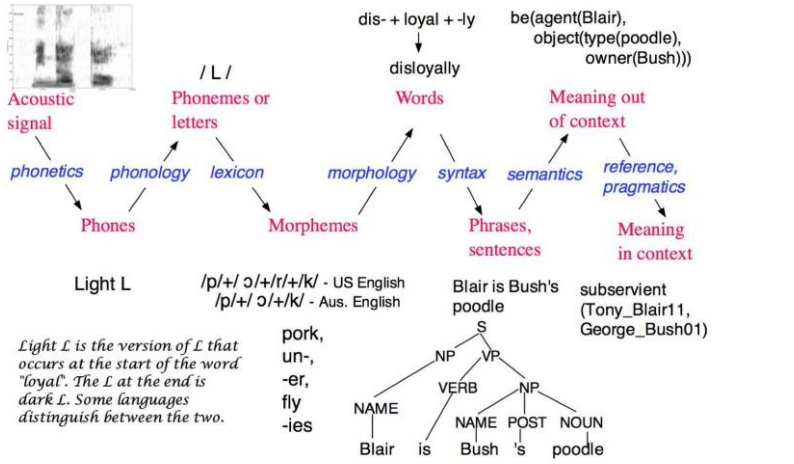
COMP9414: 人工智能讲座7a:语法和解析

韦恩-沃布克
电邮：w.wobcke@unsw.edu.au

本讲座

- 自然语言概述
- 自然语言的句法和语法

语言学景观



自然语言处理

- (自然语言的（简单）语义学
- 自然语言的语用学

- 语法
  - 语言学知识
  - 语法和解析
  - 概率分析法
- 语义学
  - 语义解释和逻辑形式
- 语用学
  - 话语处理
  - 言语行为理论
  - （口语）对话系统

新南威尔士大学
©W.Wobcke等人，2019-2022年

相关学科

- 语言学
  - 研究抽象的语言和特殊的语言
- **心理**语言学
  - 人类语言处理的心理学模型
- **神**经语言学
  - 人类语言处理的神经模型
- 逻辑
  - 研究形式推理

NLP的应用

- **聊天机器人**
  - 客户服务，例如：CBA、Amtrak、Lyft、Spotify、Whole Foods
- 个人助理
  - Siri, Alexa, Google Assistant
- **信息提取**
  - 财务报告、新闻文章
- 机器（辅助）翻译
  - 天气报告，欧盟的合同，加拿大的汉萨德
- **社会机器人技**术

中心问题--模糊性

- 自然语言表现出模糊性

  "渔夫去了银行"(词条)

  "男孩看到一个拿着望远镜的女孩"（结构）"每个学生都参加考试"（语义）

  "桌子无法通过门口，因为它太[宽/窄]了"（务实）。
- 含糊不清使人难以解释短语/句子的含义
  - 但也使推理更难定义和计算
- 通过映射到无歧义的表述来解决歧义问题

结构上的模糊性

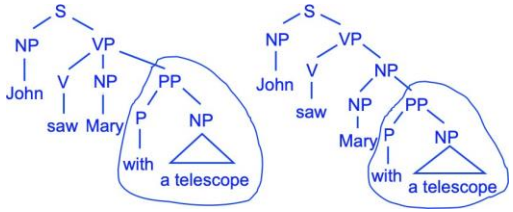
- 家庭护理机器人

"约翰用望远镜看到玛丽"

- 不同的解释→不同的表述

"约翰卖给玛丽一辆车 "和 "玛丽被约翰卖给一辆车"

- 相同的解释→相同的表述



新南威尔士大学
©W.Wobcke等人，2019-2022年

语法

- 语言学知识和语法
- **无语境语法**
- **剖析**
  - 自上而下的解析法
  - 自下而上的解析法
  - 图表解析
  - 确定性分析法
  - 概率分析法

框架（乔姆斯基）

- 描述性与规范性
  - 目标不是规定语言的使用，而是描述语言，特别是口语的实际使用方式。
- 句子与话语
  - 考虑句子（作为语词的抽象）。
- 能力与业绩
  - 注重基本的语言学知识
- 描述性与解释性的充分性

方法论

- 语法的自主性

  约翰承诺将工作

  *劝说约翰去工作

  VS

  *约翰被许诺工作（由其他人）

  约翰被说服工作（由其他人）
- 显示 "承诺 "和 "劝说 "具有不同的属性

*指不符合语法的

词汇项目（基本词汇）

- 旨在解释语言知识是如何获得的

- **公开**课
  - 名词：表示物体（例如：猫、约翰、正义）。
  - 动词：表示行动、事件（如购买、打破、相信）。
  - 形容词：表示物体的属性（如红色、大）。

  副词：表示事件的属性（例如：迅速）。
- **封**闭类**(功能**词**)**

  介词：at, in, of, on, . . .
  - 文章：the, a, an

  连词：and, or, if, then, than, . . .

新南威
尔士大
学
©W.Wobcke等人，2019-
2022年

句子形式

- 陈述式（指示式）
  - 巴特在听。
- 是/否问题（问句式）
  - 巴特在听吗？
    - 疑问句（问句）
  - 巴特什么时候在听？
- 当务之急
  听着，巴特!
- 分词
  - 如果巴特在听，他可能会听到一些有用的东西。

名词短语

- 句子的分布
  - 名词短语：作为 "主语 "出现，有一系列的 "谓语"。
    - (名词短语)吃了骨头
    - (见鸟在天
    - (相信2+2=4的说法
  - 示例
    约翰，那条狗，那条丑陋的大狗
    ，那个开红车的人。
    世界上最年长的胡须男，住在中国的最
    年长的人，.....。
- 句子不需要 "有意义"

动词短语

- 句子的分布
  - 动词短语：作为 "谓语 "出现，有一系列的 "主语
    "约翰（动词短语）。
    狗（动词短语）
    任何名词短语(动词短语)
  - 示例
    ...
- 注意（动词短语）取决于（名词短语）。

内部名词短语

- 在名词短语内
  - 主要项目（短语的头）：名词
  - 可选的说明者
    - 定语词（冠词、指示词、量词）
    - 形容词和其他名词
  - 强制性论据
    - 取决于头部（例如：首都（法国））。
  - 可选修饰语
    - 形容词短语（如：比西班牙大）。
    - 介词短语（如在公园里）。
    - 相对句（如：谁喜欢啤酒）。
  - 英语中的顺序指定符、头、修饰符

新南威尔士大学
©W.Wobcke等人，2019-2022年

内部动词短语

- 在动词短语内
  - 主要项目（短语的头）：动词
  - 可选的说明者
    - 助动词（如do, does, will, might, ...）。
    - 副词（如：快速）。
  - 强制性论据
    - 凭头（如买了（亨利）（一本书））。
  - 可选修饰语
    - 副词短语（例如，比亨利更快）。
  - 注意与名词短语的类似结构

介词短语

- 在介词短语内
  - 主要项目（短语的头）：介词
  - 强制性论据
    - (名词短语) (例如在公园里)
- 名词、动词等只是短语的头。

无语境语法

- 非顶点符号（语法类别）
- 终端符号（词条）
- 起始符号（非终端），例如（句子）。
- 重写规则
  - 非终端→非终端、终端的序列
    如：（句子）→（名词短语）（动词短语）。
- 开放性问题：英语是否无语境？

典型（小）语法

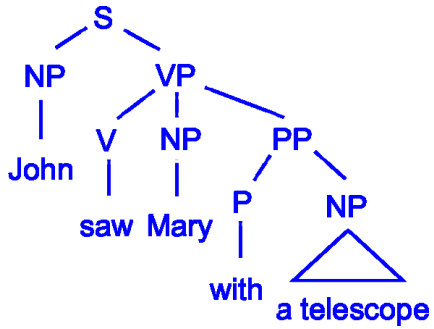
S → NP VP
NP → [Det] Adj* N [AP | PP | Rel Clause]*
VP → V [NP] [NP] PP*
AP → Adj PP
PP → P NP
Det → a | an | the | .. .
N → 约翰|公园|望远镜| .. .V →
锯子 | 喜欢 | 相信 | .. .Adj → hot
| hotter | .. .
P → in | .. .

特殊符号：* 是 "0或更多"；[. .] 是 "可选的"

新南威
尔士大
学

©W.Wobcke等人，2019-
2022年

句法结构



句法含糊 = 多于一个解析树

(最左边)例子的推导

S
⇒ NP VP
⇒N VP
⇒ John VP
⇒ John V NP PP
⇒ 约翰看到了NP PP
⇒ 约翰看到了N个PP
⇒ 约翰看到了玛丽-帕克
⇒ 约翰看到玛丽-P-NP
⇒ 约翰看到马利亚和NP在一起
＜约翰看见马利亚和底特律在一起＞
＜约翰看见马利亚和底特律在一起＞。

最右边的引申

S
⇒ NP VP
⇒ NP V NP PP
⇒ NP V NP P NP
⇒ 拒绝了，因为我不知道该怎么做。
⇒ NP V NP P Det telescope
⇒ NP V NP P 一架望远镜
⇒ . . .
⇒ . . .
⇒ . . .

剖析

⇒ 约翰看到马利亚有一个N
⇒ 约翰用望远镜看到了玛丽

⇒的意思是 "改写为"

- **目的是**计算一个句子的推导（因此是树）。
- **方法**
  - 自上而下
    - **从S开始**，应用改写规则，直到达到句子为止
  - 自下而上
    - 从句子开始，"反向"应用改写规则，直到达到S为止
  - 图表解析
    - 图表记录解析的片段和假设
    - 可以混合自上而下和自下而上的策略

新南威尔士大学
©W.Wobcke等人，2019-2022年

自上而下的解析

- 使用堆栈来记录工作假设
- 以S作为堆栈中唯一的符号开始
- 在每个步骤中
  - 使用语法规则T→RHS重写栈顶T
    即用RHS代替T（顺序相反），或
  - 将堆栈顶部的单词与句子中的下一个单词相匹配
- **在失**败时应用回溯法
- **当堆**栈为空且句子中的所有单词都匹配时，接受句子；当没有规则可供尝试时，拒绝句子
- 产生最左边的推导

例子

叠加	输入
S	约翰用望远镜看到玛丽
VP NP	约翰用望远镜看到玛丽
VP N	约翰用望远镜看到玛丽
约翰副总裁	约翰用望远镜看到玛丽
副总经理	用望远镜看到玛丽
PP NP V	用望远镜看到玛丽
PP NP 锯	用望远镜看到玛丽
PP NP	拿着望远镜的玛丽
.. .	.. .
.. .	.. .

自下而上的解析

- 使用堆栈来记录解析过的（左-右）片段
- 从堆栈空的时候开始
- 在每个步骤中
  - 使用T→RHS规则重写堆栈顶部的序列，即用T替换RHS（反向），或
  - 将字从输入端移至堆栈
- **在失**败时应用回溯法
- **当**输入为空且堆栈包含S时，接受句子；当没有更多的规则可以尝试时，拒绝句子
- 产生最右边的推导（反向）。

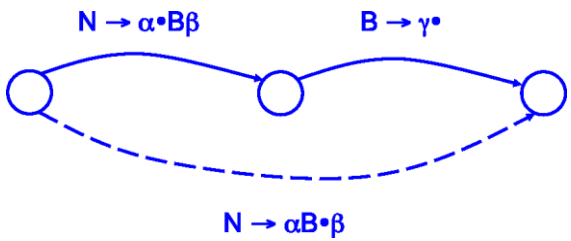
例子

叠加	输入
	约翰用望远镜看到玛丽
约翰	用望远镜看到玛丽
N	用望远镜看到玛丽
NP	用望远镜看到玛丽
NP看到	拿着望远镜的玛丽
NP V	拿着望远镜的玛丽
NP V Mary	用望远镜
NP V N	用望远镜
.. .	.. .
.. .	.. .

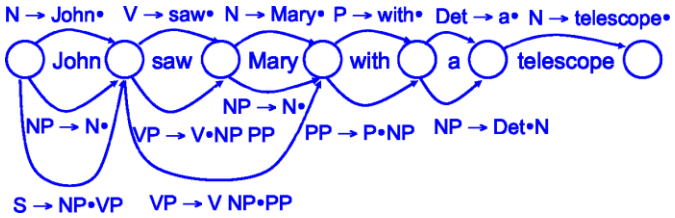
图表解析

- 使用图表来记录解析的片段和假说
- 假设N→ α –β，其中N→ α β是一条语法规则，意味着
  "试图将N解析为α β，并且到目前为止已经解析了α"
- 图表中每个词的间隙、开始和结束都有一个节点
- 每个假设在图表中都有一个弧线
- 在每个步骤中，应用基本规则
  - 如果图表有N → α – Bβ，从n_1到n_2，B → γ–，从n_2到n_3
    从n_1到n添加N → α B – β
- 当S→ α -被从头到尾加上时，接受句子
- 可以产生任何形式的推导

基本规则



示例图表



自上而下的图表解析

- 从S→–α开始，从起始节点到起始节点的所有规则S→ α
  其中S是起始符号
- 当把N → α – Bγ从n_1加到n_2
  - 同时从n_2到n_2，为每个规则B → β添加B → –β
  - 包括当α为空和n_1 = n的特殊情况
- 练习。追踪自上而下的图表解析器的例子

自下而上的图表解析

- 从每个词条的弧开始
- 当把C→ α -从n_1 加到n_2
  - 同时从n_1 到n_1 ， 为每个规则B → Cγ添加B → −Cγ
    - 练习。追踪自下而上的图表解析器的例子

比较和对比

- 自上而下的解析
  - 简单，内存效率高
  - 多次重复工作，可无限循环。
- 自下而上的解析
  - 减少重复工作，更难控制
- 图表解析
  - 内存效率低下（特别是有功能时）。
  - 没有重复工作，难以控制

确定性的解析

- **激励**

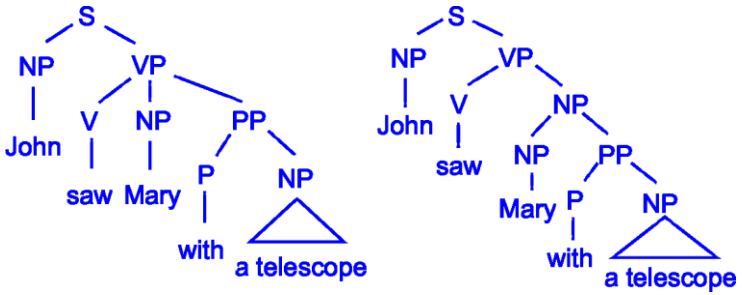
  人们不会注意到模棱两可的东西.....。
  - 但有时会遇到困难

    "马跑过谷仓的时候摔倒了""我们把所有的墙都刷成了裂缝"

    "那人把狗养在家里"
- 我们能不能像 "人类分析器 "那样做？

启发式方法

最低限度的附着力

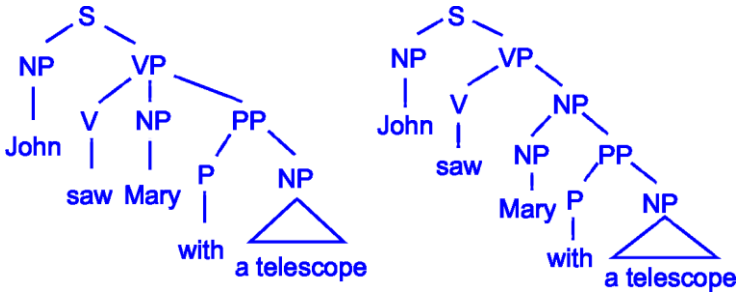


- 最大限度地减少解析树的大小

新南威
尔士大
学
©W.Wobcke等人，2019-
2022年

启发式方法

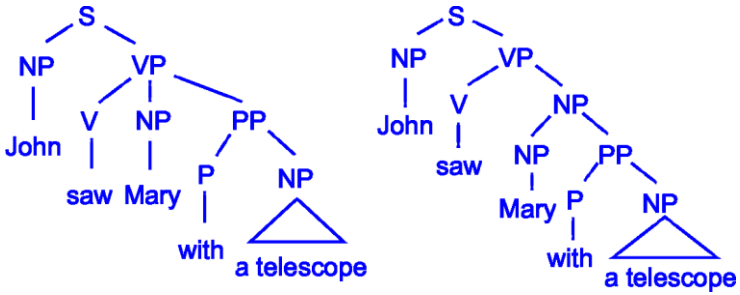
正确的协会



- 始终附着在最右边（较低）的节点上

启发式方法

词汇偏爱



- 尝试填补最常见的子分类框架

概率上下文自由语法

- 将概率与语法规则联系起来
  - 需要解析的语料库（如Penn Treebank）。
  - 计算在解析语料库句子中使用规则的次数
- 解析树的概率
  - Π_r 概率规则r * Π_w 给定类别的单词w的概率
  - 假设独立（再一次）

概率图解析

- 从语篇标记器计算的概率开始
- 应用基本规则时，将概率相乘
- 最佳优先图表解析
  - 首先审查最可能的选民（优先队列）。
    - 估计这些概率的各种方法!
  - 当添加A → α B – β时，尝试扩展到A → α B β^1 – β^2
  - 从不以低于解析的概率构造成分

摘要

- 句法知识
  - 通过分布定义的语法类别
  - 多由词条的属性决定
- 无语境语法
  - 有用且强大的形式主义
  - 相对高效的解析器
  - 在处理复杂现象时受到限制
- 剖析
  - 自上而下的方法容易理解，但效率不高
  - 自下而上的方法更有效率