

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

第1章 绪论

本章内容

- 
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

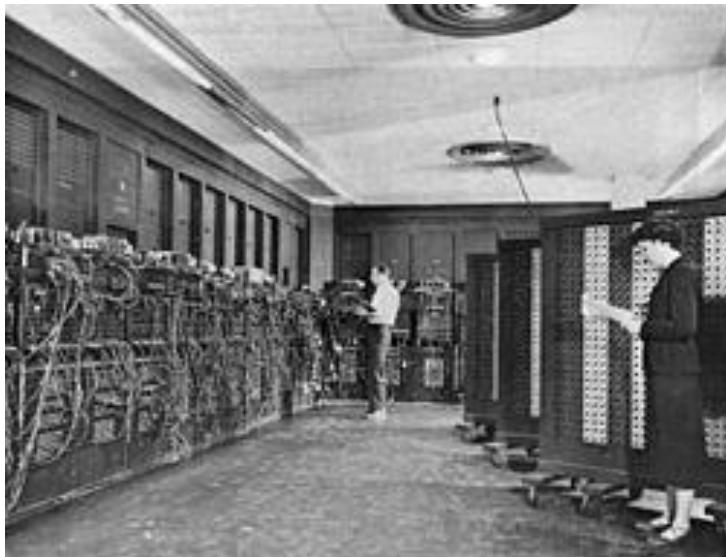
- 自然语言是人类社会发展过程中自然产生的，简称语言。
- 语言是思维的载体，是人类交流思想、表达情感最自然、最直接、最方便的工具。

“自然语言理解是人工智能皇冠上的明珠。”

- 自然语言理解(natural language understanding, NLU)
- 自然语言处理(natural language processing, NLP)
- 计算语言学(computational linguistics, CL)
- 中文信息处理(Chinese information processing, CIP)

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.

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 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(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, 1956)

Summer Research Project on **Artificial Intelligence** (Dartmouth Conference)

自然语言理解(NLU)成为人工智能研究的核心问题之一。



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，自然语言处理术语由此诞生。



1. 基本概念

◆ 术语解释

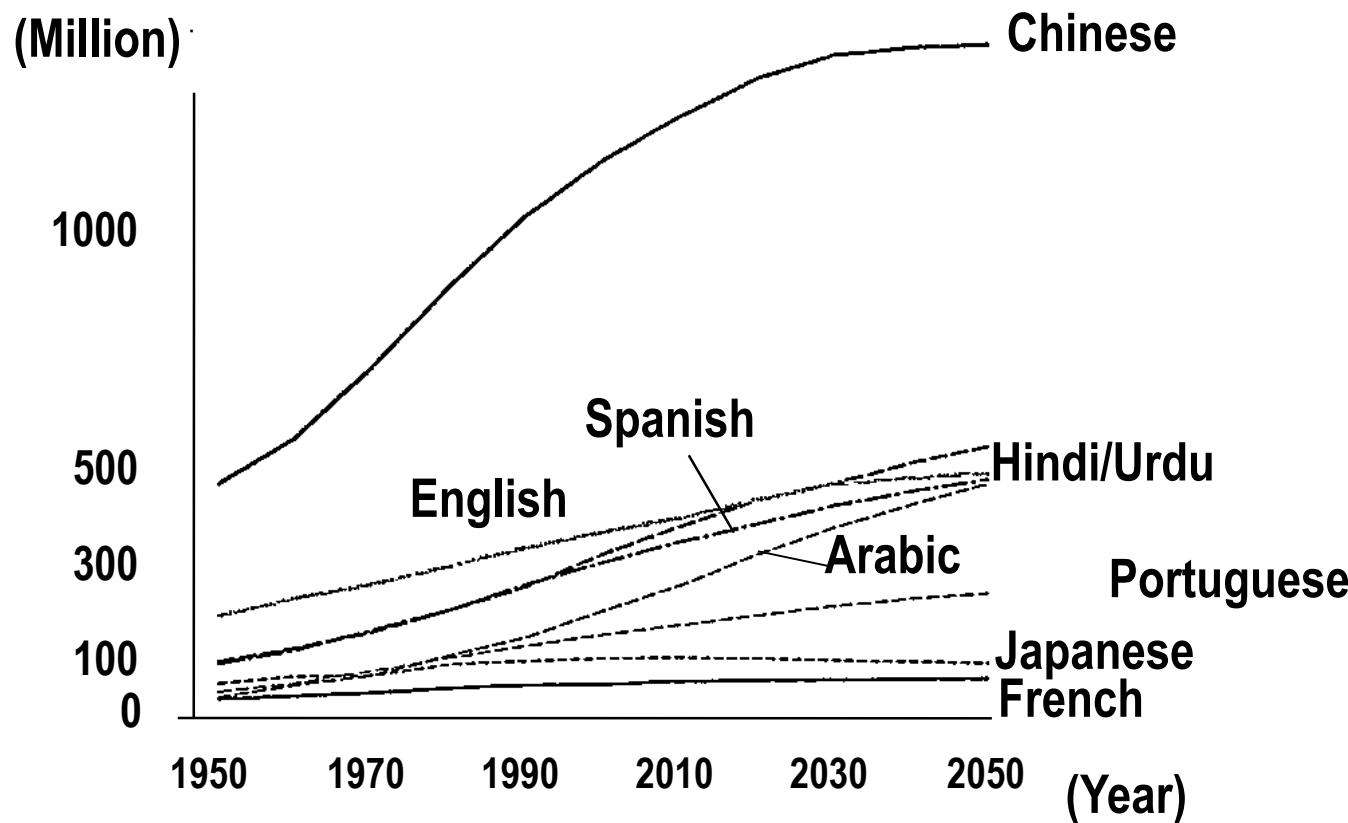
- **自然语言理解**是探索人类自身语言能力和语言思维活动的本质，研究模仿人类语言认知过程的自然语言处理方法和实现技术的一门学科。它是人工智能早期研究的领域之一，是一门在语言学、计算机科学、认知科学、信息论和数学等多学科基础上形成的交叉学科。**(宗成庆, 黄昌宁)**
- **计算语言学**是通过建立形式化的计算模型来分析、理解和生成自然语言的学科，是人工智能和语言学的分支学科。计算语言学是典型的交叉学科，其研究常常涉及计算机科学、语言学、数学等多个学科的知识。与内容接近的学科自然语言处理相比较，计算语言学更加侧重基础理论和方法的研究。**(常宝宝)**
- **自然语言处理**是研究如何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

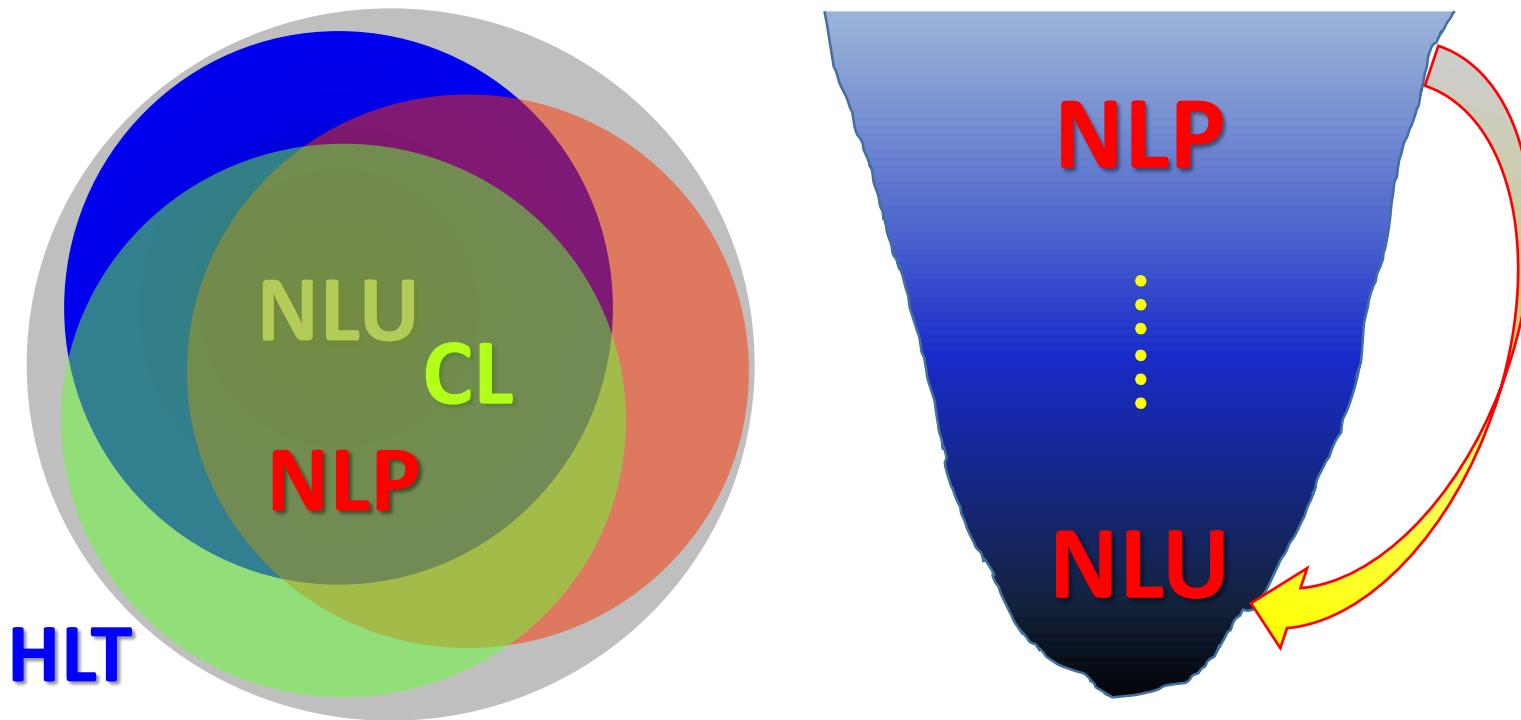
清华大学出版社, 2018.5

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

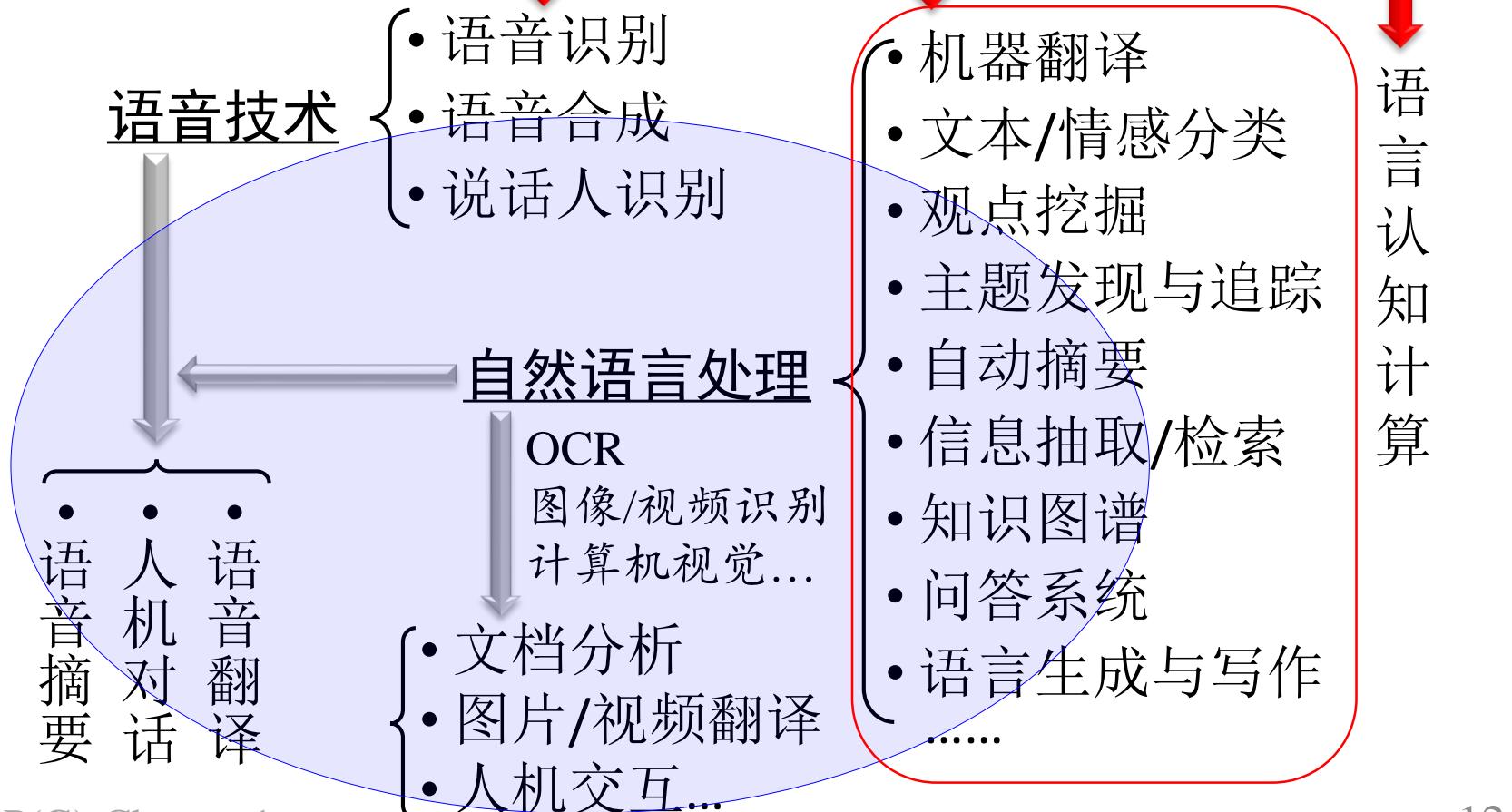
语言的两个基本属性:

语音

+

文字

+ {**神经科学**
语言心理学}



第1章 绪论

本章内容

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

◆ 现实与需求

- ❖ 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
- ❖ 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
- ❖ 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
- ❖ 面对文本**大数据**，我们面临怎样的机遇和挑战？

2. 问题挑战

“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



与138个国家签署合作文件，涉及110多种语言。

一带一路 语言铺路

—光明日报

2. 问题挑战



2. 问题挑战



2. 问题挑战



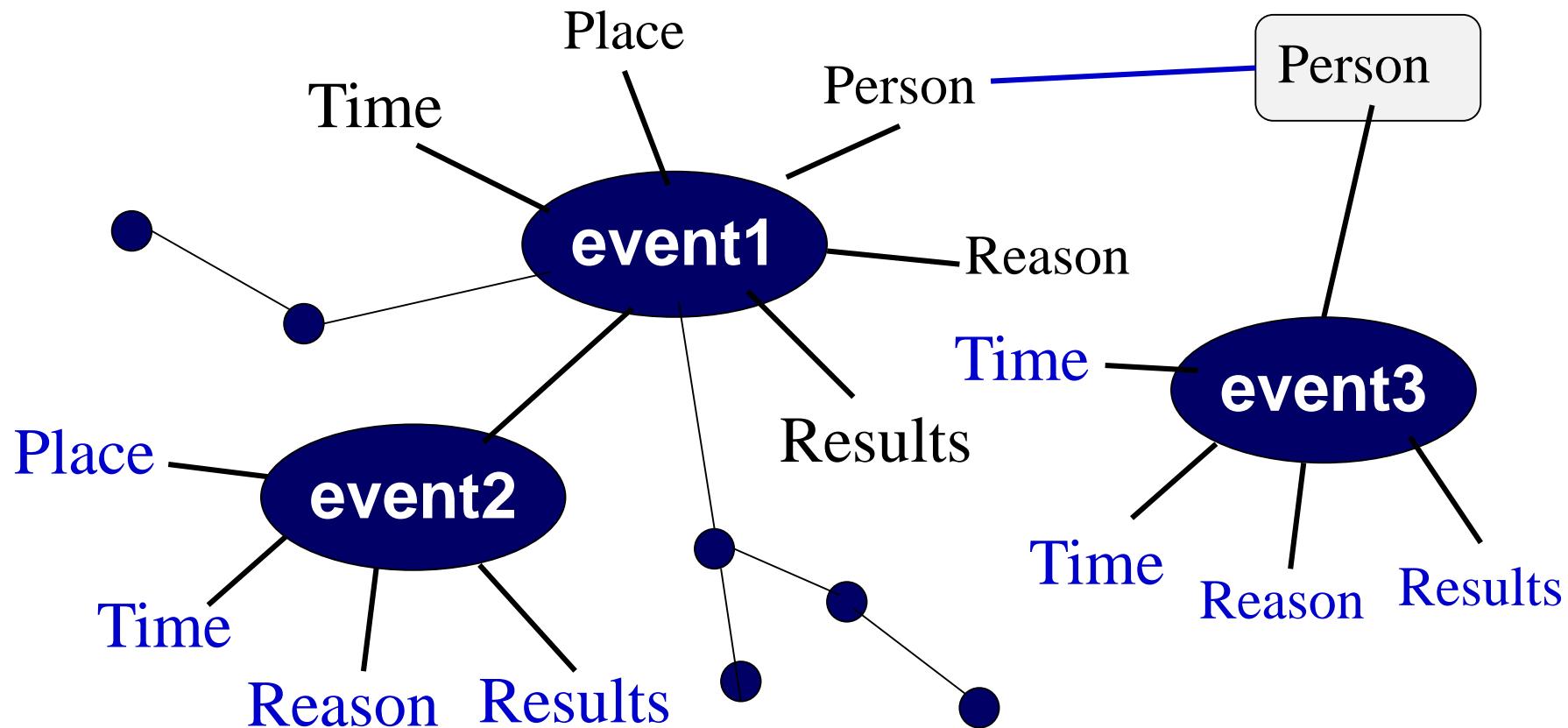
人们处在不同的国家，使用不同的语言，通过不同的手段和方式发表不同的言论(微博/微信/Twitter/Facebook、专著、论文、网页等)，千丝万缕的关系将他们联系在一起，构成一个特定的社会网络。如何发现或挖掘这种网络？如何确定不同的实体、事件和知识之间的关联？



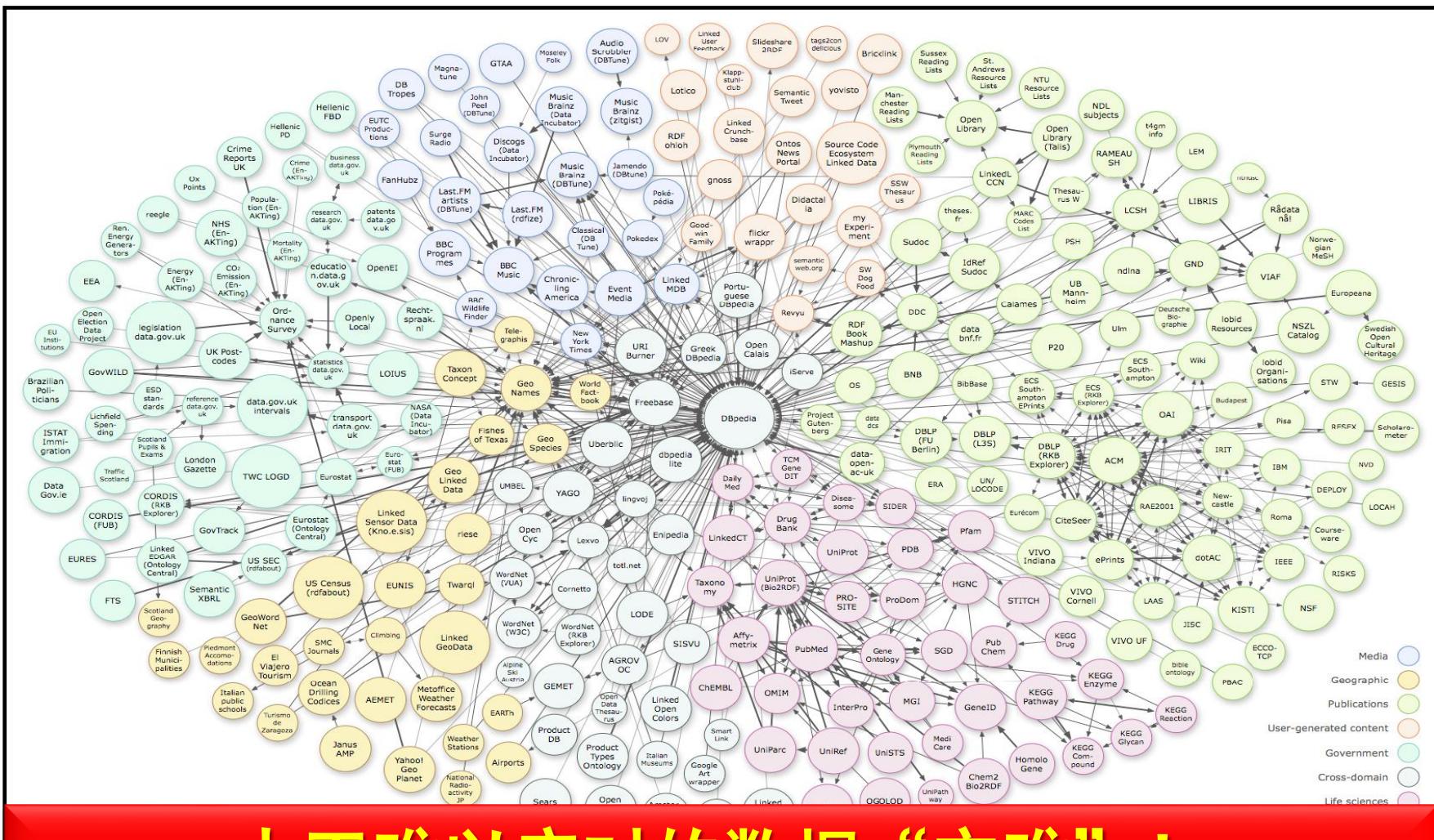
社区论坛

2. 问题挑战

人物、事件关系分析意义重大



2. 问题挑战



人工难以应对的数据“灾难”！

2. 问题挑战

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

文件(E) 查看(V) 收藏(B) 账户(U) 工具(I) 帮助(H)

后退 前进 停止 刷新 主页 恢复 收藏 历史 无痕

http://news.sina.com.cn/s/2012-03-08/233624083315.shtml

MSN.com 电台指南 Microsoft 网站 Windows Live 服务 Coling 2010 Home ACL Anthology -A...

谷歌 截图 下载 收藏正文 >

图片

文本

扫描文档

视频

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

3月1日，天涯社区、广西红豆社区等论坛出现了《“许云鹤案”再陷广西？》的帖子，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2日，天涯社区、广西红豆社区等论坛出现了《“许云鹤案”再陷广西？》的帖子，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侧停车场出来，在大门口我看见一个老人站在马路中间的隔离栏边颤颤巍巍。这时正好在雨加小雨。

热门博客

“红泥人”的奇异婚俗 地狱般的印度陨石矿场

· 买个秦始皇的棺官要多少钱(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 教育 教育 教育

解酒保肝—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... 驱动人生 20:27

CASIA NLP(G)-Chapter 1 21/68

2. 问题挑战

سۈرەتى شەتىلدىك اقشاسنا باسلۇان بىردىن - ئېرى تەگى جۇڭگولقى ادام

ماڭرىتىپىسىنىڭ 25 روپيا قاڭار اقشاسى جۇڭگولقى سۈرەتى باسلۇان

كەلە قابىتارى : خالق تۈزۈلى 16:42 2018.12.12

بىزدە



جۇڭگولقى سۈرەتى باسلۇان ماڭرىتىپىسىنىڭ 25 روپيا قاڭار اقشاسى (ماڭرىتىپىسى سۈرەت)

جۇڭگولقى ادا - ئەگى گۈڭگۈلچىك ولىكسى مېھجۇر قالاسىدەي كىچىا وتابىستان نازىبىدى، ول بۆزىنگە دەپن شەتىلدىك

قاڭار اقشاسنا سۈرەتى باسلۇان بىردىن - ئېرى تەگى جۇڭگولقى ادام، 1991 - جىلى ماڭرىتىپىش حالقى جۇڭگولقى

ماڭرىتىپىش دەلسە ئەتكەن دېڭەن دەستە ساقتاۋ ئۇشىن ونسىك سۈرەتسىن 25 روپيا قاڭار اقشاسا باسقان، سونىمەن بىرگە، جۇڭگولقى



Os mais recentes desenvolvimentos no Afeganistão são, indubitavelmente, um duro golpe para os Estados Unidos e um fracasso total da sua tentativa de transformar aquele país. A reformulação política promovida pelos EUA em solo afgão provou ser incapaz de criar raízes ao longo do tempo. A retirada súbita e desordenada dos Estados Unidos demonstra perante seus aliados a falta de confiabilidade perante os compromissos assumidos: quando os próprios interesses exigem o abandono de aliados, os EUA encontram várias razões para o fazer.



Saddam Hussein is hanged for his crimes



Former Iraqi dictator Saddam Hussein was executed in front of witnesses early this morning in central Baghdad

- Picture of corpse will be released
- Curfew in tyrant's home town
- Troops expecting renewed violence
- Body will get secret burial

A merciless finale for a savage dictator
Troops on alert | Saddam's obituary
Tyrant knew the price of losing power
Iraq deserved better | So what now?

如何让计算机自动分析自然语言文本，理解人的意图，以根据不同用户的需求实现个性化信息服务？

2. 问题挑战

◆问题与挑战

①无处不在的歧义(ambiguity)

❖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

2. 问题挑战

文章标题中的歧义比比皆是：

- ◆ 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)



2. 问题挑战

❖ 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)

2. 问题挑战

❖ 结构歧义

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



2. 问题挑战

(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 :

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

n 为句子中介词短语的个数。

2. 问题挑战

❖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，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；她厉害得简直像个母老虎；全国各族人民要像石榴籽一样紧密地团结在一起。

2. 问题挑战

❖语音歧义：同音字(词)

施氏食狮史

石室诗士施氏，嗜狮，誓食十狮。施氏时时适市视狮。十时，适十狮适市。是时，适施氏适市。施氏视是十狮，恃矢势，使是十狮逝世，氏拾是十狮尸，适石室。石室湿，氏使侍拭石室。石室拭，氏始试食是十狮尸，食时，始识是十狮尸，实十石狮尸。试释是事。

(百度百科)



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。

2. 问题挑战

❖语音歧义：多音字及韵律等歧义

(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/《幸福里的故事》/聊吧/说吧

2. 问题挑战

②大量未知语言现象

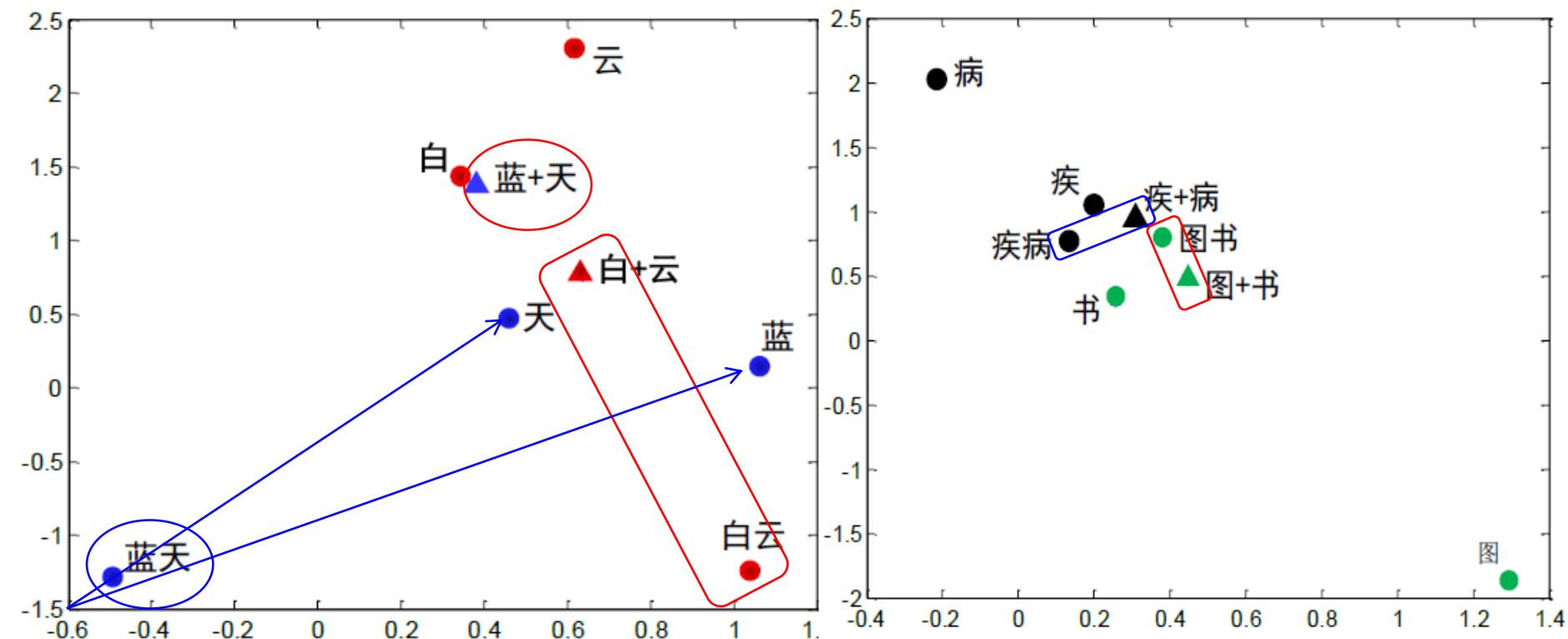
- ❖ 新词、人名、地名、术语等，如：奥特、给力、内卷；非典、新冠肺炎；夏天、高山、温馨、温泉、时光、平安、田野、边疆、程序、吉林、桂林，不来梅
 - ❖ 新含义
如：苹果、奔腾、老虎、后台
 - ❖ 新用法和新句型等，尤其在口语中或部分网络语言中，不断出现一些“非规范的”新的语句结构。如：被长工资，很中国，百度一下……

2. 问题挑战

③语义表示与计算面临太多的困难

$$\begin{cases} \text{蓝} + \text{天} \neq \text{蓝天} \\ \text{白} + \text{云} \neq \text{白云} \end{cases}$$

$$\begin{cases} \text{疾} + \text{病} \approx \text{疾病} \\ \text{图} + \text{书} \approx \text{图书} \end{cases}$$



2. 问题挑战

$$\begin{array}{c} \text{火} \\ \left[\begin{matrix} 0.15 \\ \cdot \\ \cdot \\ \cdot \\ 0.21 \end{matrix} \right] \\ ? \end{array} \quad \begin{array}{c} \text{药} \\ \left[\begin{matrix} 0.23 \\ \cdot \\ \cdot \\ \cdot \\ 0.19 \end{matrix} \right] \\ = \end{array} \quad \begin{array}{c} \text{火药} \\ \left[\begin{matrix} 0.18 \\ \cdot \\ \cdot \\ \cdot \\ 0.06 \end{matrix} \right] \end{array}$$

山 \oplus 药 = 山药 ?

语义计算不是简单的数字运算，也不是逻辑运算，更不是“赌徒式”的概率推算，而应该是可以溯源、可以举一反三和产生新概念的新的运算系统。



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: 有限的样本集合无法涵盖开放的语言现象
- 知识表示的复杂性: 语义知识(包括常识)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, 语义表示和计算困难巨大
- 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

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

◆不同的语言有其独特的问题

● 三大语系

- ❖ **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**(isolating language): 又称分析语(analytic language)，几乎没有形态变化，语法关系靠词序和虚词表示，如汉语、苗语、越南语等。

2. 问题挑战

◆ 关于“理解”的标准

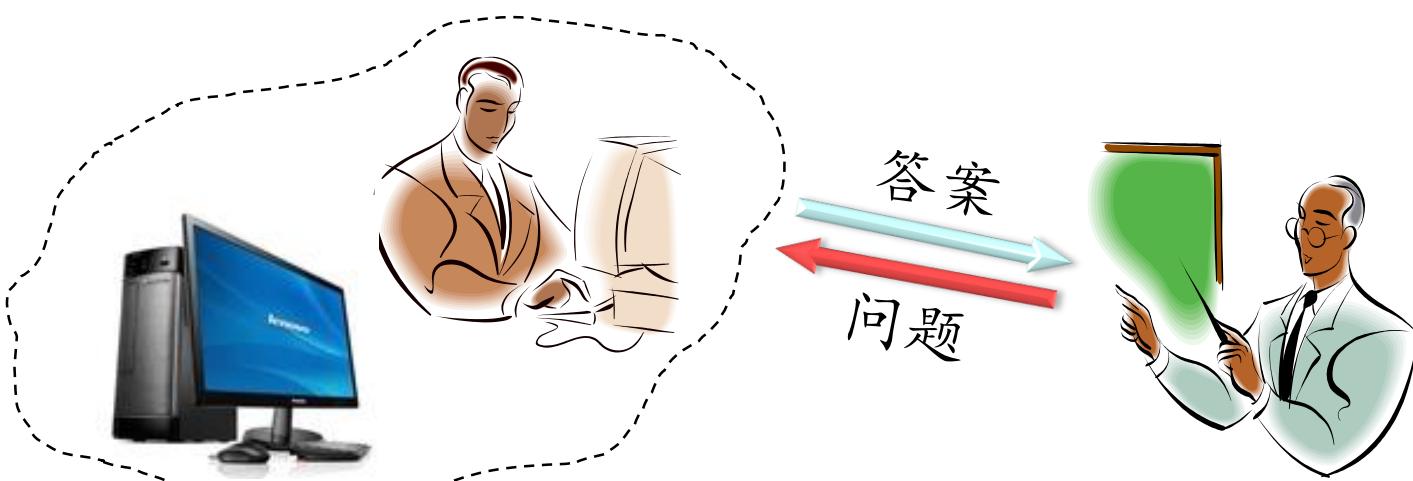
● 如何判断计算机系统的智能？

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

◆人脑理解语言是个复杂的思维过程



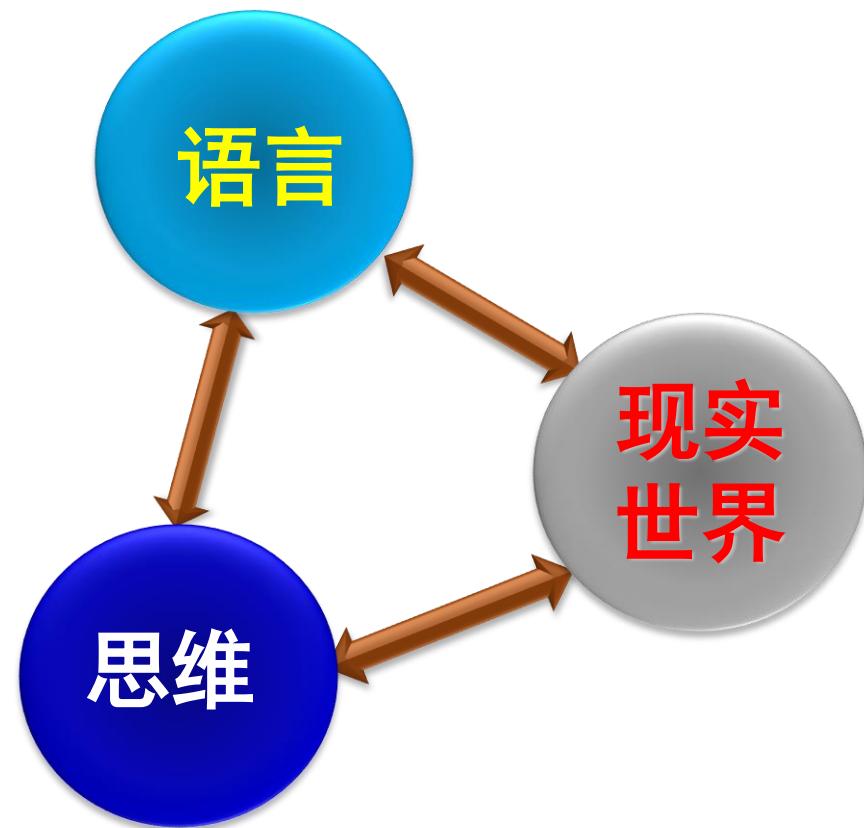
什么意思？

2. 问题挑战

◆人脑理解语言是个复杂的思维过程

- 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

... ...



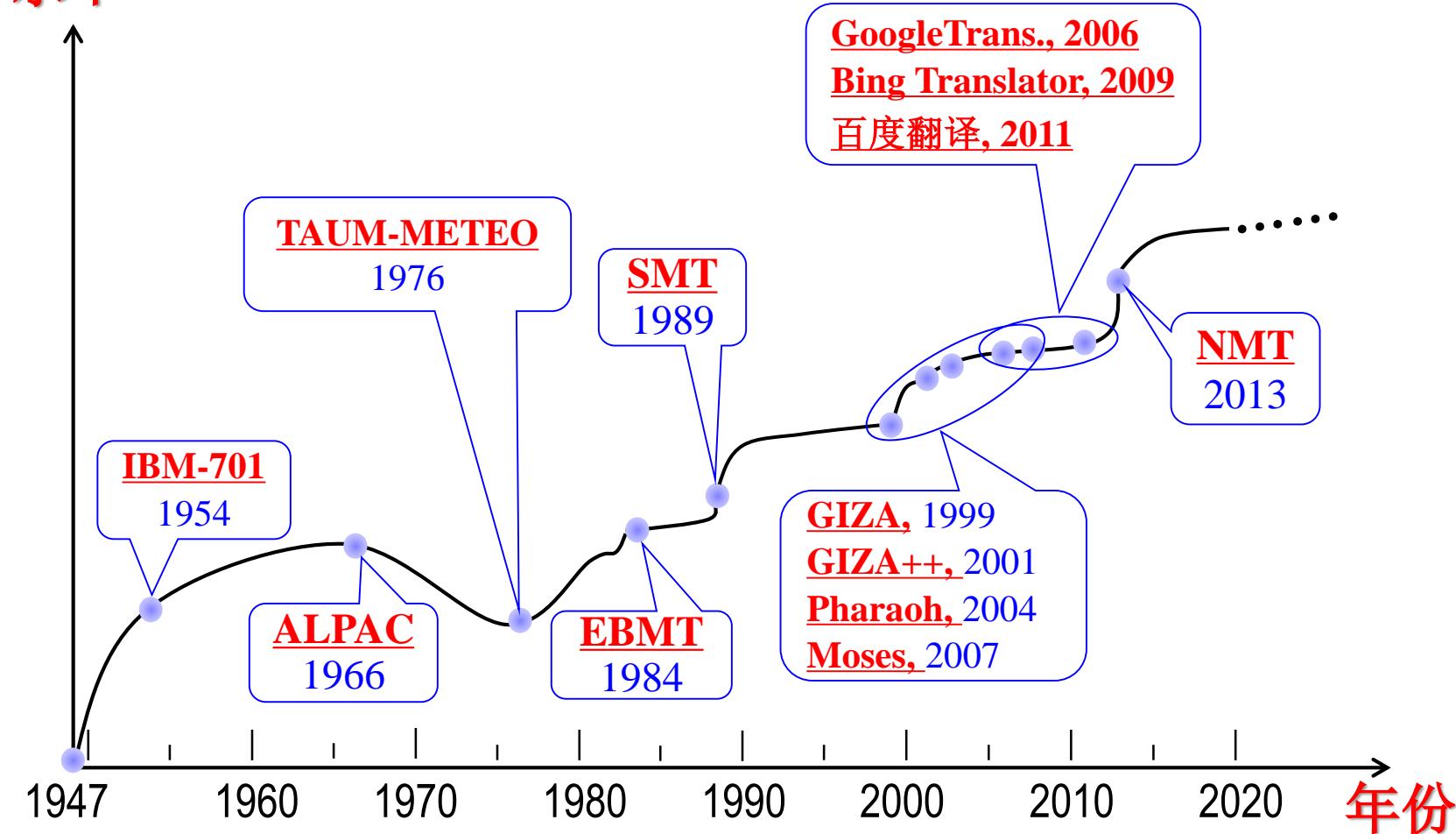
第1章 绪论

本章内容

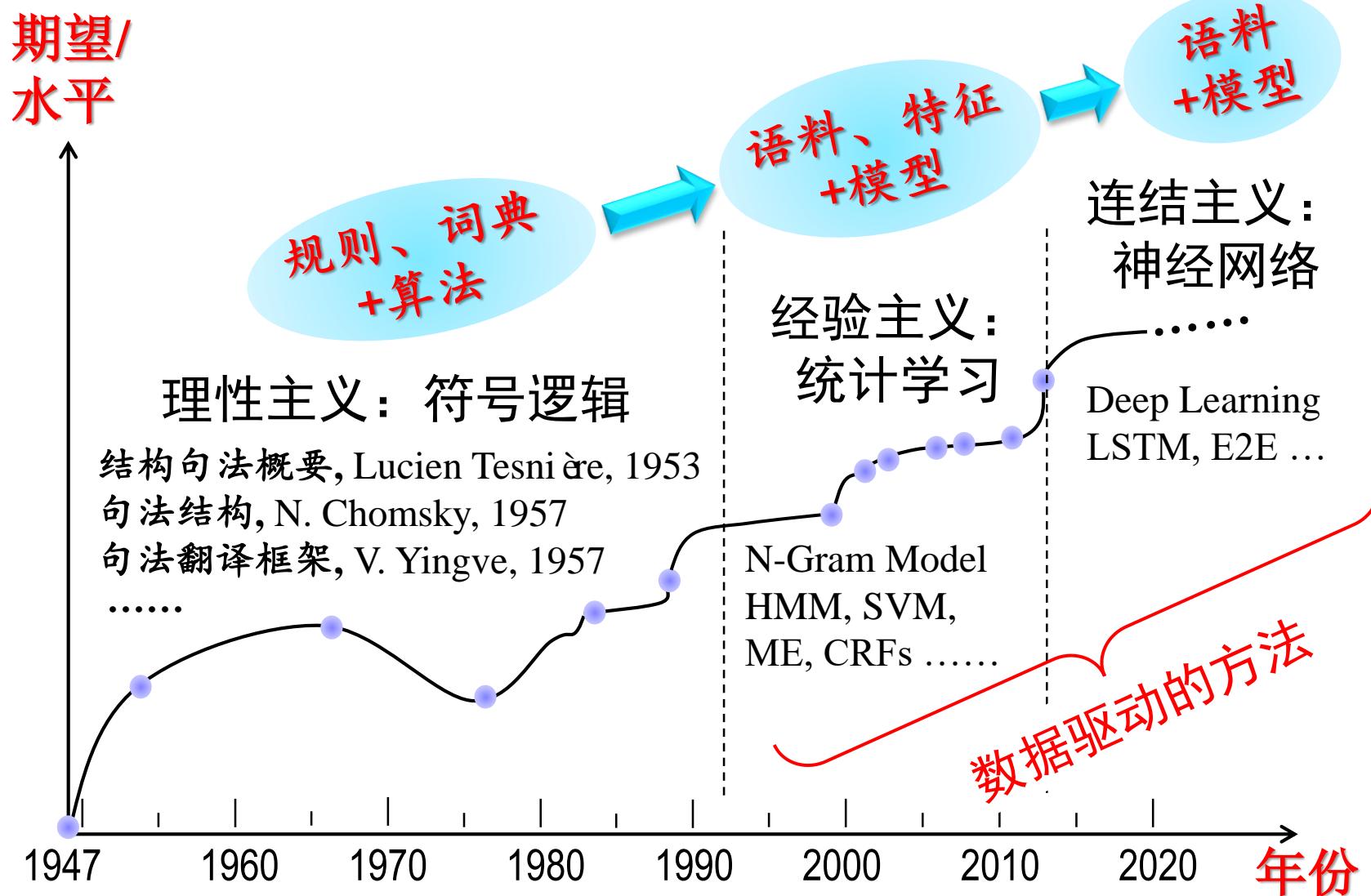
1. 基本概念
2. 问题挑战
-  3. 技术方法
4. 课程内容
5. 参考文献
6. 习题

3. 技术方法

期望/
水平



3. 技术方法



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

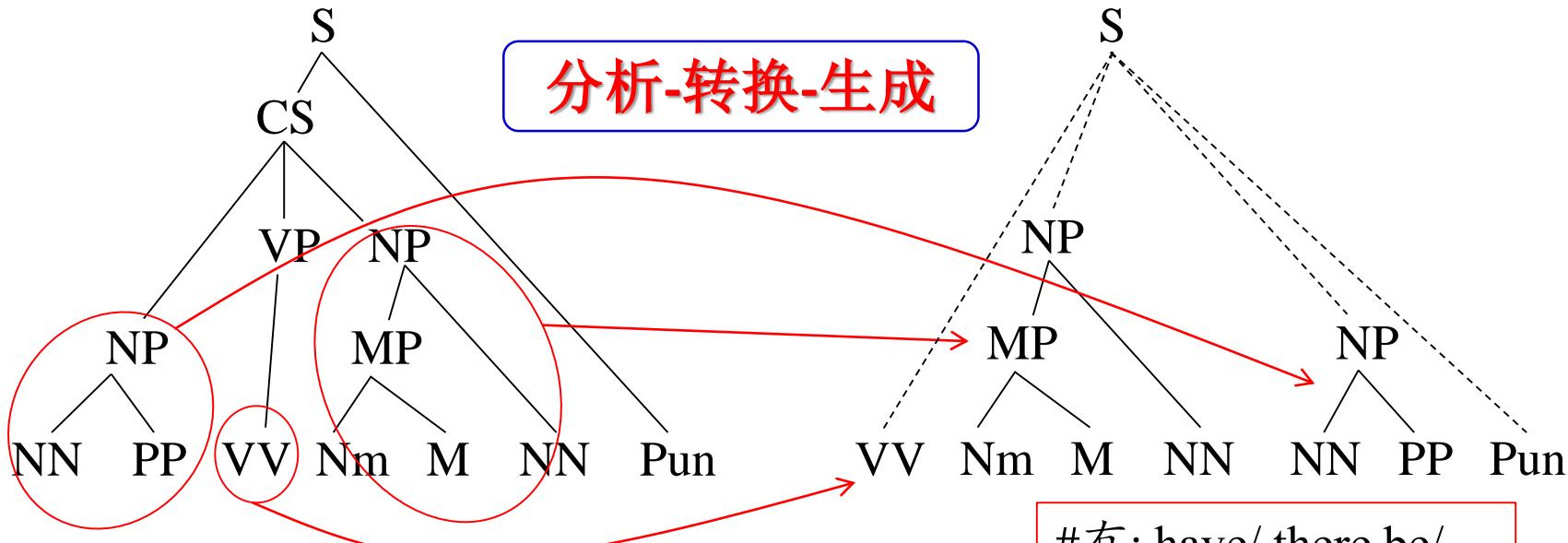
◆以机器翻译为例 ①基于规则的方法

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

#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。对于很多对语言来说，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: $T = t_1^l \equiv t_1 t_2 \cdots t_l$

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

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

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 apparatirining 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。容易与人工翻译的经验（规则、词典等）相结合。
- 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神经网络方法

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

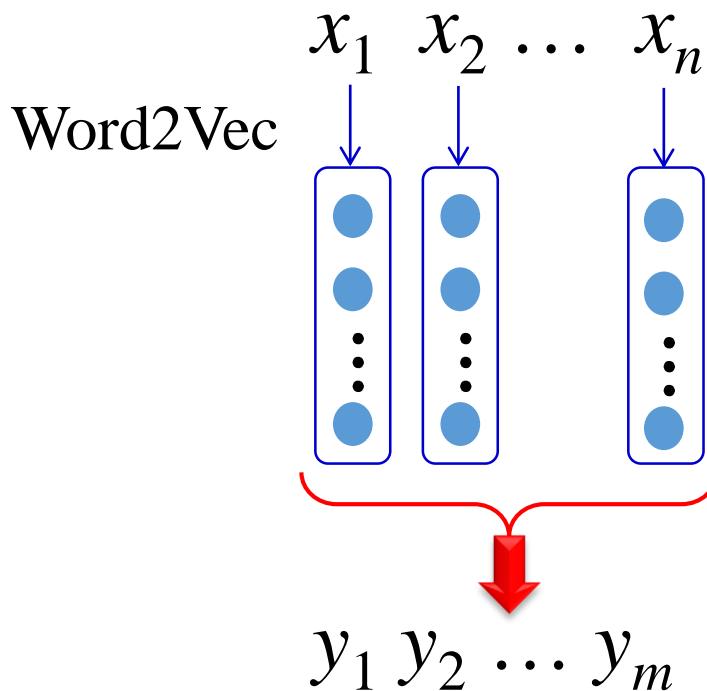
将其翻译成目标语言句子: $E = e_1^m \equiv e_1 e_2 \cdots e_m$

词汇向量化表示: Word2Vec

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

$$\text{目标函数: } L = \sum_i \log(P(e_i | C))$$

3. 技术方法



$$h_t = \tanh(Wx_t + Uh_{t-1} + b)$$

权重1 权重2 偏置
 ↓ ↓ ↓
 W U b
 t 时刻 t 时刻 t-1 时刻
 隐层向量 词向量 隐层向量

$$Score = \text{softmax}(Y)$$

端到端(End-to-end, E2E)的翻译方法

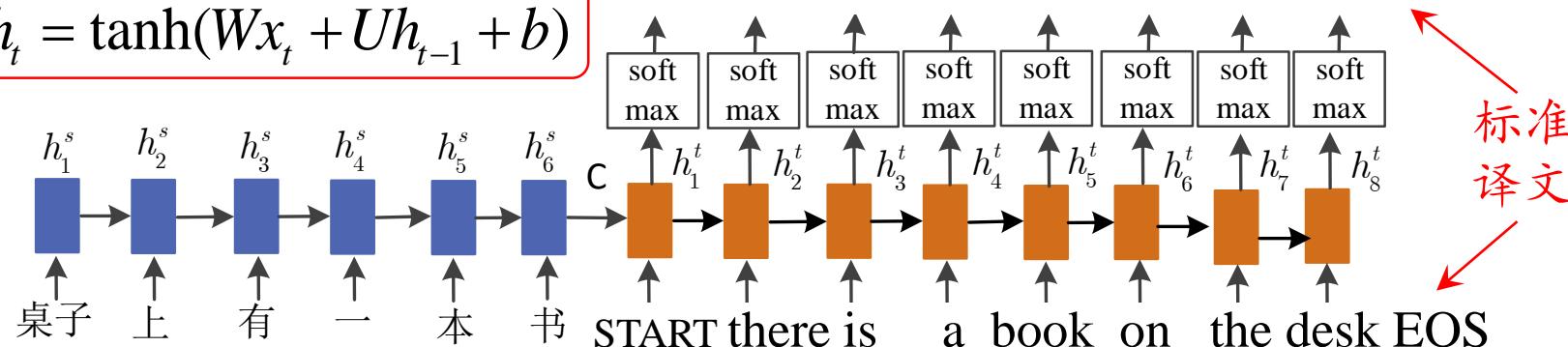
- （ 桌子上有一本书。
There is a book on the desk.
- （ 今天是星期三。
Today is Wednesday.
- （ 明天他将来北京。
Tomorrow he will come to Beijing.
-

3. 技术方法

➤ 训练阶段：利用梯度下降法，最大化标准译文的预测概率。

$$h_t = \tanh(Wx_t + Uh_{t-1} + b)$$

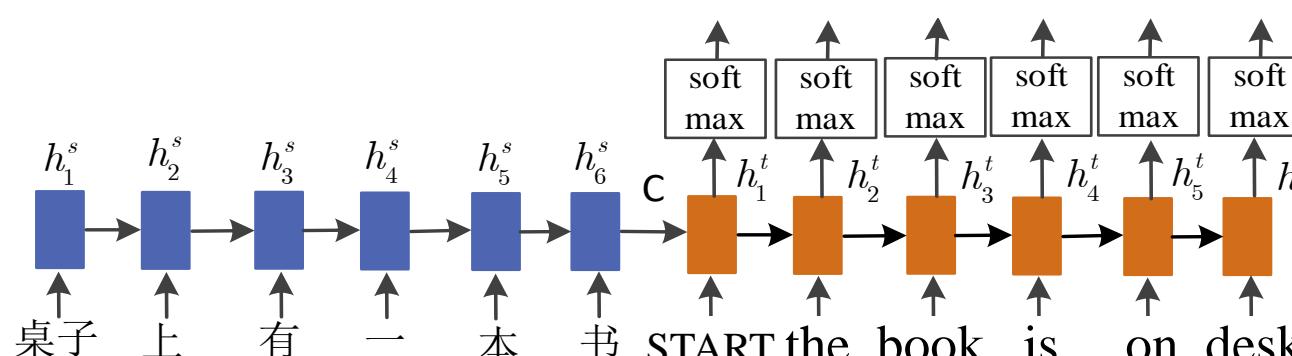
there is a book on the desk EOS



标准
译文

➤ 测试阶段：利用训练好的模型，每次选择概率最高的单词，直到预测出EOS为止。

the book is on desk EOS



模型自身预
测的最高概
率单词

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就可以建立一个机器翻译系统, 系统开发周期短。译文的流畅性较好。
- 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, 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, 可能产生语义错误、缺失或“无中生有”的译文, 翻译过程无法解释。难以与人工翻译经验(规则、词典等)相结合。

3. 技术方法

◆ 端到端的翻译模型推广应用

- 任意语言对之间的自动翻译
- 自动问答系统: Q&A
- 人机对话系统: Turn: Utterance 1 ~ Utterance 2
- 自动文摘: Document → Summary
- 文本/情感分类: Document / Sentence → Sentiment label
- 图像标注/ 看图说话: Image → Caption

.....

◆ 注意力机制(Transformer...)

◆ 预训练模型(BERT, GPT1~3 ...)

第1章 绪论

本章内容

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

◆ 技术实践

- 平日作业：方法实现/开源工具使用 + 技术报告

◆ 课程成绩

- 闭卷考试 (60%)
- 技术实践 (30%)
- 课堂考勤 (10%)



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 语义分析

第11章 篇章分析

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

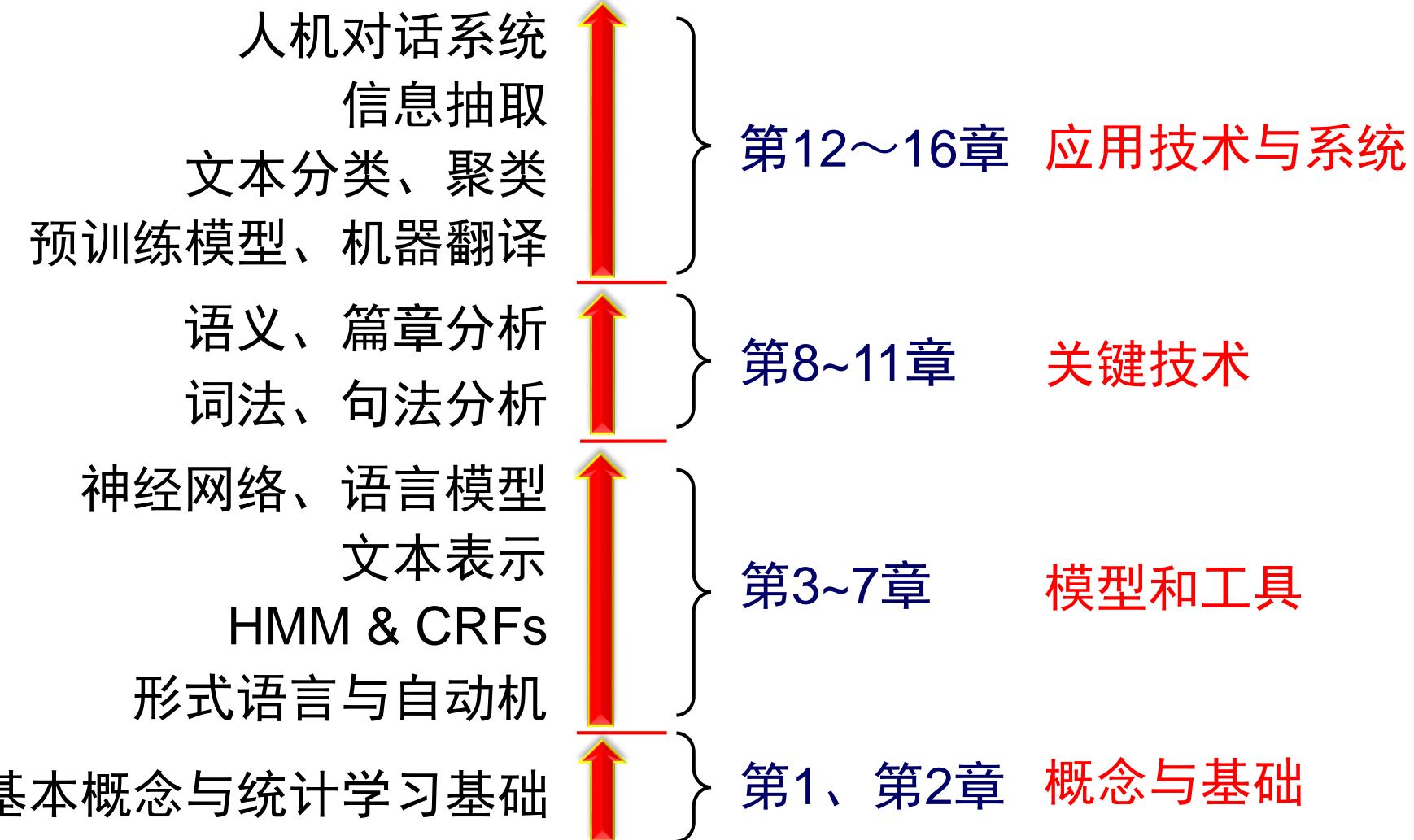
第15章 信息抽取

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



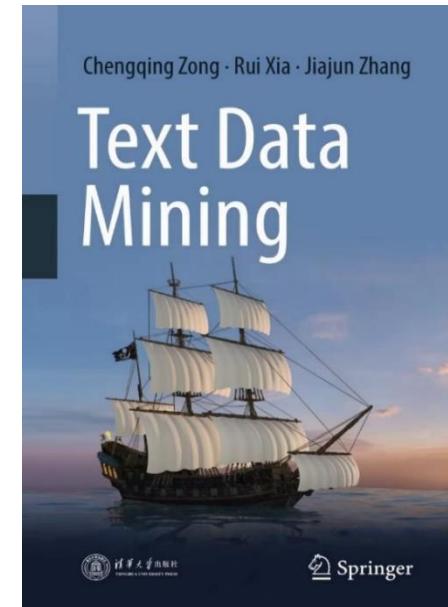
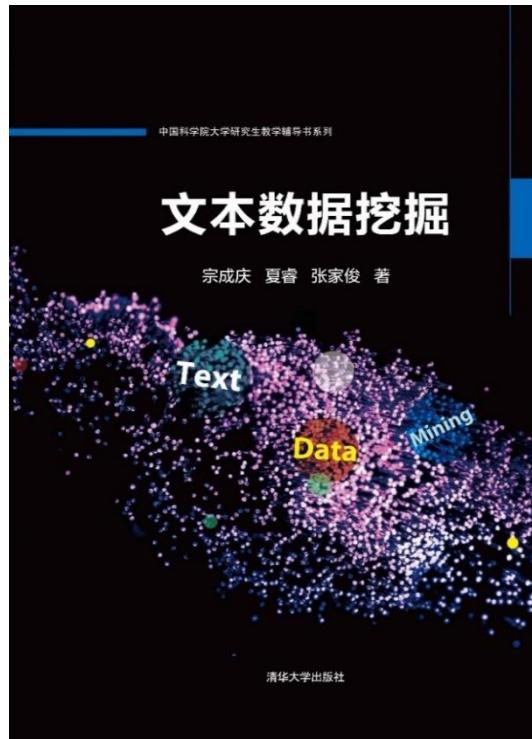
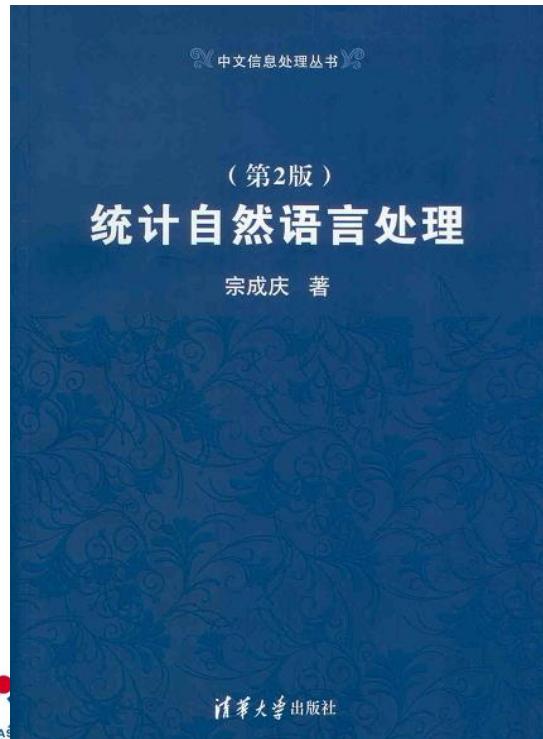
第1章 绪论

本章内容

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

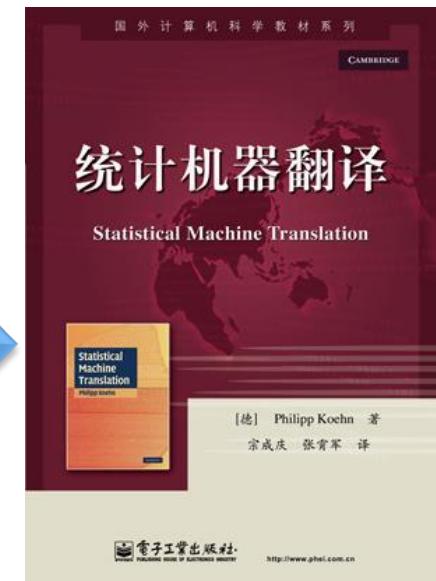
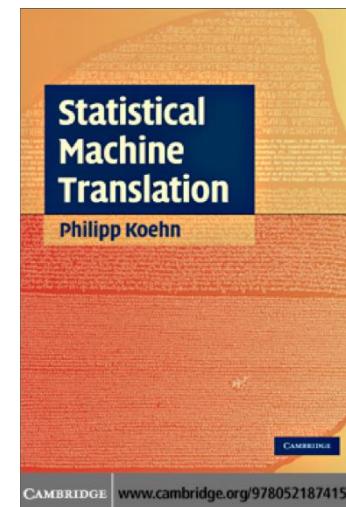
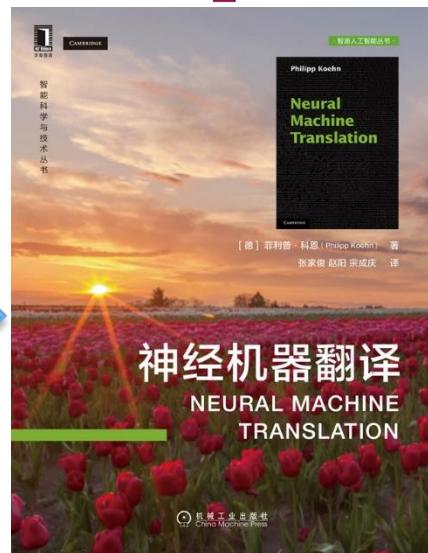
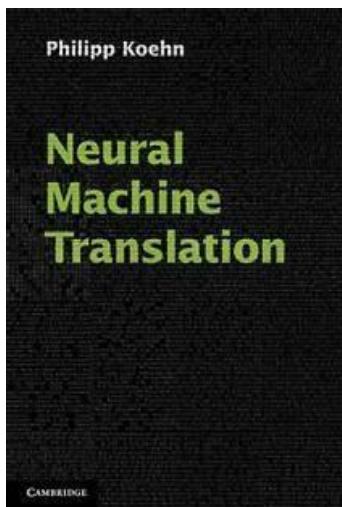
- [1] 宗成庆, 统计自然语言处理(第2版), 清华大学出版社, 2013.8
- [2] 宗成庆, 夏睿, 张家俊, 文本数据挖掘, 清华大学出版社, 2019.5
- [3] C. Zong, R. Xia, J. Zhang, Text Data Mining, Springer, May 2021
- [4] 宗成庆, 张霄军(译), 统计机器翻译, 电子工业出版社, 2012.9
- [5] 张家俊, 赵阳, 宗成庆(译), 神经机器翻译, 机械工业出版社, 2022.3



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

5. 参考文献

张家俊、赵阳、宗成庆译，神经机器翻译，机械工业出版社，2022.3



宗成庆，张霄军译，统计机器翻译，电子工业出版社，2012.9



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>)

第1章 绪论

本章内容

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

1-1. 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。

1-2. 试举例比较汉英句子的差异。

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

1-4. 试列举不少于10种自然语言处理技术应用的场景。

1-5. 通过对比测试Google 翻译系统、微软Bing翻译系统和百度等翻译系统，了解机器翻译技术的性能现状。



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...
- ◆ 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： 三本专著+两部译著； 美国3所大学的NLP课程

谢谢！

Thanks!

