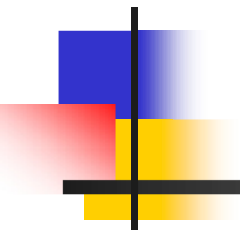
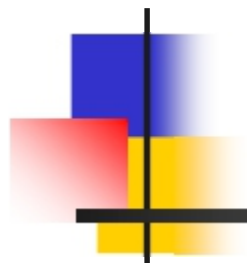


# 第4章 语料库与 语言知识库

---





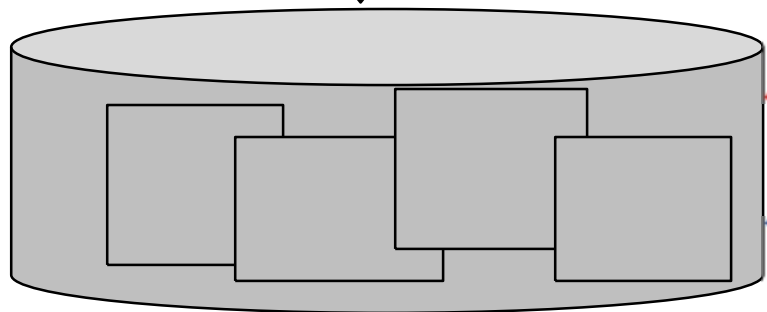
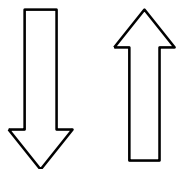
# 4.1 基本概念

## 4.1 基本概念

输入

输出

处理模块



语言数据库或知识库

大规模语言数据:

- 模型参数训练
- 评测标准

NLP中知识库包括:

- 词汇语义库
- 词法、句法规则库
- 常识库等等



## 4.1 基本概念

---

### ◆语料库(corpus)

指经科学取样和加工的大规模电子文本库，其中存放的是实际使用中真实出现过的语言材料。

### 命名实体识别

---

- [CCKS2017](#) 面向中文电子病历的医疗实体识别及属性抽取数据集
- [CCKS2018](#) 面向中文电子病历的医疗实体识别及属性抽取数据集
- [CCKS2019 数据下载](#) 面向中文电子病历的医疗实体识别及属性抽取数据集

### QA

---

- [CCIR2019](#) CCIR 2019 基于电子病历的数据查询类问答
- [cMedQA](#) 中文医学QA数据集
- [cMedQA2](#) 中文医学QA数据集
- [CMID](#) 中文医学QA意图理解数据集
- [KGQA](#) 基于医药知识图谱的智能问答系统



---

## 4.2 语料库的类型



## 4.2 语料库的类型

---

### ◆ 按语言种类划分

- 单语的
- 双语的或多语的

篇章对齐 / 句子对齐 / 结构对齐

### ◆ 是否标注？

- 具有词性标注
- 句法结构信息标注(树库)
- 语义信息标注



## 4.2 语料库的类型

### ◆ 平行语料库

平行语料库是指在两种或多种语言之间的平行采样和加工，例如，机器翻译中的双语对齐语料库

C: 早晨好！

E: **Good morning.**

C: 您能给我一杯咖啡吗？

E: **Could you give me a cup of coffee?**

... ..

C: 早晨<sup>1</sup> 好<sup>2</sup> !<sup>3</sup>

E: **Good<sup>2</sup> morning<sup>1</sup> .<sup>3</sup>**



## 4.3 典型语料库介绍





## 4.3 典型语料库介绍

---

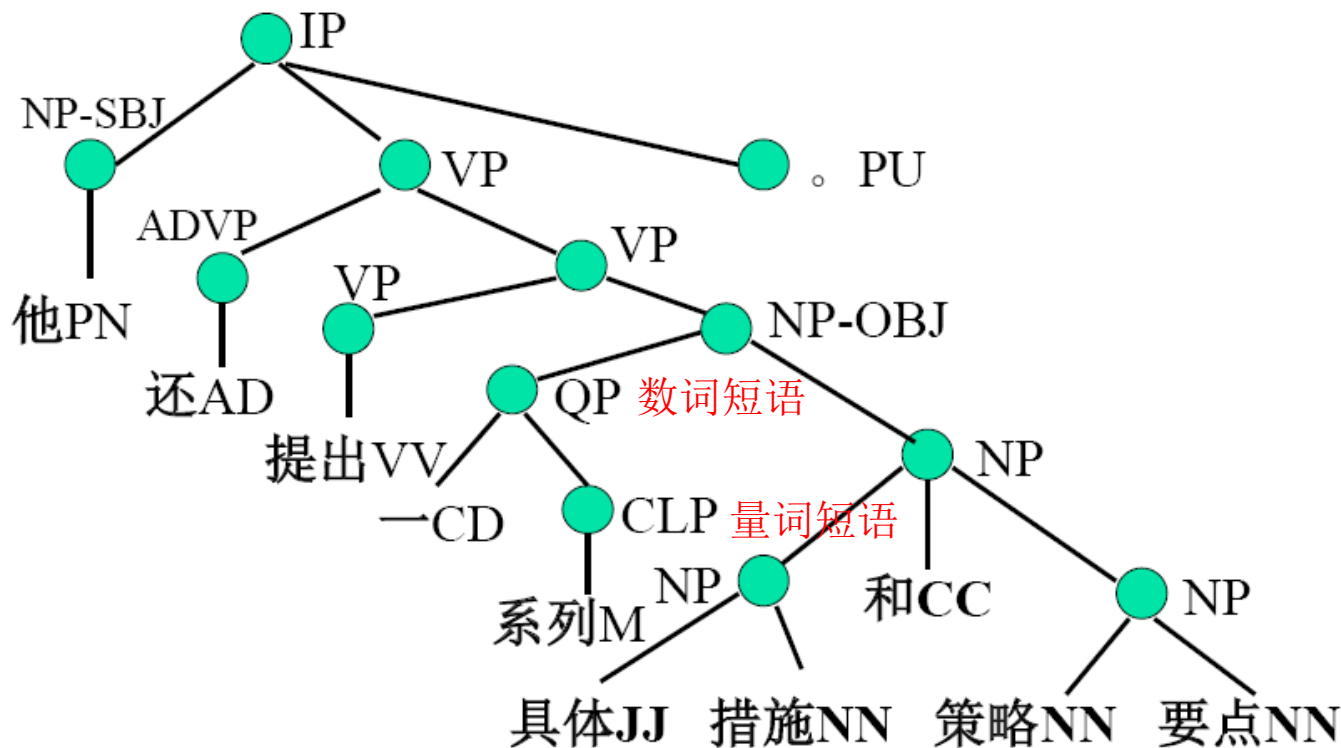
### ◆ 宾夕法尼亚大学 (UPenn) 树库(Tree Bank)

- 美国宾夕法尼亚大学计算机系 M. Marcus 教授主持
- 1993年完成约300万词次英语句子的语法结构标注
- 2000年完成第一版汉语树库，约10万词次，4185个句子
- **Chinese Tree Bank (CTB)** 中汉语词性被划分为33类，23类句法标记(Syntactic tags)

## 4.3 典型语料库介绍

◆ 例句：他还提出一系列具体措施的政策要点。

词性标注：他/PN 还/AD 提出/VV 一/CD 系列/M 具体/JJ 措施/NN 和/CC 政策/NN 要点/NN 。/PU





## 4.3 典型语料库介绍

### ◆ 北京大学开发的CLKB

- 现代汉语语法信息词典：8万词、360万语法属性描述
- 汉语短语结构规则库：600多条语法规则
- 现代汉语多级加工语料库：实现词语切分并标注词类的基本标注语料库1.5亿字
- 多语言概念词典：10万个以同义词集表示的概念
- 平行语料库：含对译的英汉句对100万
- 多领域术语库：有35万汉英对照术语



## 4.3 典型语料库介绍

---

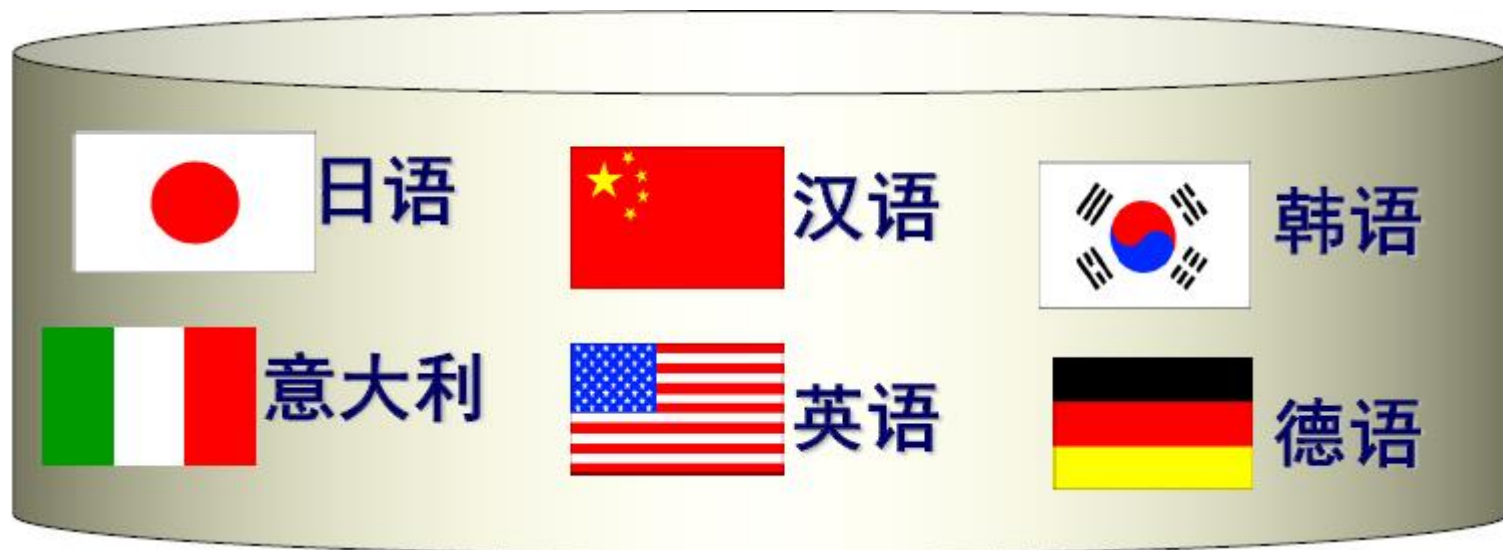
多级加工语料样例：

咱们/**r** 中国/**ns** 这么/**r** 大/**a** 的/**u** 一个/**m** 多/**a**  
民族/**n** 的/**u** 国家/**n** 如果/**c** 不/**d** 团结/**a** , /**w**  
就/**d** 不/**d** 可能/**v** 发展/**v** 经济/**n** , /**w** 人民/**n**  
生活/**n** 水平/**n** 也/**d** 就/**d** 不/**d** 可能/**v** 得到/**v**  
改善/**vn** 和/**c** 提高/**vn** 。 /**w**

## 4.3 典型语料库介绍

### ◆ 口语语料库: BTEC (Basic Traveler's Expression Corpus)

目标是开展语音翻译的国际合作研究, 开发实用的语音翻译技术



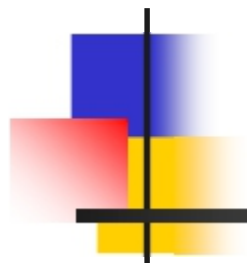
## 4.3 典型语料库介绍

### ◆CASIA-CASSIL 语音对话语料库

- 选自**15000**余段汉语电话(语音)对话录音
- 每段平均不少于**90**秒、**10**个回合(turns), 如:

场景	旅馆	餐馆	机场	全部	平均
对话个数	<b>206</b>	<b>263</b>	<b>323</b>	<b>792</b>	--
回合个数	<b>3,676</b>	<b>4,389</b>	<b>4,993</b>	<b>13,058</b>	<b>16.5</b>
话语个数	<b>7,352</b>	<b>8,778</b>	<b>9,986</b>	<b>26,116</b>	<b>33.0</b>
字数	<b>78,950</b>	<b>85,491</b>	<b>110,135</b>	<b>274,576</b>	<b>10.5</b>
词数	<b>57,800</b>	<b>44,112</b>	<b>78,368</b>	<b>180,280</b>	<b>6.9</b>

- 基于文字的对话语料



## 4.4 词汇知识库



## 4.4 词汇知识库

---

### ◆ WordNet (<http://wordnet.princeton.edu/>)

- 普林斯顿大学(Princeton University) 认知科学实验室 George A. Miller 教授领导开发。
- 开发目的: 解决词典中同义信息的组织问题
- 目前规模: 95600 英语词条, 其中, 51500 个简单词, 44100 个搭配词。70100 个词义(同义词集合)。
- 五大类词汇: 名词、动词、形容词、副词、虚词。





## 4.4 词汇知识库

---

- 特色：根据词义（而不是词形）组织词汇信息，从某种意义上讲，它是一部语义词典。
- WordNet 按语义关系组织：语义关系看作是同义词集合之间的一些指针，语义关系是双向的。

有了该知识库，我们才能回答诸如下列问题：

- Java有哪些含义？
- 当Java作为“咖啡”时，有哪些同义词？



## 4.4 词汇知识库

---

### ➤ 4 种语义关系:

- 同义关系(**synonymy**)
- 反义关系(**antonymy**)
- 上下位关系(**hypernymy/ hyponym**)或称从属/上属关系：如：{枫树}是{树}的下位，{树}是{植物}的下位。
- 部分关系(**meronymy**)或称部分/整体关系。

## 4.4 词汇知识库

➤使用wordnet的基本功能 <http://wordnetweb.princeton.edu/perl/webwn>

Word to search for:

Display Options:

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (frequency) {offset} <lexical filename> [lexical file number]  
(gloss) "an example sentence"

Display options for word: word#sense number (sense key)

### Noun

- (2){08928021} <noun.location>[15] [S:](#) (n) **Java#1 (java%1:15:00::)** (an island in Indonesia to the south of Borneo; one of the world's most densely populated regions)
- (1){07945759} <noun.food>[13] [S:](#) (n) **coffee#1 (coffee%1:13:00::)**, **java#2 (java%1:13:00::)** (a beverage consisting of an infusion of ground coffee beans) *"he ordered a cup of coffee"*
- {06913829} <noun.communication>[10] [S:](#) (n) **Java#3 (java%1:10:00::)** (a platform-independent object-oriented programming language)
  - [direct hypernym](#) / [inherited hypernym](#) / [sister term](#)
    - {06913460} <noun.communication>[10] [S:](#) (n) [object-oriented programming language#1 \(object-oriented\\_programming\\_language%1:10:00::\)](#), [object-oriented programming language#1 \(object-oriented\\_programing\\_language%1:10:00::\)](#) ((computer science) a programming language that enables the programmer to associate a set of procedures with each type of data structure) *"C++ is an object-oriented programming language that is an extension of C"*



## 4.4 词汇知识库

### ➤ WordNet 的应用

词汇消歧，语义推理，理解等。

如：机器翻译时，必须先明确Java的词义，才能准确翻译对应的中文。

例如：食堂 没 地方，我 在 饭馆 吃 了 蛋 炒饭。

“地方”的三种含义：

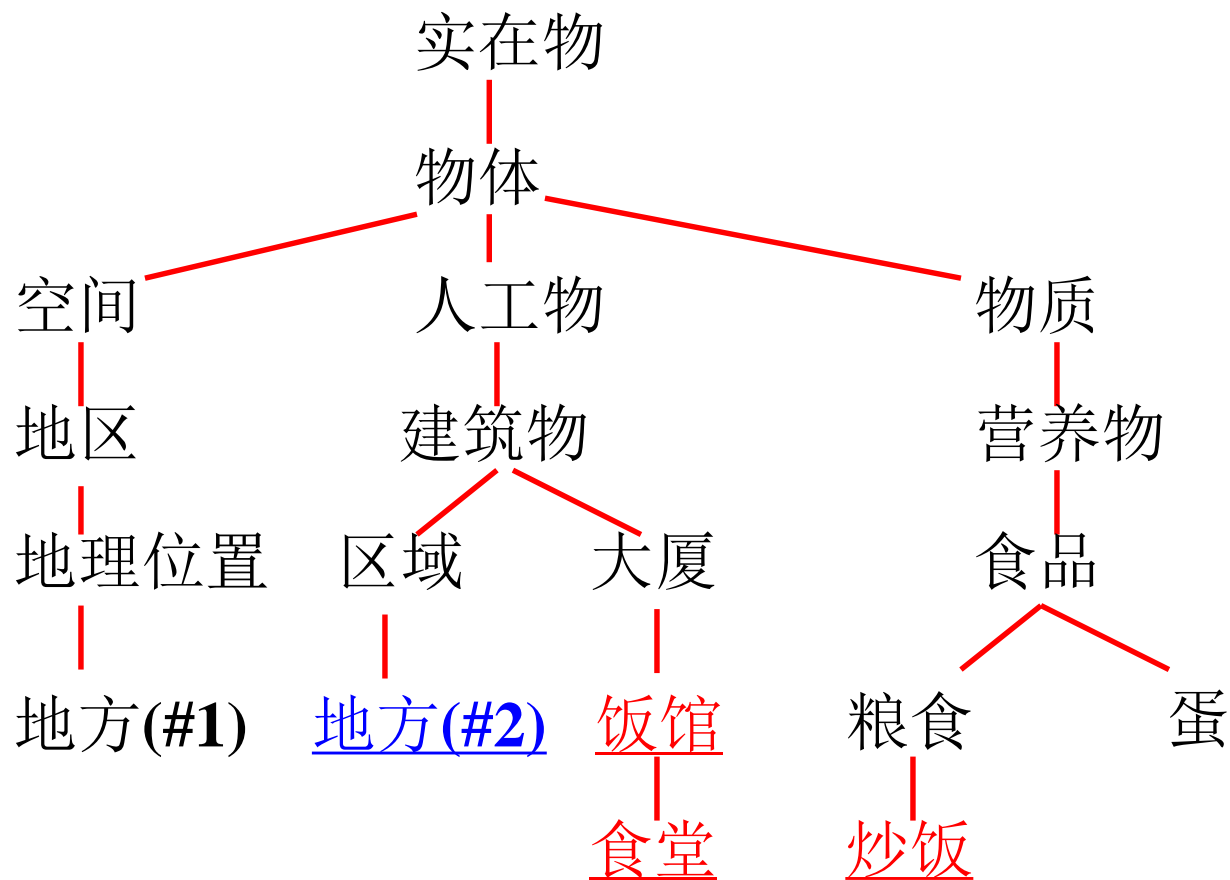
# 指地理位置      如：在祖国各个地方

# 指空间          如：没地方

# 指部分          如：他说的有些地方不对

## 4.4 词汇知识库

三个含义在两棵不同的名词语义树上，其中一个树的部分：



## 4.4 词汇知识库

### ➤使用wordnet的基本功能

使用NLTK之WordNet 接口

```
>>> dog = wn.synset('dog.n.01')  
>>> dog.hypernyms()          #上位词集合  
[Synset('domestic_animal.n.01'), Synset('canine.n.02')]  
>>> dog.root_hypernyms()     #得一个最一般的上位（或根上位）同义词集  
[Synset('entity.n.01')]  
>>> dog.hyponyms()           #下位词集合  
[Synset('puppy.n.01'), Synset('great_pyrenees.n.01'),  
Synset('basenji.n.01'), Synset('newfoundland.n.01'),  
Synset('lapdog.n.01'), Synset('poodle.n.01'),
```

应用场景：基于语义相似性，自动扩展检索词，如“水杯”



## 4.4 词汇知识库

### ➤使用wordnet的基本功能

使用NLTK之WordNet 接口

```
>>> dog = wn.synset('dog.n.01')↵  
>>> cat = wn.synset('cat.n.01')↵
```

```
>>> dog.path_similarity(cat)↵  
0.20000000000000001↵  
>>> cat.path_similarity(cat)↵  
1.0↵
```



## 4.4 词汇知识库

---

◆ 知网(**HowNet**) (<http://www.keenage.com>)

➤ 1988年由董振东教授提出：

- (1) **NLP**系统最终需要更强大的知识库的支持。
- (2) 知识是一个系统，是一个包含着各种概念与概念之间的关系，以及概念的属性与属性之间的关系的系统。





## 4.4 词汇知识库

---

◆ 知网描述了下列各种关系：

- (a) 上下位关系 (由概念的主要特征体现)
- (b) 同义关系
- (c) 反义关系
- (d) 对义关系
- (e) 部件-整体关系
- (f) 属性-宿主关系
- (g) 材料-成品关系



## 4.4 词汇知识库

### ◆ 知网描述了下列各种关系：

(h) 施事/经验者/关系主体-事件关系（由在事件前标注 \* 体现，如“医生”，“雇主”等）

(i) 受事/内容/领属物等-事件关系（由在事件前标注 \$ 体现，如“患者”，“雇员”等）

(j) 工具-事件关系（由在事件前标注 \* 体现，如“手表”，“计算机”等）

(k) 场所-事件关系（由在事件前标注 @ 体现，如“银行”，“医院”等）

(l) 时间-事件关系（由在事件前标注 @ 体现，如“假日”，“孕期”等）



## 4.4 词汇知识库

---

◆ 知网描述了下列各种关系：

(m) 值-属性关系（直接标注无须借助标识符，如“蓝”，“慢”等）

(n) 实体-值关系（直接标注无须借助标识符，如“矮子”，“傻瓜”等）

(o) 事件-角色关系（由加角色名体现，如“购物”，“盗墓”等）

(p) 相关关系（由在相关概念前标注 # 体现，如“谷物”，“煤田”等）



## 4.4 词汇知识库

---

◆词语例子：

**NO.=000001**

**W\_C=打**

**G\_C=V**

**E\_C=~酱油，~张票，~饭，去~瓶酒，醋~来了**

**W\_E=buy**

英语

**G\_E=V**

**E\_E=**

**DEF=buy|买**

概念定义



## 4.4 词汇知识库

---

**NO.=015492**

**W\_C=打**

**G\_C=V**

**E\_C=**~毛衣，~毛裤，~双毛袜子，~草鞋，~一条围  
巾，~麻绳，~条辫子

**W\_E=knit**

**G\_E=V**

**E\_E=**

**DEF=weave|**辫编



---

# Thanks

谢谢!