# Reminders

HOMEWORK 8 DUE DATE TONIGHT BY 11:59PM

QUIZ ON CHAPTERS 15&16 WILL BE RELEASED LATER THIS WEEK.

I DISTRIBUTED DATA FOR HW9 FOR EVERYONE WHO COMPLETED IRB TRAINING

# Information Extraction

JURAFSKY AND MARTIN CHAPTER 18

# Information Extraction (IE)

- Information extraction (IE), turns the unstructured text information into structured data
- Populate a relational database to enable further processing, support queries

# Template Filling

Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower cost carriers. American Airlines, a unit of AMR Corp., immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL Corp., said the increase took effect Thursday and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Denver to San Francisco.

| | | |
|---|---|---|
| FARE-RAISE ATTEMPT: | LEAD AIRLINE: | UNITED AIRLINES |
| | AMOUNT: | $6 |
| | EFFECTIVE DATE: | 2006-10-26 |
| | FOLLOWER: | AMERICAN AIRLINES |

# Steps in IE

- NER and co-reference resolution
- **relation extraction**
  - spouse-of, child-of, employer-of, partof, membership-in, located-in
- event extraction
- temporal expression normalization
- template filling

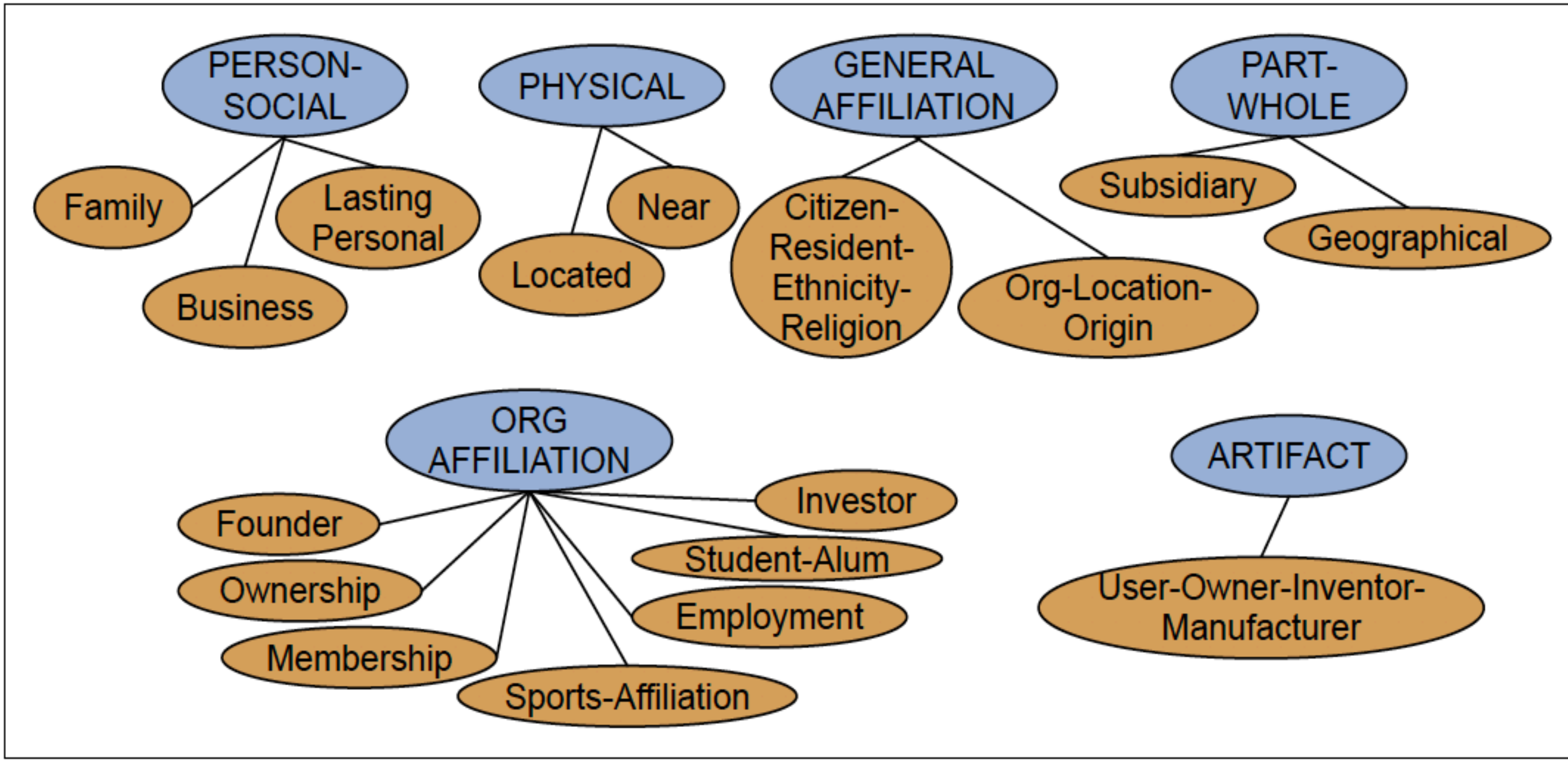
# Relation Extraction

- A relation consists of a set of ordered tuples over elements of a domain
- The domain elements corresponds to the named entities

| | |
|---|---|
| **Domain** | $\mathscr{D} = \{a, b, c, d, e, f, g, h, i\}$ |
| United, UAL, American Airlines, AMR | $a, b, c, d$ |
| Tim Wagner | $e$ |
| Chicago, Dallas, Denver, and San Francisco | $f, g, h, i$ |
| **Classes** | |
| United, UAL, American, and AMR are organizations | $Org = \{a, b, c, d\}$ |
| Tim Wagner is a person | $Pers = \{e\}$ |
| Chicago, Dallas, Denver, and San Francisco are places | $Loc = \{f, g, h, i\}$ |
| **Relations** | |
| United is a unit of UAL | $PartOf = \{\langle a, b\rangle, \langle c, d\rangle\}$ |
| American is a unit of AMR | |
| Tim Wagner works for American Airlines | $OrgAff = \{\langle c, e\rangle\}$ |
| United serves Chicago, Dallas, Denver, and San Francisco | $Serves = \{\langle a, f\rangle, \langle a, g\rangle, \langle a, h\rangle, \langle a, i\rangle\}$ |

Citing high fuel prices, [$_{\text{ORG}}$ **United Airlines**] said [$_{\text{TIME}}$ **Friday**] it has increased fares by [$_{\text{MONEY}}$ **$6**] per round trip on flights to some cities also served by lower-cost carriers. [$_{\text{ORG}}$ **American Airlines**], a unit of [$_{\text{ORG}}$ **AMR Corp.**], immediately matched the move, spokesman [$_{\text{PER}}$ **Tim Wagner**] said. [$_{\text{ORG}}$ **United**], a unit of [$_{\text{ORG}}$ **UAL Corp.**], said the increase took effect [$_{\text{TIME}}$ **Thursday**] and applies to most routes where it competes against discount carriers, such as [$_{\text{LOC}}$ **Chicago**] to [$_{\text{LOC}}$ **Dallas**] and [$_{\text{LOC}}$ **Denver**] to [$_{\text{LOC}}$ **San Francisco**].

Tim Wagner **[is a spokesman]** for American Airlines.
United **[is a unit of]** UAL Corp.
American **[is a unit of]** AMR.




| Relations | Types | Examples |
|---|---|---|
| Physical-Located | PER-GPE | **He** was in **Tennessee** |
| Part-Whole-Subsidiary | ORG-ORG | **XYZ**, the parent company of **ABC** |
| Person-Social-Family | PER-PER | **Yoko**'s husband **John** |
| Org-AFF-Founder | PER-ORG | **Steve Jobs**, co-founder of **Apple**... |

# Unified Medical Language System

| Entity | Relation | Entity |
|---|---|---|
| Injury | disrupts | Physiological Function |
| Bodily Location | location-of | Biologic Function |
| Anatomical Structure | part-of | Organism |
| Pharmacologic Substance | causes | Pathological Function |
| Pharmacologic Substance | treats | Pathologic Function |

Given a medical sentence like:

Doppler echocardiography can be used to diagnose left anterior descending artery stenosis in patients with type 2 diabetes

Extract the UMLS relation:

Echocardiography (Doppler) **Diagnoses** Acquired stenosis

# Wikipedia info boxes

**Barack Obama**

**44th President of the United States**

**In office**
January 20, 2009 – January 20, 2017

| | |
|---|---|
| **Vice President** | Joe Biden |
| **Preceded by** | George W. Bush |
| **Succeeded by** | Donald Trump |

**United States Senator from Illinois**

**In office**
January 3, 2005 – November 16, 2008

| | |
|---|---|
| **Preceded by** | Peter Fitzgerald |
| **Succeeded by** | Roland Burris |

**Member of the Illinois Senate from the 13th district**

**In office**
January 8, 1997 – November 4, 2004

| | |
|---|---|
| **Preceded by** | Alice Palmer |
| **Succeeded by** | Kwame Raoul |

**Personal details**

| | |
|---|---|
| **Born** | Barack Hussein Obama II<br>August 4, 1961 (age 58)<br>Honolulu, Hawaii, U.S. |
| **Political party** | Democratic |

# RDF Triples

- Resource Description Framework
- **RDF triple** is a tuple of
  - Entity-relation-entity, aka
  - Subject-predicate-object expression
- **Subject**: University of Pennsylvania
- **Predicate**: location
- **Object**: Philadelphia, PA

## Zachary G. Ives

FOLLOW

Professor of Computer and Information Science, University of Pennsylvania
Verified email at cis.upenn.edu - Homepage

Databases   data integration   distributed systems   web data management

| TITLE | CITED BY | YEAR |
| --- | --- | --- |
| Dbpedia: A nucleus for a web of open data<br>S Auer, C Bizer, G Kobilarov, J Lehmann, R Cyganiak, Z Ives<br>The semantic web, 722-735 | 4119 | 2007 |

DBpedia is a community effort to extract structured information from Wikipedia and to make this information available on the Web. DBpedia allows you to ask sophisticated queries against datasets derived from Wikipedia and to link other datasets on the Web to Wikipedia data. We describe the extraction of the DBpedia datasets, and how the resulting information is published on the Web for human-and machine-consumption. We describe some emerging applications from the DBpedia community and show how website authors can facilitate DBpedia content within their sites. Finally, we present the current status of interlinking DBpedia with other open datasets on the Web and outline how DBpedia could serve as a nucleus for an emerging Web of open data.

# Freebase, WordNet, other ontologies

- Freebase relations:
  - people/person/nationality
  - location/location/contains
  - people/person/place-of-birth
  - biology/organism classification
- WordNet relations:
  - is-a, instance-of
  - hypernyms/hyponyms
  - Giraffe is-a ruminant is-a ungulate is-a mammal is-a vertebrate is-a animal…

# Strategies for relation extraction

- Hand-written patterns
- Supervised machine learning
- Semi-supervised machine learning
  - Bootstrapping
  - Distant supervision
- Unsupervised machine learning

# Hearst Patterns

Agar is a substance prepared from a mixture of red algae, such as Gelidium, for laboratory or industrial use.

She suggests that the following lexico-syntactic pattern

$$NP_0 \ \textit{such as}\ NP_1 \{, NP_2 \ldots, (\textit{and} | \textit{or})\ NP_i\},\ i \geq 1$$

implies the following semantics

$$\forall NP_i,\ i \geq 1,\ \textit{hyponym}(Np_i, NP_0$$

allowing us to infer

*hyponym(*Gelidium, red algae*)*

| | |
|---|---|
| NP {, NP}* {,} (and\|or) other $NP_H$ | temples, treasuries, and other important civic buildings |
| $NP_H$ such as {NP,}* {(or\|and)} NP | red algae such as Gelidium |
| such $NP_H$ as {NP,}* {(or\|and)} NP | such authors as Herrick, Goldsmith, and Shakespeare |
| $NP_H$ {,} including {NP,}* {(or\|and)} NP | common-law countries, including Canada and England |
| $NP_H$ {,} especially {NP}* {(or\|and)} NP | European countries, especially France, England, and Spain |

# Machine learning techniques

- **Supervised**: training corpus annotated with manually annotated with fixed set of relations and entities
- **Semi-supervised**: high-precision seed patters, or seed tuples, are used to bootstrap more examples
- **Distant supervision**: start with a huge number of seeds, learn noisy pattern fields (e.g. 100k examples of `birth-place-of` from infoboxes, help learn the corresponding text patterns)

# Semisupervised Relation Extraction

**function** BOOTSTRAP(*Relation R*) **returns** *new relation tuples*

*tuples* ← Gather a set of seed tuples that have relation *R*
**iterate**
*sentences* ← find sentences that contain entities in *tuples*
*patterns* ← generalize the context between and around entities in *sentences*
*newpairs* ← use *patterns* to grep for more tuples
*newpairs* ← *newpairs* with high confidence
*tuples* ← *tuples* + *newpairs*
**return** tuples

Bootstrapping proceeds by taking the entities in the seed pair, and then finding sentences (on the web, or whatever dataset we are using) that contain both entities.

# Example

Task: Create airline/hub pairs

Seed: **Ryanair** has a hub at **Charleroi**

use this seed fact to discover new patterns by finding other mentions of this relation in our corpus

Sentences found:

Budget airline **Ryanair**, which uses **Charleroi** as a hub, scrapped all weekend flights out of the airport.

All flights in and out of **Ryanair**'s Belgian hub at **Charleroi** airport were grounded on Friday...

A spokesman at **Charleroi**, a main hub for **Ryanair**, estimated that 8000 passengers had already been affected.

Patterns extracted:

/ [ORG], which uses [LOC] as a hub /

/ [ORG]'s hub at [LOC] /

/ [LOC] a main hub for [ORG] /

# Distant Supervision

**function** DISTANT SUPERVISION(*Database D, Text T*) **returns** *relation classifier C*

**foreach** relation *R*
  **foreach** tuple (*e1*,*e2*) of entities with relation *R* in *D*
    *sentences* ← Sentences in *T* that contain *e1* and *e2*
    *f* ← Frequent features in *sentences*
    *observations* ← observations + new training tuple (*e1*, *e2*, *f*, *R*)
  *C* ← Train supervised classifier on *observations*
  **return** *C*

**R:** place of birth
**(e1, e2):** <Edwin Hubble, Marshfield>, <Albert Einstein,Ulm>, etc
**Sentences:** Hubble was born in Marshfield; Einstein, born (1879), Ulm; Hubble's birthplace in Marshfield..., etc

# Open IE

Unsupervised relation extraction

Find all strings of words that satisfy the tripe relation.

United has a hub in Chicago, which is the headquarters of United Continental Holdings.

r1: <United, has a hub in, Chicago>

r2: <Chicago, is the headquarters of, United Continental Holdings>

# Evaluation of Relation Extraction

- Supervised
  - Test sets with human annotated, gold-standard relations and computing precision, recall, and F-measure
- Semi and Unsupervised
  - Human evaluations
  - Compute precision at different levels of recall

# Temporal Expression Extraction

| Absolute | Relative | Durations |
|---|---|---|
| April 24, 1916 | yesterday | four hours |
| The summer of ’77 | next semester | three weeks |
| 10:15 AM | two weeks from yesterday | six days |
| The 3rd quarter of 2006 | last quarter | the last three quarters |

Lexical triggers for temporal expressions:

| Category | Examples |
|---|---|
| Noun | *morning, noon, night, winter, dusk, dawn* |
| Proper Noun | *January, Monday, Ides, Easter, Rosh Hashana, Ramadan, Tet* |
| Adjective | *recent, past, annual, former* |
| Adverb | *hourly, daily, monthly, yearly* |

- Temporal expression recognition
- Temporal normalization
  - mapping a temporal expression to either normalization a specific point in time or to a duration

# Event Extraction

The task of event extraction is to identify mentions of **events** in texts. An event mention is **any expression denoting an event or state** that can be assigned to a **point in time** or an **interval in time**.

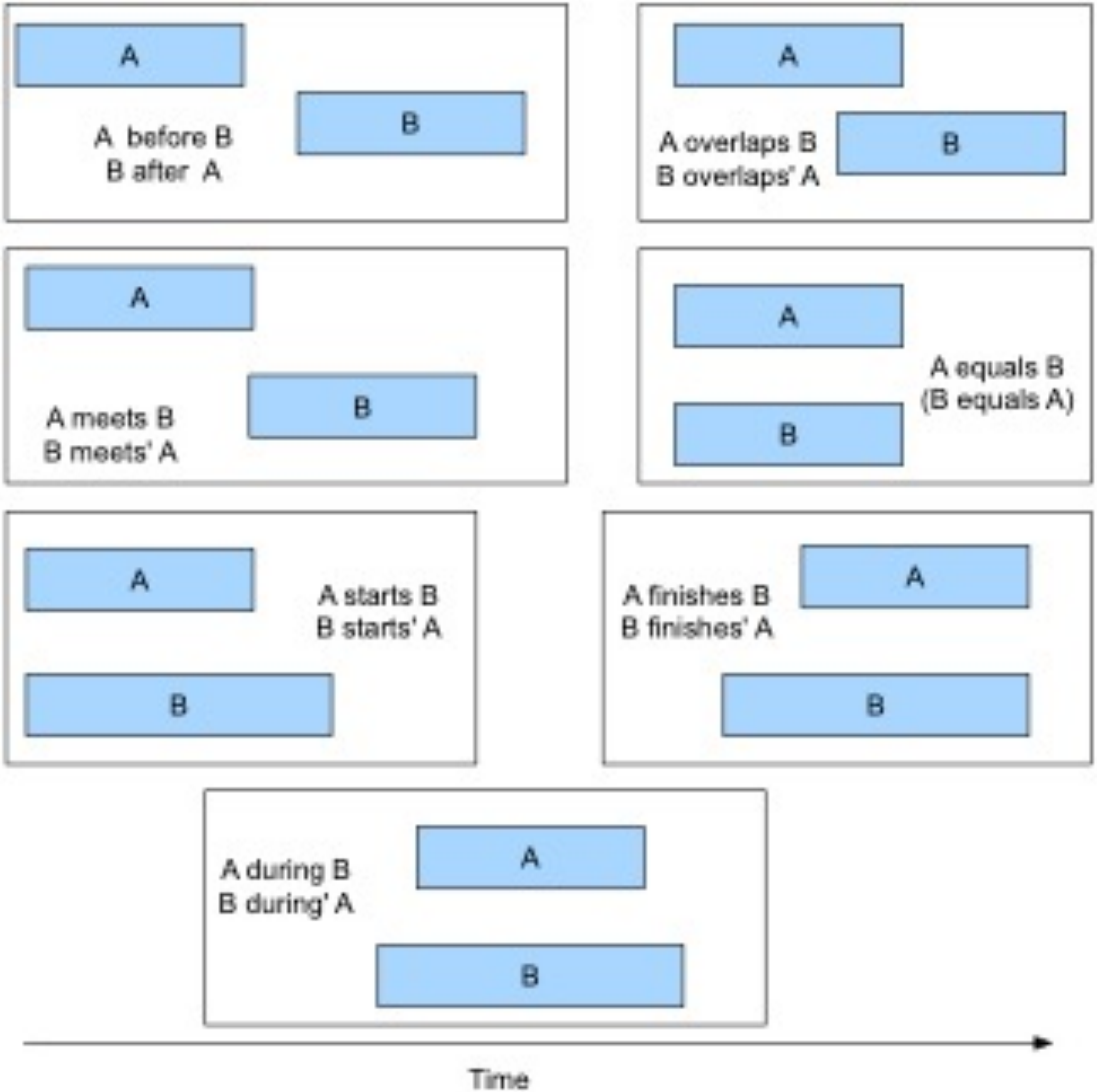
In English events are typically expressed **verbs** like exploded. However, there are also some **nouns**, like explosion, that denote an event.

Some “light verbs” like *make*, *take*, and *have* often do not denote events. Instead, the event is often expressed by the nominal direct object (*took a **flight***).

# Event Extraction

[EVENT Citing] high fuel prices, United Airlines [EVENT said] Friday it has [EVENT increased] fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit of AMR Corp., immediately [EVENT matched] [EVENT the move], spokesman Tim Wagner [EVENT said]. United, a unit of UAL Corp., [EVENT said] [EVENT the increase] took effect Thursday and [EVENT applies] to most routes where it [EVENT competes] against discount carriers, such as Chicago to Dallas and Denver to San Francisco.

Events can be classified as **actions**, **states**, **reporting events, perception events**, etc. The aspect, tense, and modality of each event also needs to be extracted.


A
B
A before B
B after A
A
B
A overlaps B
B overlaps' A
A
B
A meets B
B meets' A
A
B
A equals B
(B equals A)
A
B
A starts B
B starts' A
A
B
A finishes B
B finishes' A
A
B
A during B
B during' A
Time

# Temporal ordering of events

Delta Air Lines earnings soared 33% to a record in the fiscal first quarter, bucking the industry trend toward declining profits.

- Soaring$_{e1}$ is **included in** the fiscal first quarter$_{t58}$
- Soaring$_{e1}$ is **simultaneous with** the bucking$_{e3}$
- Declining$_{e4}$ **includes** soaring$_{e1}$

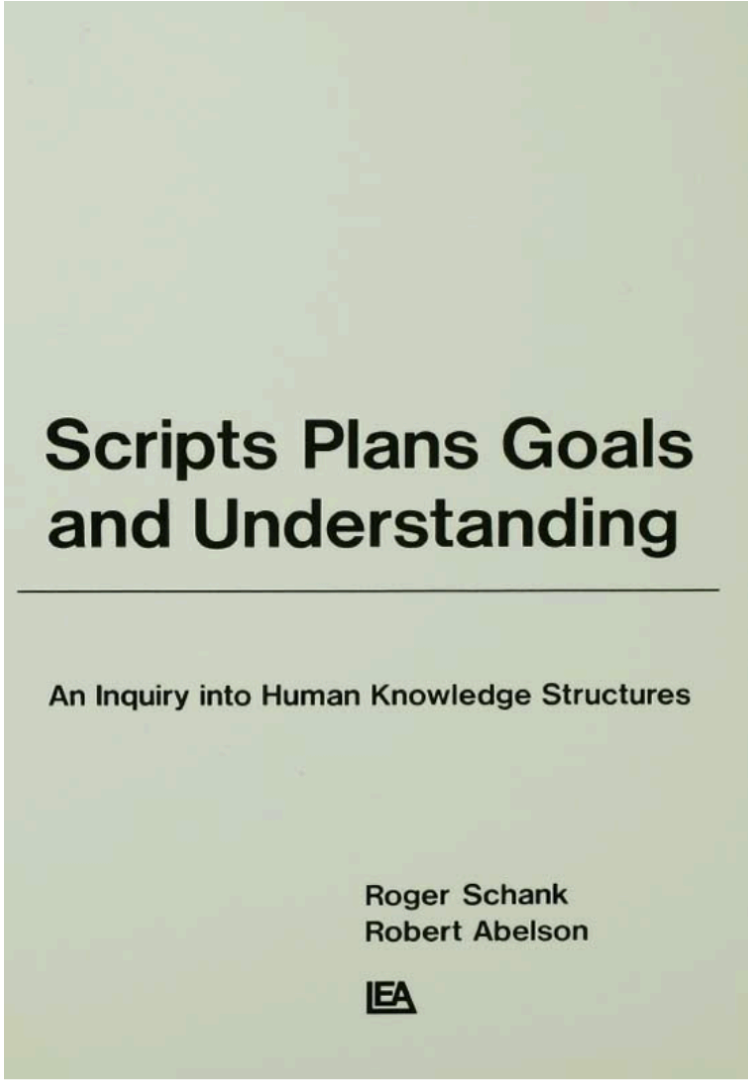
# Scripts

How do people organize all the knowledge they must have in order to understand? How do people know what behavior is appropriate for a particular situation?

**Scripts** consist of prototypical sequences of sub-events, participants, and their roles.

The strong expectations provided by these scripts can help with the classification of entities, the assignment of entities into roles and relations.

Most critically scripts can be used to draw inferences that fill in things that have been left unsaid.



Script: RESTAURANT
Track: Coffee Shop
Props: Tables
Menu
F-Food
Check
Money

Roles: S-Customer
W-Waiter
C-Cook
M-Cashier
O-Owner

Entry conditions: S is hungry.
S has money.

Results: S has less money
O has more money
S is not hungry
S is pleased (optional)

Scene 1: Entering

S **PTRANS** S into restaurant
S **ATTEND** eyes **to** tables
S **MBUILD** where to sit
S **PTRANS** S **to** table
S **MOVE** S **to** sitting position

Scene 2: Ordering

(menu on table)
S **PTRANS** menu **to** S

(W brings menu)

(S asks for menu)
S **MTRANS** signal **to** W
W **PTRANS** W **to** table
S **MTRANS** 'need menu' **to** W
W **PTRANS** W **to** menu

W **PTRANS** W **to** table
W **ATRANS** menu **to** S

S **MTRANS** food list **to CP**(S)
* S **MBUILD** choice of F
S **MTRANS** signal **to** W
W **PTRANS** W **to** table
S **MTRANS** 'I want F' **to** W

W **PTRANS** W **to** C
W **MTRANS** (**ATRANS** F) **to** C

C **MTRANS** 'no F' **to** W
W **PTRANS** W **to** S
W **MTRANS** 'no F' **to** S
(go back to *) or
(go to Scene 4 at no pay path)

C **DO** (prepare F script)
to Scene 3

# Unsupervised Learning of Narrative Event Chains

**Nathanael Chambers** and **Dan Jurafsky**
Department of Computer Science
Stanford University
Stanford, CA 94305
{natec,jurafsky}@stanford.edu

## Abstract

Hand-coded *scripts* were used in the 1970-80s as knowledge backbones that enabled inference and other NLP tasks requiring deep semantic knowledge. We propose unsupervised induction of similar schemata called *narrative event chains* from raw newswire text.

A narrative event chain is a partially ordered set of events related by a common protagonist. We describe a three step process to learning narrative event chains. The first uses unsupervised distributional methods to learn narrative relations between events sharing coreferring arguments. The second applies a temporal classifier to partially order the connected events. Finally, the third prunes and clusters self-contained chains from the space of events. We introduce two evaluations: the *narrative cloze* to evaluate event relatedness, and an *order coherence* task to evaluate narrative order. We show a 36% improvement over baseline for narrative prediction and 25% for temporal coherence.

## 1 Introduction

This paper induces a new representation of structured knowledge called **narrative event chains** (or narrative chains). Narrative chains are partially ordered sets of events centered around a common **protagonist**. They are related to structured sequences of participants and events that have been called **scripts** (Schank and Abelson, 1977) or *Fillmorean frames*. These participants and events can be filled in and instantiated in a particular text situation to draw inferences. Chains focus on a single actor to facilitate learning, and thus this paper addresses the three tasks of chain induction: *narrative event induction*, *temporal ordering of events* and *structured selection* (pruning the event space into discrete sets).

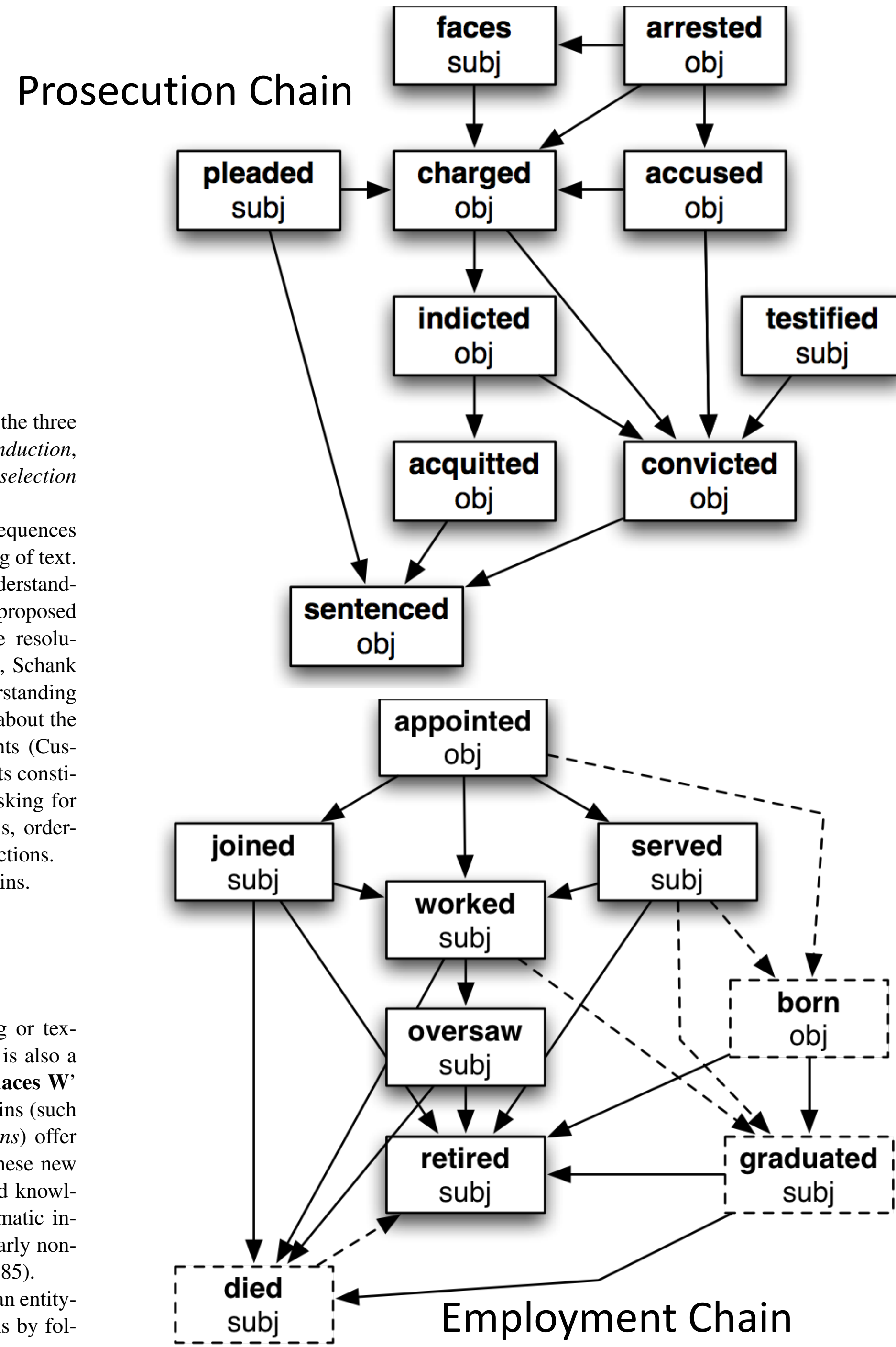
Learning these prototypical schematic sequences of events is important for rich understanding of text. Scripts were central to natural language understanding research in the 1970s and 1980s for proposed tasks such as summarization, coreference resolution and question answering. For example, Schank and Abelson (1977) proposed that understanding text about restaurants required knowledge about the Restaurant Script, including the participants (Customer, Waiter, Cook, Tables, etc.), the events constituting the script (entering, sitting down, asking for menus, etc.), and the various preconditions, ordering, and results of each of the constituent actions.

Consider these two distinct narrative chains.

| | |
|---|---|
| **_ accused X** | **W joined _** |
| **X claimed _** | **W served _** |
| **X argued** | **W oversaw _** |
| **_ dismissed X** | **W resigned** |

It would be useful for question answering or textual entailment to know that '**X denied _**' is also a likely event in the left chain, while '**_ replaces W**' temporally follows the right. Narrative chains (such as *Firing of Employee* or *Executive Resigns*) offer the structure and power to directly infer these new subevents by providing critical background knowledge. In part due to its complexity, automatic induction has not been addressed since the early non-statistical work of Mooney and DeJong (1985).

The first step to narrative induction uses an entity-based model for learning narrative relations by fol-



# Narrative Cloze Test

A **cloze test** removes words from a text and asks the participant to fill in the missing language item. Cloze tests require the ability to understand **context** and **vocabulary** in order to identify the correct language or part of speech that belongs in the deleted passages. This exercise is commonly administered for the assessment of native and second language learning and instruction.

Today, I went to the _________ and bought some milk and eggs. I knew it was going to rain, but I forgot to take my _________, and ended up getting wet on the way.

# Narrative Cloze Task

$\text{event}_1\ \text{event}_2$ __________ $\text{event}_4$ … $\text{event}_n$

What word should fill in the gap?

take → nsubj
tell → nsubj
see → nsubj
…
persuade → dobj
sail → nsubj
regard → dobj

Event vocabulary

# Template Filling

Templates represent scripts with a fixed set of slots that take slot-filler values

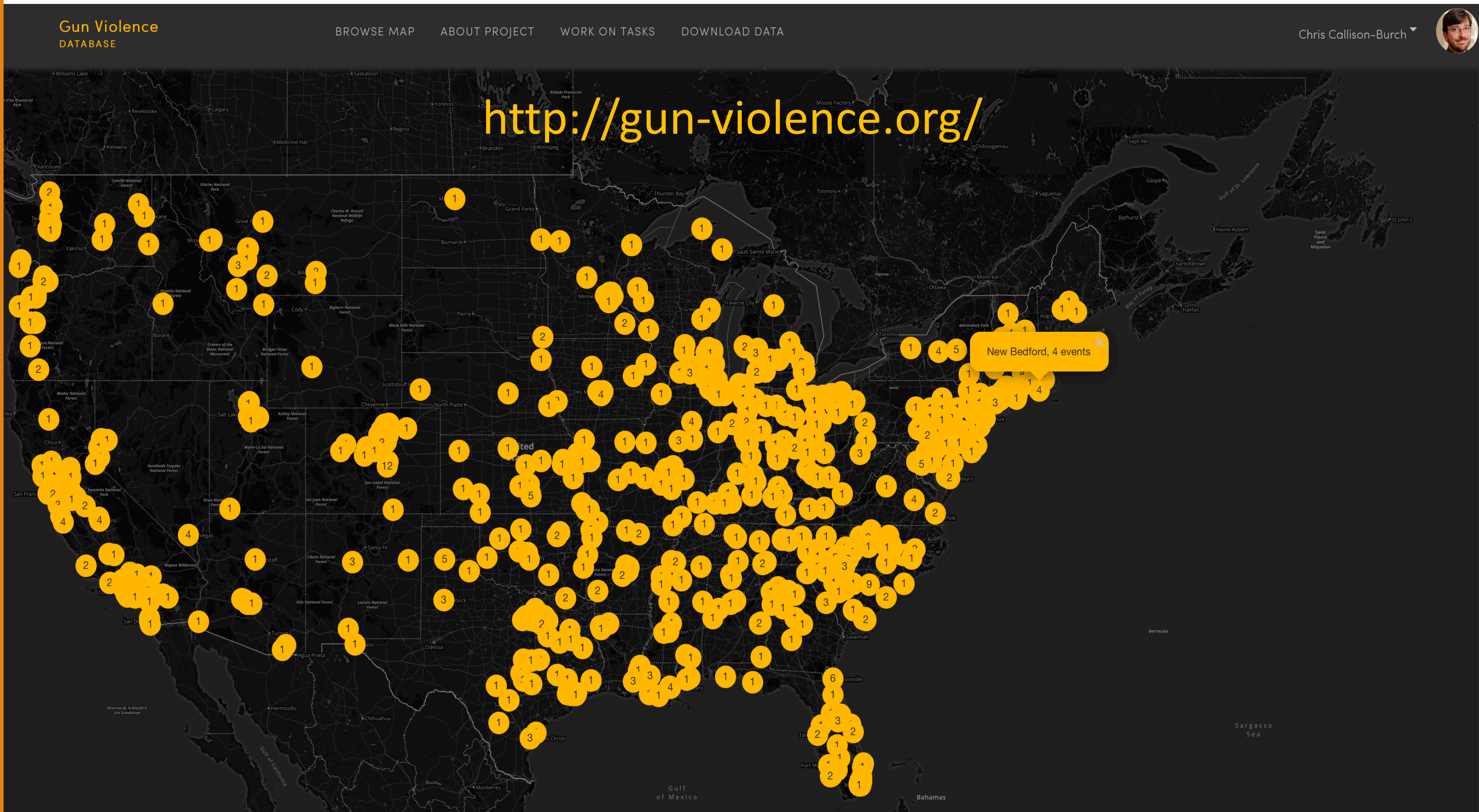
Train two separate supervised systems

1. Template recognition
2. Role-filler extraction

Most earlier systems were based on handwritten regular expressions and grammar rules.

| | | |
|---|---|---|
| FARE-RAISE ATTEMPT: | LEAD AIRLINE: | UNITED AIRLINES |
| | AMOUNT: | $6 |
| | EFFECTIVE DATE: | 2006-10-26 |
| | FOLLOWER: | AMERICAN AIRLINES |

# The Gun Violence Database




Ellie Pavlick and Chris Callison-Burch, University of Pennsylvania

# Goals of the GVDB

Collect data about gun violence in the US to facilitate **public health research**.

Draw sample from local newspapers and television stations that publish online.

Use machine learning and crowdsourcing to **extract structured data from text**.

# GVDB Template

**Chicago Police release Laquan McDonald shooting video | National News**

Three seconds. On a dashcam video clock, that's the amount of time between the moment when two officers have their guns drawn and the point when Laquan McDonald falls to the ground. The video, released to the public for the first time late Tuesday, is a key piece of evidence in a case that's sparked protests in Chicago and has landed an officer behind bars. The 17-year-old McDonald was shot 16 times on that day the video shows in October 2014. Chicago police Officer Jason Van Dyke was charged Tuesday with first-degree murder....

Incident #1053

| | |
|---|---|
| City | |
| Date | |
| Shooter | |
| Victim | |
| Victim Killed | |

Person #1014

| | |
|---|---|
| Name | |
| Gender | |
| Age | |
| Race | |

# Semantic Role Labeling

JURAFSKY AND MARTIN CHAPTER 20

# Events and their Participants

A **purchasing** event and its participants can be described by a wide variety of surface forms.

1. XYZ corporation bought the stock.
2. They sold the stock to XYZ corporation.
3. The stock was bought by XYZ corporation.
4. The purchase of the stock by XYZ corporation...
5. The stock purchase by XYZ corporation...

Commonality: there was a **purchase** event, the participants were **XYZ Corp** and some amount of **stock**, and **XYZ Corp** was **the buyer**.

**Semantic Role Labels** give a shallow semantic representation of the event and its arguments.

# Semantic Roles

Last time we discussed *neo-Davidsonian* event representations.

*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.*

$\exists e,x,y\ Opening(e) \wedge Opener(e,Pat) \wedge OpenedThing(e,y) \wedge Door(y)$

The **semantic role** of the subject of the ***break*** is ***Breaker***

The semantic of the subject of the ***open*** is ***Opener***

These **deep roles** are specific to each event.

# Thematic Roles

*Breakers* and *Openers* have something in common. They are both volitional actors, usually animate, and they have a direct causal responsibility for their events.

**Thematic roles** are a way to capture this semantic commonality between these roles. In this case, both *Breakers* and *Openers* fill the thematic role of AGENT.

AGENT is the thematic role that represents an abstract idea such as *volitional causation*.

BrokenThing and OpenedThing, are both prototypically inanimate objects that are affected in some way by the action. The semantic role for participant most directly affected by an event is THEME ;

# Thematic Roles

| Thematic Role | **Definition** |
|---|---|
| AGENT | The volitional causer of an event |
| EXPERIENCER | The experiencer of an event |
| FORCE | The non-volitional causer of the event |
| THEME | The participant most directly affected by an event |
| RESULT | The end product of an event |
| CONTENT | The proposition or content of a propositional event |
| INSTRUMENT | An instrument used in an event |
| BENFICIARY | The beneficiary of an event |
| SOURCE | The origin of the object of a transfer event |
| GOAL | The destination of an object of a transfer event |

# Thematic Roles

| Thematic Role | **Definition** |
| --- | --- |
| AGENT | *The player* kicked the ball. |
| EXPERIENCER | *Dan* has a cough and a fever. |
| FORCE | *The coronavirus* spread rapidly through the country |
| THEME | The wind blows *debris* from the street into our yard |
| RESULT | The city implemented *a stay-at-home policy for non-essential personnel* |
| CONTENT | Chris asked "*You met Rebecca at a supermarket?*" |
| INSTRUMENT | He poached catfish, stunning them with *a shocking device…* |
| BENFICIARY | Joe makes hotel reservations for *his boss*. |
| SOURCE | I flew in *from Boston*. |
| GOAL | I drove *to Portland.* |

# Verb Alternation

Semantic roles act as a shallow meaning representation that allow systems make simple inferences that aren't possible from the surface string of words, or from the parse tree.

Arguments for verbs alternate in their positions, which makes pure surface analysis difficult.

*John* *broke the window.*
AGENT THEME

*John* *broke the window with a rock.*
AGENT THEME INSTRUMENT

*The rock* *broke the window.*
INSTRUMENT THEME

*The window broke.*
THEME

*The window was broken by John.*
THEME AGENT

# Problems with Thematic Roles

Analysis of thematic roles should be useful for handling verb alternation. However, there is no single, standard set of thematic roles.

And it's quite difficult to come up with a formal definition for things like AGENT, THEME, or INSTRUMENT.

For example, there are two kinds of INSTRUMENTS, intermediary instruments that can appear as subjects and enabling instruments that cannot:

1. The cook opened the jar with the new gadget.
   The new gadget opened the jar.

2. Shelly ate the sliced banana with a fork.
   *The fork ate the sliced banana.

Different theories of thematic roles treat these differently, which causes fragmentation across theories.

# Generalized Semantic Roles

Instead of creating a more fine-grained inventory of Thematic Roles, research in NLP has shifted in the direction of coarser roles.

You may see terms like PROTO-AGENT and PROTO-PATIENT, which are generalized roles that express roughly agent-like and roughly patient-like meanings. These meanings are defined a set of heuristics.

A second direction that NLP goes in is to define semantic roles that are specific to each verb, or to a group of semantically related verbs or nouns.

Lexical resources that make use of this second direction are **PropBank** and **FrameNet**.

# PropBank

PropBank, is a resource of Penn TreeBank sentences annotated with semantic roles. It was created by Martha Palmer at UPenn.

Because defining universal thematic roles is difficult, PropBank defines semantic roles for each verb sense.

Each verb has a specific set of roles, given by numbers: Arg0, Arg1, Arg2.

In general, Arg0 represents the PROTO-AGENT, and Arg1, the PROTO-PATIENT.

Martha Palmer

# PropBank Frame File

**Agree.01**

Arg0: Agreer
Arg1: Proposition
Arg2: Other entity agreeing

Glosses to be read by humans

Example 1:
[Arg0 The group] ***agreed*** [Arg1 it wouldn't make an offer].

Example 2: [ArgM-TMP Usually] [Arg0 Chris] ***agrees*** [Arg2 with Ellie] [Arg1 on everything].

# PropBank Frame File

**Increase.01** “go up incrementally”

Arg0: Causer of increase
Arg1: Thing increasing
Arg2: Amount increased by
Arg3: Start point
Arg4: End point

Example 1: [Arg0 Big Fruit Co. ] **increased** [Arg1 the price of bananas].
Example 2: [Arg1 The price of bananas] was **increased** again [Arg0 by Big Fruit Co. ]
Example 3: [Arg1 The price of bananas] **increased** [Arg2 5% ].

# ArgMs

PropBank also has a number of non-numbered arguments called ArgMs, which represent modification meanings. These are stable across predicates, so aren’t listed with each frame file

| | | |
|---|---|---|
| **TMP** | when? | yesterday evening, now |
| **LOC** | where? | at the museum, in San Francisco |
| **DIR** | where to/from? | down, to Bangkok |
| **MNR** | how? | clearly, with much enthusiasm |
| **PRP/CAU** | why? | because ... , in response to the ruling |
| **REC** | | themselves, each other |
| **ADV** | miscellaneous | |
| **PRD** | secondary predication | ...ate the meat raw |

# FrameNet

In order to make semantic inferences about price increase events, we want to make the connection across many different verbs, not just the verb **increase.**

Example 1: [Arg1 The price of bananas] **increased** [Arg2 5% ].
Example 2: [Arg1 The price of bananas] **rose** [Arg2 5% ].
Example 3: There has been a [Arg2 5% ] **rise** in [Arg1 the price of bananas]

FrameNet is another Semantic Role Labeling project that attempts to address just these kinds of problems.

PropBank labels roles specific to an individual verb, and FrameNet labels roles are specific to a **frame**.

# Frames

What is a frame? Consider the following set of words:

*reservation, flight, travel, buy, price, cost, fare, rates, plane*

They form coherent chunk of common-sense background information concerning air travel.  The background knowledge that unites these words a **frame**.

The idea that groups of words are defined with respect to some back-ground information is widespread in AI and cognitive science.  Similar to the notion of a **script** that we saw before.

# Frame Elements

| Core Roles | |
|---|---|
| ATTRIBUTE | The ATTRIBUTE is a scalar property that the ITEM possesses. |
| DIFFERENCE | The distance by which an ITEM changes its position on the scale. |
| FINAL_STATE | A description that presents the ITEM's state after the change in the ATTRIBUTE's value as an independent predication. |
| FINAL_VALUE | The position on the scale where the ITEM ends up. |
| INITIAL_STATE | A description that presents the ITEM's state before the change in the ATTRIBUTE's value as an independent predication. |
| INITIAL_VALUE | The initial position on the scale from which the ITEM moves away. |
| ITEM | The entity that has a position on the scale. |
| VALUE_RANGE | A portion of the scale, typically identified by its end points, along which the values of the ATTRIBUTE fluctuate. |
| **Some Non-Core Roles** | |
| DURATION | The length of time over which the change takes place. |
| SPEED | The rate of change of the VALUE. |
| GROUP | The GROUP in which an ITEM changes the value of an ATTRIBUTE in a specified way. |

Frame: **Change position on a scale**

# Lexical Units

| | | | | | |
|---|---|---|---|---|---|
| **VERBS:** | dwindle | move | soar | escalation | shift |
| advance | edge | mushroom | swell | explosion | tumble |
| climb | explode | plummet | swing | fall | |
| decline | fall | reach | triple | fluctuation | **ADVERBS:** |
| decrease | fluctuate | rise | tumble | gain | increasingly |
| diminish | gain | rocket | | growth | |
| dip | grow | shift | **NOUNS:** | hike | |
| double | increase | skyrocket | decline | increase | |
| drop | jump | slide | decrease | rise | |

Lexical Units that trigger the **change position on a scale** frame

# FrameNet+: Fast Paraphrastic Tripling of FrameNet

**Ellie Pavlick[1] Travis Wolfe[2,3] Pushpendre Rastogi[2]**
**Chris Callison-Burch[1] Mark Dredze[2,3] Benjamin Van Durme[2,3]**
[1]Computer and Information Science Department, University of Pennsylvania
[2]Center for Language and Speech Processing, Johns Hopkins University
[3]Human Language Technology Center of Excellence, Johns Hopkins University

## Abstract

We increase the lexical coverage of FrameNet through automatic paraphrasing. We use crowdsourcing to manually filter out bad paraphrases in order to ensure a high-precision resource. Our expanded FrameNet contains an additional 22K lexical units, a 3-fold increase over the current FrameNet, and achieves 40% better coverage when evaluated in a practical setting on New York Times data.

## 1 Introduction

*Frame semantics* describes a word in relation to real-world events, entities, and activities. Frame semantic analysis can improve natural language understanding (Fillmore and Baker, 2001), and

**accurate**, **ambiguous**, **apparent**, **apparently**, audible, **axiomatic**, **blatant**, **blatantly**, **blurred**, **blurry**, **certainly**, **clarify**, clarity, clear, clearly, **confused**, **confusing**, **conspicuous**, **crystal-clear**, **dark**, **definite**, **definitely**, **demonstrably**, **discernible**, **distinct**, evident, **evidently**, **explicit**, **explicitly**, **flagrant**, **fuzzy**, **glaring**, **imprecise**, **inaccurate**, **lucid**, manifest, **manifestly**, **markedly**, **naturally**, **notable**, **noticeable**, **obscure**, **observable**, obvious, obviously, **opaque**, **openly**, **overt**, **patently**, **perceptible**, **plain**, **precise**, **prominent**, **self-evident**, show, show up, **significantly**, **soberly**, **specific**, **straightforward**, **strong**, **sure**, **tangible**, **transparent**, **unambiguous**, **unambiguously**, **uncertain**, unclear, **undoubtedly**, **unequivocal**, **unequivocally**, **unspecific**, **vague**, **viewable**, **visibility**, visible, **visibly**, **visual**, **vividly**, **well**,[1] **woolly**

Table 1: 81 LUs invoking the Obviousness frame according to the new FrameNet+. New LUs (bold) have been added using the method of paraphrasing and human-vetting described in Section 4.

# Semantic Role Labeling

SRL is the task of automatically finding the semantic roles of each argument of each predicate in a sentence.

Most state-of-the-art approaches to SRL use supervised machine learning, with FrameNet and PropBank providing training and test sets and defining what counts as a predicate and what the roles are.

# Primitive Decomposition of Predicates

One way of thinking about semantic is that they help us define the roles that arguments play in a **decompositional** way, based on finite lists of thematic roles.

1. Jim killed his philodendron.
   Jim did something to cause his philodendron to become not alive.

   KILL(x,y) ⇔ CAUSE(x, BECOME(NOT(ALIVE(y))))

2. John opened the door. ⇒ CAUSE(John, BECOME(OPEN(door)))

3. The door opened. ⇒ BECOME(OPEN(door))

The door is open. ⇒ OPEN(door)

# Conceptual dependency primitives

| Primitive | Definition |
|---|---|
| ATRANS | The abstract transfer of possession or control from one entity to another |
| PTRANS | The physical transfer of an object from one location to another |
| MTRANS | The transfer of mental concepts between entities or within an entity |
| MBUILD | The creation of new information within an entity |
| PROPEL | The application of physical force to move an object |
| MOVE | The integral movement of a body part by an animal |
| INGEST | The taking in of a substance by an animal |
| EXPEL | The expulsion of something from an animal |
| SPEAK | The action of producing a sound |
| ATTEND | The action of focusing a sense organ |

# Neo-Davidsonian Event with Primatives

The waiter brought Mary the check.

$\exists x,y\ Atrans(x) \land Actor(x, Waiter) \land Object(x, Check) \land To(x, Mary)$
$\land Ptrans(y) \land Actor(y, Waiter) \land Object(y, Check) \land To(y, Mary)$

# The Decompositional Semantics Initiative

*Rapid, simple, commonsensical annotations of meaning*

## *About*

The **Decompositional Semantics Initiative (Decomp)** collects and models simple, commonsensical annotations of meaning inspired by linguistic theory.

Traditional semantic annotation frameworks generally define complex, often exclusive category systems that require highly trained annotators to build. And in spite of their high quality for the cases they are designed to handle, these frameworks can be brittle to deviations from prototypical instances of a category.

http://decomp.io

# Conclusions

Semantic roles are abstract models of the role an argument plays in the event described by the predicate.

Thematic roles are a model of semantic roles based on a single finite list of roles.

Per-verb semantic role lists and proto-agent/proto-patient, are implemented in PropBank and FrameNet.

Semantic role labeling is the task of assigning semantic role labels to the constituents of a sentence.

Cool new research directions in finding event primatives.