语义角色标注

孙栩

信息科学技术学院

xusun@pku.edu.cn

http://xusun.org

大纲

- **语义角色标注**
- **PropBank与FrameNet**
- **解决方案**
  - 句法树方法
  - 序列标注方法

大纲

- 语义角色标注
- PropBank与FrameNet
- 解决方案
  - 句法树方法
  - 序列标注方法

语义

- 这些句子是否有同样的含义?

- Yesterday, Kristina hit Scott with a baseball
- Scott was hit by Kristina yesterday with a baseball
- Yesterday, Scott was hit with a baseball by Kristina
- With a baseball, Kristina hit Scott yesterday
- Yesterday Scott was hit by Kristina with a baseball
- Kristina hit Scott with a baseball yesterday

语义

- **何为语义?**
  - 哲学性问题，目前语言学领域未有定论
  - 提出了众多的语义表示方法

- **代表理论：一阶谓词逻辑[Neo-Davidsonian事件表示]**
  - 对一个事件的形式化表示(一阶谓词逻辑)，例如
    - Sasha broke the window
      - $\exists e, x, y\ Breaking(e) \wedge Breaker(e, Sasha) \wedge BrokenThing(e, y) \wedge Window(y)$
    - Pat opened the door
      - $\exists e, x, y\ Opening(e) \wedge Opener(e, Pat) \wedge OpenedThing(e, y) \wedge Door(y)$
  - 谓词需要人工定义、且无法穷尽
  - 这种表示很难分析得到，更难以进行有效推理

一阶谓词逻辑 到 语义角色标注

- **一阶谓词逻辑没有考虑语义的共性**
  - Breaker和Opener虽然对应了不同的事件，但有语义共同之处
    - 主动行动者(volitional actor)
    - 有生命的(animate)
    - 事件的直接原因(direct causal responsibility)

- **语义角色(semantic roles)**
  - 通过捕捉语义间的共性，降低分析的难度和复杂度
  - 在上一例子中，两者可以统一：
    - *Breaker* 和 *Opener* 都是 **AGENTS（施事）**
    - *BrokenThing* 和 *OpenedThing* 都是 **THEMES（客体）**

除了施事和客体
还有很多其它
类型的语义角色！

语义角色标注

- **语义角色标注** **(Semantic Role Labeling, SRL)**
  - 一种**浅层**语义分析技术
  - 确定作为谓语变元的名词性短语所扮演的语义角色

- **例子：** The student solved problems with a calculator in the classroom this morning
  - 谓语(Predicate): solved
  - 施事(Agent): the student
  - 客体(Theme): problems
  - 工具(Instrument): a calculator
  - 地点(Location): the classroom
  - 时间(Time): this morning

语义角色的类型是人工确定的，有很多不同的划分方式

语义角色标注的应用

- 语义角色标注的应用非常广泛
- 问答系统
  - 同一类问题的答案往往对应同一种语义角色
  - Who -> agent / experiencer
  - What -> force / theme / content
  - How -> instrument
  - Where -> goal / source
  - For whom -> beneficiary

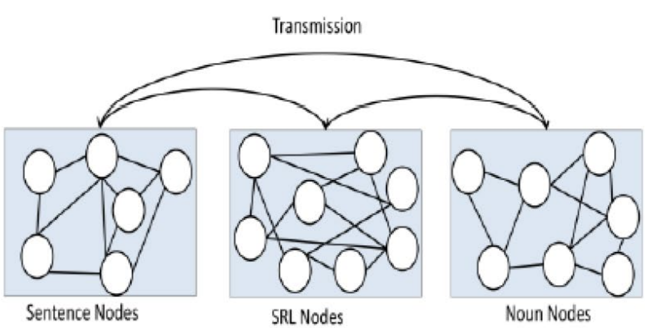
语义角色标注的应用

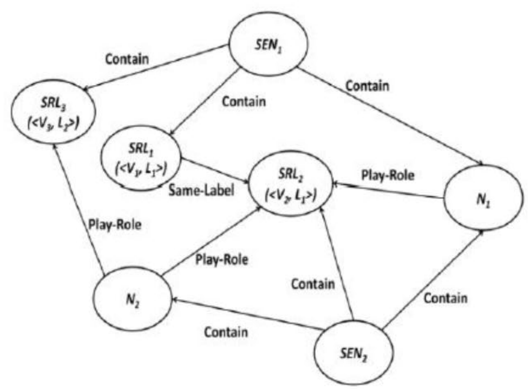
- **语义角色标注的应用非常广泛**
- **问答系统**
- **信息抽取**
  - 同一类信息往往对应同一种语义角色
  - London gold fell $4.70 to $ 308.45

Slot	Filler	Semantic Role
Product	London gold	Experiencer
Price change	-$4.70	Theme
Current price	$308.45	Goal

语义角色标注的应用

- 语义角色标注的**应用非常广泛**
- 问答系统
- 信息抽取
- **文档摘要**
  - 层级化摘要
  - 需要归纳不同文档中同一语义角色





语义角色标注的应用

- 语义角色标注的**应用非常广泛**
- 问答系统 [Hendrix et al., 1973; Shen & Lapata, 2007; Surdeanu et al., 2011]
- 信息抽取
- 文档摘要
- 知识获取
- 机器翻译 [Wilks, 1973; Liu & Gildea, 2010; Lo et al., 2013]
- 对话系统 [Bobrow et al., 1977]
- 口语理解 [Nash-Webber, 1975]
- …

语义角色

- **语义角色 (Semantic Roles)的语言学定义**
  - 一种**浅层**的语义表示
  - 语义由一句话描述的**事件(event)**表示
  - **事件**由**谓语(predicate)**表示
  - **谓语**可以携带多个**论元(arguments)**，表示与事件相关的对象
  - **语义角色**是**论元**在**事件**中充当的**抽象角色**

> 原型施事是对施事的一般化：
> 以下均是原型施事
> Tom hits the ball. (施事)
> Tom likes the ball. (Experiencer,感事)
> The sky is blue. (Theme,主事)

- **语义角色同样有多种粒度**

更具体 ⟷ 更一般

Hitter (打击者)	Agent (施事)	Proto-agent (原型施事)

题旨角色 (Thematic Role)

- **语义角色由题旨角色发展而来**
  - 最古老的语言学模型之一
    - 印度语法学家Panini [7th to 4th BCE]
  - 现代阐述
    - Fillmore的**格理论(case theory)** [1966, 1968]， Gruber [1965]
      - Fillmore受Lucien Tesnière的Éléments de Syntaxe Structurale [1959] 启发，起初称这些角色为actant [1966]后改为case
      - 中心动词与名词短语作为句法的深层结构，之间的语义关系被称为深层格
  - 示例

对依存句法在语义上的进一步细化!

Thematic Role	Definition	Example
AGENT	The volitional causer of an event	*The waiter* spilled the soup.
EXPERIENCER	The experiencer of an event	*John* has a headache.
FORCE	The non-volitional causer of the event	*The wind* blows debris from the mall into our yards.
THEME	The participant most directly affected by an event	Only after Benjamin Franklin broke *the ice*...
RESULT	The end product of an event	The city built a *regulation-size baseball diamond*...
CONTENT	The proposition or content of a propositional event	Mona asked *"You met Mary Ann at a supermarket?"*
INSTRUMENT	An instrument used in an event	He poached catfish, stunning them *with a shocking device*...
BENEFICIARY	The beneficiary of an event	Whenever Ann Callahan makes hotel reservations *for her boss*...
SOURCE	The origin of the object of a transfer event	I flew in *from Boston*.
GOAL	The destination of an object of a transfer event	I drove *to Portland*.

题旨角色的问题

- **难以建立标准的角色集合或准确定义题旨角色**
  - 粒度 与 原子性 常常冲突
  - 角色通常需要被分裂才能被准确定义

- **例如，题旨角色中的INSTRUMENTS(工具)并包含了两种类型的角色[Levin & Hovav, 2015]：**
  - 媒介工具(intermediary instruments)：**可作主语**
    - The cook opened the jar with **the new gadget**
    - **The new gadget** opened the jar
  - 赋能工具(enabling instruments): **不可做主语**
    - Shelly ate the sliced banana with **a fork**
    - ***The fork** ate the sliced banana.

语义角色的粒度

- **实际中处理的语义角色有两类**

- **更一般化的、更少角色（一般所说的语义角色）**
  - 基于原型施事、原型受事 [Dowty 1991]
  - **PropBank语料库为代表（语义角色标注所用的语料）**

- **更细粒度的、更多角色（框架语义）**
  - frames [Fillmore 1968, 1977]
  - 根据一类谓语定义特定的角色
  - **FrameNet语料库为代表**

语义角色标注的特性

语义角色与句法的关系

- 常见情况下，语义角色可以通过特定句法位置确定
  - Agent: subject
  - Patient: direct object
  - Instrument: object of with
  - Beneficiary: object of for
  - Source: object of from
- 但以上泛化规则不是绝对的，至多也只是倾向
  - The hammer hit the window （这里不是Agent，是Instrument）
  - The ball was passed to Mary from John （这里不是Agent，是Patient）
  - John went to the movie with Mary （不是instrument）
  - John bought the car for $20K. （不是受益者Beneficiary）

语义角色标注的特性

- **语义角色与选择限制(Selectional Restrictions)的关系**
  - 选择限制：比如一个动词只能跟有限的名词搭配，比如“吃手机”不太可能出现
  - 语义角色标注可以帮助解决选择限制的问题

- **例子：I want to eat *someplace nearby*.**
  - Two interpretations
    - a) sensible: eat is intransitive and *someplace nearby* is a location adjunct
    - B) speaker is Godzilla: eat is transitive and *someplace nearby* is a direct object
  - 通过语义角色标注：a > b

选择限制 与 选择倾向

- **选择限制(selectional restrictions)或选择倾向(selectional preferences)?**
- **早期，选择限制是严格约束[Katz and Fodor, 1963]**
- **很快，人们明白选择限制其实只是倾向[Wilks, 1975]**
  - 目前的语义分析还难以解决
- **例子**
  - But it fell apart in 1931, perhaps because people realized you *can’t eat gold* for lunch if you’re hungry.
  - In his two championship trials, Mr. Kulkarni *ate glass* on an empty stomach, accompanied only by water and tea.

大纲

- 语义角色标注
- **PropBank与FrameNet**
- 解决方案
  - 句法树方法
  - 序列标注方法

PropBank

- **The Proposition Bank (PropBank) [Palmer et al. 2005]**
  - 采用粗粒度的角色定义[Dowty 1991]
  - 使用原型施事(proto-agent)和原型受事(proto-patient)
- **PropBank中根据动词的词义标注以下几类论元**
  - ARG0: PROTO-AGENT
  - ARG1: PROTO-PATIENT
  - ARG2: benefactive, instrument, attribute, end state
  - ARG3: start point, benefactive, instrument or attribute
  - ARG4: end point
  - ARGM: modifiers or adjuncts of the predicate
    - TMP, LOC, DIR, MNR, ADV, …

标注示例

- 根据动词确定每个Arg的具体含义

- Predicate **accept$_1$ "take willingly"**
  - Arg0: acceptor
  - Arg1: thing accepted
  - Arg2: accepted-from
  - Arg3: attribute
- [$_{Arg0}$He] [$_{ArgM-mod}$would] [$_{ArgM-neg}$n't] **accept** [$_{Arg1}$anything of value] [$_{Arg2}$from those he was writing about].
- Predicate **kick$_1$ "drive or impel with the foot"**
  - Arg0: kicker
  - Arg1: thing kicked
  - Arg2: instrument (defaults to foot)
- [$_{Arg0}$John] tried [$_{Arg0}$*trace*] to **kick** [$_{Arg1}$the football].

PropBank

- **PropBank的标注可以很好的表示语义上的共性**
  - 0，1规律比较明显，2之后根据具体词有变化
- **示例:**

➢ Predicate $increase_1$ “go up incrementally”
  - ➢ Arg0: causer of increase
  - ➢ Arg1: thing increasing
  - ➢ Arg2: amount increased by, EXT or MNR
  - ➢ Arg3: start point （升高的起点)
  - ➢ Arg4: end point （升高的终点)

➢ [$_{Arg0}$ Big Fruit Co.] increased [$_{Arg1}$ the price of bananas].

➢ [$_{Arg1}$ The price of bananas] was increased again [$_{Arg0}$ by Big Fruit Co.]

PropBank

- **PropBank 中也包含一些名词和轻动词(light verb)**
  - 如**decision**和**make** a decision中的make

    Example Noun: *Decision*
    - Roleset: Arg0: decider, Arg1: decision…
    - "…[your$_{ARG0}$] [decision$_{REL}$] [to say look I don't want to go through this anymore$_{ARG1}$]"

    Example within an LVC: *Make a decision*
    - "…[the President$_{ARG0}$] [made$_{REL\text{-}LVB}$] the [fundamentally correct$_{ARGM\text{-}ADJ}$] [decision$_{REL}$] [to get on offense$_{ARG1}$]"

    对比make a decision
    和make a toy:
    是否是实际的制作?

- **NomBank**
  - PropBank以动词为主
  - 在PropBank的基础上进一步扩充了名词和形容词

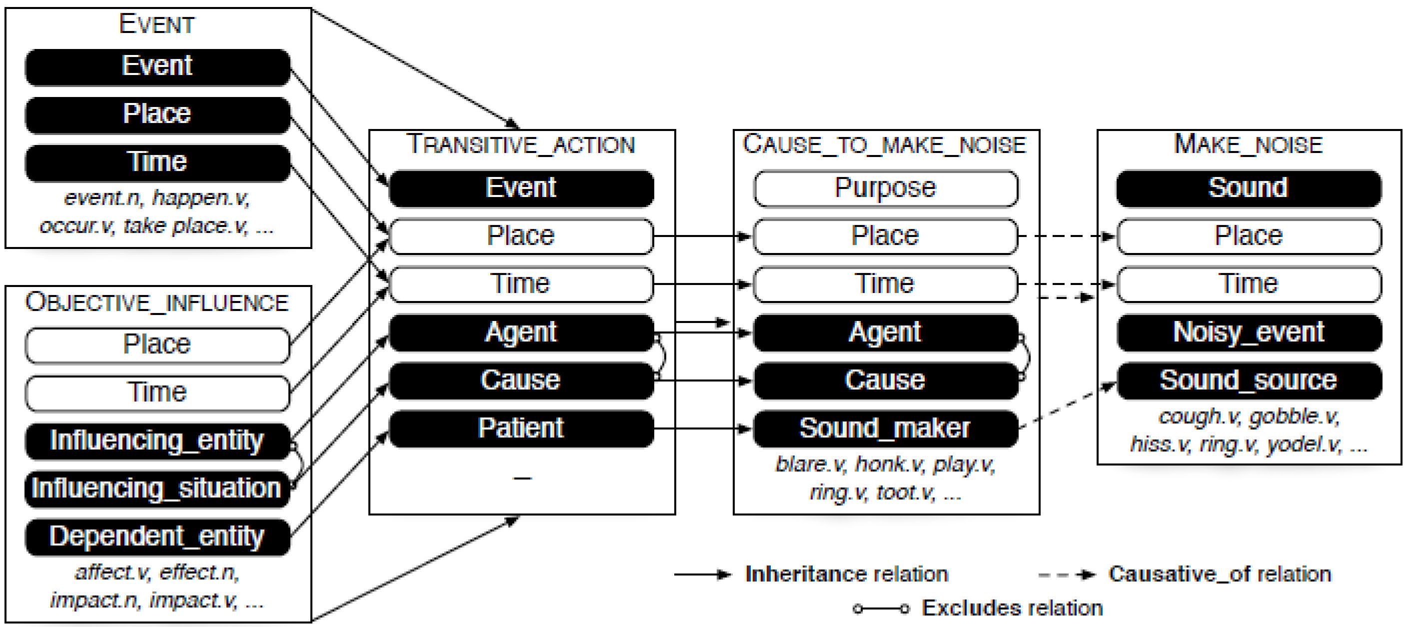
FrameNet

- **FrameNet**
  - Baker et al. 1998, Fillmore et al. 2003, Fillmore and Baker 2009, Ruppenhofer et al. 2006
- **PropBank中的角色根据动词定义**
- **FrameNet中的角色根据框架定义**
- **框架的定义**
  - 可以理解成，把同一类动词进行了聚类，这个类就是一个框架（比如“拿”、“取”可以属于一个框架；而且还确定了框架间的层级关系，比如“继承”、“原因”
  - 框架元素：A background knowledge structure that defines a set of frame-specific semantic roles, called **frame elements （就是后一页的加黑部分，黑框部分是必须的元素，白框部分是可选元素）**
  - 谓语（一般是动词，但也可以是名词）：Includes **a set of predicates** that use these roles（就是后一页的最底下的那些词）
  - 实际分析过程中，每个词都要找到其对应的框架，然后获取部分框架元素

FrameNet

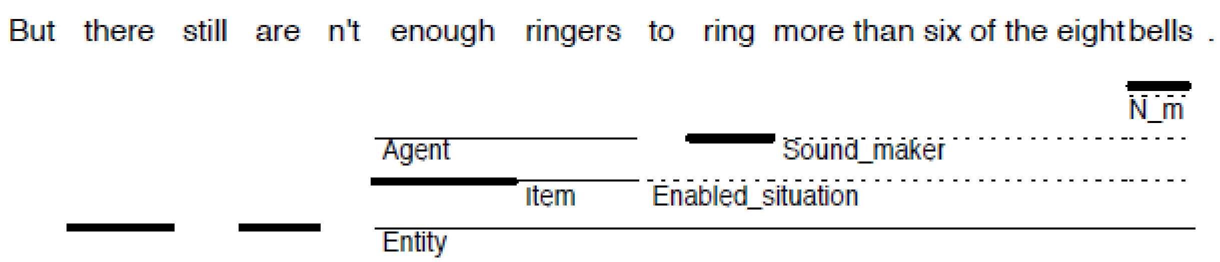
为何是Frame*Net*

- 框架通过关系相连构成网络（框架上的箭头）
- 框架元素之间同样由关系相连构成网络（加黑部分的箭头）
- 箭头来自父类，指向子类；比如“继承”是出现最多的一个类型，代表“语义的细化”



FrameNet

- **与PropBank相比，FrameNet的复杂度更高**
  - 下例，粗黑线代表单词触发了一个语义框架，一行是一个语义框架
  - 比如对于ring，左边的是agent，右边的是sound maker



NOISE_MAKERS	*bell.n*
CAUSE_TO_MAKE_NOISE	*ring.v*
SUFFICIENCY	*enough.a*
EXISTENCE	*there be.v*

Frame示例

- 框架里面除了结构化的元素和谓词，还有非结构化的自然语言解释，以下是非结构化的解释举例

- **apply heat**: situation involving a cook, food and a heating instrument
  evoked by *bake*, *blanch*, *boil*, *broil*, *brown*, *simmer*, etc.
- **change position on a scale**: situation involving the change of an items's position on a scale (the attribute) from a starting point (initial value) to an end point (final value)
  evoked by *decline*, *decrease*, *gain*, *rise*, etc.
- **damaging**: situation involving an agent that affects a patient in such a way that the patient (or some sub-region of the patient) ends up in a non-canonical state
  evoked by *damage*, *sabotage*, *scratch*, *tear*, *vandalise*, etc.

FrameNet

标注示例

- frameNet除了标注了之前说的结构化知识库，还标注了非结构化的训练语料（就像propBank的训练语料一样），以下为样例
- 但是语料还是偏少，几万句，这是frameNet准确度还是偏低的原因之一

- Verbs:
  - [$_{Cook}$ Matilde] **fried** [$_{Food}$ the catfish] [$_{HeatingInstrument}$ in an iron skillet]
  - [$_{Item}$ Colgate's stocks] **rose** [$_{Difference}$ \$3.64] to [$_{FinalValue}$ \$49.94]
- Nouns:
  - ...the **reduction** of [$_{Item}$ debt levels] to [$_{Value2}$ \$25] from [$_{Value1}$ \$2066]
- Adjectives:
  - [$_{Sleeper}$ They] were **asleep** [$_{Duration}$ for hours]

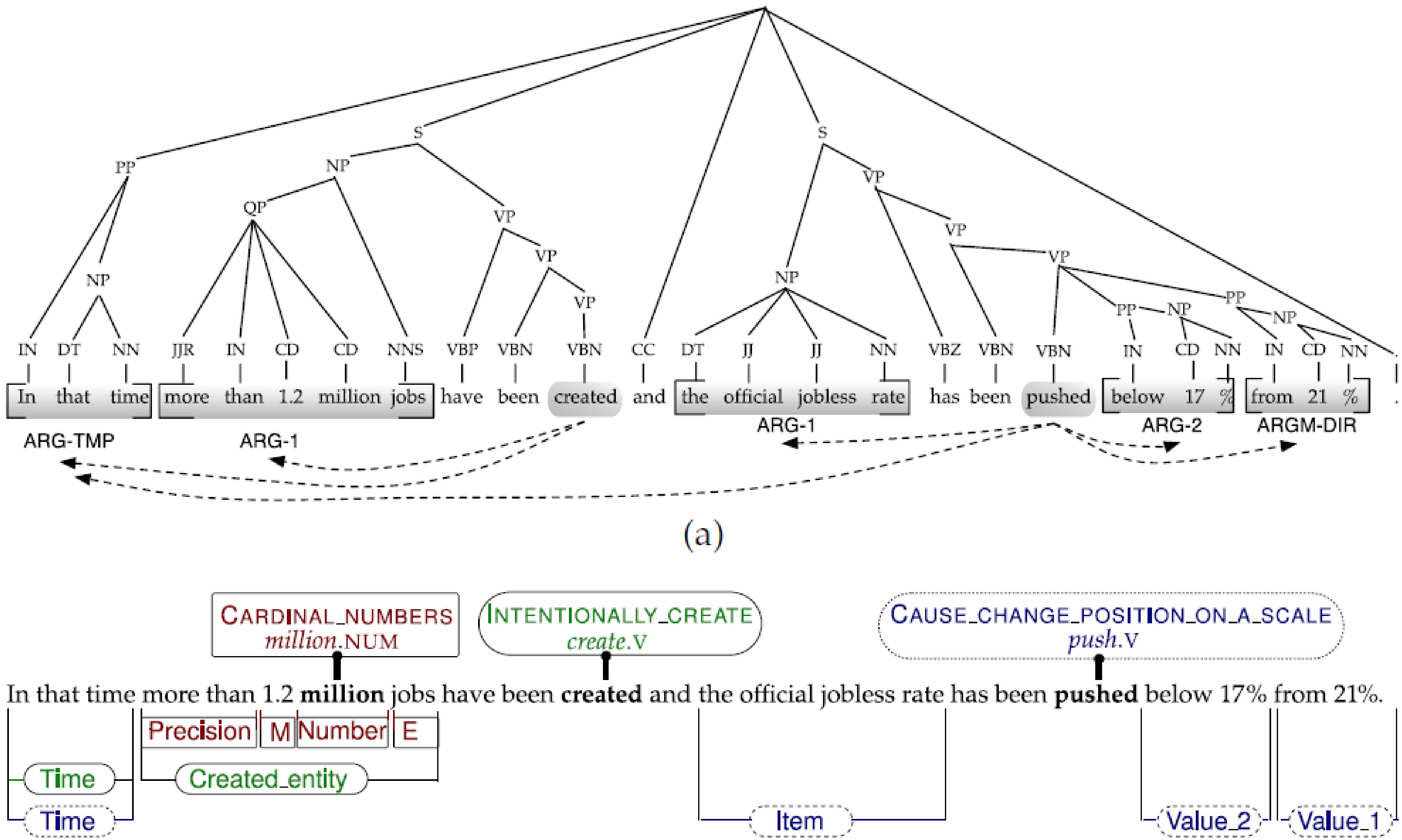
FrameNet

- **FrameNet可以更好的表示同一类事件之间的共性**
  - PropBank针对同一动词之间的共性
  - 比如以下几个句子，用了不同的动词，但是item和agent都能成功分析出

- [$_{Agent}$ Big Fruit Co.] increased [$_{Item}$ the price of bananas].
- [$_{Item}$ The price of bananas] rose [$_{Agent}$ by Big Fruit Co.]
- There has been a [$_{Difference}$ 5%] rise in [$_{Item}$ the price of bananas].

FrameNet与PropBank

- FrameNet vs. PropBank （上图是propBank，它是由句法树细化标注得到的）



大纲

- **语义角色标注**
- **PropBank与FrameNet**
- **解决方案**
  - 句法树方法
  - 序列标注方法

语义角色标注方法

- **目标：寻找句子中每个谓语的每个论元的语义角色（因为是以动词为中心）**
  - 识别谓语
  - 识别论元
  - 标定论元角色
- **对象：FrameNet vs. PropBank （上面是frameNet，下面是propBank）**

[You]	can't	[blame]	[the program]	[for being unable to identify it]
COGNIZER		TARGET	EVALUEE	REASON

[The San Francisco Examiner]	issued	[a special edition]	[yesterday]
ARG0	TARGET	ARG1	ARGM-TMP

- **两大类方法**
  - 序列标注方法
  - 句法树方法

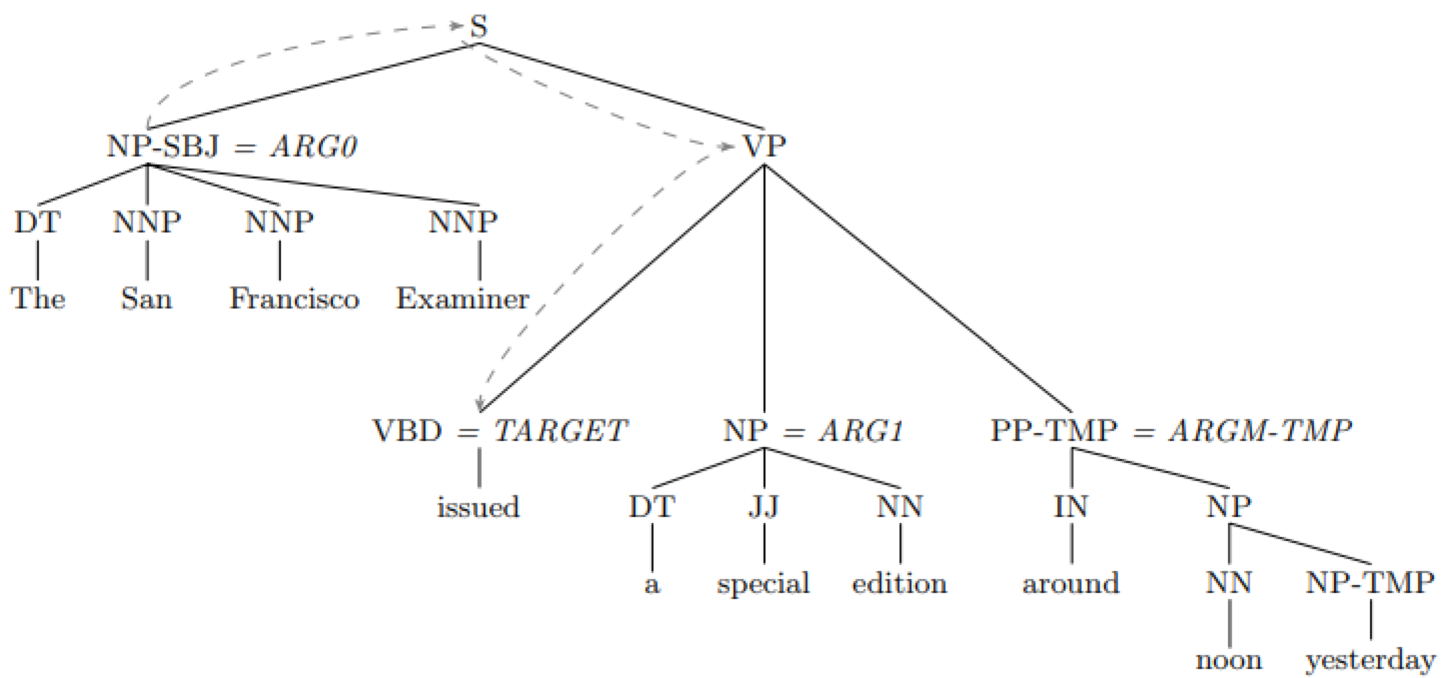
序列标注方法

- **语义角色标注视为Segmenting类的序列标注任务**
- **标签含有两个属性**
  - 边界属性：BIO，BIO2，BIOSE
  - 角色属性：Arg0, Arg1, …
- **可以使用任意序列标注模型**
- **有效的特征包括：中心词、窗口词、词性等**

- **在没有神经网络的时代，效果极差**
- **在深度学习时代，主要用LSTM进行序列标注，效果跟句法树方法相当，大概是80-85%左右**

句法树方法

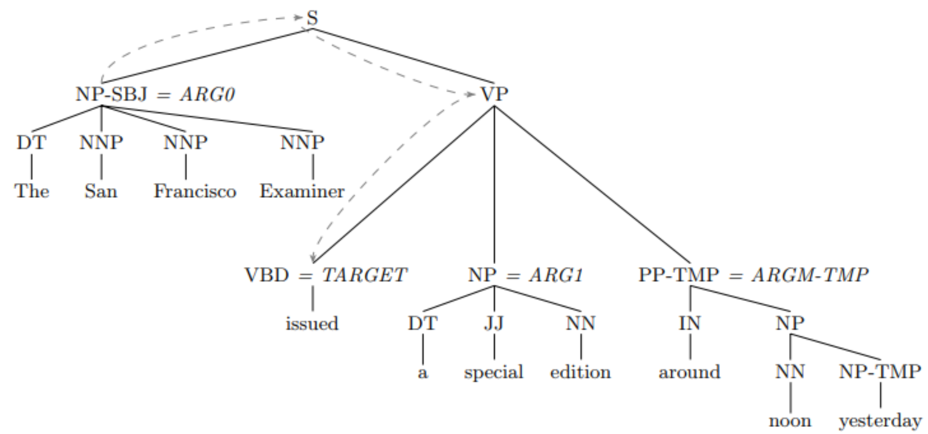
- **借助句法树完成分类任务**
  - 句法树提供了大量的语义线索
  - 下例是CFG句法分析，在句法树结构上识别arg0，arg1等



句法树方法

一个简单的算法框架

- 遍历一棵树，在每个节点上提取特征，做分类



function SEMANTICROLELABEL(*words*) **returns** labeled tree

parse ← PARSE(*words*)
for each *predicate* **in** *parse* **do**
  for each *node* **in** *parse* **do**
    featurevector ← EXTRACTFEATURES(*node*, *predicate*, *parse*)
    CLASSIFYNODE(*node*, *featurevector*, *parse*)

句法树方法

- **第一步：What is a predicate?**
- **PropBank verbs**
  - 选定所有动词
  - 可以排除light verbs（表）
- **FrameNet verbs/nouns/adjectives**
  - 选定训练数据中所有标为中心词的词

```
function SEMANTICROLELABEL(words) returns labeled tree

  parse ← PARSE(words)
  for each predicate in parse do
    for each node in parse do
      featurevector ← EXTRACTFEATURES(node, predicate, parse)
      CLASSIFYNODE(node, featurevector, parse)
```

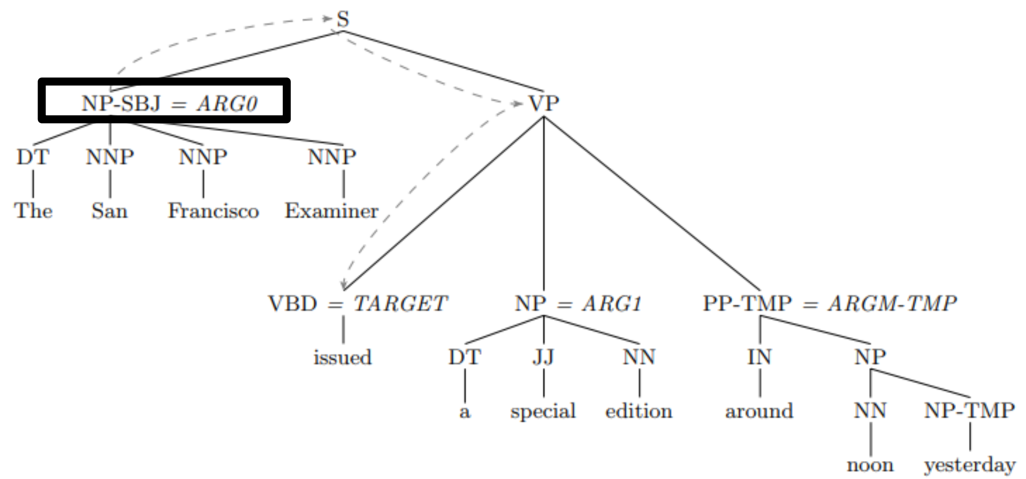
句法树方法

- **基本型Features**

```
function SEMANTICROLELABEL(words) returns labeled tree

  parse ← PARSE(words)
  for each predicate in parse do
    for each node in parse do
      featurevector ← EXTRACTFEATURES(node, predicate, parse)
      CLASSIFYNODE(node, featurevector, parse)
```

- **Headword**
  - (通过规则确定，如Examiner



- **Headword POS**
- **单词的主动、被动形态**
- **Subcategorization of predicate**
- **Named Entity type of constituent**
- **First and last words of constituent**
- **Linear position, clause w.r.t. predicate**

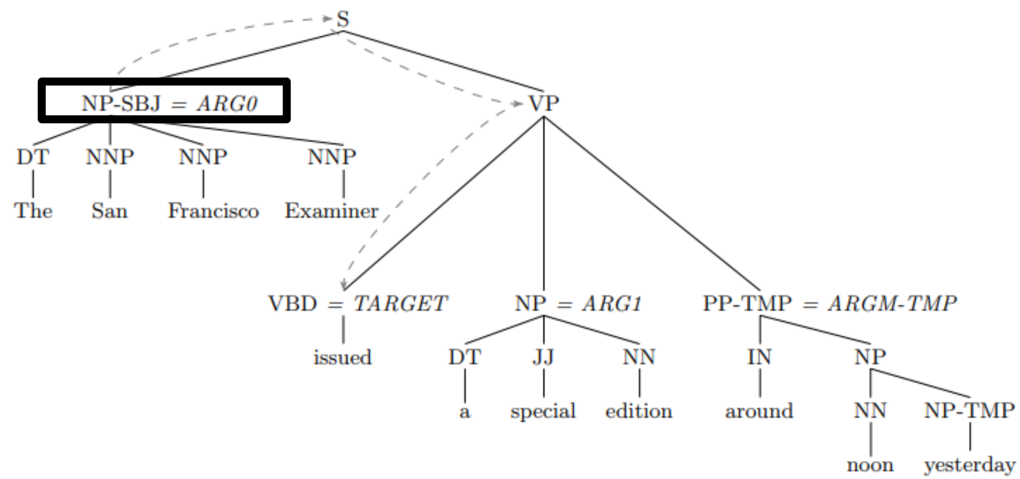
句法树方法

- 特殊型Features

```
function SEMANTICROLELABEL(words) returns labeled tree

  parse ← PARSE(words)
  for each predicate in parse do
    for each node in parse do
      featurevector ← EXTRACTFEATURES(node, predicate, parse)
      CLASSIFYNODE(node, featurevector, parse)
```

- Path
  - 从当前节点到谓语词在句法树上的路径



NP↑S↓VP↓VBD

Frequency	Path	Description
14.2%	VB↑VP↓PP	PP argument/adjunct
11.8	VB↑VP↑S↓NP	subject
10.1	VB↑VP↓NP	object
7.9	VB↑VP↑VP↑S↓NP	subject (embedded VP)
4.1	VB↑VP↓ADVP	adverbial adjunct
3.0	NN↑NP↑NP↓PP	prepositional complement of noun
1.7	VB↑VP↓PRT	adverbial particle
1.6	VB↑VP↑VP↑VP↑S↓NP	subject (embedded VP)
14.2		no matching parse constituent
31.4	Other	

句法树方法

- 分类的实现：3-step version

```
function SEMANTICROLELABEL(words) returns labeled tree

  parse ← PARSE(words)
  for each predicate in parse do
    for each node in parse do
      featurevector ← EXTRACTFEATURES(node, predicate, parse)
      CLASSIFYNODE(node, featurevector, parse)
```

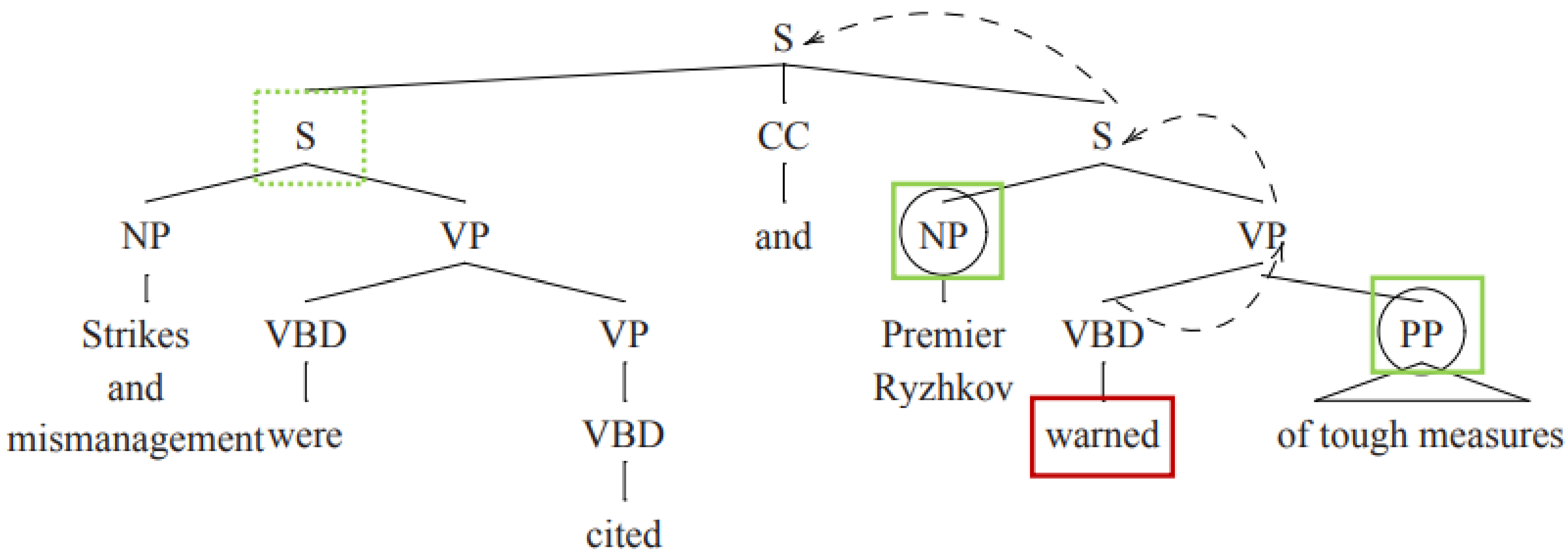
- 1，过滤：Pruning
  - Simple heuristics to prune unlikely constituents
- 2，识别是否跟谓词有关系：Identification
  - 是否问题：Binary classification of each node as an argument to be labeled or a NONE
- 3，具体是属于哪种关系：Classification
  - 多分类问题：1-of-N classification of all the constituents that were labeled as arguments

句法树方法

- **过滤的重要性：Why Pruning?**
- **大量的词都跟谓词无关：One predicate at a time, Imbalance data**
  - Very few of the nodes in the tree could possible be arguments of that one predicate
  - Positive samples vs negative samples
- **Prune the very unlikely first, and then use a classifier to get rid of the rest**

句法树方法

- 过滤的重要性： Pruning heuristics [Xue and Palmer, 2004]
- 比如下例，and代表了并列关系，如果找warned的论元，则先找兄弟节点，再找叔父节点，再找祖父节点，然后把左边的分支全部裁掉



句法树方法

- **怎么分类：先局部分类，然后re-ranking**
- **局部分类：The algorithm classifies everything locally**
- **But lots of global or joint interactions**
  - Non-overlapping
  - No Multiple identical arguments
- **重排序：通过Reranking捕捉全局的信息**
  - Possible labels -> classifier -> best global label
  - Takes all the input along with other features

句法树方法

- **FrameNet更复杂一些：还需要判断是那个框架，因为不是arg0，arg1的分类问题了，还需要判定是具体的那个框架**
  - We need an extra step to find the frame
  - Features for frame identification [Das et al, 2014]

the POS of the parent of the head word of t_i
the set of syntactic dependencies of the head word[21] of t_i
if the head word of t_i is a verb, then the set of dependency labels of its children
the dependency label on the edge connecting the head of t_i and its parent
the sequence of words in the prototype, w_ℓ
the lemmatized sequence of words in the prototype
the lemmatized sequence of words in the prototype and their part-of-speech tags π_ℓ
WordNet relation[22] ρ holds between ℓ and t_i
WordNet relation[22] ρ holds between ℓ and t_i, and the prototype is ℓ
WordNet relation[22] ρ holds between ℓ and t_i, the POS tag sequence of ℓ is π_ℓ, and the POS tag sequence of t_i is π_t

总结：语义角色标注

- **任务：who does what to whom when where how**
- **对象：thematic roles -> Frame or Proto-A/P（propBank)**
- **资源：PropBank, FrameNet, CoNLL shared tasks**
- **特性：句法线索syntactic, 选择限制selection**
- **方法**
  - Sequence labelling: very bad before DL
  - Syntactic: very good before DL
  - DL: Bi-LSTM作序列标注反而效果好

深度学习方法

- **扩展阅读：**
- **End-to-end Learning of Semantic Role Labelling Using Recurrent Neural Networks (E2E)**
  - ACL 2015
  - Jie Zhou and Wei Xu, Baidu Research
- **Deep Semantic Role Labelling: What Works and What’s Next (Deep)**
  - ACL 2017
  - Luheng He, Kenton Lee, Univ. of Washington
  - Mike Lewis, FAIR
  - Luke Zettlemoyer, Allen Institute for AI

THANKS!