自然语言处理

Natural Language Processing

课程性质：专业主干课

孔力
kongli@nnu.edu.cn

课程目标

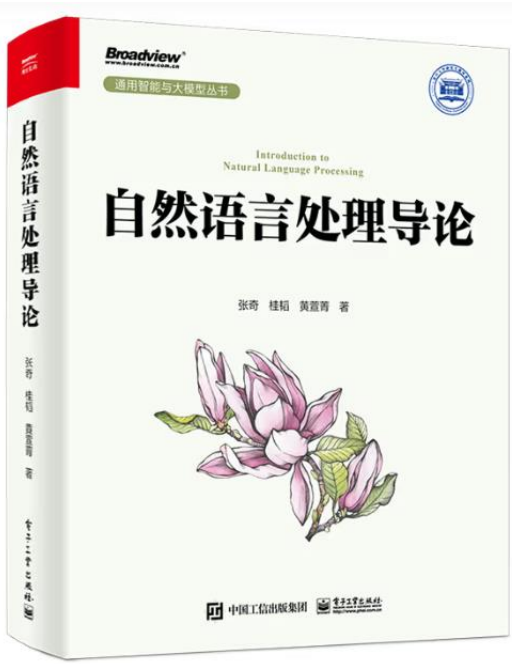
- 课程目标1：掌握自然语言处理的概念、发展历程、发展现状与发展趋势，熟悉自然语言处理中常见的基础任务和应用任务，掌握不同任务的主要评价标准，能够实现具体问题到自然语言处理任务的转化。
- 课程目标2：掌握自然语言处理的常见技术、模型、工具，特别是深度学习技术、预训练技术在文本处理中的应用。
- 课程目标3：理解我国自然语言处理技术的发展现状、与国际先进水平的差距、发展趋势和未来挑战，培养创新精神、工匠精神与爱国精神。

大纲

- 绪论
- NLP中的前馈神经网络　　介绍任务：依存分析
- 预训练1：词向量
- NLP中的循环神经网络　　介绍任务：文本分类，命名实体识别
- NLP中的卷积神经网络
- 端到端与注意力机制　　介绍任务：机器翻译等文本生成任务
- Transformer模型
- 预训练2：预训练模型
- 16次新课授课，1次复习

参考书目

- 《自然语言处理导论》，张奇，电子工业出版社
- 《自然语言处理》，刘挺 等，高等教
- 《统计自然语言处理（第2版）》，宗 社
- 斯坦福**NLP**教程：https://web.stanfor
- Neural Network Methods for Natural ，Yoav Goldberg
- 复旦邱锡鹏 神经网络与深度学习： h
- 台大李宏毅 机器学习：https://speech.ee.ntu.edu.tw/~hylee/ml/2021-spring.php
- Neural Networks and Deep Learning，Michael Nielsen



课程介绍

- 课程教学与考核

课程教学：

课堂讲授为主，辅以讨论、交流、互动；

授课内容和形式：理论知识、视频、文件、论文等

课程考核：

平时成绩10%：考勤、课堂表现；

过程性考核40%：作业、实验、测验；

期末考试50%：闭卷。

绪论

- 自然语言处理的概念
- 自然语言处理的主要研究内容
- 自然语言处理研究的困难
- 自然语言处理的研究方法

什么是自然语言

- 自然语言是指人类日常使用的语言，如汉语、英语、法语，德语，等等
- 区别于人工语言, **e.g.** 编程语言
- 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
- 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

问题的提出

- 如何让计算机实现自动的或人机互助的语言处理功能？
- 如何让计算机实现海量语言信息的自动处理、知识挖掘和有效利用？

自然语言处理
Natural Language Processing, NLP

什么是自然语言处理

- **Natural Language Processing (NLP)**
  - NLP is the branch of computer science focused on developing systems that allow computers to communicate with people using everyday language.
  - **自然语言处理要研制表示语言能力(linguistic competence)和语言表现(linguistic performance)的模型，建立计算框架来实现这样的语言模型，提出相应的方法来不断地完善这样的语言模型，根据这样的语言模型设计各种实用系统，并探讨这些实用系统的评测技术。**

    —马纳瑞斯**(Bill Manaris)**在《从人—机交互的角度看自然语言处理》

什么是自然语言处理

- 自然语言处理（Natural Language Processing, NLP）是人工智能领域的重要研究方向之一，旨在探索实现人与计算机之间用自然语言进行有效交流的理论与方法。
- 它融合了**语言学、计算机科学、机器学习、数学、认知心理学**等多学科内容，涉及从字、词、短语到句子、段落、篇章的多种语言单位，以及**处理、理解、生成**等不同层面的知识点，研究内容涉及的知识点多且复杂。
- 自20 世纪90 年代以来，自然语言处理发展迅猛，各类任务和算法和研究范式层出不穷，在搜索引擎、医疗、金融、教育、司法等众多领域展示出重要作用。

—— 张奇《自然语言处理导论》

什么是自然语言处理

- **Natural Language Processing (NLP)**
  - 自然语言处理就是利用计算机为工具对人类特有的书面形式和口头形式的自然语言的信息进行各种类型处理和加工的技术。

    —冯志伟《自然语言的计算机处理》
- 其它名称
  - 自然语言理解(Natural Language Understanding)?
  - 计算语言学(CL, Computational Linguistics)
  - 人类语言技术(Human Language Technology)

什么是自然语言处理

- NLP能为我们做什么？
  - 处理邮件
  - 精准翻译
  - 管理、摘要、汇聚知识
  - 语音交互
  - …
- 现实中相关技术达到的水平？

什么是自然语言处理

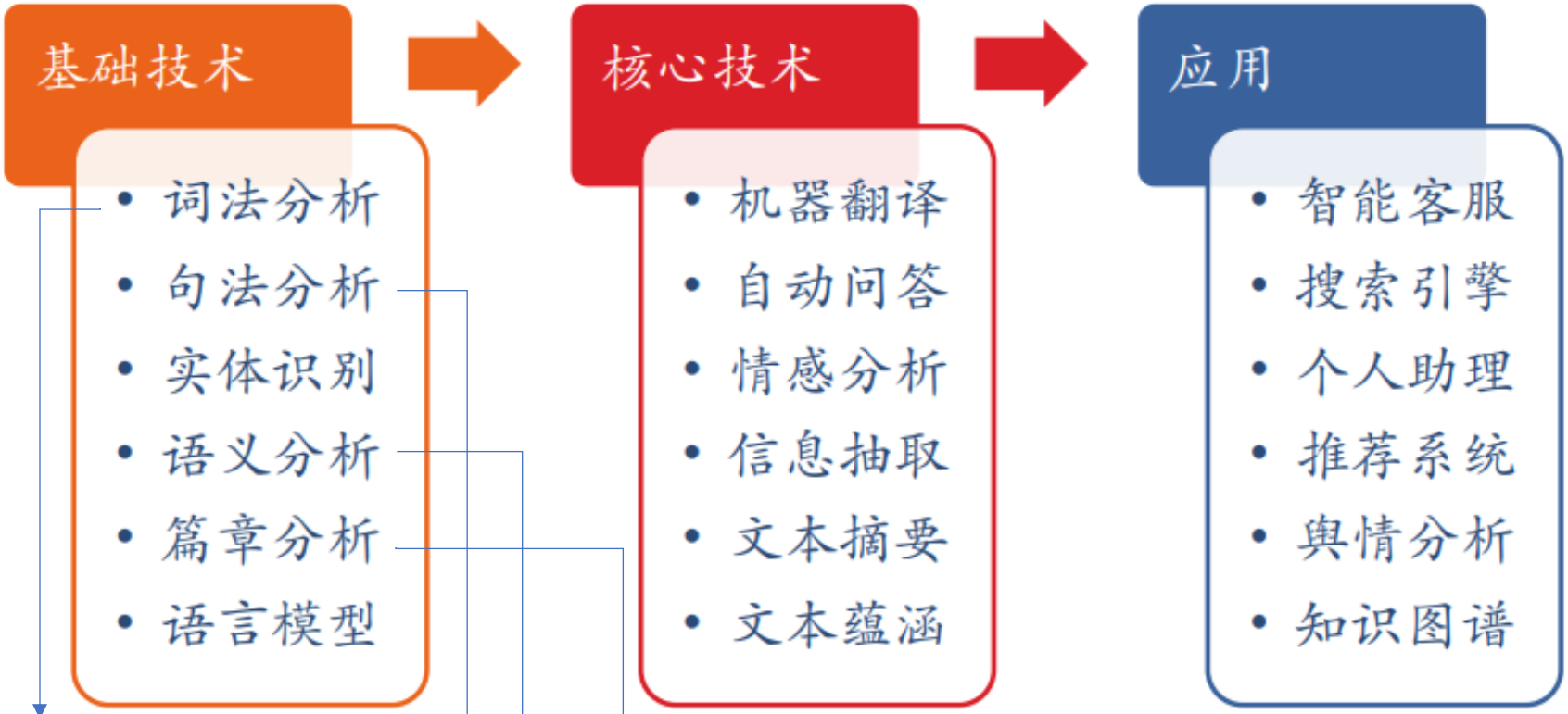
- 相关学术会议/期刊
- ACL (Annual Meeting of the Association for Computational Linguistics)
- NAACL (North American Chapter of the Association for Computational Linguistics)
- EMNLP (Conference on Empirical Methods in Natural Language Processing)
- COLING (International Conference on Computational Linguistics)
- IJCAI (International Joint Conference on Artificial Intelligence)
- AAAI (AAAI Conference on Artificial Intelligence)

- TSLP (ACM Transactions on Speech and Language Processing)
- Computational Linguistics
- TASLP (IEEE Transactions on Audio, Speech, and Language Processing)
- TAC (IEEE Transactions on Affective Computing)

绪论

- 自然语言处理的概念
- 自然语言处理的主要研究内容　　有哪些任务？
- 自然语言处理研究的困难
- 自然语言处理的研究方法

研究内容 任务层次



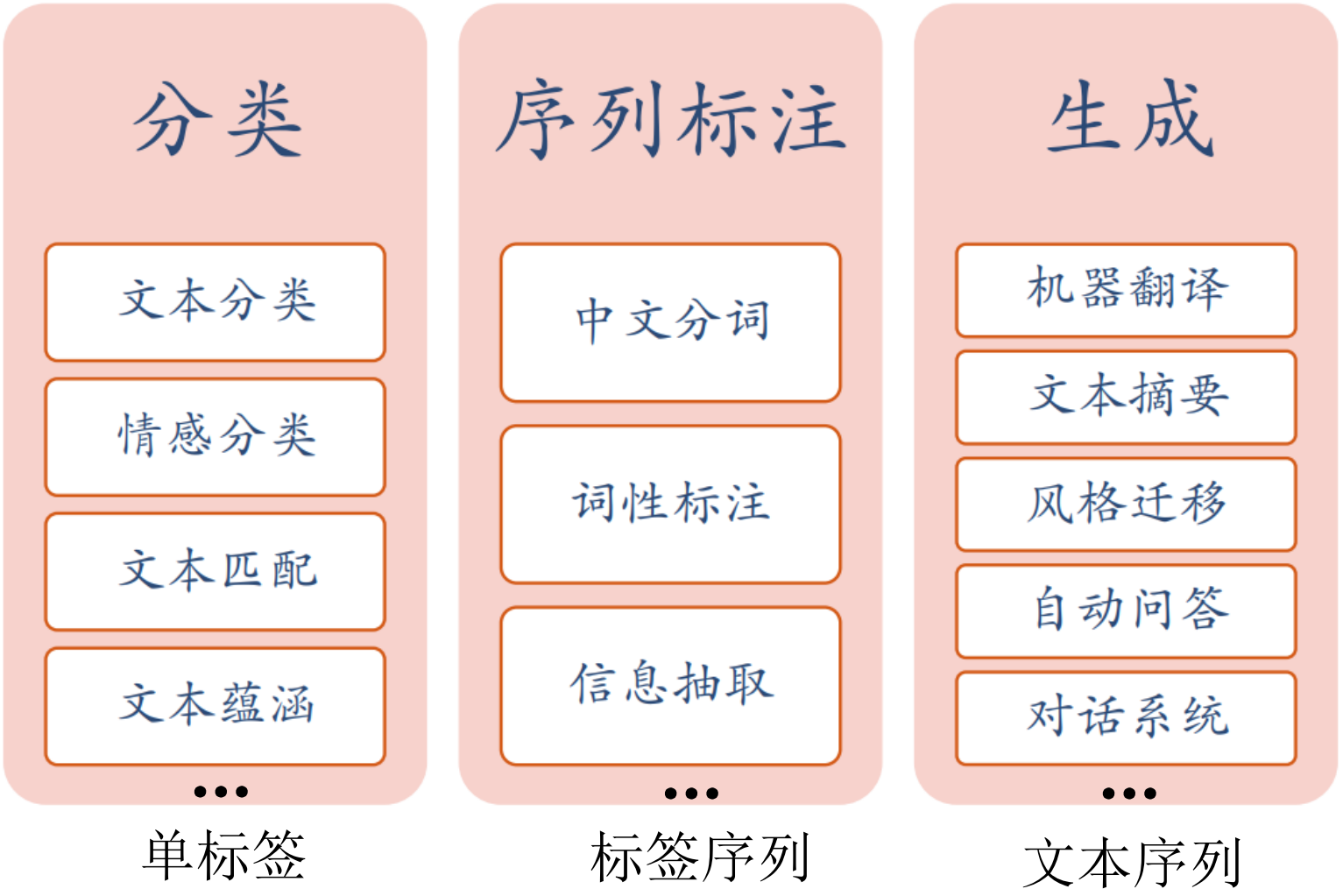
分词、词性标注

句法结构、词汇依存关系

语义消歧、语义角色标注

篇章结构分析：衔接性、连贯性

研究内容 任务形式


分类
文本分类
情感分类
文本匹配
文本蕴涵
...
序列标注
中文分词
词性标注
信息抽取
...
生成
机器翻译
文本摘要
风格迁移
自动问答
对话系统
...
单标签
标签序列
文本序列

基础任务

- 词法分析
  - 分词 **word segmentation**
    - 南京市长江大桥
    - 南京市/长江大桥？南京/市长/江大桥？
    - Even in English, characters other than white-space can be used to separate words [e.g. , . ; : ( )]

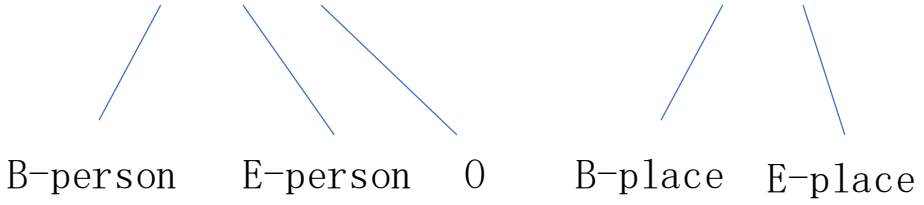
      序列标注任务

      南 京 市 长 江 大 桥
      B I I B I I I

      B:beginning
      I: inside
  - 词性标注 **part-of-speech (POS)**
    - 外商_**NN** 投资_**NN** 企业_**NN** 成为_**VV** 中国_**NR** 外贸_**NN** 重要_**JJ** 增长点_**NN**

基础任务

- 命名实体识别(Name entity recognition, NER)
  - 指识别文本中具有特定意义的实体，主要包括人名、地名、机构名、专有名词等
  - 3大类（实体类，时间类，数字类）
  - 7小类（人名、地名、组织机构名、时间、日期、货币、百分比）
  - 据新华日报，美国总统拜登在2月20日突访基辅，会见泽连斯基。



基础任务

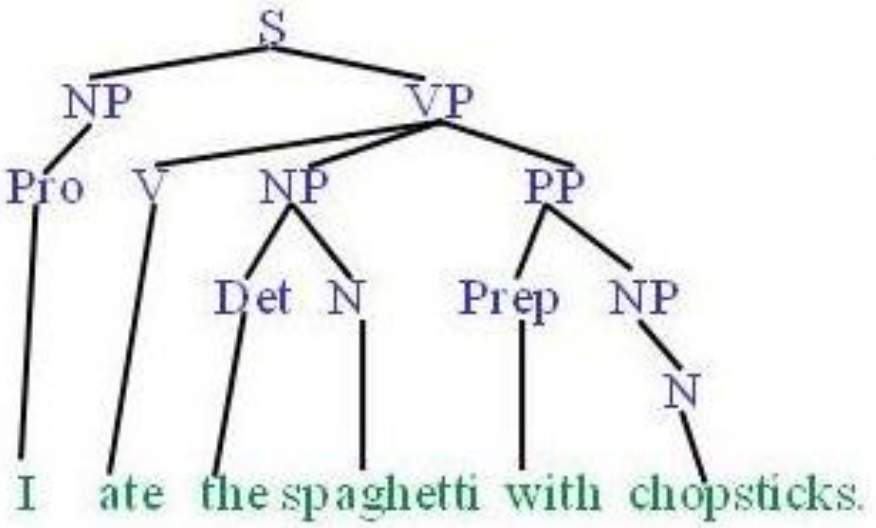
- 句法分析 **(Syntactic Parsing)**

  任务：确定句子的句法结构，或者词汇之间的依存关系

  （1）成分句法分析 constituent parsing

  - 又称句法结构分析、短语结构分析
  - 分析句子的内部结构，如成分构成、上下文关系
  - 为一个句子给出正确的句法分析树**(parsing tree)**
  - 形式化的语法规则

例：I ate the spaghetti with chopsticks.



❖ 完全句法分析：
Full parsing

基础任务

❖局部分析 partial parsing

- **aka 浅层句法分析 (shallow parsing)**
  - 只识别句子中某些结构简单的独立成分
  - **E.g. 句子中的非递归 (non-recursive) 的短语/块 (chunk)，如名词短语NP、动词短语VP等。**

外商_NN 投资_NN 企业_NN 成为_VV 中国_NR 外贸_NN 重要_JJ 增长点_NN

⇩

[NP 外商 投资 企业] [VP 成为] [NP 中国] [NP 外贸] [ADJP 重要] [NP 增长点] 。

基础任务

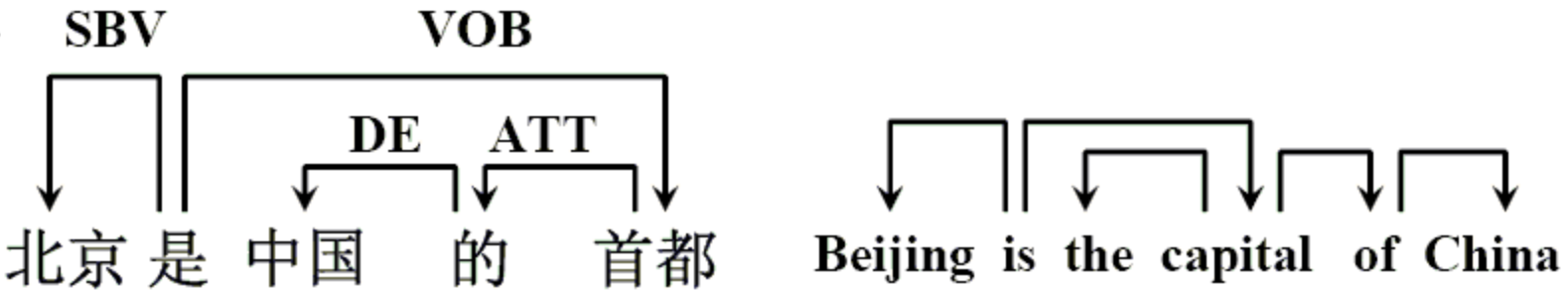
（2）依存句法分析 (Dependency Parsing)

- 现代依存语法理论的创立者是法国语言学家**Lucien Tesnière(1893-1954)**。**L. Tesnière**的思想主要反映在他**1959**年出版的《结构句法基础》。
- 在依存语法理论中，“依存”就是指词与词之间支配与被支配的关系，这种关系不是对等的，而是有方向的。处于支配地位的成分称为支配者**(governor, regent, head)**，而处于被支配地位的成分称为从属者**(dependency, modifier, subordinate)**。
- 又称依存关系分析、依存结构分析、依存分析。

基础任务

- 依存句法分析 (Dependency Parsing)

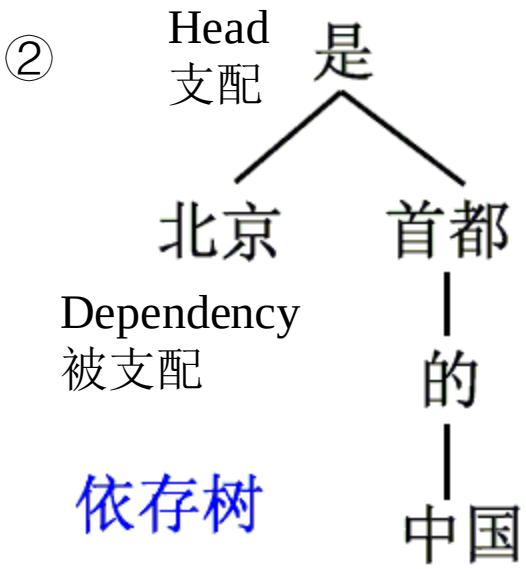
三种常用依存句法结构图示：

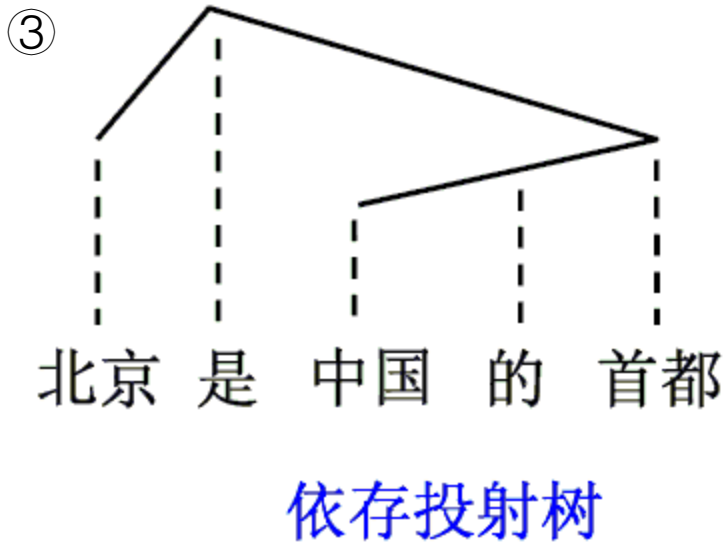
① 

两个有向图用带有方向的弧(或称边，edge)来表示两个成分之间的依存关系，支配者在有向弧的发出端，被支配者在箭头端，我们通常说被支配者依存于支配者。

基础任务

- 依存句法分析 **(Dependency Parsing)**





用树表示的依存结构，树中子节点依存于该节点的父节点。

带有投射线的树结构，实线表示依存联结关系，位置低的成份依存于位置高的成份，虚线为投射线。

基础任务

- 依存结构树的条件：

  1）单纯结点：只有终结结点，没有非终极结点（所有结点都是句子中的单词）

  2）单一父结点

  3）独根结点

  4）非交条件：树枝不能相交

  5）互斥条件：从上到下的支配关系和从左到右的前于关系互斥

冯志伟：判断从属树合格性的五个条件

基础任务

- 语义分析 **(semantic)**
- 词的层面：词义消歧 **(Word Sense Disambiguation, WSD)**
  - 从给定上下文中确定一个多义词的具体意思
    - “讲/故事”（说），“讲/卫生”（注意）
    - bank account，river bank
  - 词义消歧涉及自然语言处理的诸多应用领域，如机器翻译、信息检索、问答系统等

基础任务

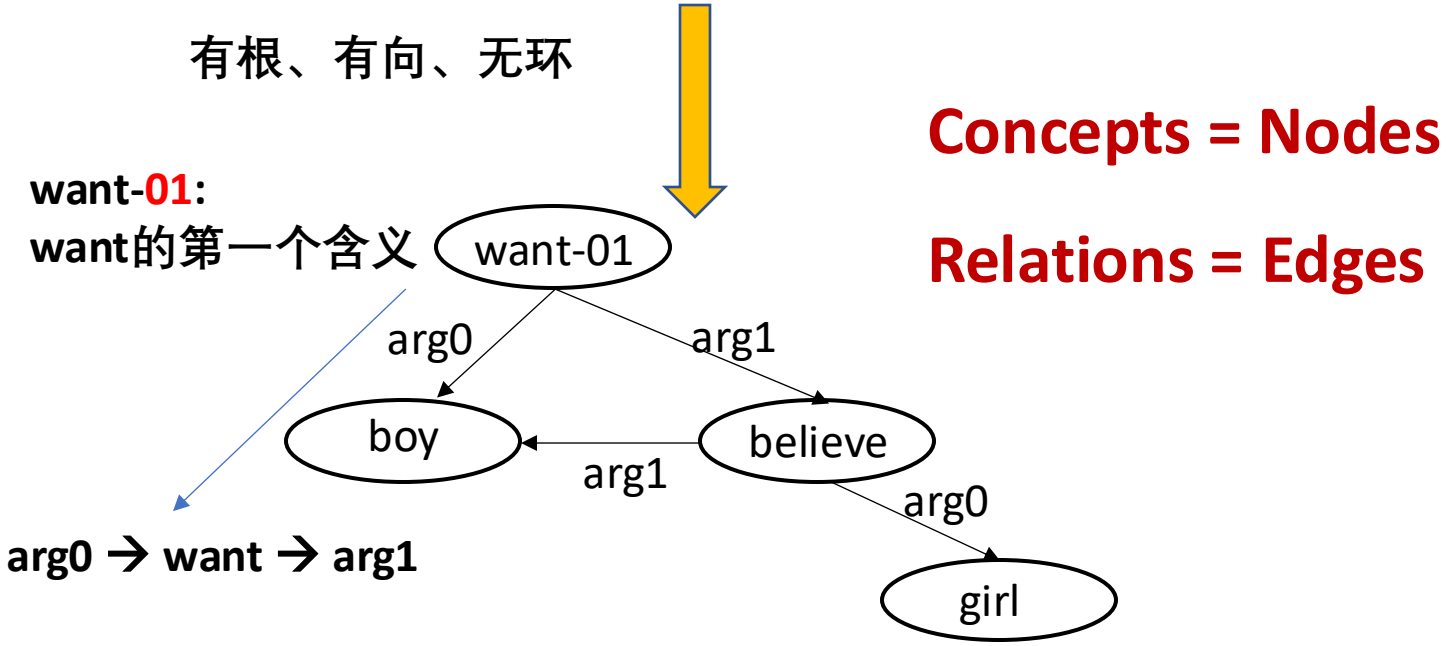
- 句子层面：语义角色标注 **(Semantic Role Labeling, SRL)**
  - 又叫 浅层语义分析**(shallow semantic parsing)**
  - 根据句子的句法结构和句中每个实词的词义，推导出能够反映这个句子意义(即句义)的某种形式化表示.
  - 谓词-论元结构，以谓词为中心
  - **E.g.:**
    - “张三昨天吃了苹果” / “苹果昨天被张三吃了”
    - 表示成语义的形式为:“吃(张三,苹果, 昨天)”
    - 吃：谓词，张三：施事，苹果：受事，昨天：发生时间

基础任务

- 语义解析（**Semantic Parsing**）
  - 把自然语言转化为机器可读可执行的逻辑语句
  - **E.g. SQL（Text-to-SQL）, python, Prolog**
  - 有哪些学生选了课？→ **select DISTINCT Sno from SC**
- 语义解析的实际应用领域：
  - **CLang: Robocup Coach Language**
  - **Geoquery: A Database Query Application**

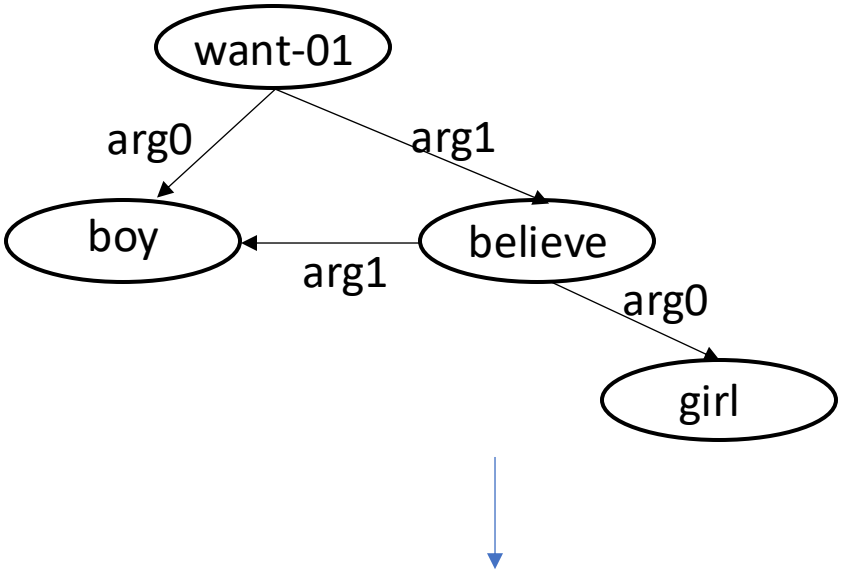
基础任务

- 抽象语义表示 **(AMR, Abstract Meaning Representation)**
- **AMR 捕捉“谁在对谁做什么”。每个句子都表示为一个有根、有向、无环的图，在边(语义关系 semantic relation) 和叶(concept) 都有标签**
- E.g.: The boy wants the girl to believe him.



基础任务

- 抽象语义表示 **(AMR)**



instance(w, want-01)将want实例化为w

```
(w / want-01
   :ARG0 (b / boy)
   :ARG1 (b2 / believe-01
             :ARG0 (g / girl)
             :ARG1 b))
```

The boy wants the girl to believe him/himself

The boy wants to believed by the girl

Goal: supporting research in:

natural language generation

summarization

machine translation

……

基础任务

- 抽象语义表示 **(AMR)**
- AMR **concepts** are either English words (“boy”), PropBank framesets (“want-01”), or special keywords.
- AMR uses approximately 100 **relations**:
  - Frame arguments, following PropBank conventions.
    :arg0, :arg1, :arg2, :arg3, :arg4, :arg5.
  - General semantic relations.
    :age, :beneficiary, :cause, :concession ……
  - Relations for quantities
    :quant, :unit, :scale
  - Relations for date-entities.
    :day, :month, :year, :weekday
  - Relations for lists.
    :op1, :op2, :op3, :op4, :op5 .....

The data are the mean ± SEM of three independent experiments.

```
(e / equal-01
      :ARG1 (d / data)
      :ARG2 (d2 / distribution-range-91
            :ARG6 (s / standard-error-of-the-mean)
            :poss (e2 / experiment-01
                  :ARG0-of (d3 / depend-01
                        :polarity -)
                  :quant 3)))
```

基础任务

- 篇章分析
  - 篇章discourse：有组织、层级性的句子序列整体
  - 分析文章的内部结构和内部关系，主要包括衔接性和连贯性：

(1) 衔接性（外部联结）：整个篇章内词汇/短语之间的关联。当语篇中一个成分的含义依赖于另一个成分的解释时，便产生了衔接关系。

衔接：指代、替换、省略（都可以看成指代）、连接、词汇衔接

√ 共指消解/指代消解（Coreference Resolution）

- 实体共指消解：苹果起诉高通公司，状告其未按照合约进行合作
- 事件共指消解：他购买了一台电脑，由单位承担这笔交易

√ 词汇衔接：重复、泛指词、相似性、可分类性和搭配。目标是构成词汇链，即一个主题下的一系列词义相关的词共同组成的词网。

未上升到以此实现语篇衔接性的分析

(2) 连贯性（内部联结）：句子/句群之间的语义关系

基础任务

- 语言模型 (language model, LM)
  - 计算一个句子的概率的模型
  - $S = w_1, w_2, \ldots, w_n$

→ $P(S)= P(w_1\ w_2\ \ldots w_n) = P(w_1)\ P(w_2|\ w_1)\ P(w_3|\ w_1w_2)\ \ldots P(w_n|w_1\ w_2\ \ldots w_{n-1})$

e.g. S = 我是一个大学生

$$\frac{count(\text{我是一个大学生})}{\text{count(我是一个大学)}}$$

→ 计算量巨大！

n元语法（n-gram）：只考虑n个词/字

应用任务

- 文本分类 (text classification)
  - 也叫文档分类（Document categorization），其目的就是利用计算机系统对文档按照一定的分类标准实现自动归类。
  - 输入：$s=\{x_1, x_2, \ldots, x_n\}$，输出：y
  - 应用：图书管理、内容管理、信息监控等。

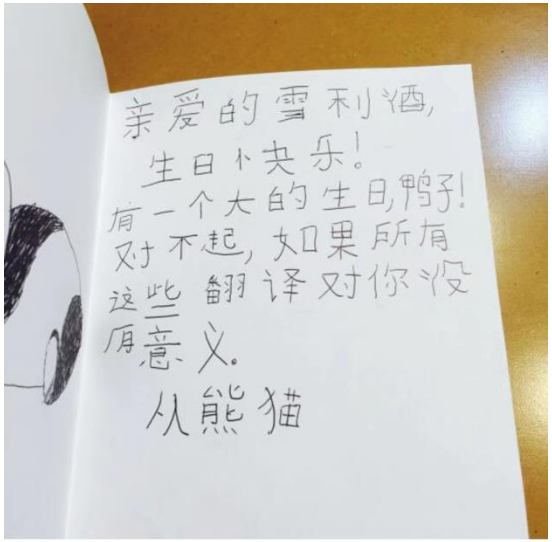
文本蕴含(text entailment)：判断后一句话（假设句hypothesis）能否从前一句话（前提句premise）中推断出来

输入：s_1+s_2，输出：$y \in$ {蕴含，矛盾} → 句子对的文档分类

目前研究热点：新的分类场景

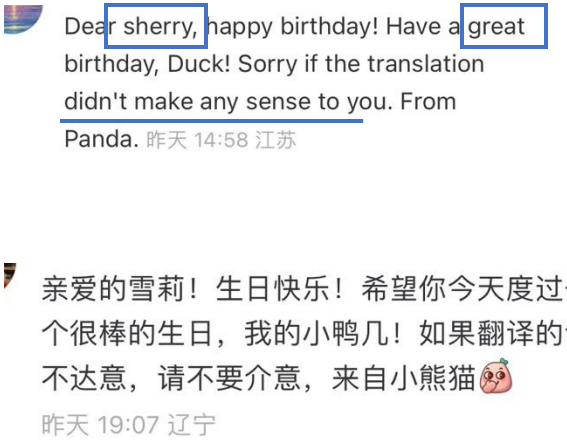
应用任务

- **机器翻译 (Machine Translation)**
  - 一种语言（source）到另一种语言（target）的自动翻译
  - 应用：文献翻译、网页辅助浏览



美国同学送我的生日贺卡

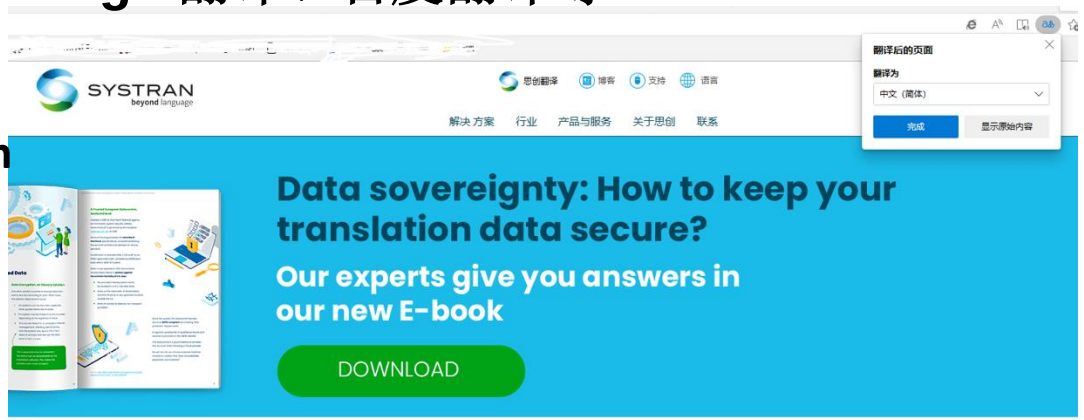
今天翻出了六年前收到的生日贺卡，是当时高中的好朋友（白人妹子）送给我的，可以看出来很努力在写中文了...



应用任务

- **机器翻译 (Machine Translation)**
  - 代表系统：Google翻译、百度翻译等

Systran在线翻译平台：
http://www.systransoft.com



无法翻译图片
存在翻译失败问题
翻译质量

- ❖ 后续重点讲解

应用任务

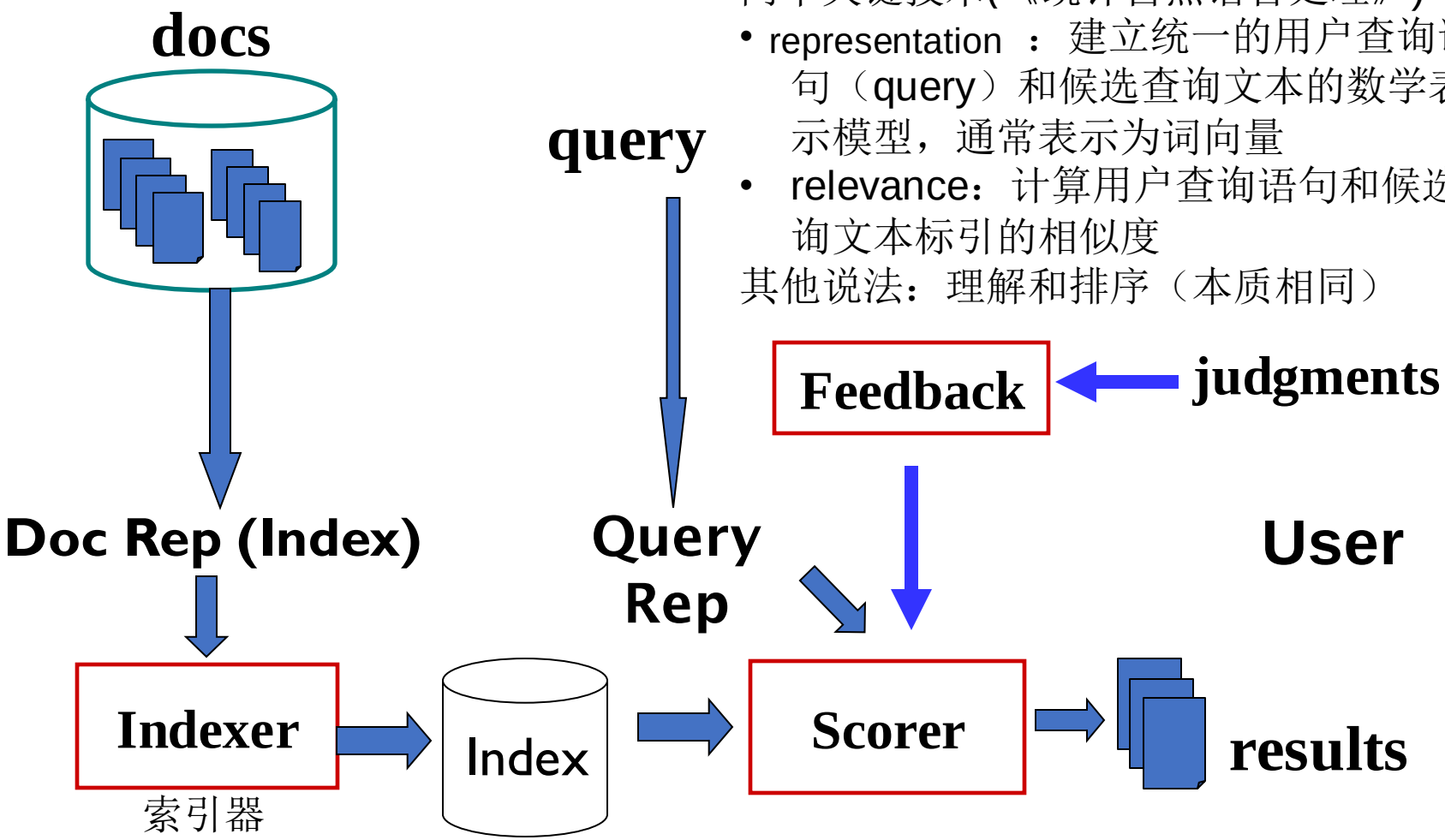
- 文本摘要 (text summarization)
  - aka 自动文摘
  - 将原文档的主要内容或某方面的信息自动提取出来，并形成原文档的摘要或缩写。
  - 应用：电子图书管理、情报获取等。
  - 抽取式（1.直接抽取句子2.抽取并简化句子构成文摘）vs 生成式/理解式（更难）
  - 单文档摘要 vs 多文档摘要
  - 单语言摘要 vs 跨语言（cross-lingual）摘要
  - 难点：正确性，重要性，简洁性，流利性

应用任务

- 信息检索 **(Information Retrieval, IR)**
  - 信息检索也称文本检索，就是利用计算机系统从大量文档中找到符合用户需要的相关信息→有用的文档
  - Text Retrieval is defined as the matching of some stated user query against useful parts of free-text records.
  - 起源：图书馆的参考咨询和文摘索引
  - 精确匹配模型（主要用于内部文本库） **vs** 文档相关匹配模型（主要用于网络搜索）
  - 应用：搜索引擎；邮件搜索；电脑文件搜索；法律知识检索…

应用任务

- IR系统的一般模式



两个关键技术(《统计自然语言处理》)

- representation ： 建立统一的用户查询语句（query）和候选查询文本的数学表示模型，通常表示为词向量
- relevance：计算用户查询语句和候选查询文本标引的相似度

其他说法：理解和排序（本质相同）

indexing：建立倒排索引的过程(《信息检索导论》)