自然语言理解

第四章 语料库与词汇知识库

宗成庆

中科院自动化研究所

模式识别国家重点实验室

cqzong@nlpr.ia.ac.cn

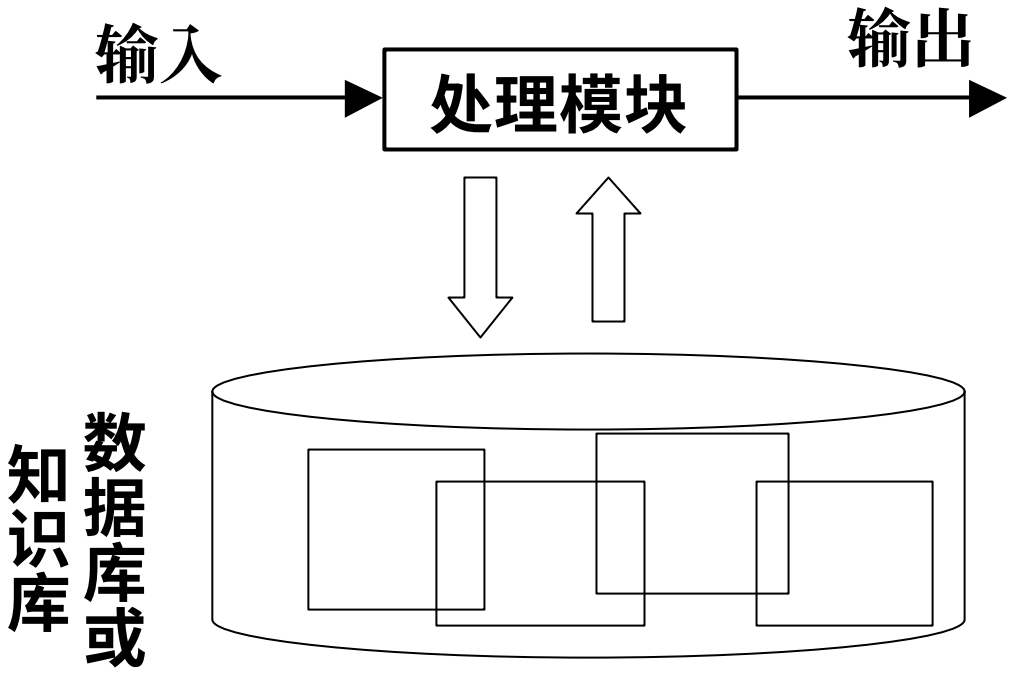
http://www.nlpr.ia.ac.cn/English/cip/cqzong.htm

No.95, Zhongguancun East Road
Beijing 100080, China

http://www.ia.ac.cn
Tel. No.: +86-10-6255 4263

第四章 语料库与词汇知识库

4.1 概述



数据库或知识库的用途：

- 用于机器学习（训练）
- 用于实际处理

NLP中知识库包括：

- 词汇知识库（词典）
- 规则库
- 常识库等

4.2 语料库语言学

❑ 概述

语料库（corpus）就是存放语言材料的仓库（数据库）。

语料库语言学（corpus linguistics）就是基于语料库进行语言学研究的一门学问。

两种解释：

- ❖ 不是新的术语：利用语料库对语言的某个方面研究。
- ❖ 是新的术语：对现行语言学理论批评，提出新理论。

一般指前者

4.2 语料库语言学

“语料库语言学已经成为语言研究的主流。基于语料库的研究不再是计算机专家的独有领域，它正在对语言研究的许多领域产生愈来愈大的影响。”

—J. Thomas 等人为祝贺语料库语言学的主要奠基人和倡导者 G. Leech 六十岁生日而出版的语料库语言学研究论文集的开场白。

—[丁信善，1998]

4.2 语料库语言学

- 语料库语言学的定义
  - 根据篇章材料对语言的研究称为语料库语言学。

    - [Aijmer, 1991]
  - 基于现实生活中语言运用的实例进行的语言研究称为语料库语言学。

    - [McEnery, 1996]
  - 以语料为语言描写的起点或以语料为验证有关语言的假说的方法称为语料库语言学。

    -[Crystal, 1991]

4.2 语料库语言学

- 语料库语言学研究的内容：
  - 语料库的建设与编纂
  - 语料库的加工和管理技术
  - 语言研究中语料库的使用
  - 语料库语言学在计算语言学中的应用

4.3 语料库技术的发展

- 语料库语言学的发展
  - ～ 20世纪50年代中期：早期的语料库语言学
    - 语料库在语言研究中被广泛使用：语言习得、方言学、语言教学、句法和语义、音系研究

4.3 语料库技术的发展

❖ **1957～20世纪80年代初期：沉寂时期**

- 1957年Chomsky 的《句法理论》及其以后一系列著作的发表，根本改变了语料库语言学的发展状况。
- Chomsky 及其转换生成语法学派批判早期的语料库研究方法：
  - 基于语料库的研究方法有误
  - 语料的不充分性

4.3 语料库技术的发展

- **20世纪80年代～：复苏与发展时期**
  - **第二代语料库相继建成：**
    - 1983**年英国**Lancaster**大学建成** Lancaster-Oslo / Bergen Corpus (LOB**语料库**): **研究英国英语，**500**语篇，每个语篇约**2000**词。**
    - **法国国家科学研究中心与美国芝加哥大学联合建成法语语料库（**Tremor de la Language Francaise, TLF**语料库）：**2000**书面法语文本，**1.5**亿词。**
    - **芬兰赫尔辛基大学建成历史英语语料库（**The Helsinki Corpus of Historical English）:850-1720**年**, 1600**万词。**

4.3 语料库技术的发展

- 1988年伦敦大学建成国际英语语料库（The International Corpus of English, ICE）：语料来自所有英语国家，各100万词，1990－1993年，口语和书面语各一半，18岁以上接受英语教育的成人。

➢ **基于语料库的研究项目增多**

1981年至1991年的11年时间里，大约有480个语料研究项目得到资助，而在1959年至1980年20多年的时间里，只有140个基于语料的研究项目。

4.3 语料库技术的发展

- 语料库技术复苏的原因

1）计算机的迅速发展；

2）转换生成语言学派对语料库语言学的批判不都正确（如指责计算机分析语料是伪技术），有的是片面的甚至是错误的（如对语料数据价值的否定）。

4.4 国内语料库研究状况

- 汉语现代文学作品语料库（1979年，武汉大学，527万字）
- 现代汉语语料库（1983年，北京航空航天大学，2000万字）
- 中学语文教材语料库（1983年，北京师范大学，106万字）
- 现代汉语词频统计语料库（1983年，北京语言学院，182万字）

4.4 国内语料库研究状况

- 1991年，中国国家语言文字工作委员会开始建立国家级大型汉语语料库，以推进汉语的词法、句法、语义和语用研究，其计划规模将达7000万汉字。
- 北京大学计算语言学研究所从1992年开始现代汉语语料库的多级加工，在语料库建设方面成绩卓著，先后建成2600万字的1998年《人民日报》标注语料库、2000万字汉字，1000多万英语单词的篇章级英汉对照双语语料库、以及8000万字篇章级信息科学与技术领域的语料库等。
- 清华大学于1998年建立了1亿汉字的语料库，着重研究歧义切分问题。

4.4 国内语料库研究状况

- 山西大学、哈尔滨工业大学、北京语言文化大学、东北大学、中科院自动化研究所和香港城市大学、台湾中央研究院等相当一批大学和研究机构都对汉语语料库的建设做出了重要贡献。
- 新疆大学、新疆师范大学、内蒙古大学、中国社科院民族研究所和西北民族大学等院所研究和开发我国少数民族语言的语料库。

4.5 语料库的类型

- 按内容构成和目的划分
  - 异质的（heterogeneous）　　—[黄昌宁，2002]

    最简单的语料收集方法，没有事先规定和选材原则
  - 同质的（homogeneous）

    与“异质”正好相反，比如美国的 TIPSTER 项目只收集军事方面的文本。

4.5 语料库的类型

◆ 系统的（Systematic）

充分考虑语料的动态和静态问题、代表性和平衡问题以及语料库的规模等问题。

◆ 专用的（specialized）

如：北美的人文科学语料库

4.5 语料库的类型

- 按语言种类划分
  - 单语的
    - （已切分的）具有词性标注
    - 句法结构信息标注（树库）
    - 语义信息标注
  - 双语的或多语的
    - 篇章对齐 / 句子对齐 / 结构对齐

两个术语：生语料，熟语料

4.5 语料库的类型

- 平衡语料库与平行语料库
  - 平衡语料库
    - 平衡语料库着重考虑的是语料的代表性与平衡性。
    - 语料采集的七项原则：语料的真实性、语料的可靠性、语料的科学性、语料的代表性、语料的权威性、语料的分布性和语料的流通性。其中，语料的分布性还要考虑语料的科学领域分布、地域分布、时间分布和语体分布等。[张普，2003]
    - 问题：（1）各个分布点所选取的语料量的科学依据是什么？（2）使用度是否已经完全真实地反映了语言的使用情况？

4.5 语料库的类型

❖ 平行语料库

➢两种含义，一种是指在同一种语言的语料上的平行，例如正在建立的“国际英语语料库”，共有20个平行的子语料库，分别来自以英语为母语或官方语言和主要语言的国家，如英国、美国、加拿大、澳大利亚、新西兰等。其平行性表现为语料选取的时间、对象、比例、文本数、文本长度等几乎是一致的。建库的目的是对不同国家的英语进行对比研究。

➢另一种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

4.5 语料库的类型

- 共时语料库与历时语料库

  - 所谓共时语料库是为了对语言进行共时研究而建立的语料库。研究大树的横断面所见的细胞和细胞关系，即研究一个共时平面中的元素与元素的关系。

  - 所谓的历时语料库是为了对语言进行历时研究而建立的语料库。研究大树的纵剖面所见的每个细胞和细胞关系的演变，即研究一个历时切面中元素与元素关系的演化。

4.5 语料库的类型

- 判断历时语料库的4条原则　　—[张普，2003]
  - 是否动态语料库：语料库必须是开放的、动态的。
  - 语料库的文本是否具有量化的流通度属性：所有的语料都应来源于大众传媒，都具有采用不同计算方法的与传媒特色相应的流通度属性。其量化的属性值也是动态的。
  - 语料库的深加工是否基于动态的加工方法：随着语料的动态采集，语料也应进行动态地加工。
  - 是否取得动态的加工结果：语料的加工结果也应是动态的和历时的。

4.6 语料库建设中的问题

- 语料库设计需要考虑的问题
  - 静态与动态

    语料库建设的另一种主张是动态的，或监督语料库（monitor corpus）：动态文本集，数据的收集通常是随遇的，而不是平衡的
  - 代表性和平衡

    一个语料库具有代表性，是指在该语料库上获得的分析结果可以概括成为这种语言整体或其指定部分的特性。

    —[Leech, 1991]

    如何达到不同部分之间的平衡？

4.6 语料库建设中的问题

- **规模**

  第一代语料库100**万词次**

  1990s 1000－2000 **万词次小型的一般语料库**

  一般而言，在保证质量的前提下应足够大。

- **语料库的管理与维护**
  - **错误修正或改善**
  - **版本升级**
  - **语料库的检索系统、分析和处理工具的维护等**

4.6 语料库建设中的问题

- 汉语语料库开发中存在的问题
  - 语料库建设的规范问题
    - 信息处理用GB13000.1 字符集汉字部件规范1997.12.5 国家语委；
    - GB12200.1-90 汉语信息处理词汇01部分：基本术语 国家技术监督局（1993）；
    - GB/T12200.2-94 汉语信息处理词汇02部分：汉语和汉字国家技术监督局（1994）；
    - GB13715 信息处理用现代汉语分词规范。

4.6 语料库建设中的问题

问题：

- 分词标准已经确定和统一？
- 词类标记集被普遍采用和遵循？
- 文本属性规范在哪里？
- … …

4.6 语料库建设中的问题

- 产权保护和国家语料库建设问题
  - 汉语语料库的知识产权包括两个方面：文本的知识产权和语料库的知识产权及其衍生产品。
  - 语料库的知识产权却没有得到保护，至今在著作权法、语言文字法、计算机软件保护等相关法规和实施条例中语料库的知识产权都是空白。

4.6 语料库建设中的问题

国家语料库的建设、开发、保护应该是一种国家行为，在信息社会和数字化生存时代，我们要把语言资源的收集、保护、开发提高到一种对待国家资源的高度来认识。国家要像对待人力资源、地矿资源、国土资源、森林资源、水源资源一样对待语言资源，语言资源是国家最重要的信息资源。语料库的建设、保护、开发要站在国家面向未来的一种战略决策高度，要作为一种对待国家资源的行为，才能得到法律的保护，纳入法制的轨道[张普，2003]。

4.7 典型语料库介绍

❑ 布朗语料库 (Brown Corpus)

- ➢ 20世纪60s，Francis 和 Kucera 在布朗（Brown）大学建立，是世界上第一个根据系统性原则采集样本的标准语料库
- ➢ 100万词规模
- ➢ 选自1961年美国人撰写出版的普通语体的文本
- ➢ 15种题材，共500个样本，每个样本不少于2000词
- ➢ 1961年布朗大学出版了当代英语词频词典
- ➢ 1970s Greene 和 Rubin 设计了TAGGIT词性标注系统（词类标记81种，上下文约束规则3300条），自动标注正确率77%。

4.7 典型语料库介绍

LLC口语语料库 (London-Lund Corpus of Spoken English)

- 1960s 伦敦大学著名语言学家Quirk组织
- 2000小时的对话和广播等口语素材
- 瑞典隆德（Lund）大学教授 Svartvik 主持录入计算机
- 英语口语调查（The Survey of Spoken English, SSE）
- SSE 于1981年完成，建成 London-Lund Corpus of Spoken English (LLC)
- 87个文本，每个文本约5000词，最终规模50万词
- 5大类：面对面交谈，电话交谈，讨论、采访、辩论，未经准备的当众评论、论证、演讲，经准备的当众演讲
- 标注：语调、节律、关键词（语段），词类、出现次数、搭配关系等

4.7 典型语料库介绍

朗文语料库 (Longman Corpus)

- 朗文语料库委员会（Longman Corpus Committee）
- January 1981- November 1990
- 设计原则：1）尊重本族语言者的直觉和语料库权威
  2）向研究人员提供语料（英国50%，美国40%，其它国家10%）
  3）书面语
- 选自1900～的20世纪英语：知识性（informative）文本60%，想象性（imaginative）文本40%
- 10个分布广泛的领域：自然和纯科学、应用科学、社会科学、世界事务等
- 2800 万词

4.7 典型语料库介绍

❑宾州(Pennsylvania)大学语料库 (UPenn Tree Bank) (http://www.cis.upenn.edu/~treebank/home.html)

- 美国宾州大学计算机系 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年完成第一版中文树库，约10万词次，4185个句子

例子：原始句子：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

4.7 典型语料库介绍

例子：(IP (NP-SBJ (PN 他))
(VP (ADVP (AD 还))
(VP (VV 提出))
(NP-OBJ (QP (CD 一)
(CLP (M 系列)))
(NP (NP (ADJP (JJ 具体)
(NP (NN 措施)))
(CC 和)
(NP (NN 政策)
(NN 要点))))))
(PU 。))

4.7 典型语料库介绍

- **北京大学语料库（ http://icl.pku.edu.cn/ ）**
  - **北大计算语言学研究所俞士汶教授主持，北大、富士通、人民日报社共同开发**
  - **《人民日报》1998年全部文本（约2600万字）**
  - **完整的词语切分和词性标注信息**

例子：

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

4.7 典型语料库介绍

❑ 台湾中研院平衡语料库

(http://rocling.iis.sinica.edu.tw/ROCLING/corpus98/)

- **台湾中央研究院平衡语料库（Sinica Corpus）：世界上第一个带有完整词类标记的汉语平衡语料库**
- **目标：500万词次汉语平衡语料库**
- **设计思想：1) 遵循台湾计算语言学会的分词标准**
  2) 采样时以自然段落为准，不看文章长度
  3) 语料采用多重分类法

4.7 典型语料库介绍

- Chinese LDC
  - 国家 973 项目资助（图象、语音、自然语言理解与知识挖掘，编号：G1998030504）
  - 语音，文字（口语，书面语）
  - 单语：分词及词性标注语料，树库语料
  - 双语：汉英句子对齐
  - 规模： 汉语通用词表 ：8－10万词
    汉语信息词典：2.5-3.0 万词
    分词词性标注语料：500万字
    汉语句法树库：100万字 … …

4.7 典型语料库介绍

- LC-STAR 项目 (NLPR-Nokia)
  - 12 国语言：加泰罗尼亚语、芬兰语、德语、希腊语、希伯来语、意大利语、汉语、俄语、西班牙语、标准阿拉伯语、土耳其语和美式英语 。
  - 目的：口语翻译
  - 规模：文本语料不少于100M words （中文不少于2500万汉字）

4.7 典型语料库介绍

- **领域：新闻** 612**万字，**19%、

  财经418**万字，**14%、
  文化娱乐 374**万字，**12%
  体育829**万字，**27%
  消费 499**万字，**16%
  个人通讯 355**万字，**12%
  共计约：3087 **万字**

4.7 典型语料库介绍

- **抽取词汇：通用词汇：**38142**个**

  专用名词(proper names)**：**

  人名：22,156**个**

  地点名：19,930**个**

  组织机构名：15,618**个**

  专用词汇：7521**个**

- **标注词性、拼音等**

4.8 词汇知识库

- **WordNet (http://wordnet.princeton.edu/)**
  - **美国普林斯顿大学**(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。（实际上** WordNet **中近包含前**4**类）**

4.8 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。如果词义{x1, x2, …}和{y1, y2, …}之间有一种语义关系R，则在{y1, y2, …}和{x1, x2, …} 之间也有语义关系R。属于这两个同义词集合的单词之间的关系也是R。

4.8 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）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。

4.8 词汇知识库

- 名词的25个独立起始概念：

{动作，行为，行动}、{自然物}、{动物，动物系}、{自然现象}、{人工物}、{人，人类}、{属性，特征}、{植物，植物系}、{身体，躯体}、{所有物}、{认知，知识}、{作用，方法}、{信息，通信}、{量，数量}、{事件}、{关系}、{直觉，情感}、{形状}、{食物}、{状态，情形}、{团体，组织}、{物质}、{场所，位置}、{时间}、{目的}

4.8 词汇知识库

- 21000个动词词形、大约8400个词义，14个文件：

照顾动词，功能动词，变化动词，认知动词，通信动词，竞争动词，消费动词，接触动词，创作动词，感情动词，运动动词，感觉动词，占用动词，社会交往动词，天气变化动词。

- 19500个形容词词形，近10000个词义

描述性形容词，参照修饰形容词，颜色形容词，关系形容词。

4.8 词汇知识库

- WordNet 的应用

  词汇消歧，语义推理，理解等。

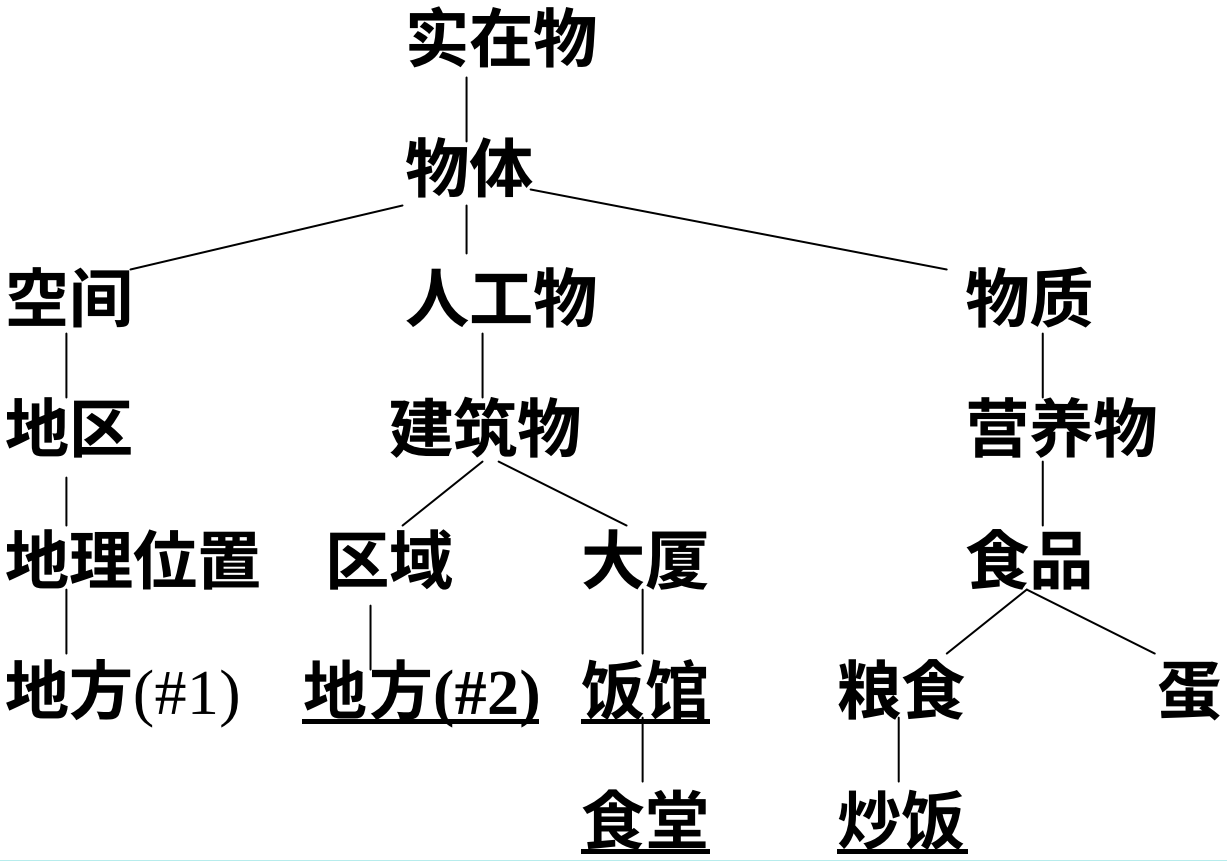
例如：食堂 没 地方，我 在 饭馆 吃 了 蛋 炒饭。

“地方”的三种意思：

#指地理位置	如：在祖国各个地方
#指空间	如：没地方
#指部分	如：他说的有对的地方

4.8 词汇知识库

3个含义在两棵不同的名词集成语义树上，其中一个树的部分：



4.8 词汇知识库

❑ 知网(HowNet)

(http://www.keenage.com/html/c_index.html)

- 1988年由董振东教授提出，基本观点：

(1) 自然语言处理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。一个人比另外一个人有更多的知识说到底是他不仅掌握了更多的概念，尤其重要的是他掌握了更多的概念之间的关系以及概念的属性与属性之间的关系。

4.8 词汇知识库

（3）关于如何建立知识库，他提出应首先建立一种可以被称为知识系统的常识性知识库。它以通用的概念为描述对象，建立并描述这些概念之间的关系。

（4）首先应由知识工程师来设计知识库的框架，并建立常识性知识库的原型。在此基础上再向专业性知识库延伸和发展。专业性知识库或称百科性知识库主要靠专业人员来完成。这里很类似于通用的词典由语言工作者编纂，百科全书则是由各专业的专家编写。

4.8 词汇知识库

❑ 知网的哲学

知网哲学的根本点是：世界上一切事物（物质的和精神的）都在特定的时间和空间内不停地运动和变化。它们通常是从一种状态变化到另一种状态，并通常由其属性值的改变来体现。试以人为例，人的生老病死是一生的主要状态。这个人的年龄（属性）一年比一年大｛属性值｝，随着年龄的增长头发的颜色（属性）变为灰白｛属性值｝。另一方面，一个人随着年龄的增长他的性格（精神）变得日益成熟｛属性值｝，他的知识（精神产品）愈益丰富｛属性值｝。基于上述，知网的运算和描述的基本单位是：万物，其中包括物质的和精神的两类，部件，属性，时间，空间，属性值以及事件。

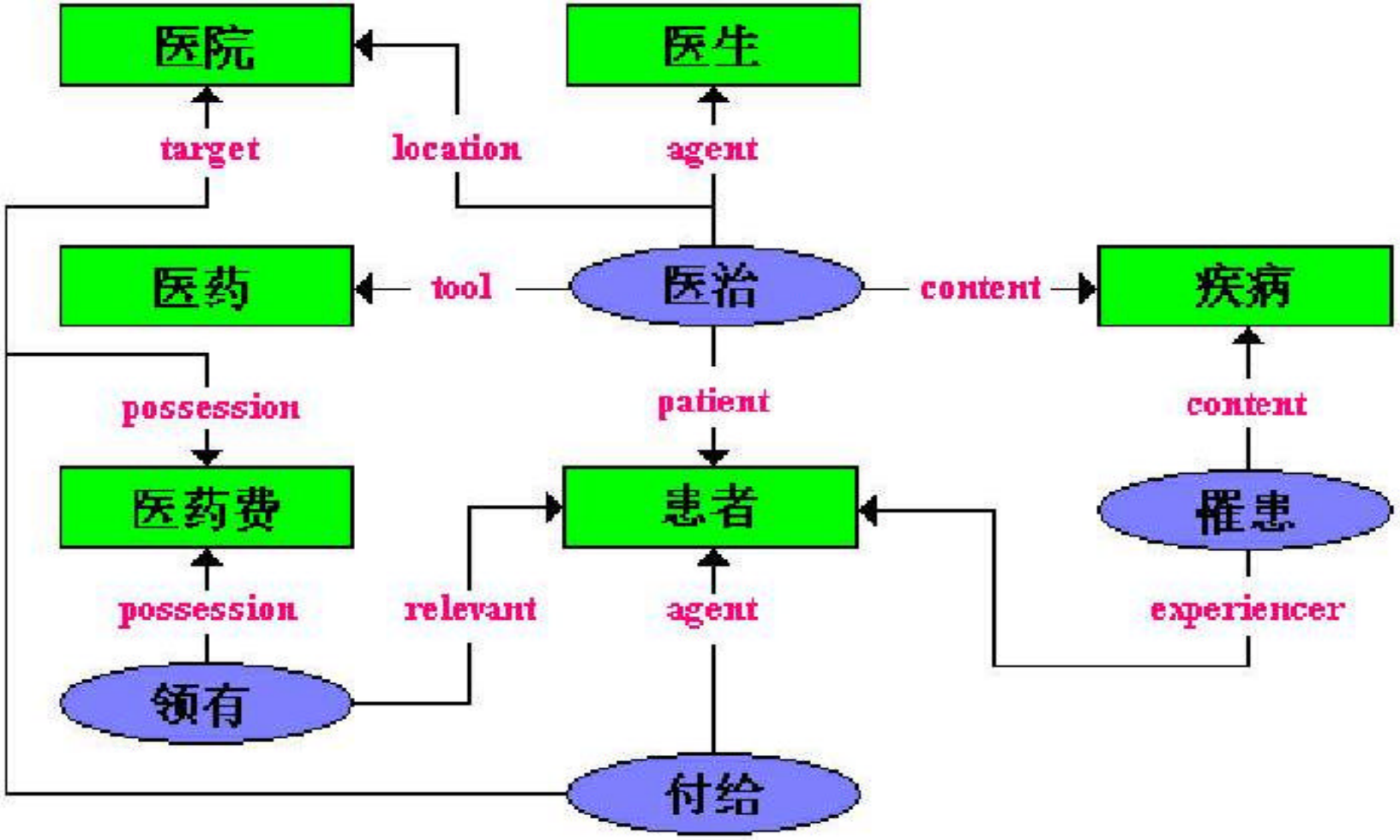
4.8 词汇知识库

❑ 知网的特色

知网作为一个知识系统，名副其实是一个网而不是树。它所着力要反映的是概念的共性和个性，例如：对于“医生”和“患者”，“人”是它们的共性。

同时知网还着力要反映概念之间和概念的属性之间的各种关系。

4.8 词汇知识库


医院
医生
target
location
agent
医药
tool
医治
content
疾病
possession
patient
content
医药费
患者
罹患
possession
relevant
agent
experiencer
领有
付给

4.8 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

4.8 词汇知识库

(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)
(m) 值-属性关系（直接标注无须借助标识符，如"蓝"，"慢"等)

4.8 词汇知识库

(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）

4.8 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
G_E=V
E_E=
DEF=buy|买

4.8 词汇知识库

❑词语例子：

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

本章小结

- 语料库语言学的基本定义、研究内容和发展历程
- 语料库类型
- 语料库建设中的基本问题
- 典型语料库
- WordNet
- 知网

习题

1. 思考一下，如果让你评价一个语料库，并给出定量的分值，你将如何建立评分方法？
2. 查阅或通过网页下载有关北京大学语料库和宾州大学语料库（UPenn Tree Bank）的文献资料，了解语料库的设计、加工过程。

Thanks

谢谢!