第五章 语料库与词汇知识库

王峰
华东师大计算机系

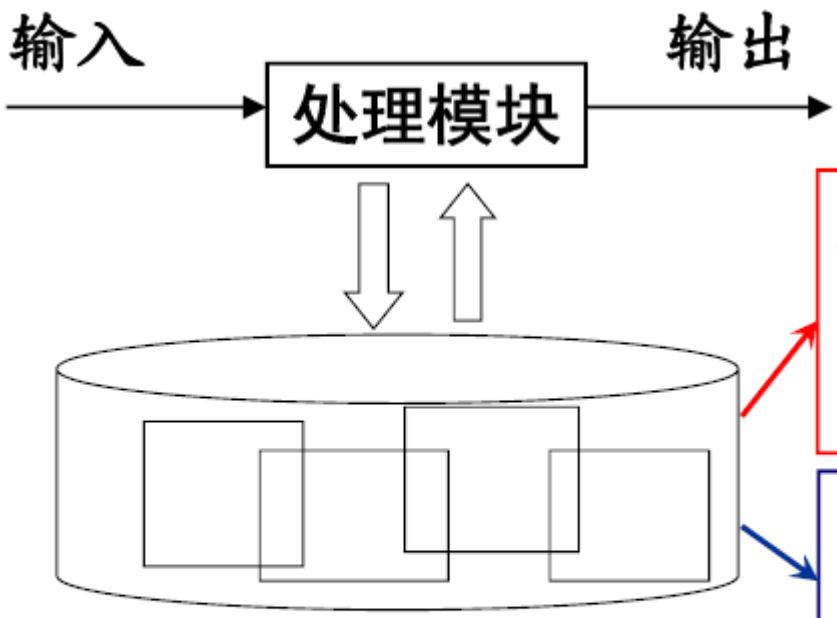
提 纲

- 语料库
  - 基本概念
  - 语料库技术的发展
  - 语料库的类型
  - 典型语料库介绍
  - 语料库的设计
  - 语料库应用
- 词汇知识库
  - Wordnet
  - Hownet

对语言知识的认识

- 代数学（理性主义）的定义方法
  - 确定性定义方法
  - 语言是由规则所定义的句子的集合
- 统计学（经验主义）的定义方法
  - 不确定性定义方法
  - 语言就是一个概率分布，又称为语言模型
  - 语言中的每一个句子都有自己的出现概率

基本概念



大规模语言数据：

- 模型参数训练
- 知识获取

NLP中知识库包括：

- 词汇语义库
- 词法、句法规则库
- 常识库等等

语料库(corpus)

- 语料库就是存放语言材料的仓库(语言数据库)。
  - 语料库中存放的是在语言的实际使用中**真实**出现过的语言材料；
  - 语料库是以电子**计算机为载体**承载语言知识的基础资源；
  - 真实语料需要**经过加工**（分析和处理），才能成为有用的资源。

语料库示例 1

北京大学计算语言所富士通人民日报标注语料库样例：

- 历史/n 将/d 铭记/v 这个/r 坐标/n :/w 北纬/b 41.1/m 度/q 、/w 东经/b 114.3/m 度/q ;/w 人们/n 将/d 铭记/v 这/r 一/m 时刻/n :/w 1998年/t 1月/t 10日/t 11时/t 50分/t 。/w

- [中国/ns 政府/n]nt 顺利/ad 恢复/v 对/p 香港/ns 行使/v 主权/n ，/w 并/c 按照/p “/w 一国两制/j ”/w 、/w “/w 港人治港/l ”/w 、/w 高度/d 自治/v 的/u 方针/n 保持/v 香港/ns 的/u 繁荣/an 稳定/an 。/w

语料库示例 2

- London-Lund英语口语语料库样例

^what a_bout a cigar\ette# .
((4 sylls))
I ^w\on't have one th/anks# - - -
^aren't you .going to sit d/own# -
^[/\m]# -
^have my _coffee in p=eace# - - -
^quite a nice .room to !s\it in ((actually))#
^\isn't it#
^y/\es# - - -

标记	含义
#	语调群的结束(end of tone group)
^	语音开始(onset)
/	上升型核心语调(rising nuclear tone)
\	下降型核心语调(falling nuclear tone)
^	先升后降型核心语调(rise-fall nuclear tone)
_	平型核心语调(level nuclear tone)
[ ]	不完整的词语和音节符号(enclose partial words and phonetic symbols)
.	标准重音(normal stress)
!	高音高于前一个音节的重音(booster: higher pitch than preceding prominent syllable)
=	高音跟前一个音节相当的重音(booster: continuance)
(( ))	不清晰的音节(unclear)
* *	同步发音(simultaneous speech)
-	一个重音单位的停顿(pause of one stress unit)

Tony McEnery & Andrew Wilson, 1996, Corpus Linguistics, p55

语料库 vs. 语言知识库

- 语料库
  - 以语言的真实材料为基础呈现语言知识
- 语言知识库
  - 专家从大量的实例中提炼、概括出来的系统的语言知识
  - 电子词典、句法规则库、词法分析规则库

语料库语言学 (Corpus Linguistics)

- 基于语料库进行语言学研究
  - 利用语料库对语言的某个方面进行研究，或者发现某些规律性知识
  - 基于语料库进行语言学研究新的理论

- 语料库语言学
  - 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]

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
  - 语料库的使用

- “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]

语料库技术的发展

- **20**世纪**50**年代中期之前：早期
  - 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、音系研究等
- **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：例句数目不够，不包含错误的/不礼貌的句子
  - Brown语料库：60-70年代
    - TAGGIT词性标注系统

语料库技术的发展

- 20世纪80年代以后：复苏与发展时期
- 特征之一：第二代语料库相继建成：
  - 法国国家科学研究中心与美国芝加哥大学联合建成法语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。
  - 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，口语和书面语各半，18 岁以上接受英语教育的成人。
- 特征之二：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。

语料库技术的发展

- 语料库技术复苏的原因
  - 计算机的迅速发展；
  - 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）。

语料库的类型

- 按内容构成和目的划分
  - 异质的 (heterogeneous)
    - 最简单的语料收集方法，没有事先规定选材原则。
  - 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。
  - 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。

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

- 生语料
- 熟语料

语料库的类型

- 平衡语料库
  - 平衡语料库着重考虑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、可靠性、科学性、代表性、权威性、分布性和流通性。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。
- 问题：
  - 各个分布点所选取的语料量的科学依据是什么？
  - 使用度是否真实地反映了语言的使用情况？

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，“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。

语料库的类型

- 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

  - **C:** 您能给我一杯咖啡吗？
  - **E: Could you give me a cup of coffee?**

  - **C:** 早晨1好2！3
  - **E: Good2 morning1.3**

语料库的类型

- 共时语料库与历时语料库
  - 共时语料库是为了对语言进行共时(同一 时段)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。
  - 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。

典型语料库介绍

第一代语料库

- Brown语料库
  - 1960年代初，美国Brown大学，100万词次， 当代美国英语，根据系统性原则采样
- LOB语料库
  - 1970年代初，英国Lancaster大学，挪威Oslo大学，挪威Bergen大学，当代英国英语
- LLC语料库
  - 1960年代初，由London大学Randolph Quirk主 持，收集2000小时的谈话和广播等口语素材并 整理成书面材料，由瑞典Lund大学J. Svartvik 主持全部录入计算机，1975年建成

- 百万词级
- 以语言研究为导向

第二代语料库

- COBUILD语料库
  - 建于1980年代，由英国Birmingham大学与Collins出版社合作完成，规模达2000万词次，基于该语料库出版的CollinsCobuild词典（1987）受到了广泛的好评
- Longman语料库
  - 建于1980年代，包括三个语料库：LLELC语料库（Longman/Lancaster英语语料库）, LSC语料库（Longman口语语料库）, LCLE（Longman英语学习语料库）。目标是编撰英语学习词典，为外国人学习英语服务，词典规模达5000万词次。

■ 千万词级
■ 词典编纂 — 应用导向

第三代语料库

- ACL/DCI语料库
- UPenn树库
  - 美国计算语言学会倡议发起“数据采集计划”（Data Collection Initiative），由宾州大学M. Liberman主持，保存语料原始文本形式以及SGML标注信息。

- LDC
- 超大规模（上亿词级）
- 标准编码体系
- 深度标注/多语种
- NLP应用

宾州大学树库 (Upenn Treebank)

- 美国Pennsylvania大学1980年代末开始发起
- 由该校计算机系M.Marcus主持
- 1993年，完成了对近300万英语词的句子语法结构标注
- 2000年发布中文树库（第一版）
  - 10万词，4185个句子，325 data files（新华社语料）
- 2004年发布中文树库 4.0版
  - 404,156 words, 664,633 Hanzi, 15,162 sentences, and 838 data files（大陆、香港、台湾语料）

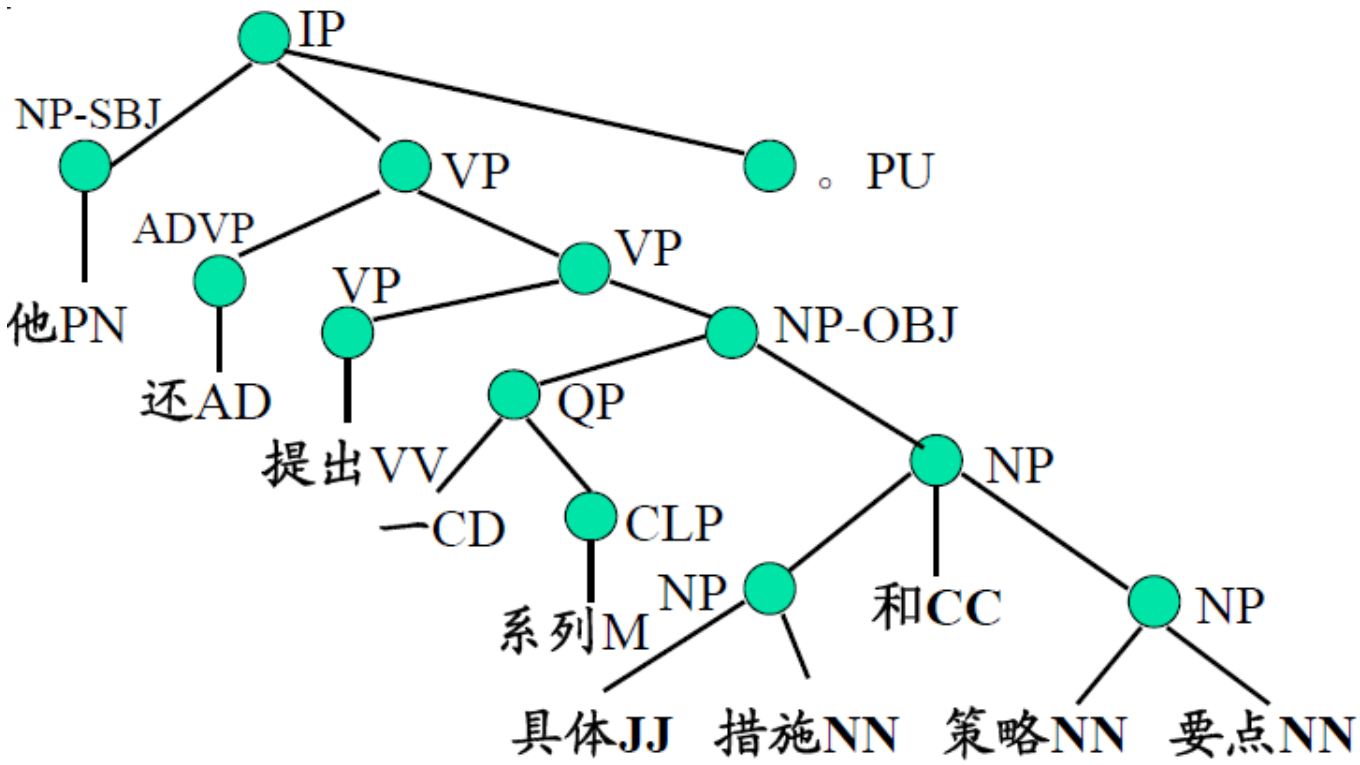
宾州大学中文树库示例

- 他还提出一系列具体措施和政策要点。
- 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

```
(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 。))
```

宾州大学中文树库示例

树结构



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，北航，2000万字)
- 中学语文教材语料库(1983年, 北师大, 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)
- 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，其计划规模达7000万汉字。
- 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。

国内语料库研究状况

- 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，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

- 山西大学、哈尔滨工业大学、北京语言大学、 东北大学、中科院自动化研究所、科技部中信 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、内蒙古师范大学、中央民族大学、社科院民族 学与人类学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。

北京大学语料库

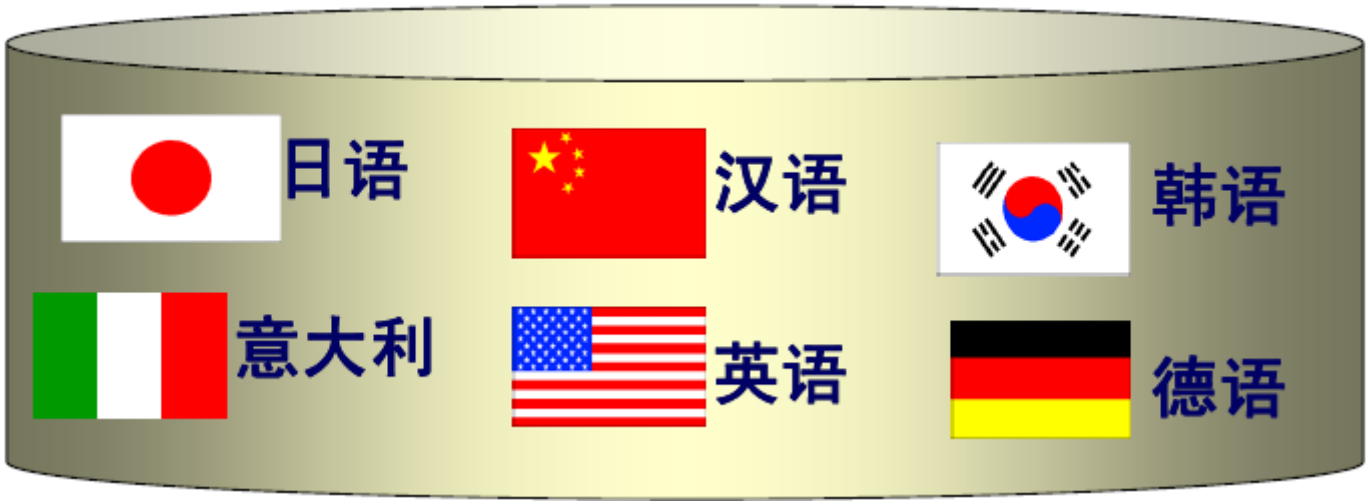
- http://icl.pku.edu.cn/
- 北大计算语言学研究所俞士汶教授主持， 北大、富士通、《人民日报》社共同开发
- 《人民日报》**1998**年全部文本 **(**约**2600**万 字**)**
- 完整的词语切分和词性标注信息
- 样例：

  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**

台湾中研院平衡语料库（Sinica Corpus）

- http://rocling.iis.sinica.edu.tw/ROCLING/corpus98/
- 世界上第一个带有完整词类标记的汉语平衡语料库
- 520万词次(789万汉字)，汉语平衡语料库
- 语料选自1990年至1996年期间出版的哲学、艺术、科学、 生活、社会和文学领域的文本
- 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。
- 2003年增加了汉英平行语料库，含 2373 个汉英平行对照文本。

口语语料库：BTEC (Basic Travelers' Expression Corpus)



20 万句 **6** 国语言对照口语句子
＋ **25** 万句 **3** 国语言对照口语句子

语料库的设计

语料库三方面	属性	值
A. 语料本身	规模	百万词级 | 千万词级 | 亿万词级 | …
	领域	政治 | 经济 | 体育 | 心理学 | …
	体裁	文学 | 应用文 | 新闻 | …
	时代	共时 | 历时
	语体	书面语 | 口语
	语种	单语 | 双语 | 多语 双语平行语料库 | 双语比较语料库
	语言层次	语音（音节，韵律） | 语法(词，句，…)
B. 语料加工	数据形式	Text文本 | HTML文本 | 数据库 | …
	编码体系	TEI标准 | 自定义编码体系 | …
	加工层次	词性 | 句法 | 语义 | 语篇 | … |双语句子对齐 | 词对齐
	加工方式	自动 | 人机互助 | 人工
C. 语料应用	应用领域	通用 | 词典编纂 | 机器翻译 | …
	辅助软件	检索工具 | 人机界面 | 数据接口 | …

语料的选取

- 精品原则
- 有影响力原则
- 随机挑选原则
- 高流通度原则
- 典型性原则
- 易于获得原则
- 具有统计样本意义原则
- 符合语言规范原则

语料库的编码体系

- SGML(标准置标语言)
  - http://www.w3.org/MarkUp/SGML/
- XML(可扩展的置标语言)

  http://www.w3.org/TR/REC-xml
- TEI (文档编码计划)

  http://www.tei-c.org/
- CES(语料库编码标准)

  http://www.tei-c.org/Applications/index-co02.html

- *冯志伟，1998，《标准通用置标语言SGML及其在自然语言处理中的应用》，载《当代语言学》1998年第4期。*

CES标准(Corpus Encoding Standard)

语料库/n 标记/n 应该/v 有/v 规范/n

语料库A: 不符合CES

```
<sample_corpora>
…
<p>
  <s>
    <w POS="n">语料库</w>
    <w POS="n">标记</w>
    <w POS="v">应该</w>
    <w POS="v">有</w>
    <w POS="n">规范</w>
  </s>
</p>
…
</sample_corpora>
```

语料库B: 符合CES

语料库的加工

语料库标注 (Annotation)

(1) 词性标记 (Part-of-speech tagging)

(2) 句法层次和范畴标记 (Grammatical parsing)

(3) 词义标记 (Word sense tagging)

(4) 篇章指代标记 (Anaphoric annotation)

(5) 韵律标记 (Prosodic annotation)

……

语料库加工工具

分类	工具名称	功能描述
文件处理工具	文本过滤器	将不同的文件格式转成为纯文本文件格式
	文本分类器	自动判别文本领域
	语料库辅助校对工具及一致性检查工具	按照语料库加工规范，对语料质量进行管 理
语言处理工具	分词与词性标注工具	对语料进行词语识别，词性标记处理
	词义标注工具	对词义进行标注
	浅层分析工具	对语块(chunk)进行标注
	句法分析工具	对句子进行完全句法分析
	双语语料对齐工具	对双语语料进行各个层级(段落、句子、小 句、词)的对齐加工

双语语料库(Bilingual Corpora)加工

- 段落对齐
- 句子对齐
- 短语对齐
- 词对齐

双语句子对齐

- 基于长度(length-based)的对齐方法 Gale & Church (1993)
  - 纯粹基于句子的长度来估计对齐可能性
  - 资源要求少，算法效率相对较高

  Church, Kenneth W. & Mercer, R. L., Introduction to the special issue on computational linguistics using large corpors, In Computational Linguistics, Vol. 19, No.1, 1993.

- 基于词(word-based)的对齐方法
  - 一般要依赖词典资源，算法效率相对较低

双语句子对齐示例

中国支持在平等参与、协商一致、求同存异、循序渐进的基础上，开展多层次、多渠道、多形式的地区安全对话与合作。

中国参加了东盟地区论坛、亚洲建立协作与建立信任措施会议、亚太安全合作理事会和东北亚合作对话会等活动，主张通过这些政府和民间讨论安全问题的重要渠道，增进各国的相互了解与信任，促进地区和平与稳定。

……

China advocates regional-security dialogue and cooperation at different levels, through various channels and in different forms.

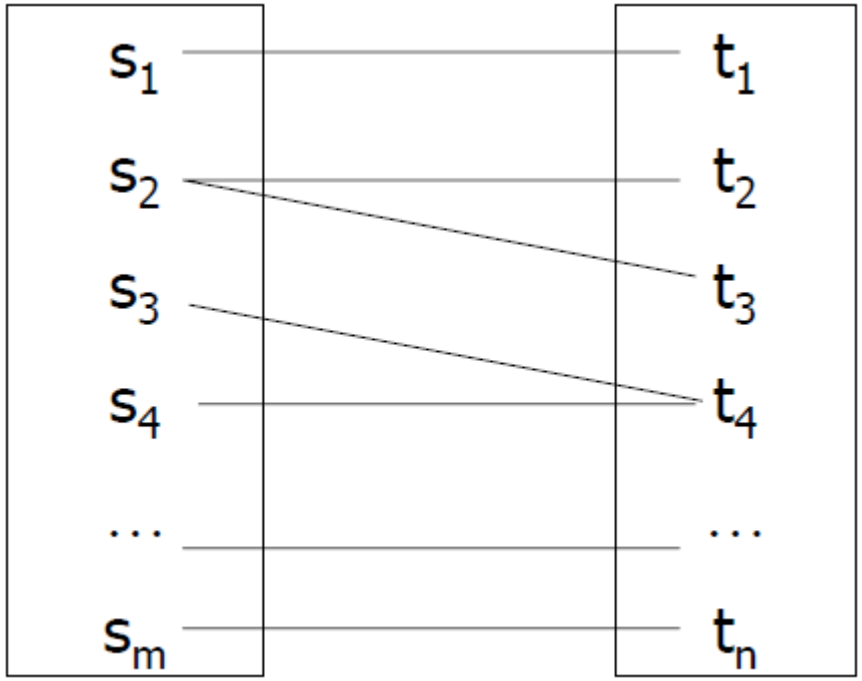
Such dialogue and cooperation should follow these principles: participation on an equal footing, reaching unanmimity through consultation, seeking common ground while reserving differences, and proceeding in an orderly way and step by step.

China has participated in the ASEAN Regional Forum (ARF), Conference on Interaction and Confidence-Building Measures in Asia (CICA), Council on Security Cooperation in Asia and Pacific Regional (CSCAP), Northeast Asia Cooperation Dialogue (NEACD) and other activities, holding that all countries should further mutual understanding and trust by discussions on security issues through these important governmental and non-governmental channels, so as to promote regional peace and stability.

……

双语句子对齐问题

- 影响对齐猜测的两个因素：
  - 配对模式
  - 句长差距



句子配对模式(Match)

- Gale & Church(1993) 定义了六种配对模式，在实际语料中的分布频度为：

句子配对模式 (Match)	出现次数	概率 P(Match)
1-0 或0-1	13	0.0099
1-1	1167	0.89
1-2 或2-1	117	0.089
2-2	15	0.011
	1312	1.00

UBS/Union Bank of Switzerland出版的经济报告，同时使用英、 法、德三种语言

句长相关性 Gale & Church(1993)

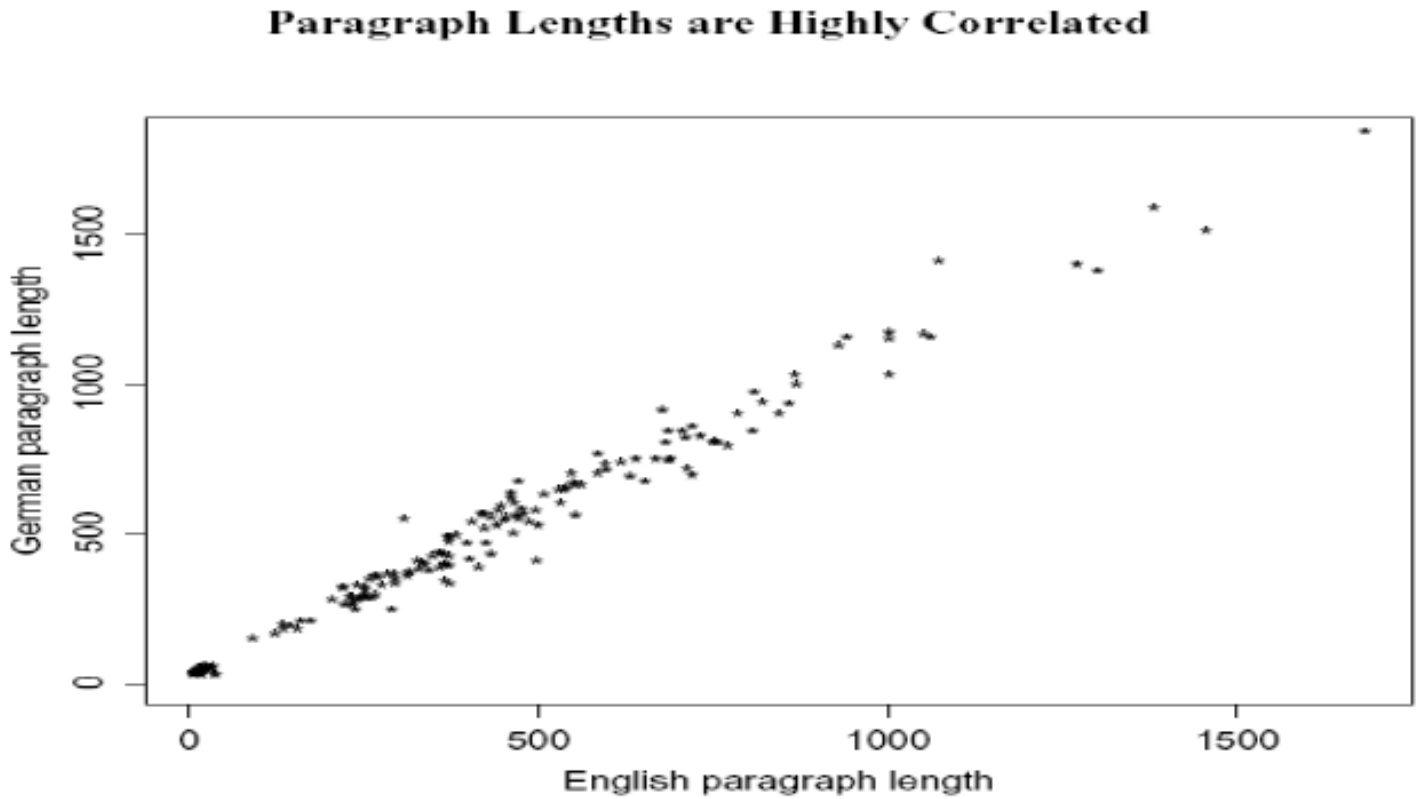


Figure 1. The horizontal axis shows the length of English paragraphs, while the vertical scale shows the lengths of the corresponding German paragraphs. Note that the correlation is quite large (.991).

语料库应用

- 支持自然语言处理应用系统开发
- 支持语言学研究和语言教学研究

语料库对NLP的支持

- 基于大规模语料库的语音识别
- 基于大规模语料库的音字转换技术(中文输入)
- 基于大规模语料库的自动文本校对技术
- 利用语料库训练HMM模型进行分词、词性标注、词义标注
- 基于语料库的句法分析
- 基于语料库的机器翻译

- 基于机器学习技术，通过语料库获取语言知识，包括搭配特征、句法规则等
- 基于语料库的语言模型训练和语法模型评价
- 支持NLP自动评测

中文音字转换

拼音串(无声调)	xue xi dian nao ji shu	
候选字串	雪 系 点 脑 机 树	共有 14×98×41×1 5×167×68 ＝ 95.8亿种可能性
	学 洗 电 闹 给 述	
	学 西 颠 挠 记 书	
	……	
候选词串	学习 电脑 级数	共有 2×1×7＝14种 可能性
	血洗 电脑 奇数	
	血洗 电脑 基数	
	……	
正确文字串	学习 电脑 技术	

基于语料库的句法分析

- 基于广义互信息的句法分析方法 (Magerman & Marcus, 1990)
  - 基于标注好词性的语料库，观察词性序列
  - 假设语法单位的边界可以通过N-gram模型，计算词性序列的互信息来判定
- 广义互信息 (Generalized Mutual Information)
  - $GMI_{(i,j)}(x_1 \cdots x_i, y_1 \cdots y_j) = \sum \frac{1}{\delta_{XY}} MI(X,Y) = \sum \frac{1}{\delta_{XY}} \frac{P(X,Y)}{P(X)P(Y)}$
  - 互信息判断的是成分之间结合的紧密程度，互信息值高的两个成分比互信息值低的两个成分之间更可能组成一个大的语法单位。
  - 采用广义互信息，可以通过扩大词长的方式，计算从2个词性标记到多个词性标记之间的广义互信息，从而通过多遍扫描，对句法层次的嵌套进行分析判定

基于语料库的句法分析

- 利用定界语法排除那些不可能构成一个语法单位的成分
  - 将那些不可能被分析为一个短语片段的词性序列列出来
  - 例如 “n p”两个词性不可能组成一个语法单位，应该被分到两个单元中

例: He directed the cortege of autos to the dunes near Santa Monica.

词性： r v det n p n p det n p n

基于语料库的句法分析

- 计算互信息，运用定界语法排除不可能的组合后，结果为
  - (He) (directed) (the cortege) (of autos) (to) (the dunes) (near Santa Monica)
- 继续进行互信息计算，直到无法插入新的句法分界标记为止，得到结果：
  - (He (directed ((the cortege) (of autos))) ((to (the dunes)) (near Santa Monica)))
- 在Brown语料库上的结果，长度≤15个词的句子，平均每句2个错误

基于双语语料库的翻译知识抽取

- 在经过句子对齐的双语语料基础上，抽取翻译等价对
  - 假设：扫描句子对齐的双语语料库，生成所有候选的翻译等价对；
  - 检验：通过统计关联的度量方法，从候选的翻译等价对中选出统计意义上较为可靠的翻译等价对
  - 利用词性等信息剔除明显不可能的对子
  - 利用关联度手段甄别候选对：点式互信息、DICE系数、χ^2统计、对数似然性度量等

- 任何存款在5年内无人认领，可视之为收受作政府用途的款项…
- A deposit which is unclaimed for 5 years may be treated as moneys received for the purposes of the government and …

基于双语语料库的翻译知识抽取

汉语	英语	χ^2统计值
附注	remark	496.47
款	subsection	496.24
废除	repeal	495.81
命令	order	493.20
豁免	exemption	490.83
附表	schedule	489.95
许可证	license	488.87
文告	proclamation	487.53
更改	vary	485.82
行政长官	Chief_Executive	484.37
财政司司长	Financial_Secretary	475.71
香港	Hong Kong	473.91
香港特别行政区	Hong Kong Special Administrative Region	448.58

基于语料库的语言研究

- 搭配的定量研究

关于搭配的描述与定义

- 搭配是重复出现的
  - “大房子”—“大手笔”
- 搭配是不可类推的（自由组合 – 受限组合）
  - “吃白菜”—“吃豆腐”—“喝西北风”
- 搭配一般具有正常的句法结构
  - “戴高帽”
- 搭配通常与领域相关
  - “语言习得”—“学说话”　　“风险投资”

搭配的量化分析 1

- 语料库：90-91年新华社新闻语料库，1000万字，710万词
- 搭配强度： 重复出现越多，搭配强度越大

$$MI(w_i, w_j) = \log_2 \frac{p(w_i, w_j)}{p(w_i)p(w_j)}$$

$$S(w_i, w_j) = \log_2 \frac{N \sum_{k=-5}^{+5} Count_k(w_i, w_j)}{Count(w_i)Count(w_j)}$$

K= -5, -4, -3, -2, -1, 1, 2, 3, 4, 5

N 表示语料库规模

K 表示wj相对于wi的位置
－表示在左，+表示在右

搭配强度的量化分析示例

- 候选搭配：(能力，弱) (能力，大)
- 通过语料库统计得到：

$Count_{-3}$ (能力, 弱) = 1, $Count_{1}$ (能力, 弱) = 3, $Count_{2}$ (能力, 弱) = 5

$Count_{-5}$ (能力, 大) = 6, $Count_{-4}$ (能力,大) = 4, $Count_{-3}$ (能力, 大) = 8

……

$Count_{1}$ (能力, 大) = 9 …… $Count_{5}$ (能力, 大) = 5

Count (能力)=2241, *Count* (弱)=177, *Count*(大)=19913

搭配强度的量化分析示例

$$S(\text{能力},\text{弱}) = \log_2 \frac{7.1\times10^6(1+3+5)}{2241\times177} = 7.33$$

$$S(\text{能力},\text{大}) = \log_2 \frac{7.1\times10^6(6+4+8+4+2+9+6+4+6+5)}{2241\times19913} = 3.10$$

同理可得：

$S(\text{能力},\text{强}) = 7.45$　　$S(\text{能力},\text{差}) = 6.63$　　$S(\text{能力},\text{小}) = 0.74$

与“能力”的搭配能力：　强 > 弱 > 差 > 大 > 小

搭配的量化分析 2

- 搭配的离散度

$$u(w_i, w_j) = \frac{\sum_{k=-n}^{n}((Count_k(w_i, w_j) - \overline{Count}(w_i, w_j))^2}{2n}$$ 方差公式

$$\overline{Count}(w_i, w_j) = \frac{\sum_{k=-n}^{n} Count_k(w_i, w_j)}{2n}$$ 均值公式

$n=5$

离散度反映两个成分共现的分布情况。
离散度越高，越可能是搭配。

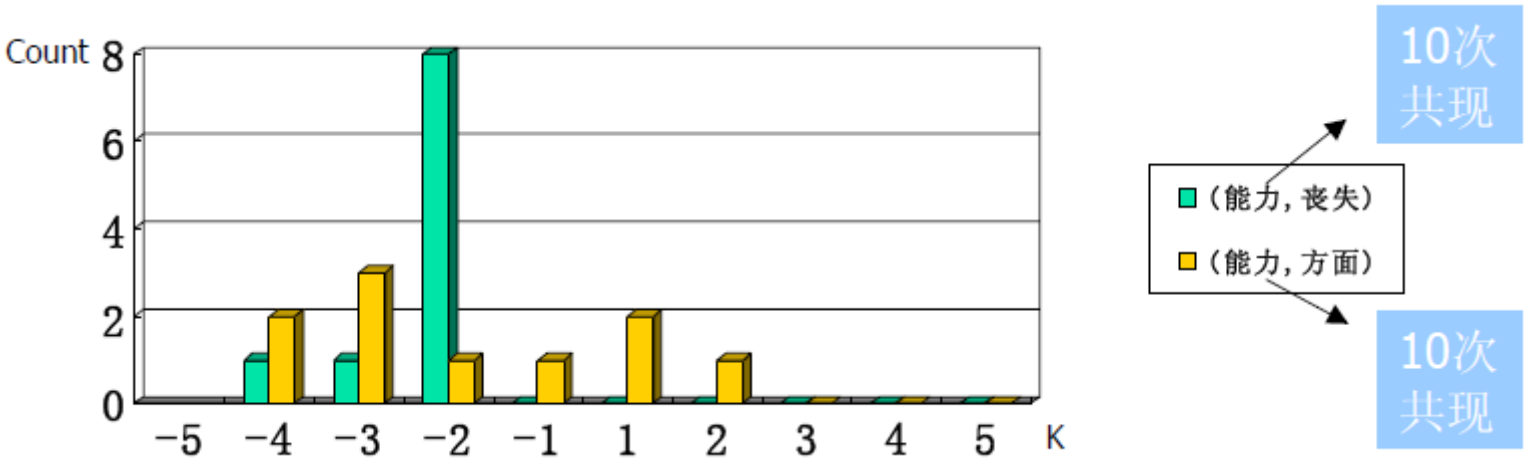
搭配离散度的量化分析示例

- 候选搭配： (能力,丧失)　　(能力,方面)
- 通过语料库统计得到：

Count$_{-4}$ (能力, 丧失) = *Count*$_{-3}$ (能力, 丧失) = 1, *Count*$_{-2}$ (能力, 丧失) = 8

Count $_{-4}$ (能力, 方面) = *Count*$_{1}$ (能力, 方面) = 2, *Count*$_{-3}$ (能力, 方面) = 3

Count$_{-2}$ (能力, 方面) = *Count*$_{-1}$ (能力, 方面) = *Count*$_{2}$ (能力, 方面) = 1



搭配离散度的量化分析示例

$$\overline{Count}(\text{能力, 丧失}) = \frac{1+1+8}{2\times 5} = 1 \qquad \overline{Count}(\text{能力, 方面}) = \frac{2+3+1+1+2+1}{2\times 5} = 1$$

$$u(\text{能力, 丧失}) = \frac{(1-1)^2 + (1-1)^2 + (8-1)^2}{10} = 5.60$$

$$u(\text{能力, 方面}) = \frac{2\times(2-1)^2 + (3-1)^2 + 3\times(1-1)^2}{10} = 1.00$$

“丧失”与“能力”构成搭配关系，而“方面”跟“能力”不构成搭配关系

搭配的量化分析 3

- 搭配的尖峰位置度量

$$Z_k(w_i, w_j) = \frac{(Count_k(w_i, w_j) - \overline{Count}(w_i, w_j)}{\sqrt{u(w_i, w_j)}}$$

$$Z_{-2}(\text{能力, 丧失}) = \frac{Count_{-2}(\text{能力, 丧失}) - \overline{Count}(\text{能力, 丧失})}{\sqrt{u(\text{能力, 丧失})}}$$

$$= \frac{8-1}{\sqrt{5.6}}$$

$$= 2.96$$

“丧失”在 -2 位置形成尖锋

“尖锋位置”的语言学含义

反映W_j与W_i可能形成的句法结构

- 能力－具有　　尖峰位置 -3, -2　　述宾结构
- 能力－差　　尖峰位置 1　　主谓结构
- 能力－提高　　尖峰位置 -4,-3; 1,3　　述宾|主谓

Z_{-3} (能力, 提高) > Z_1 (能力, 提高)

- 能力 － 吞吐　　尖峰位置 -1　　定中结构

词汇知识库

什么是知识

- Knowledge is justified true belief. -柏拉图
- 一条陈述能称得上是知识必须满足三个条件，它一定是被验证过的，正确的，而且是被人们相信的
- 知识也是人类在实践中认识客观世界（包括人类自身）的成果，它包括事实、信息的描述或在教育和实践中获得的技能。
- 知识是人类从各个途径中获得得经过提升总结与凝练的系统的认识。
- AI的核心是研究怎样用计算机易于处理的方式表示各种知识

聪明的AI vs 有学识的AI

- 人类大脑依赖所学的知识进行思考、逻辑推理、理解语言

聪明的AI

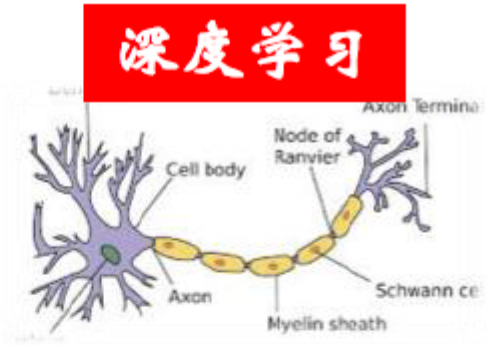
感知

识别

判断

学习

推理

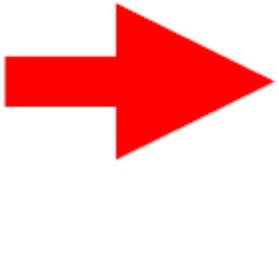


有学识的AI

思考

语言

推理



知识表示

- 知识表示（Knowledge Representation,KR）就是用计算机易于处理的方式来描述人脑的知识
- KR不是数据格式、不等同于数据结构或编程语言
- 数据与知识的区别在于KR支持推理

人的记忆偏重关联

Memex: Memory Extender

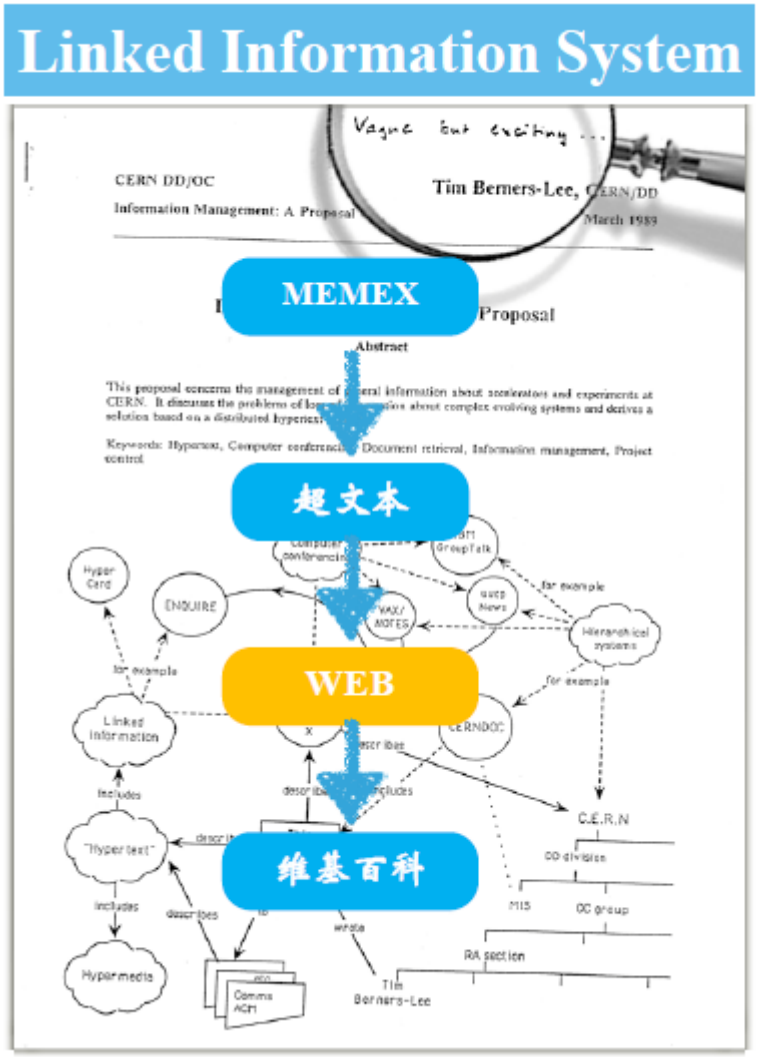
"Wholly new forms of encyclopedias will appear, ready made with a mesh of associative trails running through them, ready to be dropped into the memex and there amplified"

As We May Think, The Atlantic,1945

Vannevar Bush

National Science Foundation
Manhattan Project
First Presidential Science Advisor

Web: 以链接为中心的系统



…This is why a "web" of notes with links between them is far more useful than a fixed hierarchical system. …… Circles and arrows leave one free to describe the interrelationships between things in a way that tables, for example, do not. The system we need is like a diagram of circles and arrows, where circles and arrows can stand for anything.

Information Management: A proposal 1989.

以“链接”为中心的系统，在开放的互联网环境里面更加容易生长和扩展。这一理念逐步被人们实现，并演化发展成为今天的万维网。

万维网创始人
MIT教授
图灵奖获得者

Sir Tim Berners-Lee

语义网：从链接文本到链接数据

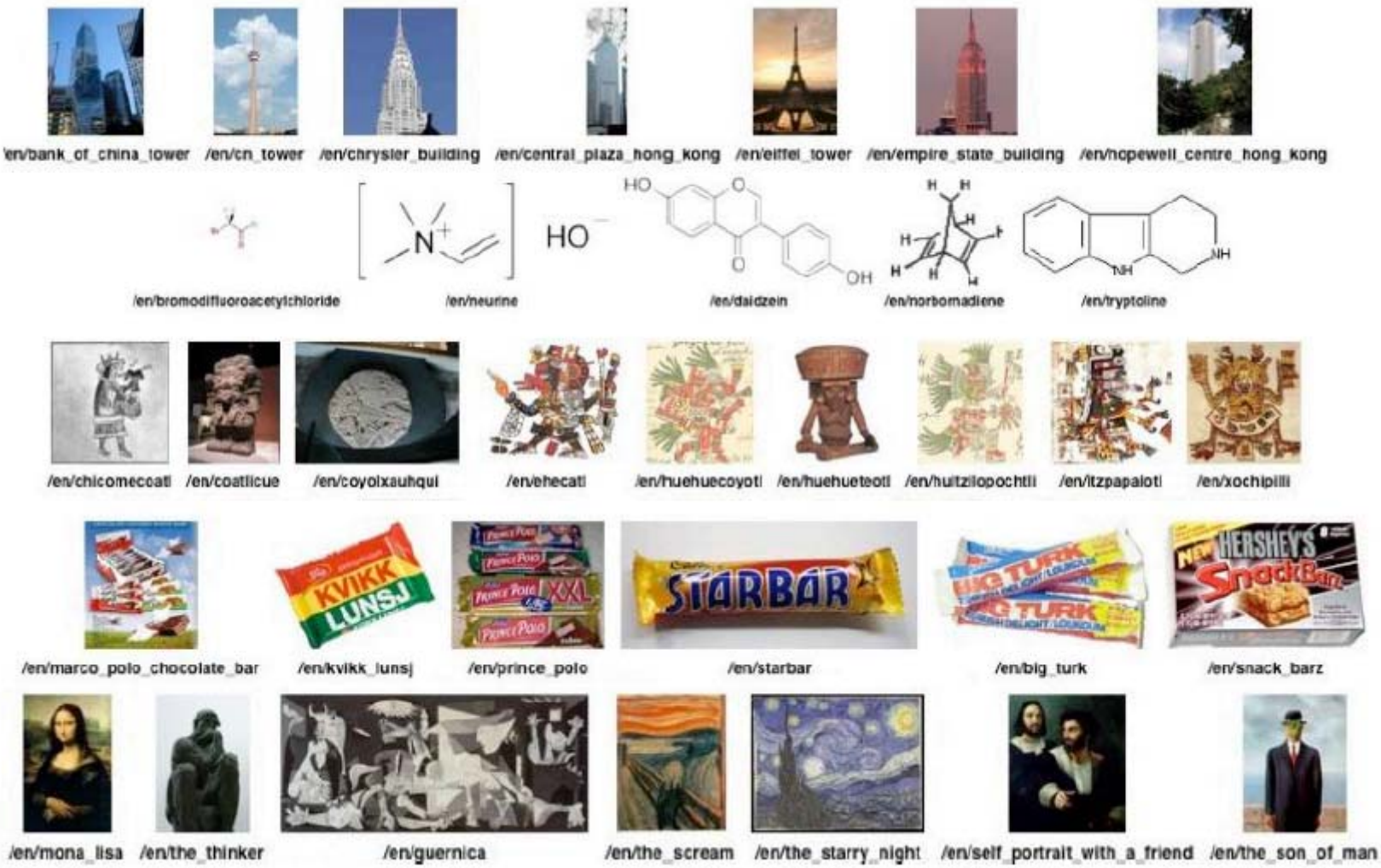
"This is a pity, as in fact documents on the web describe real objects and imaginary concepts, and give particular relationships between them but we could not process them at all...

Tim Berners-Lee，Inventor of the Web, @WWW Geneva, 1994

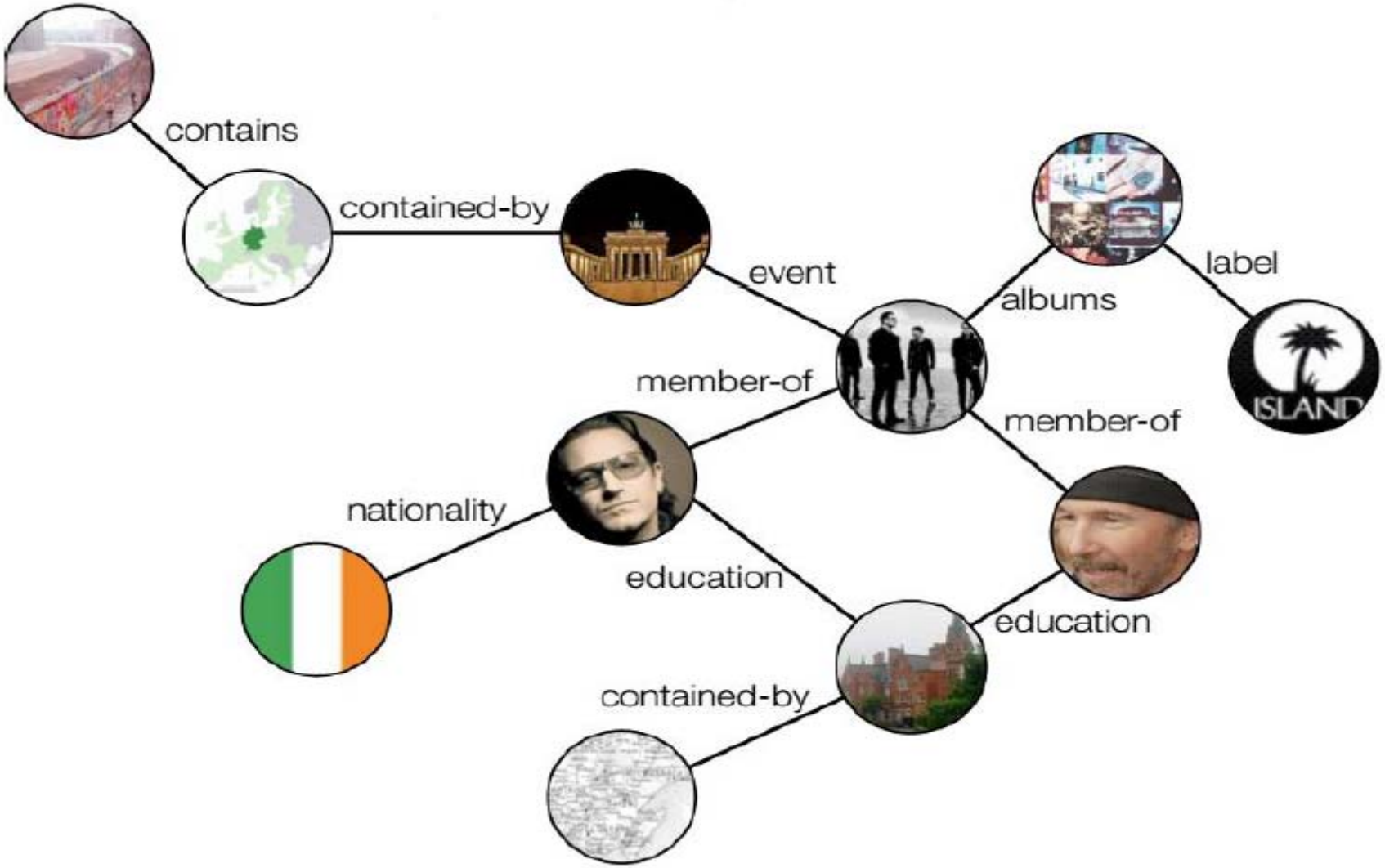
Web of Texts, Web of documents

Web of Objects, Web of Data, Web of Things

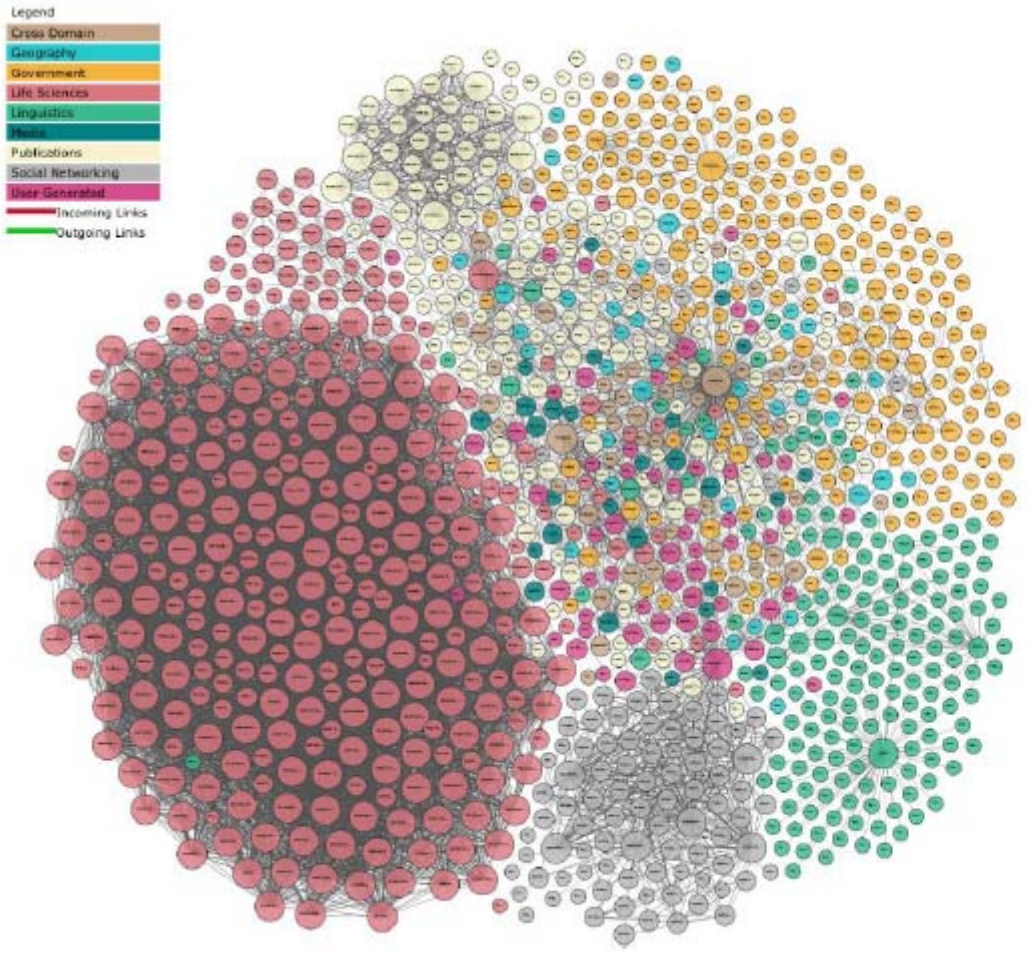
Web中有多种事物

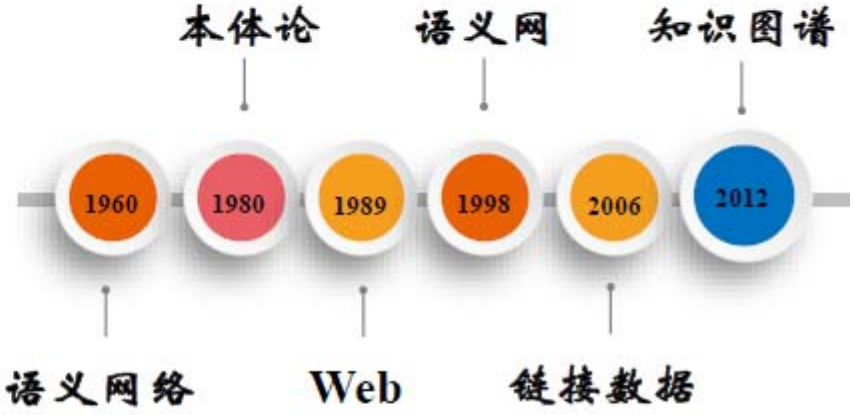

/en/bank_of_china_tower
/en/cn_tower
/en/chrysler_building
/en/central_plaza_hong_kong
/en/eiffel_tower
/en/empire_state_building
/en/hopewell_centre_hong_kong
/en/bromodifluoroacetylchloride
/en/neurine
/en/daidzein
/en/norbornadiene
/en/tryptoline
/en/chicomecoatl
/en/coatlicue
/en/coyolxauhqui
/en/ehecatl
/en/huehuecoyotl
/en/huehueteotl
/en/huitzilopochtli
/en/itzpapaloti
/en/xochipilli
/en/marco_polo_chocolate_bar
/en/kvikk_lunsj
/en/prince_polo
/en/starbar
/en/big_turk
/en/snack_barz
/en/mona_lisa
/en/the_thinker
/en/guernica
/en/the_scream
/en/the_starry_night
/en/self_portrait_with_a_friend
/en/the_son_of_man

事物间有多种类型的连接



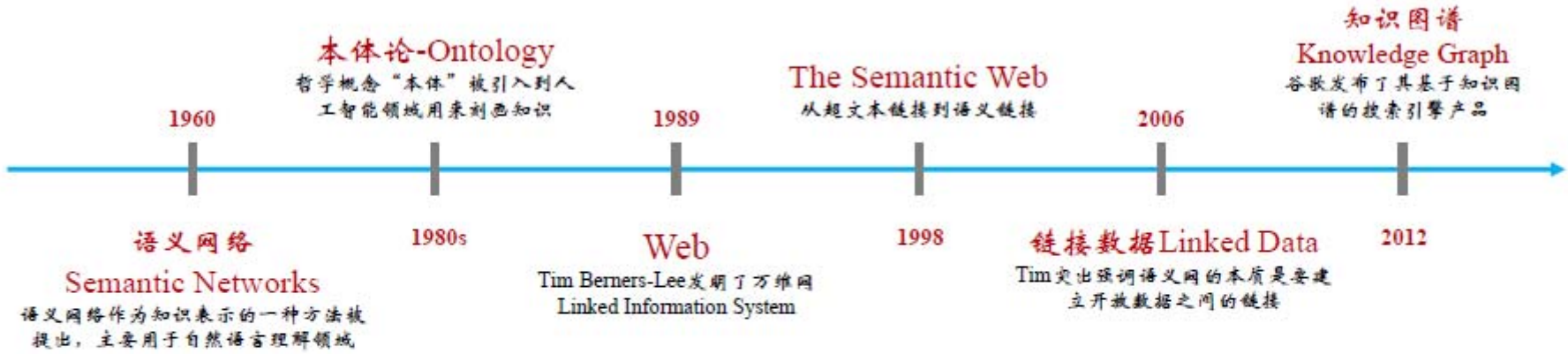
知识图谱 Knowledge Graph





"Linking Open Data cloud diagram 2017, by Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/"

知识图谱的起源



知识表示与知识库- Knowledge Representation / Knowledge Base

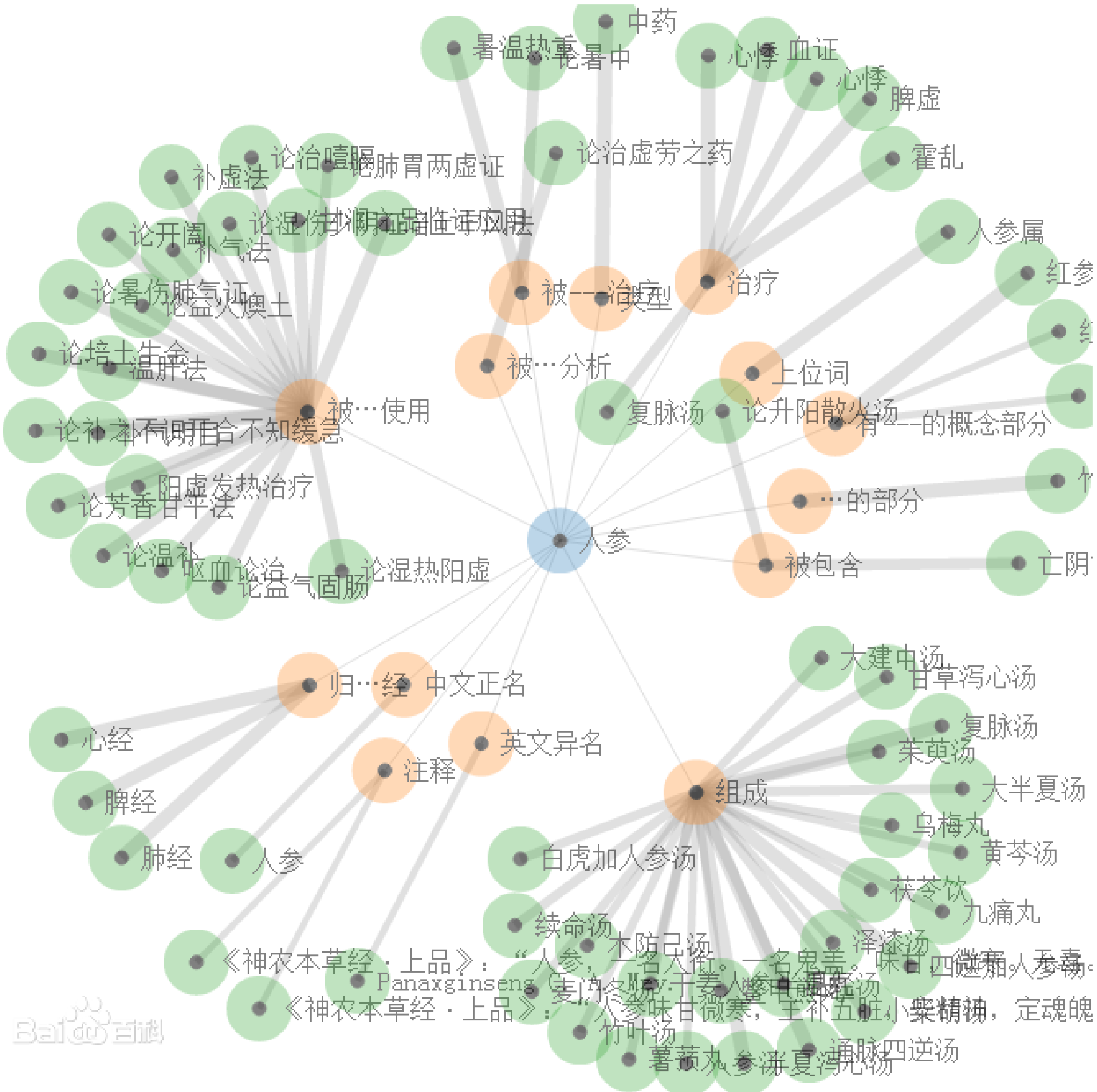
人工智能研究者陆续提出了大量知识表示的方法，如框架系统、产生式规则、描述逻辑等。

知识图谱得益于Web的发展（更多的是数据层面），有着来源于KR、NLP、Web、AI多个方面的基因。

各类知识图谱项目


Freebase
LINKINGOPENDATA
W3C SWEO Community Project
WordNet
A lexical database for English
ZHISHI.me
PKUBASE
NELL
schema.org
....the new SEO?
WIKIDATA
WEB CHILD
DBpedia
XLORE
CN-DBpedia
ConceptNet
An open, multilingual knowledge graph
Herbnet
yago
select knowledge
LinkedGeoData.org
linked life data
WEBKB

知识图谱



人参
被…使用
被…治疗
被…分析
治疗
上位词
有…的概念部分
…的部分
被包含
归…经
中文正名
英文异名
注释
组成
中药
论治虚劳之药
心悸
脾虚
霍乱
人参属
红参
复脉汤
论升阳散火汤
补虚法
论开阖
补气法
论湿热阳虚
论益气固肠
论温补
阳虚发热治疗
论芳香甘平法
温肝法
心经
脾经
肺经
大建中汤
甘草泻心汤
茱萸汤
大半夏汤
乌梅丸
黄芩汤
茯苓饮
九痛丸
白虎加人参汤
续命汤
竹叶汤
通脉四逆汤

WordNet – 词义知识库

- 自然语言处理的发展
  - 从词法分析、句法分析到语义分析
  - Web应用对“内容理解”的强烈需求
    - 智能检索，文本分类，自动文摘，语义推理，Word Sense Disambiguation，Semantic Web应用等

词义区分的研究:

- “方向”
  - (1) 指东、南、西、北等：在山里迷失了~
  - (2) 正对的位置；前进的目标：军队朝渡口的~行进
  - (3) 情势：看~做事
- “事情”
  - (1) 人类生活中的一切活动和所遇到的一切社会现象：~多，忙不过来
  - (2) 事故，差错：出~就麻烦了
  - (3) 职业，工作：在公司里找了一个~
  - “忙公司里的~” 该标(1) /(3)?

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。(实际上 WordNet 中仅包含前4类)

虚词往往没有语义

WordNet

WordNet的理论与方法

- 概念：由同义词集(Synset)来表示，概念即同义词集
  - {教师，教员，老师，先生，师傅，师爷，孩子王，臭老九，阿姨，导师，老板}
- 知识本体 (Ontology)：概念及概念之间多种语义关系，形成概念网络。
- 一个高度形式化的、通用/跨语言的词义知识表示方法
  - HowNet 对词义的内涵式定义：意在定义，关注个体，建立在义素分析及格语法上
  - WordNet对词义的外延式定义：意在区分，关注系统，建立在词义系统结构分析上
    - 词义即多种语义关系约束构成的网络结构
    - 添加新词、新义项时只需要 将特定词加入特定同义词集
- 一个可以对词义进行分析、计算的基础，一个同义词集之上的 “形-义”系统

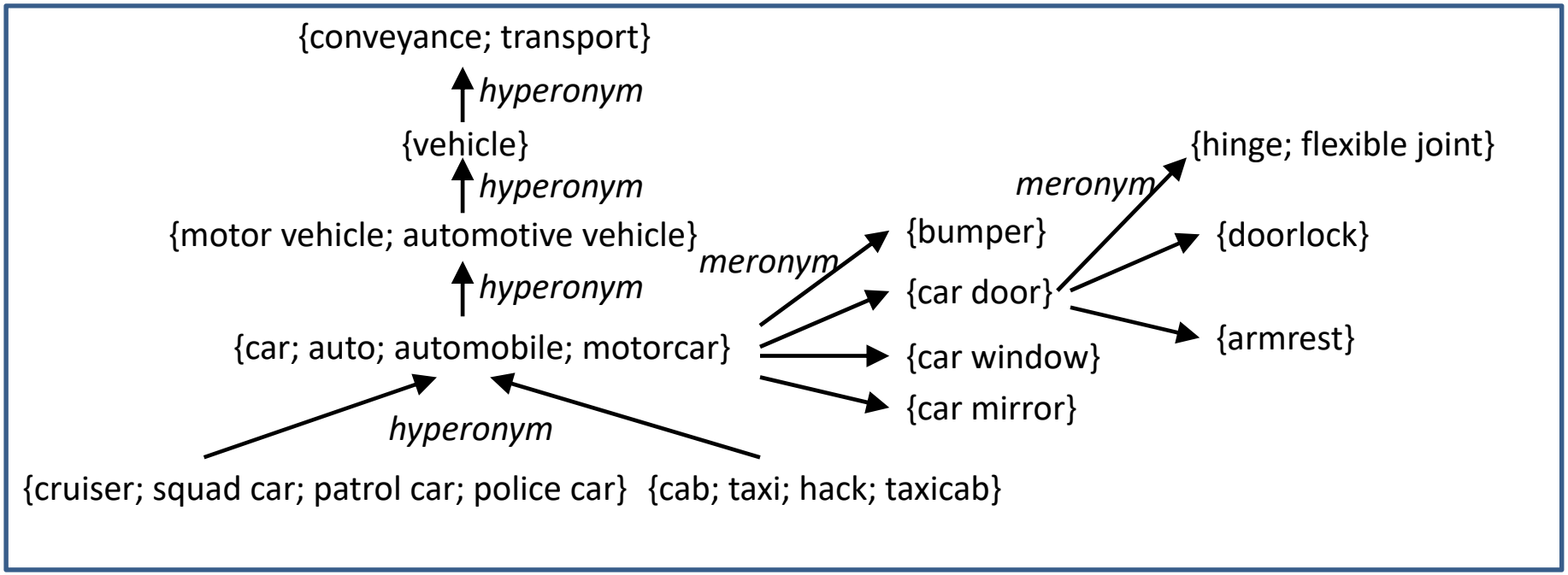
WordNet

- 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**。

WordNet

- **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。



WordNet

◆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}

- **21000**个动词词形、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。

兼容WordNet的中文概念词典（CCD)

- CCD作为双语WordNet
  - 提供汉英双语的概念对应
  - 可以直接复用现有的WordNet的理论、方法、技术
  - 全球WordNet资源建设的组成部分
- 构造双语WordNet的焦点与难点
  - 两类不同的知识体系和概念对应
    - 大规模的复杂网络结构
  - 实用、高效的双语WordNet构造模型的必要特征
    - 强调双语词典构造中的继承与转换
    - 方便双语词典的演化
- 研究与开发现状
  - 对PWN 1.6的覆盖率已经在90%以上
    - 6.6万名词概念，1.2万动词概念，2.1万修饰词概念
  - 基本满足概念对应的语义原则，语义质量持续提高

兼容WordNet的中文概念词典（CCD)

Offset	Synset	Hypernym	Hyponym	OtherRel	Definition
07632177	教师 教员 老师 先生 师傅 师爷 师长 阿姨 导师 老板 孩子王	07235322	070863320 716230407 209465072 437670727 965907297 622073411 760740109 807414251 074251800	********	以教学为职业的人
Offset	Synset	Hypernym	Hyponym	OtherRel	Definition
07331418	丈夫 先生 夫君 夫婿 爱人 老公 郎君 驸马 驸马爷	07602853	071094820 719596807 255726073 28008	********	已婚男子； 婚姻中女性一方的伴侣
07414666	先生 师傅 同志 大哥 老兄 老弟	07391044		********	对男子的一种称呼

兼容WordNet的中文概念词典（CCD)

CCD不仅仅是双语WordNet

- 反映汉语的实际情况，对中文信息处理有切实帮助
  - 对概念、概念关系的界定、调整和发展
  - 增添汉语特有/显著的语义属性和特征
    - 简称(j)，褒义(c)，中性(n)，贬义(d)，正规(r)，非正规(i)等
- 涉及复杂结构的规划和调整
  - 分类原则、概念粒度等
- 补充必要的组合关系信息
  - 自然语言处理任务和应用的客观需要
  - 搭配信息
- 关注词义区分 (Word Sense Discrimination)及语言学证据的表现
  - 面向机器并影响人群，强调尽可能自动/半自动化的词义区分

WordNet 的应用

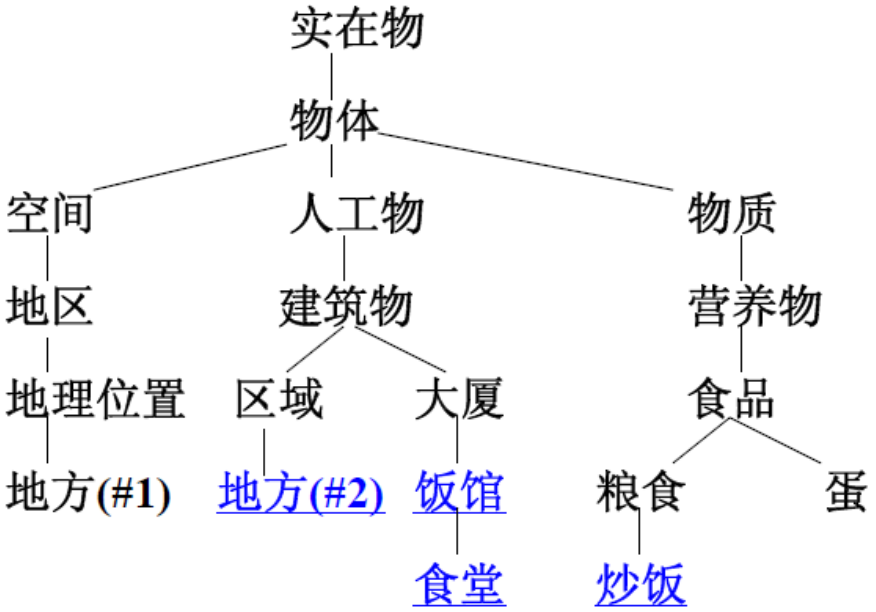
- 词汇消歧，语义推理，理解等。

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：他说的有些地方不对

- 三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



知网(HowNet)

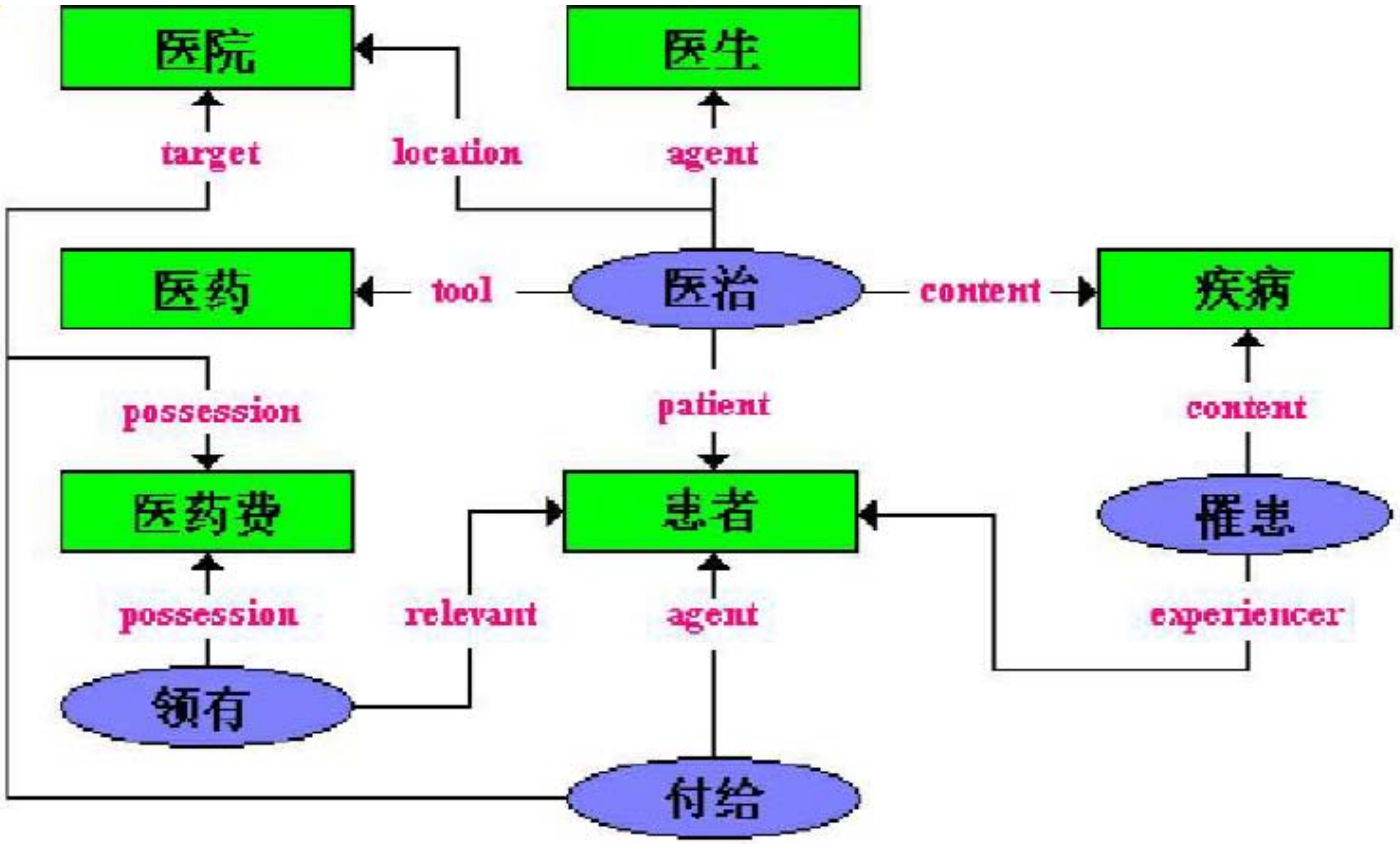
http://www.keenage.com

1988年由董振东教授提出，**4**个基本观点：

- **NLP**系统最终需要更强大的知识库的支持。
- 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。
- 关于知识库建设，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。
- 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。

知网(HowNet)

- 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。



知网(HowNet)

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
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）

知网(HowNet)

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）
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）