词性标注任务描述

- **什么叫词性?**
  - 词性又称词类，是指词的语法分类，或者说是按照其各自的语法功能的不同而分出来的类别
- **划分词类的依据**
  - 词的形态、词的语法功能、词的语法意义

英语词的分类

- 开放类（open class）
  - Nouns
    - 句法上：可作物主、可有限定词、有复数形式
    - 语义上：人名、地名和物名
  - Verbs
    - 句法上：作谓语、有几种词形变化
    - 语义上：动作、过程（一系列动作）
  - Adjectives
    - 句法上：修饰Nouns等
    - 语义上：性质
  - Adverbs
    - 句法上：修饰Verbs等
    - 语义上：方向、程度、方式、时间

- 封闭类（closed class，function words）
  - Determiners
  - Pronouns
  - Prepositions
  - Conjunctions
  - Auxiliary verbs
  - Particles（if、not、...）
  - Numerals

词性标注任务描述

- **词性标注：给某种语言的词标注上其所属的词类**
  - The lead paint is unsafe.
  - The/Det lead/N paint/N is/V unsafe/Adj.
  - 他有较强的领导才能。
  - 他/代词 有/动词 较/副词 强/形容词 的/助词 领导/名词 才能/名词。

词性标注歧义（兼类词）

- 一个词具有两个或者两个以上的词性
- 英文的**Brown**语料库中，**10.4%**的词是兼类词
  - The back door
  - On my back
  - Promise to back the bill
- 汉语兼类词
  - 把门锁上，　买了一把锁
  - 他研究与自然语言处理相关的研究工作
  - 汉语词类确定的特殊难点
- 对兼类词消歧 – 词性标注的任务

词性标注常见方法

- **规则方法：**
  - 词典提供候选词性
  - 人工整理标注规则
- **统计方法**
  - 寻找概率最大的标注序列
  - 如何建立统计模型
  - HMM方法
  - 最大熵方法
  - 条件随机场方法
  - 结构化支持向量机方法
- **基于错误驱动的方法**
  - 错误驱动学习规则
  - 利用规则重新标注词性

词性标注的性能指标

- **性能指标：标注准确率**
- **当前方法正确率可以达到97%**
- **正确率基线(Baseline)可以达到90%**
  - 基线的做法：
    - 给每个词标上它最常见的词性
    - 所有的未登录词标上名词词性

回顾

- 分词
  - 根据字典中是否存在该词来决定
  - “阵风”：根据前面是否有数词来消歧。“一/阵/风/吹/过/来”、“今天/有/阵风”
- 词性标注
  - 根据字典中的词性进行直接标注（90+%）
  - 把门锁上，　买了一把锁
- 如何对不确定的情况进行判断?

- **是否可以用分类问题进行建模?**
  - 为每个单元（字、词）预测一个标记

  - 如何确定输出?
  - 如何确定输入?
  - 如何获得样例?
  - 如何评价结果?

Neinwen Xue, Susan P. Converse Combining classifiers for Chinese word segmentation,2002
黄昌宁,中文分词十年回顾,2007
赵海 等,中文分词十年又回顾: 2007-2017

- **如何确定输出?**
  - 标记是什么?

Table IX. Tag Sets Employed in the State-of-the-Art Chinese Word Segmentation Systems

Four-tag set Low/(Xue)		Three-tag set Zhang		Two-tag set Peng/Tseng	
Function	Tag	Function	Tag	Function	Tag
Begin	B(LL)	Begin	B	Start	Start
Middle	M(MM)	Middle or end	I	Continue	NoStart
End	E(RR)				
Single	S(LR)	Single	O		

今天 有 阵风
B E S B E
B I O B I
S C S S C

- 如何表示输入？
  - 用什么样的信息来决定输出

Table VI. Basic Feature Templates Set

Code	Type	Feature	Function
(a)	Unigram	C_{-1}, C_0, C_1	Previous, current, or next character
(b)	Bigram	$C_{-1}C_0$	Previous and current characters
		C_0C_1	Current and next characters
		$C_{-1}C_1$	Previous and next characters

今	天	有	阵	风
-2	-1	0	1	2

x: C-2=今，C-1=天，C0=有，C1=阵，C2=风
y: B/S/Start

还存在什么问题?

- **决策独立性**
  - 开头为E；结尾为B；连续的E E？

- **决策关联**
  - 形容词修饰名词；副词修饰动词

今	天	有	阵	风
-2	-1	0	1	2
B	E	S	E?	B?
B	E	S	S?	S?

  - 决策也应该作为输入的依据！ (x,y)?
  - 先做哪一个决策?

序列化标注问题

- 输出为一系列相互存在关联的决策整体

X: 今 天 有 阵 风
Y: B E S B E

- 该决策Y由一系列子决策y_i构成
  - 后做的决策依赖于先做的决策
- 找到这个Y的整体是一个搜索问题
  - 自左向右
  - 自右向左
  - 从易到难
  - Coarse-to-Fine

序列化标记问题的应用

- 分词

习 近 平 向 2019 年 中 国 国 际 服 务 贸 易 交 易 会 致 贺 信

B M E S S S B E B E B E B E B M E S B E

习近平 向 2019 年 中国 国际 服务 贸易 交易会 致 贺信

- 词性标记

习近平 向 2019 年 中国 国际 服务 贸易 交易会 致 贺信

Noun P Num Q N N N N N V N

- 命名实体识别

习近平 向 2019 年 中国 国际 服务 贸易 交易会 致 贺信

S-Per O B-Ti E-Ti B-Org … M-Org … E-Org O O

习近平 向 2019 年 中国 国际 服务 贸易 交易会 致 贺信

- 评论要素识别（Sentiment Aspect Identification）

词	手机通话声音清晰，相机聚焦脸部非常准确，又把我拍得太帅了！
标注	N N S M M E N N　S E N N N N N N N N　N N N N N N N N N N

- 阅读理解
  - B-Ans
  - M-Ans
  - E-Ans
  - Others

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

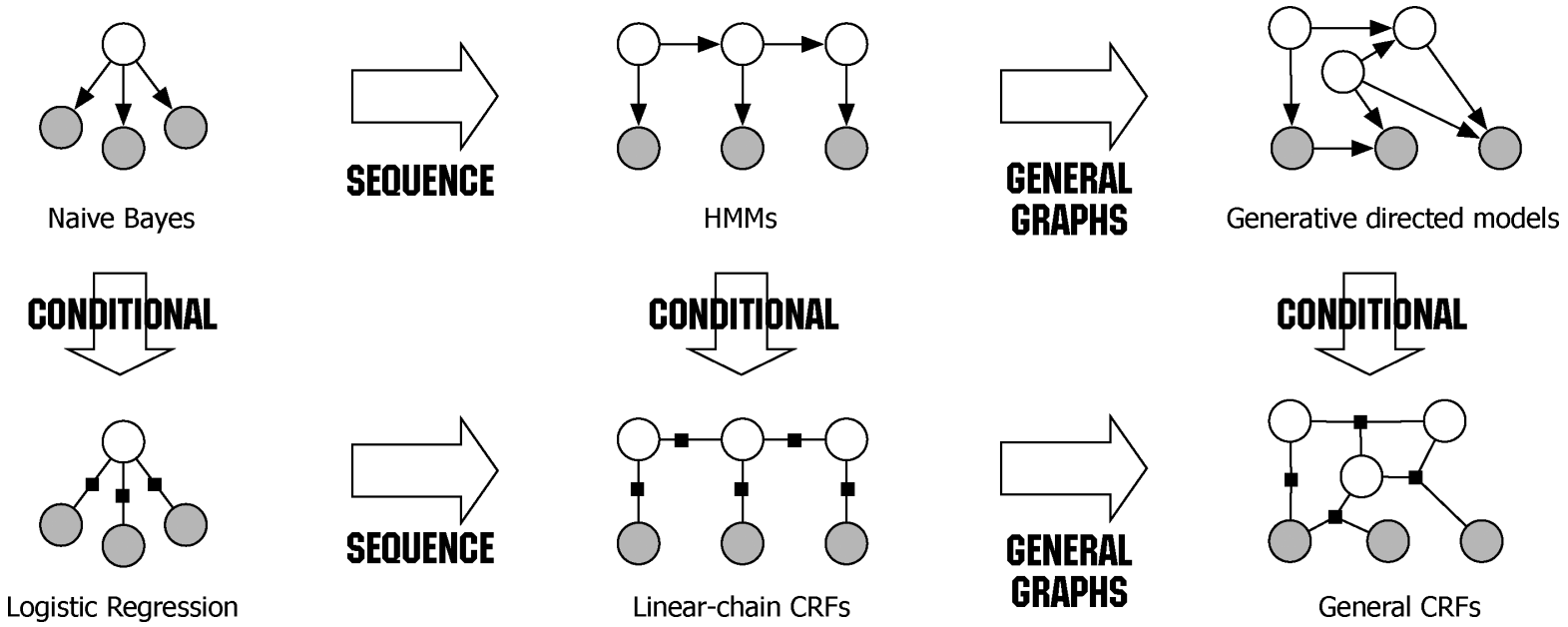
What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

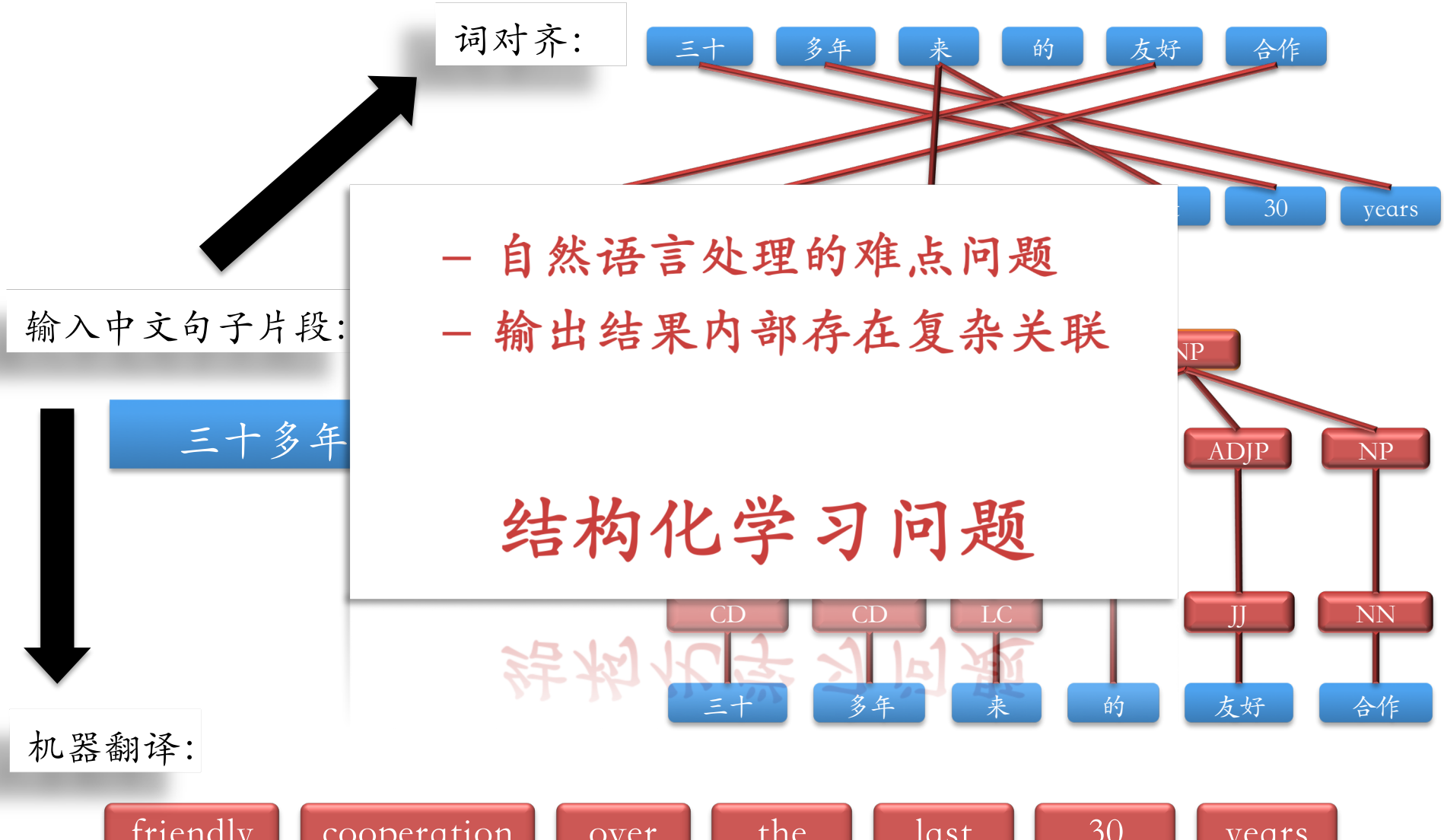
SQuAD: 100,000+ Questions for Machine Comprehension of Text Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, Percy Liang

- 概率图模型（Probabilistic Graphic Models）



Sutton, Charles, and Andrew McCallum. "An introduction to conditional random fields." Machine Learning 4.4 (2011): 267-373.

更复杂的问题：


词对齐：
三十
多年
来
的
友好
合作
30
years
输入中文句子片段：
三十多年
– 自然语言处理的难点问题
– 输出结果内部存在复杂关联
结构化学习问题
NP
ADJP
NP
JJ
NN
CD
CD
LC
三十
多年
来
的
友好
合作
机器翻译：
friendly
cooperation
over
the
last
30
years