Lecture 11:
Information Extraction and Semantic Role Labeling

Information Extraction

- **Definition:** Information extraction is the process of turning unstructured information embedded in texts into structured information (example: relational databases)

Named Entity Recognition

- The first step is usually Named Entity Recognition (NER), which we talked about earlier
  - Named entities often serve to define which entities appear in the text

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 lowercost 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.

Legend:
Organization
Person
Location

Co-reference Resolution

- A common second step is find equivalence classes of mentions that refer to the same entity

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 lowercost 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.

Relation Extraction

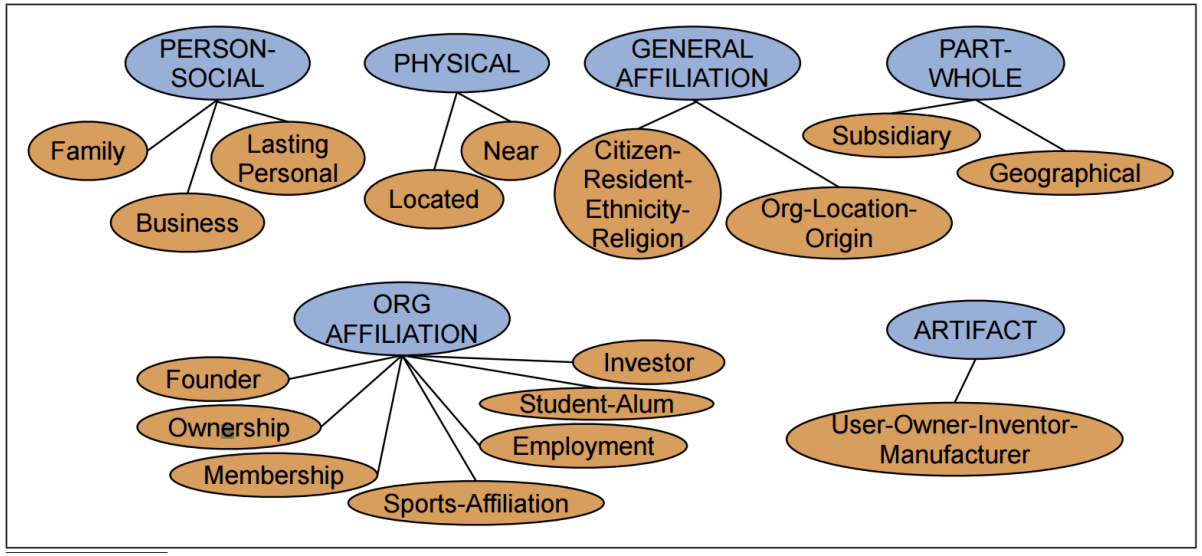
- Next on our list of tasks is to discern the relationships that exist among the detected entities

> 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 lowercost 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.

- For instance, the text tells us that Tim Wagner is a spokesman for American Airlines, that United is a unit of UAL Corp., and that American is a unit of AMR

Relation Extraction

- Often the relations are mapped to a pre-defined ontology
  - This is an example from the ACE shared task:
- For instance the aforementioned relations are instances of the PART-WHOLE relation:
  - “United is a unit of UAL Corp.”
  - “American is a unit of AMR”



ACE 2005 Multilingual Training Corpus
https://catalog.ldc.upenn.edu/ldc2006t06

Some Ontologies can be Mapped to Model-Theoretic Semantics

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	$Serves = \{\langle a, f\rangle, \langle a, g\rangle, \langle a, h\rangle, \langle a, i\rangle\}$

Relation Extraction

- There are many other ontologies of relations that have been defined
  - For example, UMLS, the Unified Medical Language System from the US National Library of Medicine has a network that defines 134 broad subject categories, such as:

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

Wikipedia Infoboxes

- Wikipedia also offers a range of relation types, drawn from the infoboxes
- For example, the Wikipedia infobox for Brad Pitt includes structured facts like *Occupation = "Actor"/"Producer"* or *Born = "Shawnee, Oklahoma"*
- These facts can be turned into relations like *born-in* or *occupied_as*
- These relations can complement relations extracted from text, or serve as features for them
- *Wikidata* (https://www.wikidata.org) is a knowledge base based on Wikipedia

Brad Pitt

Pitt at the premiere of *Fury* in Washington, D.C., October 2014

Born	William Bradley Pitt December 18, 1963 (age 53) Shawnee, Oklahoma, U.S.
Occupation	Actor • producer
Years active	1987–present
Works	Filmography
Home town	Springfield, Missouri
Spouse(s)	Jennifer Aniston (m. 2000; div. 2005) Angelina Jolie (m. 2014; separated 2016)
Children	6
Relatives	Douglas Pitt (brother)

Using Patterns to Discover Relations

- The earliest and still common approach for relation extraction is the use of lexico-syntactic patterns

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

Hearst, M. Automatic Acquisition of Hyponyms from Large Text Corpora, *Proceedings of the Fourteenth International Conference on Computational Linguistics*, 1992

Using Patterns to Discover Relations

- More modern approaches use additional features to define the patterns, such as named entity constraints
- For instance, if our goal is to answer questions about “who holds what office in which organization?”, we can use patterns like the following:

PER, POSITION of ORG:
George Marshall, Secretary of State of the United States

PER (named|appointed|chose|etc.) PER Prep? POSITION
Truman appointed Marshall Secretary of State

PER [be]? (named|appointed|etc.) Prep? ORG POSITION
George Marshall was named US Secretary of State

Using Patterns to Discover Relations

- More accurate patterns still are ones that use syntactic information
- For instance, symmetric patterns can be used to discover synonymy or relatedness between entities

from **X** to **Y**

X and **Y**

X or **Y**

neither **X** nor **Y**

X as well as **Y**

Using Patterns to Discover Relations

- Syntactic information can filter a lot of noise
- What are the *Xs* and *Ys* in these cases:
  - "when they go to Austria, they like walking in the woods **as well as** skiing"
  - "apricots **and** other vegetables **and** fruit"
  - "Sandy is a Republican **and** proud of it"

from **X** to **Y**

X and **Y**

X or **Y**

neither **X** nor **Y**

X as well as **Y**

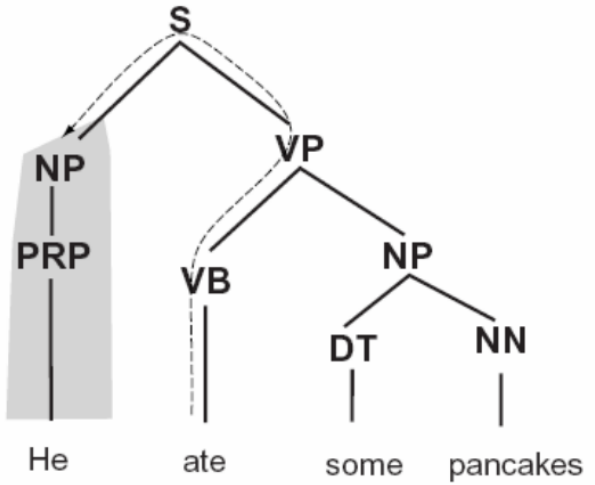
Distant Supervision for Relation Extraction

- Hand-labeled data with relation labels is expensive to produce

- However, available resources such as Wikipedia infoboxes, have a great many relations that are in structured format
  - Wikipedia articles contain sentences that express these relations
  - This huge amount of examples allows us to define rich features

Path Features

- Path features:
  - Encode the path between two nodes, through the direction of the edges and their labels
- Similar features exist for dependency parses
  - We will see a bit of this in Exercise 4



Path	*Description*
VB↑VP↓PP	PP argument/adjunct
VB↑VP↑S↓NP	subject
VB↑VP↓NP	object
VB↑VP↑VP↑S↓NP	subject (embedded VP)
VB↑VP↓ADVP	adverbial adjunct
NN↑NP↑NP↓PP	prepositional complement of noun

Examples of Rich Features

...Hubble was born in Marshfield...
...Einstein, born 1879, Ulm...
...Hubble's birthplace in Marshfield...

PER was born in LOC
PER, born *, LOC
PER's birthplace is LOC

American Airlines, a unit of AMR, immediately matched the move, spokesman **Tim Wagner** said

Constituent path	$NP \uparrow NP \uparrow S \uparrow S \downarrow NP$
Base phrase path	$NP \rightarrow NP \rightarrow PP \rightarrow NP \rightarrow VP \rightarrow NP \rightarrow NP$
Typed-dependency path	$Airlines \leftarrow_{subj} matched \leftarrow_{comp} said \rightarrow_{subj} Wagner$

Schematized Algorithm for 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*

Evaluation of Relation Extraction

- Semi-supervised methods are much more difficult to evaluate than supervised methods
  - They extract **new** relations from the web or a large text
  - As methods use very large amounts of text, it is impossible to test them in a sand box with a small pre-annotated gold standard
- Evaluation is therefore done by:
  - Computing precision
  - If the system can rank its output, we can measure precision for different output sizes (e.g., precision for 100 relations, 1000 relations etc.)
  - Extrinsic evaluation by testing how well these relations help, say, question answering

Poon and Domingos, *Unsupervised Semantic Parsing*, EMNLP 2009

Open Information Extraction

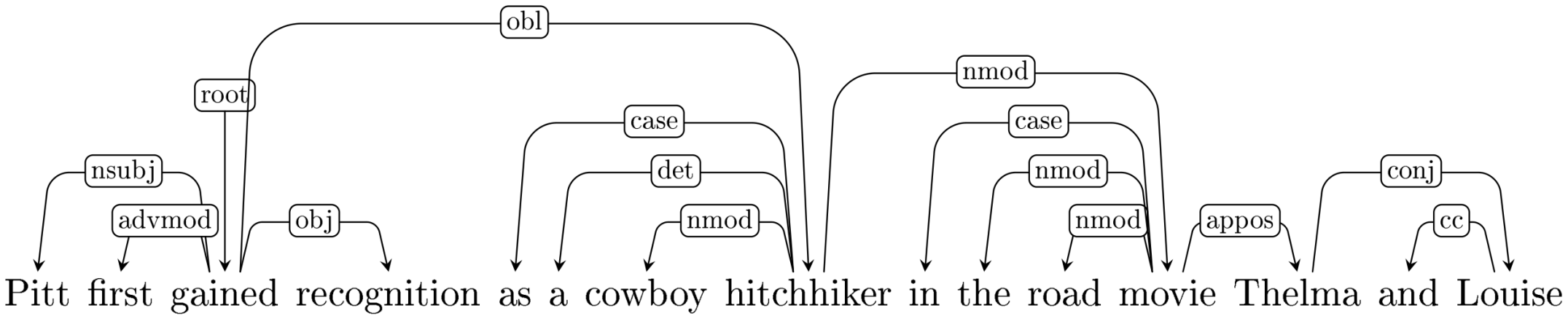
- Open Information Extraction (OpenIE) seeks to discover information from plain text
  - One way is to apply syntactic parsing and extract the relations it finds between named entities
  - Consider this example:

    Pitt first gained recognition as a cowboy hitchhiker in the road movie **Thelma and Louise** (1991). His first leading roles in big-budget productions came with the dramas **A River Runs Through It** (1992) and **Legends of the Fall** (1994), and **Interview with the Vampire** (1994).

  - What's the relation between "Pitt" and "Thelma and Louise"? Between "A River Runs Through It" and "Interview with the Vampire"?

Open Information Extraction

- The dependency path between *Pitt* and *Thelma and Louise* is *nsubj*↑*obl*↓*nmod*↓*appos*



- This is an indirect syntactic relation
  - It would be useful to have a representation that directly encodes the **semantic** relations between the participants

Syntax doesn't always align w/ Semantics

Mary opened the door.
The door opened.

John slices the bread with a knife.
The bread slices easily.
The knife slices easily.

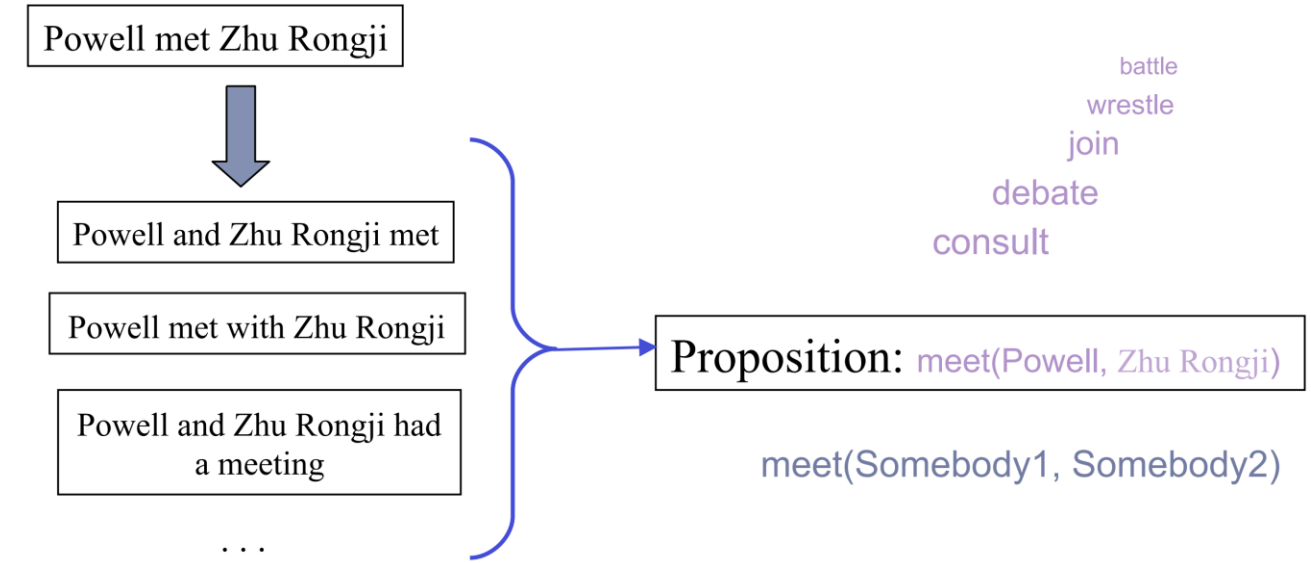
Mary loaded the truck with hay.
Mary loaded hay onto the truck.
The truck was loaded with hay (by Mary).
Hay was loaded onto the truck (by Mary).

John got Mary a present.
John got a present for Mary.
Mary got a present from John.

Examples by Nathan Schneider

Semantic Role Labeling (SRL)

- The task of *Semantic Role Labeling* aims to represent a sentence in terms of its events or proposition
- Events consist of a main relations, participants and secondary relations



When Powell met Zhu Rongji on Thursday they discussed the return of the spy plane.

meet(Powell, Zhu) discuss([Powell, Zhu], return(X, plane))

Semantic Role Labeling

- Informally - determining **who** did **what** to **whom**, **where**, **when** and **how**

- More formally – **identifying**, for each predicate, its set of arguments and **establishing the semantic role** of each of them
  - Roles are usually taken from a pre-defined ontology

- Produces a flat (not hierarchical) structure for each of the predicates

SRL Example

Acceptor Modal Negation **Predicate** Thing Accepted

He would n't **accept** anything of value

from those he was writing about.

Accepted From

Semantic Roles

- Semantic role:
  - An underlying relationship an argument has with the main verb in a clause
  - This relation is (ideally) invariant to paraphrases
- Many semantic role lists (or ontologies) in the literature
- In this lecture – the prevailing approaches in NLP

Annotation Schemes of Semantic Roles

- We will briefly survey the three main representation approaches for SRL, which differ in:
  1. Organizing principles
  2. Scope (what is considered an event)
  3. Granularity
  4. Consistency within and across verbs

SRL Schemes: FrameNet

- Organized by *frames*: a schematized event type
- A frame is invoked by frame elements, generally words or morphemes
  - They constitute the anchor of the scene and determine what it is about
- Semantic roles are defined per frames
- Example:
  - The *Judgment* frame.
  - Frame elements: *admire, appreciate, value*…

Judge Evaluee Reason

She **blames** the government for failing to do enough to help

https://framenet.icsi.berkeley.edu/fndrupal

FrameNet Judgement Frame Example

Judgment

Lexical Unit Index

Definition:

A Cognizer makes a judgment about an Evaluee. The judgment may be positive (e.g. respect) or negative (e.g. condemn), and this information is recorded in the semantic types Positive and Negative on the Lexical Units of this frame. There may be a specific Reason for the Cognizer's judgment, or there may be a capacity or Role in which the Evaluee is judged.

This frame is distinct from the Judgment_communication frame in that this frame does not involve the Cognizer communicating his or her judgment to an Addressee.
JUDGMENT: She ADMIRED Einstein for his character.

JUDGMENT_COMMUNICATION: She ACCUSED Einstein of collusion.

Currently, however, some lexical units and annotation for both remain in this frame.

FEs:

Core:

Cognizer [Cog] **Semantic Type:** Sentient	The Cognizer makes the judgment. This role is typically expressed as the External Argument (or in a by-PP in passives). The boss APPRECIATES you for your diligence. The boss is very APPRECIATIVE of my work.
Evaluee [Eval]	Evaluee is the person or thing about whom/which a judgment is made. With verbs this FE is typically expressed as Object: The boss APPRECIATES you for your diligence.
Expressor [Exr]	Expressor is the body part or action by a body part that conveys the judgment made by the Cognizer. She viewed him with an APPRECIATIVE gaze.
Reason [Reas] **Semantic Type:** State_of_affairs	Typically, there is a constituent expressing the REASON for the Judge's judgment. It is usually a for-PP, e.g. I ADMIRE you for your intellect.

SRL Schemes: PropBank

- Predicate-specific core roles, with adjunct roles (such as temporal description, locations, manner adverbs, negation) are shared across predicates
  - So each predicate (e.g., *blame*) has core arguments which are indexed A0-A5
  - Non-core predicates are marked *AM-**

https://verbs.colorado.edu/verb-index

PropBank includes Basic Predicate Sense Disambiguation

- **Decline.01** – "go down incrementally"
  - *Arg1:* Entity going down
  - *Arg2:* Amount gone down by
  - *Arg3:* Start point
  - *Arg4:* End point

[*A1* Its net income] declining [*A2* 42%] [*A4* to $121 million] [*AM-TMP* in the first 9 months of 1989]

- **Decline.02** – "demure, reject"
  - *Arg1:* Agent
  - *Arg2:* Rejected Thing

[*A1* A spokesman] declined [*A2* to elaborate]

SRL Schemes: VerbNet

- A hierarchical verb lexicon
- Builds on *Levin Classes* (Beth Levin, 1993)
  - The hypothesis is that verbs that have a similar syntactic distribution, also have a shared meaning component
  - Each category lists a set of syntactic environments in which the members of the category may appear
- The semantic roles are shared across frames (only 22 roles are used through the lexicon)
- Examples of roles: Agent, Patient, Theme, Oblique, Beneficiary

SRL Schemes: VerbNet (Example)

No Comments

judgment-33-1-1

Members: 7, Frames: 1

Members

ACCLAIM (FN 1; WN 1)
DOUBT (FN 1; G 2)
HAIL (FN 1; WN 1; G 1)
HERALD (FN 1; WN 2)
JUDGE (WN 4; G 2)
LAUD (FN 1; WN 1)
PRAISE (FN 1; WN 1; G 1)

Roles

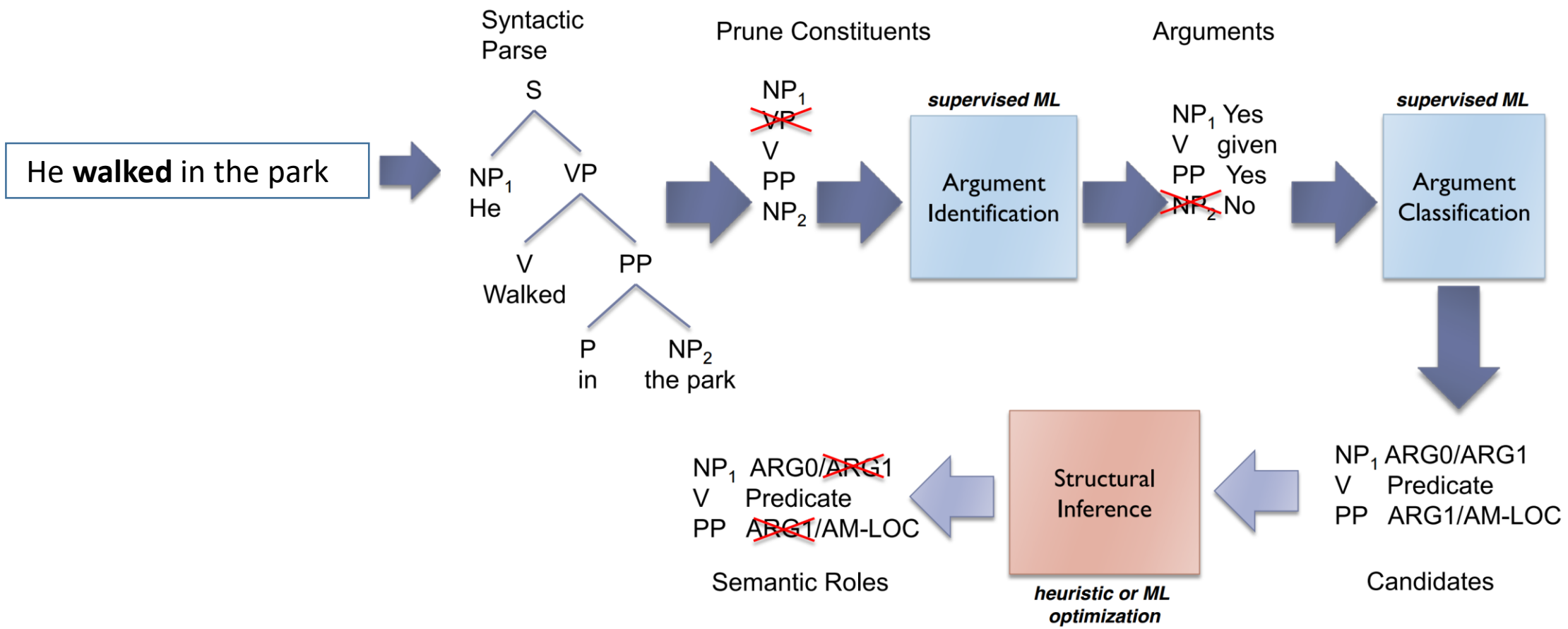
NO ROLES

Frames

NP V NP TO BE NP

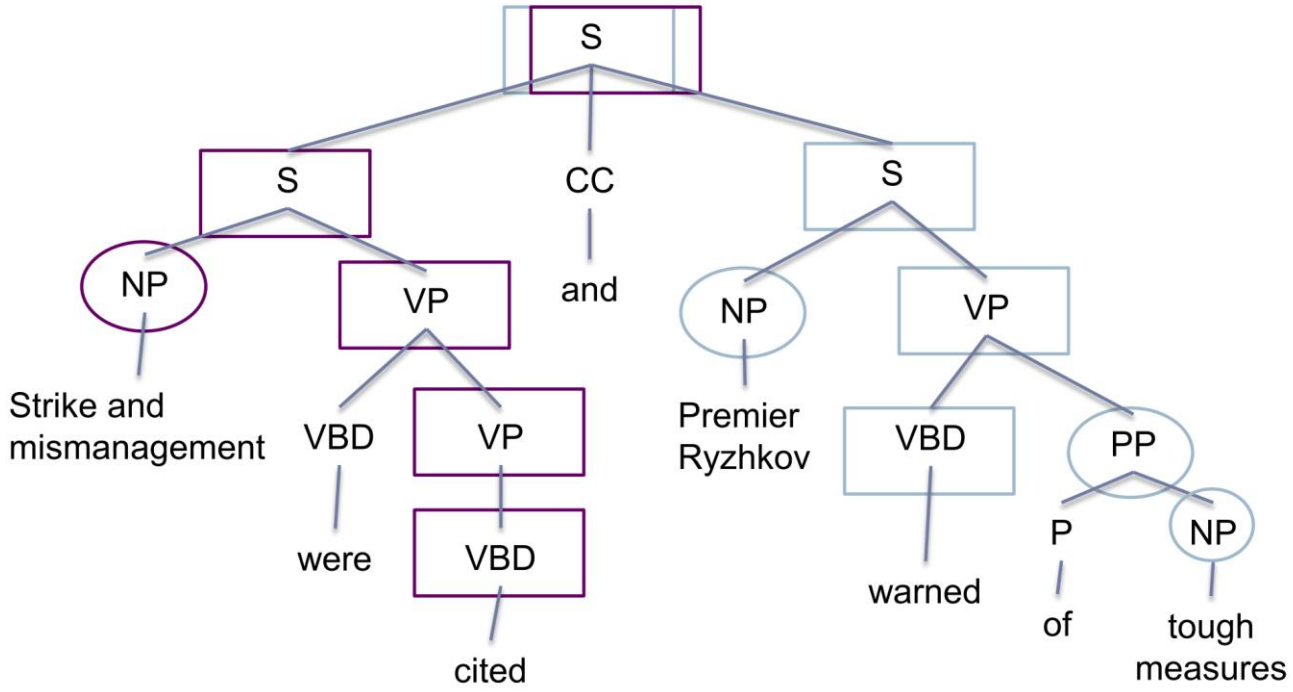
EXAMPLE "I judged him to be a good man."
SYNTAX AGENT V THEME ATTRIBUTE <+SMALL_CLAUSE>
SEMANTICS DECLARE(DURING(E), AGENT, THEME, ATTRIBUTE)

SRL Parsing: A Common Architecture



Pruning Algorithm

- For the predicate and each of its ancestors, collect their sisters unless the sister is coordinated with the predicate
- If a sister is a PP also collect its immediate children



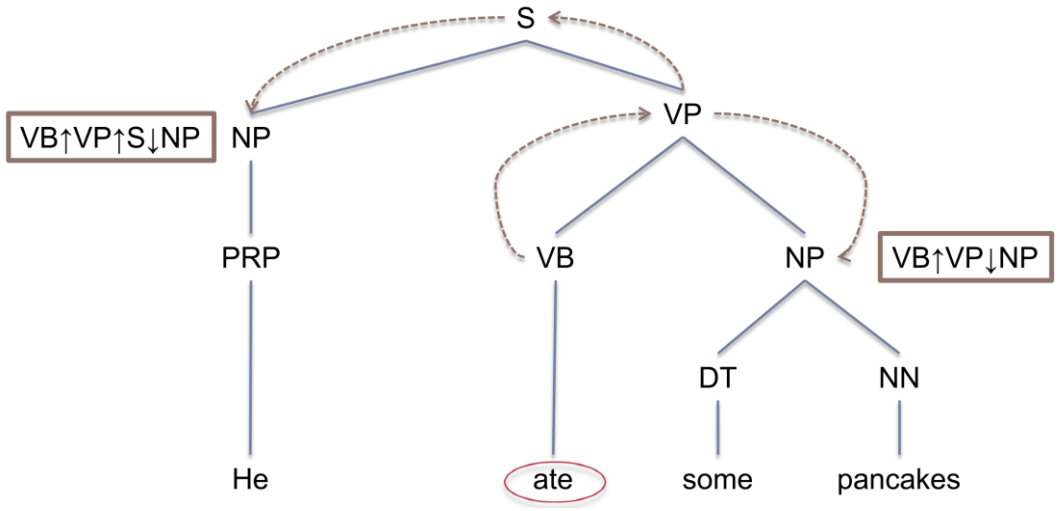
ML for Argument Identification/Labeling

Training:

1. Extract features from sentence, syntactic parse, and other sources for each candidate constituent
2. Train statistical ML classifier to identify arguments
3. Extract features same as or similar to those in step 1
4. Train statistical ML classifier to select appropriate label for arguments

Commonly Used Features

1. Phrase type of the consistuent
2. Parse tree path; **BUT:**
   1. Parsers aren't perfect:
   2. Data sparseness: thousands of possible values
3. Subcategorization frame
   - Such as, *subj V obj, subj V in+NP etc.*
   - Intuition: Knowing the number of arguments to the verb constrains the possible set of semantic roles



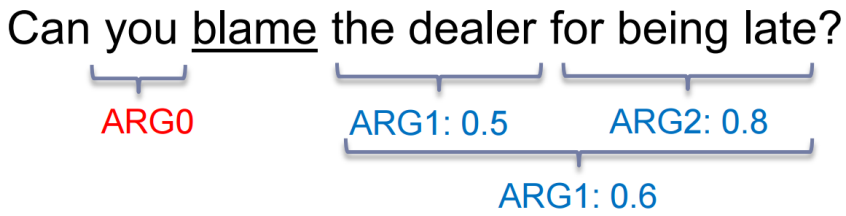
Commonly Used Features

4. Position relative to the predicate (right/left)
5. Voice: identify if the verb is in passive
6. Head word: head of constituent
7. Named entities: which are in the clause?
8. Verb cluster (according to distributional clustering): Similar verbs share similar argument sets
9. First/last word of constituent
10. Previous role

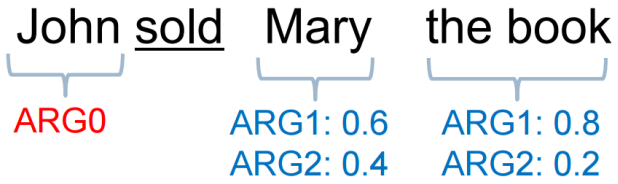
Structural Inference

- The third and final stage of classification
- Takes all arguments into account to end up with a globally consistent label assignment

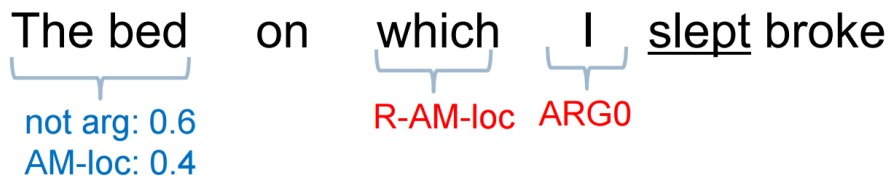
- Arguments should not overlap



- Numbered arguments (arg0-5) should not repeat



- R-arg[type] and C-arg[type] should have an associated arg[type]



Structural Inference Methods

1. Optimize log probability of label set

The probability of assigning a label l_i to the *i-th* argument

$$max_{l_1,\ldots,l_N} \frac{1}{N} \sum_i \log(p(i, l_i))$$

- Formulate into integer linear programming (ILP) problem

2. Alternative: re-rank top label sets that conform to constraints
   - Choose n-best label sets for each argument
   - Train sequence classifiers (e.g., CRF) to predict the most likely sequence

SRL Evaluation

- Recall, Precision, F-score
- Variants when evaluating in SRL:
  - Arguments: Full span (CoNLL-2005), Headword only (CoNLL-2008)
- Predicates:
  - Given (CoNLL-2005)
  - System Identifies (CoNLL-2008)
  - Verb and nominal predicates (CoNLL-2008)

SRL Evaluation

Gold Standard Labels	SRL Output	Full	Head
Arg0: John	Arg0: John	+	+
Rel: mopped	Rel: mopped	+	+
Arg1: the floor	Arg1: the floor	+	+
Arg2: with the dress … Thailand	Arg2: with the dress	-	+
Arg0: Mary	Arg0: Mary	+	+
Rel: bought	Rel: bought	+	+
Arg1: the dress	Arg1: the dress	+	+
Arg0: Mary		-	-
rel: studying		-	-
Argm-LOC: in Thailand		-	-
Arg0: Mary	Arg0: Mary	+	+
Rel: traveling	Rel: traveling	+	+
Argm-LOC: in Thailand		-	-

John mopped the floor with the dress Mary bought while studying and traveling in Thailand.

Evaluated on **Full** Arg Span

Precision

P = 8 correct / 10 labeled = 80.0%

Recall

R = 8 correct / 13 possible = 61.5%

Evaluated on **Head** word Arg

Precision

P = 9 correct / 10 labeled = 90.0%

Recall

R = 9 correct / 13 possible = 69.2%

Back to the Relation Extraction Example

Pitt first gained recognition as a cowboy hitchhiker in the road movie **Thelma & Louise** (1991)

gained recognition

Theme: Pitt

Role: as a cowboy hitchhiker

Means: Thelma & Louise