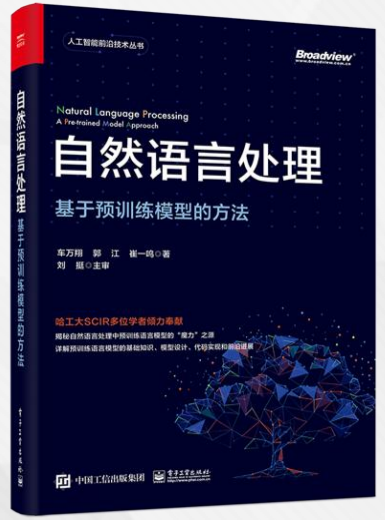


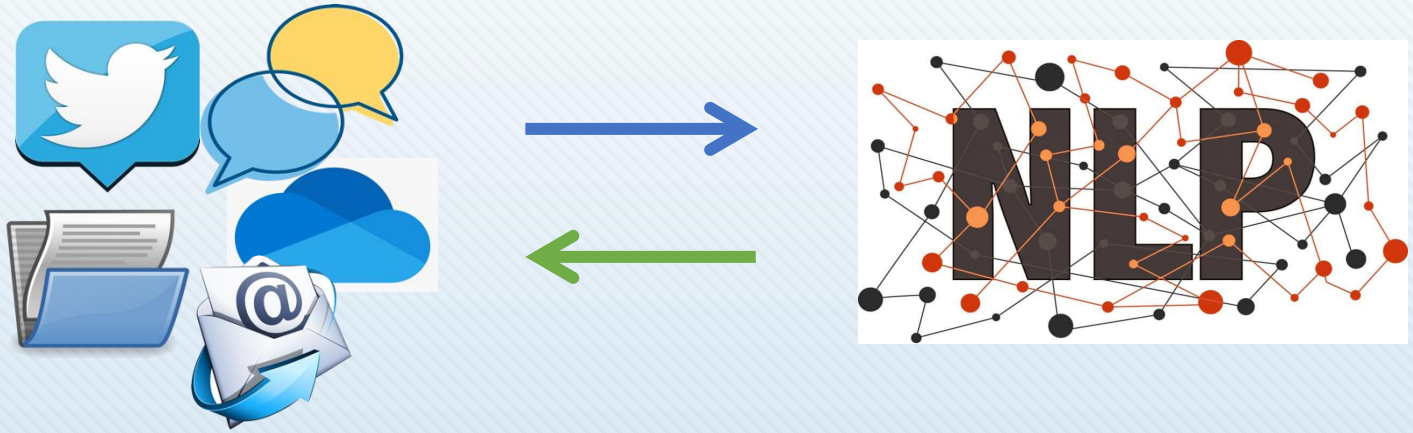
绪论

车万翔、郭江、崔一鸣

社会计算与信息检索研究中心
哈尔滨工业大学

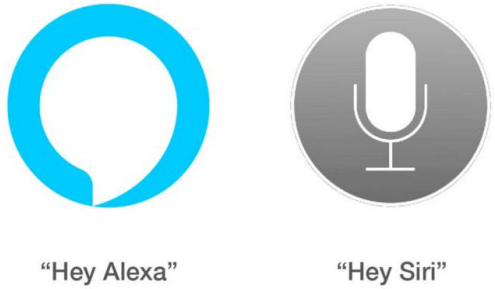
什么是自然语言处理?

- 语言是**思维的载体**，是人类交流思想、表达情感最自然、最方便的工具
  - 人类历史上大部分知识是以语言文字形式记载和流传的
- 自然语言指的是人类语言，特指**文本符号**，而非语音信号
- 自然语言处理（Natural Language Processing，NLP）
  - 用计算机来**理解**和**生成**自然语言的各种理论和方法



自然语言处理的代表性应用

机器翻译

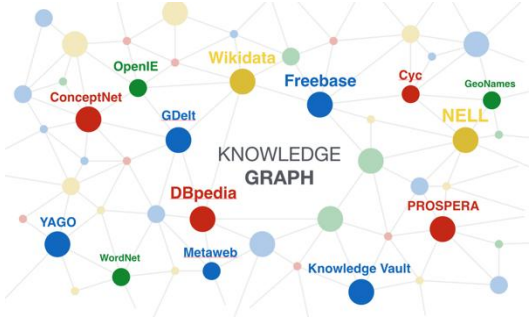


智能助手

文本校对

舆情分析

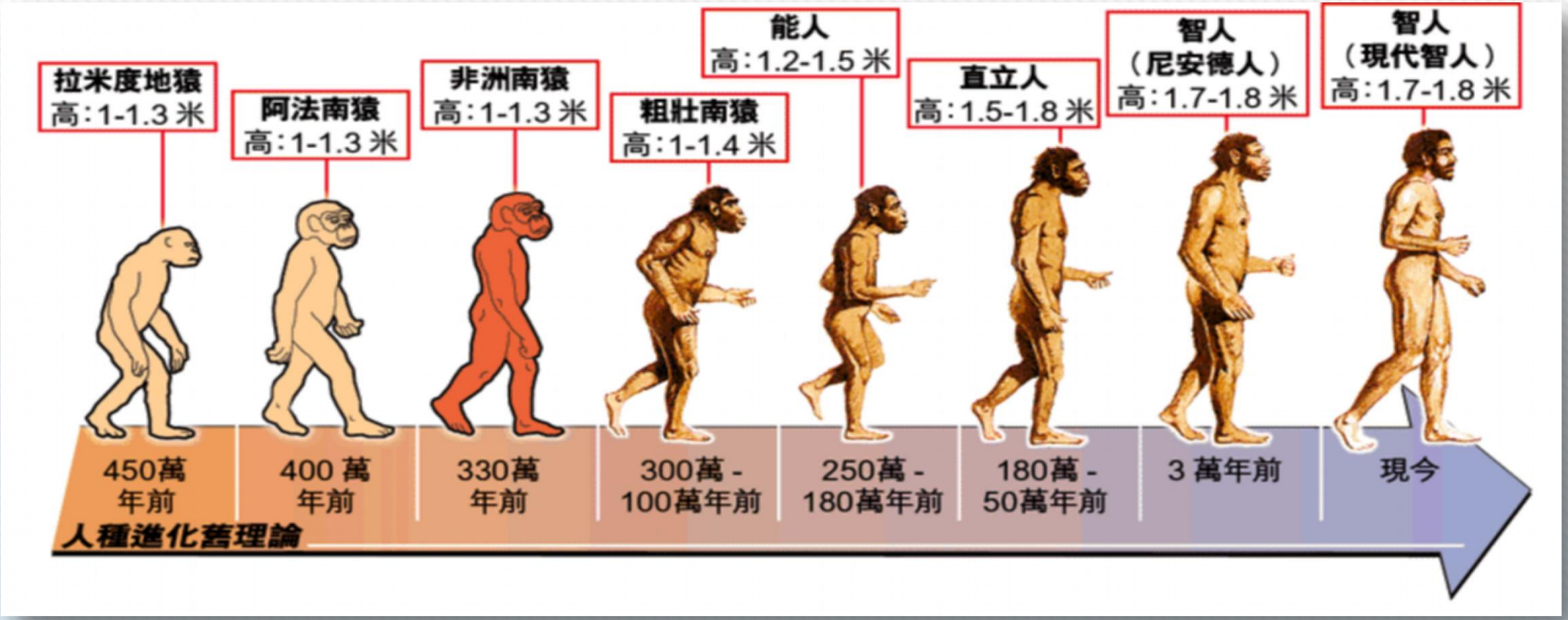
智能教育



知识图谱

自然语言处理属于认知智能任务

- **认知智能**是人类和动物的主要区别之一
  - 需要更强的抽象和推理能力



自然语言处理的难点与特点

领导：“你这是什么意思？”

阿呆：“没什么意思，意思意思。”

领导：“你这就不够意思了。”

阿呆：“小意思，小意思。”

领导：“你这人真有意思。”

阿呆：“其实也没有别的意思。”

领导：“那我就不好意思了。”

阿呆：“是我不好意思。”

“人工智能皇冠上的明珠”

- 自然语言处理成为**制约人工智能取得更大突破和更广泛应用的瓶颈**

“深度学习的下一个前沿课题是自然语言理解”

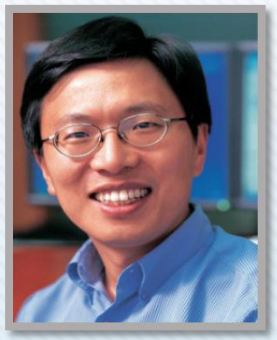
Yann LeCun
图灵奖得主、Facebook AI 负责人

“深度学习的下一个大的进展应该是让神经网络真正理解文档的内容”

Geoffrey Hinton
图灵奖得主、深度学习之父

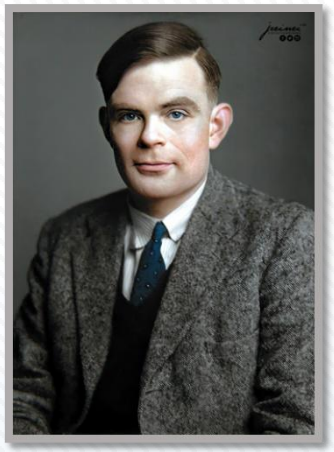
“如果给我10亿美金，我会用这10亿美金建造一个NASA级别的自然语言处理研究项目”

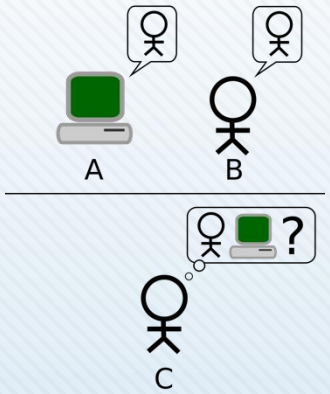
Michael I. Jordan
美国双院院士、世界知名机器学习专家



“下一个十年，懂语言者得天下”

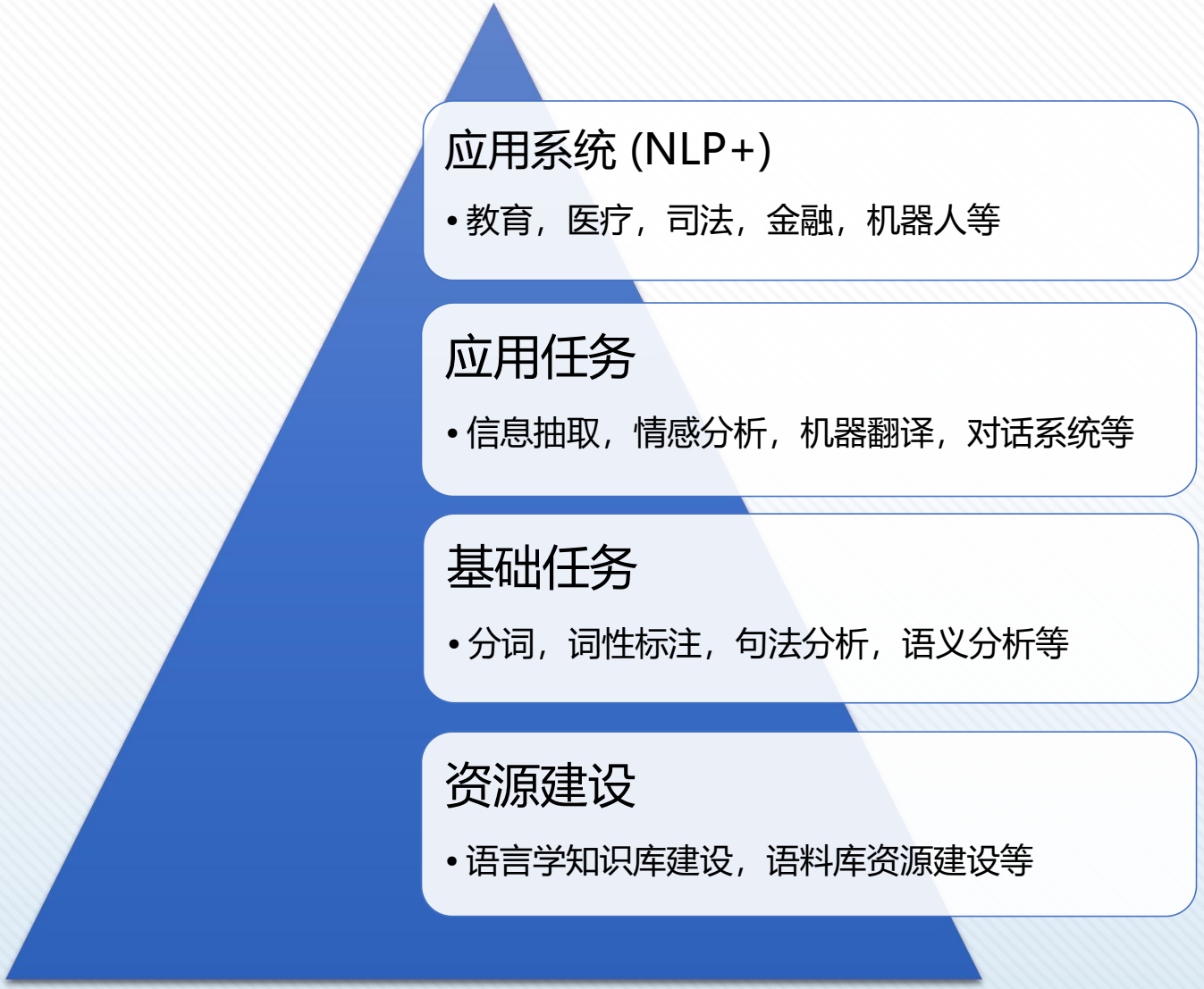
沈向洋
美国工程院院士、微软前全球执行副总裁



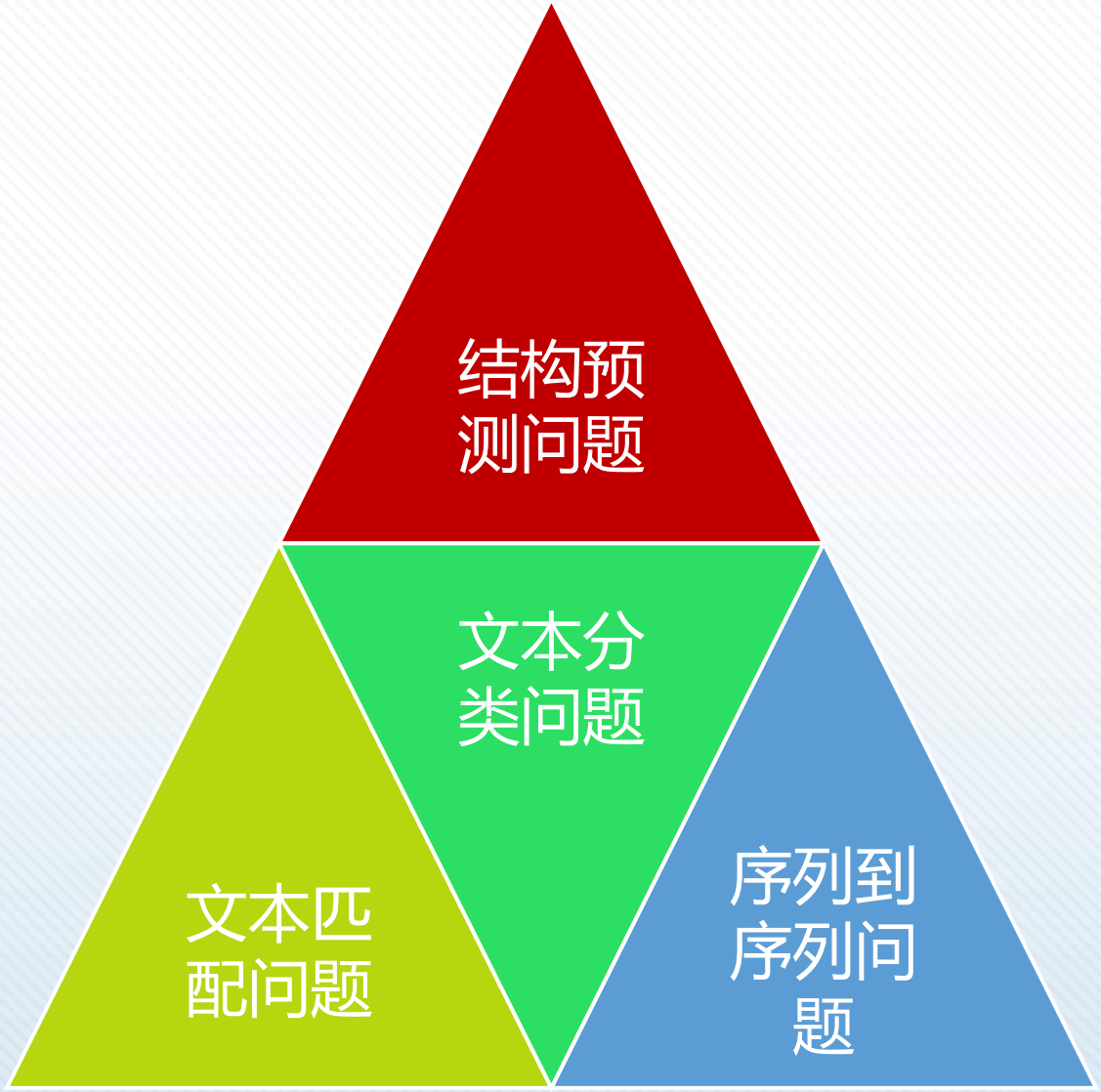


图灵测试

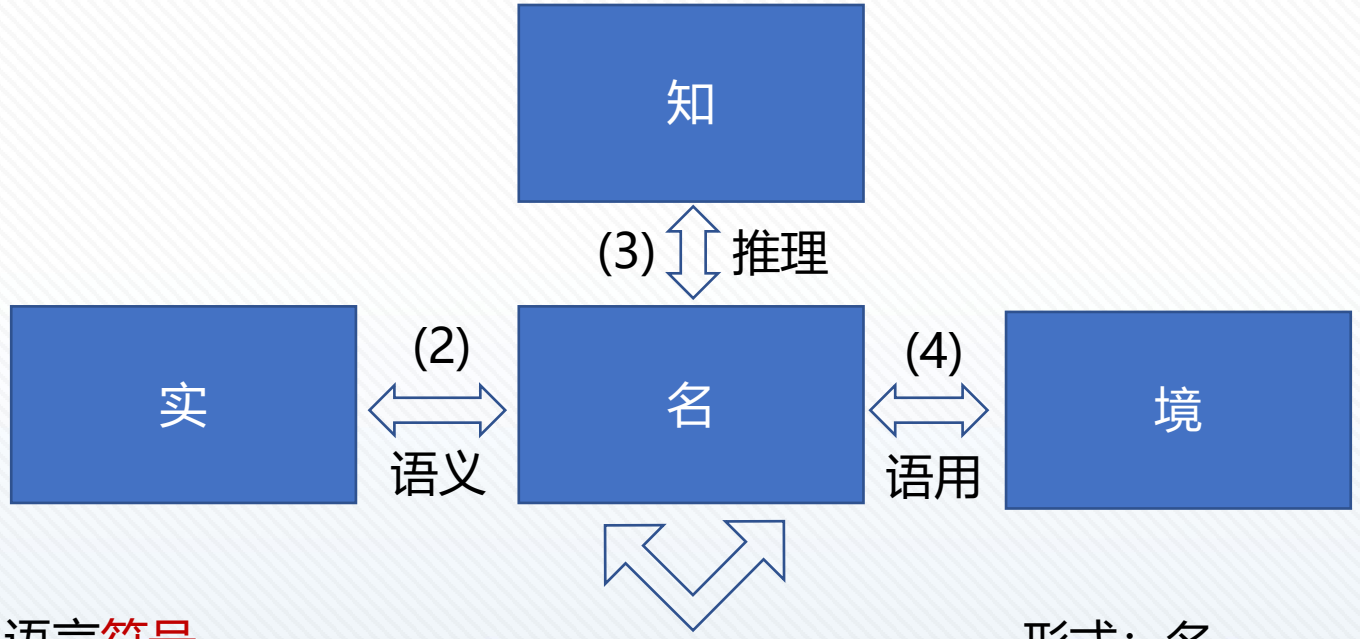
自然语言处理任务层级


应用系统 (NLP+)
• 教育，医疗，司法，金融，机器人等
应用任务
• 信息抽取，情感分析，机器翻译，对话系统等
基础任务
• 分词，词性标注，句法分析，语义分析等
资源建设
• 语言学知识库建设，语料库资源建设等

归结为四个基本问题



自然语言处理研究对象与层次



名：语言符号
实：客观事实、主观意见
知：知识
境：语言所处环境

形式：名
语义：名+实
推理：名+实+知
语用：名+实+知+境

“层次 x 任务”二维表

	分类	解析	匹配	生成
形式	文本分类	词性标注 句法分析	搜索	机械式文摘
语义	情感分析	命名实体识别 语义角色标注	问答	机器翻译
推理	隐式情感分析		文本蕴含	写故事结尾
语境	反语			聊天

语言模型

- 语言模型（Language Model，LM）
  - 描述一段自然语言的概率或给定上文时下一个词出现的概率
    - $P(w_1, \cdots w_l)$，$P(w_{l+1}|w_1, \cdots w_l)$
  - 以上两种定义等价（链式法则）

$$P(w_1 w_2 \cdots w_l) = P(w_1)P(w_2|w_1)P(w_3|w_1 w_2)\cdots P(w_l|w_1 w_2 \cdots w_{l-1})$$

$$= \prod_{i=1}^{l} P(w_i|w_{1:i-1})$$

- 广泛应用于多种自然语言处理任务
  - 机器翻译（词排序）
    - P(the cat is small) > P(small the is cat)
  - 语音识别（词选择）
    - P(there are **four** cats) > P(there are **for** cats)

中文分词

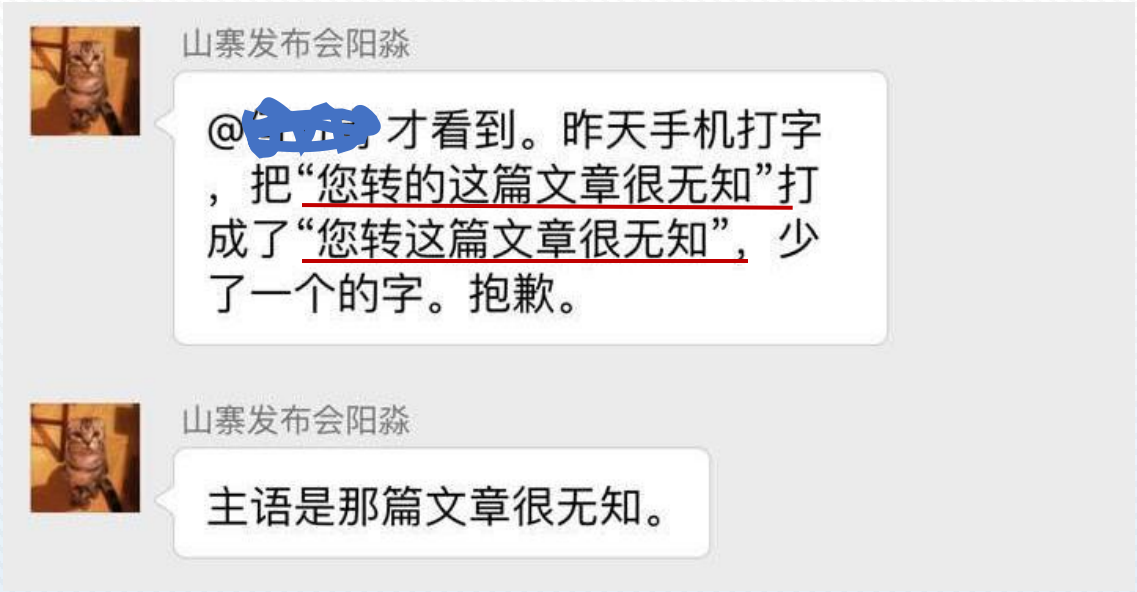
- 词（Word）
  - 最小的能独立使用的音义结合体
  - 以汉语为代表的汉藏语系，以阿拉伯语为代表的闪-含语系中不包含明显的词之间的分隔符
- 中文分词是将中文字序列切分成一个个单独的词
- 分词的歧义
  - 如：严守一把手机关了
    - 严守一/ 把/ 手机/ 关/ 了
    - 严守/ 一把手/ 机关/ 了
    - 严守/ 一把/ 手机/ 关/ 了
    - 严守一/ 把手/ 机关/ 了
    - ……

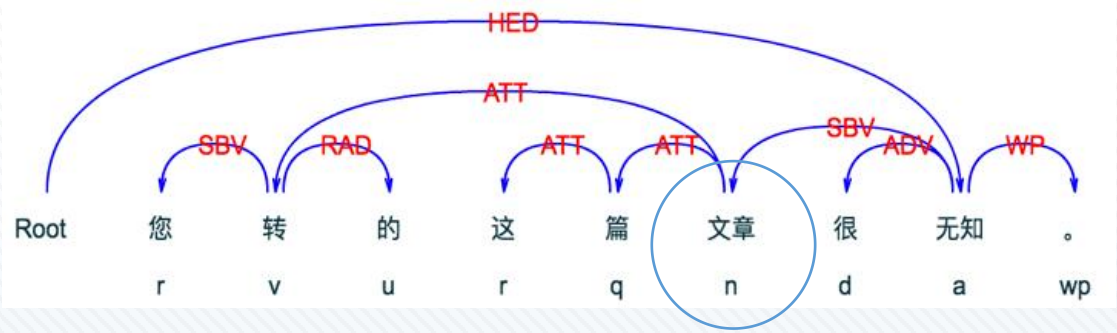
子词切分（Subword）

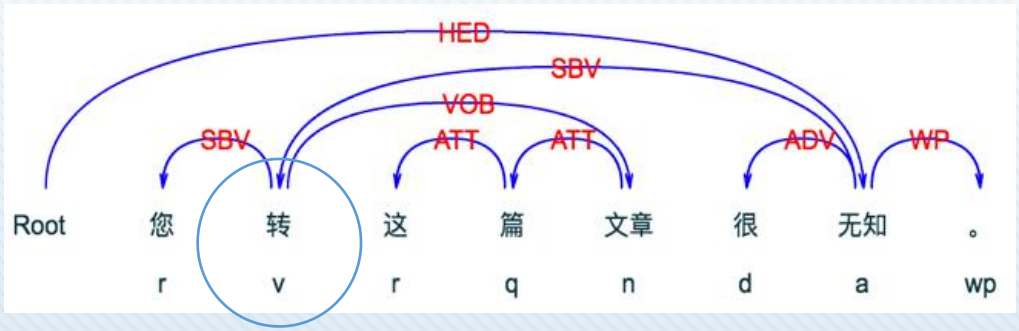
- 以英语为代表的印欧语系语言，是否需要进行分词？
- 这些语言词形变化复杂
  - 如：*computer*、*computers*、*computing*等
- 仅用空格切分的问题
  - 数据稀疏
  - 词表过大，降低处理速度
- 子词切分
  - 将一个单词切分为若干连续的片段（子词）
  - 方法众多，基本原理相似
    - 使用尽量长且频次高的子词对单词进行切分

句法分析

- 分析句子的句法成分，如主谓宾定状补等
- 将词序列表示的句子转换成树状结构







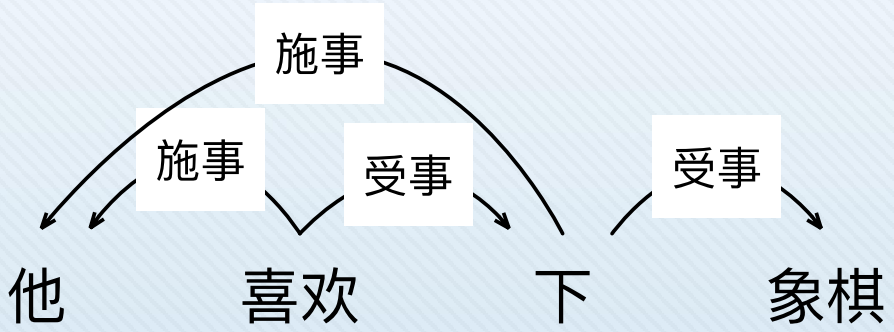
http://ltp.ai/demo.html

语义分析

- 词义消歧（Word Sense Disambiguation，WSD）
- 语义角色标注（Semantic Role Labeling，SRL）
  - 也称谓词论元结构（Predicate-Argument Structure）

输入	他	喜欢	下	象棋	。
输出 1	施事	谓词	受事		
输出 2	施事		谓词	受事	

- 语义依存图（Semantic Dependency Graph）



信息抽取

- 信息抽取（Information Extraction，IE）
  - 从非结构化的文本中自动提取结构化信息
- 输入
  - *10月28日，AMD宣布斥资350亿美元收购FPGA芯片巨头赛灵思。这两家传了多年绯闻的芯片公司终于走到了一起。*

信息抽取子任务	抽取结果
命名实体识别	公司名：AMD 公司名：赛灵思
关系抽取	赛灵思 $\xrightarrow{\text{从属}}$ AMD
时间表达式抽取	*10 月 28 日*
时间表达式归一化	*10 月 28 日* → *2020 年 10 月 28 日*
事件抽取	事件：收购 时间：*2020 年 10 月 28 日* 收购者：AMD 被收购者：赛灵思 收购金额：*350 亿美元*

情感分析

- 情感分析（Sentiment Analysis）
  - 个体对外界事物的态度、观点或倾向性，如正面、负面等
  - 人自身的情绪（Emotion），如喜怒哀惧等
- 输入
  - *这款手机的屏幕很不错，性能也还可以。*

情感分析子任务	分析结果
情感分类	褒义
情感信息抽取	评价词：不错；可以 评价对象：屏幕；性能 评价搭配：屏幕 ⇔ 不错；性能 ⇔ 可以

问答系统

- 问答系统（Question Answering，QA）
  - 用户以自然语言形式描述问题，从异构数据中获得答案
- 根据数据来源的不同，问答系统可以分为4种主要的类型
  - 检索式问答系统
    - 答案来源于固定的文本语料库或互联网，系统通过查找相关文档并抽取答案完成问答
  - 知识库问答系统
    - 回答问题所需的知识以数据库等结构化形式存储，问答系统首先将问题解析为结构化的查询语句，通过查询相关知识点，并结合知识推理获取答案
  - 常问问题集问答系统
    - 通过对历史积累的常问问题集合进行检索，回答用户提出的类似问题
  - 阅读理解式问答系统
    - 通过抽取给定文档中的文本片段或生成一段答案来回答用户提出的问题

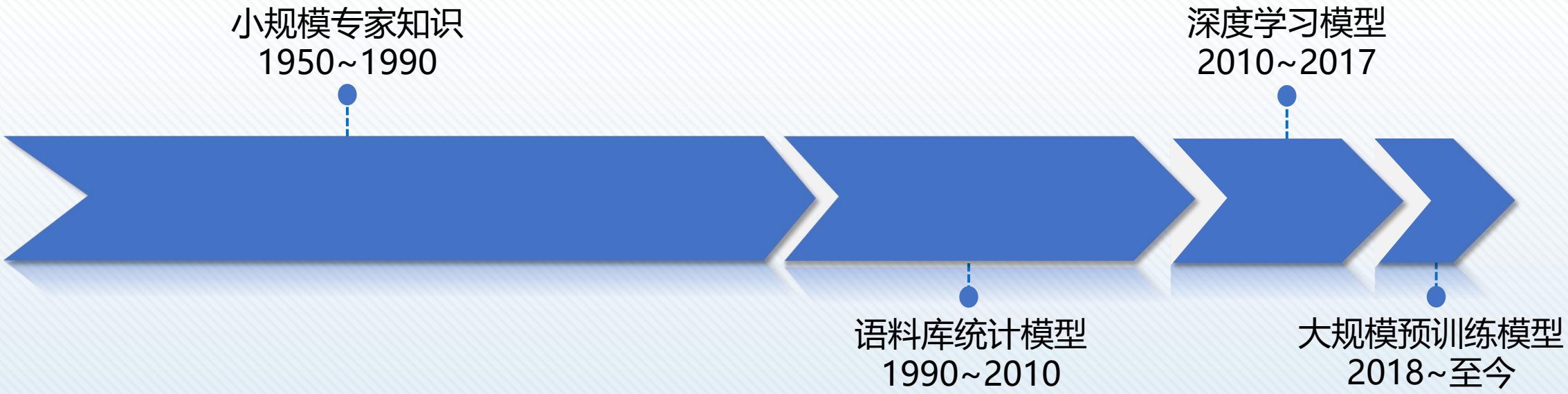
更多任务

- 机器翻译（Machine Translation，MT）
- 对话系统（Dialogue System）

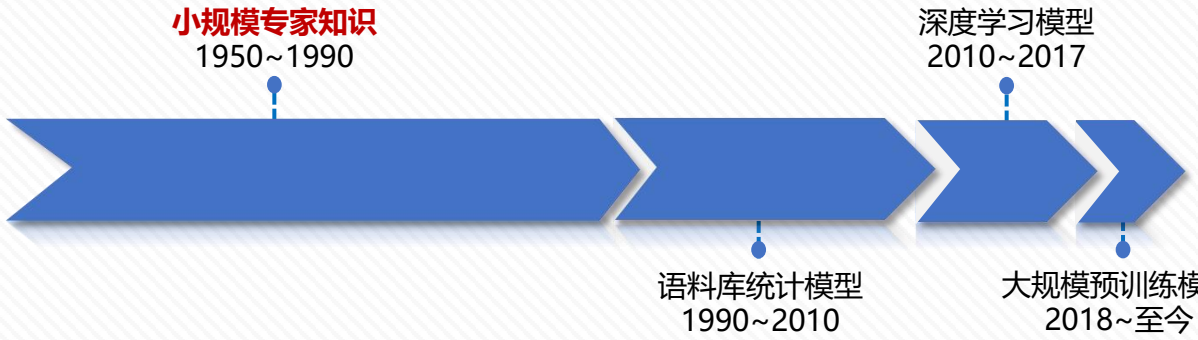
	任务型 Task	聊天 Chat	知识问答 Knowledge	推荐 Recommen-dation
目的	完成任务或动作	闲聊	知识获取	信息推荐
领域	特定域（垂类）	开放域	开放域	特定域
以话轮数评价	越少越好	话轮越多越好	越少越好	越少越好
应用	虚拟个人助理	娱乐、情感陪护	客服、教育	个性化推荐
典型系统	Siri、Cortana、Google Assistant、度秘	小冰、笨笨	Watson、Wolfram Alpha	阿里小蜜

自然语言处理技术发展历史

- 语义在计算机内部是如何**表示**的?
- 根据表示方法的不同，自然语言处理技术经历了**四次范式变迁**



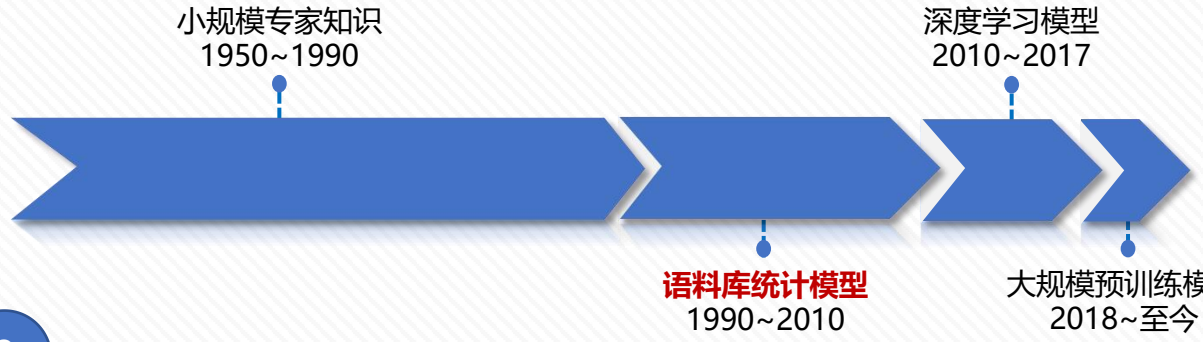
基于符号（字符串）表示的专家知识

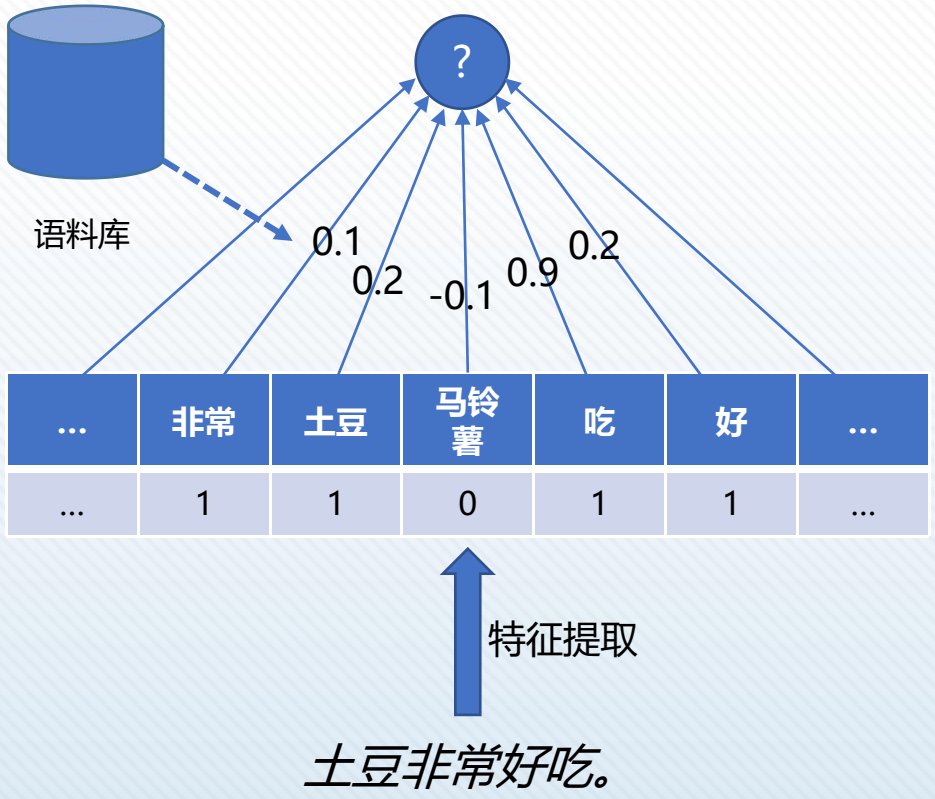


- “*土豆非常好吃。*”的情感倾向性？
  - 如果：出现褒义词“好”“喜欢”等
  - 那么：结果为褒义
  - 如果：出现“不”
  - 那么：结果倾向性取反

- 优点
  - 符合人类的直觉
  - 可解释、可干预性好
- 缺点
  - 知识完备性不足
  - 需要专家构建和维护
  - 不便于计算

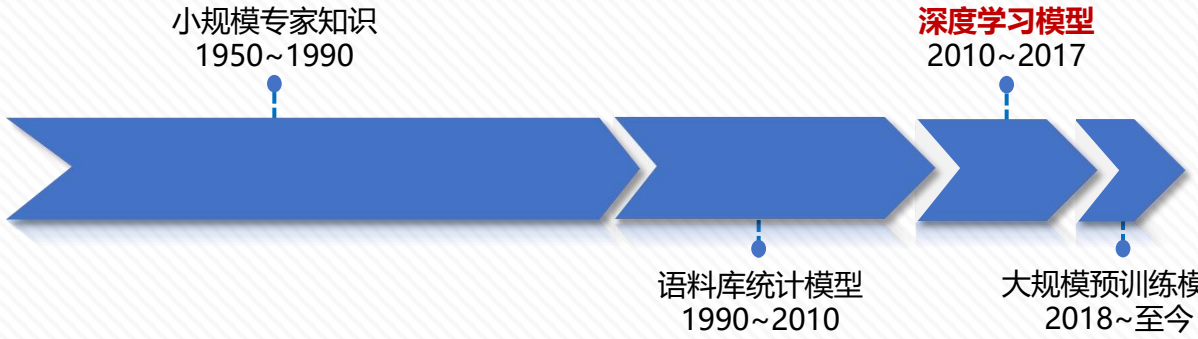
基于向量表示的统计模型





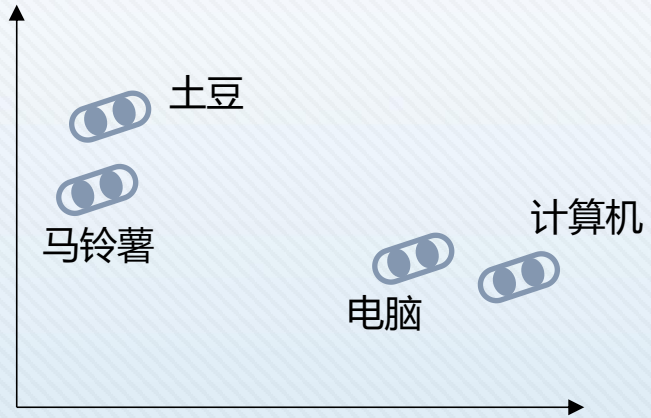
- 使用**高维、离散、稀疏**的向量表示词
  - 维度为词表大小，其中只有一位为1，其余为0
    - **土豆**：[0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, ...]
    - **马铃薯**：[0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, ...]
- 缺点
  - 无法处理“**多词一义**”的现象

基于嵌入表示的深度学习模型



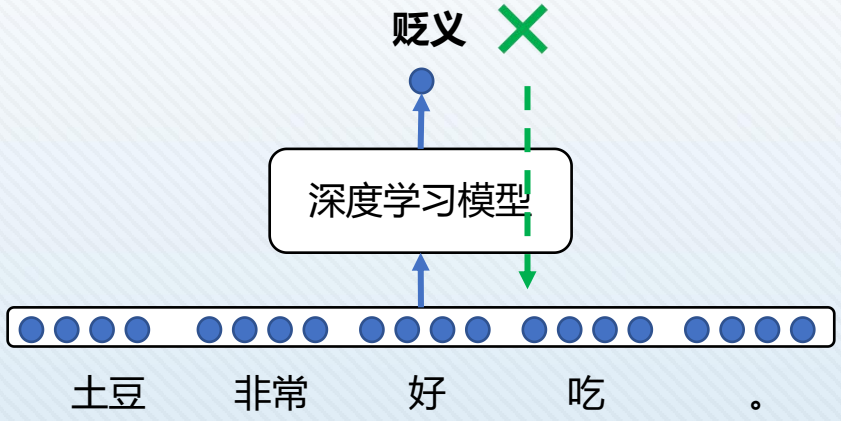
词嵌入（Word Embedding）

- 直接使用一个**低维、连续、稠密**的向量表示词（Bengio等2003）

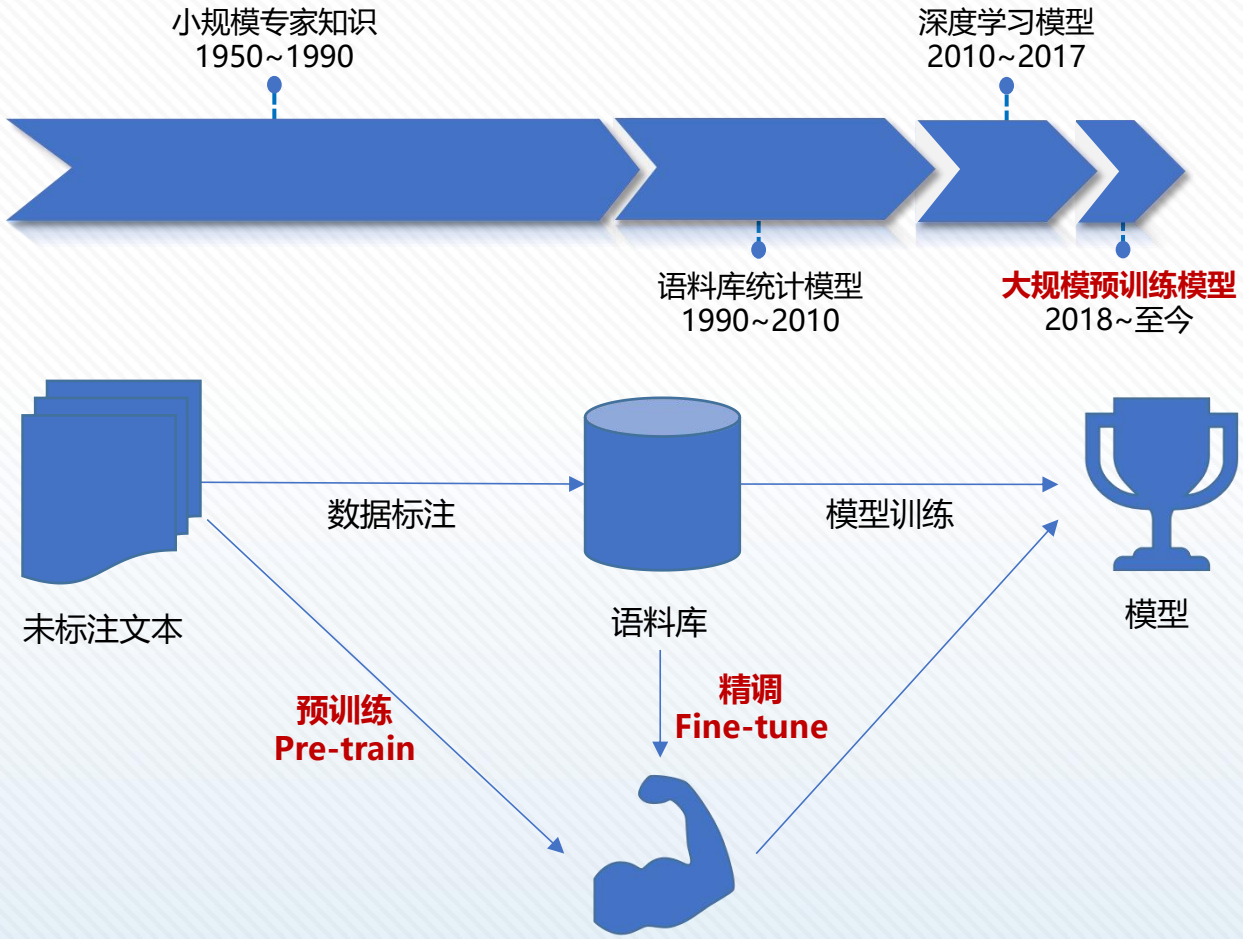


词嵌入表示的赋值方法

- 通过优化在**下游任务**上的表现自动学习



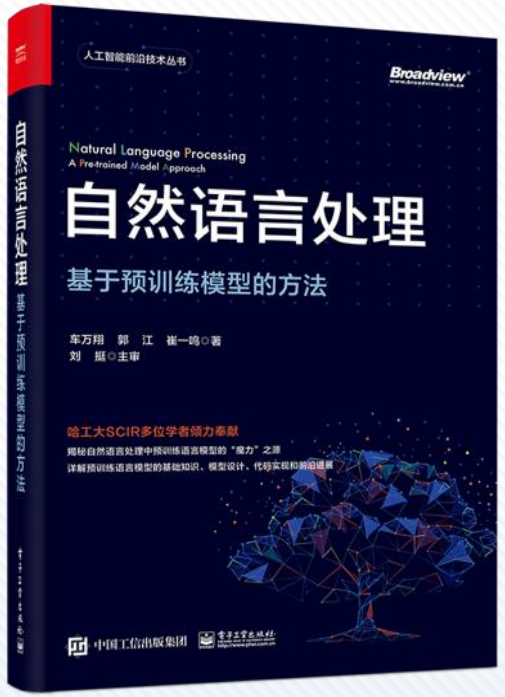
预训练模型获得更好的表示



预训练 + 精调 = 自然语言处理新范式

参考书

- 《自然语言处理：基于预训练模型的方法》
- 出版社：电子工业出版社
- 作者：车万翔，郭江，崔一鸣 著；刘挺 主审
- 书号：ISBN 978-7-121-41512-8
- 出版时间：2021.7
- 网购链接
  - https://item.jd.com/13344628.html
- 书中代码
  - https://github.com/HIT-SCIR/plm-nlp-code



谢 谢!

理解语言，认知社会

以中文技术，助民族复兴

长按二维码，关注哈工大SCIR

微信号：HIT_SCIR