自然语言处理——绪论

杨沐昀

语言技术研究中心
哈尔滨工业大学

课程信息

- 自然语言处理（Natural Language Processing，NLP）
- 课时安排（共32学时）
  - 授课：2 学时 x 12 次（第 2-7 周，周一7-8节、周四5-6节）
  - 实验：4 学时 x　2 次（第 6-7 周，周一9-12 节）
- 考核方式

考核环节	所占分值	考核与评价细则
实验	40%	两个实验项目，每个20分。
期末考试	60%	闭卷考试，采用填空、选择、简答、分析、综合等题型。

课程大纲（1）

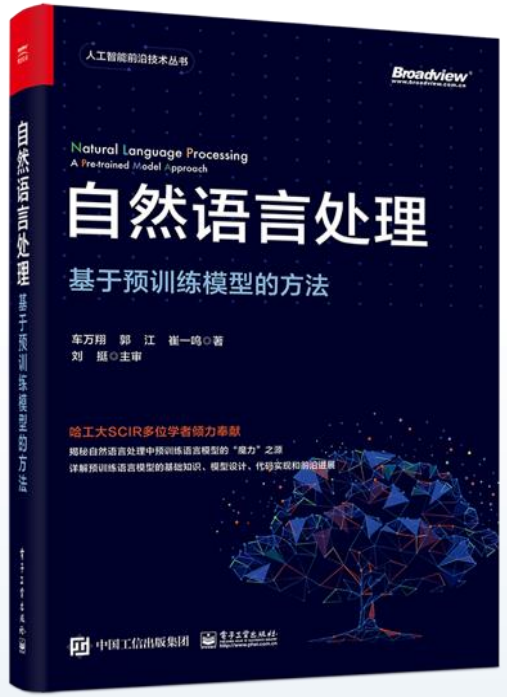
序号	教学内容	学时
1	**绪论** 自然语言处理的发展历史；自然语言处理的主要研究内容和现状、以及所面临的挑战。	2
2	**自然语言处理基础** 文本预处理方法、子词切分方法	2
3	**基础工具集与常用数据集** NLTK、LTP、Spacy等常用NLP基础工具集和各种NLP任务的常用数据集的介绍	2
4	**自然语言处理中的深度学习基础** 全连接网络、卷积神经网络、循环神经网络、注意力机制、Transformer模型	2
5	**语言模型** 语言模型的定义、N元文法模型、神经网络语言模型、基于Transformer的语言模型	2
6	**预训练词向量** 获取静态、动态词向量的常用方法、词向量的评价方法	2
7	**语言模型的预训练** 模型结构：编码器-解码器、单独编码器、单独解码器	2

课程大纲（2）

序号	教学内容	学时
8	**语言模型的微调** 针对分类、序列标注、抽取式问答、语言生成等NLP任务的微调方法	2
9	**提示学习** 语境学习、思维链	2
10	**基于人类反馈的强化学习** 指令微调、强化学习的基本思想、基于人类反馈的强化学习的原理和应用方法	2
11	**大模型的应用** 工具学习、模型压缩	2
12	**自然语言处理的未来发展趋势** 大模型的发展方向、多模态、低资源语言处理	2
13	**实验1：从头实现语言模型**	4
14	**实验2：类ChatGPT通用对话系统**	4

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



课程联系方式

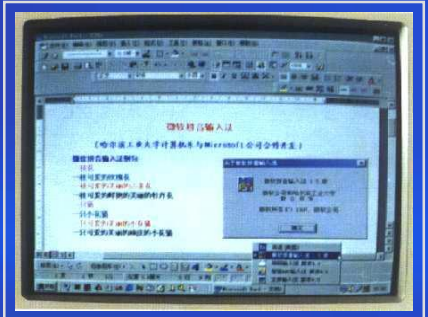
- 课程QQ群：
  - 群号：924975115
  - 群名：24春自然语言处理
  - 入群请实名
    - 助教按选课名单通过
    - 群昵称：姓名-学号
  - 课件会上传到群里

教师介绍

- 杨沐昀（教授/博士生导师）
  - 计算学部语言技术研究中心
  - Research Center on Language Technology
    - 机器智能与翻译研究室
- 研究方向：自然语言处理、信息检索
  - 主页：http://homepage.hit.edu.cn/yangmuyun
- 联系方式
  - Email: yangmuyun@hit.edu.cn

语言技术研究中心

- 1980's全面开展中文信息处理研究
  - 语句级汉字输入系统—微软拼音
  - 大陆第一个汉英机器翻译系统
- 2000.6：哈工大-微软机器翻译技术联合实验室
- 2004.6：哈工大-微软自然语言处理及语音技术联合实验室
- 2004.11：教育部-微软语言语音联合重点实验室



微软拼音 - -
基于语句级的拼音输入系统

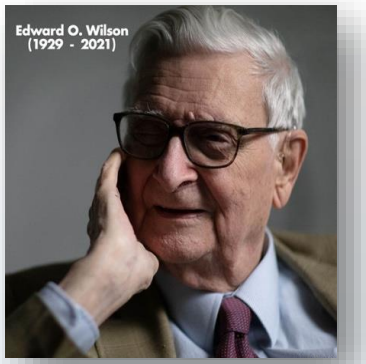
汉英机器翻译系统多次获奖

学科带头人: 李生教授

- 中国中文信息学会理事长（2011~2016）
- 国际计算语言学会终身成就奖
  - 首位华人获奖者、亚洲第2位获奖者
- 中文信息学会终身成就奖
- 汉英机器翻译开拓者
  - 1985年国内最早汉英机器翻译研究之一
  - 1989年国内第一个通过鉴定的汉英翻译系统

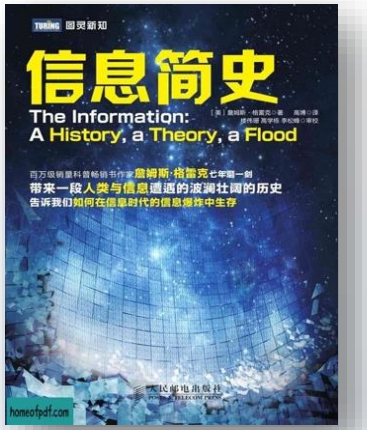
什么是自然语言处理?

- **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



"语言是继真核细胞之后最伟大的进化成就"

——社会生物学之父
爱德华•威尔逊



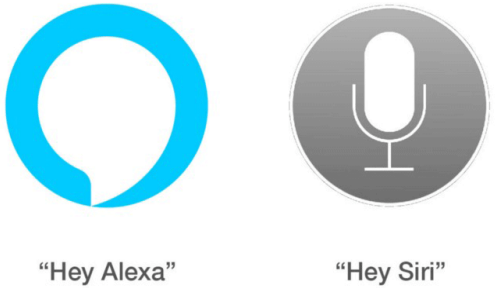
"语言本身就是人类有史以来最大的技术发明"

——**詹姆斯•格雷克**
《信息简史》

- **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

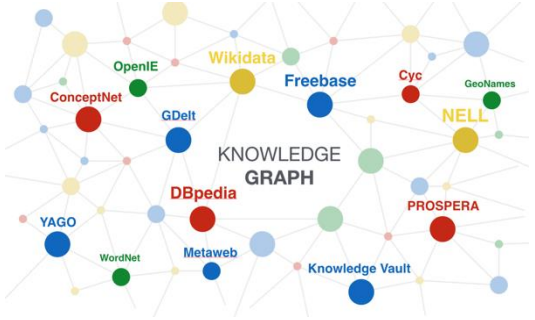


智能助手

文本校对

舆情分析

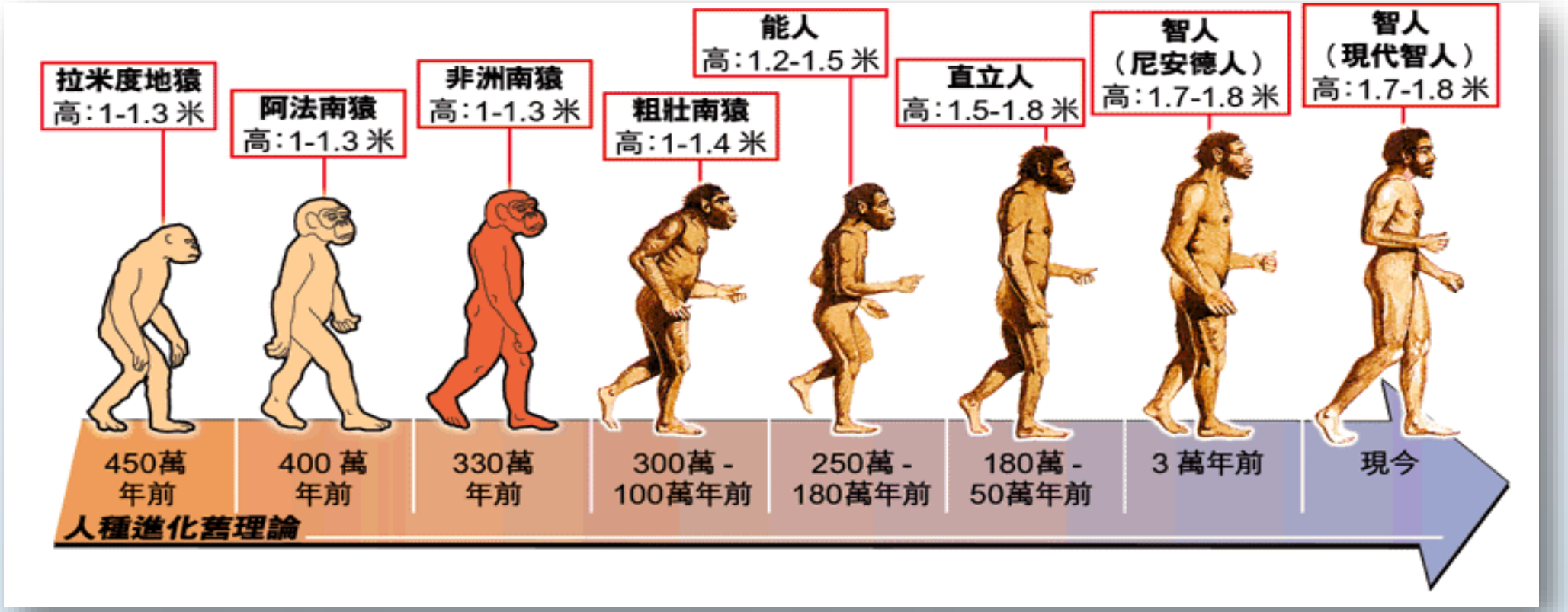
智能教育



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。”

自然语言处理的难点与特点

- “凡规则必有例外”——语言学的尴尬
  - 吃面包 √　吃食堂 √　　喝酒吧?!
  - 咬死猎人的狗—— 谁死了?
  - 了得|了不得　掉地上|掉地下　烟头|烟屁股
    完败|完胜　我可想死你了|你可想死我了
- “这样也行？！”——数学家的无奈
  - 他喝了一杯水。　　他一杯水喝了。
  - 他把一杯水喝了。　　一杯水被他喝了。
  - 一杯水他喝了。　　水他喝了一杯。
  - 枯藤老树昏鸦、古道西风瘦马
  - 研表究明，汉字的序顺并不定一能影阅响读，比如
    发这现里的字全是都乱的。

最佩服的两支球队

中国有两个体育项目大家根本不用看，也不用担心。

一个是乒乓球，一个是男足。

前者是“**谁也赢不了！**”，

后者是“**谁也赢不了！**”

最佩服的也是这两支球队，乒乓球队和男足。一支是“**谁也打不过**”，另一支是“**谁也打不过**”，

——这汉语的表达也是醉了！

“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”

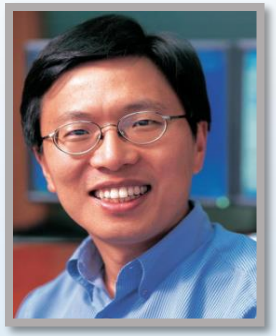
Yann LeCun
图灵奖得主、Facebook AI 负责人

“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

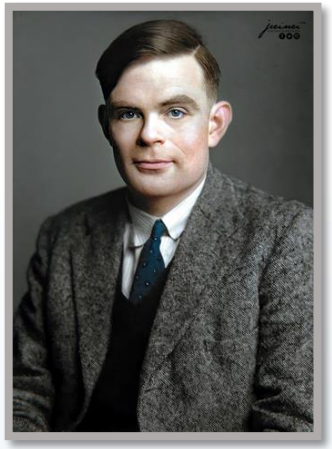
“如果给我10亿美金，我会用这10亿美金建造一个NASA级别的自然语言处理研究项目”

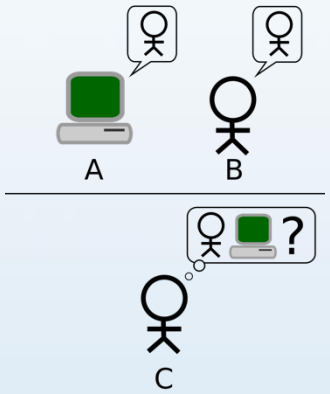
Michael I. Jordan
美国双院院士、世界知名机器学习专家



“下一个十年，懂语言者得天下”

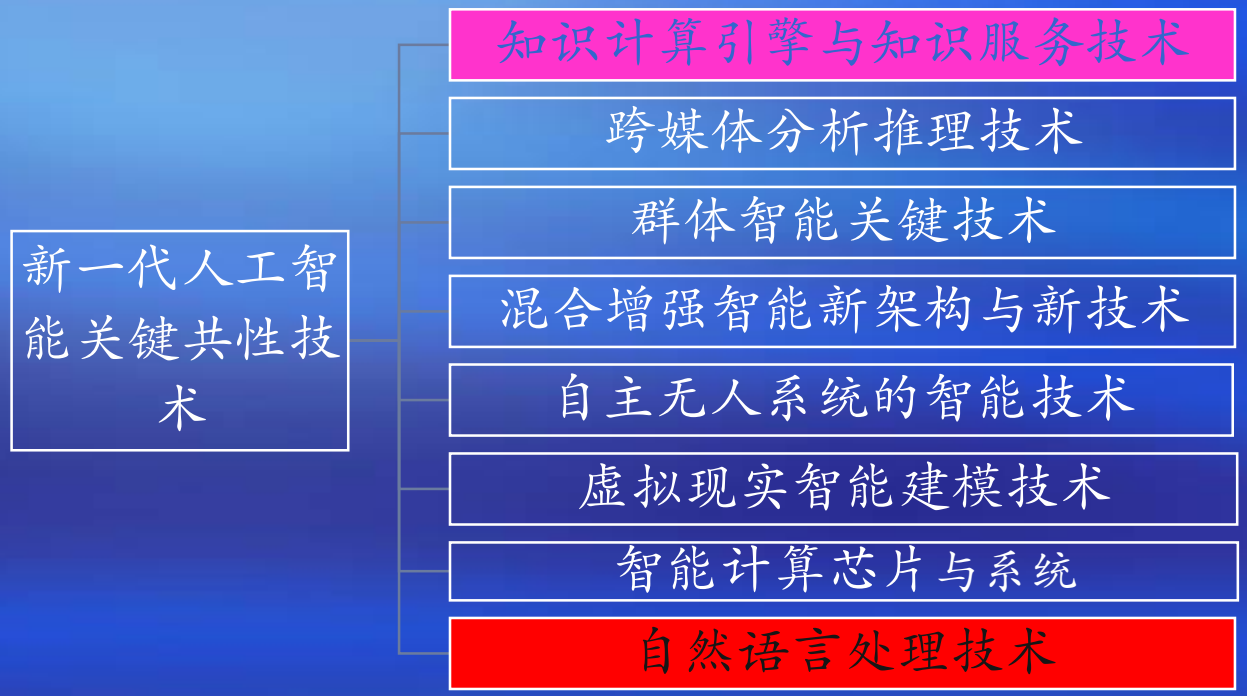
沈向洋
美国工程院院士、微软前全球执行副总裁



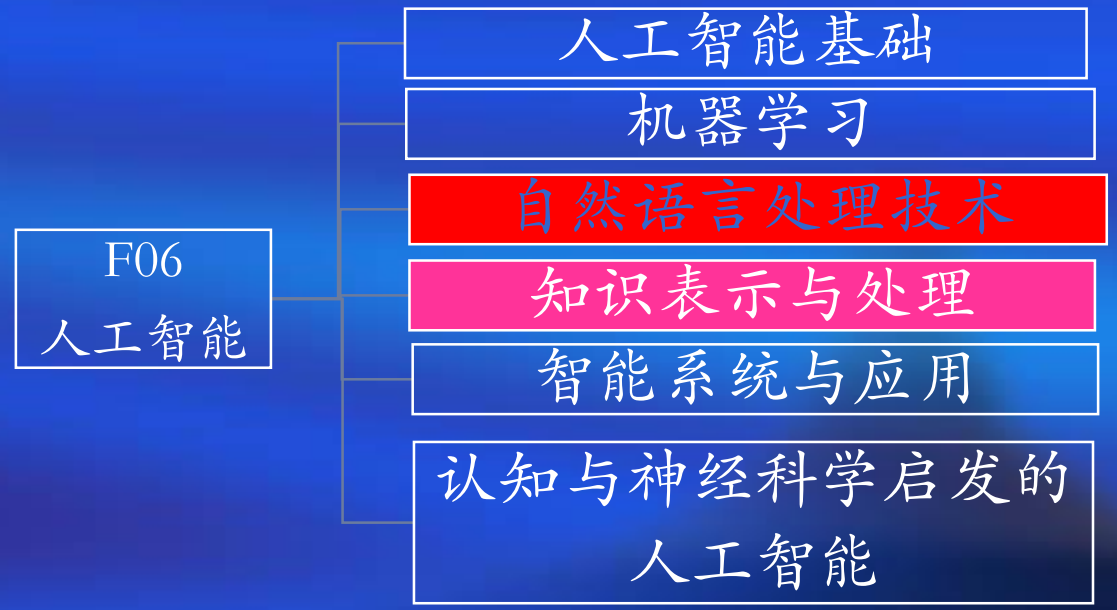


图灵测试

自然语言处理：国家需求

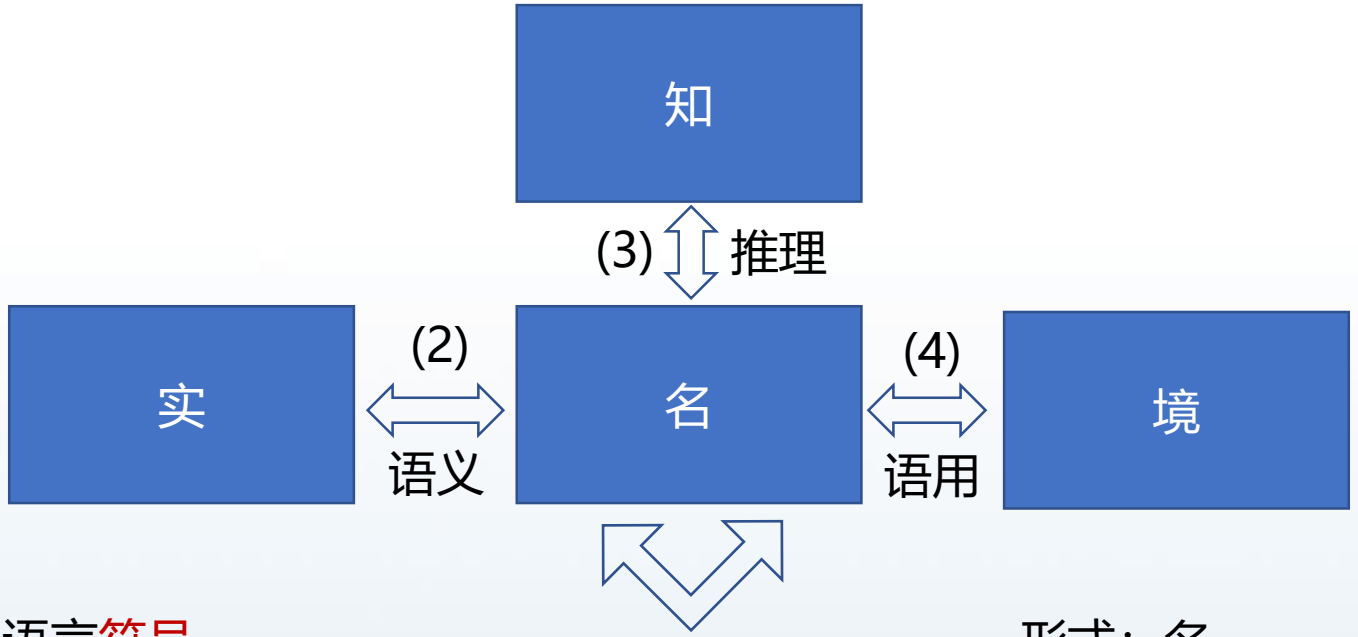


国务院 《新一代人工智能发展规划》，2017



国家自然科学基金申请代码：信息学部

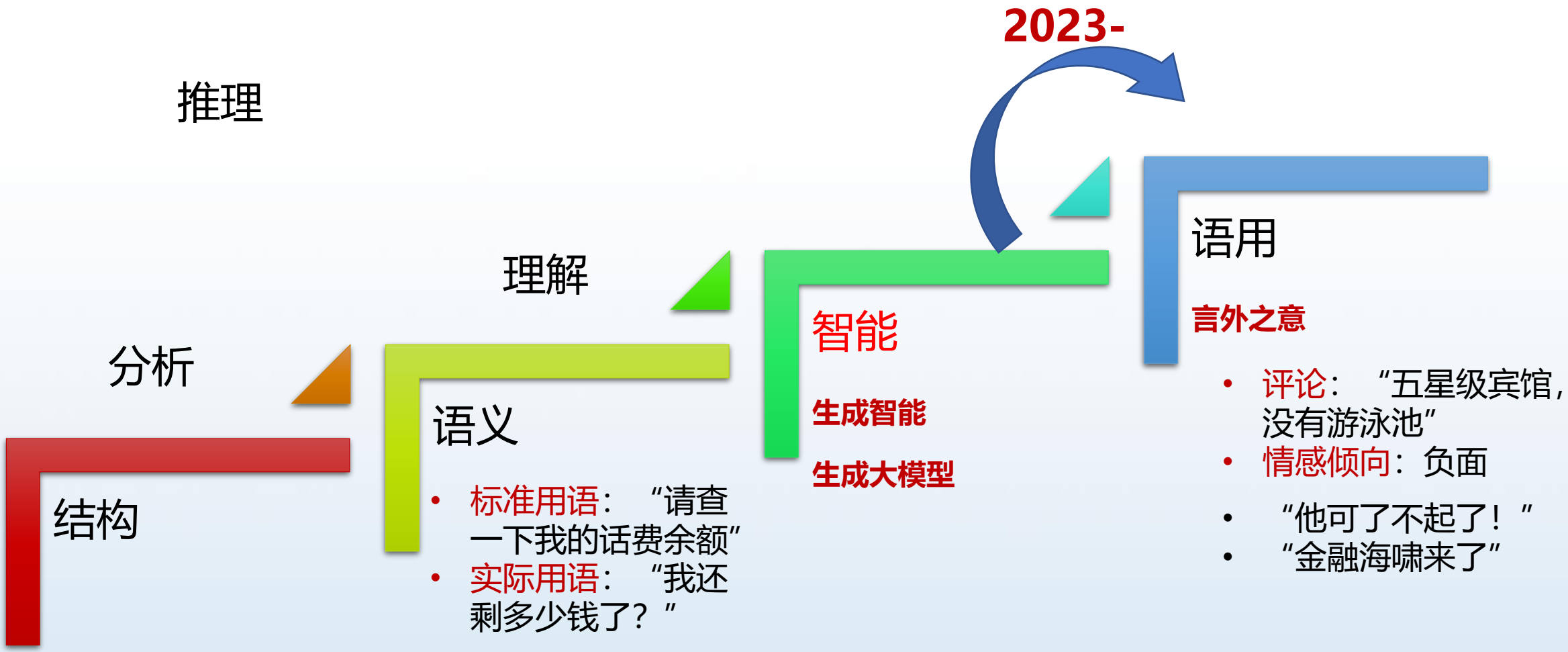
自然语言处理研究对象与层次



名：语言符号
实：客观事实、主观意见
知：知识
境：语言所处环境

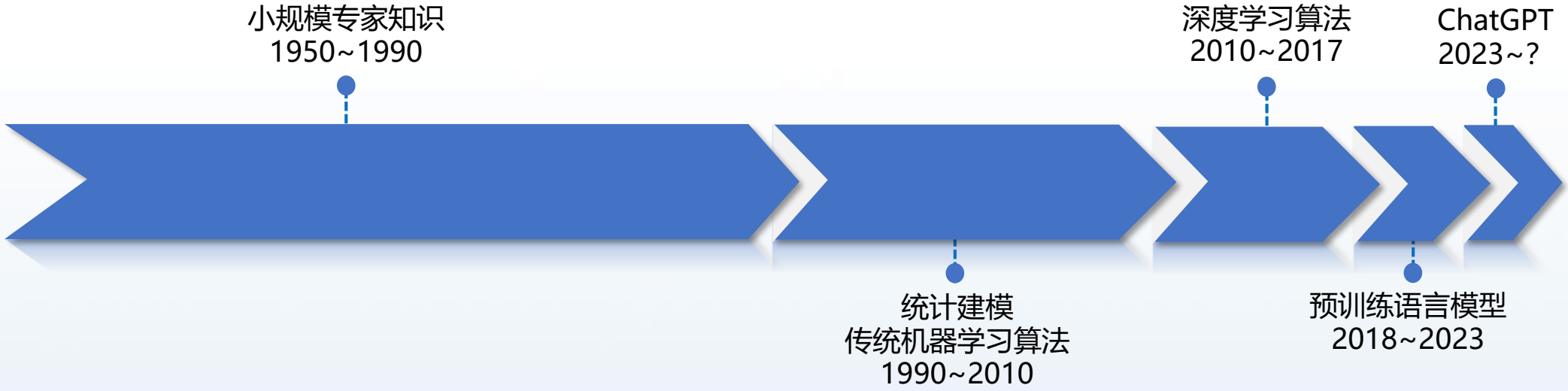
形式：名
语义：名+实
推理：名+实+知
语用：名+实+知+境

自然语言处理的未来

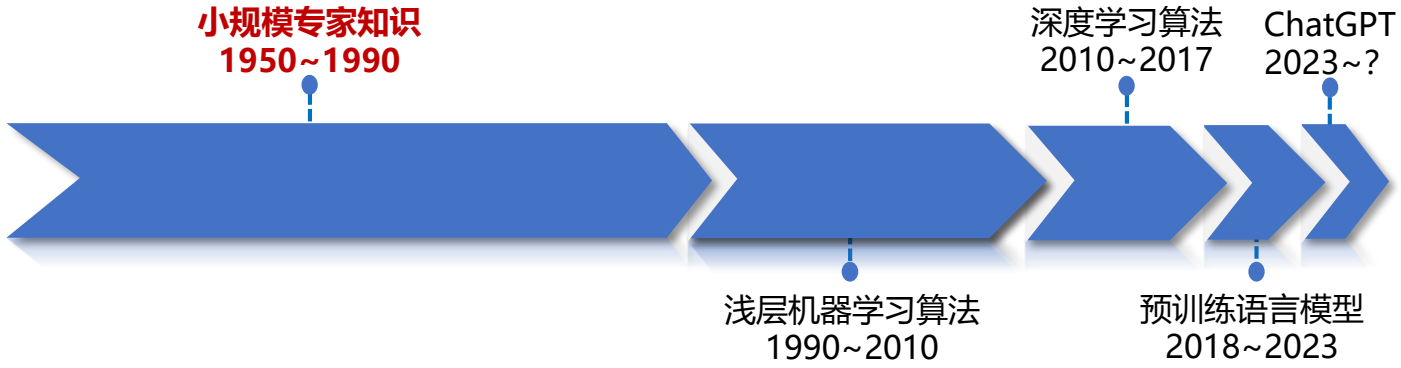

推理
2023-
分析
结构
理解
语义
• 标准用语：“请查一下我的话费余额”
• 实际用语：“我还剩多少钱了？”
智能
生成智能
生成大模型
语用
言外之意
• 评论：“五星级宾馆，没有游泳池”
• 情感倾向：负面
• “他可了不起了！”
• “金融海啸来了”

自然语言处理的发展历史

- 自然语言处理技术经历了**五次范式变迁**



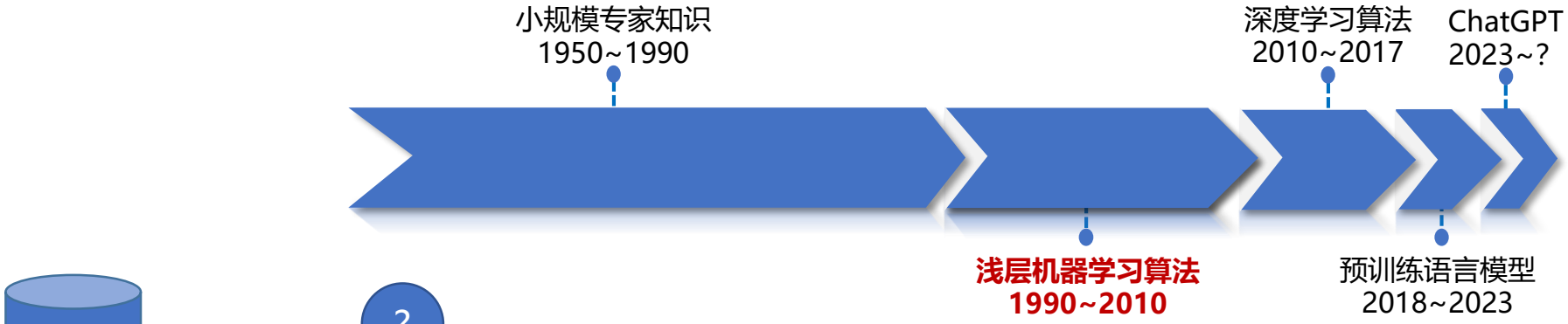
基于符号表示的专家知识

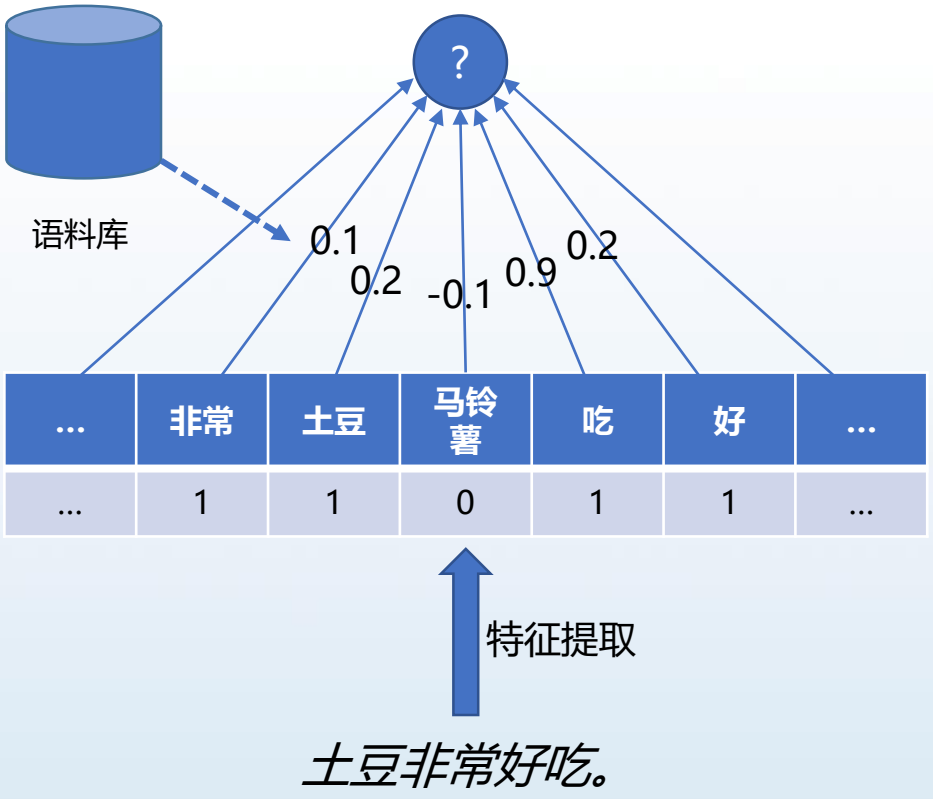


- “*土豆非常好吃。*”的情感倾向性？
  - 如果：出现褒义词“好”“喜欢”等
  - 那么：结果为褒义
  - 如果：出现“不”
  - 那么：结果倾向性取反

- 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

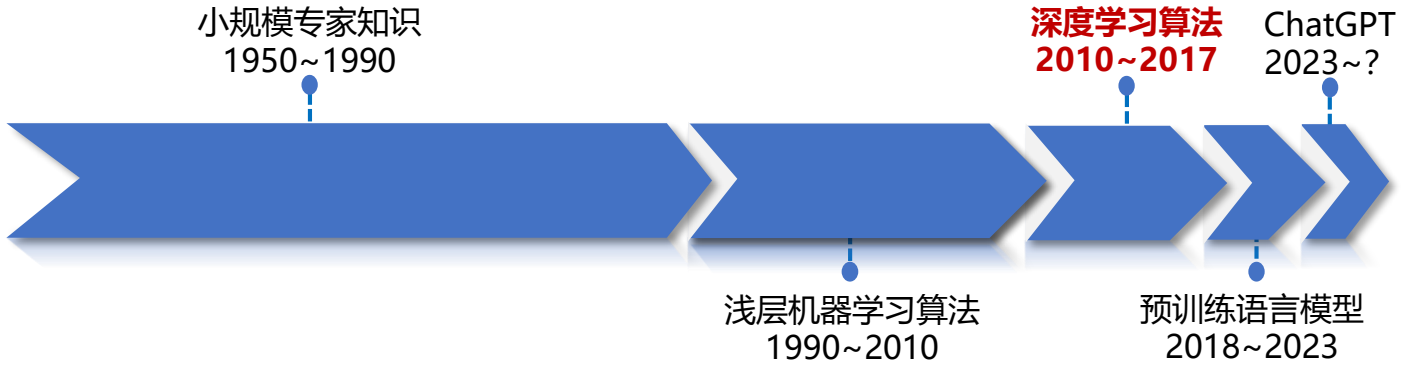
基于向量表示的浅层机器学习





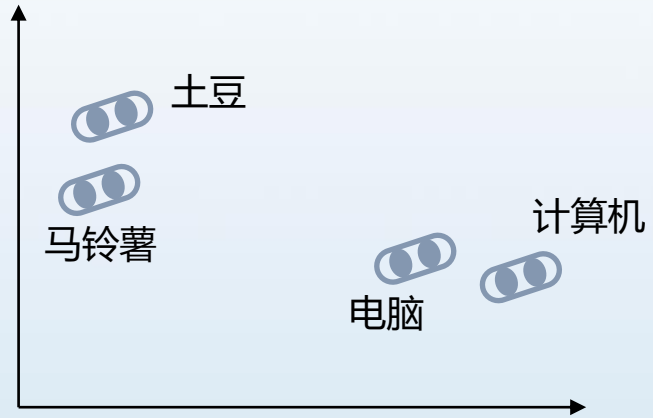
- 使用**高维、离散、稀疏**的向量表示词
  - 维度为词表大小，其中只有一位为1，其余为0
    - **土豆**：[0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, ...]
    - **马铃薯**：[0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, ...]
- 缺点
  - 无法处理“**多词一义**”的现象

基于嵌入表示的深度学习



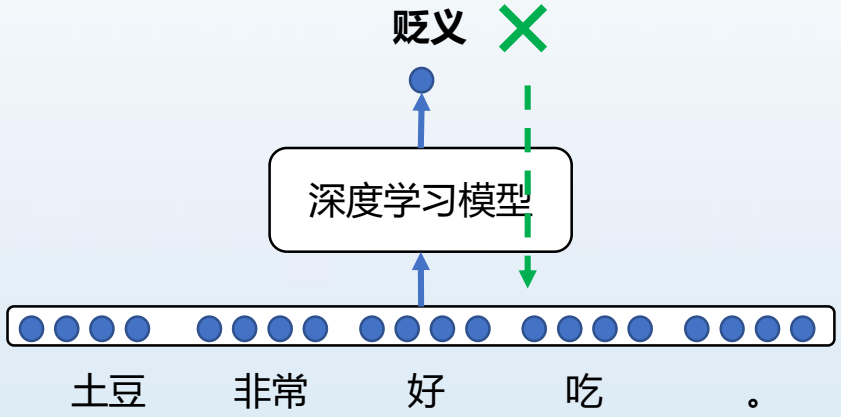
词嵌入（Word Embedding）

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

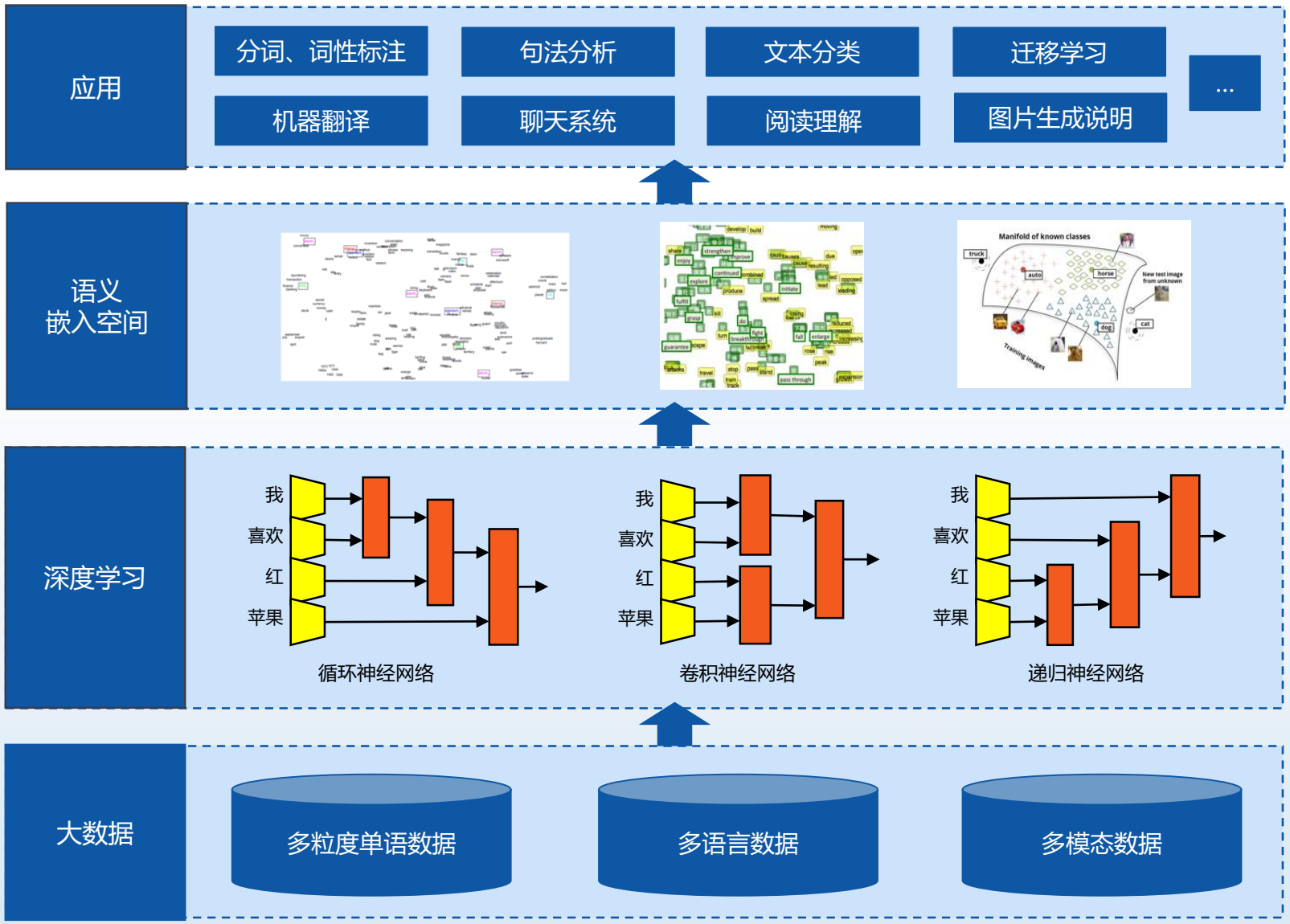


词嵌入表示的赋值方法

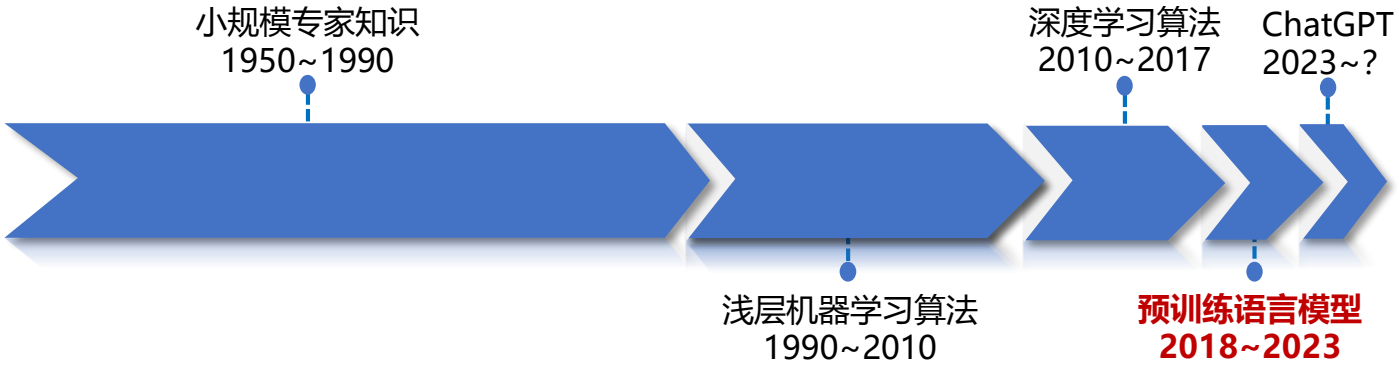
- 通过优化在**下游任务**上的表现自动学习

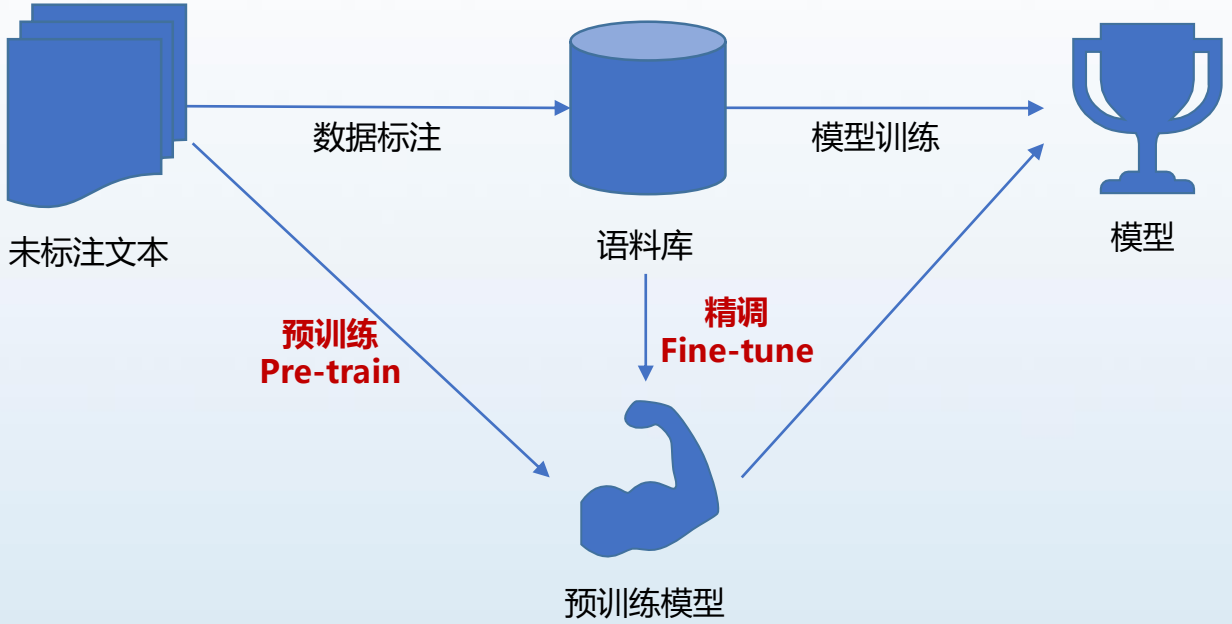


各种语言单元的统一嵌入表示

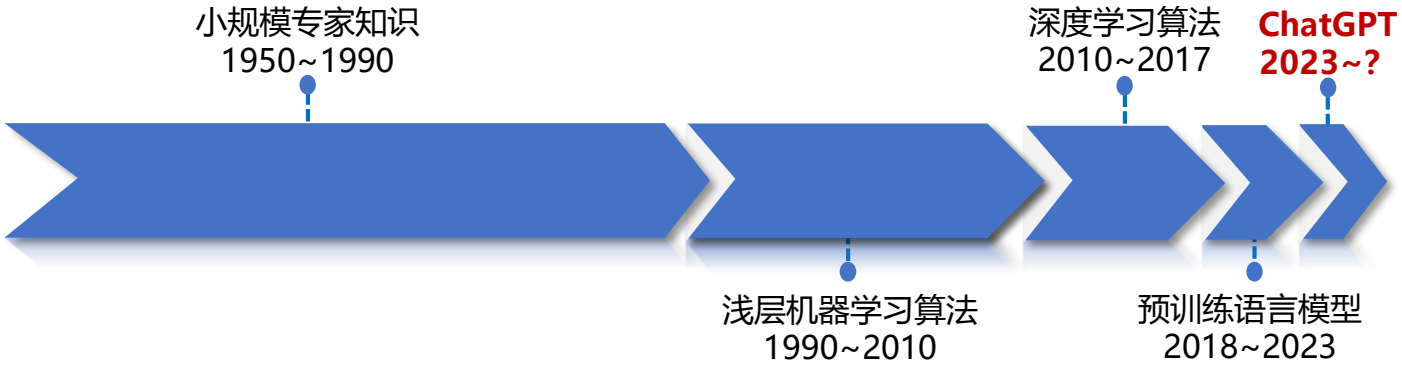


预训练语言模型


小规模专家知识
1950~1990
浅层机器学习算法
1990~2010
深度学习算法
2010~2017
ChatGPT
2023~?
预训练语言模型
2018~2023


数据标注
模型训练
未标注文本
语料库
模型
预训练
Pre-train
精调
Fine-tune
预训练模型

ChatGPT 的表现更加惊艳



71 烤箱和铅笔哪个更重?

一般来说，烤箱会比铅笔重得多。烤箱是一种电器，通常用于烘焙食物或加热食物，因此它通常会比铅笔重得多。铅笔是一种小巧的文具，通常用于写字或画画，因此它通常不会很重。

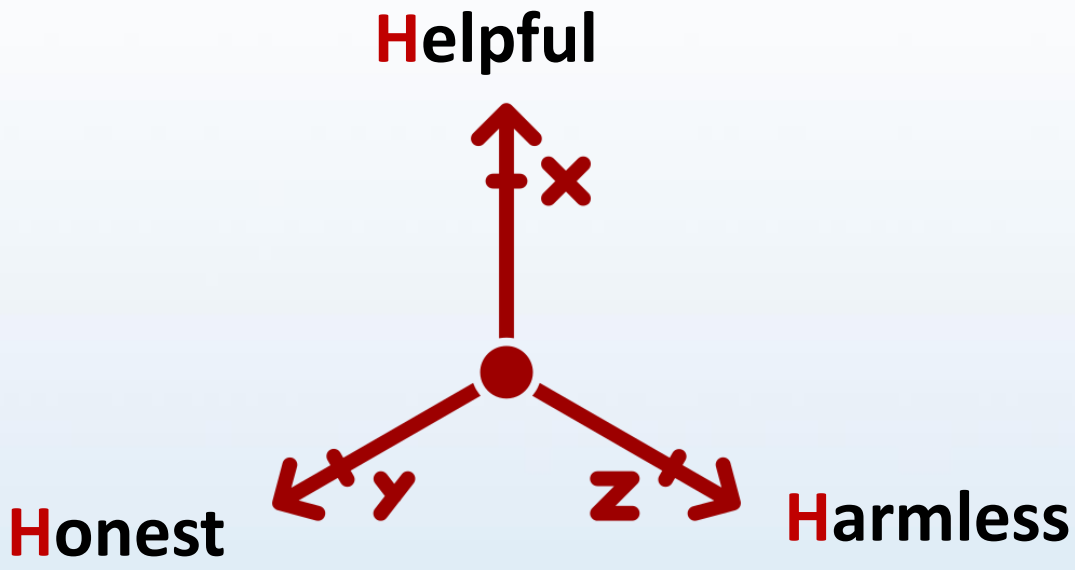
71 我的脚有几只眼睛?

人的脚并没有眼睛。人的脚是由脚趾、脚踝和脚跟组成的，它们可以帮助我们维持平衡，走路和进行其他活动。人的脚趾上有许多小骨骼和肌腱，使我们能够抓地、走路和平衡。人的脚上并没有眼睛。

https://chat.openai.com/

将大模型向人类期望对齐

- **Helpful**: 尽可能帮助人类完成任务，以提高工作效率
- **Honest**: 为人类的问题生成真实的答案
- **Harmless**: 遵守人类伦理道德，不产生偏见或有害的回复



尾声：自然语言处理参考资料

- NLP论文汇总网站
  - http://www.aclweb.org/anthology; arXiv.org
- 主要国际会议
  - ACL/NAACL, EMNLP, SIGIR, AAAI/IJCAI, COLING, HLT, EACL, MT Summit
  - INTERSPEECH, ICASSP
- 主要期刊
  - Computational Linguistics, TACL, Natural Language Engineering, Information Retrieval, Information Processing and Management, ACM TALIP, ACM TSLP
- 学术界主要研究机构
  - Berkeley, Columbia, Stanford, CMU, JHU, Brown, UMass, MIT, UPenn, USC/ISI, Illinois, etc.
  - Toronto, Edinburgh, Cambridge, Sheffield, Saarland, NUS, and many others
  - 清华大学、中科院、北京大学、哈尔滨工业大学、复旦大学、南京大学…
- 工业界主要研究机构
  - Google, MSR, Facebook, OpenAI, IBM

自然语言处理研究

- 主要国际会议
  - ACL/NAACL, EMNLP, SIGIR, AAAI/IJCAI, COLING, HLT, EACL, MT Summit
  - INTERSPEECH, ICASSP
- 主要期刊
  - Computational Linguistics, TACL, Natural Language Engineering, Information Retrieval, Information Processing and Management, ACM Transactions on Information Systems, ACM TALIP, ACM TSLP
- 学术界主要研究机构
  - Berkeley, Columbia, Stanford, CMU, JHU, Brown, UMass, MIT, UPenn, USC/ISI, Illinois, Michigan, UW, Maryland, etc.
  - Toronto, Edinburgh, Cambridge, Sheffield, Saarland, Trento, Prague, QCRI, NUS, and many others
  - 清华大学、中科院、北京大学、哈尔滨工业大学、复旦大学、南京大学…
- 工业界主要研究机构
  - Google, MSR, Yahoo!, IBM, SRI, BBN, MITRE, AT&T Labs
- NLP论文汇总网站
  - http://www.aclweb.org/anthology

谢 谢!