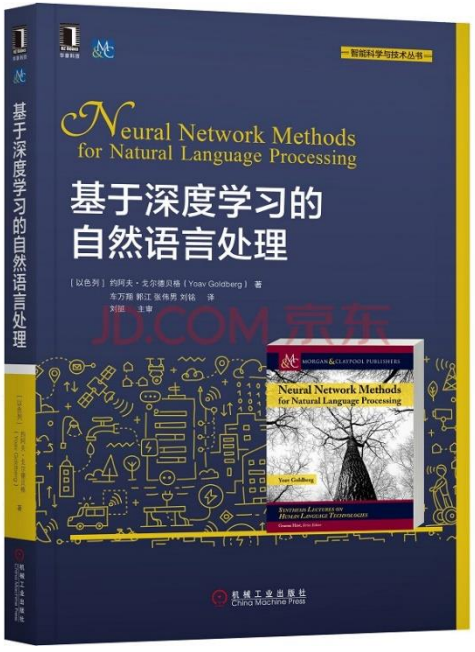


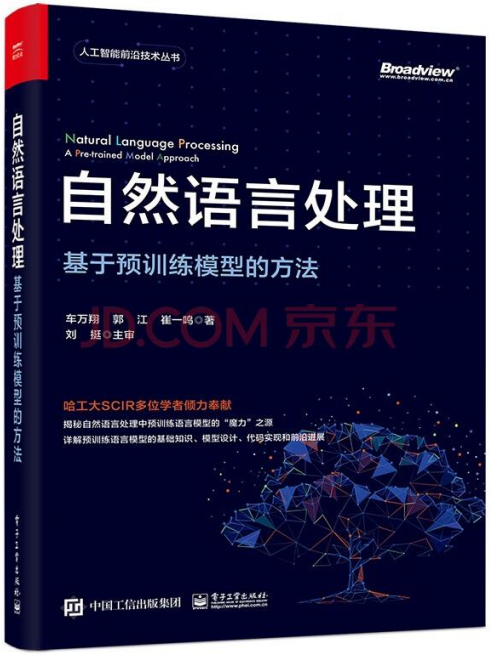


1.9 参考文献

- 基于深度学习的自然语言处理，邓力，刘洋等著，李轩涯，卢苗苗，赵玺，计湘婷译， 清华大学出版社，2020.
- 基于深度学习的自然语言处理，Yoav Goldberg著，车万翔，郭江，张伟男，刘铭译. 机械工业出版社, 2018.
- 自然语言处理：基于预训练模型的方法，车万翔，郭江，崔一鸣 著。电子工业出版社，2021.







1.9 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)

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、句法、语义到语用、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
- 参考文献

Thanks