第10章 语义分析

北京市海淀区中关村东路95号
邮编：100190

电话： +86-10-6255 4263
邮件： cqzong@nlpr.ia.ac.cn

10.1 概述

10.1 概述

- **语义计算的任务**：解释自然语言句子或篇章各部分(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。

10.1 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) 她说“这人真恶心！”

(5) 这种人算男人吗？！

10.2 语义理论简介

10.2 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。

问题：对于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？

10.2 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。

10.2 语义理论简介

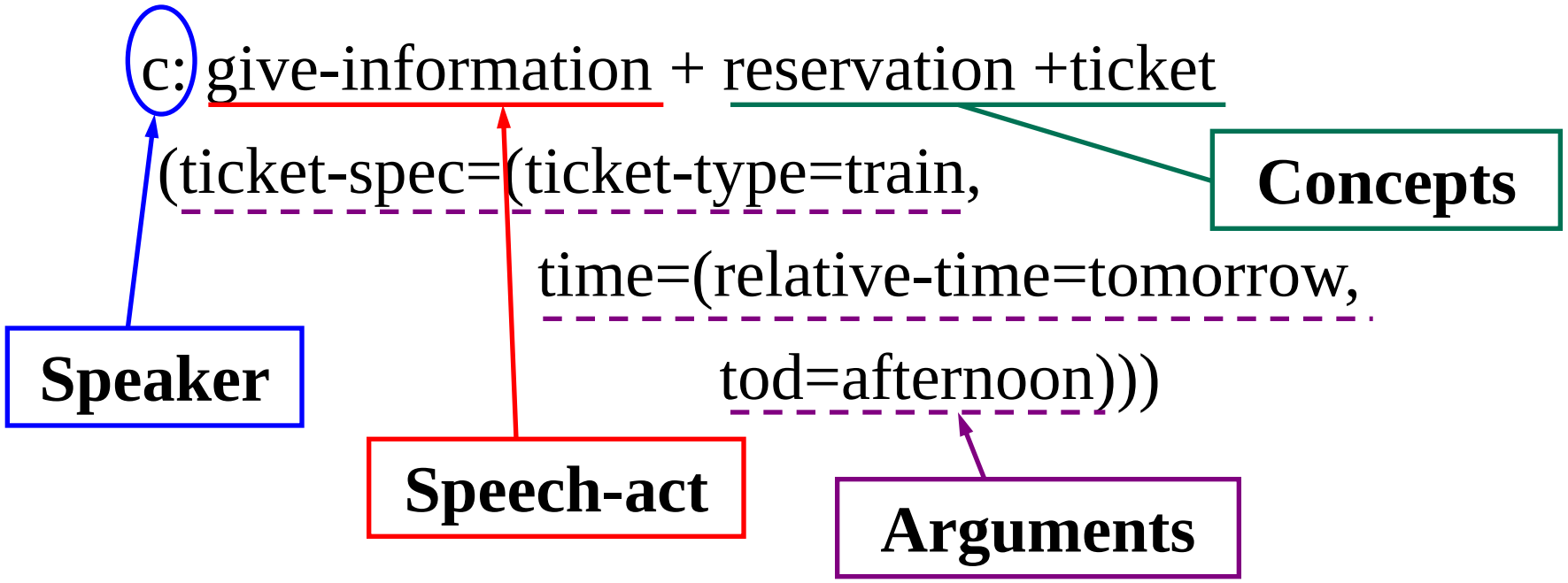
❒说话者的意图作为意义

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。

10.2 语义理论简介

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。

10.2 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。

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。

10.2 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。

10.2 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。

优点：对上下文无关部分的语义描写很有效。

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。

10.2 语义理论简介

❒情景语义学

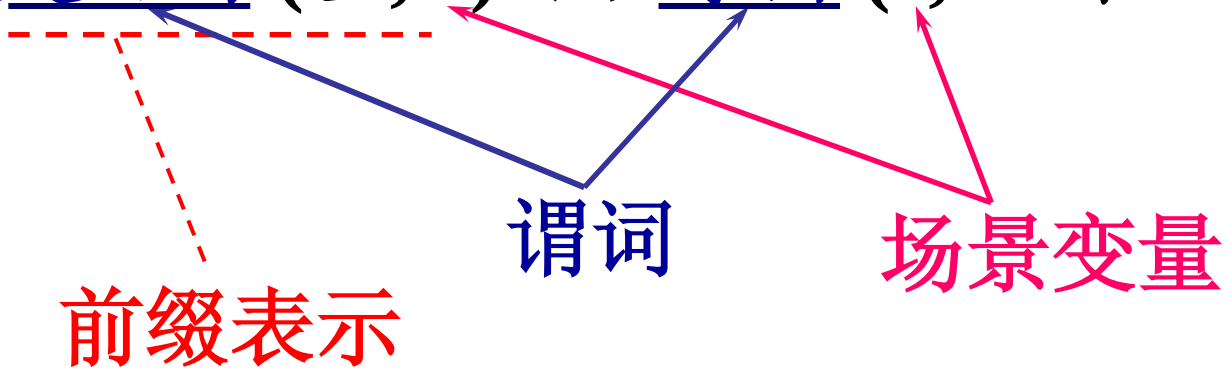
该理论认为句子的语义不仅和逻辑意义有关，而且和句子被使用的场景有关。

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。

10.2 语义理论简介

例如：雪是白的：

是白的 (雪, e) ∧ 等同 (e, 地球世界)



10.2 语义理论简介

❒模态逻辑

起源于20世纪80年代初，AI。如：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来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

10.3 格语法

10.3 格语法

❒背景

格语法(Case Grammar) 是美国语言学家 Charless J. Fillmore 于1966年提出的。代表作：

- 1966, Towards a modern theory of case
- 1968, The case for case
- 1971, Some problems for case grammar

10.3 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。

10.3 格语法

❒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。

10.3 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。

10.3 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

10.3 格语法

❒格语法的三条基本原则：

(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。

10.3 格语法

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

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 → 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。

10.3 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;

(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.

10.3 格语法

(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.

10.3 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.

10.3 格语法

(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。

10.3 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。

10.3 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]]]

10.3 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。

10.3 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。

10.3 格语法

(3) 识别可选格

(4) 判断句子的情态 Modal

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(谓词)，从而获得格框架。

10.3 格语法

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

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]**	

与格，通常表示动词的间接宾语。

run 的中性格像一个物理实体或组织，如：

John ran the machine.

He ran the corporation.

10.3 格语法

CASE

[Agentive: the young athlete

Locative: Los Angeles

Neutral: the young athlete

[Modal

[Tense: Future (将来时)

MOOD: Declarative (陈述语气)

Time: next week]]]

10.3 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。

10.4 语义网络

10.4 语义网络

❒背景

语义网络(semantic network)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。1977年美国 AI 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。

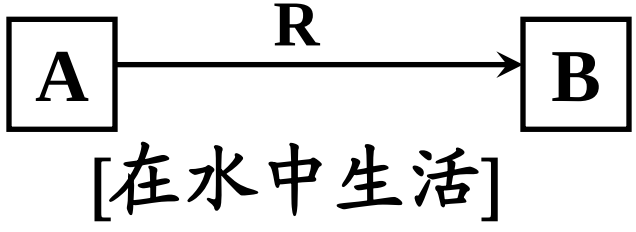
10.4 语义网络

❒语义网络的概念

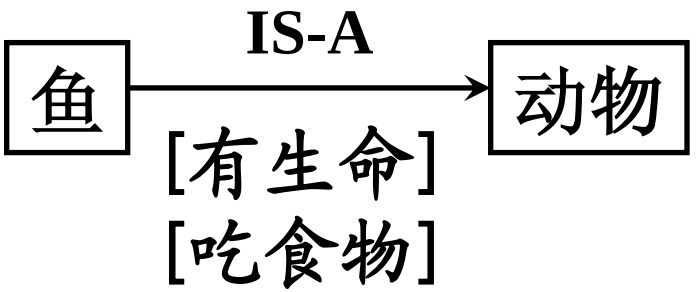
语义网络通过由概念和语义关系组成的有向图来表达知识、描述语义。

- 有向图：图的结点表示概念，图的边表示概念之间的关系。
- 边的类型：(1) “是一种”： A到B的边表示 “A是B的一种特例”；(2) “是部分”： A到B的边表示 “A是B的一部分”；……

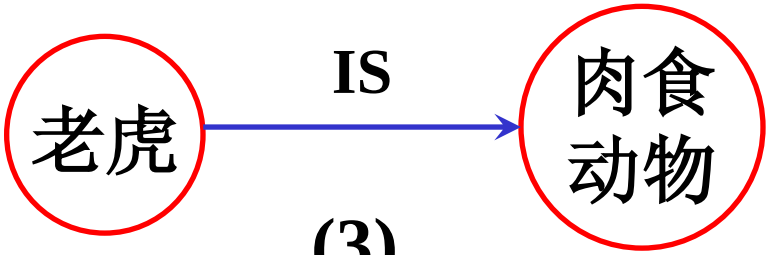
10.4 语义网络



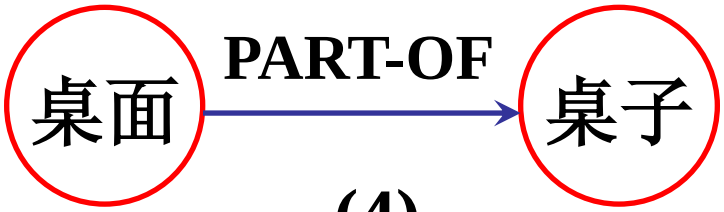
(1)



(2)



(3)



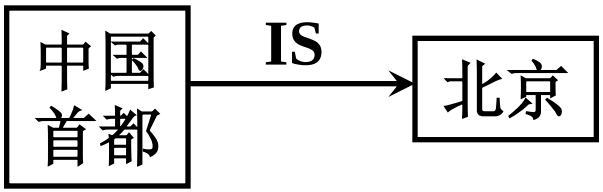
(4)

10.4 语义网络

❒语义网络的概念关系

语义网络各概念之间的关系，主要由 **IS-A, PART-OF, IS, COMPOSED-OF, HAVE, BEFORE, LOCATED-ON** 等谓词表示。

- **IS-A**：表示“具体－抽象”关系
- **PART-OF**：表示“整体－构件”关系
- **IS**：一个结点是另一个结点的属性



10.4 语义网络

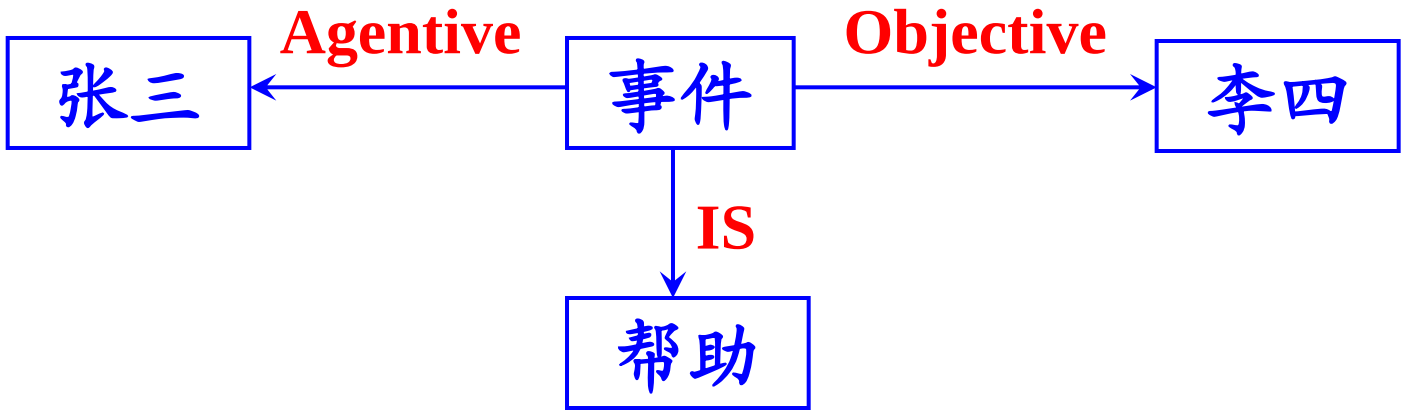
- **HAVE**：表示“占有、具有”关系
- **BEFORE/AFTER/AT**：表示事物间的次序关系
- **LOCATED-ON/UNDER/AT**：表示事物之间的位置关系

10.4 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。

例如：张三帮助李四。

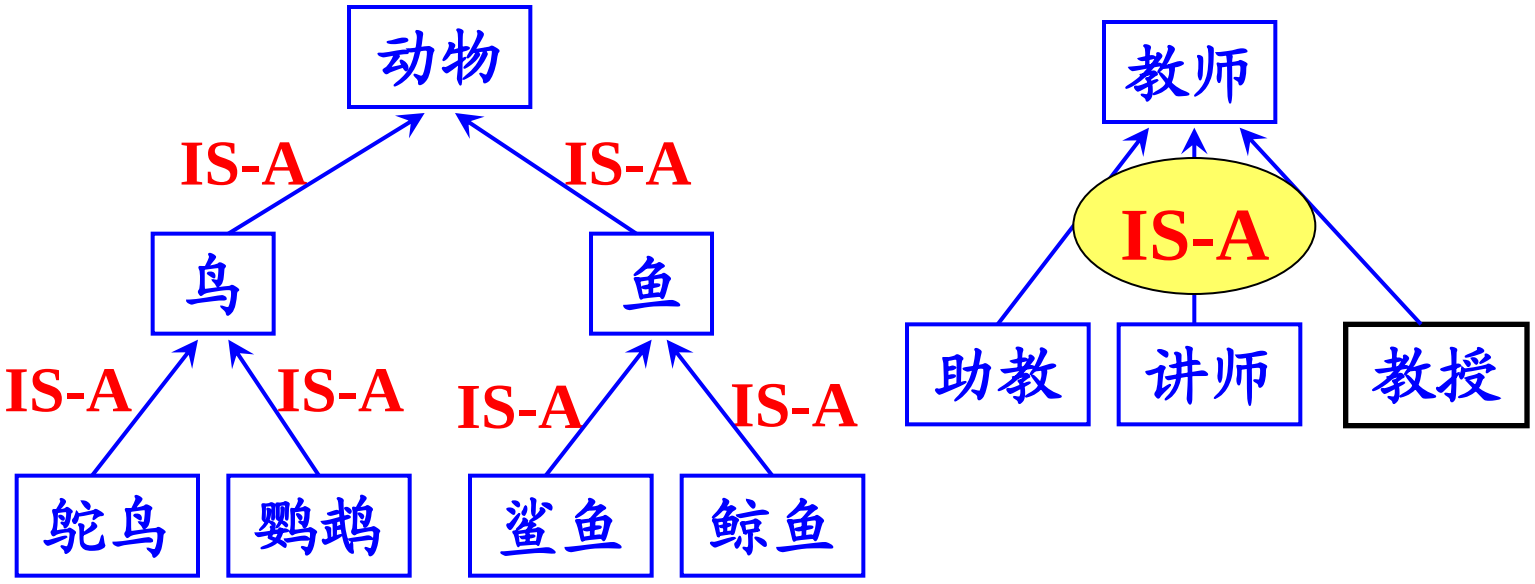


10.4 语义网络

❒事件的语义关系

(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。



10.4 语义网络

(3) 推论关系：由一个概念推出另一个概念。

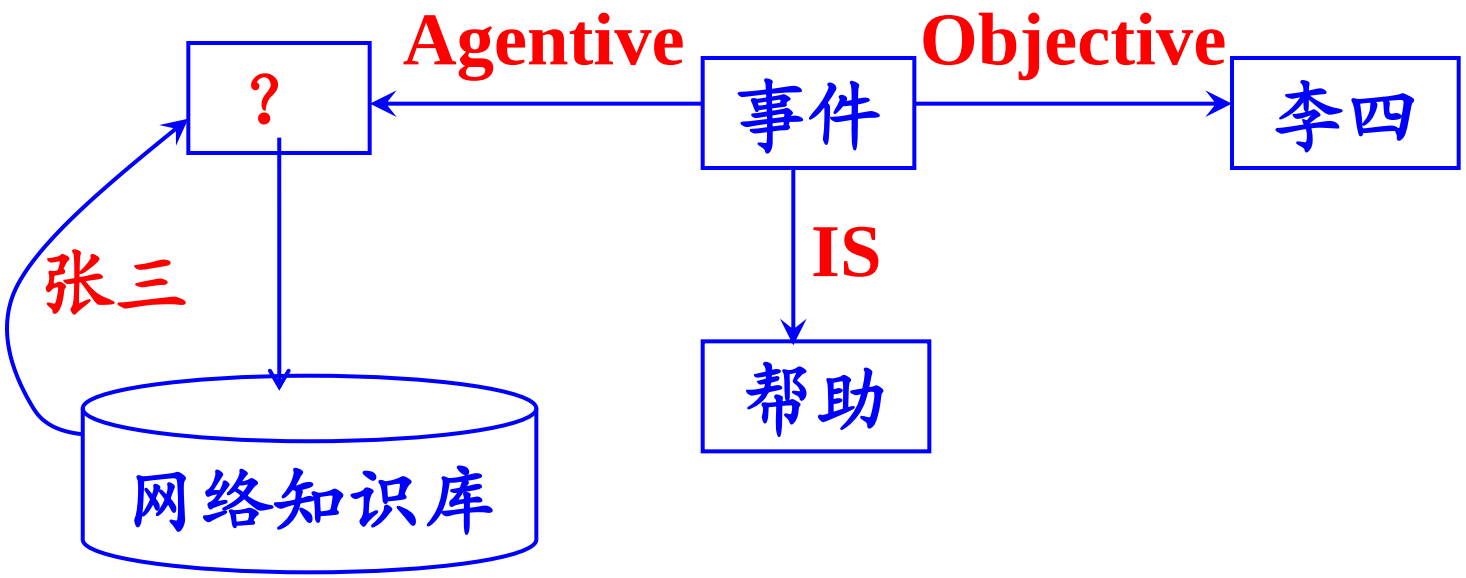
(4) 时间、位置关系：事实发生或存在的时间、位置。

10.4 语义网络

- 基于语义网络的推理、分析

(1)根据提出的问题构成局部网络;

(2)用变量代表待求的客体。



10.4 语义网络

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。

问题：如何在语义网络中表示和区分词的内涵和外延？

10.5 概念依存理论

10.5 概念依存理论

❒背景

Schank 和他的同事在70年代提出概念依存理论(Concept Dependence, CD)。

- 1975, Conceptual Information
- 1977, Scripts, Plans, Goals and Understanding

10.5 概念依存理论

❒CD 理论的组成：三个层次之一：动作基元

(1) 在概念依存层次：规定了一组动作基元，其他动作是由这些动作基元组合而成的。如：抓(Grasp)、移动(Move)、传送(Trans)、去(Go)、推(Propel)、吸收(Ingest)、撞击(Hit)等。

(2) 关于精神世界的概念：心传(MTrans)、概念化(Conceptualize)、心建(MBuild)。

(3) 关于手段或工具：闻(Smell)、看(Look-at)、听(Listen-to)、说(Speak)。

10.5 概念依存理论

◆三个层次之二：剧本

用来描写遇到一些常见场景或场合时所采取的一些固定的成套的动作。如:

(a) A推购物车或拿购物筐;

(b) A根据购物单或随意选购一些物品B;

(c) A把选购好的B给收帐员算帐、付款。

10.5 概念依存理论

◆三个层次之三：计划

计划中的每一步都是一个剧本，如，外出旅游的安排：

(a) 出门前的准备;

(b) 搭乘交通工具到目的地;

(c) 找住宿地点安顿下来;

(d) 在旅游地游玩;

(e) 若还未尽兴，转 (b)，否则，转 (f);

(f) 搭乘交通工具回家。

10.5 概念依存理论

❒依据CD 理论理解语言

一般文章中一些动作的细节被忽略，计算机难以发现事件、人物、地点等各种指代之间的联系，而CD 理论试图建立这种联系，正确描述常识，并利用基本动作推理。

该理论对限定领域内的特定应用比较有效。

缺陷：对常识的描写过于刻板和定式。

10.6 词义消歧

10.6 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 / …

10.6 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：一个词的不同语义一般发生在不同的上下文中。
- ◆基于词典信息的消歧方法

10.6 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。

◆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。

10.6 词义消歧

例子:

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

也就是说，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***。

10.6 词义消歧

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 $T_1, T_2, \ldots, T_m$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。

10.6 词义消歧

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 $T_1, T_2, \ldots, T_m$ 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*) 找到的一种改进的划分 P'，使 P' 与 Q 的互信息最大。

10.6 词义消歧

根据互信息的定义：

$$I(P;Q)=\sum_{x\in P}\sum_{y\in Q}p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$

算法终止的条件自然是互信息 $I(P;Q)$不再增加或者增加甚少。

10.6 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.

10.6 词义消歧

- 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

参见本讲义第2章2.3节。

10.6 词义消歧

- 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。

10.6 词义消歧

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。

10.6 词义消歧

(2) 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。

10.6 词义消歧

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。

10.6 词义消歧

(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。

10.6 词义消歧

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”。

10.6 词义消歧

(4) Yarowsky 消歧算法

基本思想：基于词典的词义消歧算法都是分别处理每个出现的歧义词，且对歧义词有两个限制：

- 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。

10.6 词义消歧

在 Yarowsky 消歧算法中的处理方法：

(1) 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。

(2) 对于第二个约束，Yarowsky (1995) 采用基于自举 (bootstrapping) 的(半监督) 学习技术。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。

10.6 词义消歧

❒无监督的词义消歧方法

H. Schütze (1998) 提出的上下文分组辨识 (context-group discrimination) 方法是无监督的词义消歧方法的典型代表。

与(Gale, 1992) 方法类似，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 $p(v_j \mid s_i)$ 。

10.6 词义消歧

但是，在该方法中参数 $p(v_j \mid s_i)$ 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。

10.6 词义消歧

为了解决这个问题，H. Schütze (1992) 对词汇集中的每一个词w 定义了关联向量(associate vector)，该向量为w 的平均上下文。

$$A(w)=\sum_{i=1}^{n}\delta(w_k,w^j)\langle c_k^1,c_k^2,\cdots,c_k^w\rangle$$

上标表示词汇集中的词形(type), 如 w^j表示词汇集中的第j 个词；下标表示一个词在语料库中的一次具体使用，简称为“词用(token)”，w_k表示语料库中的第k 个词；n为词的个数，即语料库大小；c_k^j为词形w^j 出现在w_k 的上下文中的次数；$\delta(x,y)$ 为Kronecker函数。

10.6 词义消歧

关于该工作的详细介绍请参阅:

[Schütze, 1992a] Schütze, Hinrich. 1992. Context Space. In *Working Notes of the AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, Menlo Park, CA. AAAI Press. Pages 113-120.

[Schütze, 1992b] Schütze, Hinrich. 1992. Word Sense Disambiguation with Sublexical Representation. In *Proceedings of the 1992 AAAI Workshop on Statistically-based Natural Language Programming Techniques*. Pages 100-104.

10.6 词义消歧

严格地讲，利用完全无监督的消歧方法进行词义标注是不可能的，因为词义标注毕竟需要提供一些关于语义特征的描述信息。但是，词义辨识 (word sense discrimination) 却可以利用完全无监督的机器学习方法实现。

10.7 语义角色标注

10.7 语义角色标注

❑语义角色标注(semantic role labeling, SRL)的任务

自动语义角色标注方法是近几年来国际研究的热点，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]$_{Agent}$ [昨天]$_{Time}$ [在北京]$_{Location}$ [讨论]$_{Pred}$ 了 [方案]$_{Patient}$。

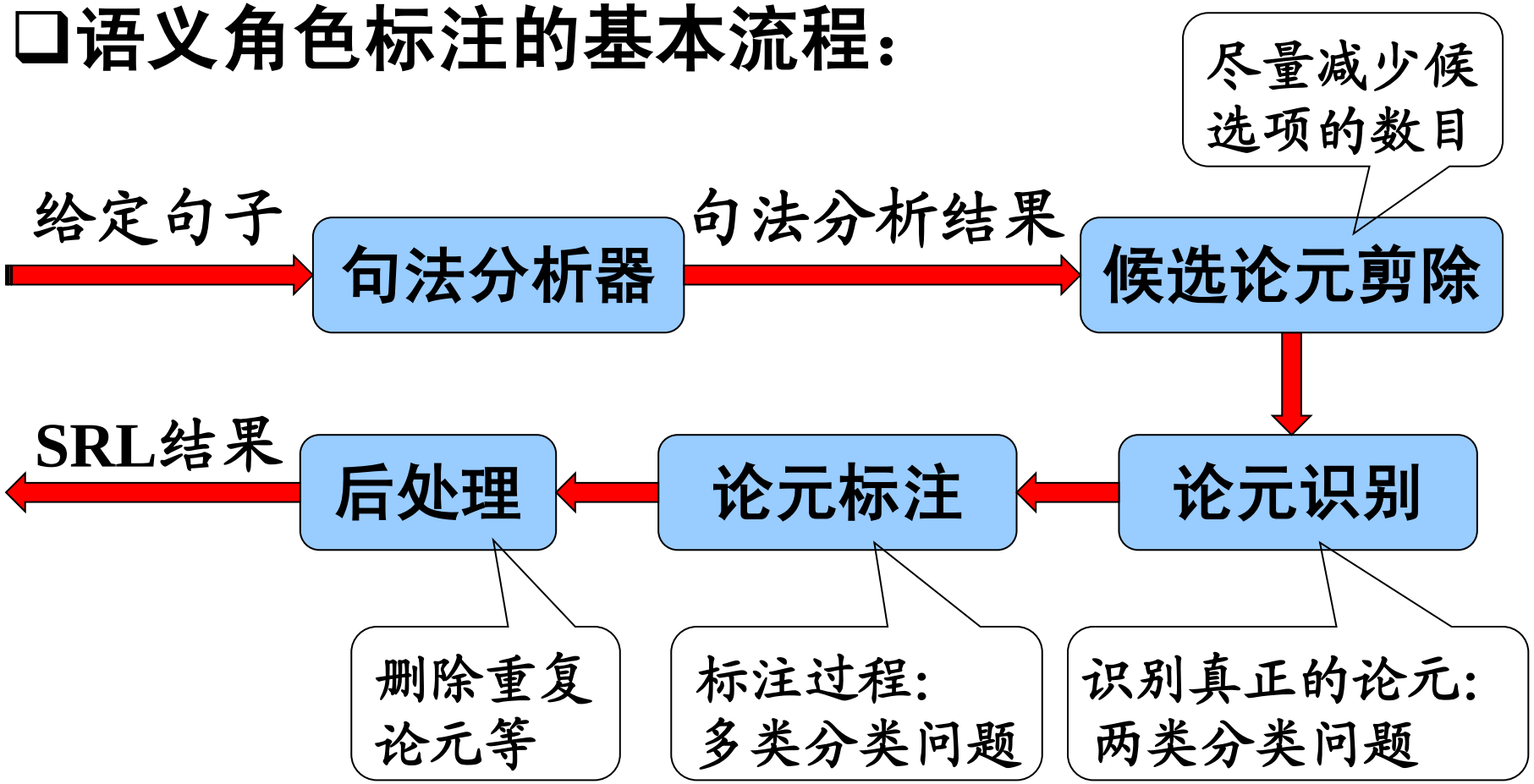
语义角色标注一般是在句法分析的基础上进行的。

10.7 语义角色标注

- SRL的主要用途：
  - 信息抽取、自动文摘、机器翻译等
- 目前用于SRL研究的主要资源有：
  - 框架网(FrameNet)
  - 英语命题库(Proposition Bank, PropBank)
  - 英语名词命题库(NomBank)

10.7 语义角色标注

❑语义角色标注的基本流程：



10.7 语义角色标注

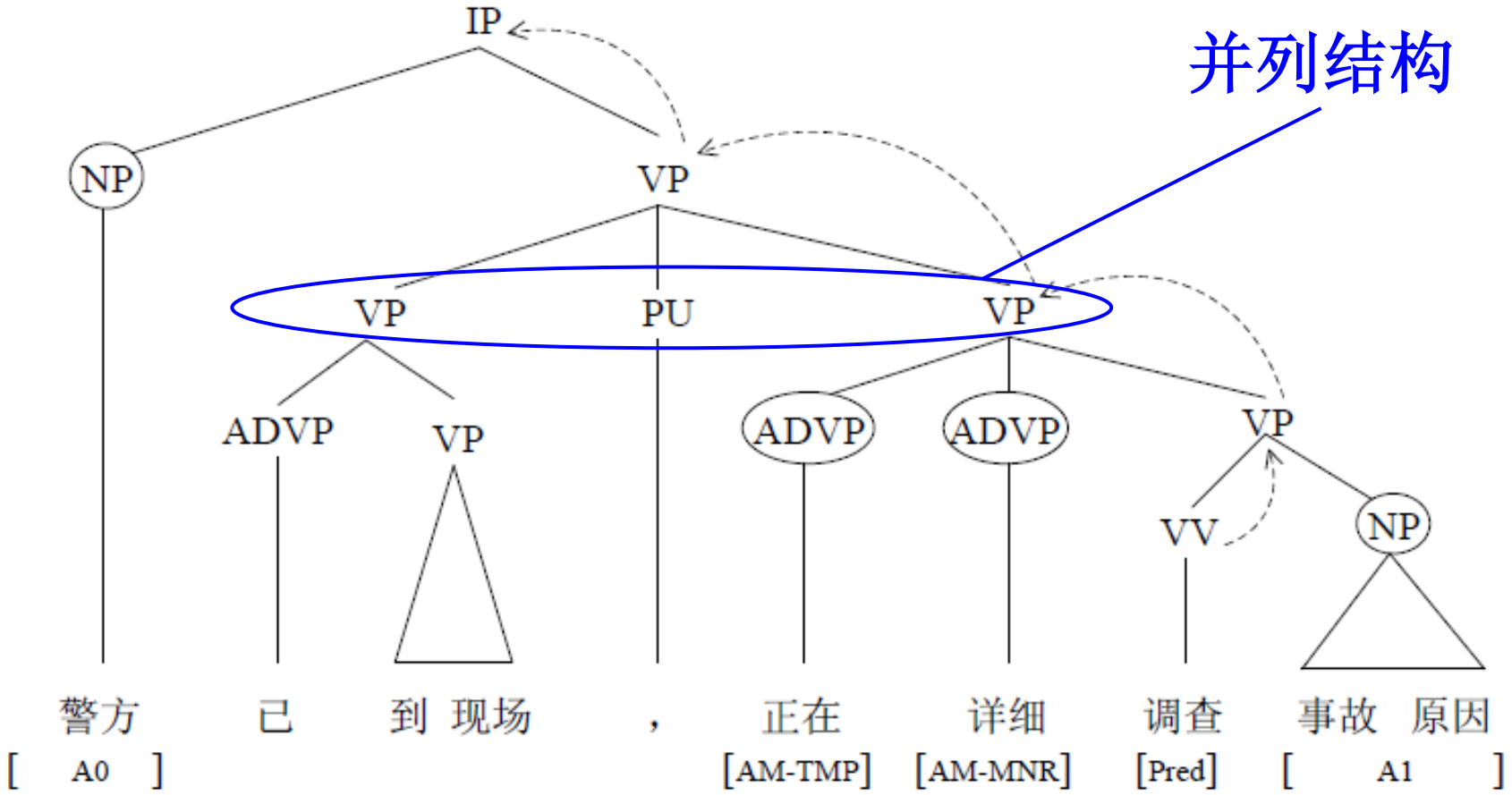
1. 基于短语结构句法分析的SRL方法(Xue and Palmer, 2004)

◆剪除方法：

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，在将它的所有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。

10.7 语义角色标注


IP
并列结构
NP
VP
VP
PU
VP
ADVP
VP
ADVP
ADVP
VP
VV
NP
警方 已 到 现场 , 正在 详细 调查 事故 原因
[A0] [AM-TMP] [AM-MNR] [Pred] [A1]

10.7 语义角色标注

◆论元识别和标注：

在论元识别和标注阶段，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(predicate): 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
- 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

10.7 语义角色标注

- 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：** 最大熵、SVM、感知机等。

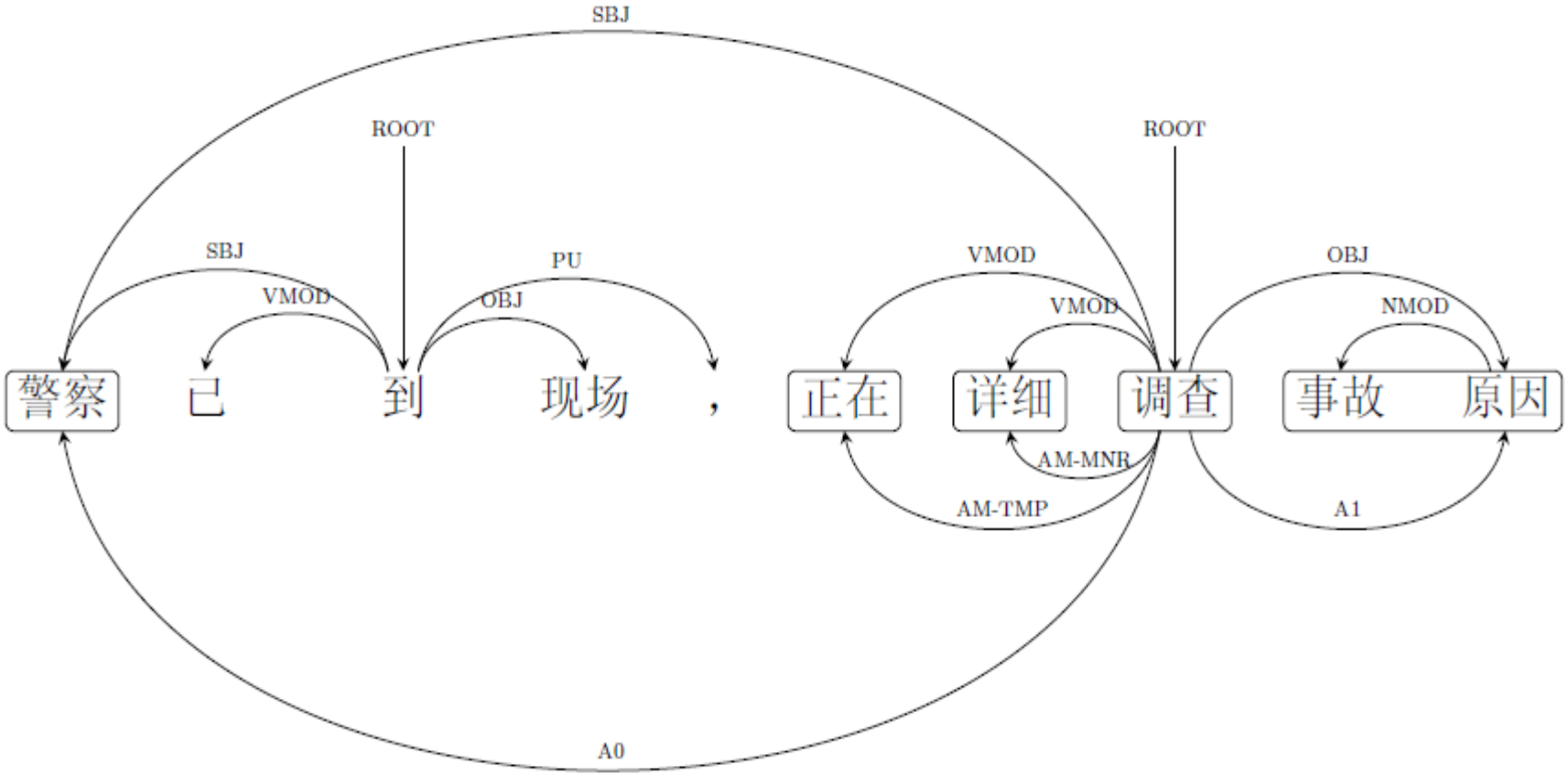
10.7 语义角色标注

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：

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和一个语义角色标签。但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。

10.7 语义角色标注


SBJ
ROOT
ROOT
SBJ
VMOD
PU
OBJ
VMOD
VMOD
OBJ
NMOD
警察
已
到
现场
，
正在
详细
调查
事故
原因
AM-MNR
AM-TMP
A1
A0

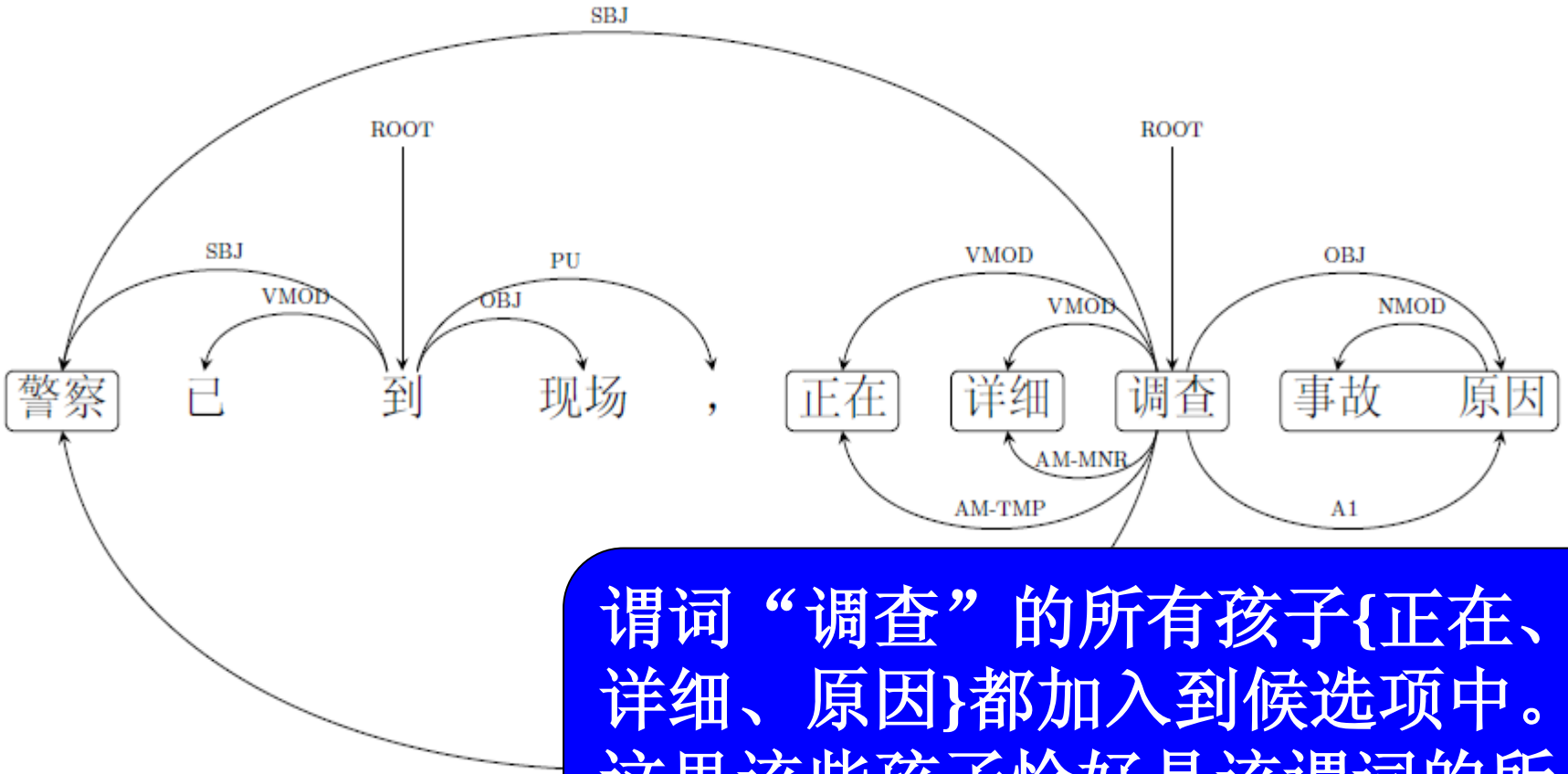
10.7 语义角色标注

◆剪除方法：

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。

10.7 语义角色标注



谓词“调查”的所有孩子{正在、详细、原因}都加入到候选项中。这里该些孩子恰好是该谓词的所有论元。

10.7 语义角色标注

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，无论是论元识别还是论元标注，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

10.7 语义角色标注

- 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↑。
- 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。

10.7 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

- **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IBO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。

10.7 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]	

10.7 语义角色标注

- 其他方法：
  - 多种方法的融合策略
  - 基于深度信念网络(deep belief network, DBN)的SRL方法

10.7 语义角色标注

- 现有方法存在的主要问题：
  - 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、汉语：F1值大约为：70％左右(68%～76%)。

本章小结

- 语义分析的基本任务及其面临的困难
- 语义计算研究概括及常见的语义理论
- 格语法(定义、格框架约束分析)
- 语义网络(概念、关系、语义网络表示、事件的语义关系、基于语义网络的推理分析)
- CD 理论(三个层次：基本动作、剧本、计划)
- 词义消歧(规则方法、统计方法、词典法)
- 语义角色标注的基本概念和方法

习题

1. 阅读有关 HowNet 和HNC 理论的文献，了解相关工作及其《同义词词林》在自然语言处理中的应用。
2. 了解蒙塔格语法(Montague Grammar)。
3. 阅读有关词义消歧的论文，了解词义消歧的相关工作。
4. 阅读有关语义角色标注的论文，了解相关工作。

Thanks