自然语言处理

Meng Yang

http://sdcs.sysu.edu.cn/content/2970

SUN YAT-SEN University

Course Materials

- **Prescribed textbook:**
  - 主要参考斯坦福大学CS224n课程讲义和Princeton COS597G课程讲义
- **Other references:**
  - 《模式识别》
  - 《深度学习》
- **Lecture notes will be available online**

课程企业微信群

自然语言处理2024

此群是企业内部群聊，仅企业成员可扫码加入

课程学习形式

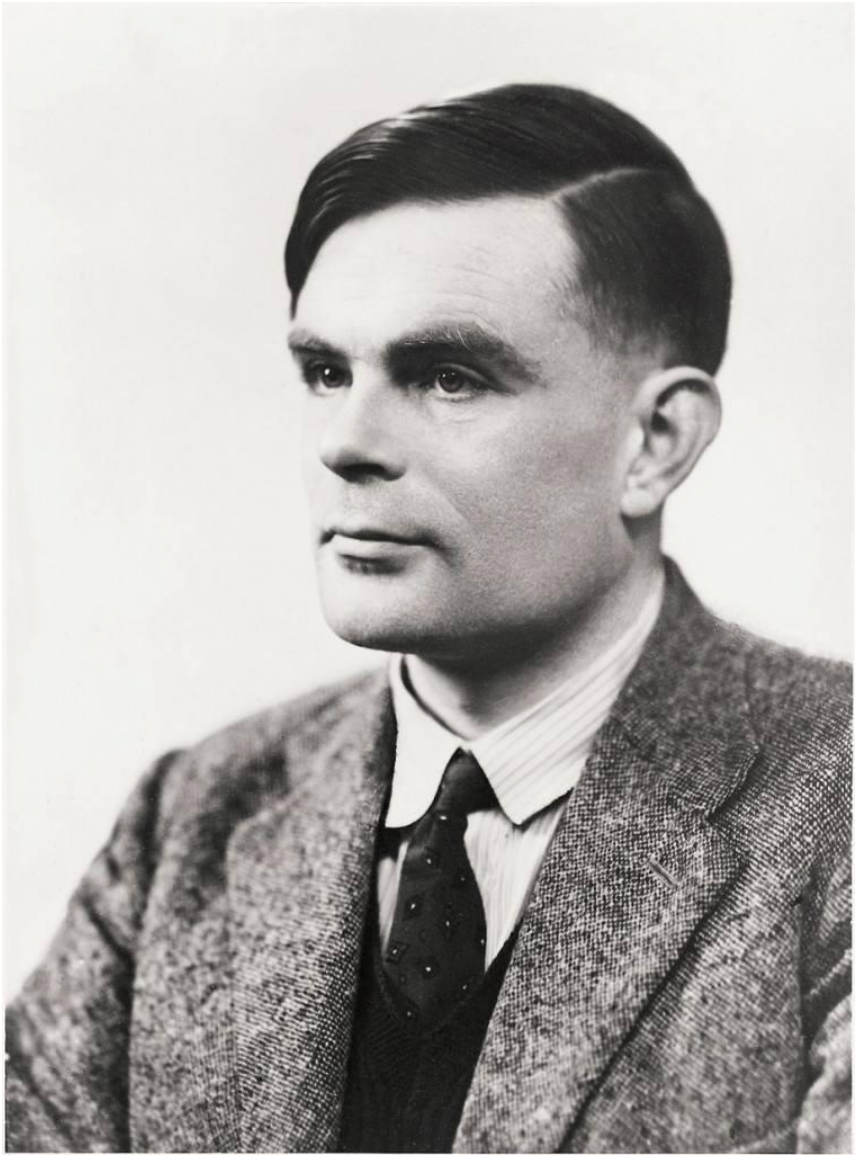
- 课堂讲授
- 课堂提问
- 研讨课
- 前排就坐

- 编程作业
- 期末大作业

- 在线讨论反馈

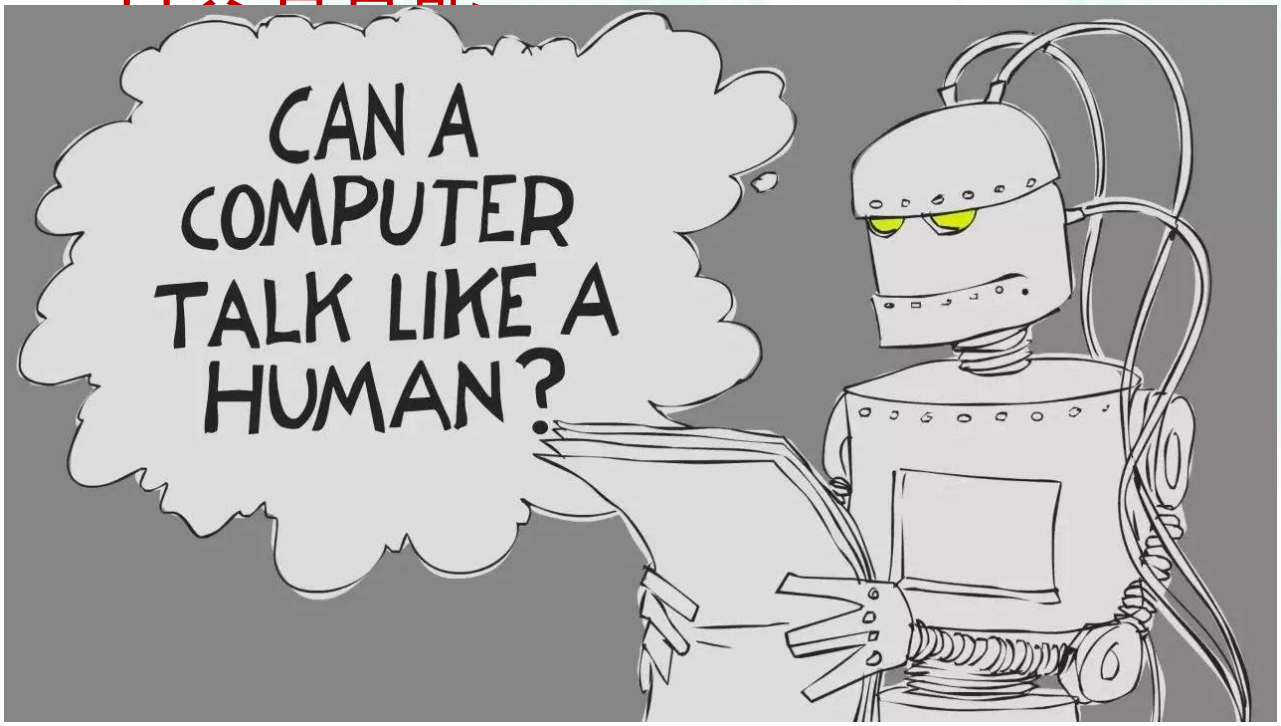
教学内容

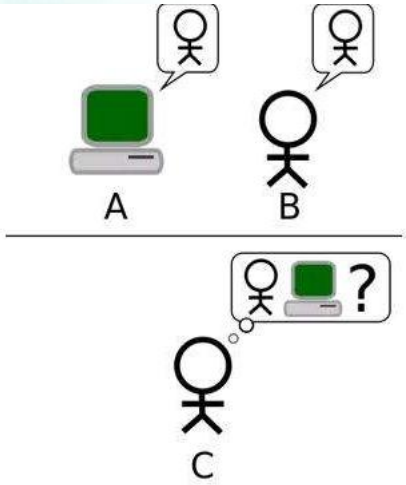
- 词向量
- 神经网络与反向传播
- 序列模型基础
- 注意力机制（Attention）
- Seq2Seq模型与Transformer
- 预训练模型与BERT
- 大语言模型技术
- GPT系列技术研讨
- 下游任务应用（如问答系统等）



图灵测试

- 如何知道一个系统是否具有智能？
- 1950年计算机科学家图灵提出了著名的“图灵测试”。
- 通过人和机器之间的自然语言对话来判断机器是否具有智能





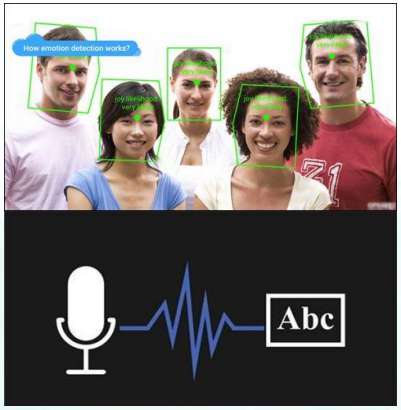
Announcement for all

- 期末成绩（60%）+ 平时成绩（40%）
- 期末成绩=大作业（算法+报告+编程实现效果）
- 平时成绩 = 理论测试 + 作业 + 考勤
- 考勤分：常规考勤 + 提问考勤
- 杜绝抄袭，按时提交
- 认定为抄袭者，此次作业0分
- 不按时提交者，此次这业0分
- 考勤不到者，按照学校规定处理

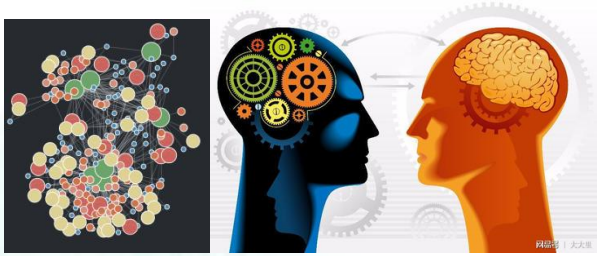
第一部分：NLP基础

人工智能发展

运算智能



感知智能



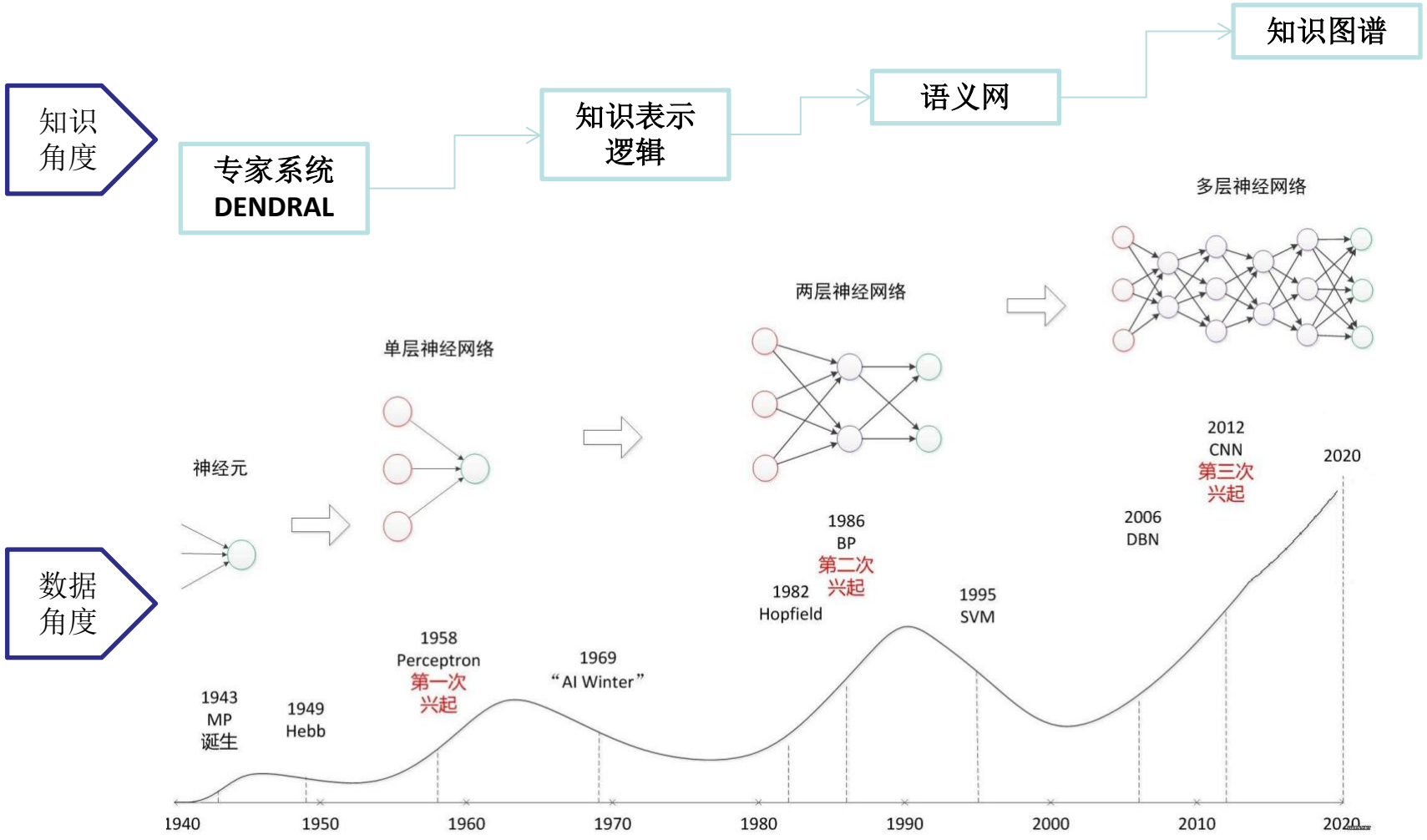
认知智能

掌握知识、进行推理

能理解会思考

知识引导+数据智能

人工智能发展


知识
角度
专家系统
DENDRAL
知识表示
逻辑
语义网
知识图谱
数据
角度
神经元
单层神经网络
两层神经网络
多层神经网络
1943
MP
诞生
1949
Hebb
1958
Perceptron
第一次
兴起
1969
“AI Winter”
1982
Hopfield
1986
BP
第二次
兴起
1995
SVM
2006
DBN
2012
CNN
第三次
兴起
2020
1940
1950
1960
1970
1980
1990
2000
2010
2020

NLP是认知智能的核心

“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”

深度网络之父：Geoffrey Hinton

“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。”

机器学习专家、美国双院院士
Michael I. Jordan

“深度学习的下一个前沿课题是自然语言理解。”

Facebook人工智能负责人：Yann LeCun

“下一个十年，懂语言者得天下”

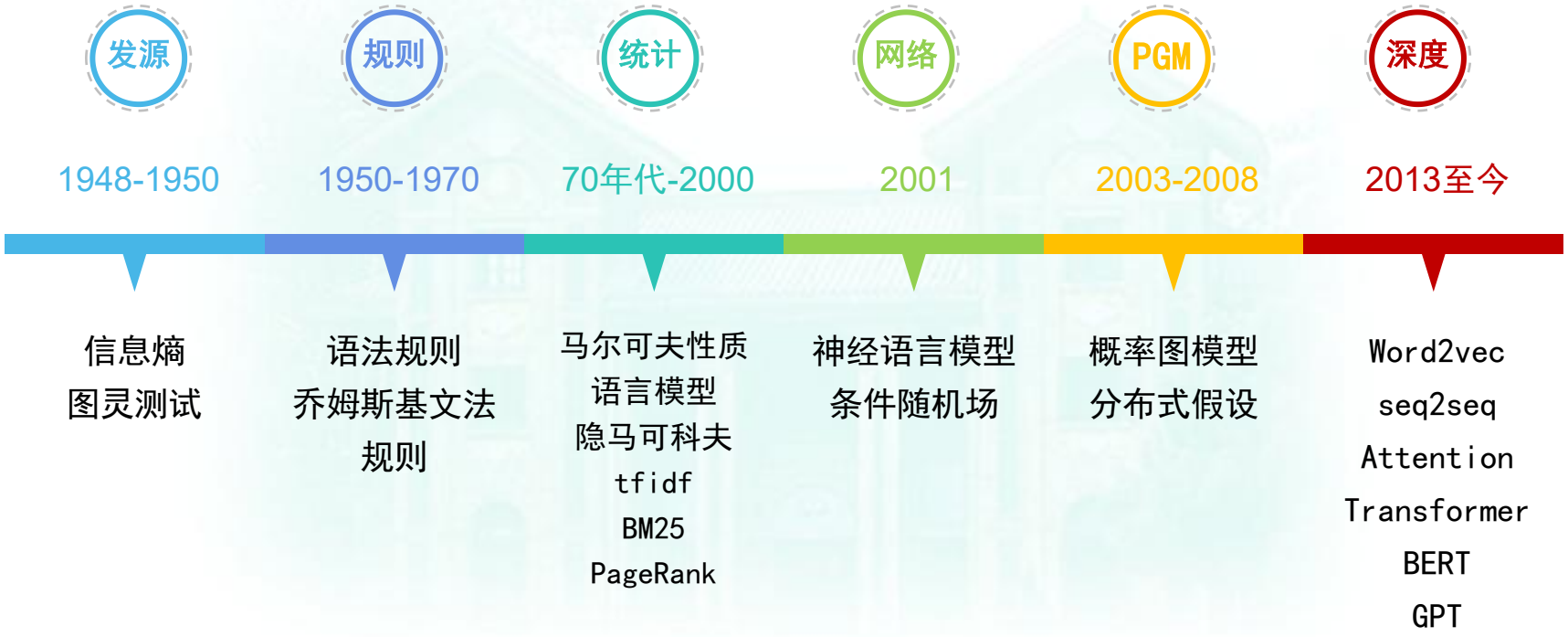
微软全球执行副总裁：沈向洋

刘挺，哈工大，中文信息处理前沿技术进展

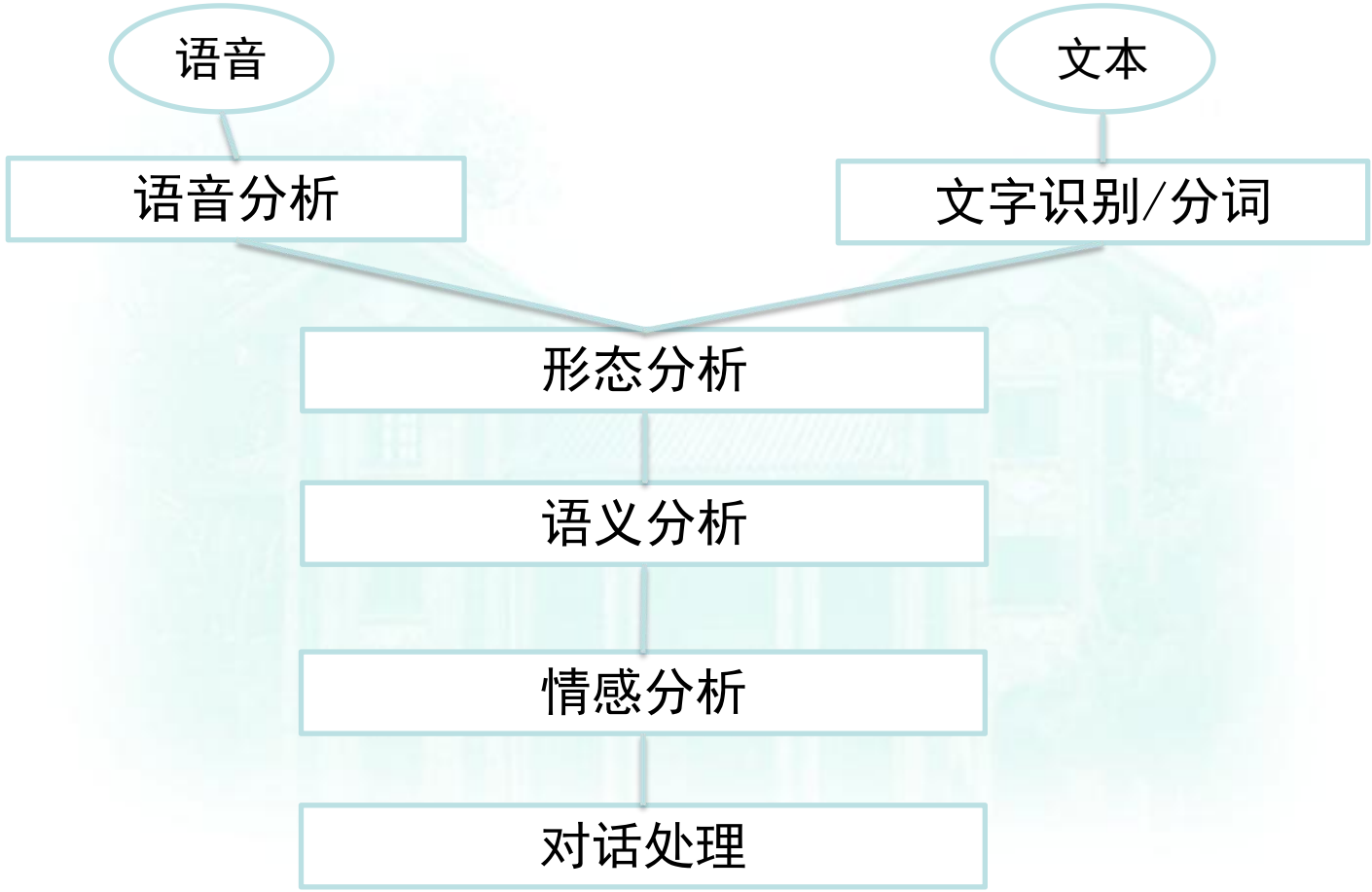
什么是NLP

- 自然语言处理（Natural Language Processing，NLP）是指用计算机对语言信息进行处理的方法和技术。
- NLP是一门交叉学科。
  - 计算机科学
  - 人工智能
  - 逻辑学
- 目标是让计算机能够处理和理解自然语言并实现一些有用的任务。
  - 智能助理（Siri，Google Assistant，Facebook M，Cortana...）
  - 机器翻译
  - …

NLP发展


发源
1948-1950
信息熵
图灵测试
规则
1950-1970
语法规则
乔姆斯基文法
规则
统计
70年代-2000
马尔可夫性质
语言模型
隐马可科夫
tfidf
BM25
PageRank
网络
2001
神经语言模型
条件随机场
PGM
2003-2008
概率图模型
分布式假设
深度
2013至今
Word2vec
seq2seq
Attention
Transformer
BERT
GPT

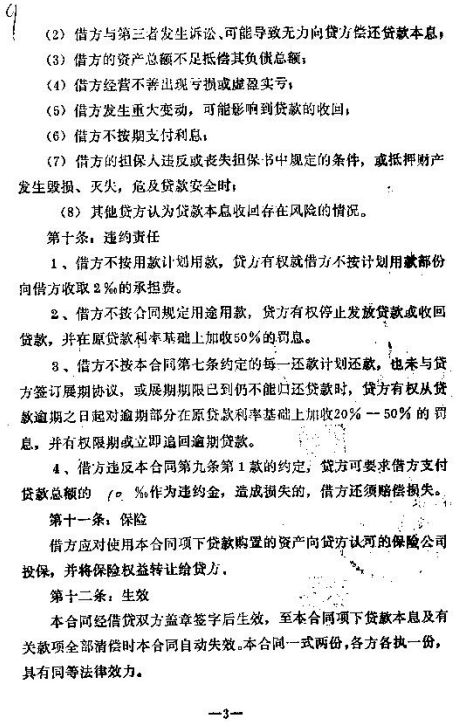
NLP的层次



NLP的层次

- （文字识别）百度OCR中文识别API：

https://console.bce.baidu.com/ai/?fromai=1#/ai/ocr/overview/index



（2）借方与第三者发生诉讼、可能导致无力向贷方偿还贷款本息；
（3）借方的资产总额不足抵偿其负债总额；
（4）借方经营不善出现亏损或虚盈实亏；
（5）借方发生重大变动，可能影响到贷款的收回；
（6）借方不按期支付利息；
（7）借方的担保人违反或丧失担保书中规定的条件，或抵押财产发生毁损、灭失，危及贷款安全时；
（8）其他贷方认为贷款本息收回存在风险的情况。
第十条：违约责任
1、借方不按用款计划用款，贷方有权就借方不按计划用款部份向借方收取2‰的承担费。
2、借方不按合同规定用途用款，贷方有权停止发放贷款或收回贷款，并在原贷款利率基础上加收50%的罚息。
3、借方不按本合同第七条约定的每一还款计划还款，也未与贷方签订展期协议，或展期期限已到仍不能归还贷款时，贷方有权从贷款逾期之日起对逾期部分在原贷款利率基础上加收20%—50%的罚息，并有权限期或立即追回逾期贷款。
4、借方违反本合同第九条第1款的约定，贷方可要求借方支付贷款总额的 (0. %作为违约金，造成损失的，借方还须赔偿损失。
第十一条：保险
借方应对使用本合同项下贷款购置的资产向贷方认可的保险公司投保，并将保险权益转让给贷方。
第十二条：生效
本合同经借贷双方盖章签字后生效，至本合同项下贷款本息及有关款项全部清偿时本合同自动失效。本合同一式两份，各方各执一份，具有同等法律效力。
—3—

〇'〈2) 借方与第三者发生诉讼、可能导致无力向贷方偿还贷款本息，〈3) 借方的资产总额不足抵偿其负债总额;(4〉 借方经营不善出现亏损或虚盈实亏;〕〈5) 借方发生重大变动，可能影响到贷款的收回;(6) 借方不按期支付利息;〈7) 借方的担保人违反或丧失担保书中规定的条件，或抵押财产发生毁损、灭失，危及贷款安全时;_翼(8〉其他贷方认为贷款本息收回存在风险的情况。^第十条= 违约责任1、借方不按用款计划用款，贷方有权就借方不按计划用蒜部份向借方收取2%。的承担费。'`,''2`借方不按合同规定用途用款，贷方有权停止发敌黛款或收回贷款，并在原贷款利率基础上加收跚%的罚息。``)、二8、_借方不按本合同第七条约定的每六还款计划还款{拖蹇与贷方签订展期协议，或展期期限已到仍不能归还贷款时，贷芳有权从贷'款逾期之日起对逾期部分在原贷款利率基础上加喀芒%一50% 的罚息，并有权限期或立即追回逾期贷款。言…薯翼薯趸'4、借方违反本合同第九条第1款的约定蓁『'〕贷圭可要求借方支付贷款总额的…缠[滩作为违约金，造成损失的，借方还须赔偿埠朱镳第十一条'保险＿＿＿.氛蔓〈『借方应对使用本合同项下贷款购置的资产向贷辕洮萼鳟葆痊公司投保，并将保险权益转让给贷方'"'_"羁寺_第十二条=生效藩锗〉…翼'本合同经借贷双方盖章签字后生效，至本杳茴项下贷款本息及有关款项全部清偿时本合同自动失效。本合同一式两份，各方各执一份，具有同等法律效力。—3—

NLP的层次

- （语音识别）科大讯飞语音识别API：

  https://www.xfyun.cn/services/voicedictation

- TTS(语音合成) http://speech.diotek.com/en/text-to-speech-demonstration.php

- （形态分析）NLTK工具包的词干提取：

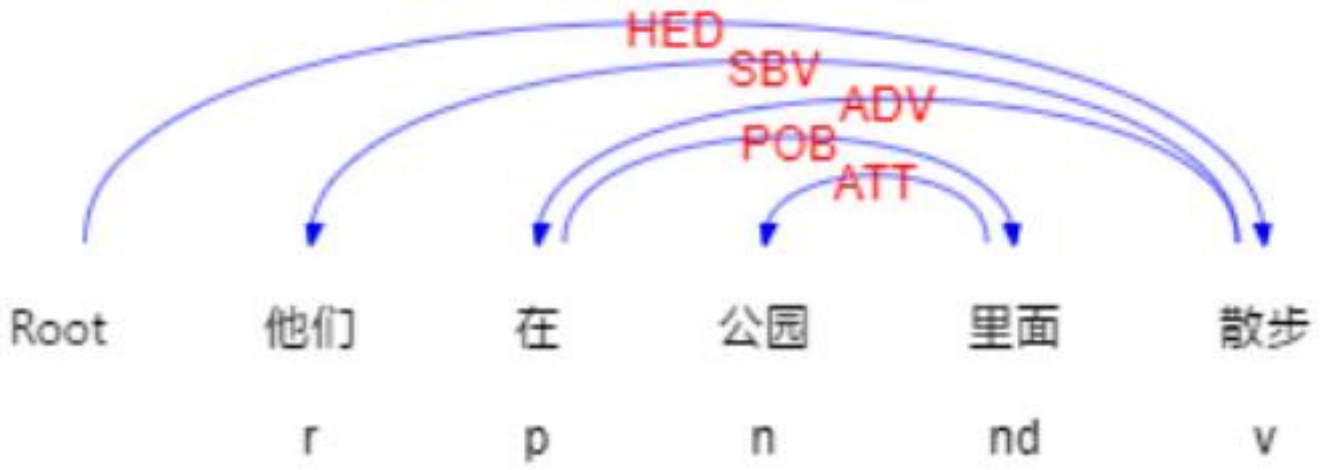
  http://www.nltk.org/

```
>>>from nltk.stem.porter import PorterStemmer
>>>stem = PorterStemmer()

>>>word = 'playing'
>>>stem.stem(word)

'play'
```

NLP的层次

- （分词、句法分析）哈工大NLP平台LTP：

  http://ltp.ai/



NLP的层次

- （情感分析）讯飞的情感分析API：

https://www.xfyun.cn/services/emotion-analysis

NLP的层次

- （对话处理）任务导向型（task-oriented）

对话系统（智能客服、助理）：

NLP技术的一些应用

- 拼写检查　（Word自动拼写检查）
- 机器翻译　（Google、Baidu翻译）
- 自动摘要

  （Text rank抽取式方法、end to end的生成式方法）
- 文本分类和信息过滤

  （特征提取的方法、深度学习的方法）
- 信息检索　（Google学术搜索）
- 信息抽取和文本挖掘

  （机器阅读理解、LDA主题分析）
- 情感分析　（基于情感词典的方法、深度学习的方法）
- 问答系统　（各类智能客服、助理,chatgpt）

NLP的主要任务

- 语言分析：分析语言表达的结构和含义
  - 词法分析：形态还原、词性标注、命名实体识别、分词等
  - 句法分析：组块分析、结构分析、依存分析
  - 语义分析：词义、句意（逻辑关系）、上下文（指代、实体关系）
- 语言生成：从某种内部表示生成语言表达
- 多语言处理：语言之间的对应、转换（机器翻译、跨语言检索）
- 不同的应用对以上任务有不同需求

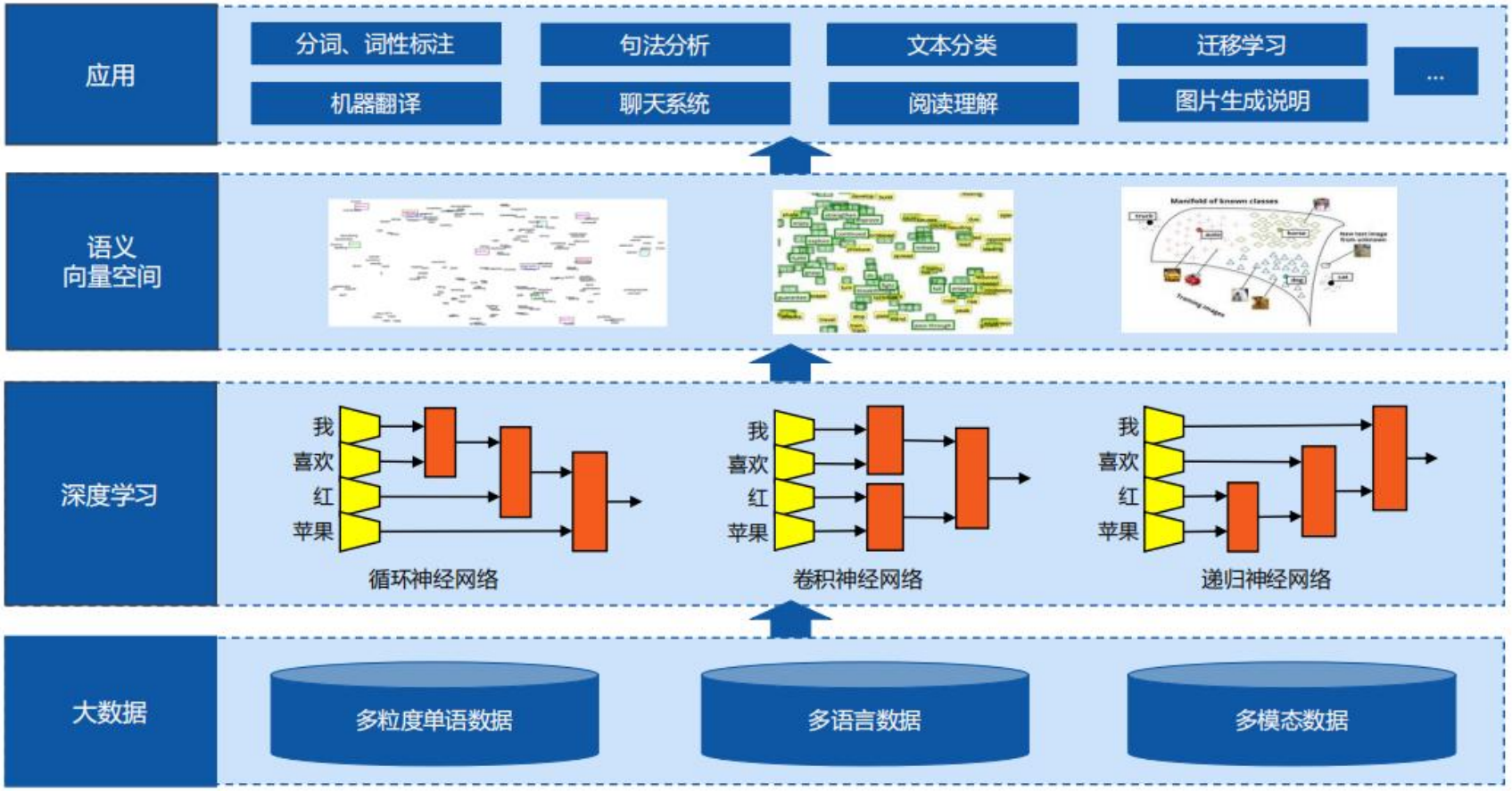
NLP的主要实现方法

- 基于规则的理性方法
  - 基于规则的知识表示和推理（符号计算）
  - 强调人对语言知识的理性整理（知识工程）
  - 受计算语言学理论指导
- 基于语料库的经验方法
  - 以大规模语料库为语言知识基础
  - 利用统计学习和基于神经网络的深度学习方法自动获取和运用隐含在语料库中的知识
  - 学习到的知识体现为一系列模型参数
- 混合方法
  - 理性方法处理效率比较高，但是鲁棒性差、知识扩充困难
  - 经验方法鲁棒性较好，但是缺乏对语言学知识的深入描述和应用、处理效率低
  - 结合两种方法的优点

NLP的难点

- 歧义
  - 有限的词汇和规则表达复杂、多样的对象
- 语言知识的表示、获取和运用
- 成语和惯用语的处理
- 对语言的灵活性和动态性的处理
  - 同一个意图的不同表达，包括包含错误语法的习惯用语
  - 语言在不断地变化，新词的出现
- 上下文和世界知识（常识）的利用和处理

深度学习：目前NLP所采用的主要技术手段



刘挺，哈工大，中文信息处理前沿技术进展