第六章 语义分析

王 峰
华东师大计算机系

主要内容

1. 语义简介

2. 语义知识的内容及其形式化表示
   - 词的语义
   - 句子语义

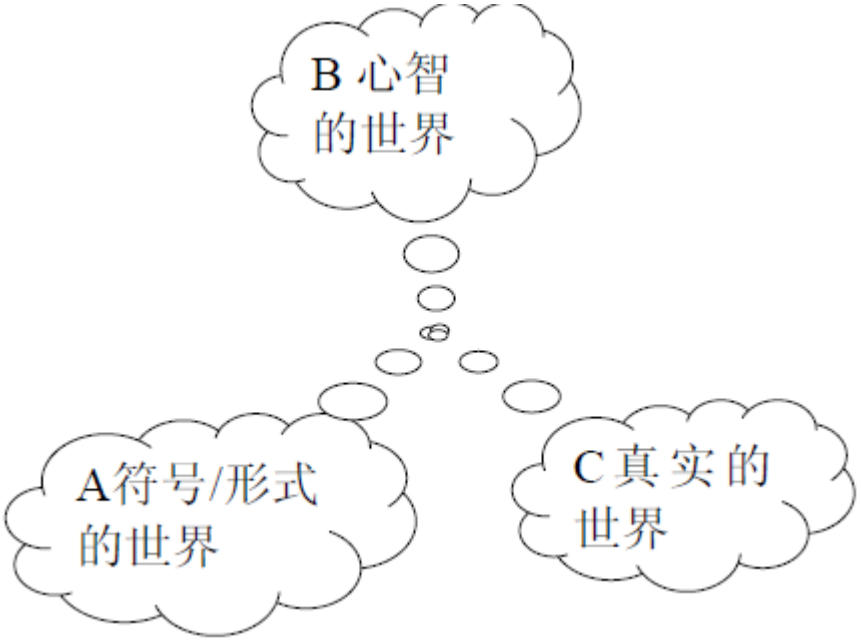
3. 语义知识的应用

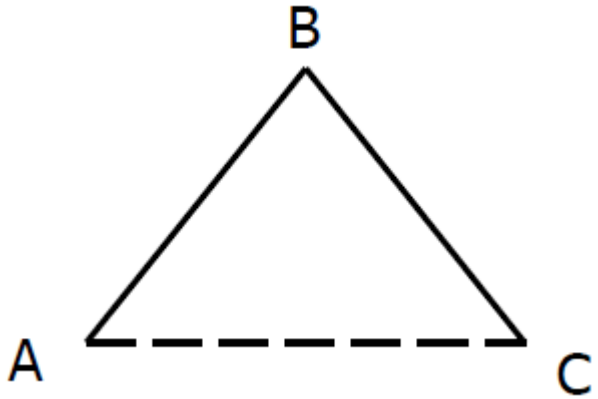
1 语义分析

- 语义计算的任务：解释自然语言句子或篇章各个部分(词、词组、句子、段落、篇章)的意义。

- 面临的困难：
  - 自然语言句子中存在大量的歧义，涉及指代、同义/多义、量词的辖域、隐喻等；
  - 同一句子对于不同的人来说可能有不同的理解；
  - 语义计算的理论、方法、模型尚不成熟。

从形式到意义：引入语义知识的必要性





A是对B的抽象，B是对C的抽象；
A通过B，与C发生间接的联系

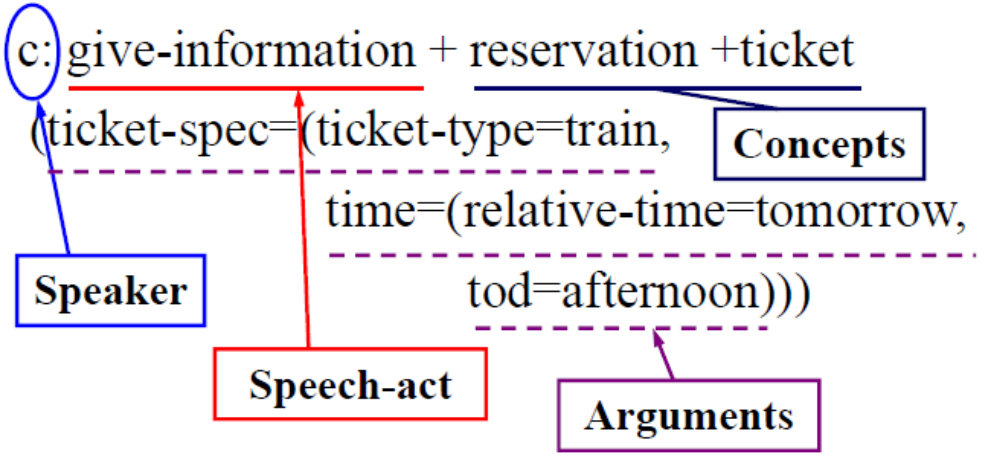
Ogden,C.K. & I.A.Richards, 1923，The Meaning of Meaning，Routledge & Kegan Paul, London.

语义理论简介

- 词的指称作为意义
  - 该理论认为，词或词组的意义就是它们在现实世界上所指的事物。那么计算语义学的任务就是将词或词组与世界模型中的物体对应起来。
  - 常用的现实世界模型假设世界上存在各种物 体，包括人。
  - 缺陷：对于复杂的问题这种定义无法处理。
    - 神仙，鬼，妖怪
- 心理图像、大脑图像或思想作为意义
  - 该理论认为，词或词组的意义就是词或词组在 人心理上或大脑中所产生的图像。
  - 缺陷：在计算机中把心理图像有效地表示出来并不是一件容易的事情；而且，不一定所有的 词义都有清晰的心理图像。

语义理论简介

- 说话者的意图作为意义
  - 该理论试图解释语言中一种被称为言语行为**(Speech Acts)** 的现象。
  - 说话者把自己的话语当作行为希望听者理 解、作出反应。这种意义被认为是独立于逻辑意义之外的。
  - 例如：我想预订明天下午的火车票。
  - 缺陷：意图的定义和划分困难。



语义理论简介

- 过程语义
  - 该理论认为，句子的语义定义为接受该句后所执行的程序或者所采取的某种动作。
  - 优点：简单明了，对于计算机智能应用系统来说，这种定义在某种程度上是有效的。
  - 缺陷：对于语言本身缺乏解释，且句子的语义常常和应用连接紧密，缺乏独立性。
- 词汇分解学派
  - 该理论把句子的语义基于它所含有的词和词组的意义之上，而词的意义则基于一组有限特征，这组特征通常称为语义基元。这样，只要给出一组语义基元和一些操作符，就可以把句子的语义描述出来。类似于化学中的元素学说。
  - 缺陷：语义基元的定义、分解标准等不好把握， 基元和组合操作的合理性直接影响句子语义描写的准确性。

在符号世界的内部看意义

- 对于一种语言中的两个表达形式，人们一般可以判断二者之间是否具有某种关系，比如同义关系，两个表达式所对应的命题之间的逻辑蕴含关系，等等。**(语义知识)**

- 意义 := 符号/形式变换

I. 在一种语言内进行的符号变换
   A. 张三打了李四 → B. 李四被张三打了

II. 在不同语言之间进行的符号变换
   A. 张三用手打了李四 → B. Zhang San hit Li Si with his hand

III. 在不同性质的符号系统之间进行的符号变换

 → B. 这是残疾人通道

A的意思是B

Paraphrase – 解释意思

If you paraphrase someone or paraphrase something that they have said or written, you express what they have said or written in a different way.

— Collins COBUILD Dictionary

para- “beside” + phrazein “to tell", to tell in other words

—http://www.etymonline.com/

意义：符号之间无止尽的变换关系

A　　B

“我买了辆车” → 意思1：我付钱从某处购买了一辆车

意思2：我拥有了一辆车

意思3：我可以使用这辆车

意思4：我的钱都花完了

……

A 意味着 B {B1,B2,B3,B4,...}

一个形式可以变换为其他多个形式

例

1. 这个**编辑**很不错。 → Editor
2. 他每年要**编辑**一百万字的书。 → Edit

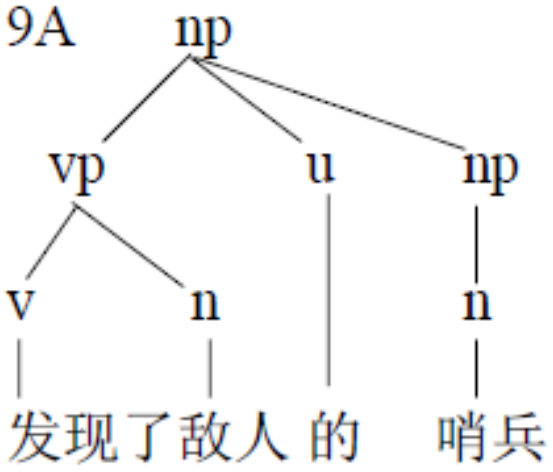
3. 这小伙子是干警察的好**材料** → 人选
4. 把这个小伙子的**材料**送人事部 → 资料
5. 这是一种新型**材料** → 原料

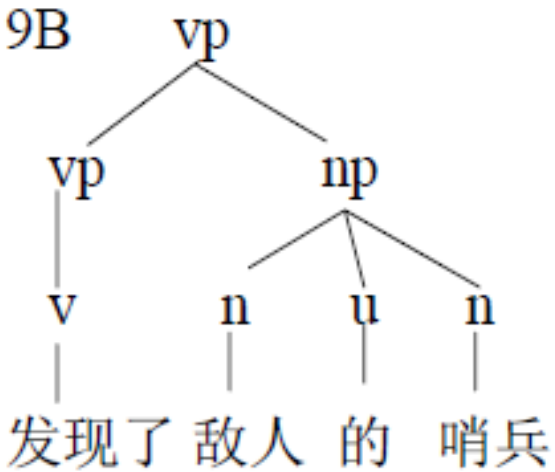
6. 编辑部有**许多读者** → 编辑部有许多人
7. 这本书有**许多读者** →× 这本书有许多人

一个形式可以变换为其他多个形式

例　发现了敌人的哨兵

vp　　np　de　np





多个符号形式可以变换为同一个形式

- 例1

  学一食堂 供应 西餐 吗？
  学一食堂 卖 西餐 吗？
  学一食堂 提供 西餐 吗？
  学一食堂 有 西餐 吗？
  学一食堂 做 西餐 吗？
  学一食堂 经营 西餐 吗？

  供应(学一食堂，西餐)

- 例2

  学一食堂 抵制 西餐 吗？
  学一食堂 反对 西餐 吗？
  学一食堂 对西餐说不 吗？

  ¬供应(学一食堂，西餐)

2 语义知识的内容及其形式化表示

- 关于词义聚类关系的知识
  - ✓语义特征/义素描述
  - ✓语义分类树
  - ✓语义关系网

词的语义

- 关于词义组合关系的知识
  - ✓**配价理论**
  - ✓**格语法**
  - ✓元结构理论
  - ✓框架语义学理论

句子语义

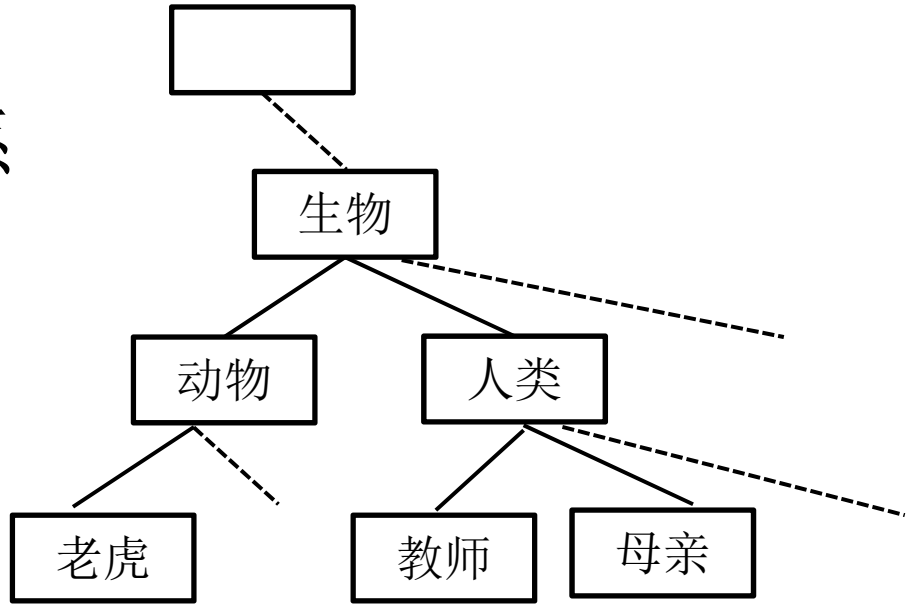
2.1 关于词义聚类关系的知识

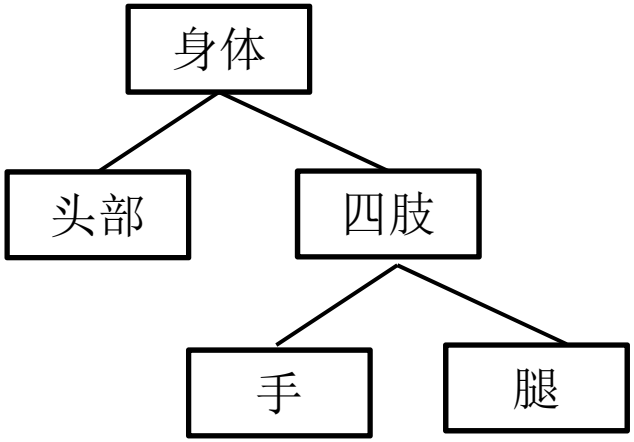
- 词义，是对词代表的概念描述。
- 词义聚类关系：同义(近义)关系，反义关系，同位关系，上/下位关系，整体-部分关系，……

词**义知识的类型**	**代表性的语义知识库/工程项目**
语义分类树	905语义工程
语义关系网	HowNet WordNet/ 北大ICL-CCD MindNet
语义特征/义素的描述	HowNet

语义分类树

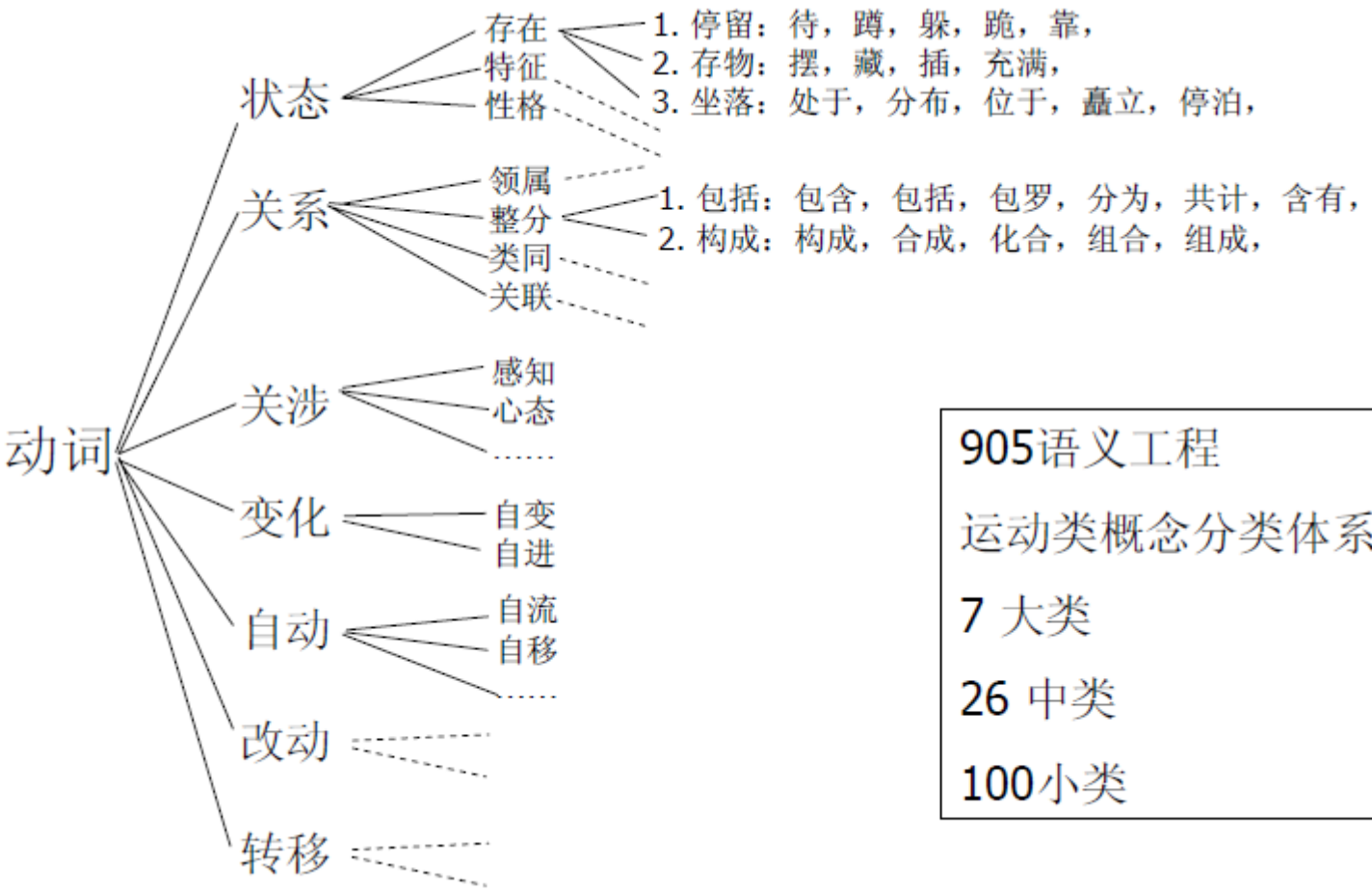
上下位关系





整体-部分关系

语义分类树



905语义工程

运动类概念分类体系：

7 大类

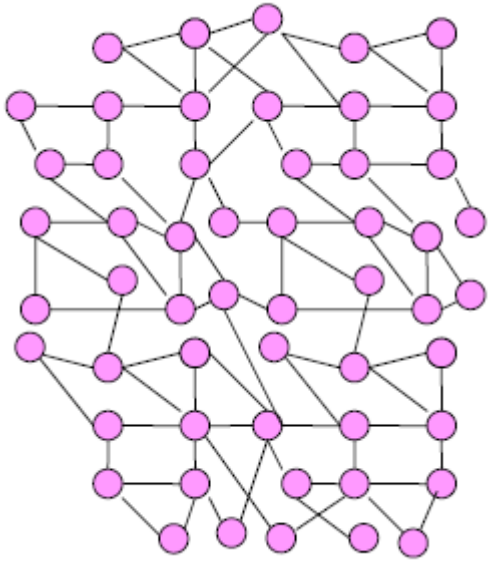
26 中类

100小类

语义关系网

Attribute属性	Goal目标	Possessor领有者
Cause原因	Hypernym上位	Purpose意图
Co-Agent联合施事	Location场所	Size大小
Color颜色	Manner方式	Source源点
Deep_Object深层宾语	Material材料	Subclass子类
Deep_Subject深层主语	Means方法	Synonym同义
Domain领域	Modifier修饰语	Time时间
Equivalent同位	Part部分	User使用者

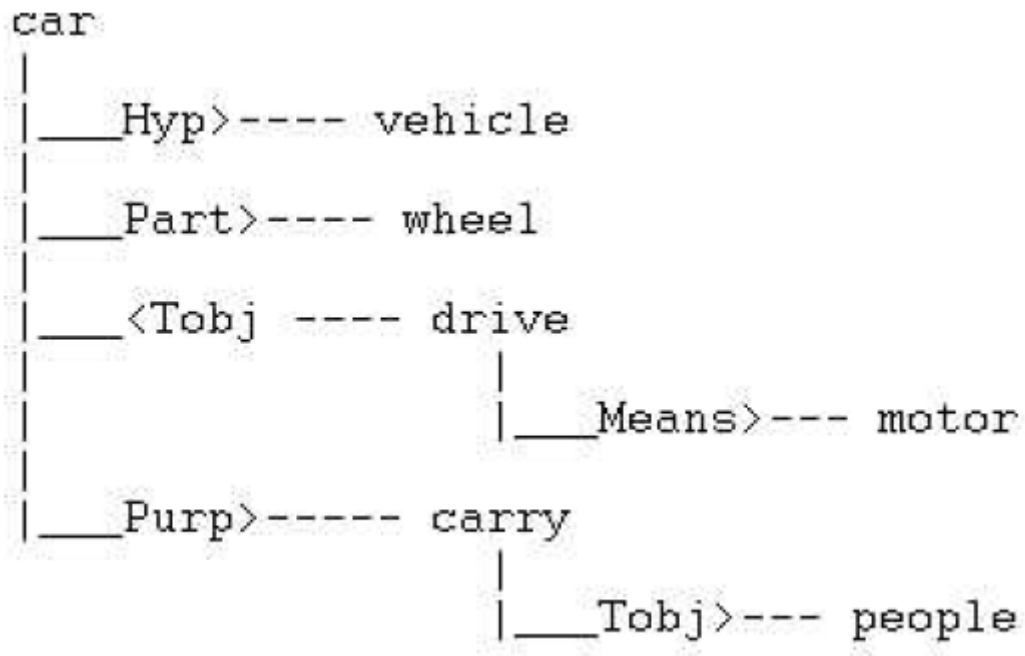
MindNet



http://stratus.research.microsoft.com/mnex/InputPath.aspx?l=e&d=d

语义关系网的自动获取

car: a vehicle with 3 or usually 4 wheels and driven by a motor, esp. one for carrying people



Stephen D.Richardson, William B.Dolan, Lucy Vanderwende, MindNet: acquiring and structuring semantic information from text

特征结构表示语义知识

((词语：警察)(词义类：人)(成年：是)(职业：是)(亲属：否))
((词语：叔叔)(词义类：人)(职业：否)(亲属：是))

- 描述词在汉语系统中的地位

- 帮助判断字符串的合法性

例：

  - 警察叔叔，医生叔叔，护士阿姨，工人伯伯
  - 叔叔警察，叔叔医生，阿姨护士，伯伯工人 ×

语义特征（义素）描述

- 表示同义关系和反义关系

n个义素

义素 / 词语	婚姻	成年	男性	人
光棍	－	＋	＋	＋
寡妇	－	＋	－	＋
女童	－	－	－	＋
男婴	－	－	＋	＋
单身汉	－	＋	＋	＋

2^n个词语

语义特征（义素）描述

n个义素 →

	对象	接触	环境	工具	方式	…
炒	荤、素、面	间接	固体	…	…	
烤	荤、面、素	直接	固体	…	…	
炸	荤、面、素	间接	油	…	…	
煮	荤、面、素	间接	水	…	…	
蒸	面、荤、素	间接	汽	…	…	
炖	…	…	…			
…						

↓ 2^n个词语

语义特征（义素）描述

- 语义特征的性质
  - 概念义（客观义）
  - 附加义（感情色彩义/主观义）

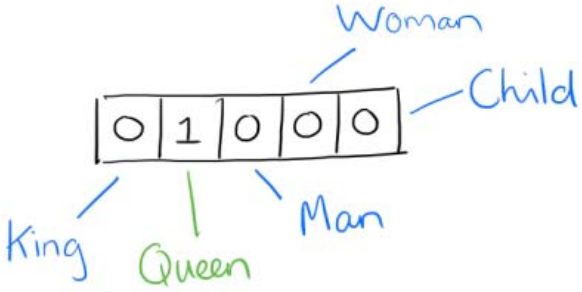
- 获取语义特征（义素）的方式：
  - 参考词典释义，提取共同项/对立项特征；
  - 考察词语在句法格式中使用时的变换差异和对比差异，对具有相同/相反的句法行为的词语，提取共同项/对立项特征。

Word2Vec

- google在2013年公布的开源工具包
- 用来将word vector化，映射每个词到一个向量，可用来表示词对词之间的关系
- 词向量具有良好的语义特性，是表示词语特征的常用方式。
- 词向量每一维的值代表一个具有一定的语义和语法上解释的特征。所以，可以将词向量的每一维称为一个词语特征。

词向量

- 用一个向量的形式表示一个词
- One-hot表示
  - 首先，统计出语料中的所有词汇
  - 然后对每个词汇编号，针对每个词建立V维的向量，向量的每个维度表示一个词
  - 对应编号位置上的维度数值为1，其他维度全为



Queen: [0,1,0,0,0]
King: [1,0,0,0,0]

词向量

One-hot表示的缺点

- 任意两个词之间都是孤立的，无法表示语义层面上词汇之间的相关信息
  - 任意两个词向量的相似度都为0
- 稀疏向量，维度大

词向量

- 分布式表示 Dristributed representation
- 词嵌入Word Embedding
- 思路：
  - 通过训练，将每个词都映射到一个较短的词向量上来
  - 所有的这些词向量就构成了向量空间，进而可以用普通的统计学的方法来研究词与词之间的关系。
  - 词向量维度大小：训练时自己指定

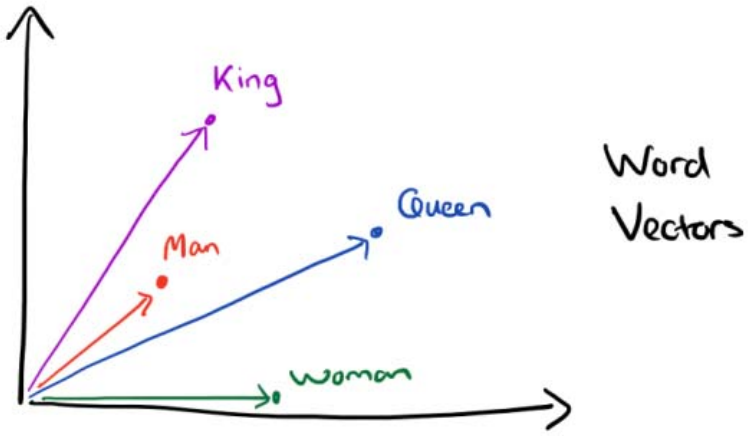
词向量

- 例如：将词汇表里的词用"Royalty", "Masculinity", "Femininity“, 和"Age"4个维度来表示

词向量

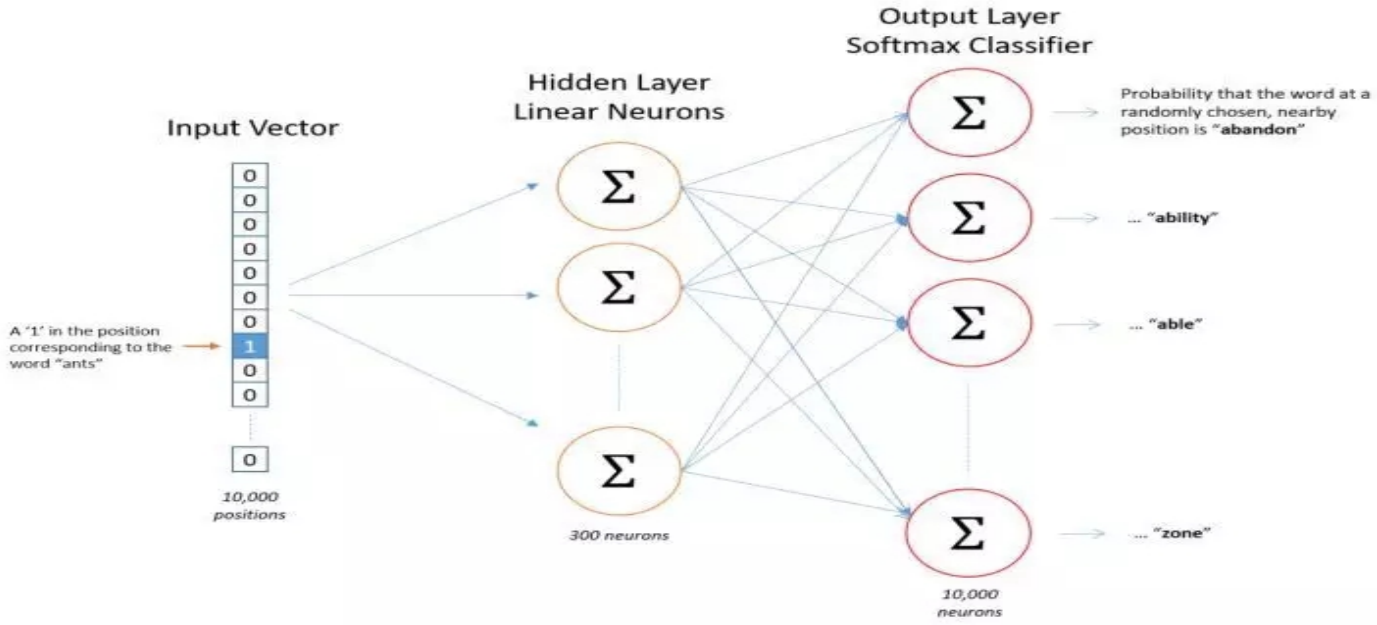
- 有了用Dristributed representation表示的较短的词向量，我们就可以较容易的分析词之间的关系了
- 比如我们将词的维度降维到2维，用下图的词向量表示我们的词时，可以发现

$$\overrightarrow{King} - \overrightarrow{Man} + \overrightarrow{Woman} = \overrightarrow{Queen}$$



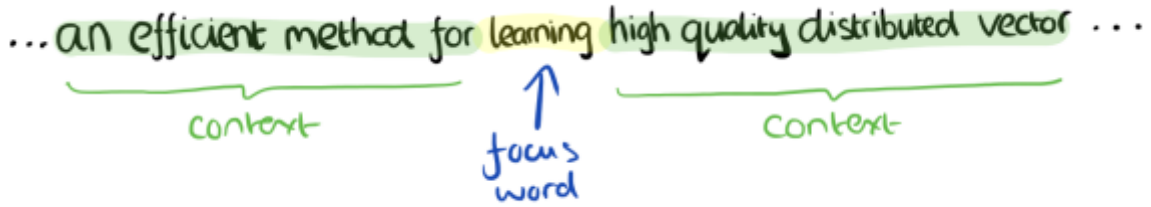
词向量训练

- 神经网络语言模型
- 一般是一个三层的神经网络结构（当然也可以多层），分为输入层，隐藏层和输出层(softmax层)
- 词向量是训练神经网络时候的隐藏层参数或者矩阵
- CBOW(Continuous Bag-of-Words)
- Skip-Gram



词向量训练

- CBOW模型
- 输入是某一个特征词的上下文相关的词对应的词向量
- 输出就是这特定的一个词的词向量
- 假设上下文大小取值为4，上下文对应的词有8个，前后各4个，这8个词是模型的输入
- 输出是所有词的softmax概率（训练的目标是期望训练样本特定词对应的softmax概率最大），对应的CBOW神经网络模型输入层有8个神经元，输出层有词汇表大小个神经元



词向量训练

- 隐藏层的神经元个数我们可以自己指定。通过DNN的反向传播算法，我们可以求出DNN模型的参数，同时得到所有的词对应的词向量。
- 当有新的需求，要求出某8个词对应的最可能的输出中心词时，可以通过一次DNN前向传播算法并通过softmax激活函数找到概率最大的词对应的神经元即可

词向量训练

- Skip-Gram模型
- 和CBOW的思路是相反的，即输入是特定的一个词的词向量，而输出是特定词对应的上下文词向量。
- 还是上面的例子，我们的上下文大小取值为4， 特定的这个词"Learning"是我们的输入，而这8个上下文词是我们的输出

词向量训练

- 在Skip-Gram的例子里，输入是特定词， 输出是softmax概率排前8的8个词
- 对应的Skip-Gram神经网络模型输入层有1个神经元，输出层有词汇表大小个神经元。隐藏层的神经元个数我们可以自己指定。
- 通过DNN的反向传播算法，可以求出DNN模型的参数，同时得到所有的词对应的词向量。
- 当有新的需求，要求出某1个词对应的最可能的8个上下文词时，可以通过一次DNN前向传播算法得到概率大小排前8的softmax概率对应的神经元所对应的词即可。

Word2Vec

- DNN模型的这个处理过程非常耗时
- 词汇表一般在百万级别
- Word2Vec也使用了CBOW与Skip-Gram来训练模型与得到词向量
- 两种加速方法
  - Negative Sample
  - Hierarchical Softmax
- 学习到的是根据共现信息得到的单词的表达

词的含义也可以用词在哪些文档中出现过来体现，将TF-IDF**反**过来

2.2 关于词义组合关系的知识

目标：描述任意词语之间的组合语义关系

语义知识理论	**代表性的语义知识库/工程项目**
配价语法	ILD，北大配价语义词典
格语法/论元结构理论	现代汉语述语动词机器词典，Propbank
框架语义学	FrameNet

2.2.1 配价语法

- 配价（valence）这一概念借自化学。目的是说明一个动词能跟多少个名词性成分发生组合关联。

游泳：	[某人] 游泳	1价动词
吃：	[某人] 吃 [某食物]	2价动词
送：	[某甲] 送 [某乙] [某物]	3价动词

- 动名语义组合关系：
  1. 可以跟动词组配的名词性成分的个数 – ***论元数***
  2. 跟动词组配的名词论元的类型 – ***论旨角色***
  3. 动词对其论旨角色的选择限制

配价数

- 语义：动作行为跟 x 类事物有意义联系。
- 句法：动词周围有 x个空位安放跟它有意义联系的名词。

V	零元（价）动词
__V	一元（价）动词
__V __	二元（价）动词
__V __ __	三元（价）动词
……	X 元（价）动词

张三发现了一个秘密	发现(X, Y)
张三告诉了李四一个秘密	告诉(X, Y, Z)

形容词、名词的配价

- 形 — 名 配价
  - 优秀　[某人] 优秀　　1价形容词
  - 友好　[某甲] 对 [某乙] 友好　　2价形容词
- 名 — 名 配价
  - 质量　[某物] 的 质量　　1价名词
  - 态度　[某甲] 对 [某乙]的态度　2价名词

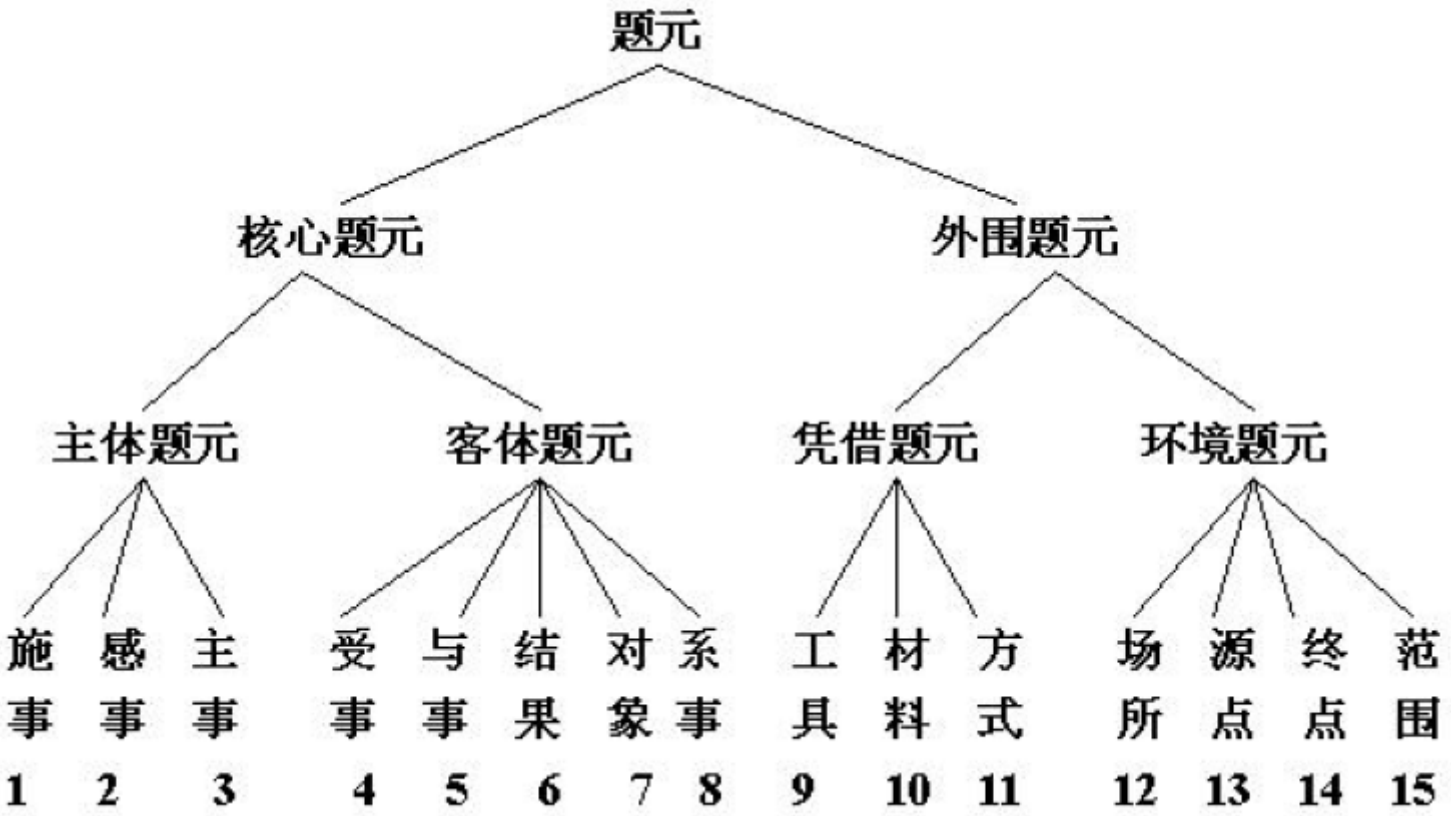
论元类型 — 论旨角色

反映动词所能结合的名词的不同语义角色

- 施事：动作的发出者
- 受事：动作所涉及、影响的内容和对象
- 与事：因动作受益或受损的人
  - 张三告诉了李四一个秘密
  - 施事：张三；受事：秘密；与事：李四
- 工具：张三吃大碗
- 处所：张三睡沙发
- 结果：张三考研究生
- ……

语义上的角色和论旨角色似乎没有那么多的句法上的关系

汉语动词的论旨角色层级系统



对论旨角色的选择限制(selectional restriction)

- 对许多动词，跟它搭配的名词性成分不是任意的
- 动词对其论旨角色进行选择限制

吃	施事论元	[+动物]	张三吃西瓜
	受事论元	[+固体][+食物][+药物]	* 张三吃思想

想	对象论元	[+人]	想妈妈	思念
	对象论元	[+抽象物]	想答案	思索

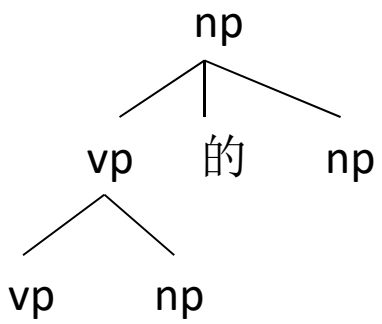
- 以特征结构描述动词与名词组配关系的语义知识，并记录在词典中

[词语：　发现
施事：　[语义类：人类]
受事：　[语义类：事物]]

语义组配关系与句法结构对应

- 结合句法的产生式规则系统，附加语义合一条件。

- {R1} vp → vp np :: $.内部结构=述宾，%vp.受事=%np
- {R2} np → vp 的 np :: IF %vp.内部结构=述宾，THEN %vp.施事 =%np ENDIF



- 发现宝藏的人
- 知道敌军的意图
  - “意图”的语义类不是”人类”，不能跟”知道”的特征结构做合一运算，即不能作为”知道”的”施事”，不满足R2的约束

$$\begin{bmatrix} \text{词语：} & \text{发现} \\ \text{施事：} & [\text{语义类：人类}] \\ \text{受事：} & [\text{语义类：事物}] \end{bmatrix}$$

$$\begin{bmatrix} \text{词语：} & \text{知道} \\ \text{施事：} & [\text{语义类：人类}] \\ \text{受事：} & [\text{语义类：事物}] \end{bmatrix}$$

配价理论小结

配价描述的内容	取值（具体的描述方式）
配价数	1. 动词跟 x 类名词有语义联系； 2. 动词周围有 x 个空位放置名词；（x 一般取值 0-3 ）
论旨角色	1. 跟动词有语义关联的 x 类名词分属哪些类型（*T1, … Ti*）； *2. Ti* 能够出现在动词周围的哪些空位上。
动词对其论旨角 色的选择限制	充当*Ti*的名词要满足哪些条件？ 1. 语法（形式）特征 | 语义属性（类别）特征； 2. 包容性条件 | 排除性条件。

把字句和被字句

2.2.2 格语法

- 格语法**(Case Grammar)** 是美国语言学家 **Charless J. Fillmore** 于**1966**年提出的。
- 基本观点
  - **C. J. Fillmore** 指出：诸如主语、宾语等语法关系实际上都是表层结构上的概念，在语言的底层，所需要的不是这些表层的语法关系，而是用施事、受事、工具、受益等概念所表示的句法语义关系。这些句法语义关系，经各种变换之后才在表层结构中成为主语或宾语。

格语法

区别于“所有格”之类的

- 格语法**(Case Grammar)**中的格是“深层格”，它是指句子中体词**(**名词、代词等**)**和谓词**(**动词、形容词等**)** 之间的及物性关系**(transitivity)** ,如： 动作和施事者的关系、动作和受事者的关系等。 这些关系是语义关系，它是一切语言中普遍存在的现象。

- 这种格是在底层结构中依据名词与动词之间的句法语义关系确定的，这种关系一经确定就固定不变，不管经什么操作、在表层结构中处于什么位置、与动词形成什么语法关系，底层上的格与任何具体语言中的表层结构上的语法概念，如主语、宾语等，没有对应关系。

格语法

例如：

(1) The door opened.

(2) The key opened the door.

(3) The boy opened the door.

(4) The door was opened by the boy.

(5) The boy opened the door with a key.

- **the boy**：施事格
- **the door**：客体格（受事格）
- **the key**：工具格

格语法

格语法的三条基本原则：

(1) $S \rightarrow M + P$

- 句子 **S** 可以改写成情态**(Modality)**和命题**(Proposition)** 两大部分，情态部分包括否定、时、式、体以及其他被理解为全句情态成分的状语。
- 命题牵涉到动词和名词短语、动词和内嵌小句之间的关系，动词是句子的中心，名词短语按其特定的格属关系依附于该动词。

格语法

- (2) $P \rightarrow V + C_1 + C_2 + \cdots + C_n$
  - 命题 P 都可以改写成一个动词 V 和若干个格C。动词是广义上的动词，包括：动词、形容词、甚至包括名词、副词和连词。
- (3) $C \rightarrow K + NP$
  - K 为格标，是各种格范畴在底层结构中的标记，可以有各种标记形式，如：前置词、后缀词、词缀、 零形式等。

格语法

C. J. Fillmore 在**1968**年的论文中认为，命题中的格包括**6**种：

1. 施事格**(Agentive)**：动作的发生者；
2. 工具格**(Instrumental)**：对动作或状态而言作为某种因素而牵涉到的无生命的力量或客体。
3. 承受格**(Dative)**：由动词确定的动作或状态所影响的有生物。如，**He is tall.**
4. 使成格**(Factitive)**：由动词确定的动作或状态所形成的客体或有生物。或理解为：动词意义的一 部分的客体或有生物。如：**John dreamed a dream about Mary.**
5. 方位格**(Locative)**：由动词确定的动作或状态的 处所或空间方位。如：**He is in the house.**
6. 客体格**(Objective)**：由动词确定的动作或状态所影响的事物。如：**He bought a book.**

格语法

后来 **Fillmore** 在语言分析时又增加了一些格：

7. 受益格**(Benefactive)**：由动词确定的动作为之服务的有生命的对象。如：**He sang a song for Mary.**
8. 源点格**(Source)**：由动词确定的动作所作用到的事 物的来源或发生位置变化过程中的起始位置。 如：**He bought a book from Mary.**
9. 终点格**(Goal)**：由动词确定的动作所作用到的事物的终点或发生位置变化过程中的终端位置。如：**I sold a car to Mary.**
10. 伴随格**(Comitative)**：由动词确定的与施事共同 完成动作的伴随者。如：**He sang a song with Mary.**

***格的数目和名称并不是确定的。

格语法

用格语法分析语义：格框架约束分析

- 格框架表示
  - 格框架中可以有语法信息，也可以有语义信息，语义信息是整个格框架最基本的部分。
  - 一个格框架可由一个主要概念和一组辅助概念组成，这些辅助概念以一种适当定义的方式与主要概念相联系。
  - 一般地，在实际应用中，主要概念可理解为动词，辅助概念理解为施事格、受事格、处所格、工具格等语义深层格。

格语法

例: **In the room, he broke a window with a hammer.**

```
[BREAK
    [ case-frame:
        [agentive: HE
         objective: WINDOW
         instrumental: HAMMER
         locative: ROOM ]
        [MODALs:
            time: past
            voice: active ]]]
```

格语法

分析的基础

- 词典中记录动词的格框架和名词的语义信息。
- 对于动词：规定它们所属的必备格、可选格或禁用格，同时填充这些格的名词的语义条件。
  - 如：《动词用法词典》把名词按其与动词格的关系分为**14**类：受事、结果、对象、工具、方式、处所、时间、目的、原因、致使、施事、同源、等同、杂类。
- 对于名词：填充语义信息，建立名词语义分类体系。

格语法

分析步骤

1. 判断待分析词序列中主要动词，在动词词典中找出该词的格框架；
2. 识别必备格：
   - 如果格带有位置标志，则从指定位置查找格的填充物；
   - 如果格带有语法标志，则在这个分析的词序列中查找语法标志，进入相应的填充；
   - 如果格框架还需要其它必备格，查找其它名词的语义信息，按格框架的语义信息要求进行相应的填充。
3. 识别可选格；
4. 判断句子的情态 **Modal**。

格语法

格框架分析可以和句法分析结合起来：

a) 句法分析：判断出句子的动词、名词短语、介词短语等；

b) 查找动词的格框架与名词短语、介词短语的格关系，并进行相应的填充。

- 必须首先找到动词，从而获得格框架。

格语法

CASE

[Agentive: the young athlete
Locative: Los Angeles
Neutral: the young athlete
[Modal
[Tense: Future
MOOD: Declarative
Time: next week]]]

格语法

格语法描写汉语的局限性

- 汉语的一些无动句、流水句、连动句、紧缩、动补、省略等结构，无法或不必用一个统率全句的模式来描述，其中连动句和兼语句尤为突出。
- 例如:
  - **(1)** 他拿了书就上楼去了。
  - **(2)** 我们选他当班长。

3 语义知识的应用

- 帮助判断一个形式是否合语法；
- 帮助做句法结构分析（消歧）；
- 词义消岐（Word Sense Disambiguation）。

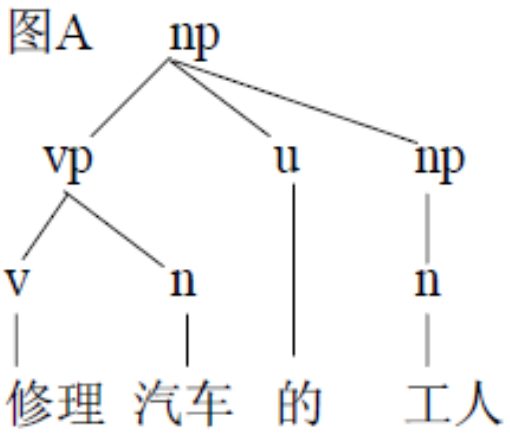
3.1 帮助判断句子是否合乎语法

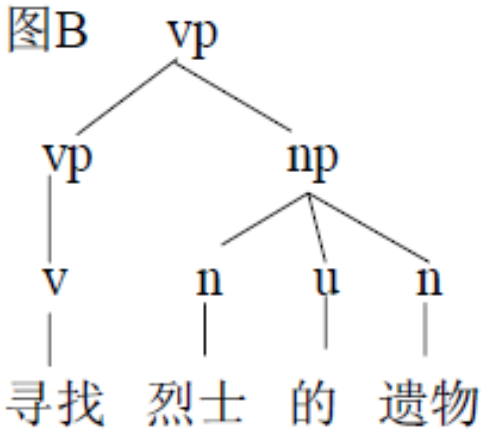
1a. 手枪比步枪更难使用。
1b. 手枪比武器更难使用。

2a. 这酒不香，这酒是醇香。
2b. 这酒不是香，这酒是醇香。

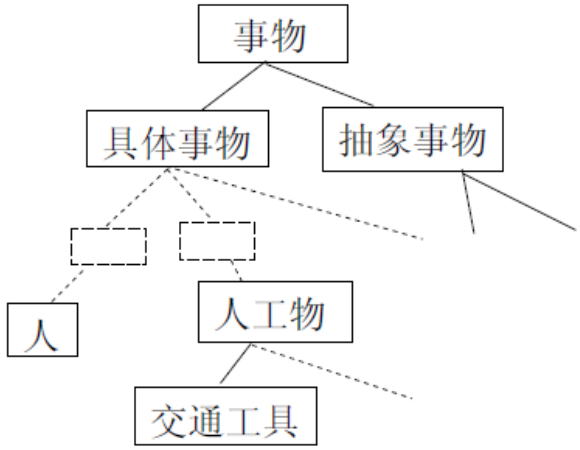
3.2 帮助做句法结构分析

A 修理汽车的工人　　B 寻找烈士的遗物　　v + n + 的 + n





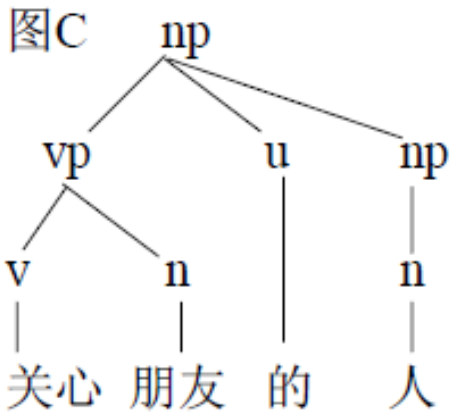
v + n[交通工具] + 的 + n[人]　　*v + n[人] + 的 + n[具体事物]*



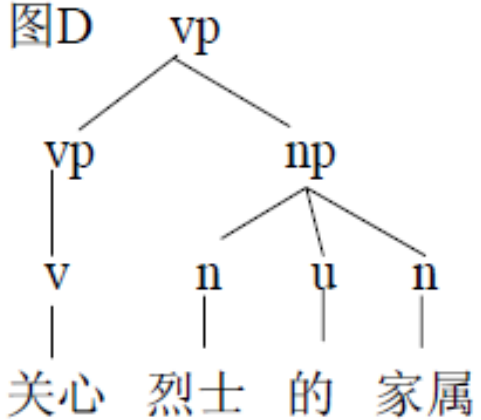
修理 {[施事:人][受事:人工物]}
寻找 {[施事:人][受事:具体事物]}
工人 [语义类:人]
汽车 [语义类:交通工具]
遗物 [语义类:具体事物]
烈士 [语义类:人]

帮助做句法结构分析

C. 关心 朋友 的 人



D. 关心 烈士 的 家属



v + n + 的 + n

v + 人 + 的 + 人

v+人[+亲属关系]+的+人　　v+人+的+人[+亲属关系]

烈士 [语义类：人 / 配价数：0]

朋友 [语义类 ： 人 / 语义特征： 亲属关系 / 配价数 ： 1]

家属 [语义类 ： 人 / 语义特征： 亲属关系 / 配价数 ： 1]

- 从词类到语义类到语义特征，是不断对结构模式进行细化的一个过程

3.3 词义消歧

词义消歧问题 (word sense disambiguation, WSD)

例：

英文： bank: 银行/ 河岸 ； plant: 工厂/ 植物

汉语：打： play/ take/ dial/ weave …

基本方法

- 基于词典信息的消歧方法
- 统计机器学习消歧方法
  - 有监督学习方法
    - **基于互信息的消歧方法**
    - 基于贝叶斯分类器的词义消歧方法
    - 基于最大熵的词义消歧方法
  - 无监督学习方法
  - ➢ 基本思路：一个词的不同语义一般发生在不同的上下文中

基于词典的词义消歧方法

(1)基于语义定义的消歧

- 基本思想：词典中词条本身的定义作为判断其语义的条件
- 例如，cone 在词典中有两个定义：一个是指 “松树的球果”，另一个是指“用于盛放其他东西的锥形物，比如，盛放冰激凌的锥形薄饼”。
- 如果在文本中，“树(tree)” 或者 “冰 (ice)与cone出现在相同的上下文中，那么，cone的语义就可以确定了，tree对应cone的语 义1，ice 对应cone的语义2。

基于词典的词义消歧方法

(2)基于义类辞典的消歧

- 基本思想：多义词的不同义项在使用时往往具有不 同的上下文语义类，即通过上下文的语义范畴可以判断多义词的使用义项。
- 如 *crane* 的两个词义“鹤”和“起重机”分别属于语义类“ANIMAL”和“MACHINERY”。不同的语义类往往具有不同的上下文环境，如：经常 表示“ANIMAL”语义类的共现词语为“species、 family、eat”等，而表示“MACHINE”语义类的共 现词语则为“tool、engine、blade”等。因此，只要 确定多义词的上下文词的义类范畴，就确定了多义词的词义。

基于词典的词义消歧方法

(3)基于双语词典的消歧

- 基本思想：需要消歧的语言称为第一语言，把需要借助的另一种语言称为第二语言。建立多义词***x*** 与相关词 ***y*** 之间的搭配关系；然后，在第二种语言的语料库中统计对应 ***x*** 不同词义的翻译与相关词***y***的翻译同现的次数，同现次数高的搭配对应的义项即为消歧后的词义。
- 例如：单词 *plant* 有两个含义：“植物” 和 “工厂”。
- 当对 *plant* 进行词义消歧时，需要首先识别出含有 *plant* 的短语，如：*manufacturing plant*；然后，在汉语语料库中搜索与这个短语对应的汉语短语实例。由于*manufacturing* 的汉语翻译“制造”只和“工厂”共现，因此，可以确定在这个短语中*plant*的词义为“工厂”。而短语 *plant life* 在汉语翻译中，“生命(*life*)”与“植物”共现的机会更多，因此，可以确定在短语 *plant life* 中*plant* 的词义为“植物”。

基于统计的消歧方法

总体思路：通过建立分类器，利用划分多义词的上下文类别的方法来区分多义词的词义。

基于互信息的消歧方法 (Brown *et al.*, 1991)

- 基本思想：假设我们有一个双语对齐的平行语料库，以法语和英语为例，通过词语对齐模型每个法语单词可以找到对应的英语单词，一个多义的法语单词在不同的上下文中对应多种不同的英语翻译。

例子：

- *Prendre une mesure* -> to take a measure
- *prendre une décision* -> to make a decision

- 法语动词 ***prendre*** 可以被翻译成 **to take**，也可以被翻译成 **to make**，这取决于它所带的宾语是 ***mesure*** 还是 ***décision***。

基于统计的消歧方法

- 可以把一个多义的法语单词的英语译词看作是这个法语单词的语义解释，而决定法语多义词语义的条件看作是语义指示器**(indicator)**，如：前面例子中法语单词***prendre***所带的宾语。因此，只要我们知道了多义词的语义指示器，也就确定了该词在特定上下文中的语义。这样，多义词的词义消歧问题就变成了语义指示器的分类问题。

- 假设 $T_1, T_2, \cdots, T_m$ 是多义法语词的翻译**(**或语义**)**， $V_1, V_2, \cdots, V_n$ 是指示器可能的取值。

基于统计的消歧方法

- 利用 **Flip-Flop** 算法来解决指示器分类问题**(**假设多义法语词只有两个语义**)**：

1. 随机地将 $T_1, T_2, \cdots, T_m$ 划分为两个集合$P = \{P_1, P_2\}$;
2. 执行如下循环：
   a) 找到 $V_1, V_2, \cdots, V_n$ 的一种划分$Q = \{Q_1, Q_2\}$，使 Q_i与 P_i 之间的互信息最大;
   b) 找到一种改进的划分 P'，使P'与Q的互信息最大。

互信息定义：$I(P,Q) = \sum_{x \in P} \sum_{y \in Q} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$

算法终止的条件是互信息$I(P,Q)$不再增加或增加甚少

基于统计的消歧方法

- 一旦指示器的取值划分确定了，词义消歧就变成了如下简单的过程：
  1. 对于出现的歧义词确定其指示器值V_i；
  2. 如果V_i在Q_1中，指定该歧义词的语义为语 义**1**，如果在 Q_2 中，指定其语义为语义**2**。

如果一个词有多个歧义的话，扩展算法请见

- Peter F. Brown, Stephen A. Della Pietra et al., A Statistical Approach to Sense Disambiguation in Machine Translation, *Proc. DARPA Workshop on Speech and Natural Language*, 1991, pp 146－151.

小结

- 语义知识跟句法知识的差别主要在于知识颗粒度的不同。从某种意义上说，语义知识就是细化了的句法知识。

- 跟语法范畴的提取是基于聚合和组合两种基本关系一样，人们也是从这两个角度出发去获取语义知识的。并且语义知识的形式化表达也可采用特征结构及合一的形式。