

自然语言处理

宗成庆

中国科学院自动化研究所

cqzong@nlpr.ia.ac.cn



问题提出

◆什么是语言？

语言是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。

《现代汉语词典》，商务印书

语言可以被看成一个抽象的**数学系统**。

—吴蔚天，1994


◆什么是自然语言？

自然语言是人类社会发展过程中自然产生的语言，特指人类使用的语言。

◆自然语言是可计算的吗？如何判断计算机是否“理解”了语言？……

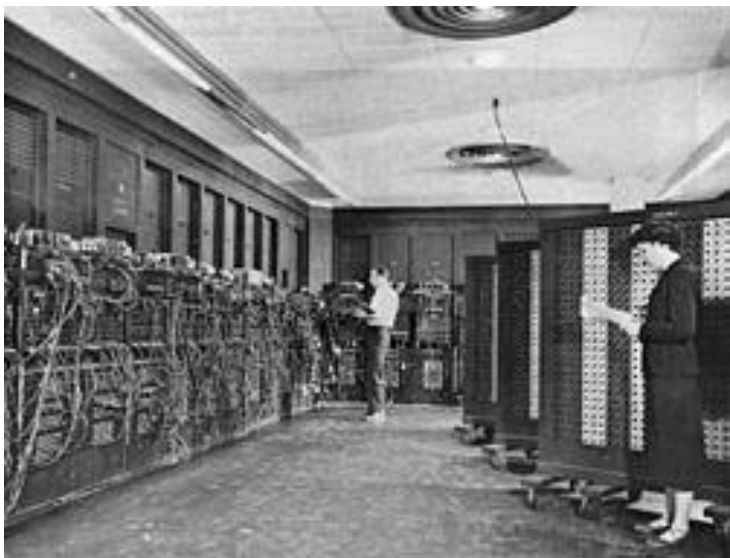


第1章 绪论

- 
1. 基本概念
 2. 问题挑战
 3. 技术方法
 4. 课程内容
 5. 参考文献
 6. 习题

1. 基本概念

◆ 学科产生



1946年，世界上第一台计算机ENIAC诞生。



Warren Weaver

- ✧ 信息论先驱
- ✧ 1920至1932年威斯康星大学数学教授
- ✧ 1932至1955年担任 Rockefeller Institute 自然科学部主任



A. D. Booth

- ✧ 数学物理学家
- ✧ 1947年3月至9月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. 基本概念



诺伯特·维纳 (N. Wiener)
(1894-1964)

[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.

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

difficulties because of multiple meanings, etc., I have wondered if it were unthinkable to design a computer which would translate. Even if it would translate only scientific material (where the semantic difficulties are very notably less), and even if it did produce an inelegant (but intelligible) result, it would seem to me worth while.

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 a linguist and expert on computers, do you think it is worth thinking about?

Cordially,

**July 1949:
“Translation” 备忘录**

Professor Norbert Wiener
Massachusetts Institute of Technology
Cambridge 39, Massachusetts

WW:AEB



1. 基本概念

- 自W. Weaver 和 A. D. Booth 提出机器翻译(machine translation, MT)概念后, 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。系统只有250条俄语词汇, 6 条语法规则, 可以翻译简单的俄语句子。
- 随后10 多年里, 机器翻译研究在国际上出现热潮。
- 与此同时, 人机接口、自动文摘、信息检索等以语言技术为核心的相关研究随之兴趣, 学科萌芽逐渐产生。

1. 基本概念



达特茅斯学院 (成立于1769年)



左起：摩尔、麦卡锡、明斯基、赛弗里奇、所罗门诺夫

人工智能夏季研讨会(大茅斯会议(Dartmouth Conference), 1956)
Summer Research Project on **Artificial Intelligence (AI)**

自然语言理解成为人工智能研究的核心问题之一。
(natural language understanding, NLU)



1. 基本概念

- 1962年国际计算语言学学会(Association for Computational Linguistics, ACL)成立。
- 1965年国际计算语言学委员会(International Committee on Computational Linguistics, ICCL)成立。
- 1964年，美国科学院成立语言自动处理咨询委员会(Automatic Language Processing Advisory Committee, ALPAC)，调查机器翻译的研究情况，并于1966年11月公布了一个题为“语言与机器”的调查报告，简称 ALPAC 报告。计算语言学术语首次正式出现在官方发布的学术报告里。
- 1970 ~ 80S，随着计算机网络的快速发展和普及，以研发实用技术和系统为目标的语言工程应运而生，自然语言处理 (natural language processing, NLP)术语由此诞生。

“自然语言处理是人工智能皇冠上的明珠。”



1. 基本概念

◆ 术语解释

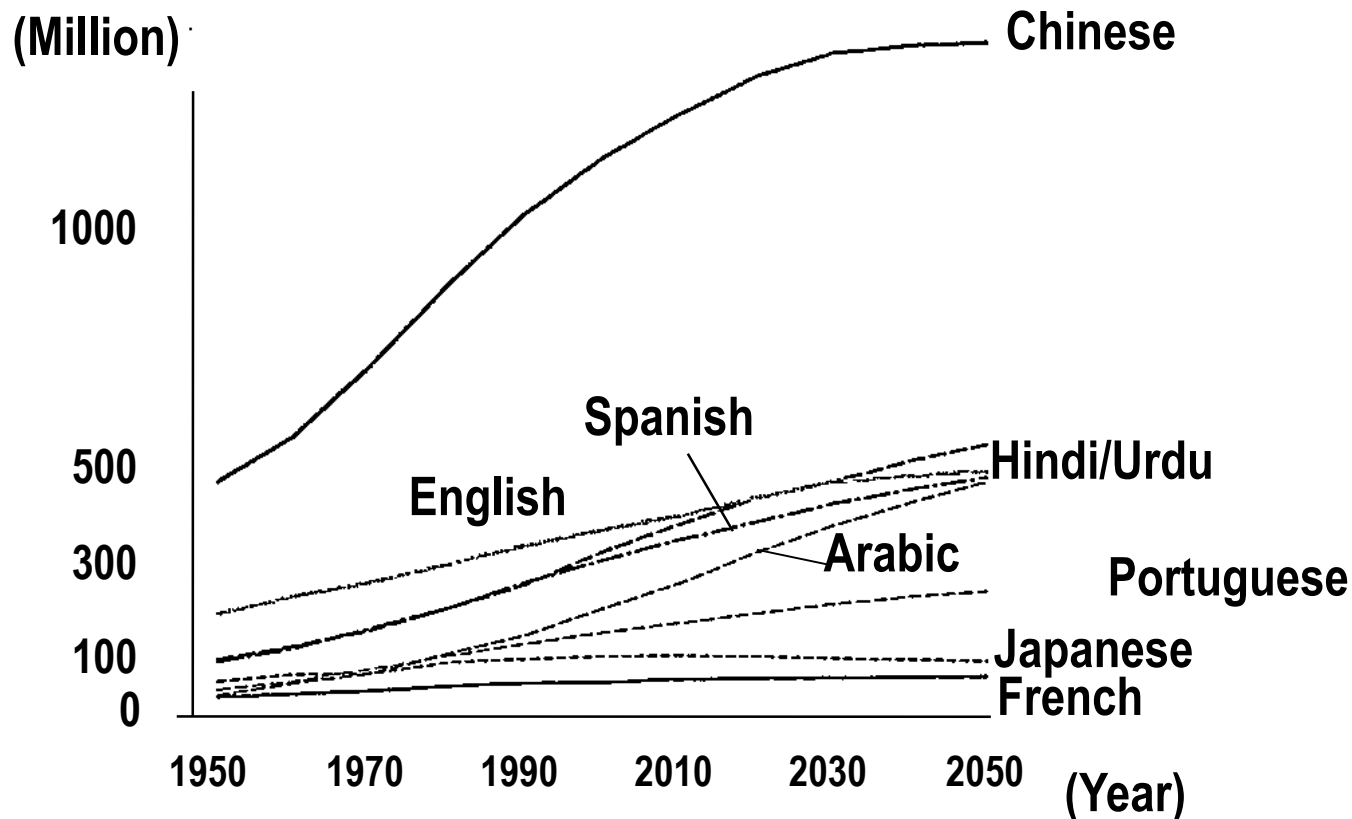
- **自然语言理解**是探索人类自身语言能力和语言思维活动的本质，研究模仿人类语言认知过程的自然语言处理方法和实现技术的一门学科。它是人工智能早期研究的领域之一，是一门在语言学、计算机科学、认知科学、信息论和数学等多学科基础上形成的交叉学科。**(宗成庆，黄昌宁)**
- **计算语言学**是通过建立形式化的计算模型来分析、理解和生成自然语言的学科，是人工智能和语言学的分支学科。计算语言学是典型的交叉学科，其研究常常涉及计算机科学、语言学、数学等多个学科的知识。与内容接近的学科自然语言处理相比较，计算语言学更加侧重基础理论和方法的研究。**(常宝宝)**
- **自然语言处理**是研究如何利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。**(宗成庆，黄昌宁)**

《计算机科学技术百科全书》清华大学出版社，2018.5

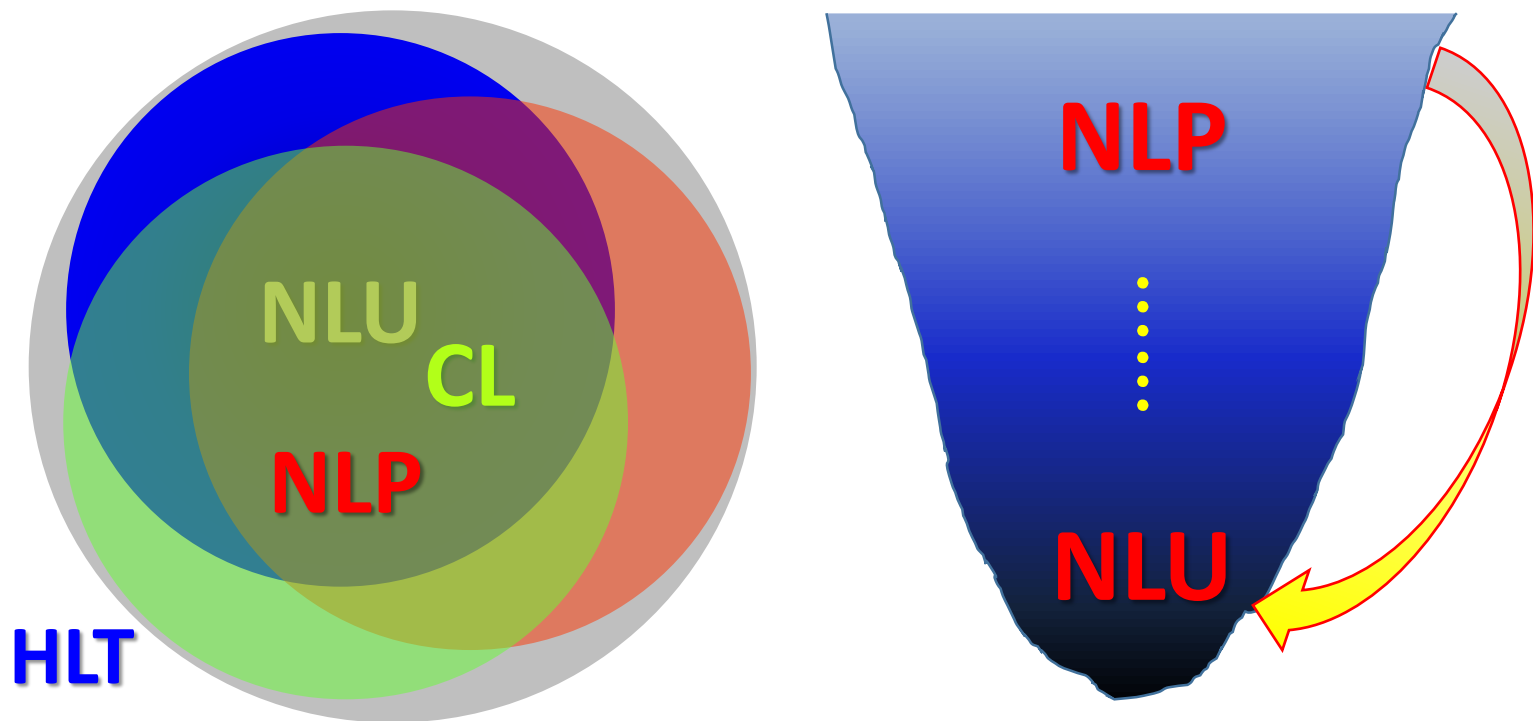


1. 基本概念

上个世纪70/80年代，随着自然语言处理术语的出现，产生了主要以中文（尤指汉语）为处理对象的中文信息处理(Chinese information processing, CIP)技术。



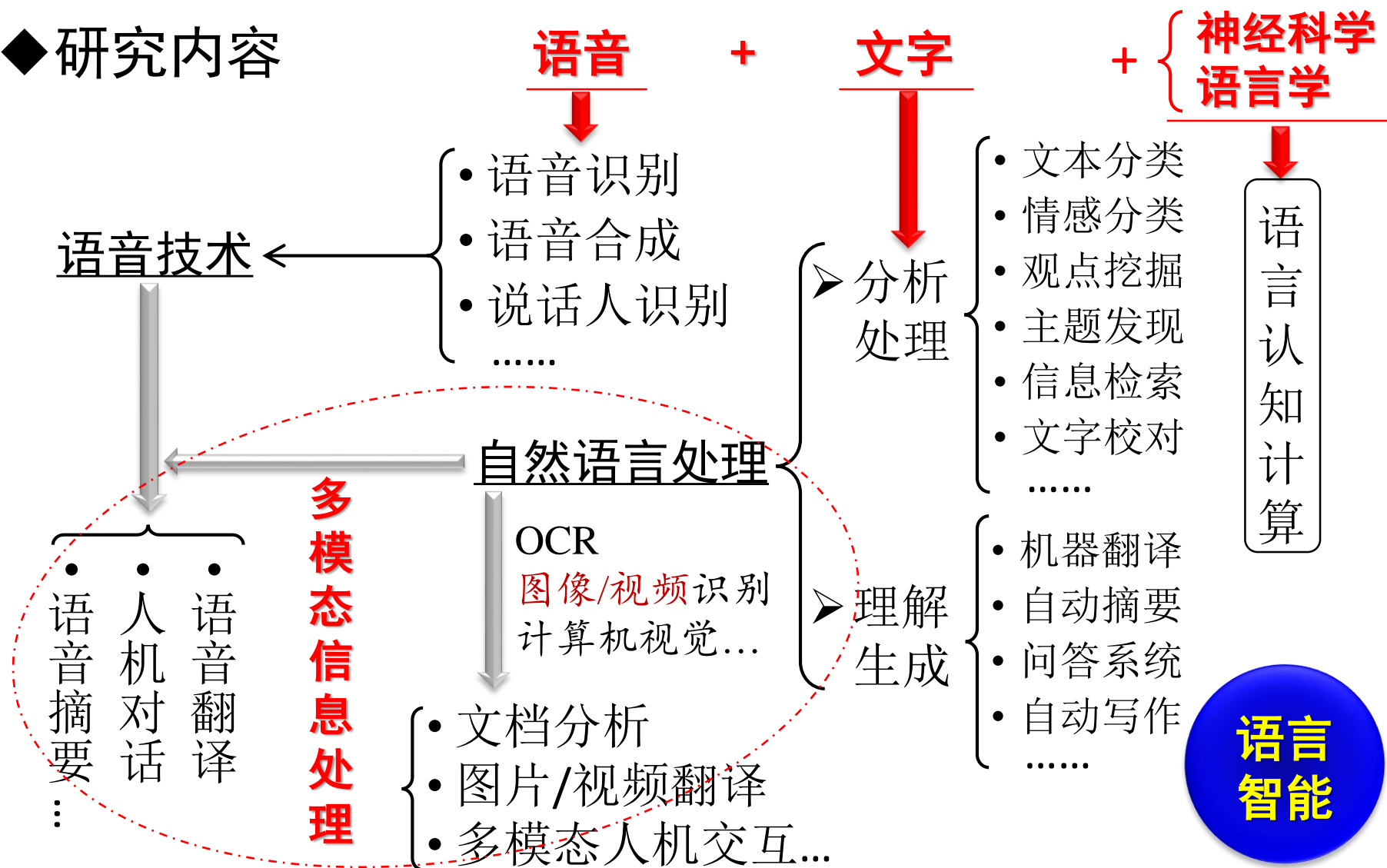
1. 基本概念



- NLU: natural language understanding (1956s)
- CL: computational linguistics (1960s)
- NLP: natural language processing (1970~80s)
- HLT: human language technology (1980s)

1. 基本概念

◆ 研究内容





1. 基本概念

◆应用需求

- ❖ 全世界正在使用的语言有**4000**（7000）多种
- ❖ 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
- ❖ 截至2023年6月，中国网民规模达**10.79亿**，互联网普及率达**76.4%**。 —中国互联网络信息中心发布了第52次《中国互联网络发展状况统计报告》，2023年8月28日
- ❖ 任意时间、任意地点、任意语言的无障碍自由通讯是人类社会的梦想
- ❖ 面对文本**大数据**，我们面临怎样的机遇和挑战？

1. 基本概念

“语言是了解一个国家最好的钥匙”

—习总书记2015年在全英孔子学院和孔子课堂年会开幕式的讲话

已与**152个**国家、**32个**国际组织签署了合作文件，涉及**110多种**语言。
(截止到2023年6月底)

◆ HLT广泛应用

- 国家与公共安全
- 国民经济发展
- 社会生活服务
- 国际交流与合作



“一带一路” 国家战略



第1章 绪论

1. 基本概念

 2. 问题挑战

3. 技术方法

4. 课程内容

5. 参考文献

6. 习题

2. 问题挑战

(1) 门把手弄坏了

门 / 把 / 手 / 弄 / 坏 / 了

门把手 / 弄 / 坏 / 了

门 / 把手 / 弄 / 坏 / 了

(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



✧ 上大学子烛光追思钱伟长 (<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)

❖ 词法歧义(morphological ambiguity) 无处不在



2. 问题挑战

- (3) 喜欢乡下的孩子。
- (4) 关于鲁迅的文章。
- (5) 重要的书籍和手稿。
- (6) 今天吃**馒头**；今天吃**食堂**；今天吃**大碗**。
- (7) 写文章/ 写毛笔/ 写黑板

❖ 句法歧义(syntactic ambiguity)比比皆是



2. 问题挑战

❖ 语义歧义(semantic ambiguity)

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，例如：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**；她厉害得像个**母老虎**；全国各族人民要**像石榴籽一样**紧密地团结在一起。

❖ 语用歧义(pragmatic ambiguity)



2. 问题挑战

❖ 大量未知语言现象

- 新词、人名、地名、术语等, 如: 奥特、给力、内卷; 非典、新冠肺炎; 夏天、高山、温馨、温泉、时光、平安、田野、边疆、程序、吉林、桂林, 不来梅
布莱尔
- 新含义
如: 苹果、奔腾、老虎、后台
- 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 ……

2. 问题挑战

◆ 问题归纳

- 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、新的语义和非规范语法等现象无处不在
- 始终面临的数据不充分性：样本的有限性与语言的动态性
- 知识表示的复杂性：知识(包括常识)的模糊性、关联性和隐喻表达的隐蔽性及人类联想的多变性
- 不同语言语义概念的差异性：语义概念和文化的差异化

如何从大量复杂多样的不确定性中寻找确定性结论



2. 问题挑战

◆ ChatGPT 已经解决了自然语言处理问题吗？

NW 请介绍一下鸡兔同笼问题。

鸡兔同笼问题是一个经典的数学问题，问题的形式通常是这样的：如果有一个笼子里有鸡和兔子，总共有 n 只脚，问笼子里有多少只鸡和多少只兔子。这道题目的答案通常可以用一个简单的方程来推导出来：如果 c 只鸡和 r 只兔子，那么 $c + r = n$ ，并且 $2c + 4r = 2n$ 。由于鸡和兔子的脚数之和是已知的，所以我们可以用这两个方程来求解 c 和 r 的值。例如，如果 $n = 10$ ，那么 $c + r = 10$ ，并且 $2c + 4r = 20$ ，我们可以用第一个方程来推出 $c + r = 10$ ，然后用第二个方程来推出 $c = 5$ 和 $r = 5$ 。这样，我们就可以得到答案：笼子里有 5 只鸡和 5 只兔子。

NW 鲁迅和周树人是同一个人吗？

不，鲁迅和周树人不是一个人。鲁迅（原名周樟寿）和周树人是两位中国现代著名的作家，他们都有重要的文学成就，但是他们是两个不同的人。……

NW 你知道“林黛玉倒拔垂杨柳”的故事吗？

一本正经地胡说八道

是的，我了解“林黛玉倒拔垂杨柳”的故事。这个故事出自中国古典文学名著《红楼梦》……在故事中，林黛玉在园中倒拔垂杨柳的情节发生在贾府的花园里……

2. 问题挑战

◆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？



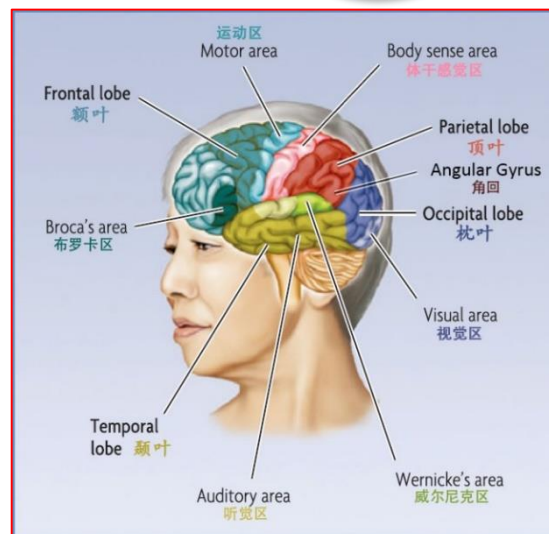
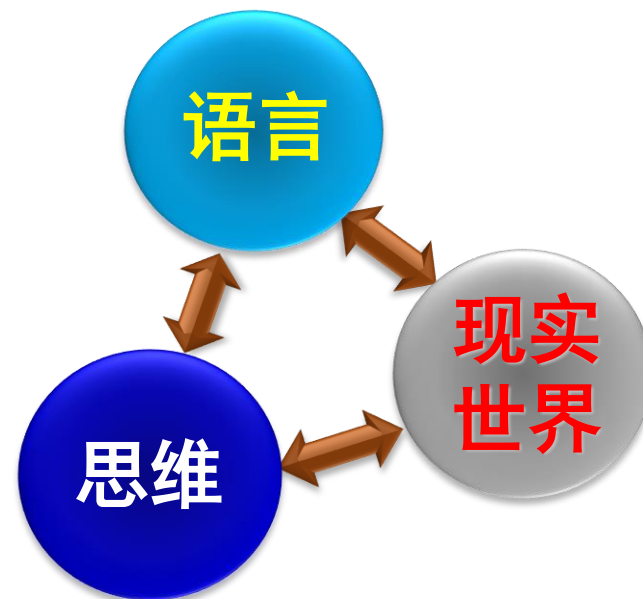
图灵设计的“模仿游戏” — 图灵实验(Turing test)

2. 问题挑战

◆ 人脑是如何理解语言的？


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

... ..



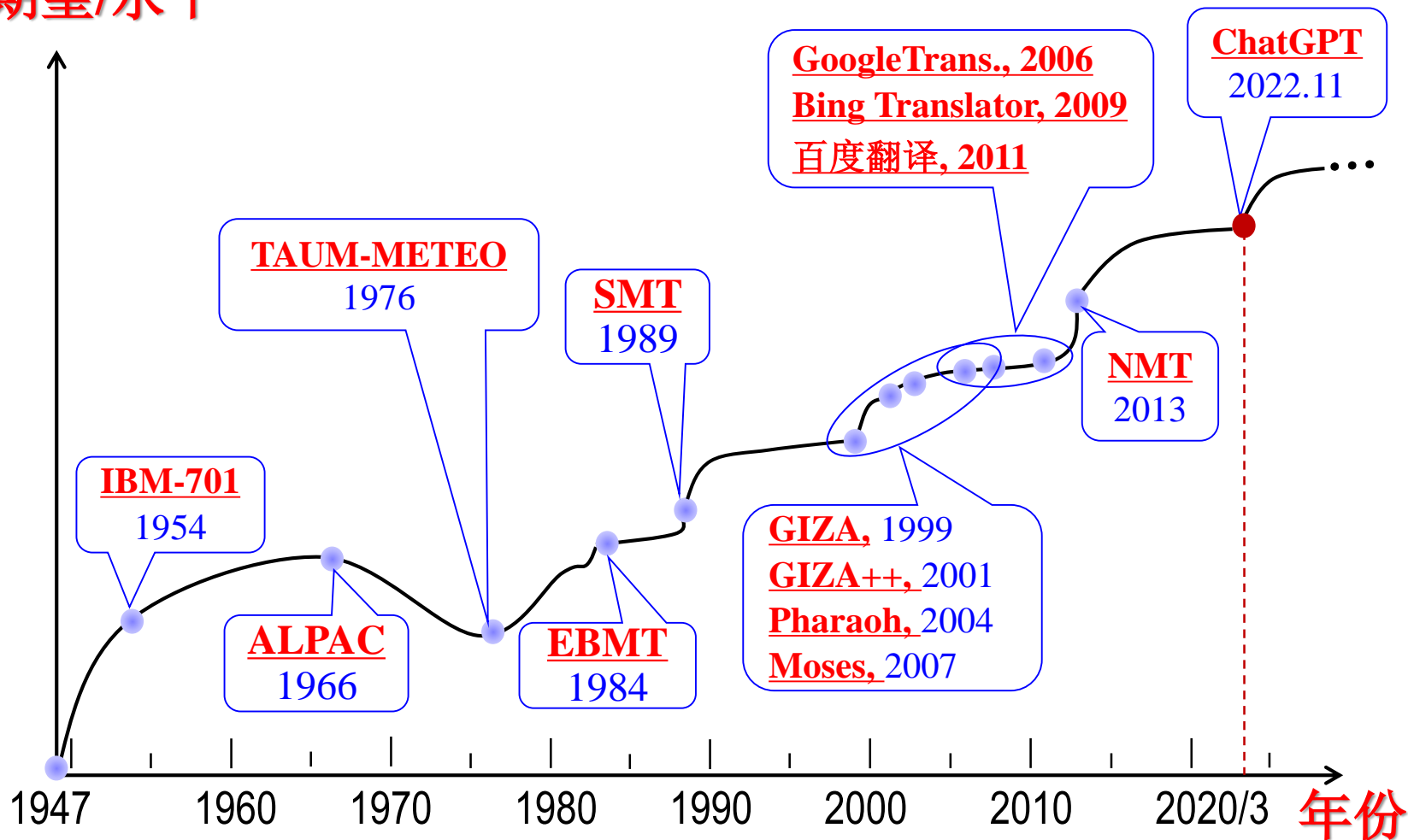


第1章 绪论

1. 基本概念
2. 问题挑战
-  3. 技术方法
4. 课程内容
5. 参考文献
6. 习题

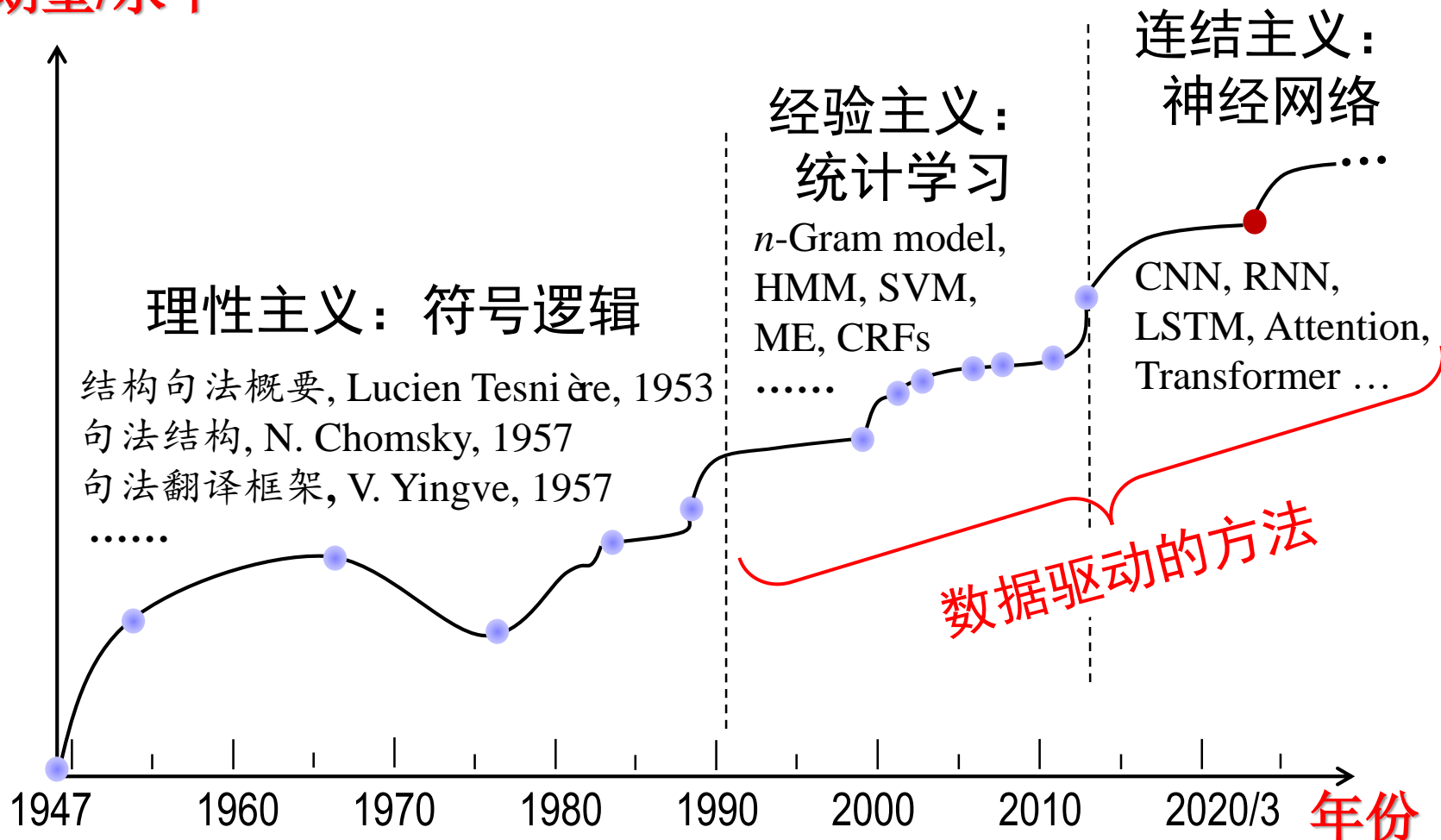
3. 技术方法

期望/水平



3. 技术方法

期望/水平





3. 技术方法

- ◆ **理性主义**：通过对一些代表性语句或语言现象的研究得到对人的语言能力的认识，归纳语言使用的规律，以此分析、推断测试样本的预期结果。
- **问题求解思路**：基于规则的分析方法建立符号处理系统
 - **设计规则**： $N + N \rightarrow NP$
 - **标注词典**： $\#工作, N(uc); V;$
 - **推导算法**：归约、推导、歧义消解方法...

知识库 + 推理系统 \rightarrow NLP 系统



3. 技术方法

- ◆ **经验主义**：利用大规模真实语言数据，借助人的帮助(标注数据和筛选特征等)，统计发现语言使用的规律及其可能性(概率)大小，以此为依据计算预测测试样本的可能结果。统计单元是离散事件(词、短语、词性等)。
- **问题求解思路**：基于大规模真实数据建立计算模型
 - **收集标注语料**：真实性、代表性、标注...
 - **统计建型**：模型的复杂性、有效性、参数训练...

语料收集、标注 + 统计模型 → NLP 系统



3. 技术方法

- ◆ **连结主义**：利用大规模真实语言数据，统计发现语言使用的规律及其可能性(概率)大小，以此为依据计算预测测试样本的可能结果。统计单元采用连续的实数空间表示(向量)。
- **问题求解思路**：基于大规模真实数据建立计算模型
 - **收集语料**：真实性、代表性...
 - **统计建型**：参数训练...

语料收集 + 神经网络 → NLP 系统

3. 技术方法

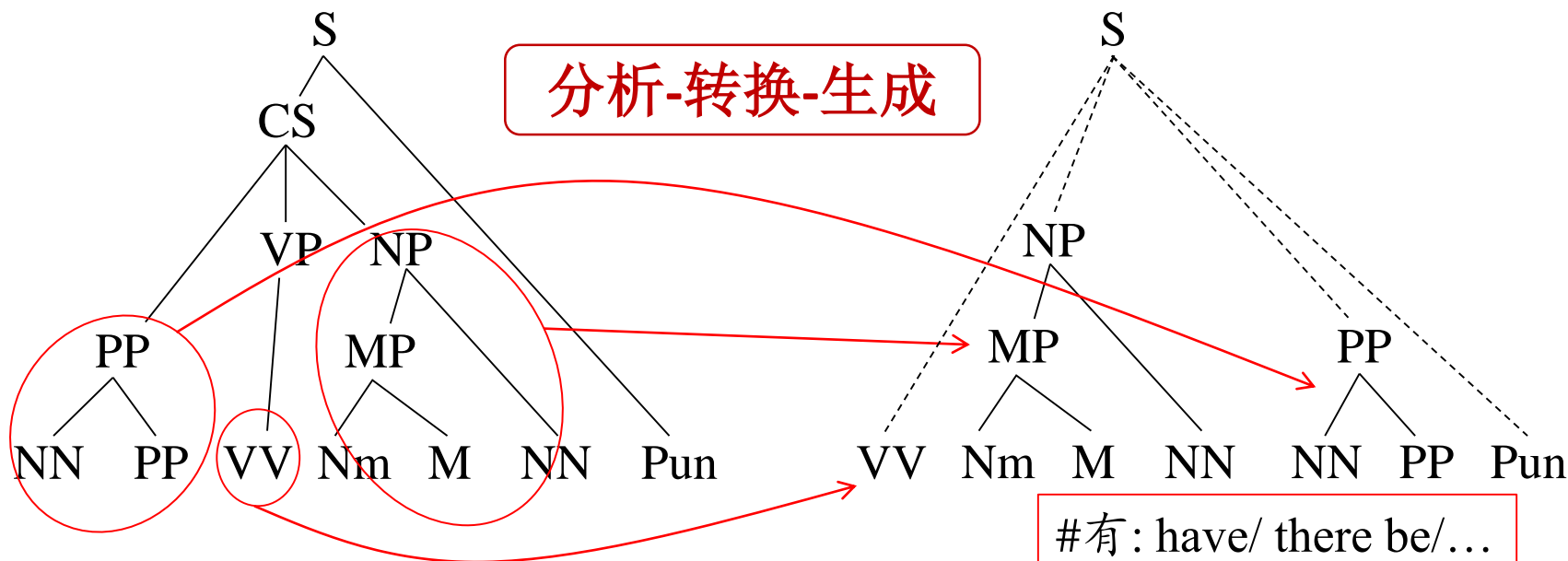
◆以机器翻译为例 ①基于规则的方法

给定源语言句子：桌子上有一本书。

(a)分词与词性标注：桌子/NN 上/PP 有/VV 一/Num 本/M 书/NN 。/Pun

(b)句法结构分析：

(c)结构转换：



(d)译文生成：There is a book on the desk.

#有: have/ there be/...
#一: one/ a
#书: book
.....

3. 技术方法

● 方法评价：

- **优点：**可以较好地保持原文的结构，产生的译文结构与源文的结构关系密切，在源文语言现象已知、句法结构规范且词汇歧义不复杂的情况下，具有很好的处理能力，可以得到较好的译文，且分析、转换和生成的每一步都是可追踪、可溯源的。
- **弱点：**需要人去编写规则，工作量大，主观性强，一致性难以保障，不利于系统扩充，非规范语言现象的处理能力差。对于很对语言来说，难以找到熟悉该语言的规则和词典编写人员。系统开发周期长，领域、语种等可移植性差。

3. 技术方法

② 统计翻译方法



源语言句子: $S = s_1^m \equiv s_1 s_2 \cdots s_m$

目标语言句子: $T = t_1^l \equiv t_1 t_2 \cdots t_l$

$$p(T | S) = \frac{p(T) \times p(S | T)}{p(S)}$$

$$\hat{T} = \arg \max_T \frac{p(T) \times p(S | T)}{p(S)}$$

语言模型

Language model, LM

翻译模型

Translation model, TM



3. 技术方法

➤ 双语平行句对

merkezdiki d ölet apparatliri bilen jaylardiki d ölet apparatlrining 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 b ölg ünchilik 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。

.....

3. 技术方法

● 方法评价：

- 优点：一般不需要对源语言句子进行深层次的分析，甚至可以对源语言没有任何基本的知识，只要有足够多的高质量双语言句对就可以建立一个机器翻译系统。系统开发周期短。
- 弱点：对于很多语言对来说，难以收集到大规模高质量的双语句对。句法结构复杂的源语言长句的译文质量差，译文与原文的语义一致性无法保证。尤其当测试集与训练集（领域、风格等）差异较大，且出现生词时，译文质量大幅度降低。



3. 技术方法

③ 基于神经网络的翻译方法

给定源语言句子: $C = c_1^l \equiv c_1 c_2 \cdots c_l$

将其翻译成目标语言句子: $E = e_1^m \equiv e_1 e_2 \cdots e_m$

词汇向量化表示: Word2Vec

基于 Transformer 的端到端的翻译框架/ 基于生成式语言模型

$$P(e_i) \approx P(e_i | \underline{e_1 \cdots e_{i-1}}, C)$$

目标函数:
$$L = \sum_i \log(P(e_i | C))$$

3. 技术方法

● 方法评价：

- 优点：不需要对源语言句子进行分析，甚至可以对源语言没有任何基本的知识，只要有足够多的高质量双语言句对和词典、大规模单语言训练样本和针对翻译任务的人工提示指令，就可以建立一个机器翻译系统，开发周期短。译文流畅性好，对缩略语和术语等有较好的翻译结果。
- 弱点：对于很多语言对来说，难以收集到大规模高质量的双语句对、词典和翻译提示指令。存在流畅性好但语义错误、缺失或“无中生有” / “幻觉”的译文。



第1章 绪论

1. 基本概念
2. 问题挑战
3. 技术方法
- ➡ 4. 课程内容
5. 参考文献
6. 习题



4. 课程内容

◆ 课堂讲授

- 基本概念
- 方法、模型和算法
- 应用系统

◆ 课程实践

- 项目作业：方法/系统实现 + 技术报告

◆ 课程成绩

- 闭卷考试 (60%)
- 课程实践 (40%)



4. 课程内容

其余各章:

第2章 统计学习基础

第3章 形式语言与自动机

第4章 N元语法模型

第5章 HMM与CRFs

第6章 神经网络与语言模型

第7章 文本表示

第8章 汉语分词与词性标注

第9章 句法分析

第10章 篇章分析

第11章 语义分析

第12章 预训练模型

第13章 机器翻译

第14章 文本分类与聚类

第15章 信息抽取

第16章 人机对话系统

课程总结与展望

共计：57+3 学时



4. 课程内容



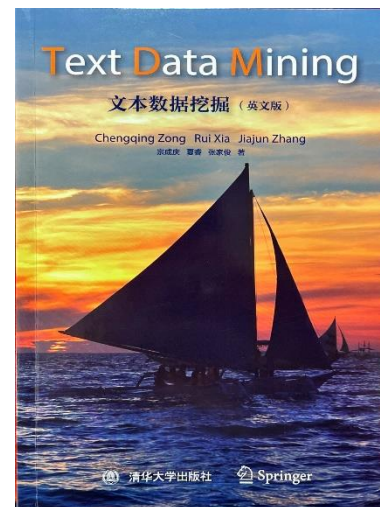
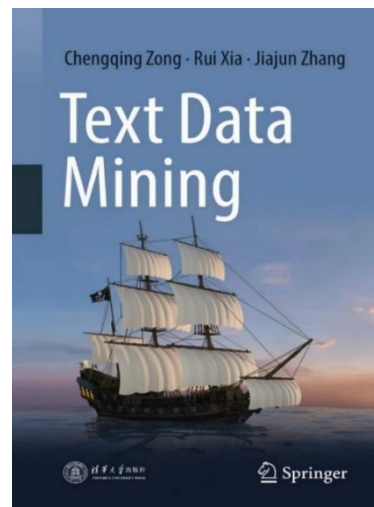
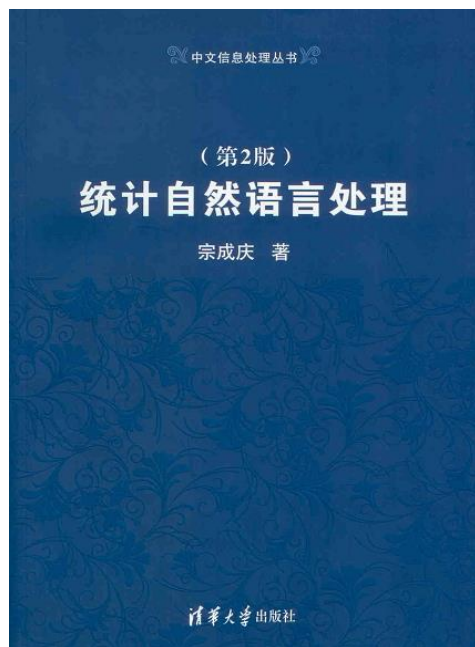


第1章 绪论

1. 基本概念
2. 问题挑战
3. 技术方法
4. 课程内容
- ➡ 5. 参考文献
6. 习题

5. 参考文献

- [1] 宗成庆, 夏睿, 张家俊, 文本数据挖掘(第2版), 清华大学出版社, 2022.11
- [2] C. Zong, R. Xia, J. Zhang, Text Data Mining, Springer/ Tsinghua Univ. Press, 2021/ 2022
- [3] 宗成庆, 统计自然语言处理(第2版), 清华大学出版社, 2013.8



<https://www.springer.com/gp/book/9789811600999> (下载网址)

5. 参考文献

- [4] 张家俊, 赵阳, 宗成庆译, 神经机器翻译, 机械工业出版社, 2022.3
- [5] 宗成庆, 赵阳, 百度飞桨, 自然语言处理基础与大模型, 清华大学出版社
(待版)





5. 参考文献

◆美国几所大学开设的NLP课程

University	Instructors	Websites
Columbia University	Michael Collins	http://www.cs.columbia.edu/~cs4705/
CMU	Alan W. Black David R. Mortensen	http://demo.clab.cs.cmu.edu/NLP/
MIT		http://web.mit.edu/6.863/www/fall2012/
Stanford University		http://online.stanford.edu/course/natural-language-processing



5. 参考文献

● Stanford School of Engineering (2021.7)

➤ Natural Language Understanding

(CS224U: <https://online.stanford.edu/courses/cs224u-natural-language-understanding>)

What you will learn ➤ 春季为高年级本科生和低年级研究生开设的。

- Lexical semantics
- Distributed representations of meaning
- Relation extraction
- Semantic parsing
- Sentiment analysis
- Dialogue agents

➤ Natural Language Understanding

(XCS224U: <https://online.stanford.edu/courses/xcs224u-natural-language-understanding>)

What you will learn ➤ 为成人进修开设的。

- Distributed word representations
- Relation extraction with distant supervision
- Natural language inference
- Evaluation methods and metrics
- Contextual word representations (including updated coverage of BERT, RoBERTa, ELECTRA, and XLNet)
- Supervised sentiment analysis
- Grounded language understanding
- Semantic parsing

Time Commitment

Expect to commit 8-12 hours/week for the duration of the 10-week program.



5. 参考文献

➤ Natural Language Processing with Deep Learning

(CS224N: <https://online.stanford.edu/courses/cs224n-natural-language-processing-deep-learning>)

What you will learn ➤ 冬季。

- Computational properties of natural languages
- Co-reference, Q&A, and machine translation
- Processing linguistic information
- Syntactic and semantic processing
- Modern quantitative techniques in NLP
- Neural network models for language understanding tasks

➤ Natural Language Processing with Deep Learning

(XCS224N: <https://online.stanford.edu/courses/xcs224n-natural-language-processing-deep-learning>)

What you will learn

- Computational properties of natural languages
- Neural network models for language understanding tasks
- Word vectors, syntactic, and semantic processing
- Co-reference, question answering, and machine translation
- Transformers and pre-training

Time Commitment

Expect to commit 10-14 hours/week for the duration of the 10-week program.

➤ Spoken Language Processing:

(CS224S: <https://cs.stanford.edu/courses/schedules/2021-2022.spring.php>)



本章小结

- ◆ 基本概念： NLU, CL, NLP, HLT, CIP
- ◆ 学科的产生与发展： 1947, 1966, 1980s, 1990s, 2013, 2022...
- ◆ 研究内容： 语音技术, NLP, 认知语言计算, 多模态...
- ◆ 问题与挑战： 从词法、句法、语义到语用 ...
- ◆ 基本方法： 理性主义、经验主义和连结主义方法
- ◆ 课程内容： 基本概念 - 基础工具 - 关键技术 - 应用系统
- ◆ 参考文献： 四本专著+一本译著； 美国部分大学的NLP课程



第1章 绪论

1. 基本概念
2. 问题挑战
3. 技术方法
4. 课程内容
5. 参考文献

 6. 习题



6. 习题

1. 请说明如下句子有多少种不同的含义？
 - (1) He drew one card.
 - (2) 咬死猎人的狗。
 - (3) 鸡不吃了。
2. 任意选取一批英文句子，分析其词性的兼类情况。
3. 试举例比较汉英句子的结构差异。
4. 请列举不少于10种自然语言处理技术应用的场景和用途。
5. 通过对比测试微软Bing翻译、百度翻译和腾讯翻译等系统，了解机器翻译技术的性能现状。
6. 试用 ChatGPT 或 “文心一言” (<https://yiyan.baidu.com/>) 等，分析系统回复存在的问题。



Sir, why you are here ?

- 为什么今晚我要坐在这里？
- 为什么我要选听这门课？
- 为什么我要选学这个专业？
- 我为什么选择国科大？

.....

你是来吃蛋糕还是学做蛋糕？

作为国科大的学生，应该具有强烈的求知欲和好奇心，对于任何一件事情不止步于问题的表面，而要探索其本质，熟知其来龙去脉，不仅局限于掌握已有的知识和使用现成的工具，更要思考如何改变现状和创造未来，不要只想做个“吃蛋糕”的人。



谢谢!

Thanks!

