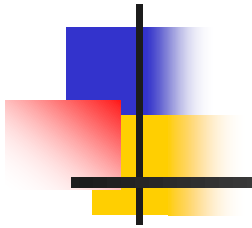


第一讲 绪论：什么是计算语言学

詹卫东

zwd@pku.edu.cn

提纲

0　引子

1. 计算语言学的研究内容
2. 计算语言学的研究方式
3. 计算语言学的应用领域
4. 计算语言学的发展简史

0 引子

我们可以期待，总有一天机器会同人在所有的智能领域里竞争起来。但是，如何开始呢？这是一个很难决定的问题。许多人以为最好从下棋之类的极为抽象的活动开始，不过，还有一种办法也应加以考虑，就是为机器配备容易买到的传感器，然后教这种机器理解并且说英语。这可以仿效通常小孩子学话的方式进行，比如以手指物称名等。我不能确定到底哪个出发点更好，但应该都值得一试。

—— A. M. Turing, *Computing Machinery and Intelligence*, *Mind* , Vol.59, 1950.
阿兰. 图灵 《计算机器与智能》

交谈

甲：听说吴妈打赢了阿Q。

乙：不错，阿Q确实被吴妈打败了。

甲：这个结果有些出人意料。

乙：阿Q是大意失荆州，怪不得别人。

图灵测试（Turing Test）

- 一个封闭的小屋里面有两间房A、B，房中各有一人；
- 其中一间房中是男人（M），一间房中是女人（W）；
- 屋外有一个人C；
- C向A，B房内的人提出各种问题；
- C根据来自A，B的回答判断A，B中谁是M，谁是W；
- A的任务是欺骗C，B的任务是帮助C；
- C如果判断出A和B的性别，则C获胜，否则A获胜。

《模仿游戏》（Imitation Game）

重温“图灵测试”：写在ChatGPT发布一周年
https://mp.weixin.qq.com/s/hgJ1qgnzy3iZaqcj4U81Xg

图灵测试（Turing Test）

- 一个封闭的小屋里面有两间房A、B，房中各有一人；
- 其中一间房中是人（H），一间房中是机器（M）；
- 屋外有一个人C；
- C向A，B房内的实体提出各种问题；
- C根据来自A，B的回答判断A，B中谁是H，谁是M；
- 如果C无法判断，则证明“机器有思考能力”。
（标准：测试时间5分钟，且人类正确率低于70%）

《模仿游戏》（Imitation Game）

Eliza问答系统

（by Joseph Weizenbaum, MIT, early 1960s）

Young woman: Men are all alike.

Eliza: In what way?

Young woman: They're always bugging us about something specific or other.

Eliza: Can you think of a specific example?

Young woman: Well, my boyfriend made me come here.

Eliza: Your boyfriend made you come here?

Young woman: He says I'm depressed much of the time.

Eliza: I am sorry to hear you are depressed.

Young woman: It's true. I'm unhappy.

Eliza: Do you think coming here will help you not to be unhappy?

Young woman: I need some help, that much seems certain.

Eliza: What would it mean to you if you got some help?

Young woman: Perhaps I could learn to get along with my mother.

Eliza: Tell me more about your family.

Young woman: My mother takes care of me.

Eliza: Who else in your family takes care of you?

Young woman: My father.

Eliza: Your father.

Young woman: You are like my father in some ways.

Eliza: What resemblance do you see?

Searle的“Chinese Room”1980s

哲学家约翰·西尔勒(John R. Searle)否认可以模拟大脑算法的计算机具有意识，他在上个世纪80年代给出了一个著名的理想实验“**西尔勒的中文屋子**”来支持他的论点。

实验的要旨是这样的：

一个用中文讲的笑话，中国人听了会哈哈大笑。如果用一台可以执行复杂算法的计算机根据这个笑话的词句进行一系列运算，得出的结果正确，也是“哈哈大笑”，于是计算机用它自己的方式笑了几声，我们是否可以认为这台计算机“理解”了这个笑话呢？

类似的，用一大群不懂中文的美国人取代计算机，他们拿着铅笔和纸重复计算机所做的一切，因为算法很复杂，可能要全美不懂中文的美国人算上一年才得到了结果“哈哈大笑”，他们派一个代表出来笑了几声。虽然反应很慢，但他们和一个中国人做得一样好，不过，这样仍然无法认定这群美国人“理解”了这个中文笑话。

Searle, John. R. (1980) Minds, brains, and programs. In *Behavioral and Brain Sciences* 3 (3): 417-457

关于语言，可以问些什么？

（1）人用来交际的“语言”具有什么样的性质？这些性质又是如何影响交际过程的？

（2）人用来交际的“语言”跟机器可以“理解”的语言有什么样的关系？

（3） 人是如何运用“语言”进行交际的？

（4）人运用语言进行交际的过程是否可以描述为一个机械的过程？

（5） 什么叫做“理解”一种语言？

……

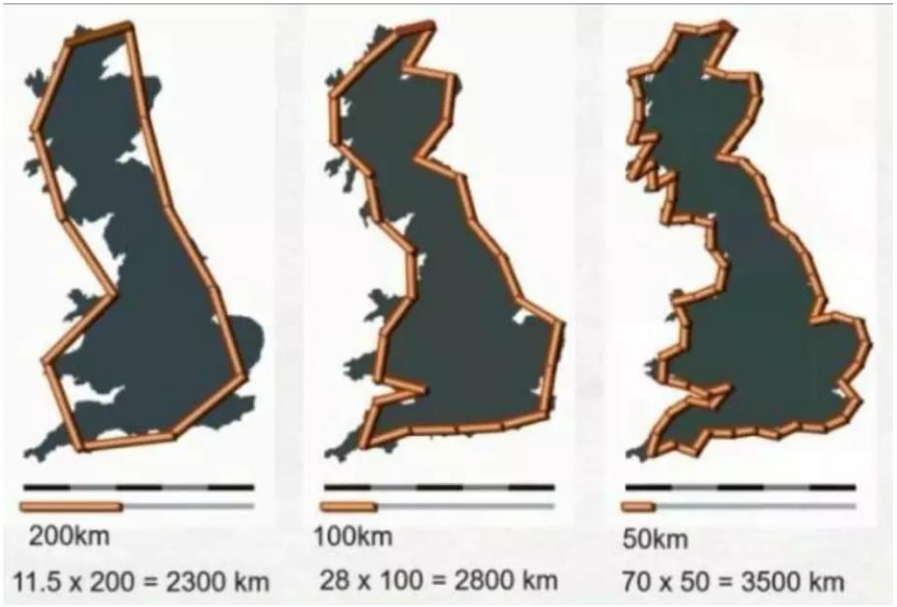
机器语言 vs. 自然语言

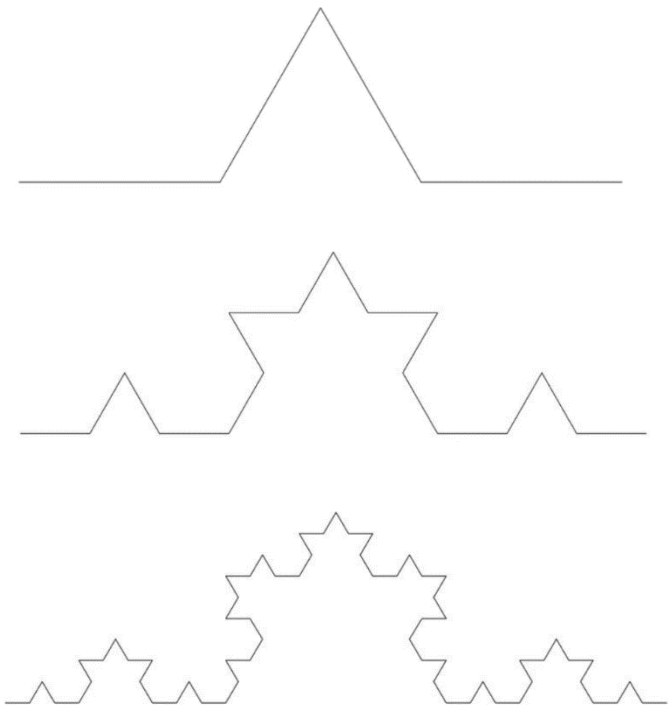
```
#include <stdio.h>

void main ()
{
        int x=1, y=2, z;
        z = x+++y;          x = 2; y = 2; z = 3
        z = x+++++y;
        z = x++ + ++y;      x = 3; y = 3; z = 5
        z = x+++ +y;        x = 4; y = 3; z = 6
        z = x + (++y);      x = 4; y = 4; z = 8

        printf("z=%d\n",z);  z = 8
}
```

人们以为他对她有“**意思**”，于是，建议他对她“**意思意思**”。他说，他没那种“**意思**”。她则反问，[illegible]是什么“**意思**”。大伙[illegible]有的觉得很有“**意思**”，有的则认为真没“**意思**”。

封闭？开放？

- 英国的海岸线有多长？





Benoit Mandelbrot, 1967, How Long Is the Coast of Britain, Science, Vol.156, No.3775, pp.636-638.

https://zhuanlan.zhihu.com/p/150584539

定义

计算语言学（**Computational Linguistics**）指的是这样一门学科，它通过建立形式化的数学模型，来分析、处理自然语言，并在计算机上用程序来实现分析和处理的过程，从而达到以机器来模拟人的部分乃至全部语言能力的目的。

1 计算语言学的研究内容

- 从计算的角度来研究语言的性质
- 将语言作为计算对象来研究相应的算法

1.1 从计算角度研究语言

所谓从计算的角度来看语言的性质，就是要求将人们对语言的结构规律的认识以精确的、形式化的、可计算的方式呈现出来，而不是像其他语言学研究那样，在表述语言的结构规律时一般采用非形式化的表达形式。

例子

- 张三赶跑了李四
- 张三把李四赶跑了
- 李四被张三赶跑了
- 吴妈以前很喜欢阿Q的理论
- * 吴妈把阿Q的理论以前很喜欢
- * 阿Q的理论被吴妈以前很喜欢

语法规律

1）汉语中的一个基本句型是：

P_0: X ＋ 动词 ＋ Y

2）P_0可以变换为“把”字句或“被”字句

P_1: X+把+Y+动词

P_2: Y+被+X+动词

3）有些时候P_0可以变换为P_1,P_2;

有些时候P_0不可以变换为P_1,P_2;

1.2 将语言作为计算对象

所谓将语言作为计算对象来研究相应的**算法**，是研究如何以机械的、规定了严格操作步骤的程序来处理语言对象（主要是自然语言对象，当然也可以是形式语言对象），包括一个语言片断（比如词组、句子或篇章）中大小语言单位的识别，该语言片断的结构和意义的分析（自然语言理解），以及如何生成一个语言片断来表达确定的意思（自然语言生成），等等

算法（algorithm）

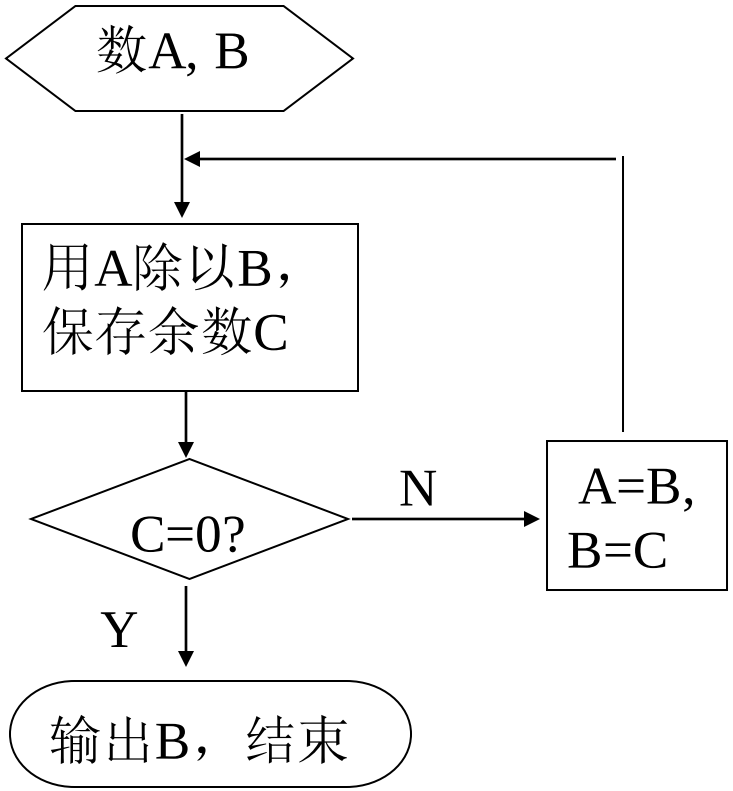
（1）通用性：算法是针对一类问题的，而不仅仅是用于解决某一个具体问题。

（2）机械性：算法的每一个步骤都是机械的，确定的。

（3）有限性：算法必须在有限步内结束。

（4）离散性：算法的输入数据及输出数据都是离散的符号。

一个算法实例：求最大公约数



算法的时间效率和空间效率

- $30 \times (1 + 0.06)^{20} = \ ?$

引自 赵瑞清、孙宗智 编著 《计算复杂性概论》 气象出版社1989年版

算法1：

令 a = 1.06

$30 \times a \times \ldots \times a$

（20个a）

20次乘法

算法2：

令 a = 1.06

1. $a^2 = a \times a$
2. $a^4 = a^2 \times a^2$
3. $a^8 = a^4 \times a^4$
4. $30 \times a^8 \times a^8 \times a^4$

6次乘法，1个寄存器

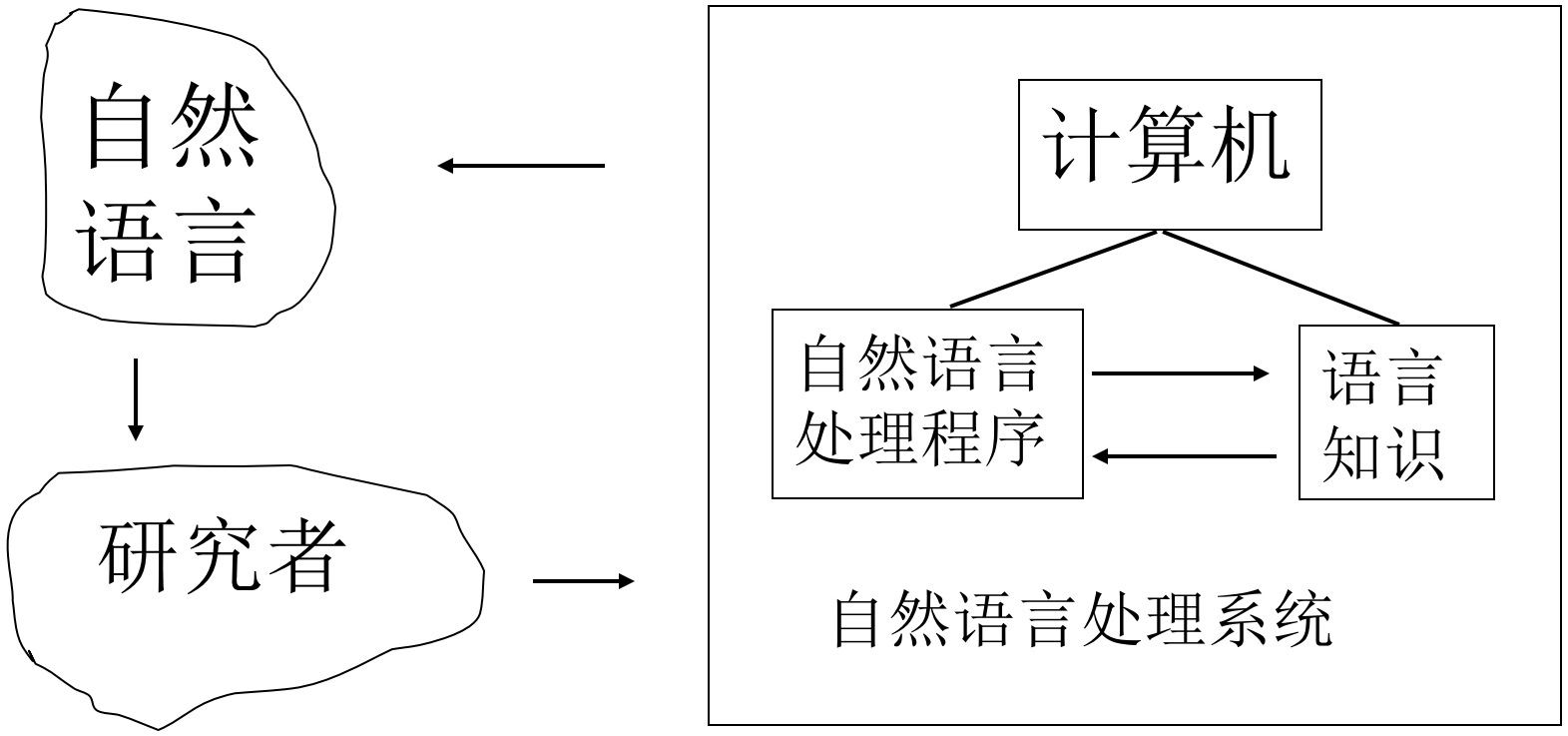
算法3：

令 a = 1.06

1. $a^2 = a \times a$
2. $a^4 = a^2 \times a^2$
3. $a^5 = a^4 \times a$
4. $a^{10} = a^5 \times a^5$
5. $a^{20} = a^{10} \times a^{10}$
6. $30 \times a^{20}$

6次乘法，无寄存器

2 计算语言学的研究方式



动态视角（流程）

S1： 研究者以特定的方式对自然语言（NL_0）的规律进行抽象，以计算机能够处理的形式来表述关于自然语言的规律——得到语言知识K；

S2： 针对特定的语言知识表示形式，研制适合的分析和处理算法；

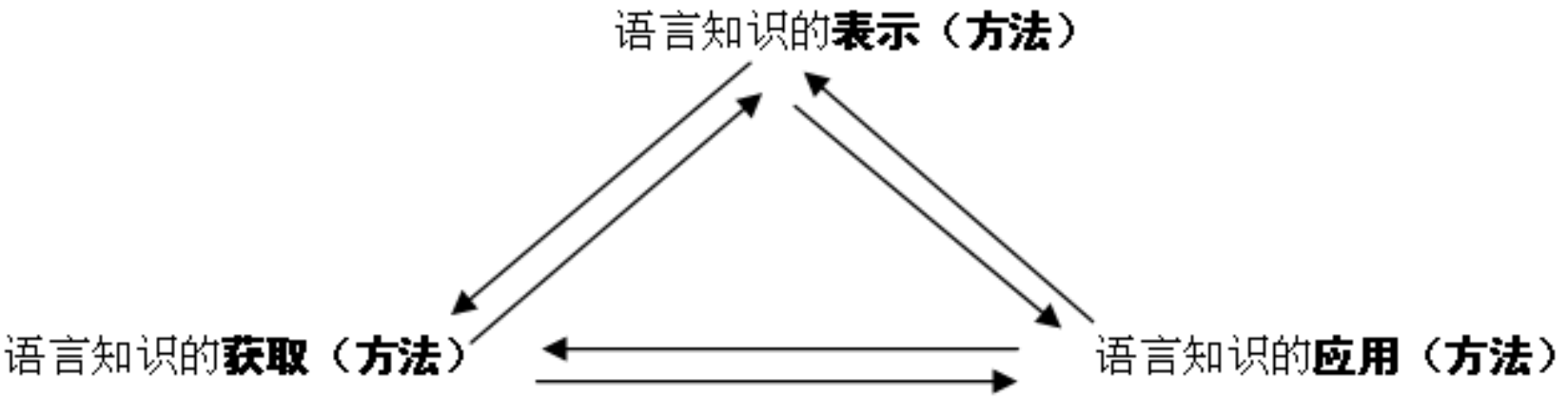
S3： 根据算法编制计算机可执行的自然语言处理程序P。这样的程序加上语言知识，加上计算机硬件系统，共同构成一个自然语言处理系统（NLPs）；

S4： 用这样一个自然语言处理系统对自然语言NL_0进行分析处理，根据反馈的结果调整原来的设计，改进NLPs。

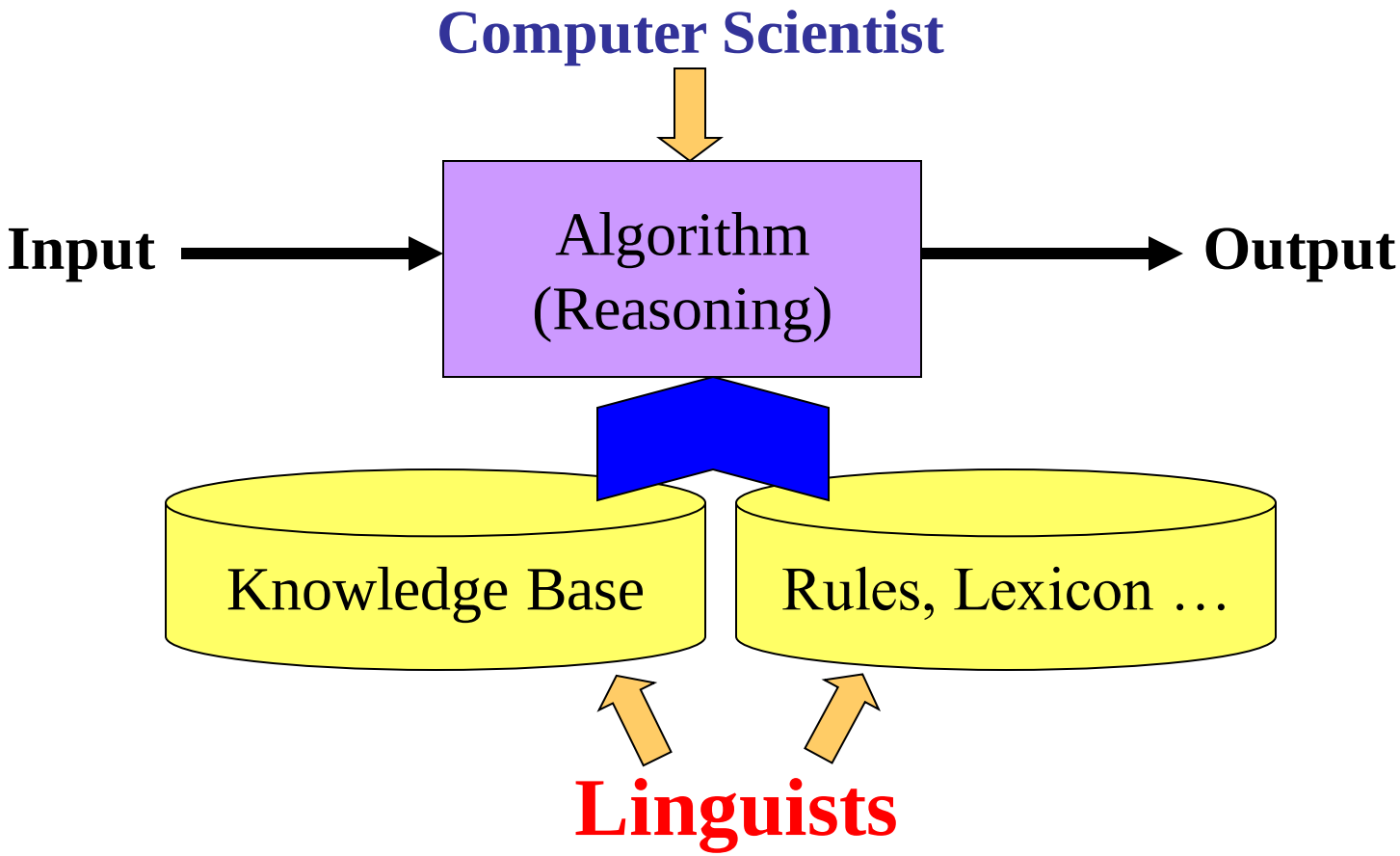
静态视角（模块）

- 语言对象　　语音　字　词　词组　句子　篇章
- 语言知识　　音系　形态　句法　语义　语篇
- 处理程序　stemmer　annotator　parser　translator …

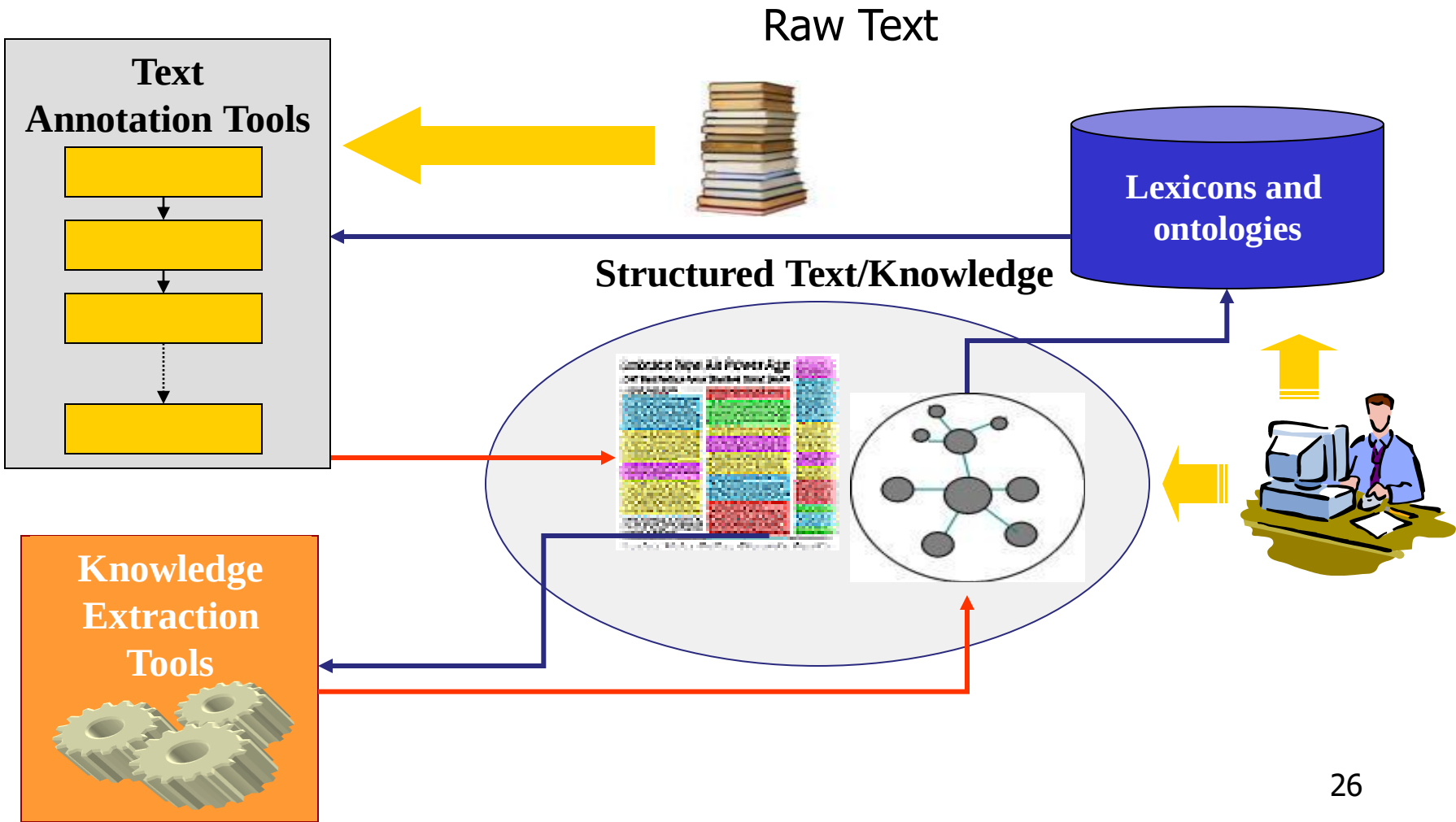
从语言知识的角度看计算语言学的框架



语言知识的获取方式（1）人工构造



语言知识的获取方式（2）知识挖掘



3 计算语言学的应用领域

- 机器翻译（Machine Translation）
- 文本分类（Text Classification）
- 信息检索（Information Retrieval）
- 信息提取（Information Extraction）
- 语音合成（Speech Synthesis）
- 语音识别（Speech Recognition）
- 人机接口（Human-Machine Interface）
- ……

机器翻译的例子

- 原文: The spirit is willing, but the flesh is weak
  译文：酒是好的，但肉是馊的
  （心有余而力不足）
- 原文：How are you？
  译文：怎么是你？

  原文：How old are you?
  译文：怎么老是你？

- 原文: The pen was in the box.
  译文：钢笔在盒子里。

- 原文：The box was in the pen.
  译文： 盒子在钢笔里。
  （盒子在围栏里。）

机器翻译的例子

- 原文：梦之缘（民宿）

  译文1：dream of margin

  译文2：dream edge

  译文3：the edge of the dream

- 原文： 斗地主我总是斗不过他。

  译文1： I can't always beat him.

  译文2： I always fight him with the landlord.

  译文3： I always fight the landlord.

http://fanyi.youdao.com/

https://translate.google.cn

https://cn.bing.com/translator/

文本检索的例子

- 在 Internet 或数字图书馆上
  - 输入词、短语或句子
  - 检索相应的文档
  - 例子：

    和服 Search

- Question
  - ➢如何得到想得到的结果？

未经中文分词处理时的检索结果

1.
电信运营商和服务提供商
采用奥维通的移动WiMAX解决方案,运营商和服务提供 商可以提供各种个人宽带服务 ……

2.
关于做好党员联系和服务群众工作的意见
做好党员联系和服务群众工作，要以马克思列宁主义、毛泽东思想、邓小平理论和“三个代表”重要……

3.
Guangzhou bomei leather co.,ltd
站长信息和服务中心:斗破苍穹 阴阳冕 九鼎记 凡人修仙传 猎国 九转金身决……

4.
关于商品和服务实行明码标价的规定
根据《中华人民共和国价格法》修订的《关于商品和服务实行明码标价的规定》，……

5.
Technical Support
利盟中国面向行业,办公和家庭提供彩色激光,黑白激光,喷墨,和多功能一体打印机及相关耗材和服务,是业届领先的打印解决方案的开发制造商。

……

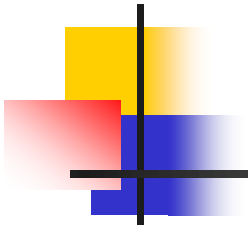
4 计算语言学的发展简史

- 1950 — 1960
- 1960 — 1970
- 1970 — 1990
- 1990 — 2010
- 2013
- 2018
- 2022

- Warren Weaver(1949)
- Turing Test(1950)
- The first MTs(1954)
- ALPAC(1964-1966)
- Searle's Chinese Room(1980)
- The first PC version of MTs(early 1980s)
- MT is available on the Web(1994)
- ……
- Word2Vec
- BERT / GPT
- ChatGPT / GPT-4
- ……

— 1950s：萌芽期

- 1933: 法国的 Georges Artsrouni & 俄国的 Peter Trojanskij建议:构建 机器多语言词典；
- 1946-1947:美国的Andrew Booth 和 Warren Weaver,提出了机器翻译的设想.
- 1950s: Yehoshua Bar-Hillel(MIT): 1952年举办了 1st MT会议，会上，Leon Dostert (Georgetown Univ.)建议开发演示系统，以吸引基金的投资.
- 1955年，第一个演示系统在 IBM & Georgetown Univ.开发成功，包含 250 个词和 6 条句法规则，实现了 Russia — English的翻译；

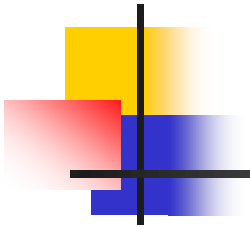


1950s：理论基础建设

- 形式语言 (Chomsky, Kleene, Backus).
  - 语言描述的形式化：对语言按复杂程度分类，对不同类语言进行形式化描述；
  - 语言处理的形式化：对不同类语言进行自动识别和分析
- 概率与信息论方法
  - 语言的理解作为解码：Shannon的噪音管道模型；
  - 使用概率方法：将语言的产生看成随机过程；

1966：低谷期/ALPAC 报告

- Automatic Language Processing Advisory Committee (ALPAC) (1964, USA)
- ALPAC 报告的内容 (1966) :
  - “There is no immediate or predictable prospects of useful machine translation”—— Ends funding MT.
  - Only support fundamental research in CL
- ALPAC 报告所指出的MT前景黯淡的主要原因:
  - 遇到了语义障碍 （Semantic Barrier）
  - 需要真实世界的知识（Bar-Hillel的评论）



1970s：恢复期

- 1971: W. Woods: Lunar, IR (ATN)
- 1972: T. Winograd: SHRDLU (Lisp)
- 1973: Schank: Concept Dependency (CD Theory)
  - MARGIE 1975: (Meaning, Analysis, Response Generation and Inference on English) on CD: NLU

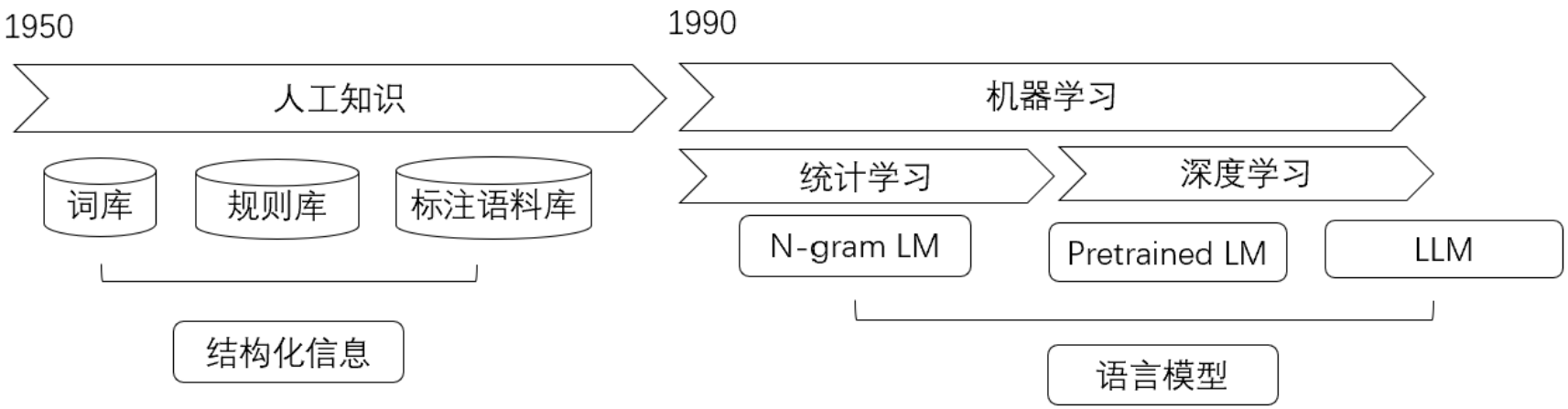
1980s：发展期

- 强调知识的重要性
  - 人工智能的发展
  - 日本的第5代计算机（知识处理系统）
  - 语言知识与语言分析的分离
- 语义知识的表示与知识库构建：CYC，Wordnet等；
- 机器翻译在受限领域获得成功：加拿大Meteo；
- 句法语义和词汇语义理论的蓬勃发展：
  - GPSG (Generalized PSG: Gazdar),
  - GB (Governing and Binding: N. Chomsky) ,
  - FUG (Functional Unification Grammar: M. Kay)
  - …

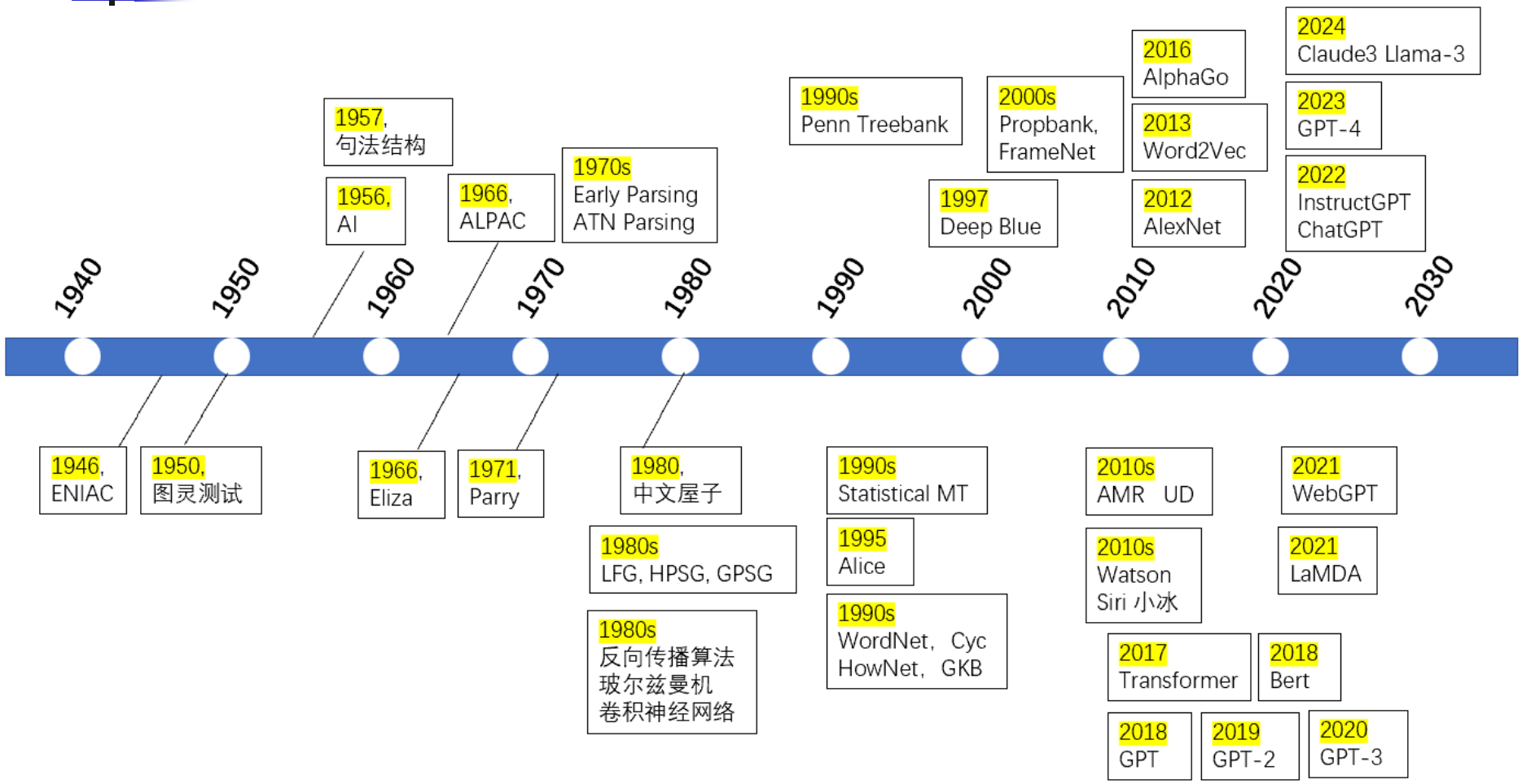
1990s：经验主义方法再度受重视

- IBM's P.Brown (1988-2nd TMI—Theoretical & Methodological Issues in MT, 1990-CL—Computational Linguistics,1993-CL): 倡导机器翻译的统计方法；
- 基于统计与语料库的方法逐步取得支配地位；
- 强调大规模真实语料；
- 强调机器学习与知识自动获取的重要性；
- 语音识别逐步实用化；
- 开展了大规模的评测；

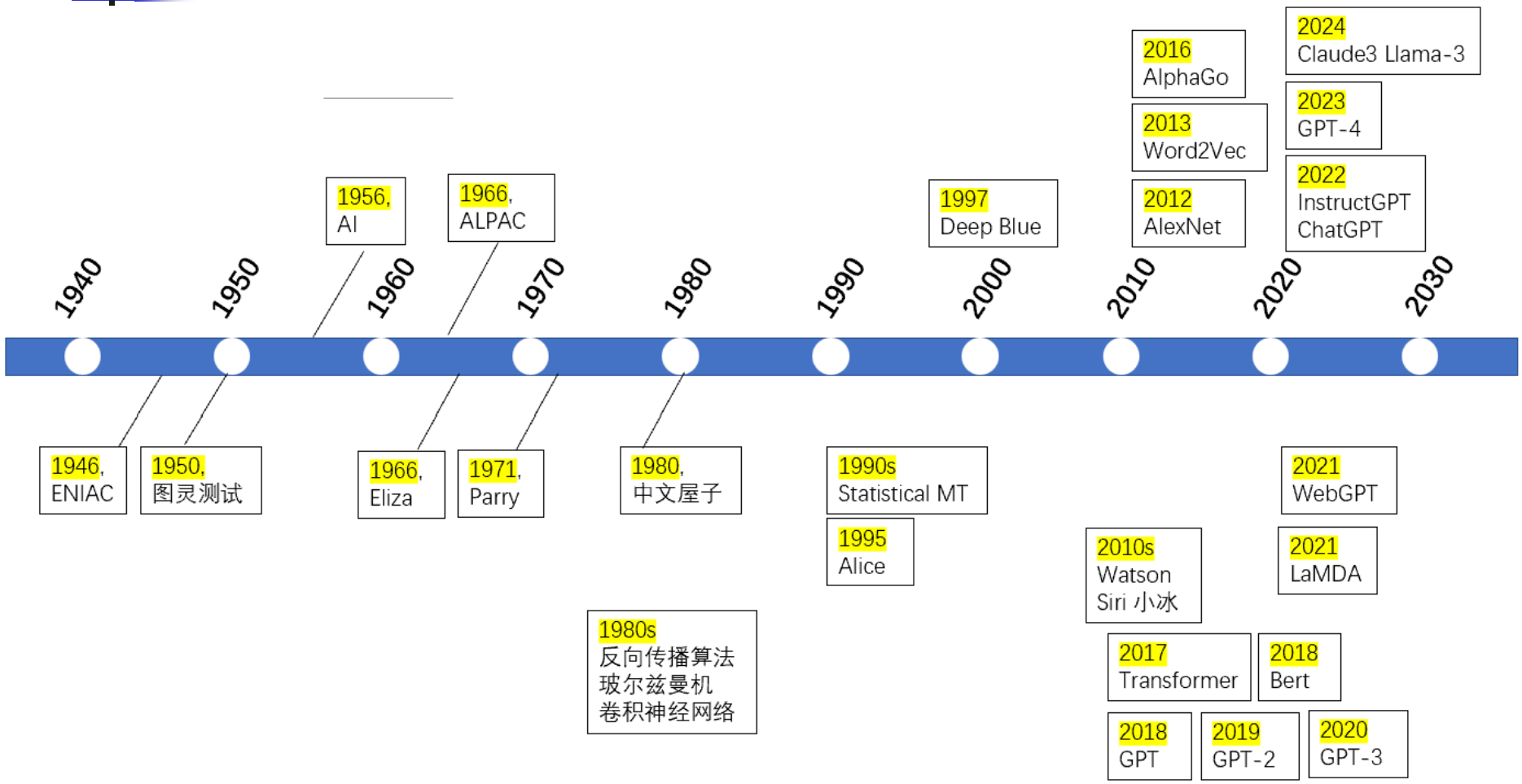
人工智能/自然语言处理/人机对话发展历程



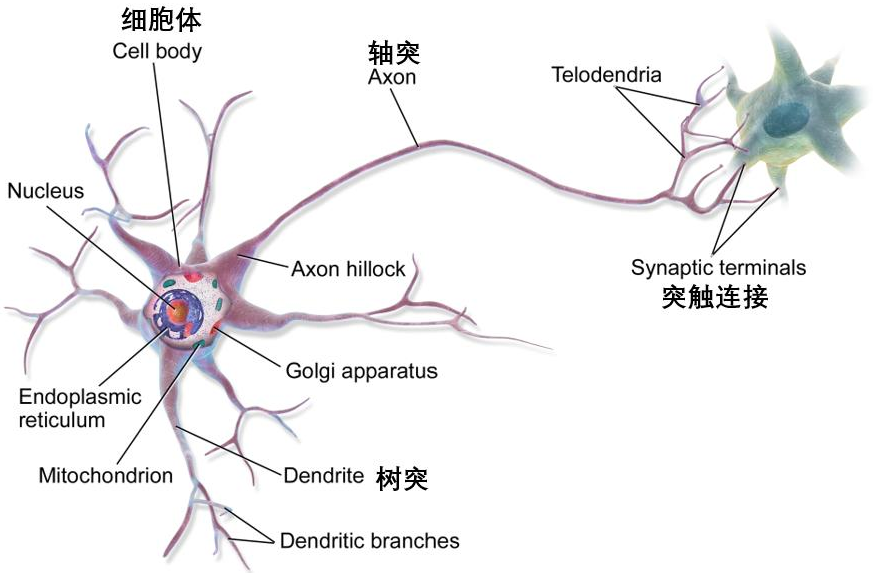
人工智能/自然语言处理/人机对话发展历程



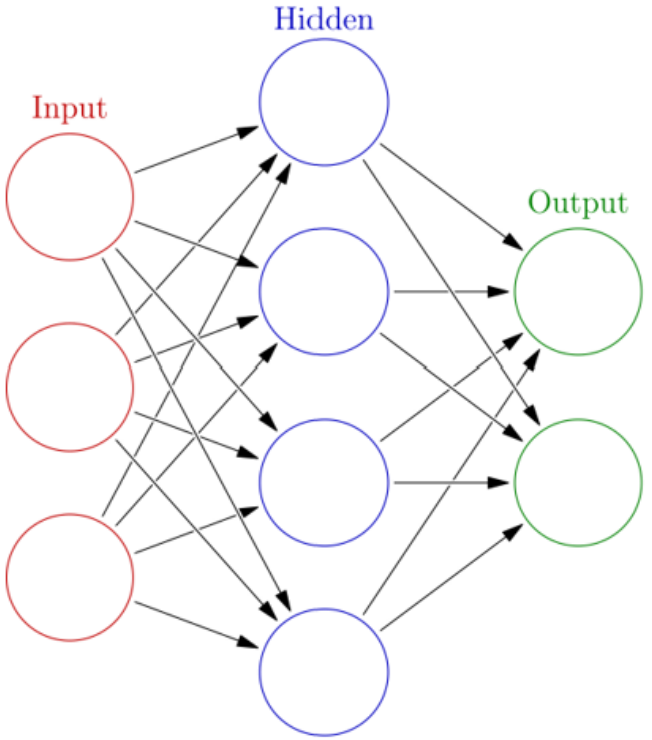
人工智能/自然语言处理/人机对话发展历程



21世纪：人工神经网络

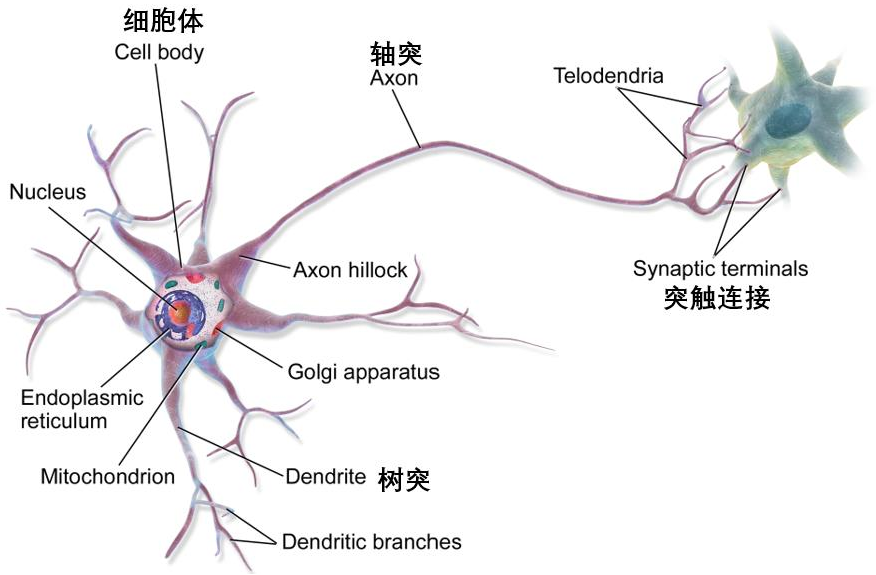


https://en.wikipedia.org/wiki/Neuron

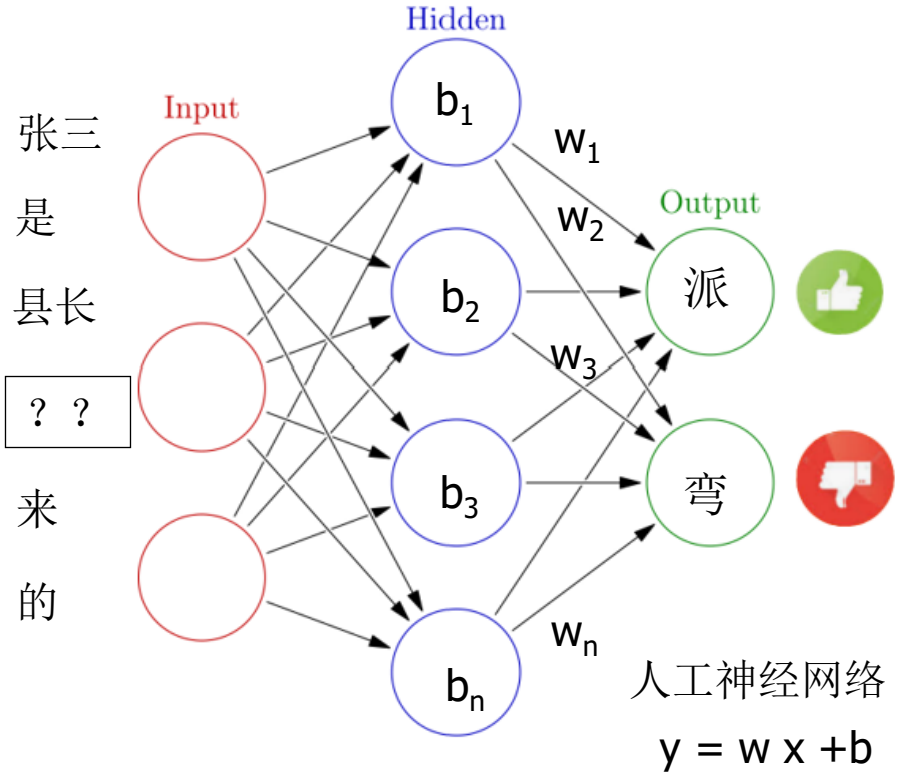


https://en.wikipedia.org/wiki/Artificial_neural_network

21世纪：人工神经网络



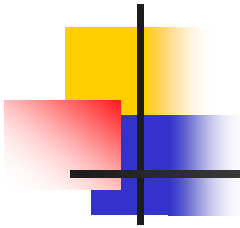
https://en.wikipedia.org/wiki/Neuron



https://en.wikipedia.org/wiki/Artificial_neural_network

21世纪：强调数据资源

- 各类电子词典
  - 词义词典：WordNet, HowNet, CCD
  - 语法词典：现代汉语语法信息词典
  - 语义角色：FrameNet, VerbNet
- 各种语料库
  - 英语：词性、实体、角色等，LDC
  - 汉语：分词+词性标注，汉语词义，汉语拼音 等
  - 多语言（用于机器翻译）：词对应，短语对应，句子对应，等；
- 海量数据（无标注语料/自然标注语料）



21世纪：强调评测

- SIGHAN（汉语 Special Interest Group of HAN）
- NIST (National Institute of Standard and Technology)
  - TREC （Text Retrieval Conference）
  - Open MT （NIST Open Machine Translation Evaluation）
  - MUC （Message Understanding Conference）
  - TDT （Topic Detection & Tracking）
  - DUC/TAC （Document Understanding Conference）
（Text Analysis Conference）
- ……

评测举例：WSC

Winograd Schema Challenge (2011)

The city councilmen refused the demonstrators a permit because they [feared / advocated] violence. Who [feared/advocated] violence?

Answers: The city councilmen / The demonstrators.

市政府拒绝给示威者颁发游行许可证，因为[担心/鼓吹]暴力事件。 谁[担心/鼓吹]暴力事件?

答案: 市政府/示威者

The trophy doesn't fit into the brown suitcase because it's too [small / large]. What is too [small/large]?

Answers:The suitcase / The trophy.

奖杯无法放进到棕色的箱子里，因为它太[小/大]了。 什么东西太[小/大]了?

答案: 箱子/奖杯

评测举例： WINOGRANDE (2019)

		Twin sentences	Options (**answer**)
(1)	a	The trophy doesn't fit into the brown suitcase because **it**'s too *large*.	**trophy** / suitcase
	b	The trophy doesn't fit into the brown suitcase because **it**'s too *small*.	trophy / **suitcase**
(2)	a	Ann asked Mary what time the library closes, *because* **she** had forgotten.	**Ann** / Mary
	b	Ann asked Mary what time the library closes, *but* **she** had forgotten.	Ann / **Mary**
? (3)	a	The tree fell down and crashed through the roof of my house. Now, I have to get **it** *removed*.	**tree** / roof
	b	The tree fell down and crashed through the roof of my house. Now, I have to get **it** *repaired*.	tree / **roof**
? (4)	a	The lions ate the zebras because **they** are *predators*.	**lions** / zebras
	b	The lions ate the zebras because **they** are *meaty*.	lions / **zebras**

remove —— tree
repair —— roof

predator —— lion
meaty —— zebras

结语：“名实之辩”

- 计算语言学（Computational Linguistics）
- 计量语言学（Quantitative Linguistics）
- 数理语言学（Mathematical Linguistics）

- 自然语言理解（Natural Language Understanding）
- 自然语言处理（Natural Language Processing）
- 人类语言技术（Human Language Technology）

结语：计算语言学与理论语言学

（1）一个句子的结构和意义是什么
（如何呈现/表示一个句子的结构和意义）？

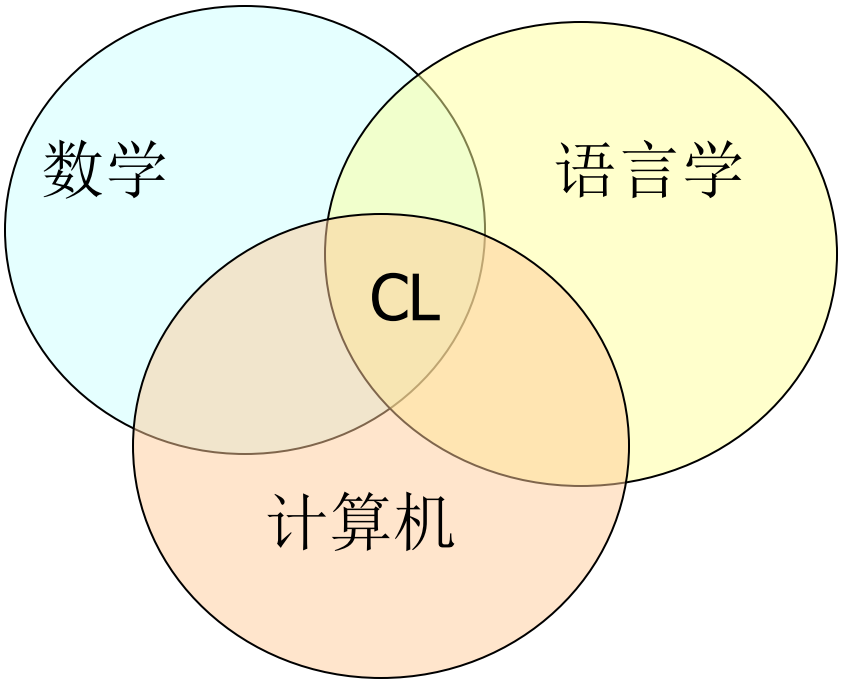
（2）如何通过“计算”，得到一个句子的结构和意义?

著名的计算语言学家John Nerbonne这样写到，“在语言学和计算语言学之间存在着理论任务的自然分野，大致说来，语言学的责任是描述语言，而计算语言学提供算法和用于计算的体系结构。基于这种观点，这两个理论领域因为其共同关注的对象——语言——而发生紧密关系。”

John Nerbonne, 1996, Computational Semantics – Linguistics and Processing, Shalom Lappin, ed. The Handbook of Contemporary Semantic Theory, Chapter 17, Blackwell Publishers. (外语教育与研究出版社2001年引进出版, pp461-462）

结语：

- 计算语言学是一个多学科交叉的领域
- 计算语言学是一个年轻的学科



目标：

语言学知识

→ 结构化
→ 形式化
→ 数据化
→ 可视化

进一步阅读文献

1. Hubert L. Dreyfus(1979), What Computers Can't Do, Harper & Row, Publishers, 1979 （计算机不能做什么）（宁春岩译本）
2. Hubert L. Dreyfus(1992), What Computers Still Can't Do: A Critique of Artificial Reason, MIT Press, Revised Edition, 1992.
3. Roger Penrose (1989), The Emperor's New Mind: Concerning Computers, Minds, and The Laws of Physics, Oxford University Press, 1989（皇帝新脑）
4. 袁毓林（1993）自然语言理解的语言学假设，《中国社会科学》1993.1
5. 宁春岩（1985）自然语言理解中的几个根本问题，《语言研究》1985.2
6. 詹卫东（2000）80年代以来汉语信息处理研究述评，《当代语言学》2000.2
7. 詹卫东（2010）计算语言学与中文信息处理研究近年来的发展综述(2004-2008)，中国语言学年鉴（2004-2008）未刊。

复习思考题

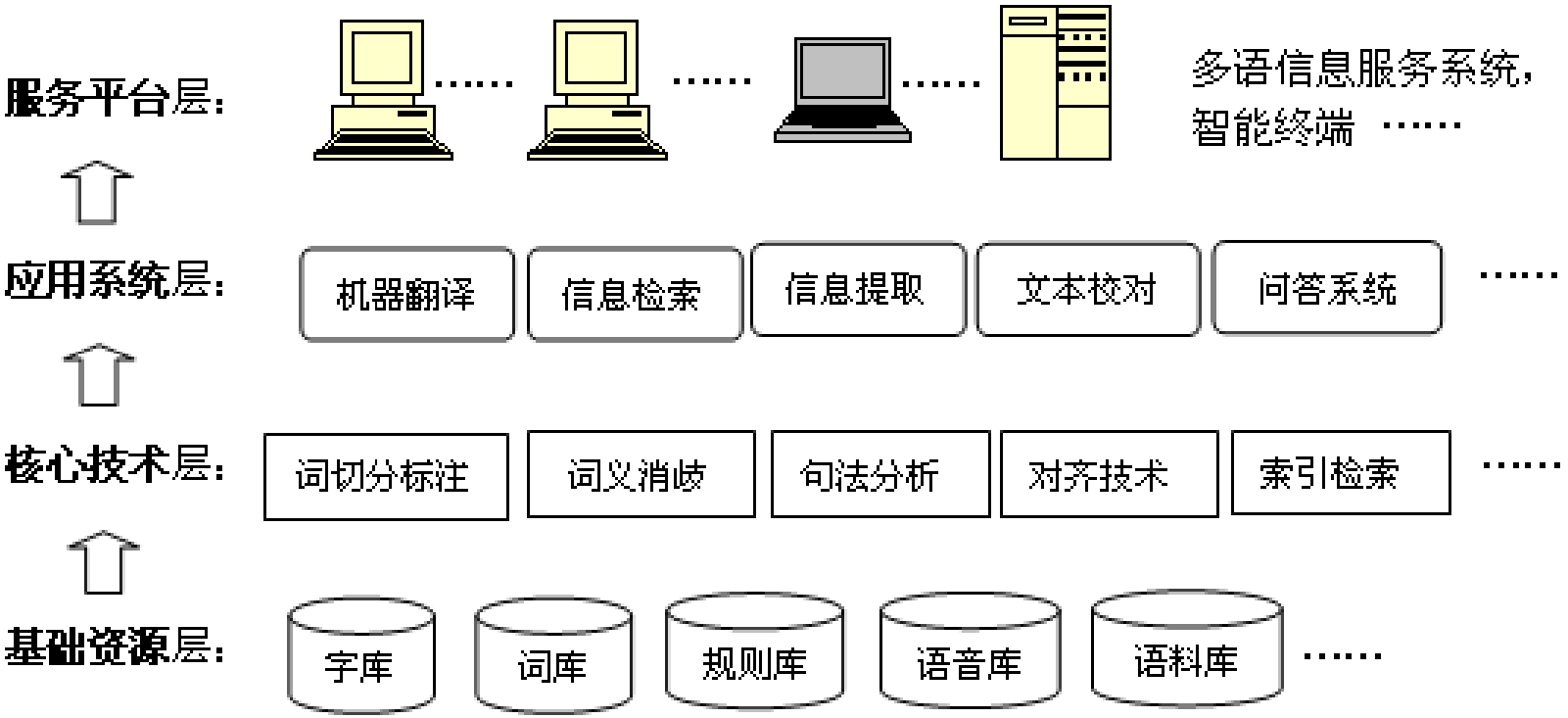
- 计算语言学是一个什么性质的学科？
- 人类的语言有哪些性质？
- 请举出一个算法的例子。
- 你对把“计算机”称作“电脑”有何评论？
- 人脑，电脑（机器）之间有不可逾越的鸿沟吗？

附录：自然语言处理的宏观架构（1）

对象 / 任务 / 对象	**书面文本** [视觉符号]		**口语语音** [听觉符号]	
处理符号的意义	文本理解 文本生成	[机器翻译 信息检索… 文本摘要 问答系统…]	语音识别 语音合成	[口语翻译… 口语问答…]
处理符号的形式	汉字输入、存储、输出 篇章版式分解与生成		语音信号采集、 波形特征抽取、波形生成	

根据符号性质的差异对中文信息处理的对象进行分类

附录：自然语言处理的宏观架构（2）



根据语言单位性质的差异对中文信息处理的对象和技术层级进行分类

附录：自然语言处理的宏观架构（3）

？？？