《互联网数据挖掘》本科生课程

自然语言处理基础

万小军

北京大学计算机科学技术研究所
语言计算与互联网挖掘研究室
http://www.icst.pku.edu.cn/lcwm/

2017年10月10日

自然语言处理概述

如果人类身体每部分的生长和与该部分动作相关的大脑皮层区域的大小成比例，那么人类将的长相如图所示。

基本概念

- 语言
  - 广义上：一套共同采用的沟通符号、表达方式与处理规则；
  - 自然语言 vs. 动物语言 vs.电脑语言

- 自然语言
  - 指自然地随文化演化的语言，是人类交流和思维的主要工具
    - 英语、汉语、日语、藏语等
  - 不包括编程语言等为计算机而设的“人造”语言
    - C/C++、Java、Perl、Python、C#等

思考：鸟类的语言是自然语言吗？

鸟儿也能“唱歌” 科学家发现鸟类懂“语法”

http://www.sina.com.cn 2006年04月28日16:05 海峡都市报

N康娟

长期以来，科学家都认为语法是将人类与动物区别开来的标志性技能之一。但最新的研究却发现，一只普通的鸟也能掌握简单语法。

据美联社4月26日报道，美国加利福尼亚大学圣迭哥分校（UCSD）的心理学专家金特纳领导的研究小组对11只八哥进行了1个月的培训，最后的结果令他们感到非常惊讶，9只八哥学会了区分一般鸣叫和包含一个“分句”的鸣叫。而科学家相信，这些八哥很可能会把听到的“分句”鸣叫模仿出来，这样它们的鸣叫就不再是单一的“句子”，而是悠扬有致的“歌声”了。当然这些分句和句子都是用鸟类语言表达的。

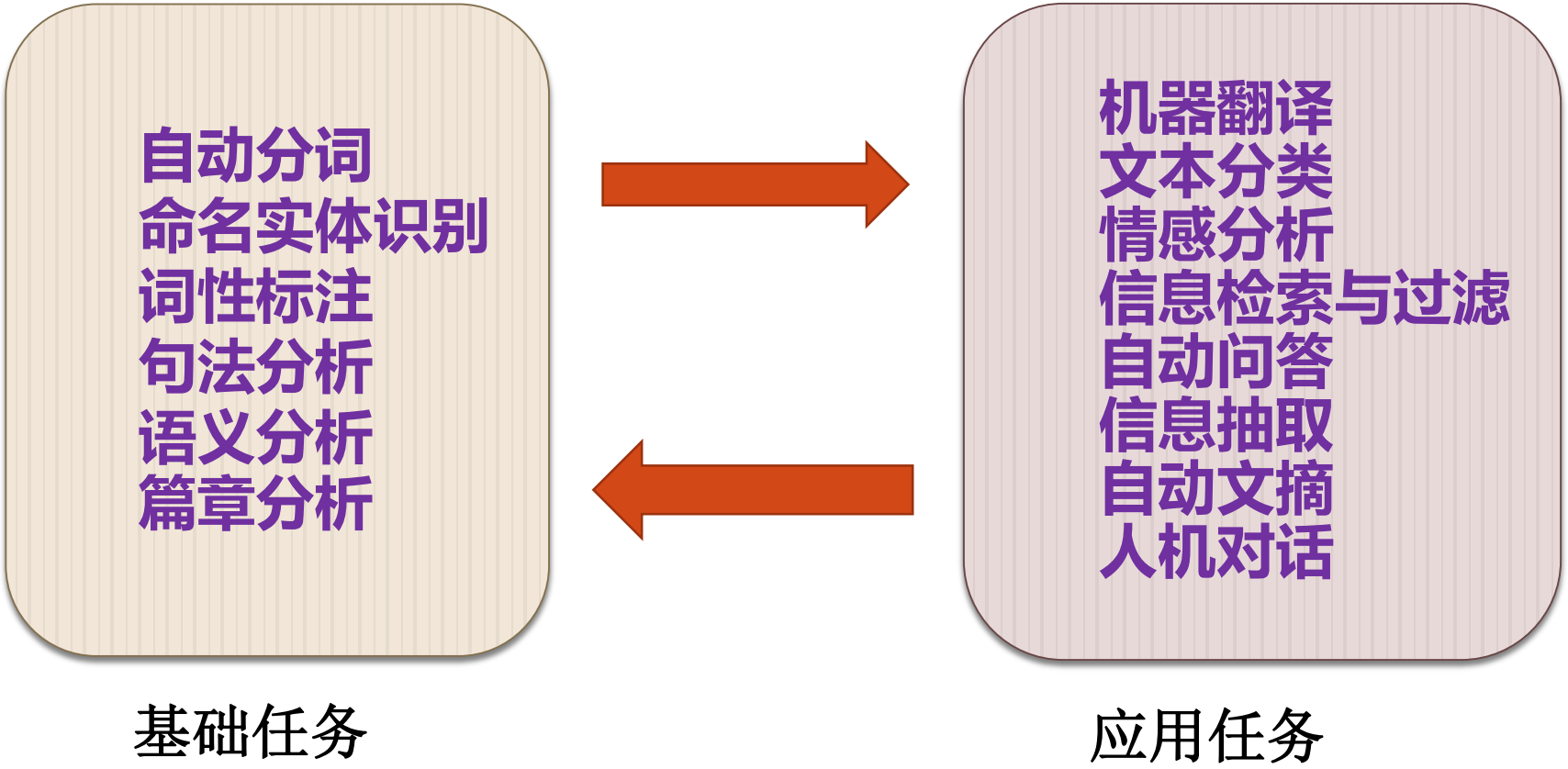
语言学家多年来都深信，人类语言技能的关键在于基本语法。只有人类才能辨认出分句，而动物不具备这种能力。通过这个实验，UCSD大学的认知科学教授埃尔曼表示，目前还没有区别人类和动物的唯一指标。哈佛大学的豪泽教授也认为，“一些认知手段是人类与动物所共有的。”

基本概念

- **自然语言处理**
  - 又称**自然语言理解**，是人工智能和语言学领域的分支学科。
  - 利用计算机为工具对人类特有的书面形式和口头形式的自然语言的信息进行各种类型处理和加工的技术。

- **自然语言生成**
  - 利用计算机为工具自动生成可理解的自然语言文本的技术。

基本任务


自动分词
命名实体识别
词性标注
句法分析
语义分析
篇章分析
机器翻译
文本分类
情感分析
信息检索与过滤
自动问答
信息抽取
自动文摘
人机对话
基础任务
应用任务

例子

原始句子

警察正在详细调查事故原因

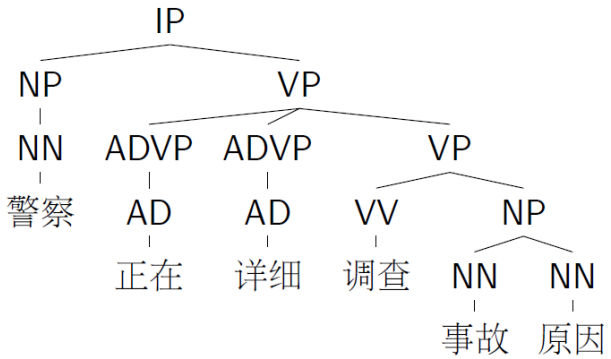
分词结果

警察 / 正在 / 详细 / 调查 / 事故 / 原因

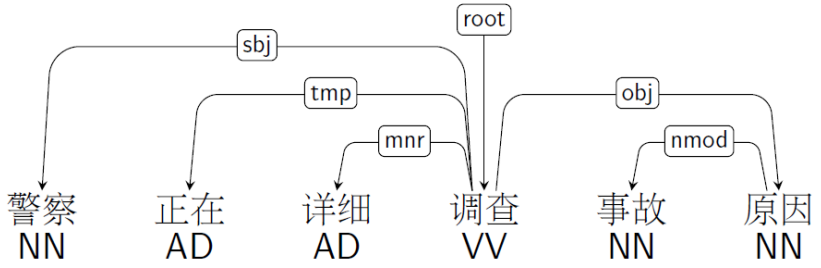
词性标注结果

警察/NN 正在/AD 详细/AD 调查/VV 事故/NN 原因/NN

短语结构树



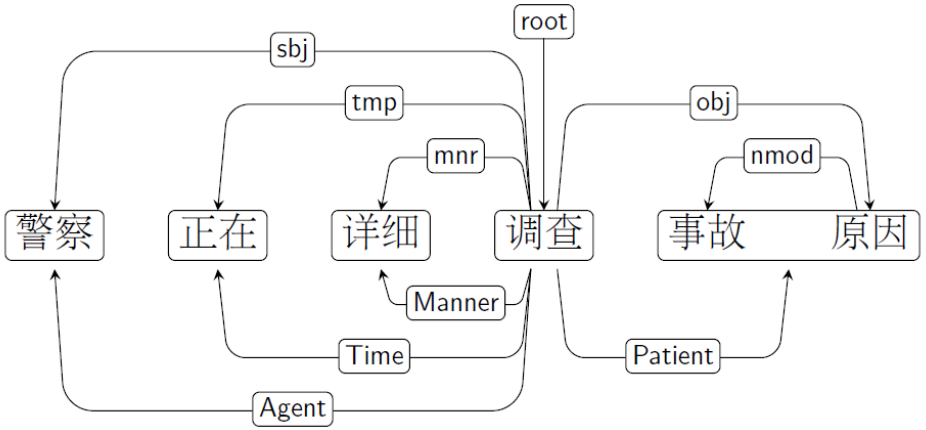
依存结构树



事件结构

Event:	调查
Agent:	警察
Time:	正在
Manner:	详细
Patient:	事故原因

句法语义依存关系图



基本方法

- 理性主义方法
  - 研究人的语言知识结构，人工编汇语言知识 + 推理系统
  - 符号处理系统
- 经验主义方法
  - 直接研究实际的语言数据，从大量的语言数据中获得语言的知识结构
  - 基于语言数据的计算方法
- 理性主义方法与经验主义方法的融合
  - 融合方法

规则与统计共舞，语言随计算齐飞。

Computational Linguistics - Rules dance with numbers, Language soars with information.

代表性应用系统

- 统计机器翻译系统：谷歌翻译
- 智能问答与对话系统：IBM沃森、苹果Siri、微软小娜、各类服务机器人
- 知识图谱： 谷歌知识图谱

我们何时能引领潮流？

日本要发明会高考的机器人

2011年12月21日 14:34:22

来源： 环球时报 新华微博 【字号：大 中 小】【打印】

据日本newspost网站20日报道，日本国立情报学研究所已决定加快人工智能技术发展，力争在10年内发明出“具有完全人类智能的机器人”，使其通过东京大学的入学考试。

日本国立情报学研究所本月中旬召开了一次人工智能技术研究会议，会议的中心课题是“让机器人考进东京大学”，即研究出能够全面应对人类的知识考试内容，包括可以应答理解性、开放性题目的机器人。

报道说，智能机器人的发展趋势将是完全模拟人类行动和思维。通过人类的综合性考试，将成为机器人技术发展的一项重要标志。但有日本民众笑称：“东京大学的考试并不好对付，机器人很快会知道。” （卢昊）

成绩不高 日本人工智能机器人“小东”弃考东大

作者： 来源：中国新闻网 发布时间：2016/11/14 13:56:35

中新网11月14日电 据日媒报道，本月14日，力争通过东京大学入学考试的人工智能机器人“小东”(东ROBO君)开发小组14日宣布，在挑战大型补习学校的大学入学统一考试模拟考试后，获得的所有学科总计标准分(日本称偏差值)为57，和去年基本持平。据报道，其物理的标准分从2015年的47大幅增至59，而数学则降低，未达到东大合格线。因为理解题目意思的阅读能力有限，研究小组今后将不把考入东大作为目标，而是转为改善记述式考试成绩等研究。

“小东”在统考中参加了语文、数学、世界史等5教科8科目的考试。语文满分为200分，其得分为96，标准分从去年的45上升到了50。另一方面，去年标准分均超过64的数1A和数2B分别获得58、56的成绩，表现不佳。

另外，记述式考试中，理科数学标准分为76，文科数学标准分为68，成绩优良。去年分别为44和59。

据了解，日本国立信息学研究所等研究小组力争最晚在2021年度通过东大入学考试，于2011年开始这一项目。

虽然成绩不断提高，但根据迄今为止研究的结果，开发小组认为难以达到考入东大的水平。今后，还将推进数学的记述式考试成绩改善研究以及成果运用于儿童教育的研究。

NLP技术在应用中的需求

Tasks

Dependency on NLP

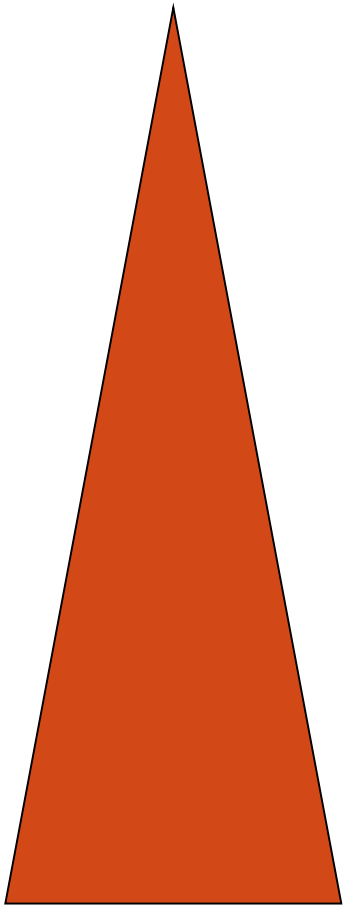
“Easier” & More “workarounds”

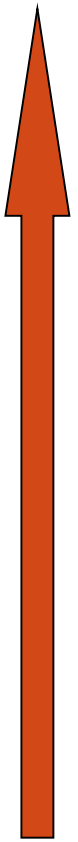
Classification/ Retrieval

Summarization/ Extraction/ Mining

Translation/ Dialogue

Question Answering





自然语言处理为什么如此之难？

- 自然语言与生俱有的歧义问题
  - 分词

“能穿多少穿多少”

Ambiguity

冬天： “能穿[多少]穿[多少]”

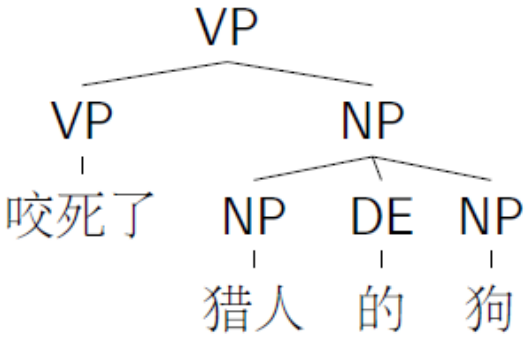
夏天： “能穿[多/少]穿[多/少]”

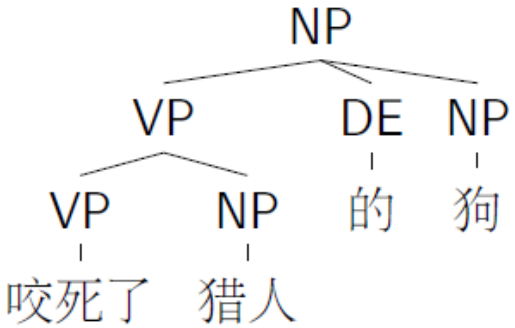
自然语言处理为什么如此之难？

- 自然语言与生俱有的歧义问题
  - 句法

“咬死了猎人的狗”

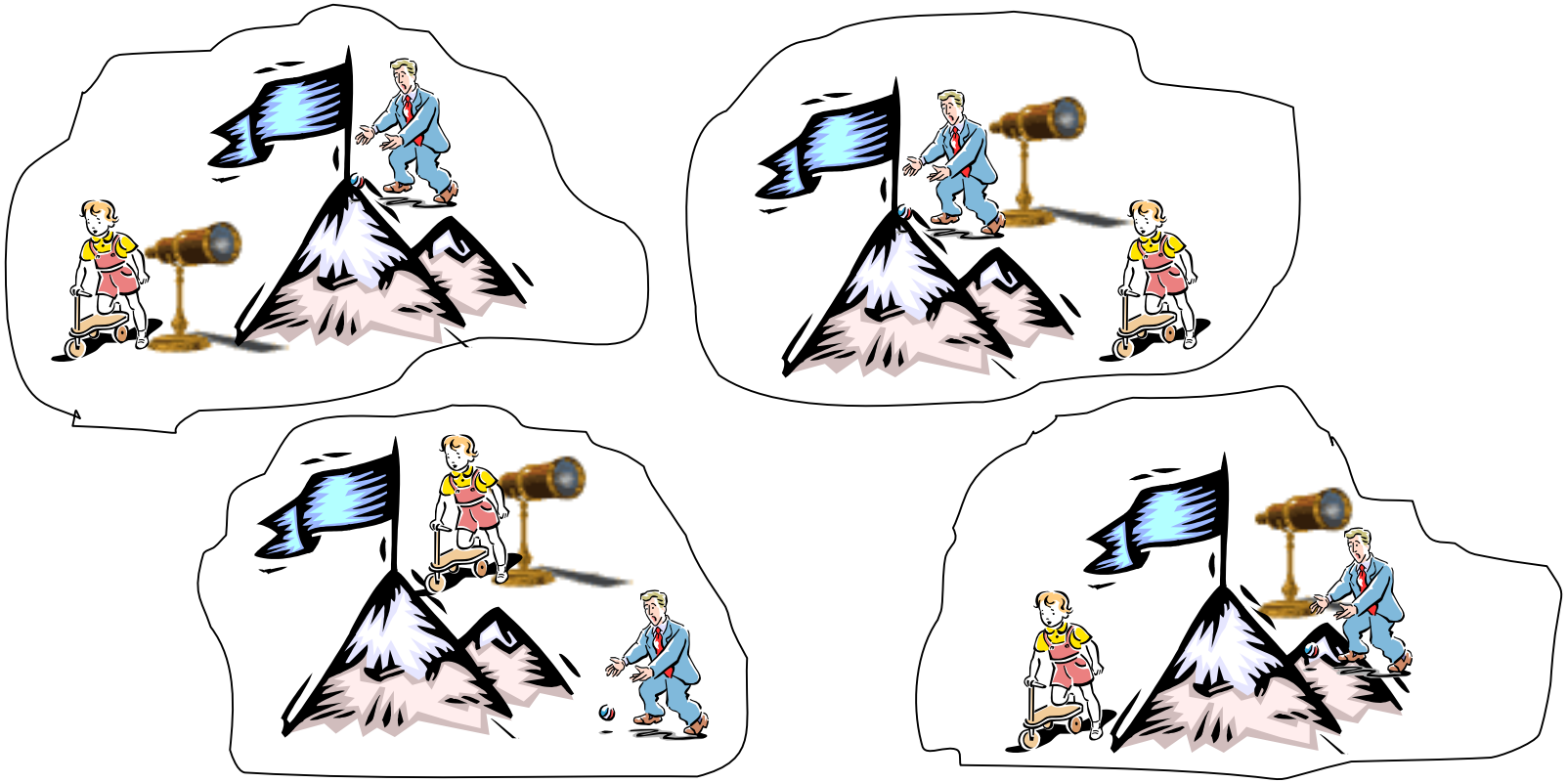
Ambiguity





The boy saw the man on the mountain with a telescope

PP attachment



Ambiguity

自然语言处理为什么如此之难？

- 自然语言与生俱有的歧义问题
  - 语义

他说："她这个人真有意思(funny)"。她说："他这个人怪有意思的(funny)"。于是人们以为他们有了意思(wish)，并让他向她意思意思(express)。他火了："我根本没有那个意思(thought)"！她也生气了："你们这么说是什么意思(intention)"？事后有人说："真有意思(funny)"。也有人说："真没意思(nonsense)"。

——《生活报》1994. 11. 13. 第六版

Ambiguity

自然语言处理为什么如此之难？

- 自然语言与生俱有的歧义问题
  - 语义

曾经喜欢一个人，如今喜欢一个人。

Ambiguity

自然语言处理为什么如此之难？

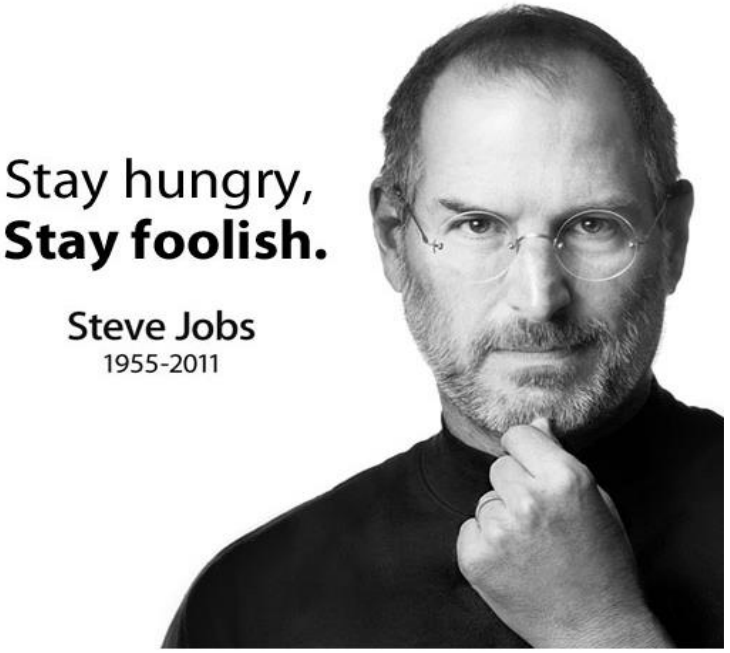
- 自然语言与生俱有的歧义问题
  - 语用

该来的没来。

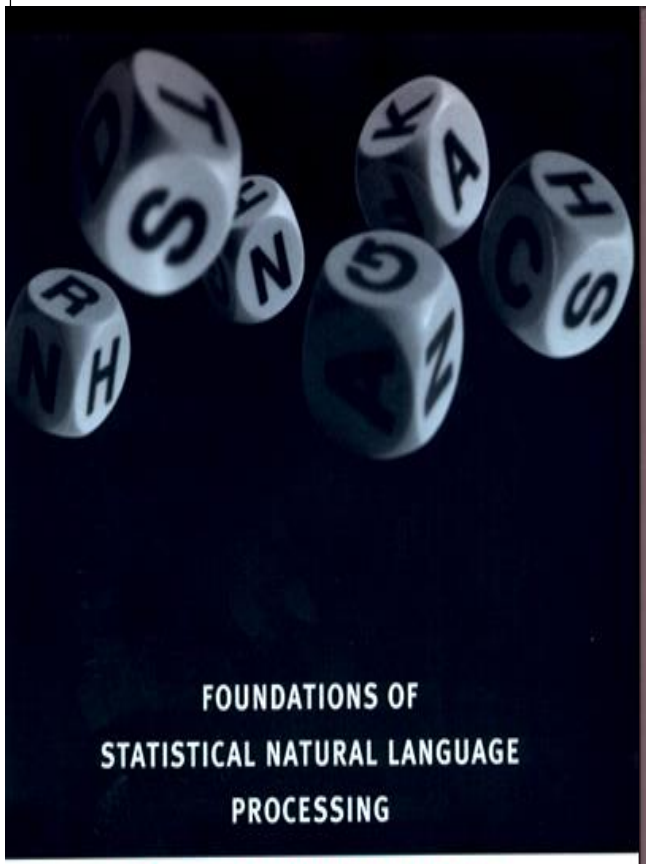
Ambiguity

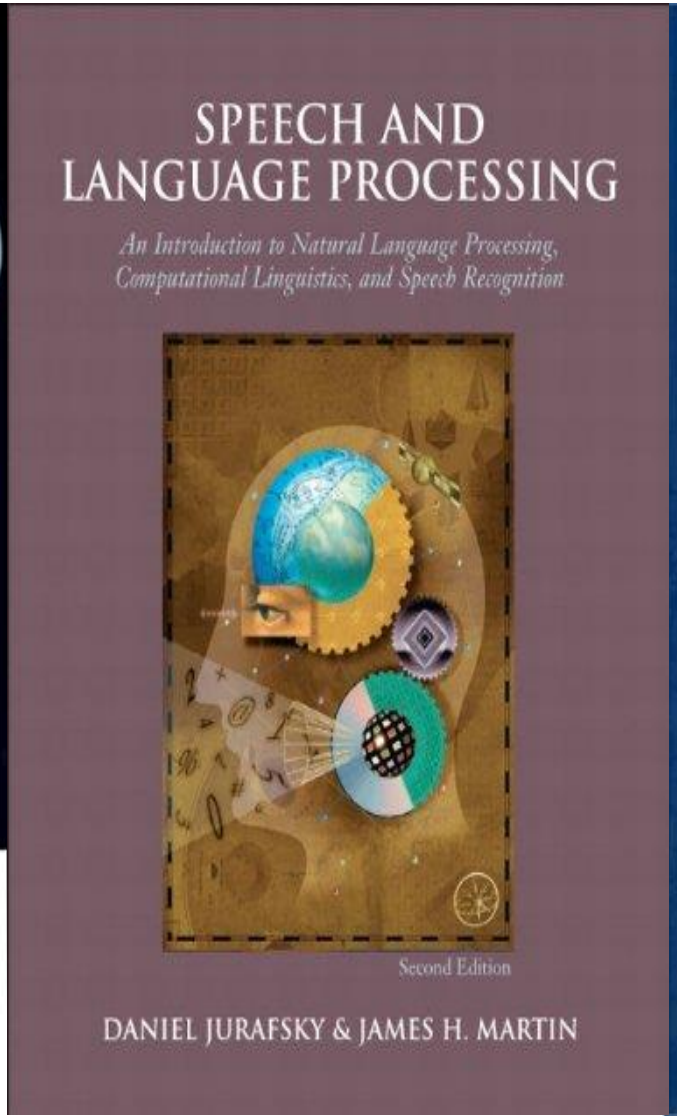
自然语言处理为什么如此之难？

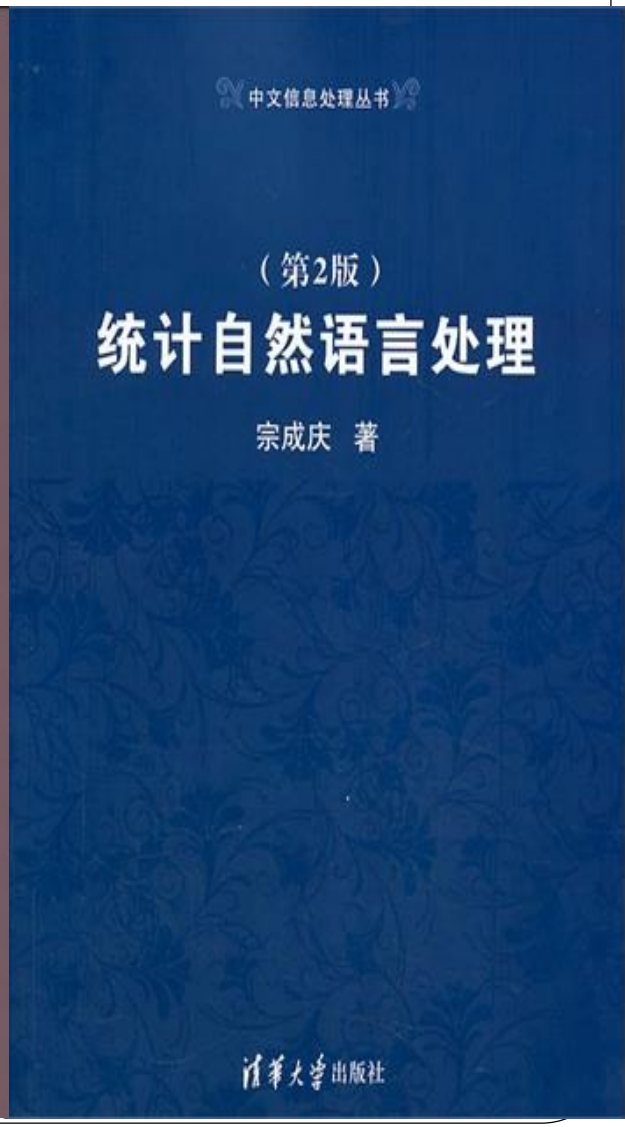
- 自然语言与生俱有的歧义问题



推荐教材


FOUNDATIONS OF
STATISTICAL NATURAL LANGUAGE
PROCESSING
CHRISTOPHER D. MANNING AND
HINRICH SCHÜTZE


SPEECH AND
LANGUAGE PROCESSING
An Introduction to Natural Language Processing,
Computational Linguistics, and Speech Recognition
Second Edition
DANIEL JURAFSKY & JAMES H. MARTIN


中文信息处理丛书
（第2版）
统计自然语言处理
宗成庆 著
清华大学出版社

汉语自动分词

分词

- **一般认为：词是最小的、能够独立运用的、有意义的语言单位。**
- **来自微博的分词实例**

【/点评团/消费/】/好久/没/吃/过/这么/划算/的/团购/了/?/这/也/是/我/有史以来/最/成功/的/一/次/团购/?/感谢/大众/点评/环境/很/日本/?

汉语分词的挑战

- 词和词组的边界模糊
  - “对随地吐痰者给予处罚” =》词 or 短语 ？
- 新词
  - 虽然一般的词典都能覆盖大多数的词语，但有相当一部分的词语不可能穷尽地收入系统词典中，这些词语称为未登录词或新词
  - 分类：
    - 专有名词：中文人名、地名、机构名称、外国译名、时间词
    - 重叠词：“高高兴兴”、“研究研究”
    - 派生词：“*一次性*筷子”
    - 与领域相关的术语：“互联网”
    - 网络用语：打酱油、犀利哥、清二、…
- 切分歧义

切分歧义

- 中文分词之交集型切分歧义
  - 汉字串AJB被称作交集型切分歧义，如果满足AJ、JB同时为词(A、J、B分别为汉字串)。此时汉字串J被称作交集串。
    - [例] “结合成分子”
      - 结合 | 成 分|子 |
      - 结合|成|分子|
      - 结 | 合成 |分子|
    - [例] “美国会通过对台售武法案”

切分歧义

- 组合型切分歧义
  - 汉字串AB被称作组合型切分歧义，如果满足条件：A、B、AB同时为词
    - [例]组合型切分歧义：“起身”
    - 他站|起|身|来。
    - 他明天|起身|去北京。
- 真歧义
  - 存在两种或两种以上的真实存在的切分形式
  - 乒乓球|拍卖|完|了
  - 乒乓|球拍|卖|完|了

分词方法

- 简单的模式匹配
  - 正向最大匹配、逆向最大匹配法、双向匹配法
- 基于规则的方法
  - 最少分词算法
- 基于统计的方法
  - 统计语言模型分词、串频统计和词形匹配相结合的汉语自动分词、无词典分词

分词方法

- **正向最大匹配分词(FMM)**

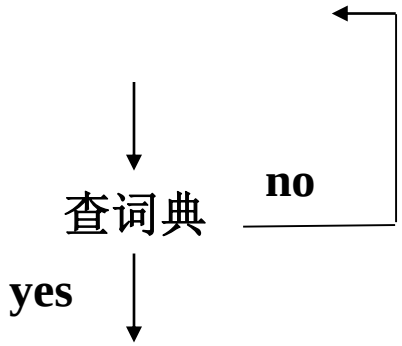
1. 设自动分词词典中最长词条所含汉字个数为I；
2. 取被处理材料当前字符串序数中的I个字作为匹配字段，查找分词词典。若词典中有这样的一个I字词，则匹配成功，匹配字段作为一个词被切分出来，转6；
3. 如果词典中找不到这样的一个I字词，则匹配失败；
4. 匹配字段去掉最后一个汉字，I--；
5. 重复2-4，直至切分成功为止；
6. I重新赋初值，转2，直到切分出所有词为止。

分词方法

- **正向最大匹配分词(FMM)**

输入字串：

TmpWord：



输出词串：　时间/　就/　是/　生命/

分词方法

- 逆向最大匹配分词(BMM)
  - 分词过程与FMM方法相同，不过是从句子(或文章)末尾开始处理，每次匹配不成功时去掉的是前面的一个汉字
  - 实验表明：逆向最大匹配法比最大匹配法更有效，错误切分率为1 / 245

分词方法

- 双向匹配法(BM)
  - 比较FMM法与BMM法的切分结果，从而决定正确的切分
    - 大颗粒度词越多越好，非词典词和单字词越少越好
  - 可以识别出分词中的交叉歧义

FMM的结果：
“我们/在野/生动/物/园/玩”

BMM的结果：
“我们/在/野生动物园/玩”

更优！！

分词方法

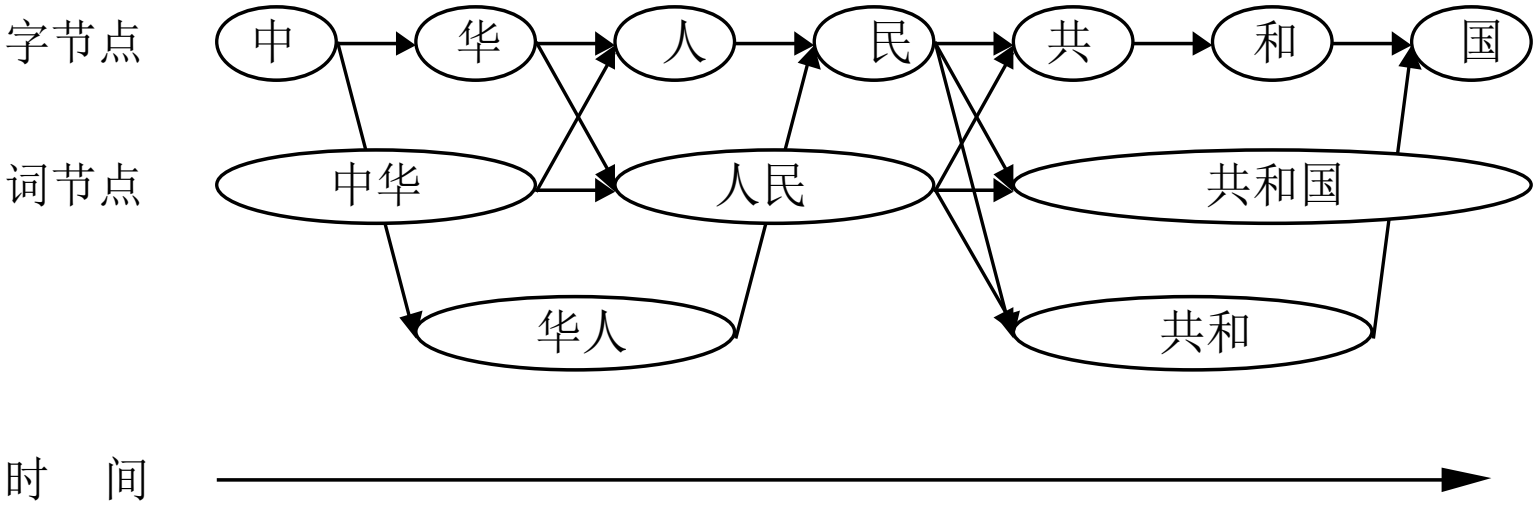
- 基于统计的词网格分词
  - 第一步是候选词网格构造：利用词典匹配，列举输入句子所有可能的切分词语，并以词网格形式保存
  - 第二步计算词网格中的每一条路径的权值，权值通过计算图中每一个节点（每一个词）的一元统计概率和节点之间的二元统计概率的相关信息
  - 根据图搜索算法在图中找到一条权值最大的路径，作为最后的分词结果

可利用不同的统计语言模型计算最优路径
具有比较高的分词正确率
算法时间、空间复杂性较高

分词方法

- **基于统计的词网格分词**

字串“中华人民共和国”的切分词网格



分词方法

- 基于字分类的分词
  - 赋予每个汉字一个相对于词的位置类别
    - B: 当前字是一个词的首字，且该词为多字词
    - E: 当前字是一个词的尾字，且该词为多字词
    - I: 当前字是一个词的中间字
    - S: 当前字组成一个单字词

赵	紫	阳	总	理	的	秘	密	日	记
B	I	E	B	E	S	B	E	B	E

可利用统计模型获得最优的类别序列：例如序列标注模型。

分词工具

- 多种工具、各有优缺点
- 效果通常在95%以上

- http://blog.csdn.net/hello9050/article/details/7889658

词性标注

词性标注(POS Tagging)

- 为句子中的每个词语标注词性(part-of-speech marker).

John saw the saw and decided to take it to the table.
NNP VBD DT NN CC VBD TO VB PRP IN DT NN

- 可看做是词法分析的关键任务，也可看做是句法分析的最低层次

- 对后续句法分析、词义消岐等任务非常有用

英文POS集合

- 原始Brown语料使用了87个POS tags
- 目前NLP领域常用的是Penn Treebank set，包含45 tags

英文POS集合

- Noun (person, place or thing)
  - Singular (NN): dog, fork
  - Plural (NNS): dogs, forks
  - Proper (NNP, NNPS): John, Springfields
  - Personal pronoun (PRP): I, you, he, she, it
  - Wh-pronoun (WP): who, what
- Verb (actions and processes)
  - Base, infinitive (VB): eat
  - Past tense (VBD): ate
  - Gerund (VBG): eating
  - Past participle (VBN): eaten
  - Non 3rd person singular present tense (VBP): eat
  - 3rd person singular present tense: (VBZ): eats
  - Modal (MD): should, can
  - To (TO): to (to eat)

英文POS集合

- Adjective (modify nouns)
  - Basic (JJ): red, tall
  - Comparative (JJR): redder, taller
  - Superlative (JJS): reddest, tallest
- Adverb (modify verbs)
  - Basic (RB): quickly
  - Comparative (RBR): quicker
  - Superlative (RBS): quickest
- Preposition (IN): on, in, by, to, with
- Determiner:
  - Basic (DT) a, an, the
  - WH-determiner (WDT): which, that
- Coordinating Conjunction (CC): and, but, or,
- Particle (RP, 小品词): off (took off), up (put up)

词性标注中的歧义

- **“Like” can be a verb or a preposition**
  - I like/VBP candy.
  - Time flies like/IN an arrow.

- **“Around” can be a preposition, particle, or adverb**
  - I bought it at the shop around/IN the corner.
  - I never got around/RP to getting a car.
  - A new Prius costs around/RB $25K.

词性标注中的歧义

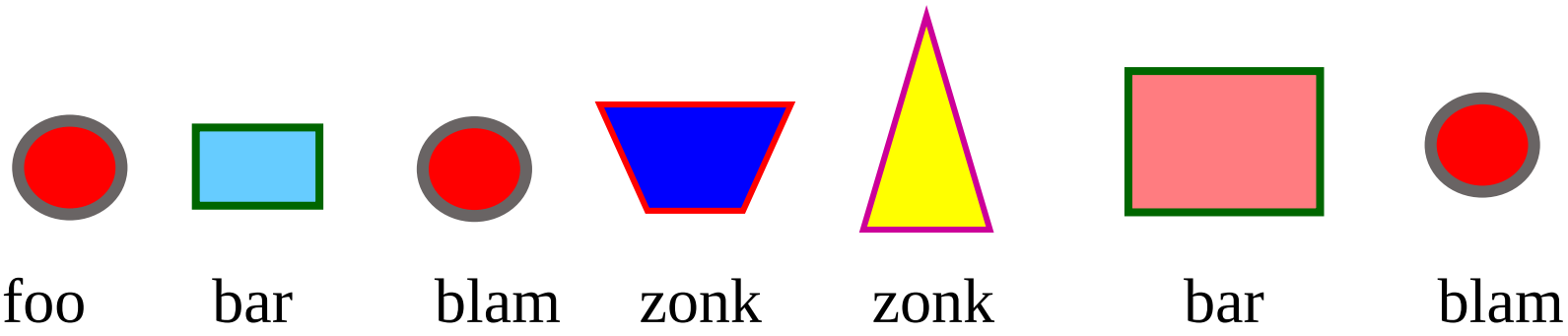
- 英文中的歧义程度 (基于Brown corpus)
  - 11.5% 的词形(word type)有歧义
  - 40%的词例(word token)有歧义
- Penn Treebank中人工标注词性的平均不一致性为3.5%
- Baseline: 为每个词选择该词最常用的词性标记，准确性（accuracy）达到90%

词性标注方法

- 基于规则(Rule-Based): 人工基于词汇与其他语言知识构造标注规则
- 基于学习(Learning-Based): 基于人工标注语料（例如Penn Treebank）进行训练
  - 统计模型(Statistical models): Hidden Markov Model (HMM), Maximum Entropy Markov Model (MEMM), Conditional Random Field (CRF)
  - 规则学习(Rule learning): Transformation Based Learning (TBL)
- 总体上, 基于学习的方法更加有效

看作序列标注问题(Sequence Labeling)

- 很多NLP问题都可看做序列标注问题
- 为序列中每个符号赋予一个标签
- 符号标签依赖于其他符号的标签，尤其是相邻符号的标签 (not i.i.d)



Training data (Annotated text)

This	*sentence*	*serves*	*as*	*an*	*example*	*of*
Det	N	V1	P	Det	N	P
annotated	*text…*					
V2	N					

"This is a new sentence" → **POS Tagger** → *This is a new sentence* / Det Aux Det Adj N

考虑所有的可能性，
快速选择具有最大概率的序列。

This	*is*	*a*	*new*	*sentence*
Det	Det	Det	Det	Det
…	…			
Det	Aux	Det	Adj	N
…	…			
V2	V2	V2	V2	V2

$$p(w_1,\ldots,w_k,t_1,\ldots,t_k) = \begin{cases} p(t_1 \mid w_1)\ldots p(t_k \mid w_k)\, p(w_1)\ldots p(w_k) \\ \prod_{i=1}^{k} p(w_i \mid t_i)\, p(t_i \mid t_{i-1}) \end{cases}$$

Partial dependency

集成序列中多个相互依赖的个体分类的不确定性，统一确定最可能的全局标签判断

HMM POS Taggers

- 使用标注POS的语料库能够很方便地基于有监督学习构建HMM模型
- 给定一个新的未标注文本（词语序列），使用Viterbi 算法快速预测最优词性标签序列
- 现代POS标注器，包括HMM在内，准确性为96-97% (for Penn tagset trained on about 800K words)

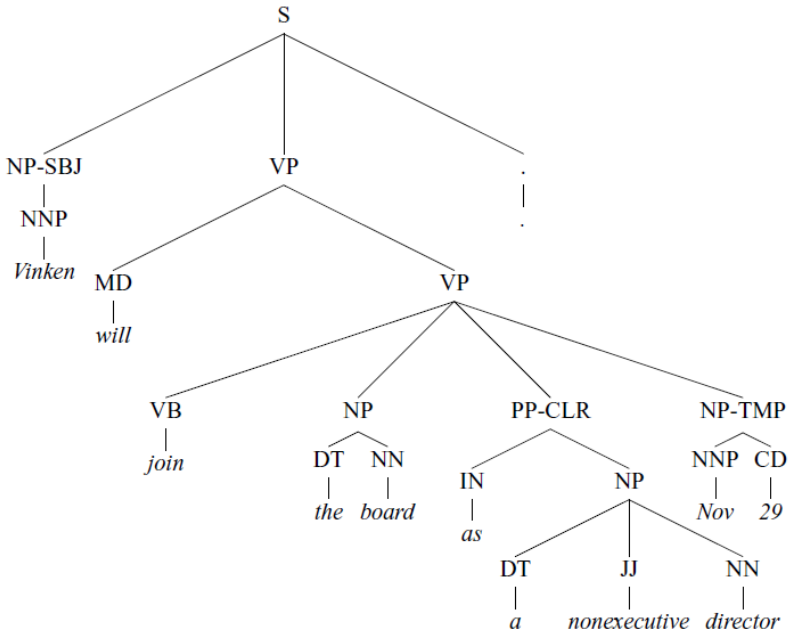
中文词法分析效果

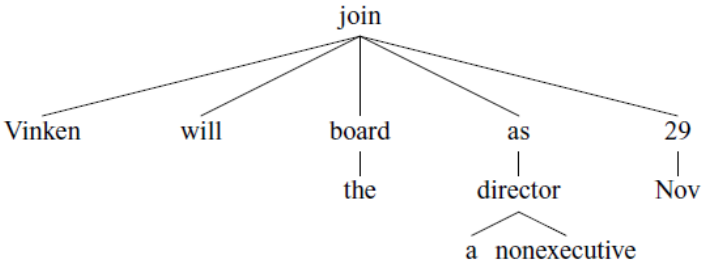
- 中文分词总体水平(F值)达到95%左右
  - 主要分词错误由新词造成
  - 命名实体识别效果偏低，尤其是机构名识别
  - 效果跟文本类型有关
- 中文词性标注总体水平(Accuracy)超过90%

句法分析

句法分析类型

- 句法/成分/短语结构分析 vs. 依存关系分析





- 完全分析 vs. 局部分析/浅层分析
  - 浅层分析：如组块分析(Chunking)
    - 找出句子中非递归的名词短语与动词短语等
    - [NP I] [VP ate] [NP the spaghetti] [PP with] [NP meatballs].

基于序列标注的组块分析

- 对于每类短语一般用三个标签对每个词进行标记
  - B (Begin)：表示是目标短语的起始词
  - I (Inside)：表示属于目标短语的一部分，但不是起始词
  - O (Other)：表示不属于目标短语的一部分
- 最好的Chunking结果一般F值 >90%
- NP chunking举例
  - He reckons the current account deficit will narrow to only # 1.8 billion in September.

Begin Inside Other

Context Free Grammars (CFG)

- N:非终结符集合(或者变量集)
- Σ：终结符集合(与N无交集)
- R：重写规则/产生式集合 (a set of productions or rules), 形式为 A→β, 其中 A 为非终结符，β为来自 (Σ∪ N)*中的符号串
- S: 一个特别的非终止符，称为初始符 (start)

Simple CFG for ATIS English

Grammar

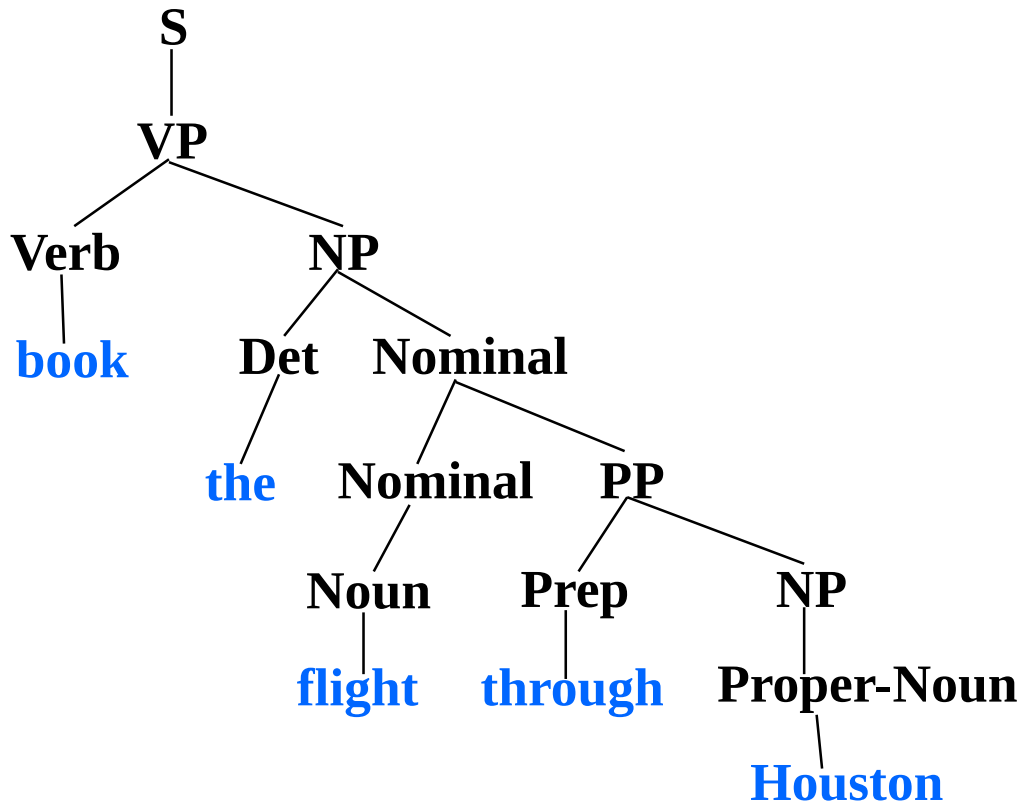
S → NP VP
S → Aux NP VP
S → VP
NP → Pronoun
NP → Proper-Noun
NP → Det Nominal
Nominal → Noun
Nominal → Nominal Noun
Nominal → Nominal PP
VP → Verb
VP → Verb NP
VP → VP PP
PP → Prep NP

Lexicon

Det → the | a | that | this
Noun → book | flight | meal | money
Verb → book | include | prefer
Pronoun → I | he | she | me
Proper-Noun → Houston | NWA
Aux → does
Prep → from | to | on | near | through

句子生成

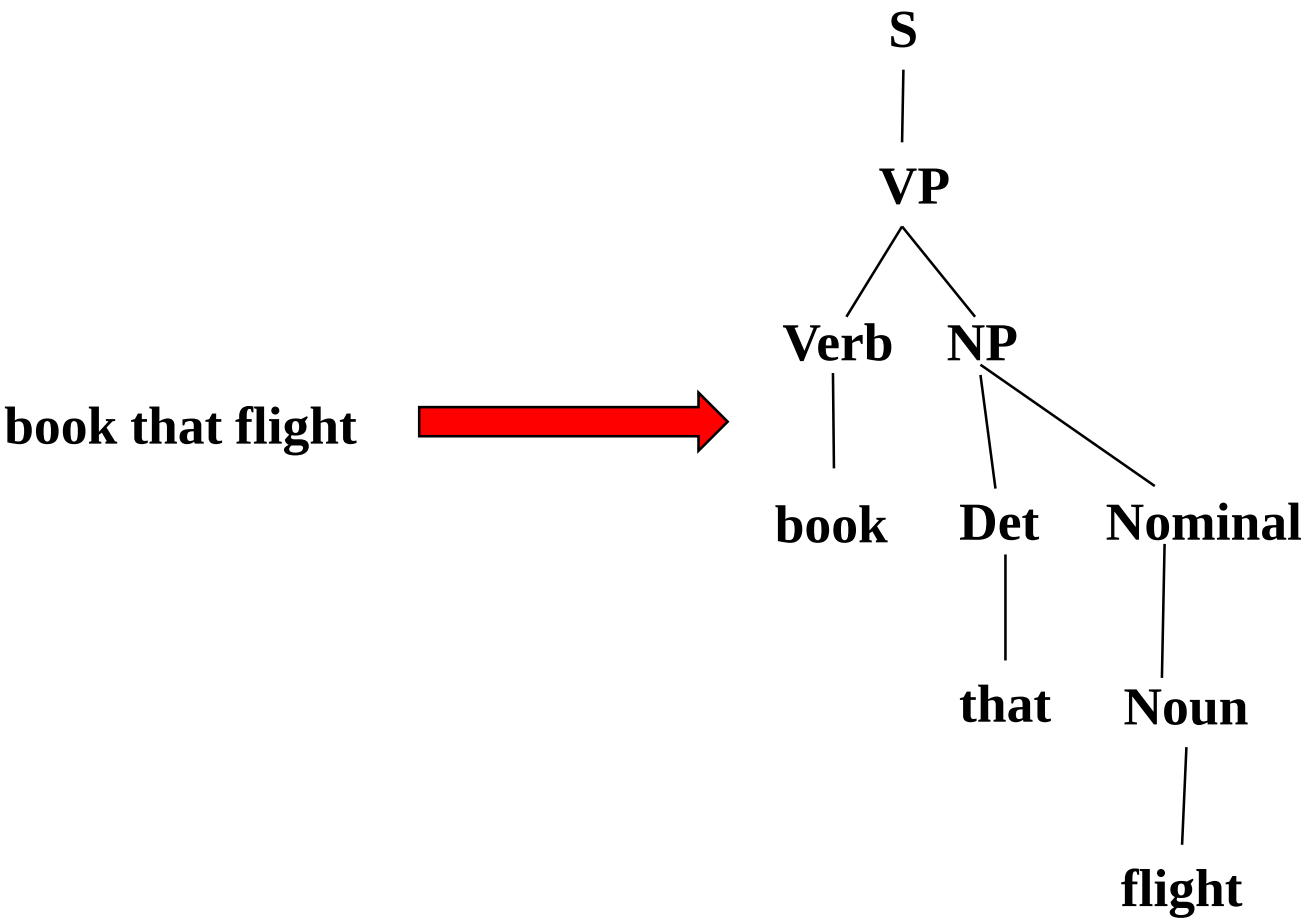
- 一个句子的生成过程：通过递归地将句法规则左侧起始符号改写为右侧生成符号，直到只剩下终结符为止



句法结构分析(Parsing)

- 给定一串终结符号和一个CFG，确定该符号串是否能够被该CFG所生成
  - 同时为该符号串返回句法树
- 必须进行搜索以获得句法树推导
  - Top-Down Parsing: 从初始符开始
  - Bottom-up Parsing: 从符号串中的终结符开始

Parsing Example



句法结构歧义

- 一个句子可能对应多个句法树，典型歧义包括
  - 连接歧义(Attachment ambiguity)
    - E.g. I shot an elephant in my pajamas.
  - 并列歧义(Coordination ambiguity)
    - E.g. old men and women
  - 名词短语括号歧义(Noun-phrase bracketing ambiguity)
    - E.g. complete peace plan

基于动态规划的句法分析算法

- CKY (Cocke-Kasami-Younger) algorithm
  - 基于自底向上分析，需要对句法进行规范化
- Earley parser
  - 基于自顶向下分析，不需要句法规范化，但更加复杂
- Chart parser
  - 融合自顶向下与自底向上搜索

统计句法分析

- 使用语法概率模型为每棵句法树计算概率值
  - 选择概率最大的句法树
- 允许使用有监督学习和无监督学习得到句法分析模型

Simple PCFG for ATIS English

Grammar	Prob	
S → NP VP	0.8	
S → Aux NP VP	0.1	+ 1.0
S → VP	0.1	
NP → Pronoun	0.2	
NP → Proper-Noun	0.2	+ 1.0
NP → Det Nominal	0.6	
Nominal → Noun	0.3	
Nominal → Nominal Noun	0.2	+ 1.0
Nominal → Nominal PP	0.5	
VP → Verb	0.2	
VP → Verb NP	0.5	+ 1.0
VP → VP PP	0.3	
PP → Prep NP	1.0	

Lexicon

Det → the | a | that | this
0.6 0.2 0.1 0.1

Noun → book | flight | meal | money
0.1 0.5 0.2 0.2

Verb → book | include | prefer
0.5 0.2 0.3

Pronoun → I | he | she | me
0.5 0.1 0.1 0.3

Proper-Noun → Houston | NWA
0.8 0.2

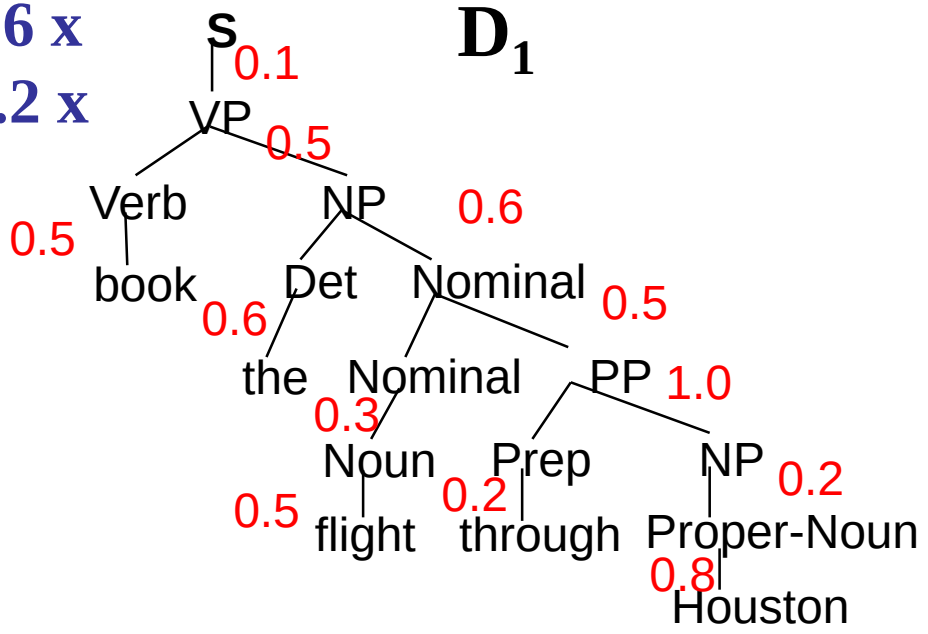
Aux → does
1.0

Prep → from | to | on | near | through
0.25 0.25 0.1 0.2 0.2

句法推导概率

- 假设句法树上每个产生式的选择是独立的
- 句法推导的概率为所有产生式概率之乘积

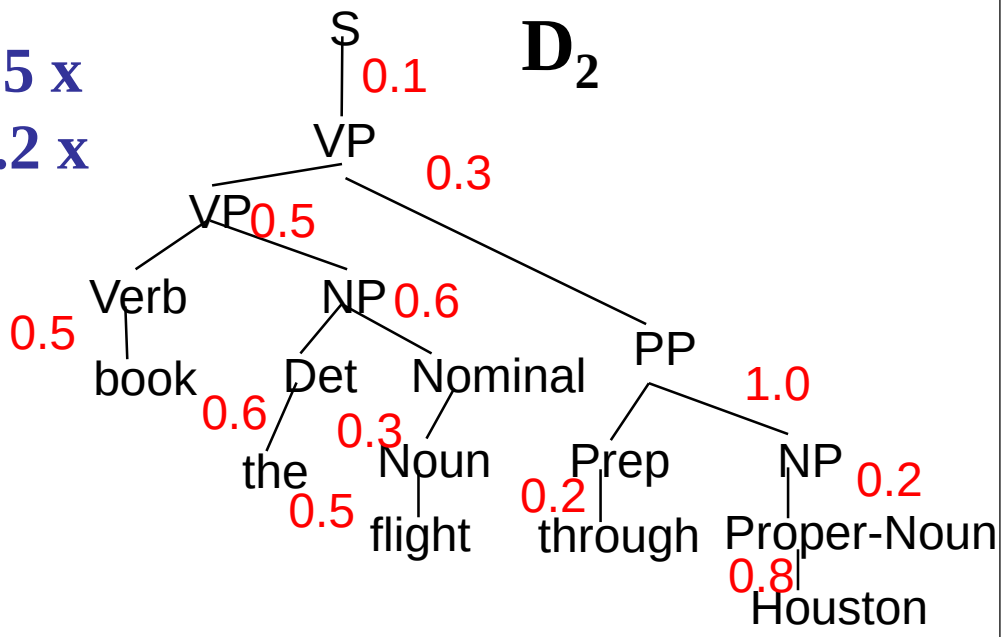
$$P(D_1) = 0.1 \times 0.5 \times 0.5 \times 0.6 \times 0.6 \times 0.5 \times 0.3 \times 1.0 \times 0.2 \times 0.2 \times 0.5 \times 0.8 = 0.0000216$$



句法消岐

- 为句子选择可能性（概率）最大的句法树

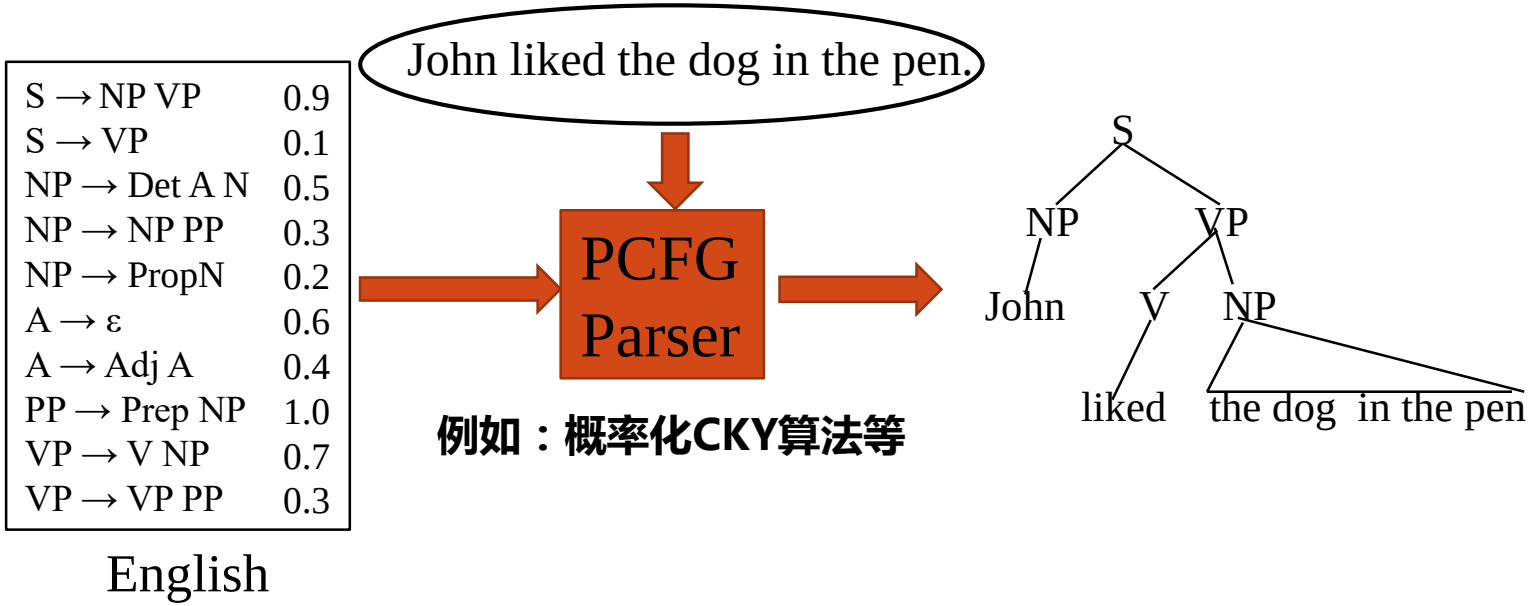
$$P(D_2) = 0.1 \times 0.3 \times 0.5 \times 0.6 \times 0.5 \times 0.6 \times 0.3 \times 1.0 \times 0.5 \times 0.2 \times 0.2 \times 0.8 = 0.00001296$$



$$P(D_1) > P(D_2)$$

最佳句法树推导

- 为一个句子确定最佳的句法树
- 类似Viterbi算法，为给定句子高效率确定可能性最大的句法树



中文句法分析效果

- 短语结构分析的总体水平(基于短语匹配的F值)超过80%
- 依存分析的总体水平（无标记依存正确率）接近90%，带标记的依存正确率会下降

Acknowledgements

- Some slides were taken or adapted from related slides written by Raymond Mooney, Lucas Champollion, Rohit Kate, Scott Yih, Kristina Toutanova, Stina Ericsson, George A. Miller, Cosmin Adrian Bejan, Marian Olteanu, Giuseppe Carenini, Pu Wang, Keith Trnka, Danushka Bollegala, Ted Pedersen, Rada Mihalcea, Chengqing Zong, etc. Thank them for sharing their slides.

QA: wanxiaojun (AT) pku.edu.cn