第11章 语义分析

宗成庆
中国科学院自动化研究所
cqzong@nlpr.ia.ac.cn

本章内容

1. 概述
2. 语义网络
3. 词义消歧
4. 语义角色标注
5. 习题
6. 附录：语义理论和格语法

1. 概述

◆语义计算的任务

解释自然语言的句子或篇章各部分(词、词组、句子、段落、篇章)的含义。

◆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、在不同的语用场景下可能有不同的理解；
- 人脑对语义理解的认知过程尚不清楚，语义计算的理论、方法和模型尚未建立。

1. 概述

◆例子

(1) I bought a car with four wheels.

I bought a car with four dollars.

(2) These boys will be dedicated persons.

These boys will be denied license.

(3) 这件事情让我感到很头疼。

(4) 这个人真恶心！

(5) 他这种人也算个男人？！

(6) 简直是个饭桶！

1. 概述

①	有你的好果子吃！	没你的好果子吃！
②	中国队大胜美国队。	中国队大败美国队。
③	沙子进眼睛了。	眼睛进沙子了。
④	东西掉地上了。	东西掉地下了。
⑤	注意危险！	注意安全！
⑥	你做梦吧！	你别做梦了！
⑦	好热闹啊！	好不热闹啊！
⑧	一会儿他就来了。	不一会儿他就来了。
⑨	别乱丢烟头。	别乱丢烟屁股。
⑩	以前只想一个人。	现在只想一个人。
⑪	中国足球谁也打不过。	中国乒乓球谁也打不过。
⑫	夏天能穿多少穿多少，冬天能穿多少穿多少。	

1. 概述

◆挑战

(撇开歧义、隐喻等复杂问题)

- 基本的语义单元是什么？
- 语义表示的标准是什么？
- 语言产生和演化的神经基础是什么？
- 人脑的语言认知机理是什么？
- 词义如何组合产生新的概念？
- 分布式语义表示的合理性解释是什么？

……

本章内容

1. 概述
2. 语义网络
3. 词义消歧
4. 语义角色标注
5. 习题
6. 附录：语义理论和格语法

2. 语义网络

◆背景

语义网络(semantic network)是由美国心理学家 M. R. Quilian 于1968年在研究人类联想记忆时提出的。1977年美国学者 G. Hendrix 提出了分块语义网络的思想，把语义的逻辑表示与“格语法”结合起来，把复杂问题分解为几个较为简单的子问题，每个子问题用一个语义网络表示，把自然语言理解的研究向前推进了一步。

2. 语义网络

◆语义网络表示

语义网络通过由实体、概念或动作、状态及语义关系组成的有向图来表达知识、描述语义。

- **有向图：**图的结点表示实体或概念，图的边表示实体或概念之间的关系。
- **边的类型：**

(1)是一种抽象(IS-A)：A到B的边表示“A是B的一种特例”;

(2)是一部分(PART-OF)：A到B的边表示“A是B的一部分”;

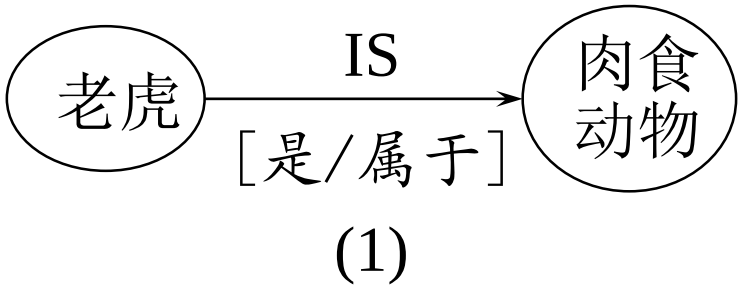
(3)是属性(IS)：A到B的边表示“A是B的一种属性”；

(4)拥有/占用(HAVE)：A到B的边表示“A拥有B”;

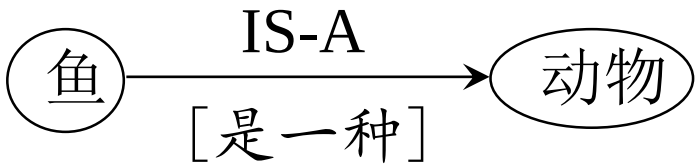
(5)次序在先/前(BEFORE)：A到B的边表示“A在B之前”;

……

2. 语义网络



(1)



(2)

(3)

(4)

(5)

(6)

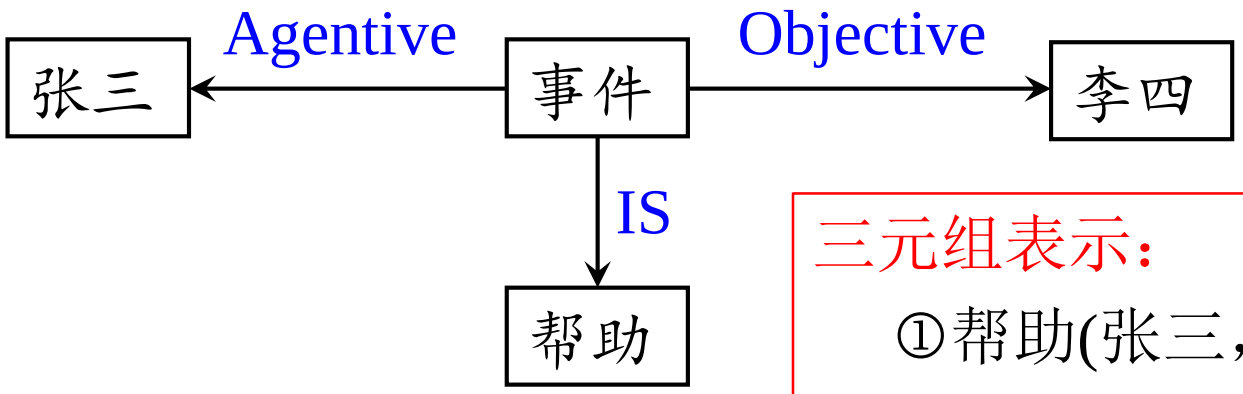
通常实体、概念或属性等采用不同形状的节点表示。

2. 语义网络

◆事件的语义网络表示

当语义网络表示事件时，结点之间的关系可以是施事、受事、时间等。这里所说的“事件”指某个具体的动作或状态，并非我们日常生活中所说的事件。

例如：张三帮助李四。



三元组表示：

①帮助(张三，李四)

或者：

②(张三，帮助，李四)

2. 语义网络

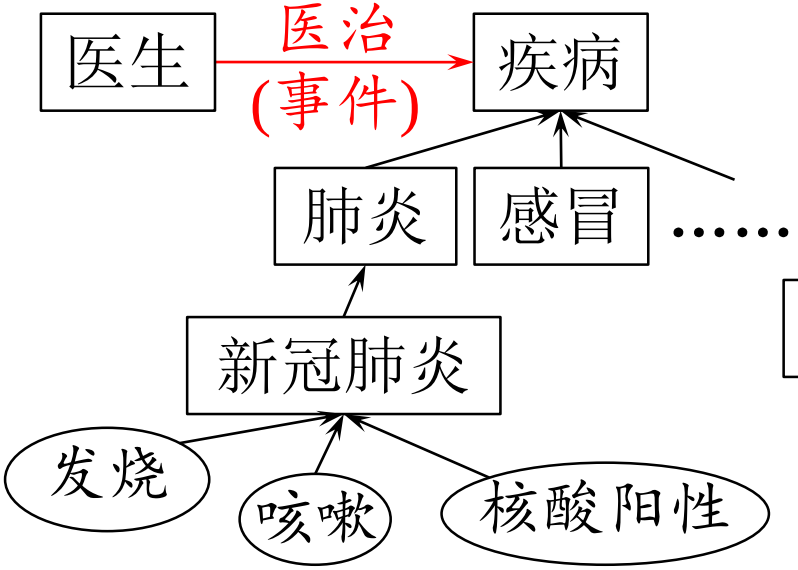
◆事件的语义表示

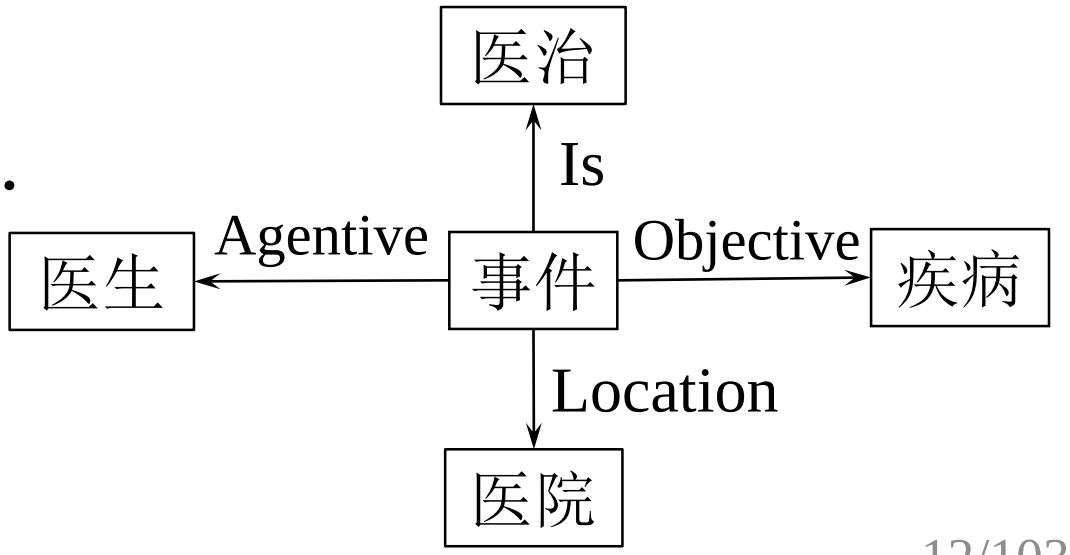
(1)分类关系：事物之间的类属关系。

(2)聚焦关系：多个下位概念构成一个上位概念。

(3)推论关系：由一个概念推出另一个概念。

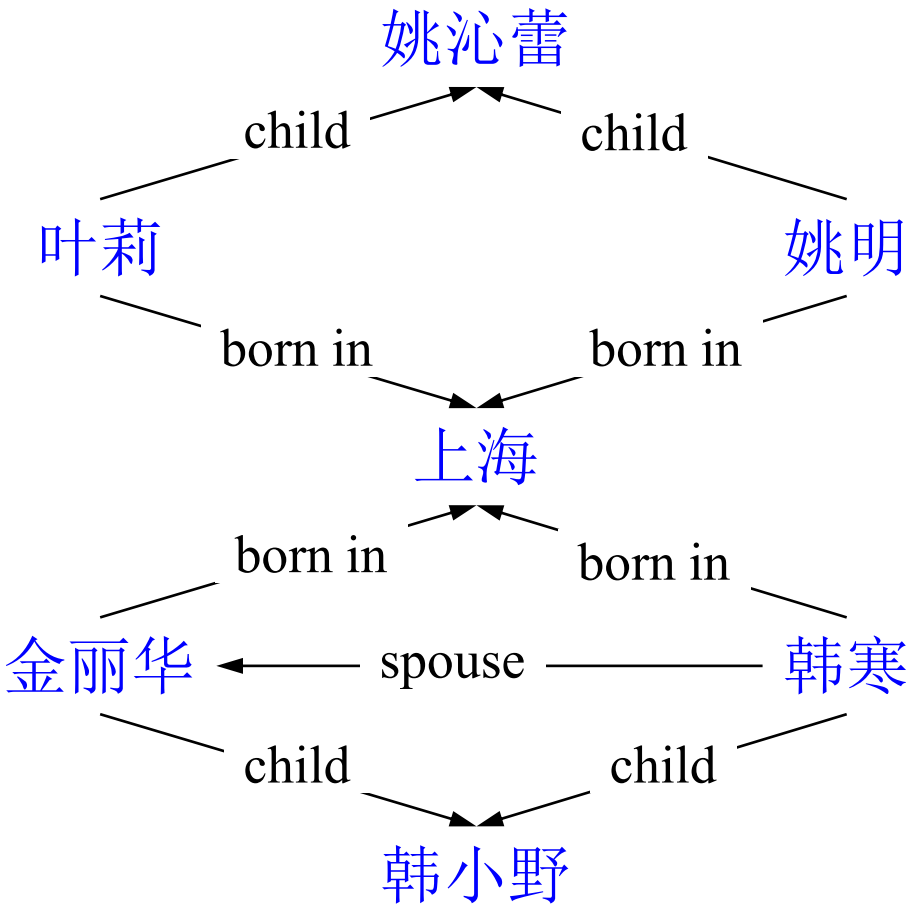
(4)时间、位置关系：事实发生或存在的时间、位置。

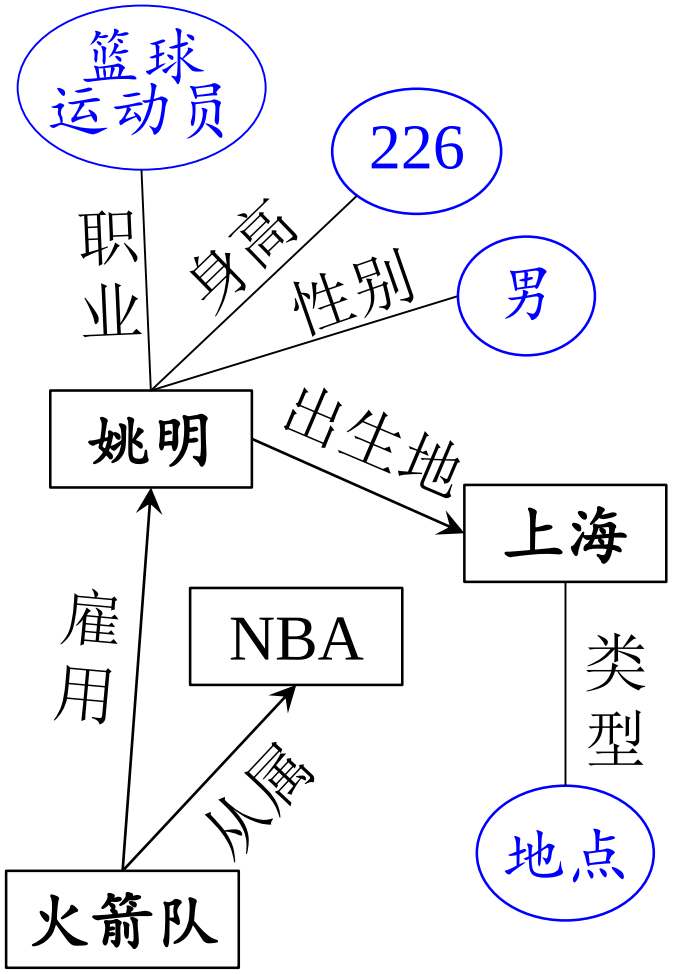




2. 语义网络

◆知识图谱


姚沁蕾
child
child
叶莉
姚明
born in
born in
上海
born in
born in
金丽华
spouse
韩寒
child
child
韩小野


篮球运动员
226
职业
身高
性别
男
姚明
出生地
上海
雇用
NBA
类型
从属
地点
火箭队

2. 语义网络

◆知网(HowNet) (http://www.keenage.com)

由董振东教授创立。

董振东教授曾在黑龙江大学担任英语老师，军事科学院研究员、机器翻译研究组组长，中国软件公司语言工程实验室主任，新加坡国立大学系统科学研究院 研究员。自上世纪七十年代开始从事机器翻译研究，1987年成功开发了我国第一个商品化机器翻译系统原型“科译1号”。1980年代研究创立知网。

董振东
(1937.4 – 2019.2.28)

2. 语义网络

- 知网的4个基本观点：

(1)NLP系统最终需要更强大的知识库支持。

(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。

(3)知识库建设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。

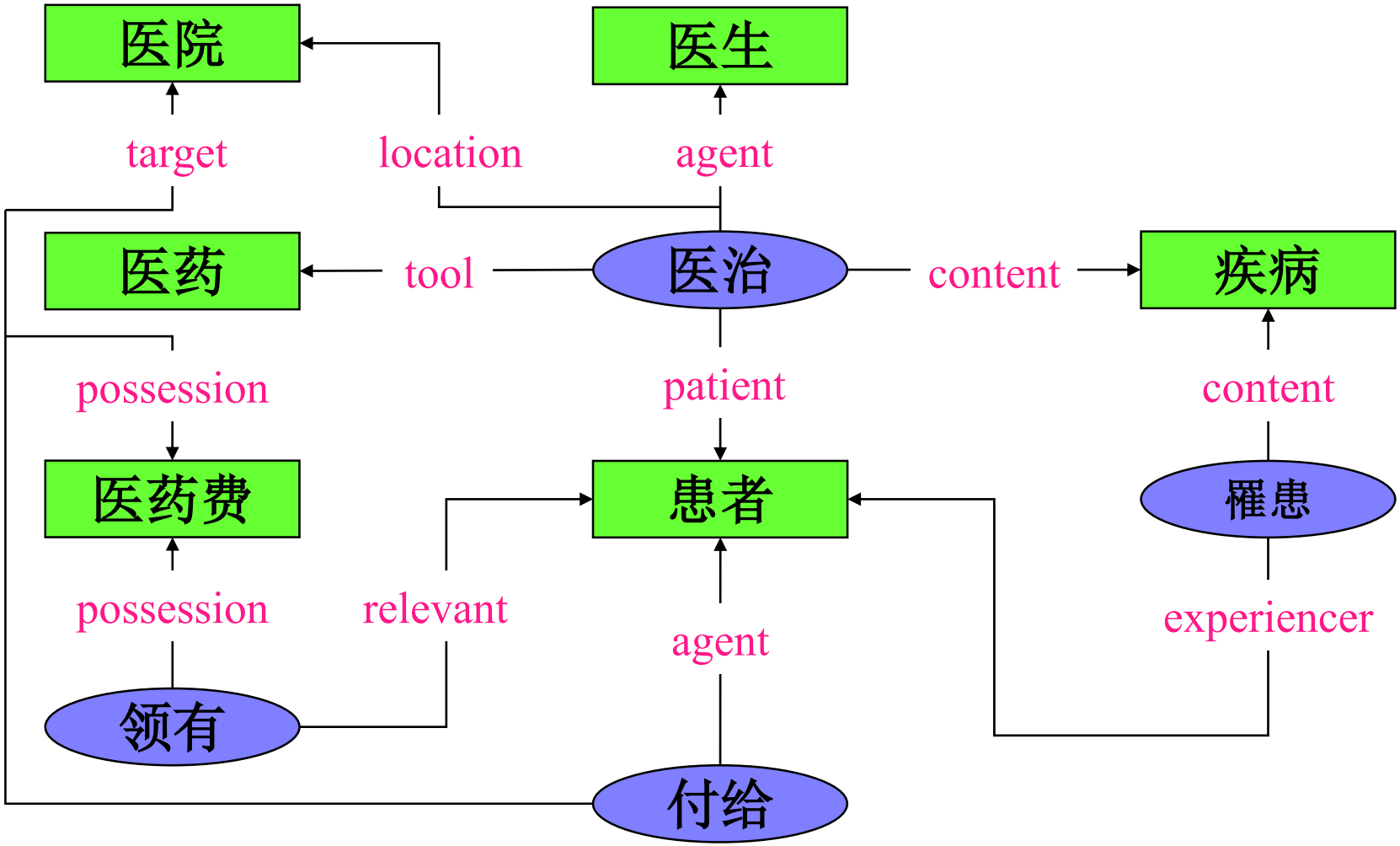
2. 语义网络

- 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。

2. 语义网络


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

2. 语义网络

◆概念层次网络

由中科院声学所黄曾阳研究员提出。

1935 -，湖北蕲春人

黄曾阳教授于1958年毕业于北京大学物理系，长期从事水声学研究和信号处理工作，现主要研究领域是自然语言理解。他兼有中国传统语言学、物理学、信号处理、哲学等方面的功底，经潜心研究创立了面向自然语言理解的概念层次网络理论，简称HNC(Hierarchical Network of Concepts)。

2. 语义网络

◆统一语言学

由高庆狮院士提出。

1934.7.17- 2011.5.15
福建厦门人

高庆狮院士于1953年以数学100分、物理97分的优异成绩考入北京大学数学力学系，1957年毕业后进入中科院计算所任工作，至1994年。1965-1970中国第一颗人造卫星地面计算控制中心早期设计负责人之一，成为两弹一星任务中我国第一台具有分时中断系统的晶体管计算机—“功勋计算机”的体系结构设计负责人。

1973年5月提出了实现向量语言的纵横加工向量机和亿次/秒巨型机的设计方案，并研制了我国第一台向量机。1980年代开始从事机器翻译、人类智能及其模拟和应用以及网络安全研究。

1980年当选中国科学院学部委员（院士）。

2. 语义网络

◆问题

词义
- 内涵：词本身的意义，是对词代表的概念描述。
- 外延：词所指代的物体。

- 如何在语义网络中表示和区分词的内涵和外延？
- 如何确定语义网络的完备性？
- 如何确定概念、实体、关系划分的粒度？
- 知识图谱与预训练语言模型的关系是什么？
- 知识图谱如何用于端到端的模型？

……

本章内容

1. 概述
2. 语义网络
3. 词义消歧
4. 语义角色标注
5. 习题
6. 附录：语义理论和格语法

3. 词义消歧

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

例如：

英文： bank: 银行/ 河岸

plant: 工厂/ 植物

汉语： 打：play/ take/ dial/ weave …

包：package/ guarantee / …

3. 词义消歧

◆基本方法

●早期基于规则的消歧方法

●统计机器学习消歧方法

➢有监督学习方法

➢无监督学习方法

基本思路：不同词义一般发生在不同的上下文中。

●基于词典信息的消歧方法

3. 词义消歧

◆有监督的词义消歧方法

- 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。

(1)基于互信息的消歧方法 (Brown *et al.*, 1991)

- 基本思想：

假设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。

3. 词义消歧

例子：

(a) *prendre une mesure* → to take a measure

(b) *prendre une décision* → to make a decision

也就是说，法语词 *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*。

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，而决定法语词语义的条件看作是语义指示器(indicator)，如该例子中法语单词 *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。

3. 词义消歧

假设 $T_1, T_2, \ldots, T_m$是一个多义法语词的英语译文(或语义)，$V_1, V_2, \ldots, V_n$ 是指示器可能的取值。利用 Flip-Flop 算法来解决指示器分类问题(假设法语词只有两个语义)：

(1) 随机地将 $T_1, T_2, \ldots, T_m$ 划分为两个集合 $P=\{P_1, P_2\}$

(2) 执行如下循环：

(*a*) 找到 $V_1, V_2, \ldots, V_n$ 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。

3. 词义消歧

一旦指示器的取值划分确定了，词义消解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。

如果法语词有多个歧义，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.

3. 词义消歧

(2) Yarowsky 消歧算法

基本思路：分别处理每个出现的歧义词，对所有的歧义词有两个基本限制：

- **每篇文本只有一个意义**：在任意给定的文本中，目标词的词义具有高度的一致性；
- **每个搭配只有一个意义**：目标词和周围词之间的相对距离、词序和句法关系，为目标词的意义提供了很强的一致性的词义消歧线索。

基于贝叶斯分类器的词义消歧方法

3. 词义消歧

Yarowsky 消歧算法：

(a)对于第一个约束，如果一个给定的多义词第一次出现时使用某个义项，那么，它在后面出现时也很有可能使用这个义项。

(b)对于第二个约束，采用基于自举(bootstrapping) 的(半监督)学习方法。搭配特征依据如下比率排序：

$$\frac{p(s_{k_1} \mid f)}{p(s_{k_2} \mid f)}$$

两个义项与特征同现的次数之比。

其中，s_{k_i} 为词义，f 为搭配特征。

3. 词义消歧

(3)基于最大熵的词义消歧方法

参见本课程讲义第2章。

3. 词义消歧

◆无监督的词义消歧方法

无监督的词义消歧方法的通常做法是，对于一个具有k个义项的词w，估计使用义项$s_i(k \geq i \geq 1)$ 的上下文中出现词v_j 的概率，即 $p(v_j|s_i)$。只是在该方法中参数 $p(v_j|s_i)$的估计不是根据有标注的训练语料，而是在无标注的语料上进行，开始时随机地初始化参数，然后根据EM算法重新估计该概率值。

主要问题在于：很多同义词的同一个意义出现的上下文往往有很大的差异，因此，很难保证同一个意义的上下文被划分到同一个等价类中。

3. 词义消歧

◆基于词典信息的消歧方法

(1) 基于词典中的词条解释进行消歧(单语言)

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，如盛放冰激凌的锥形薄饼”。如果在文本中，“树(tree)”或者“冰(ice)”分别与cone同现时，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。

3. 词义消歧

(2) 基于双语词典的消歧

基本思想：待消歧的语言称为第一语言，需借助的语言称为第二语言。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”。对plant 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”。

3. 词义消歧

(3) 基于义类辞典(thesaurus)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。

例如，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语境，如经常表示“ANIMAL”语义类的共现词为“species, family, eat”等，而表示“MACHINE”语义类的共现词则为“tool, engine, blade”等。因此只要确定多义词的上下文词的义类范畴，就可确定多义词的词义。

3. 词义消歧

- 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)

3. 词义消歧

- **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$\{x_1, x_2, \ldots\}$ 和$\{y_1, y_2, \ldots\}$之间有一种语义关系R，则在$\{y_1, y_2, \ldots\}$和$\{x_1, x_2, \ldots\}$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。

3. 词义消歧

- 4 种语义关系：
  - 同义关系(synonymy)
  - 反义关系(antonymy)
  - 上下位关系(hypernym/ 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(meronym)或称部分/整体关系。

3. 词义消歧

- 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}

3. 词义消歧

- 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。

3. 词义消歧

➢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：

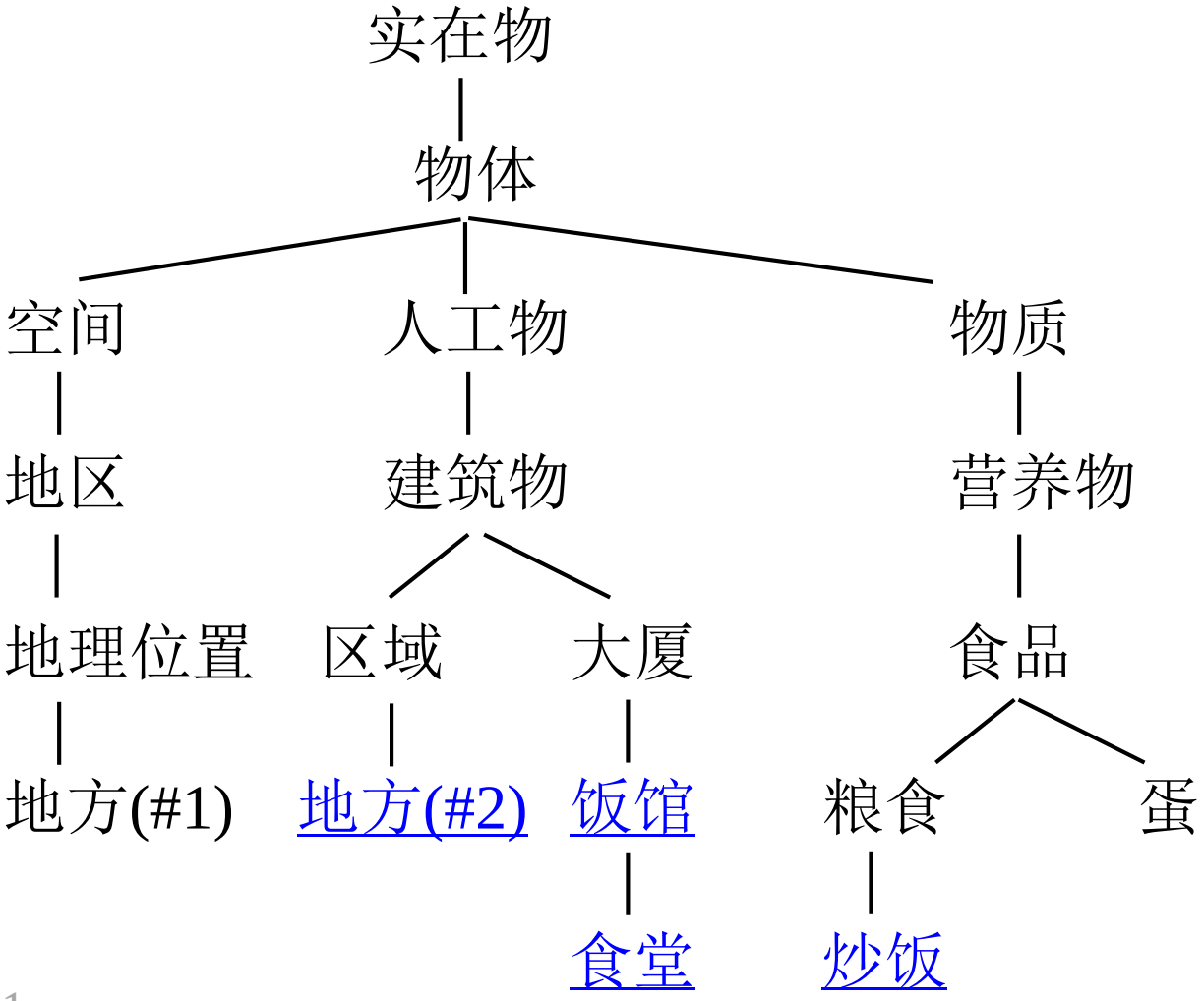
#指地理位置　　如：在祖国各个地方

#指空间　　如：没地方

#指部分　　如：他说的有些地方不对

3. 词义消歧

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：



本章内容

1. 概述
2. 语义网络
3. 词义消歧
4. 语义角色标注
5. 习题
6. 附录：语义理论和格语法

4. 语义角色标注

◆语义角色标注任务

自动语义角色标注(semantic role labeling, SRL)是NLP领域研究的热点之一，其基本任务是以句子为分析单位，以句子中的谓词为核心，分析句子中的其他成分与谓词之间的关系。如：

[他们]$_{\text{Agent}}$ [昨天]$_{\text{Time}}$ [在北京]$_{\text{Location}}$ [讨论]$_{\text{Pred.}}$ 了 [方案]$_{\text{Patient}}$。

语义角色标注一般是在句法分析的基础上进行的。

4. 语义角色标注

◆用于SRL的主要资源

- LDC命题库(PropBank)
- LDC名词命题库(NomBank)
- 框架网(FrameNet)

4. 语义角色标注

(1) 动词命题库(Proposition Bank, PropBank)

起初是在 Upenn 英语树库(English Treebank)的基础上增加语义信息后构建的“命题库”，其基本观点认为：树库仅提供句子的句法结构信息，对于计算机理解人类语言是不够的。因此，PropBank 的目标是对原树库中的句法节点标注上特定的论元标记 (argument label)，使其保持语义角色的相似性。

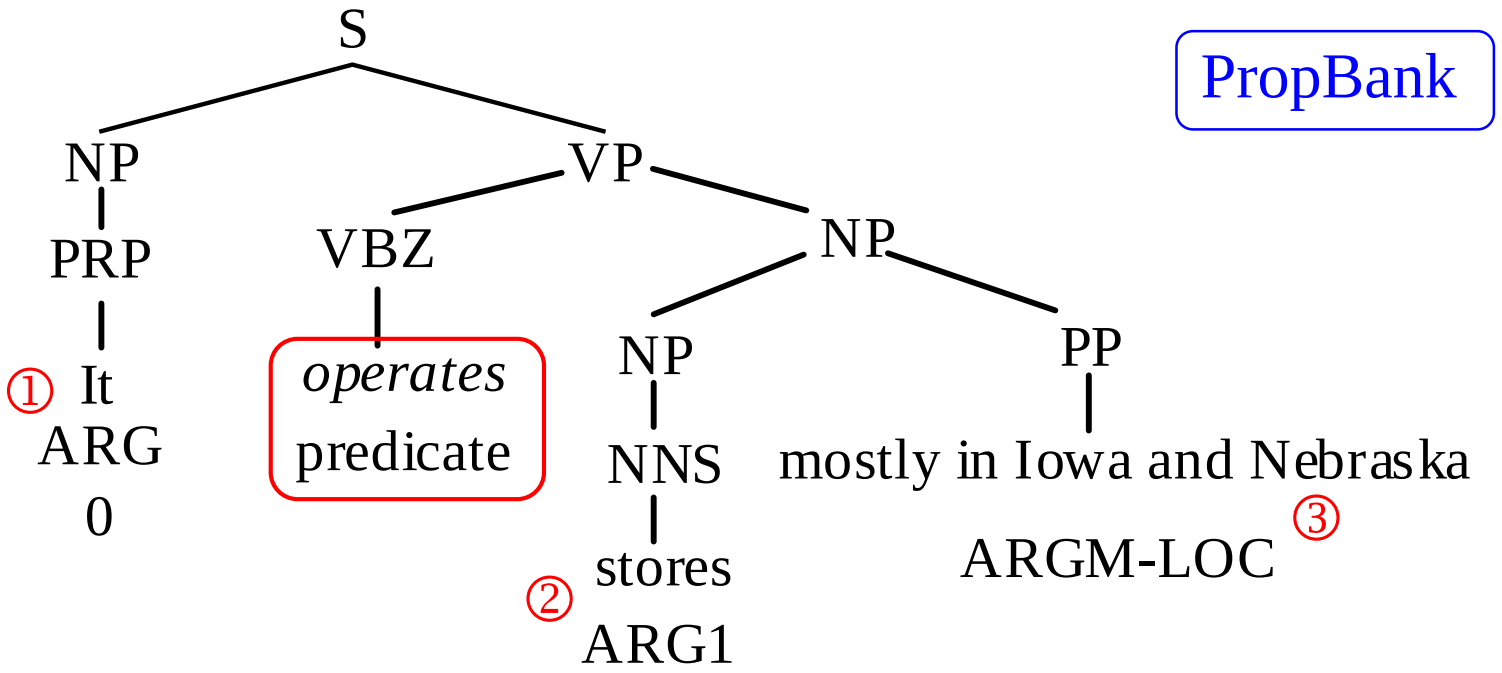
4. 语义角色标注

例如，*John broke the window*.

- 事件是 “打碎 (breaking event)”
- *John* 为事件的 制造者 (instigator)
- *window*为 受事者 (patient)
- 窗户被打碎 (broken window)为事件的结果

4. 语义角色标注

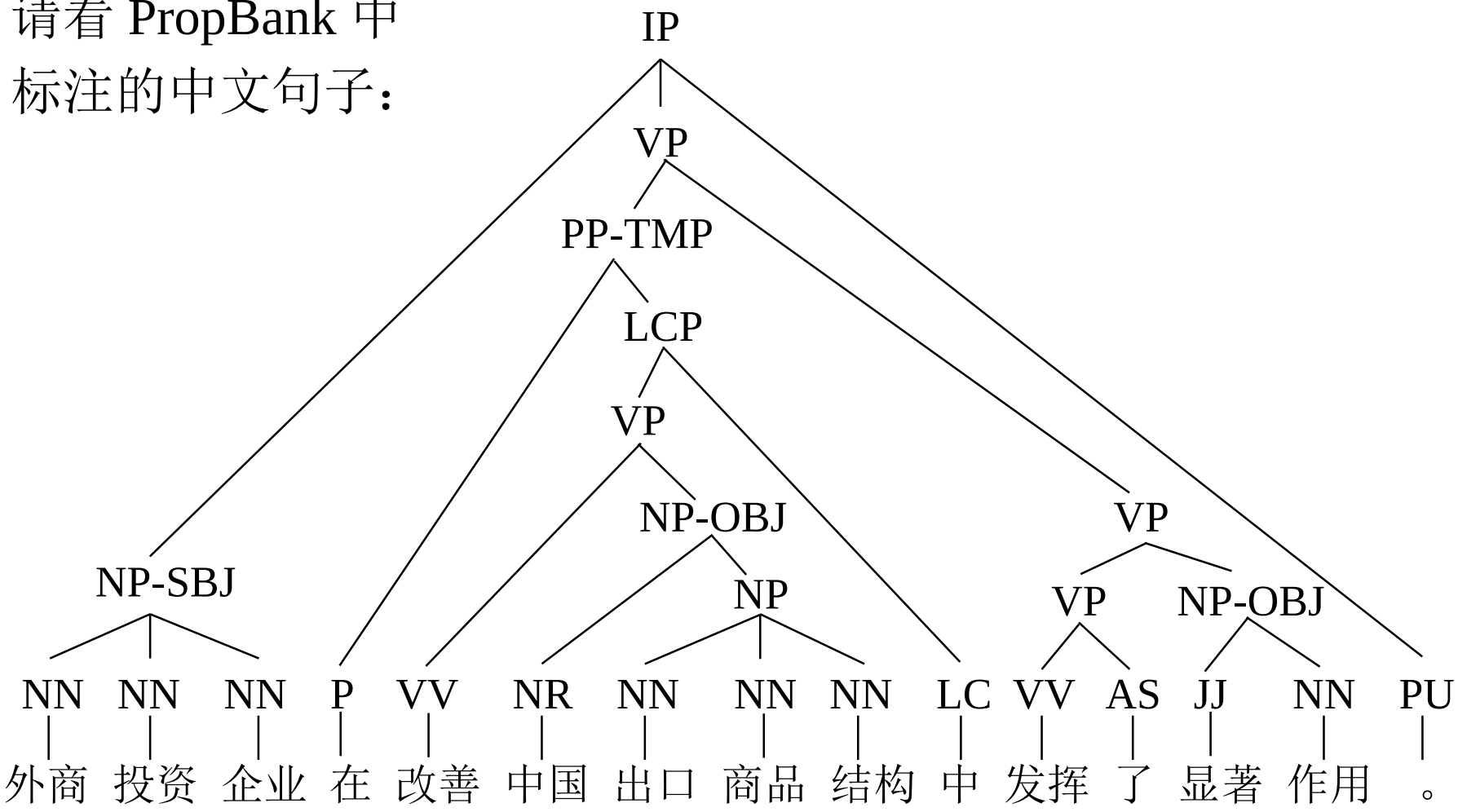
例：It operates stores mostly in Iowa and Nebraska.



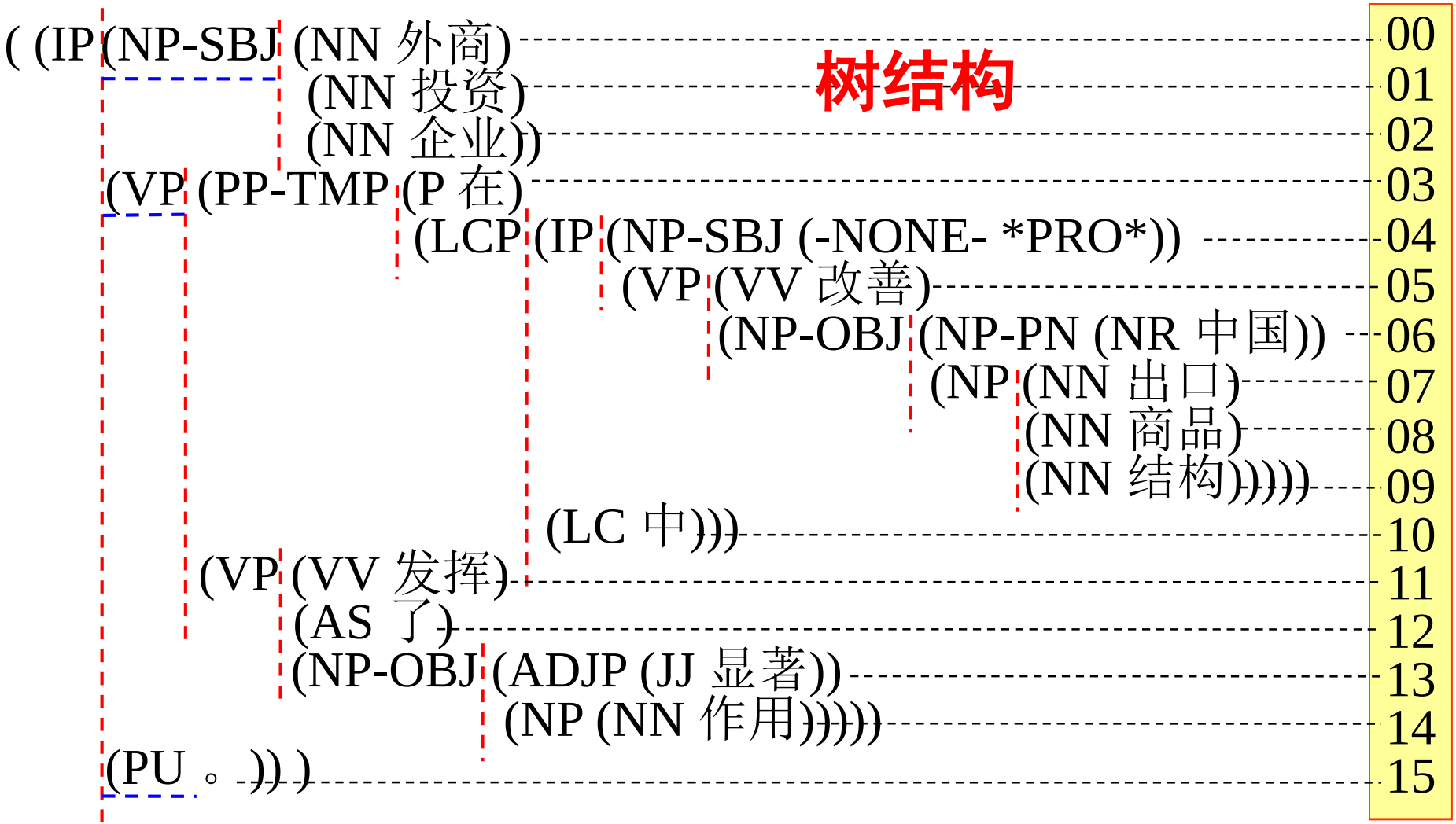
谓词是operates，它有3个论元：It (ARG0)、stores (ARG1)、mostly in Iowa and Nebraska (ARGM-LOC)，这三个语义角色都处在句法树的某个节点上。

4. 语义角色标注

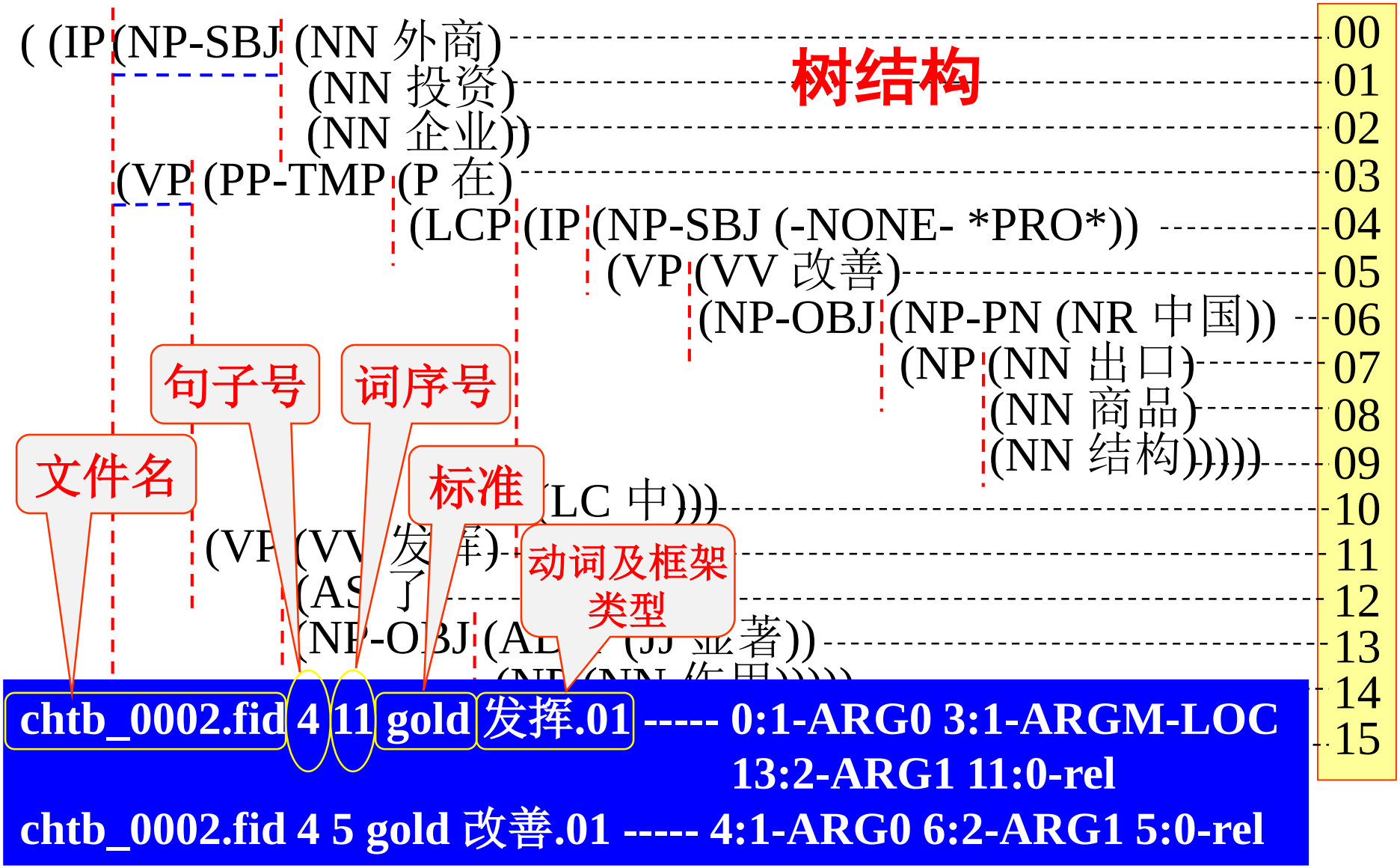
请看 PropBank 中标注的中文句子：



4. 语义角色标注


((IP (NP-SBJ (NN 外商)
(NN 投资)
(NN 企业))
(VP (PP-TMP (P 在)
(LCP (IP (NP-SBJ (-NONE- *PRO*))
(VP (VV 改善)
(NP-OBJ (NP-PN (NR 中国))
(NP (NN 出口)
(NN 商品)
(NN 结构)))))
(LC 中)))
(VP (VV 发挥)
(AS 了)
(NP-OBJ (ADJP (JJ 显著))
(NP (NN 作用)))))
(PU 。)))
00
01
02
03
04
05
06
07
08
09
10
11
12
13
14
15
树结构

4. 语义角色标注


((IP (NP-SBJ (NN 外商)
(NN 投资)
(NN 企业))
(VP (PP-TMP (P 在)
(LCP (IP (NP-SBJ (-NONE- *PRO*))
(VP (VV 改善)
(NP-OBJ (NP-PN (NR 中国))
(NP (NN 出口)
(NN 商品)
(NN 结构)))))
(LC 中)))
(VP (VV 发挥)
(AS 了
(NP-OBJ (AD
(JJ 显著))
00
01
02
03
04
05
06
07
08
09
10
11
12
13
14
15
树结构
句子号
词序号
文件名
标准
动词及框架类型
chtb_0002.fid 4 11 gold 发挥.01 ----- 0:1-ARG0 3:1-ARGM-LOC 13:2-ARG1 11:0-rel
chtb_0002.fid 4 5 gold 改善.01 ----- 4:1-ARG0 6:2-ARG1 5:0-rel

4. 语义角色标注

Frameset: f1

ARG0: *agent*

ARG1: *influence, utility, specialty, etc.*

例句：今年中国银行仍将继续发挥其在支持外商投资企业方面的主渠道作用。

ARG0: 中国银行

ARG1: 其在 *PRO* 支持外商投资企业方面的主渠道作用

ARGM-TMP: 今年

ARGM-ADV: 仍

ARGM-ADV: 将

REL: 发挥

4. 语义角色标注

Frameset: f2

ARG0: *exerter*

ARG1: *potential, influence, utility, etc.*

例句：马尔默队的马·尼尔松和斯·尼尔松都发挥出了较高的水平，……

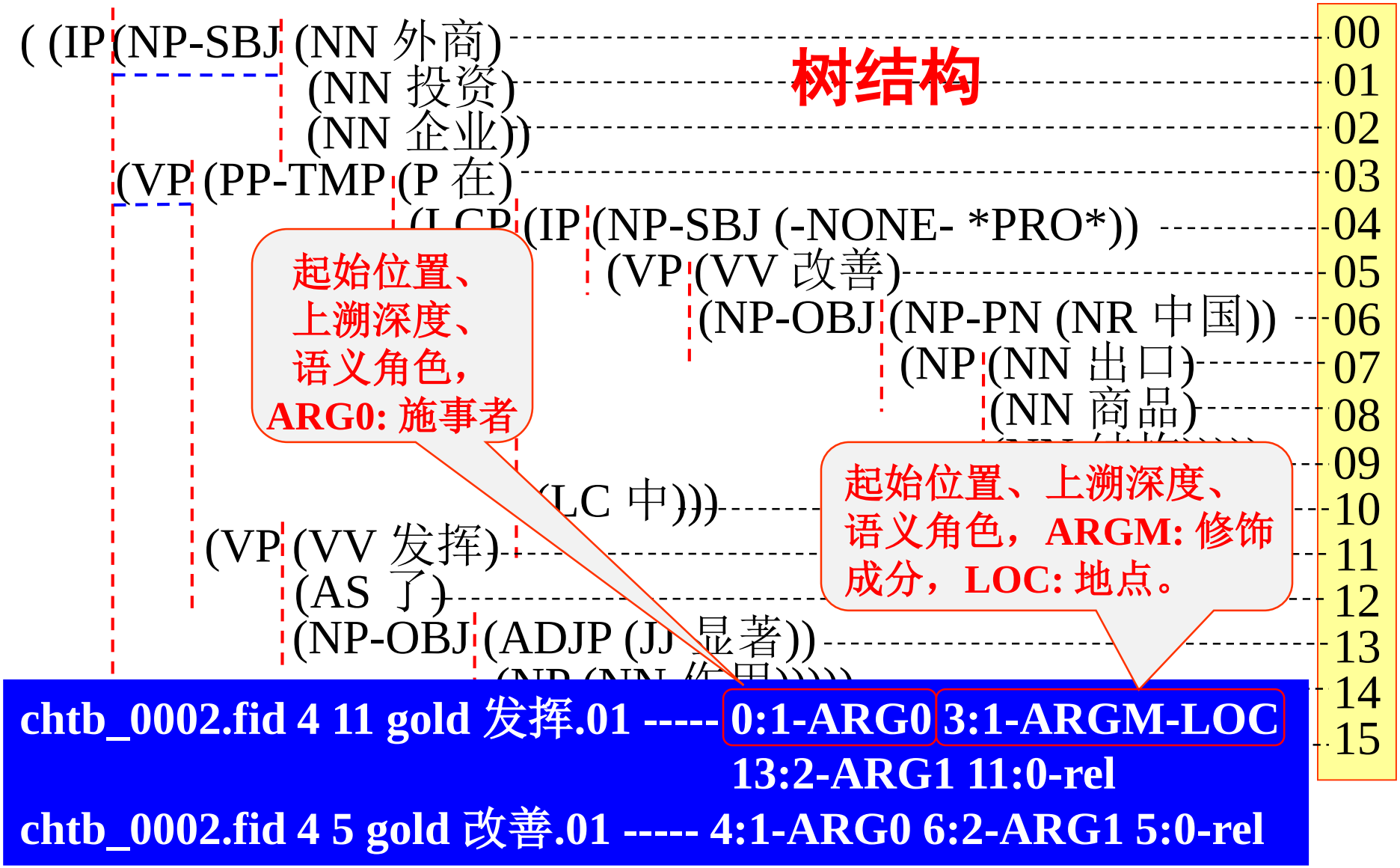
ARG0: 马尔默队的马·尼尔松和斯·尼尔松

ARG1: *OP* *T*-1 较高的水平

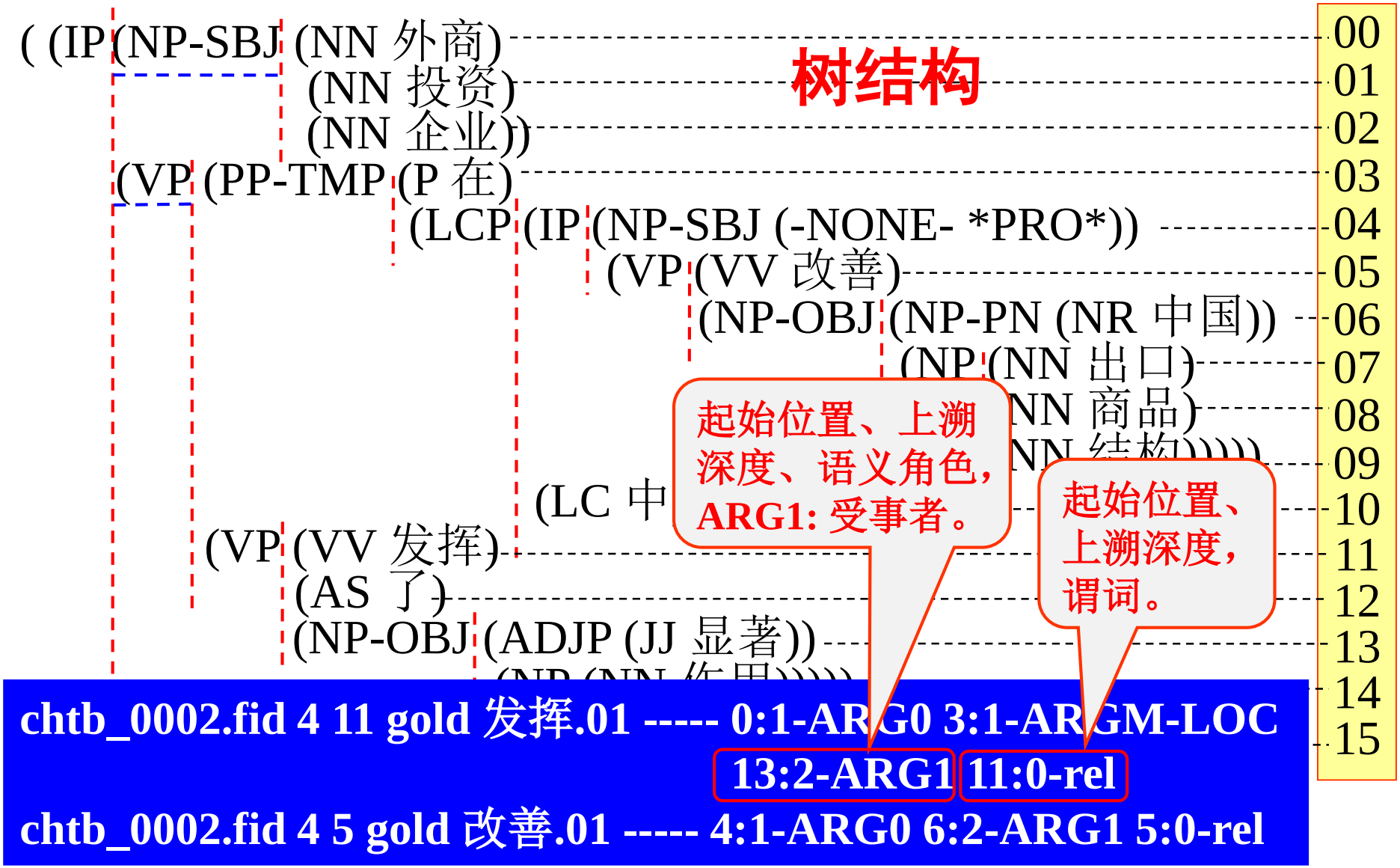
ARGM-ADV: 都

REL: 发挥

4. 语义角色标注


((IP (NP-SBJ (NN 外商)
(NN 投资)
(NN 企业))
(VP (PP-TMP (P 在)
(LCP (IP (NP-SBJ (-NONE- *PRO*))
(VP (VV 改善)
(NP-OBJ (NP-PN (NR 中国))
(NP (NN 出口)
(NN 商品)
(LC 中)))
(VP (VV 发挥)
(AS 了)
(NP-OBJ (ADJP (JJ 显著))
00
01
02
03
04
05
06
07
08
09
10
11
12
13
14
15
树结构
起始位置、上溯深度、语义角色，ARG0: 施事者
起始位置、上溯深度、语义角色，ARGM: 修饰成分，LOC: 地点。
chtb_0002.fid 4 11 gold 发挥.01 ----- 0:1-ARG0 3:1-ARGM-LOC 13:2-ARG1 11:0-rel
chtb_0002.fid 4 5 gold 改善.01 ----- 4:1-ARG0 6:2-ARG1 5:0-rel

4. 语义角色标注


((IP (NP-SBJ (NN 外商)
(NN 投资)
(NN 企业))
(VP (PP-TMP (P 在)
(LCP (IP (NP-SBJ (-NONE- *PRO*))
(VP (VV 改善)
(NP-OBJ (NP-PN (NR 中国))
(NP (NN 出口
NN 商品)
NN 结构))))
(LC 中
(VP (VV 发挥)
(AS 了)
(NP-OBJ (ADJP (JJ 显著))
00
01
02
03
04
05
06
07
08
09
10
11
12
13
14
15
树结构
起始位置、上溯深度、语义角色，ARG1: 受事者。
起始位置、上溯深度，谓词。
chtb_0002.fid 4 11 gold 发挥.01 ----- 0:1-ARG0 3:1-ARGM-LOC 13:2-ARG1 11:0-rel
chtb_0002.fid 4 5 gold 改善.01 ----- 4:1-ARG0 6:2-ARG1 5:0-rel

4. 语义角色标注

(2)名称命题库(Nominalization Bank, NomBank)

- NomBank 是 PropBank 的孪生项目，它和 PropBank 标注的都是同一批树库，区别在于NomBank 标注的是树库中名词的词义和相关的论元信息。
- The goal is to annotate each “markable” NP, marking the head, its arguments and “allowed” adjuncts in the style of PropBank.
- 英文命题库将宾州树库(Treebank)中的《华尔街日报》语料和一部分布朗语料(Brown Corpus)进行了人工的语义角色标注。
- 汉语命题库则源自新华社语料。

4. 语义角色标注

◆SRL的两类语义角色

- 与具体谓词直接相关的，这些角色用ARG0，ARG1，…，ARG5表示，如ARG0 通常表示动作的施事者，ARG1表示动作的影响(受事者/宾语(object))等，ARG2-ARG5 对于不同的谓语动词会有不同的语义含义；
- 起修饰作用的辅助性角色，其角色标签都以ARGM开头，常见的有表示时间的角色ARGM-TMP，表示地理位置的角色ARGM-LOC，表示一般性修饰成分的角色 ARGM-ADV等。

4. 语义角色标注

中文语义角色标注的语料库主要有Chinese PropBank (CPB)和Chinese NomBank。它们都是在中文树库(CTB)的句法成分中加入了人工标注的语义角色信息，也把语义角色分为两类：①核心语义角色ARG0，ARG1，…，ARG4，如ARG0表示动作的施事者，ARG1表示受事者；②起修饰作用的辅助性角色，其角色标签都以ARGM开头，如ARGM-TMP表示时间，ARGM-LOC表示地点等。

➢ **CPB 3.0:** https://catalog.ldc.upenn.edu/LDC2013T13

语义角色	角色描述
ARG0	施事者
ARG1	受事者
ARG2	范围或程度
ARG3	动作起点
ARG4	动作结束点
ARGM-ADV	状语
ARGM-BNF	受益者
ARGM-CND	条件
ARGM-DIR	方向
ARGM-LOC	地点
ARGM-MNR	方式
ARM-PRP	目的
ARGM-TMP	时间
ARGM-TPC	主题
ARGM-PRD	次谓词

4. 语义角色标注

例(1): Carl Bernstein's book about Watergate

REL = book (RELATIONAL_NOUN)

ARG0 = Carl Bernstein's

ARG1 = about Watergate

例(2): students' knowledge of two-letter consonant sounds

ARG0 = students

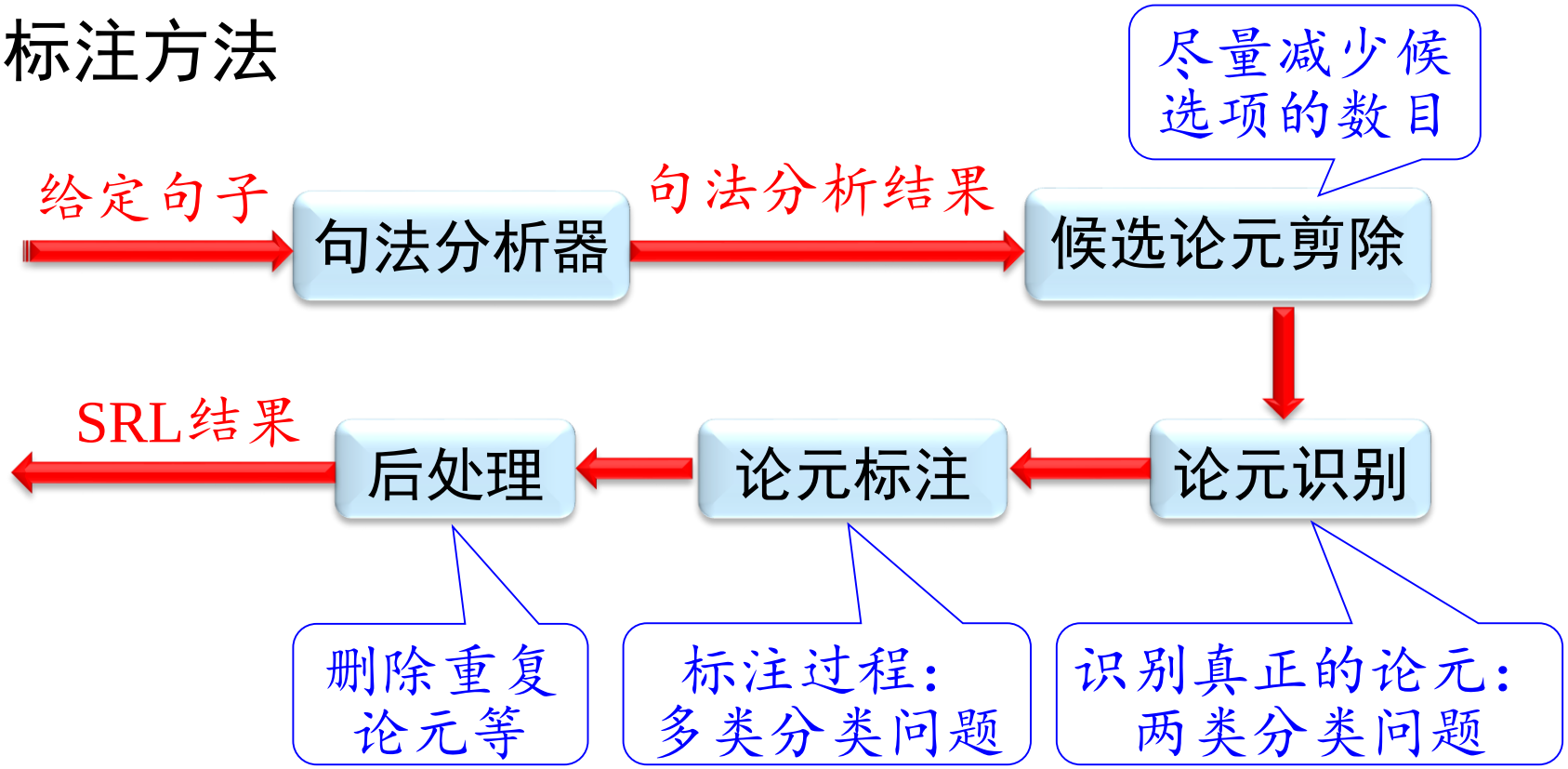
REL = knowledge

ARG1 = two-letter consonant sounds

NomBank, https://nlp.cs.nyu.edu/meyers/NomBank.html

4. 语义角色标注

◆标注方法



(1)基于短语结构树　　(2)基于依存关系　　(3)基于语块

4. 语义角色标注

(1) 基于短语结构句法分析的SRL方法(Xue and Palmer, 2004)

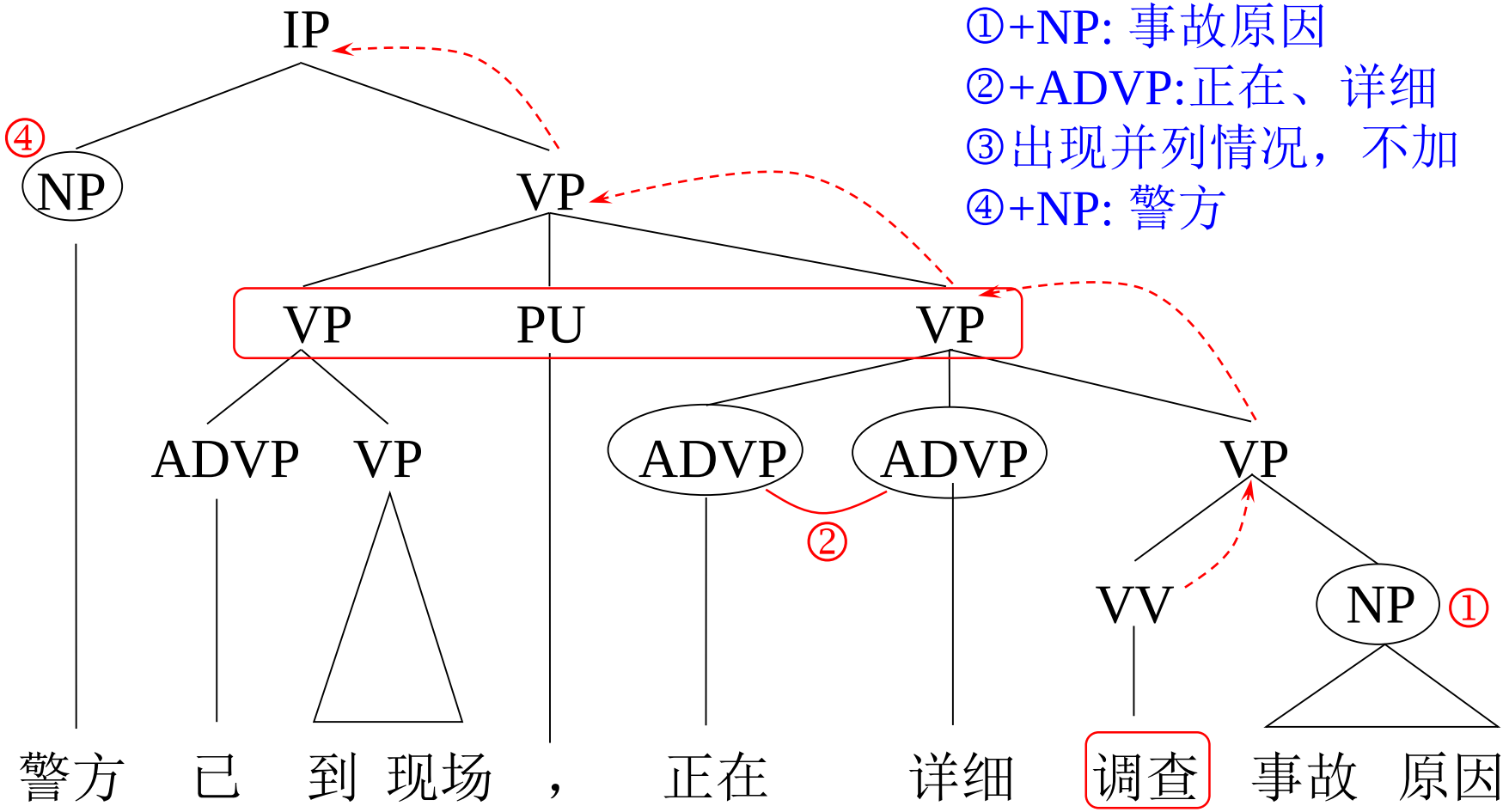
- 候选论元剪枝

第1步：**将谓词作为当前节点**，依次考察它的兄弟节点：如果一个兄弟节点和当前节点在句法结构上**不是并列的**(coordinated)关系，则将它作为候选项。如果该兄弟节点的句法标签是PP(介词短语)，则将它所有的子节点也都作为候选项。

第2步：将当前节点的父节点设为当前节点，重复第1步的操作，直至当前节点是句法树的根节点。

4. 语义角色标注

候选论元集合{ }
①+NP: 事故原因
②+ADVP:正在、详细
③出现并列情况，不加
④+NP: 警方

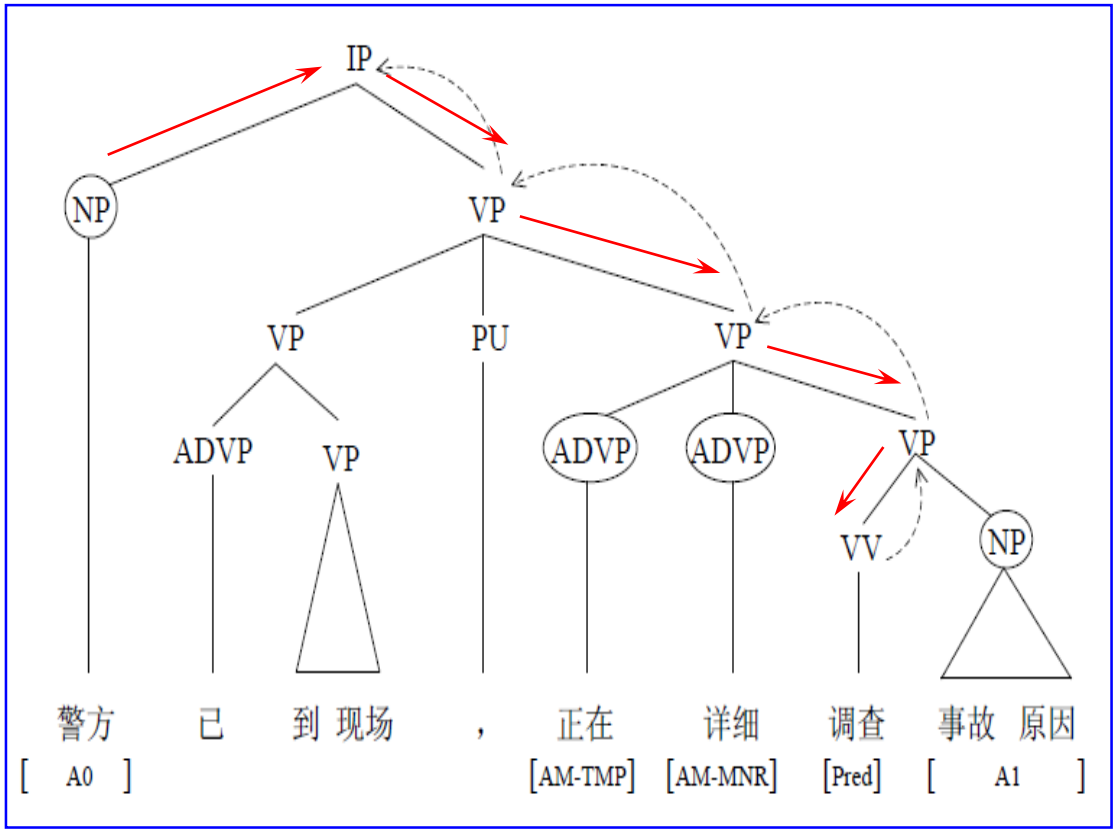


4. 语义角色标注

● 论元识别和标注

论元识别和标注看作一个分类问题，在这一阶段最重要的工作是为分类器选择有效的特征。常用的一些有效特征有：

- 谓词本身
- 路径(path): 句法树上从论元到谓词的路径，



如上面图中的A0 论元到谓词的路径就是：

NP↑IP↓VP↓VP ↓VP↓VV

4. 语义角色标注

- 短语类型(phrase type): 论元所对应的句法树节点的句法标签
- 位置(position): 论元出现在谓词之前还是之后
- 语态(Voice): 谓词是主动语态还是被动语态
- 中心词(Head Word): 论元的中心词及其词性
- 从属类别(Sub-categorization): 展开谓词父节点的上下文无关规则，如前面图中谓词的从属类别就是

  VP →ADVP ADVP VP
- 论元的第一个和最后一个词
- 组合特征(Combination features): 谓词＋中心词，谓词+ 短语类型等。

- **分类器：贝叶斯、最大熵、SVM、感知机等。**

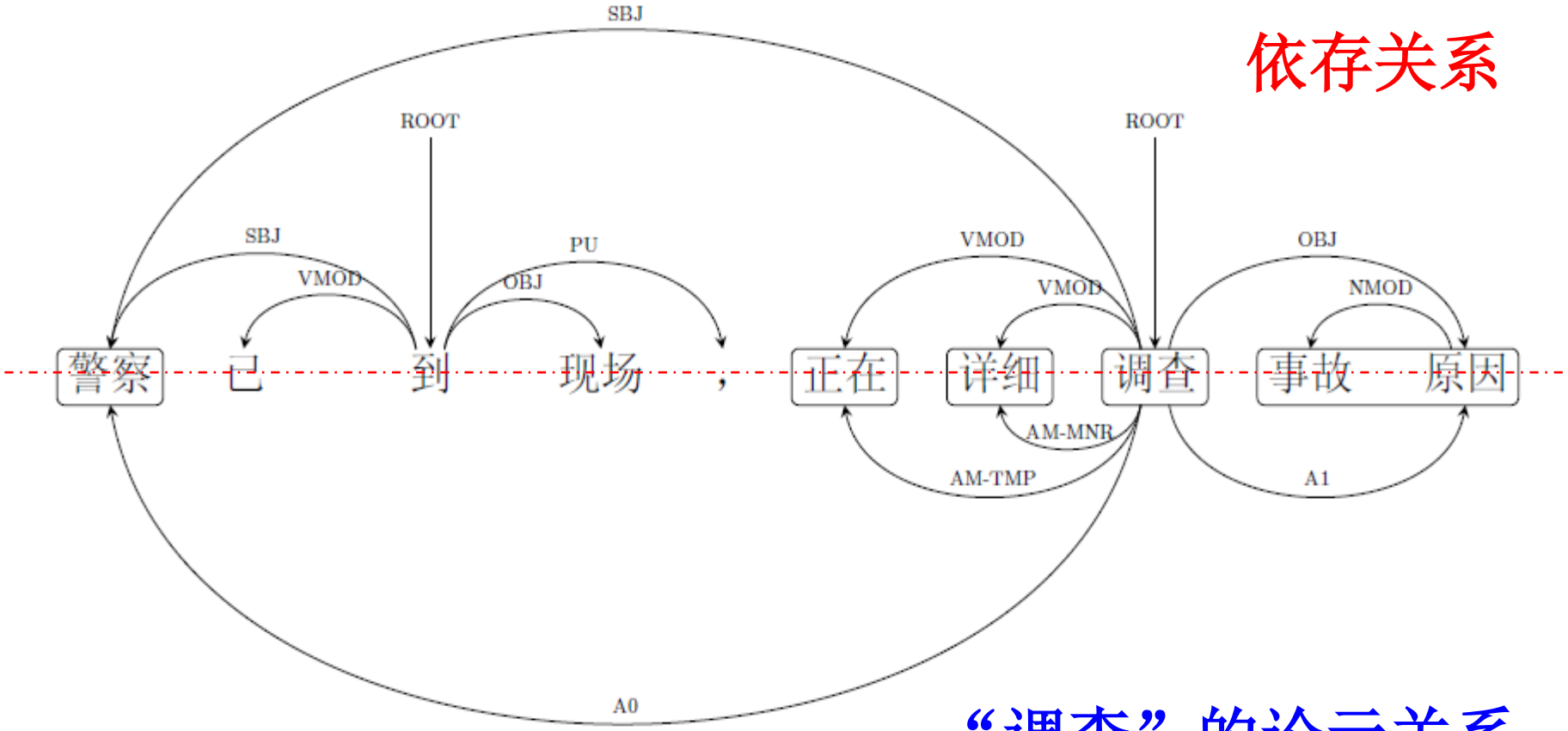
4. 语义角色标注

(2) 基于依存关系的SRL方法

- **与基于短语结构句法分析的SRL方法的区别在于：**

基于短语结构句法分析的语义角色标注方法中，一个论元被表示为连续的几个词（短语）和一个语义角色标签。而在基于依存句法分析的语义角色标注中，一个论元被表示为一个中心词和一个语义角色标签。因此，在这种方法中，谓词论元关系可以表示为谓词与论元的中心词之间的关系。

4. 语义角色标注



依存关系

“调查”的论元关系

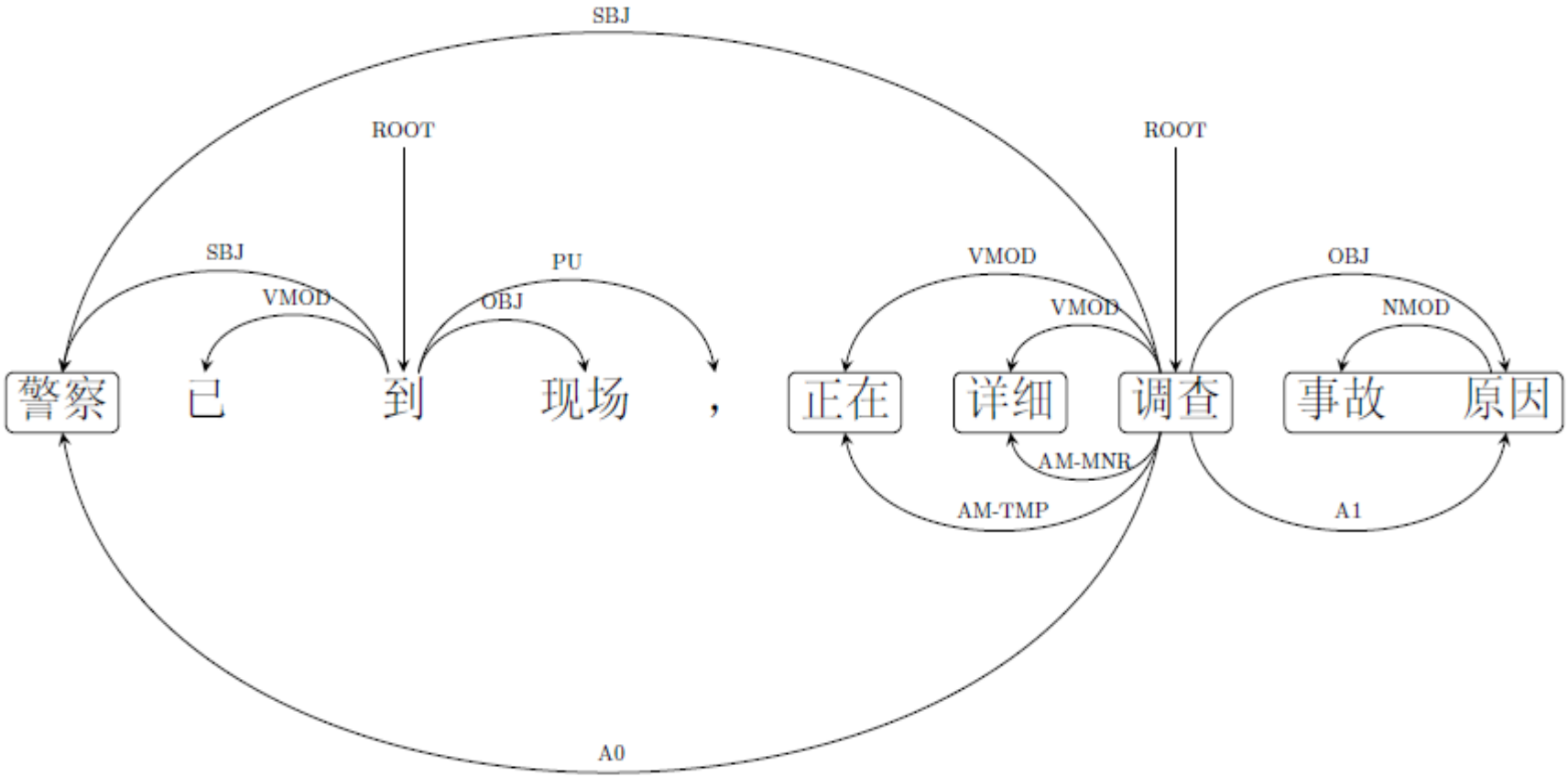
4. 语义角色标注

- 实现方法
  - 确定候选论元

第1步：**将谓词作为当前节点**，将它所有的孩子都作为候选项；

第2步：将当前节点设为它的父节点，重复第1步的操作，直到当前节点是依存句法树的根节点。

4. 语义角色标注



将谓词“调查”的所有孩子节点“正在，详细，原因，警察”都加入到候选项中。该例中恰巧这些孩子节点是谓词的所有论元。

4. 语义角色标注

➢论元的识别和标注

从上述过程可以看出，基于依存句法的语义角色标注最终都是在判断谓词和候选的词之间的关系。于是，无论是论元的识别还是论元的标注，其核心都是判断一对词之间的关系。论元识别和论元标注都被作为分类问题。几种最常用的特征包括：

- 谓词(predicate): 谓词本身及其词根
- 谓词的词义: 谓词在语料中的词义类别
- 谓词词性(predicate POS): 谓词的词性
- 谓词父节点的词及词性
- 谓词与其父节点之间的依存关系类别
- 依存关系路径(relation path): 依存句法树上从候选词到谓词的路径；例如上图中从“事故”到谓词的路径就是NMOD↑OBJ↑

4. 语义角色标注

- 位置(position): 论元出现在谓词之前还是之后
- 语态(voice): 谓词是主动语态还是被动语态
- 从属类别(dependency sub-categorization): 谓词的所有孩子对它的依存关系，如上图中谓词“调查”的依存从属类别是SBJ_VMOD_VMOD_OBJ
- 候选词本身
- 候选词最左边和最右边的孩子的词与词性
- 候选词左边和右边最近的兄弟的词与词性

- **分类器：贝叶斯、最大熵、SVM、感知机等。**

4. 语义角色标注

(3) 基于语块分析的SRL方法

用语块分析(Chunking)的结果来进行语义角色标注。谓词—论元关系的表示方法与基于短语句法分析中的表示方法相同，每一个论元都表示为连续的几个词，将语义角色标注作为一个序列标注任务。

- 基本思路：将语义角色标注作为一个序列标注问题来解决。一般采用BIO(分别表示：开始、属于、不属于)的方式来定义序列标注的标签集，将不同的语块赋予不同的标签。

  不需要剪除候选论元，论元识别和标注同时进行。

4. 语义角色标注

- 举例：

句子	警察	已	到现场	，	正在	详细	调查	事故	原因
语块	[NP]	[ADVP]	[VP]		[ADVP]	[ADVP]	[VP]	[NP]	[NP]
序列	B-A0	O	O		B-AM-TMP	B-AM-MNR	B-V	B-A1	I-A1
角色	[A0]				[AM-TMP]	[AM-MNR]	[V]	[A1]	

- 其他方法：
  - 多种方法的融合策略
  - LSTM+CRFs
  - E2E 词序列到标记序列的翻译方法

4. 语义角色标注

- 已有方法的主要问题
  - 对句法分析器性能的严重依赖性
  - 领域适应能力差
- 基本性能
  - 英语和汉语SRL标注的F1值大约为75％左右，一般不超过85%。

本章小结

◆概述

语义分析的基本任务及其面临的困难。

◆语义网络

概念、关系、语义网络表示、事件的语义关系。

◆词义消歧

基于规则的方法、基于统计的方法、基于词典法

◆语义角色标注方法

- 基本任务
- 已有的资源 (PropBank, NomBank)
- **基本方法：**基于句法结构，基于依存关系，基于语块。统计分类；序列标注。

本章内容

1. 概述
2. 语义网络
3. 词义消歧
4. 语义角色标注
5. 习题
6. 附录：语义理论和格语法

5. 习题

1. 请查阅有关词义消歧的论文，了解最新研究进展。
2. 请查阅神经网络用于语义角色标注的论文，了解最新进展。
3. 请设计实现一种基于神经网络端到端的语义角色标注方法。
4. 请下载并调试运行和对比多种词向量学习工具。
5. 请阅读有关 HowNet 理论的文献，学习其思想方法。
6. 请查阅《同义词词林》在自然语言处理中应用的相关文献。
7. 请查阅格语法相关文献，学习其基本思想。

本章内容

1. 概述
2. 语义网络
3. 词义消歧
4. 语义角色标注
5. 习题
6. 附录：语义理论和格语法

6. 附录

附录1：语义理论简介

附录2：格语法

附录1：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，包括人和其他动物。

问题：对于抽象、复杂的问题这种定义无法处理。

启明星/暮星→金星；神仙？鬼？妖怪？

附录1：语义理论简介

◆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。

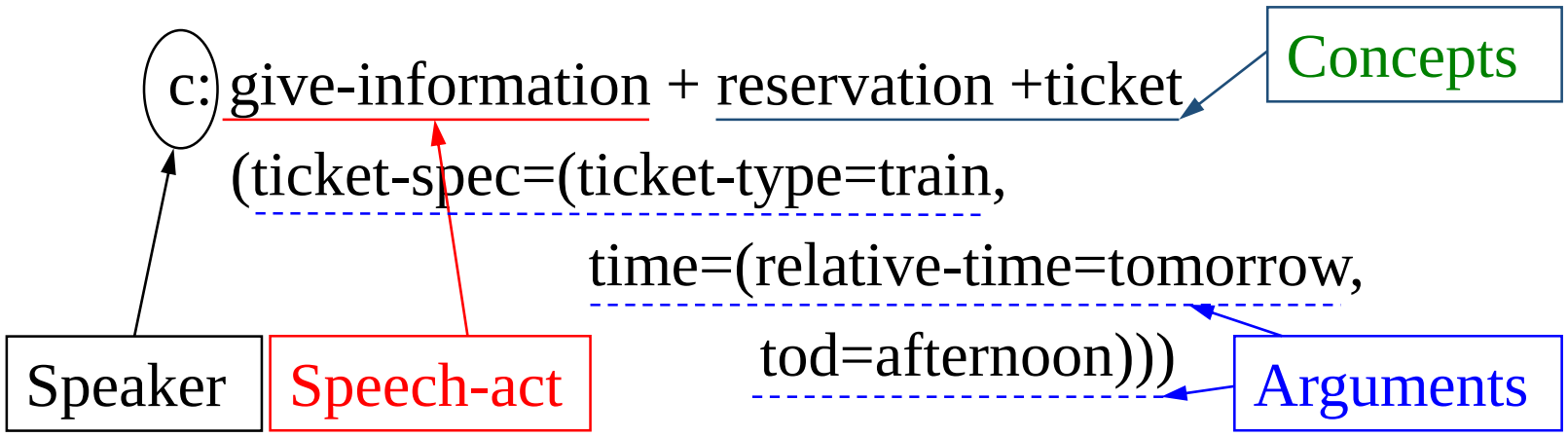
问题：在计算机中把心理图像有效地表示出来并不是一件容易的事情，而且不一定所有的词义都有清晰的心理图像，即使有，如何处理和心理图像，也是一个无解的问题。

附录1：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。



问题：意图的定义、划分和表示是困难的。

附录1：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。

这种方法简单、明了，对于计算机应用系统来说，在某种程度和限定领域内是有效的。

问题：对于语言本身缺乏解释，且句子的语义与应用之间的连接过于紧密，缺乏独立性。

附录1：语义理论简介

◆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。

问题：语义基元的定义、分解标准等难以把握，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，而且如何定义“操作”也是个困难的问题。

附录1：语义理论简介

◆条件真理模型

该理论以**谓词逻辑**为基础，句子的语义定义为它所对应的命题或谓词在全体模型（或世界）中的真伪。

例如：“雪是白的”为真，当且仅当在这个世界上雪是白的。

该方法对于上下文无关的语义描写可能有效。

问题：对时间、场景有关的语言现象不能很好地描述。不能很好地解释一句多义的问题。

附录1：语义理论简介

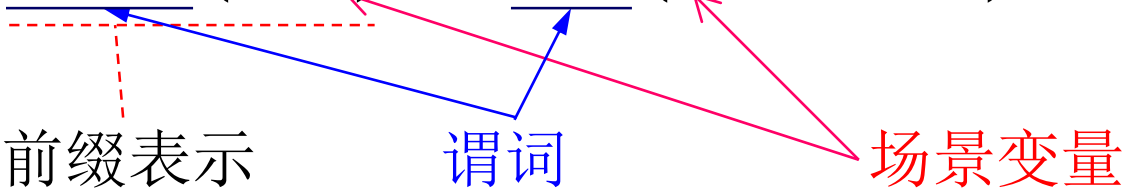
◆情景语义学

该理论认为句子的语义不仅和逻辑意义有关，而且与句子被使用的场景有关。

类似于条件真理模型，但在语义表达式中引入一些与场景相关的变量，如事件变量、时间变量等，并用逻辑“与”算子对这些变量加以限制。

例如：雪是白的

是白的 (雪, e) ∧ 等同 (e, 地球世界)



附录1：语义理论简介

◆模态逻辑

在20世纪80年代初的AI研究中提出了一系列类似的理论，包括缺省逻辑、时态逻辑、真值维护系统等。这类逻辑是试图用一套公理系统刻画现实世界和自然语言中常见的一些现象。这类现象从哲学上说就是一般性和特殊性的矛盾。

问题：“公理系统”总是刻画世界普遍成立的一般性的真理，难以涵盖特殊情况下的特殊事实。

例如：鸟会飞

企鹅不会飞

6. 附录

附录1：语义理论简介

附录2：格语法

附录2：格语法

◆背景

格语法(Case Grammar) 是美国语言学家 Charless J. Fillmore 于1966年提出来的。主要著作包括：

- 《关于现代的格理论》(Towards a modern theory of case), 1966
- 《格辨》 (The case for case) (代表作)，1968
- 《格语法的某些问题》(Some problems for case grammar), 1971
- 《再论格辨》(The Case for Case Reopened)，1977

附录2：格语法

◆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。

◆格的定义

格语法中的格是“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。

附录2：格语法

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。

例如：(1) The door is 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

附录2：格语法

◆格语法的三条基本规则

(1) S→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。

附录2：格语法

(2) $P \rightarrow V + C_1 + C_2 + \ldots C_n$

命题 P 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。

附录2：格语法

◆格表

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。He wrote a paper.

(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。He called her by mobile phone.

(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 was attacked. 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.

附录2：格语法

(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.

附录2：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: He 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.

格的数目和名称没确定。

附录2：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，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。

附录2：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]]]

附录2：格语法

- 分析的基础

构造一部词典，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。

例如，《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。

附录2：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)。

附录2：格语法

格框架分析通常与句法分析结合起来：

(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(谓词)，从而获得格框架。

附录2：格语法

The young athlete will be running in Los Angeles next week.

(1)从词典中查找 run 的格框架：

Verb: run

Case-Frame [

Neutral	-required（中性格）
Dative	-not allowed
Locative	-optional
Instrumental	-not allowed
Agentive	-required]

中性格类似一个物理实体或组织，例如：
John ran the machine.
He ran the corporation.

与格通常表示动词的间接宾语。

(2)确定情态(modals)

附录2：格语法

分析结果：

CASE

[Agentive: the young athlete

Locative: Los Angeles

Neutral: the young athlete

[Modals

[Tense: Future (将来时)

MOOD: Declarative (陈述语气)

Time: next week]]]

附录2：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。

谢谢！

Thanks!