Information Extraction and Entity/Event Semantics

CS 585, Fall 2016
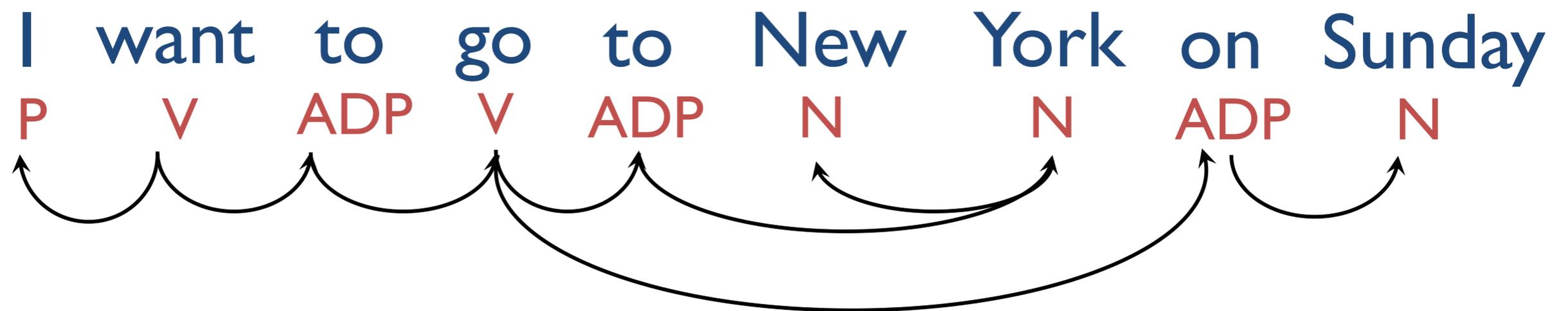
Introduction to Natural Language Processing
http://people.cs.umass.edu/~brenocon/inlp2016

Brendan O'Connor

College of Information and Computer Sciences
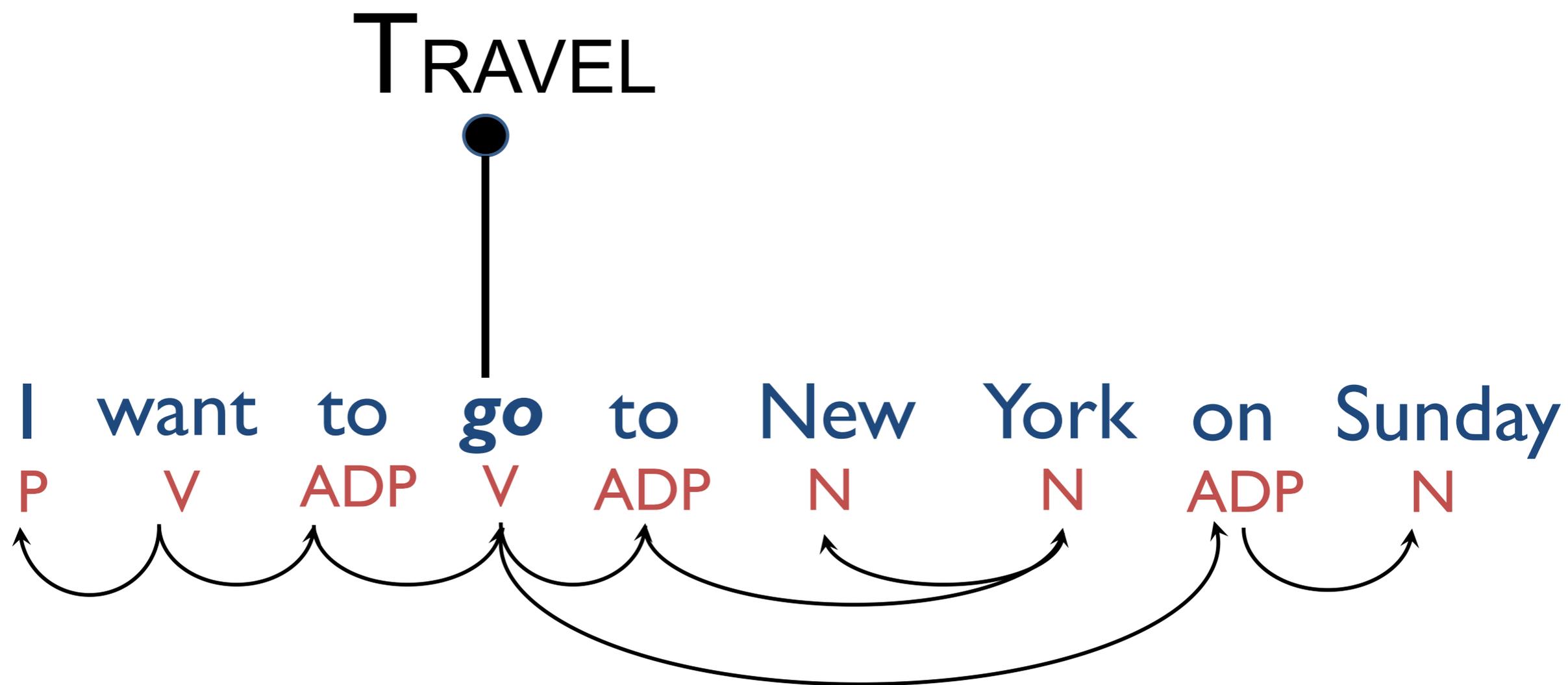University of Massachusetts Amherst

Meaning

- Lexical semantics: individual words/phrases
  - KBs, embeddings, etc.
- But much much more ...
  - Logical semantics [e.g. questions as database queries ... theorem proving ...]
- "Shallow" semantics: predicates, arguments
  - who did what to whom?
  - I bought a car from him <=> he sold me a car
- Practical examples: Information Extraction
- Major subtasks
  - Entities and coreference
    - I saw Bob, and **he** said hi
  - Time and Events

I want to go to New York on Sunday

P   V   ADP   V   ADP   N   N   ADP   N

Natural Language Understanding
Shallow Semantics: *Frames* and *Roles*

TRAVEL

I want to **go** to New York on Sunday
P V ADP V ADP N N ADP N

[Slides: *Dipanjan Das*]

Natural Language Understanding
Shallow Semantics: *Frames* and *Roles*

T RAVEL

Traveler

Goal

Time

I want to **go** to New York on Sunday

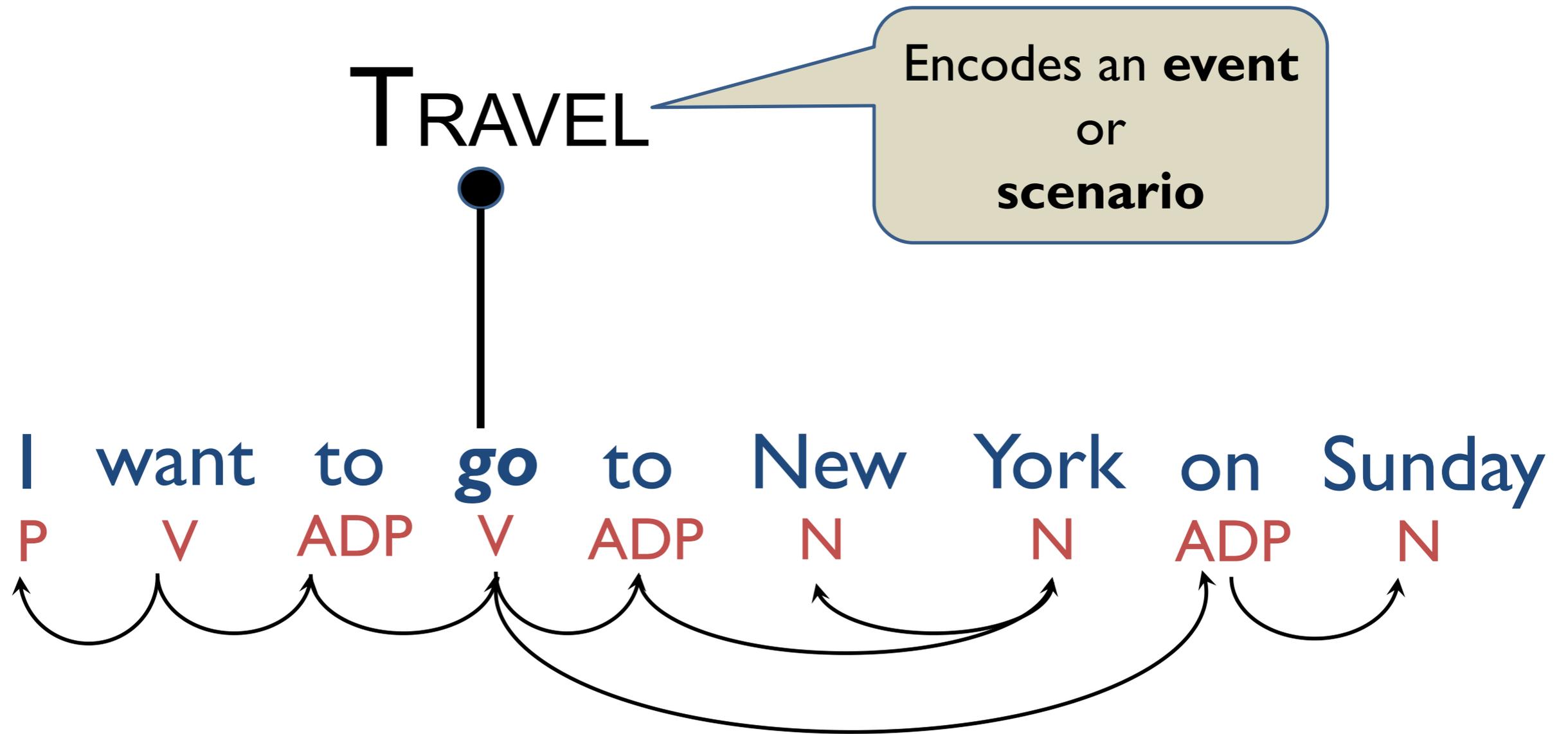P   V   ADP   V   ADP   N   N   ADP   N

[Slides: *Dipanjan Das*]

Natural Language Understanding
Shallow Semantics: *Frames* and *Roles*

[Slides: *Dipanjan Das*]

Thursday, December 1, 16
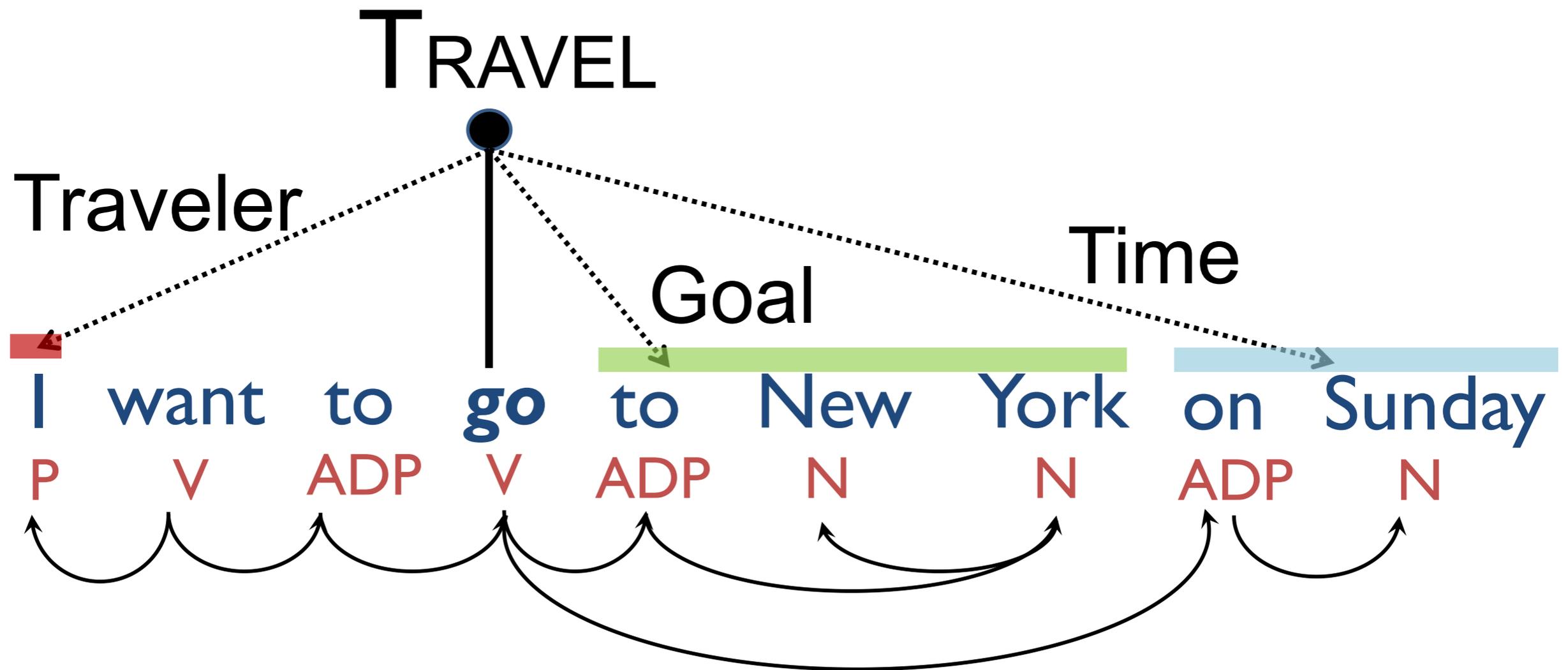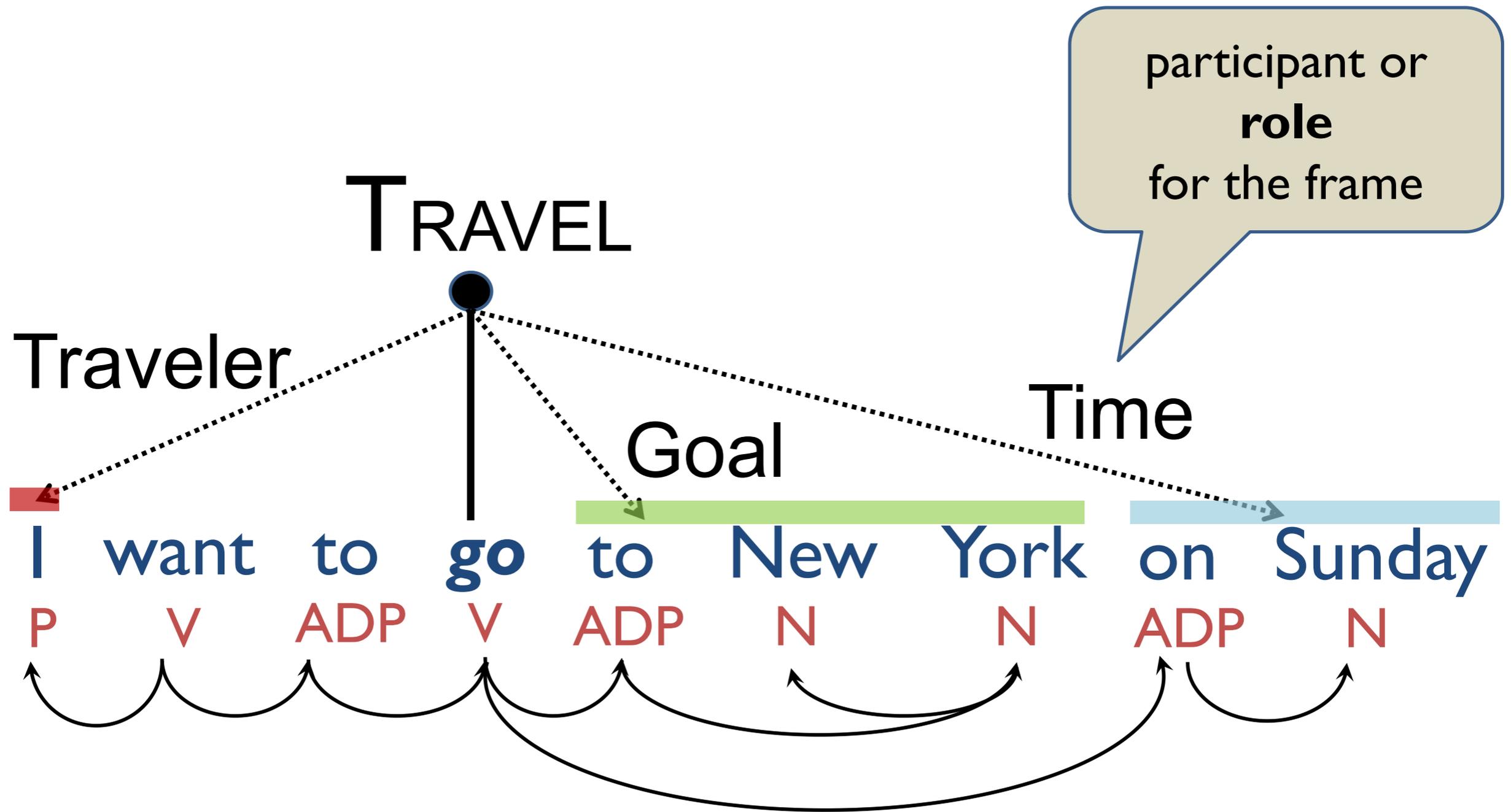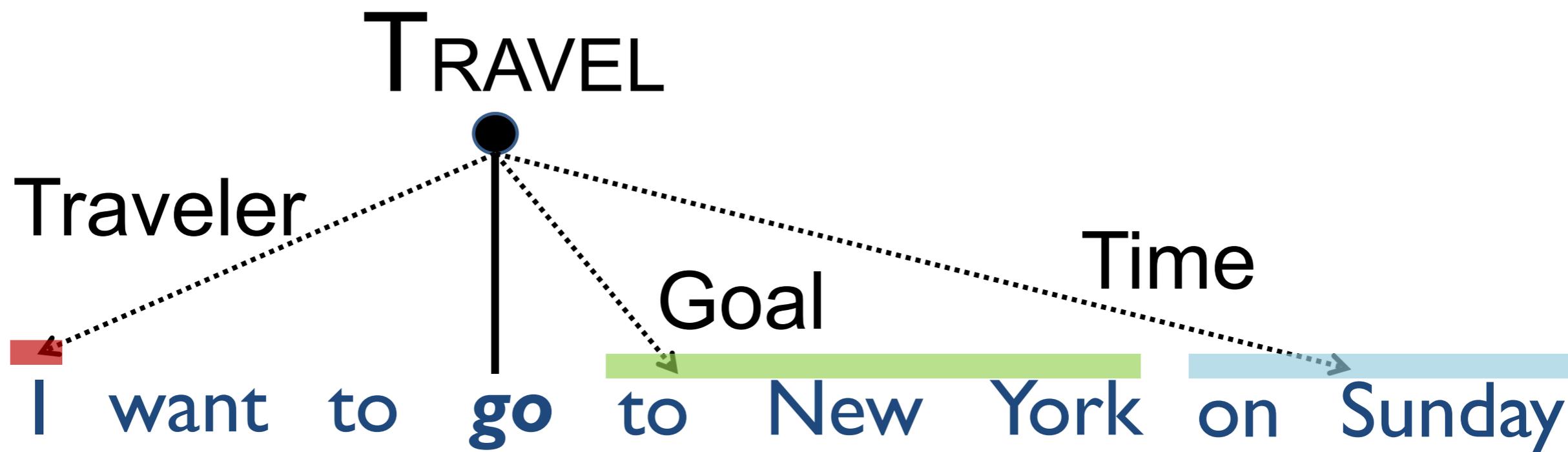
Natural Language Understanding
Shallow Semantics: *Frames* and *Roles*

TRAVEL

Traveler

Goal

Time

I want to **go** to New York on Sunday

[Slides: Dipanjan Das]

Shallow Semantics: *Frames* and *Roles*
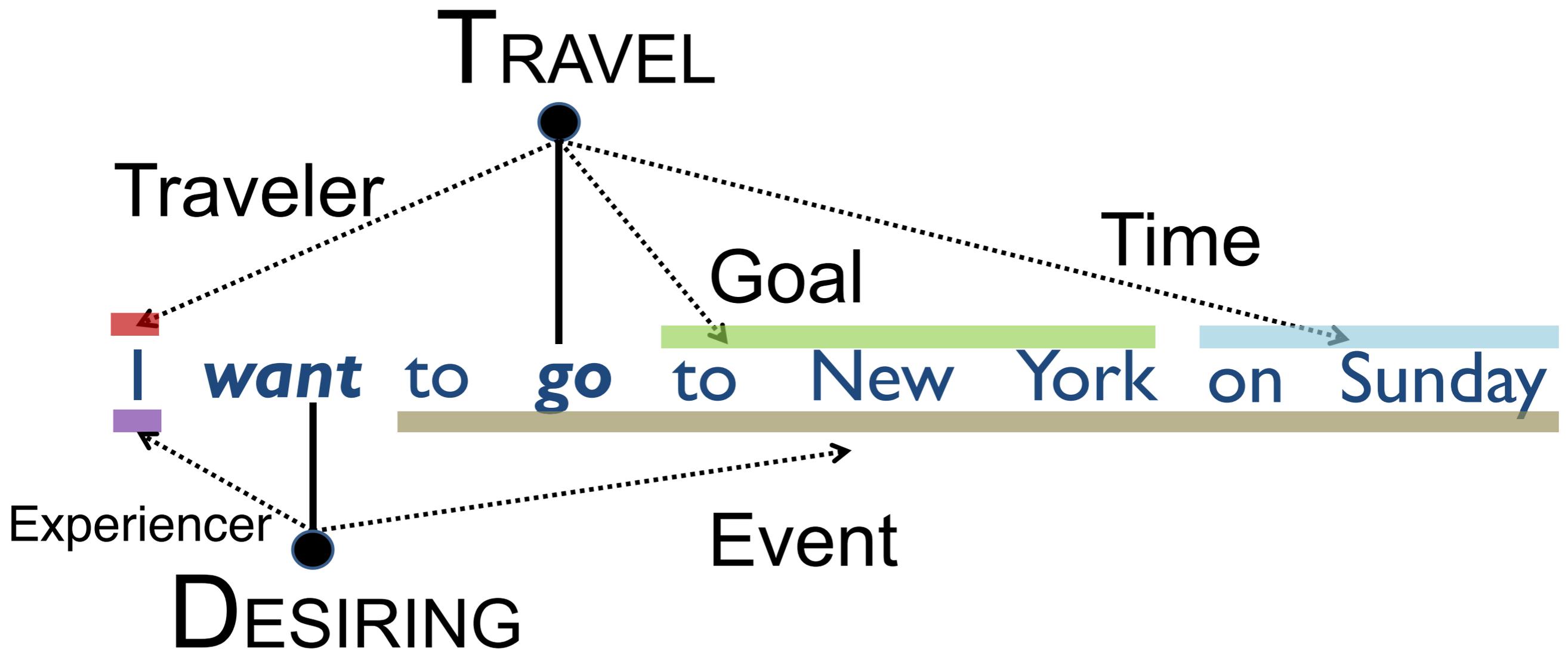
[Slides: Dipanjan Das]

Application: event analysis in international relations

- Analyze time-series of friendly vs. hostile country-country interactions, coded from newswire text

  - Manual coding (~1960's): hire people to read thousands of articles (inconsistencies!)

  - Machine coding (KEDS) -- rule-based S-V-O or S-V-PP extraction *[Phil Schrodt (1993, 1994... 2011)]*

- Various current efforts: ICEWS, OEDA, etc.

Application: event analysis in international relations

EXAMPLES OF WEIS EVENT CODES

11. REJECT

111	Turn down proposal; reject protest demand; threat
112	Refuse; oppose; refuse to allow

12. ACCUSE

121	Charge, criticize, blame, disapprove
122	Denounce, denigrate, abuse

13. PROTEST

131	Make complaint (not formal)
132	Make formal complaint or protest

17. THREATEN

171	Threat without specific negative sanctions
172	Threat with specific nonmilitary negative sanctions
173	Threat with force specified
174	Ultimatum: threat with negative sanctions and time

18. DEMONSTRATE

181	Non-military demonstration; walk out on
182	Armed force mobilization, exercise and/or display

Table 2
WEIS Coding of 1990 Iraq-Kuwait Crisis

Date	Source	Target	WEIS Code	Type of Action
900717	IRQ	KUW	121	CHARGE
900717	IRQ	UAE	121	CHARGE
900723	IRQ	KUW	122	DENOUNCE
900724	IRQ	ARB	150	DEMAND
900724	IRQ	OPC	150	DEMAND
900725	IRQ	EGY	054	ASSURE
900727	IRQ	KUW	160	WARN
900731	IRQ	KUW	182	MOBILIZATION
900801	KUW	IRQ	112	REFUSE
900802	IRQ	KUW	223	MILITARY FORCE

Application: event analysis in international relations

Figure 1
USA Actions Towards USSR, 1948-1978
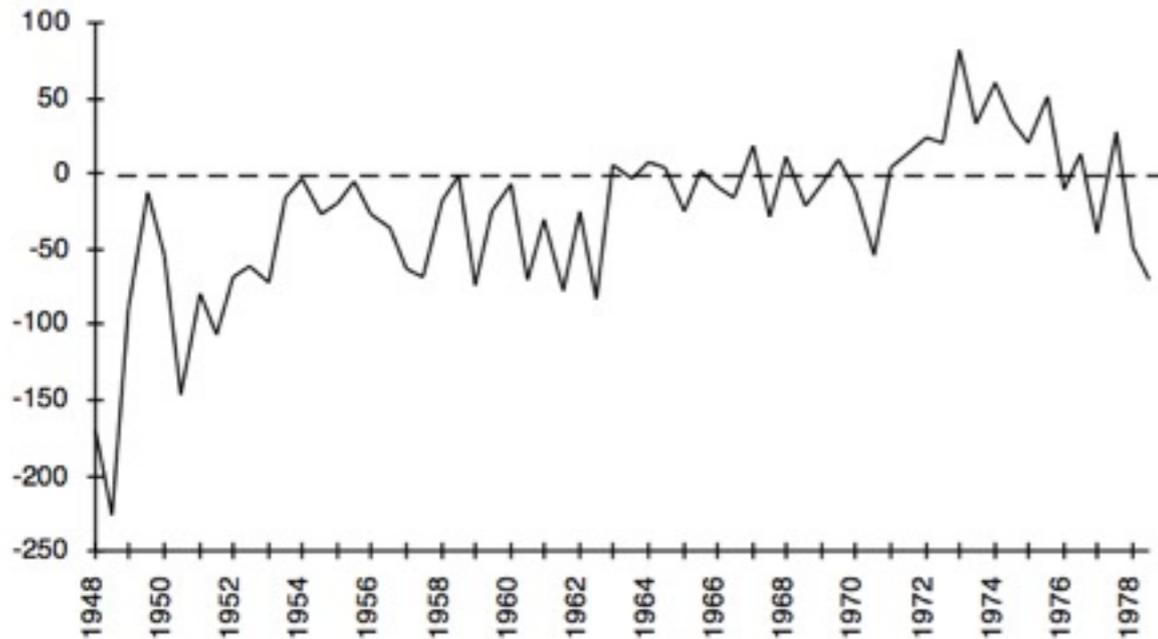
Figure 2
Israel-Palestinian interactions, 1982-1992
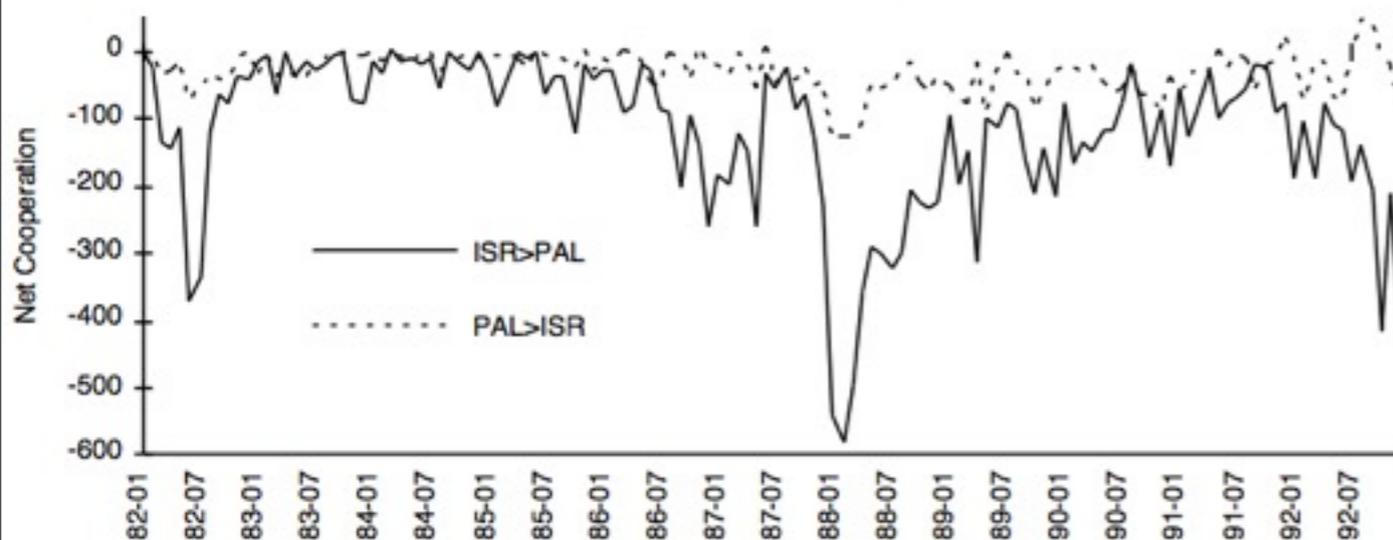
Crosscorrelation of Arms Transfers and
International Cooperation from Recepient to Supplier

<- before sale [years]   after sale ->
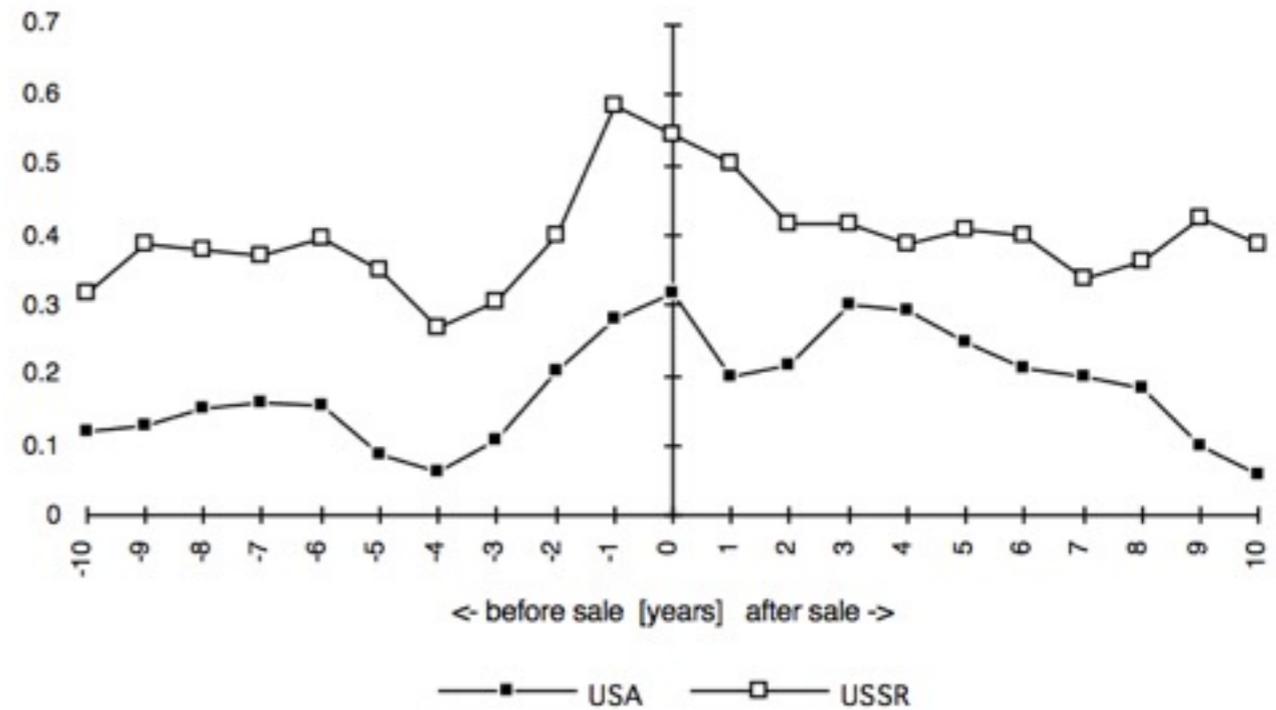
(These graphs are from manual coding; IE evaluations in Schrodt and Gerner 1994, King and Lowe 2001)

Message Understanding Conferences (MUC)

- Bakeoff format: shared task, dataset, hidden test set for competitive evaluation

- Different domains – involving specific events

  - (1987) MUC-1: Fleet operations

  - (1991-2) MUC-3, 4: Terrorist activities in Latin America

  - (1993-7) Corporate Joint Ventures, Microelectronic production, Negotiation of Labor Disputes, Airplane crashes, and Rocket/Missile Launches

- ACE (1999-2008) – Automated Content Extraction

MUC Template-Filling IE

Input: text

San Salvador, 19 Apr 89 (ACAN-EFE) – [TEXT] Salvadoran President-elect Alfredo Cristiani condemned the terrorist killing of Attorney General Roberto Garcia Alvarado and accused the Farabundo Marti National Liberation Front (FMLN) of the crime.

...

Garcia Alvarado, 56, was killed when a bomb placed by urban guerrillas on his vehicle exploded as it came to a halt at an intersection in downtown San Salvador.

...

Vice President-elect Francisco Merino said that when the attorney general's car stopped at a light on a street in downtown San Salvador, an individual placed a bomb on the roof of the armored vehicle.

...

According to the police and Garcia Alvarado's driver, who escaped unscathed, the attorney general was traveling with two bodyguards. One of them was injured.

Output: extract an event record ("terrorist attack") with the following attributes:

Incident: Date
Incident: Location
Incident: Type
Perpetrator: Individual ID
Perpetrator: Organization ID
Perpetrator: Organization
        Confidence
Physical Target: Description
Physical Target: Effect
Human Target: Name
Human Target: Description

MUC Template-Filling IE

Input: text

San Salvador, 19 Apr 89 (ACAN-EFE) – [TEXT] Salvadoran President-elect Alfredo Cristiani condemned the terrorist killing of Attorney General Roberto Garcia Alvarado and accused the Farabundo Marti National Liberation Front (FMLN) of the crime.

. . .

Garcia Alvarado, 56, was killed when a bomb placed by urban guerrillas on his vehicle exploded as it came to a halt at an intersection in downtown San Salvador.

. . .

Vice President-elect Francisco Merino said that when the attorney general's car stopped at a light on a street in downtown San Salvador, an individual placed a bomb on the roof of the armored vehicle.

. . .

According to the police and Garcia Alvarado's driver, who escaped unscathed, the attorney general was traveling with two bodyguards. One of them was injured.

Output: extract an event record ("terrorist attack") with the following attributes:

Incident: Date	- 19 Apr 89
Incident: Location	El Salvador: San Salvador (city)
Incident: Type	Bombing
Perpetrator: Individual ID	"urban guerrillas"
Perpetrator: Organization ID	"FMLN"
Perpetrator: Organization Confidence	Suspected or Accused by Authorities: "FMLN"
Physical Target: Description	"vehicle"
Physical Target: Effect	Some Damage: "vehicle"
Human Target: Name	"Roberto Garcia Alvarado"
Human Target: Description	"attorney general": "Roberto Garcia Alvarado" "driver" "bodyguards"

FASTUS System

- Hobbs, Appelt, Bear, Israel, Kameyana, Stickel, Tyson 1997, "A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text."

  - From SRI, for early-90's MUC

- Pipeline of hand-built patterns **--** but statistically guided development

- Great case study: realistic end-to-end system, with clear architecture, formalisms, and *engagement with the data*

  - Example of how to build a rule-based NLP system **--** useful skill in a pinch

Pipeline
(finite-state transducers)

Text

1. Complex Words

2. Basic Phrases

3. Complex Phrases

4. Domain Events

5. Merging Structures

Structure

<u>Syntax steps</u>

Names, multiwords...

NPs, verb groups, phrase structure...

<u>Domain-specific semantics</u>

Event Patterns

<Company/ies> <Set-up> <Joint-Venture>
with <Company/ies>

↓

Relationship:	TIE-UP
Entities:	"Bridgestone Sports Co."
	"a local concern"
	"a Japanese trading house"
Joint Venture Company:	–
Activity:	–
Amount:	–

<Produce> <Product>

↓

Activity:	PRODUCTION
Company:	–
Product:	"golf clubs"
Start Date:	–

Bridgestone Sports Co. said Friday it has set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be shipped to Japan.

The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and "metal wood" clubs a month.

(5/5) Merge Structures

Activity:	PRODUCTION
Company:	–
Product:	"golf clubs"
Start Date:	–

Activity:	PRODUCTION
Company:	"Bridgestone Sports Taiwan Co."
Product:	–
Start Date:	DURING: January 1990

Relationship:	TIE-UP
Entities:	"Bridgestone Sports Co."
	"a local concern"
	"a Japanese trading house"
Joint Venture Company:	–
Activity:	–
Amount:	–

Relationship:	TIE-UP
Entities:	–
Joint Venture Company:	"Bridgestone Sports Taiwan Co."
Activity:	–
Amount:	NT$20000000

Bridgestone Sports Co. said Friday it has set up a **joint venture** in Taiwan with **a local concern** and **a Japanese trading house** to **produce golf clubs** to be shipped to Japan.

The joint venture, **Bridgestone Sports Taiwan Co.**, **capitalized** at **20 million new Taiwan dollars**, will start **production** in **January 1990** with production of 20,000 iron and "metal wood" clubs a month.

(5/5) Merge Structures

```
Activity:              PRODUCTION
Company:               –
Product:               "golf clubs"
Start Date:            –
```

```
Activity:              PRODUCTION
Company:               "Bridgestone Sports Taiwan Co."
Product:               –
Start Date:            DURING: January 1990
```
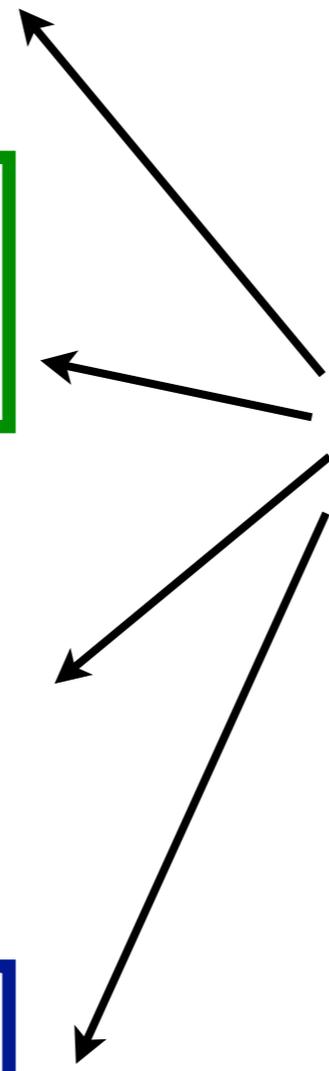
```
Relationship:          TIE-UP
Entities:              "Bridgestone Sports Co."
                       "a local concern"
                       "a Japanese trading house"

Joint Venture Company: –
Activity:              –
Amount:                –
```

```
Relationship:          TIE-UP
Entities:              –
Joint Venture Company: "Bridgestone Sports Taiwan Co."
Activity:              –
Amount:                NT$20000000
```

(5/5) Merge Structures

Activity:	PRODUCTION
Company:	–
Product:	"golf clubs"
Start Date:	–

Activity:	PRODUCTION
Company:	"Bridgestone Sports Taiwan Co."
Product:	–
Start Date:	DURING: January 1990

Relationship:	TIE-UP
Entities:	"Bridgestone Sports Co."
	"a local concern"
	"a Japanese trading house"
Joint Venture Company:	–
Activity:	–
Amount:	–

Relationship:	TIE-UP
Entities:	–
Joint Venture Company:	"Bridgestone Sports Taiwan Co."
Activity:	–
Amount:	NT$20000000

Decide identity coreference through name-matching and type compatibility; if arguments are coreferent, merge events

Activity:	PRODUCTION
Company:	"Bridgestone Sports Taiwan Co."
Product:	"iron and 'metal wood' clubs"
Start Date:	DURING: January 1990

(5/5) Merge Structures

Activity:	PRODUCTION
Company:	–
Product:	"golf clubs"
Start Date:	–

Decide identity coreference through name-matching and type compatibility; if arguments are coreferent, merge events

Activity:	PRODUCTION
Company:	"Bridgestone Sports Taiwan Co."
Product:	–
Start Date:	DURING: January 1990

Activity:	PRODUCTION
Company:	"Bridgestone Sports Taiwan Co."
Product:	"iron and 'metal wood' clubs"
Start Date:	DURING: January 1990

Relationship:	TIE-UP
Entities:	"Bridgestone Sports Co."
	"a local concern"
	"a Japanese trading house"
Joint Venture Company:	–
Activity:	–
Amount:	–

Relationship:	TIE-UP
Entities:	"Bridgestone Sports Co."
	"a local concern"
	"a Japanese trading house"
Joint Venture Company:	"Bridgestone Sports Taiwan Co."
Activity:	–
Amount:	NT$20000000

Relationship:	TIE-UP
Entities:	–
Joint Venture Company:	"Bridgestone Sports Taiwan Co."
Activity:	–
Amount:	NT$20000000

Empirical Rule-based NLP

- Originally FASTUS was just a preprocessor for a more complex system. It was too slow, they threw it out **--** deadline pressure

- Hours vs Minutes runtime on development set **--** much faster development iterations

Empirical Rule-based NLP

- Originally FASTUS was just a preprocessor for a more complex system. It was too slow, they threw it out **-- deadline pressure**

- Hours vs Minutes runtime on development set **--** much faster development iterations

January: Designed FASTUS
Jan-May: Development
May 6: First test of the FASTUS system on a blind test set of 100 terrorist reports, which had been withheld as a fair test, and we obtained a score of **8% recall and 42% precision**.
   At that point we began a fairly intensive effort to hill-climb on all 1300 development texts then available, doing periodic runs on the fair test to monitor our progress. This effort culminated in a score of **44% recall and 57% precision** in the wee hours of **June 1**, when we decided to run the official test. The rate of progress was rapid enough that even a few hours of work could be shown to have a noticeable impact on the score. Our scarcest resource was time, and our supply of it was eventually exhausted well before the point of diminishing returns.
We were thus able, **in three and a half weeks, to increase the system's F-score by 36.2 points, from 13.5 to 49.7**.

- Rule-based: poor linguistic coverage
  - But, when there's no training data, rule-based systems are still useful

- Brittle: very narrow and domain-specific
  - Efforts for broad coverage: e.g. FrameNet
  - But, general-purpose ontologies are very difficult

- Many bigger problems!
  - Entities
  - Events
- Problems are still hard! (50-90% accuracy...)

21

Entities and coreference

Noun phrase reference

Barack Obama nominated Hillary Rodham Clinton as his secretary of state. He chose her because she had foreign affairs experience.

Referring expressions reference discourse entities e.g. real-world entities

23

Noun phrase reference

Barack Obama nominated Hillary Rodham Clinton as his secretary of state. He chose her because she had foreign affairs experience.

Referring expressions reference discourse entities e.g. real-world entities

24

Noun phrase reference

http://harrypotter.wikia.com/wiki/Harry_Potter

Harry James Potter (b. 31 July, 1980) was a half-blood wizard, the only child and son of James and Lily Potter (née Evans), and one of the most famous wizards of modern times ... Lord Voldemort attempted to murder him when he was a year and three months old ...

Referring expressions reference discourse entities
e.g. real-world entities
(... or non-real-world)

25

Terminology

Harry James Potter (b. 31 July, 1980) was a half-blood wizard, the only child and son of James and Lily Potter (née Evans), and one of the most famous wizards of modern times ... Lord Voldemort attempted to murder him when he was a year and three months old ...

an **Entity** is a ~real-world object
    ("HARRY_POTTER_CONCEPT")
Mentions a.k.a. *referring expressions*
    14 NPs are underlined above (are they all referential?)
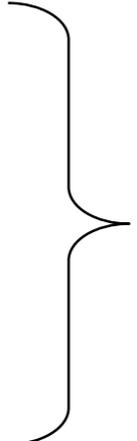**Coreference resolution**: find which mentions refer to the same entity.
I.e. cluster the mentions into **entity clusters**.

Applications: text inference, search, etc.
    - Who tried to kill Harry Potter?

Kinds of Reference

- Referring expressions
  - *John Smith*
  - *President Smith*
  - *the president*
  - *the company's new executive*

  More common in newswire, generally harder in practice

- Free variables
  - Smith saw *his pay* increase

- Bound variables
  - The dancer hurt *herself.*
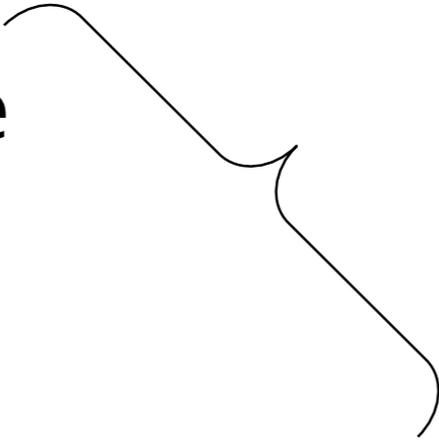
  More interesting grammatical constraints, more linguistic theory, easier in practice

  "anaphora resolution"

[slides: Chris Manning]

Syntactic vs Semantic cues

- State-of-the-art coref uses with the first three

Syntactic vs Semantic cues

- Lexical cues
  - I saw a house. The house was red.
  - I saw a house. The other house was red.

- State-of-the-art coref uses with the first three

Syntactic vs Semantic cues

- Lexical cues
  - I saw a house. The house was red.
  - I saw a house. The other house was red.
- Syntactic cues
  - John bought himself a book.
  - John bought him a book.

- State-of-the-art coref uses with the first three

Syntactic vs Semantic cues

- Lexical cues
  - I saw a house. The house was red.
  - I saw a house. The other house was red.
- Syntactic cues
  - John bought himself a book.
  - John bought him a book.
- Shallow semantic cues
  - John saw Mary. She was eating salad.
  - John saw Mary. He was eating salad.

- State-of-the-art coref uses with the first three

Syntactic vs Semantic cues

- Lexical cues
  - I saw a house. The house was red.
  - I saw a house. The other house was red.
- Syntactic cues
  - John bought himself a book.
  - John bought him a book.
- Shallow semantic cues
  - John saw Mary. She was eating salad.
  - John saw Mary. He was eating salad.
- Deeper semantics (world knowledge)
  - The city council denied the demonstrators a permit because they feared violence.
  - The city council denied the demonstrators a permit because they advocated violence.

- State-of-the-art coref uses with the first three

Features for Pronominal Anaphora Resolution

- Preferences:
  - Recency: More recently mentioned entities are more likely to be referred to
    - John went to a movie. Jack went as well. He was not busy.
  - Grammatical Role: Entities in the subject position is more likely to be referred to than entities in the object position
    - John went to a movie with Jack. He was not busy.
  - Parallelism:
   - John went with Jack to a movie. Joe went with him to a bar.

[slides: Chris Manning]

Features for Pronominal Anaphora Resolution

- Preferences:
  - Verb Semantics: Certain verbs seem to bias whether the subsequent pronouns should be referring to their subjects or objects
    - John telephoned Bill. He lost the laptop.
    - John criticized Bill. He lost the laptop.
  - Selectional Restrictions: Restrictions because of semantics
    - John parked his car in the garage after driving it around for hours.

- Encode all these and maybe more as features

[slides: Chris Manning]

Antecedent selection model

[NULL]

Hary Potter was a wizard. Lord Voldemort attempted to murder him.
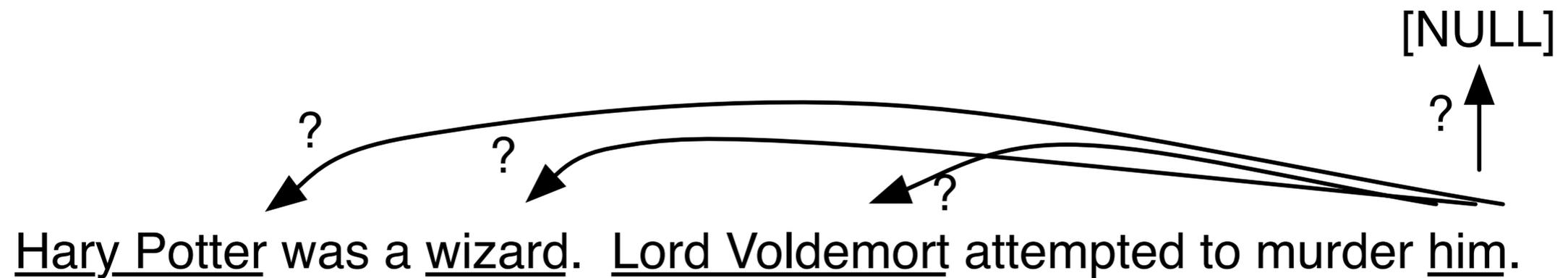
- View as antecedent selection problem: which previous mention do I corefer with?
  - Makes most sense for pronouns, though can use model for all expressions

Events and time

Thursday, December 1, 16

Temporal relations

- Goal: extract events from text and have on timeline -- or at least a partial order

- Example: TimeBank/TempEval annotations
- Temporal relations
  - Hold between events and/or temporal expressions
  - "Before <u>1993</u>, she <u>attended</u>..."
    - => *BEFORE(attend, 1993-01-01)*
  - "I <u>got</u> in a car and <u>drove</u>"
    - => *BEFORE(car, drove)*
- Aspectual/subordinate/factive relations
  - Hold between events
  - *"I <u>doubt</u> I <u>left</u> them there"*

33

TimeBank-Dense

There were four or five people inside, and they just **started firing**

Ms. Sanders was **hit** several times and was **pronounced dead** at the scene.

The other customers **fled**, and the police **said** it did not **appear** that anyone else was **injured.**

Nate Chambers et al.;
TimeBank-Dense data and CAEVO system

34

	In Washington	
tl2:head[includes] tl3:head[identity]	today	[t53 functionInDocument=NONE temporalFunction=true anchorTimeID=t0 type=DATE value=1998-01-14]
	, the Federal Aviation Administration	
tl2:relto[includes] tl5:head[after]	released	[e1 class=OCCURRENCE polarity=POS tense=PAST aspect=NONE pos=VERB]
	air traffic control tapes from the night the TWA Flight eight hundred	
tl5:relto[after] tl6:relto[is_included] tl11:head[identity]	went	[e2 class=OCCURRENCE polarity=POS tense=PAST aspect=NONE pos=VERB]
	down. There's nothing new on why the plane	
tl6:head[is_included] tl10:relto[identity]	exploded	[e3 class=OCCURRENCE polarity=POS tense=PAST aspect=NONE pos=VERB]
	, but you cannot miss the moment. ABC's Lisa Stark has more. There was no hint of trouble in the last	
tl8:relto[after] tl20:head[begins]	conversation	[e10 class=OCCURRENCE polarity=POS tense=NONE aspect=NONE pos=NOUN]
	between controllers and TWA pilot Steven Snyder TWA eight hundred climb	
tl7:relto[after]	maintain	[e13 class=OCCURRENCE polarity=POS tense=PRESENT aspect=NONE pos=VERB]
	one five thousand. TWA's eight hundred heavy climb and	
tl7:head[after]	maintain	[e15 class=OCCURRENCE polarity=POS tense=PRESENT aspect=NONE pos=VERB]
	one five thousand leavin- three thousand. But	
tl20:relto[begins] tl21:relto[ends]	a minute and a half	[t58 type=DURATION functionInDocument=NONE value=PT1M30S temporalFunction=false]
	later, a pilot from a nearby flight	
tl8:head[after] tl12:relto[after] tl21:head[ends]	calls	[e18 class=REPORTING polarity=POS tense=PRESENT aspect=NONE pos=VERB]
	in. Ah, we just	
	saw	[e19 class=PERCEPTION polarity=POS tense=PAST aspect=NONE pos=VERB]
	an	
tl9:head[is_included] tl10:head[identity] tl17:relto[identity]	explosion	[e20 class=OCCURRENCE polarity=POS tense=NONE aspect=NONE pos=NOUN]
	up ahead of us here about sixteen thousand feet or something like that. It just	

35

Temporal expressions

- ## Absolute
  - *June 11, 1989*; the *Summer of 2002*
- ## Relative
  - *early in the morning, Monday, in recent days, few days ago, two weeks from next Tuesday, next September, the current month, last year, a decade ago.*
- ## Durations
  - *three months, two years*

- ## Semantic encoding standard

36

Temporal Relations

- BEFORE, AFTER, DURING
  - R(evt, time)
  - R(evt, evt)
- Logical implications: e.g. transitivity

Relation	Illustration	Interpretation
$X < Y$ $Y > X$		X takes place before Y
$X \, \mathrm{m} \, Y$ $Y \, \mathrm{mi} \, X$		X meets Y (*i* stands for *inverse*)
$X \, \mathrm{o} \, Y$ $Y \, \mathrm{oi} \, X$		X overlaps with Y
$X \, \mathrm{s} \, Y$ $Y \, \mathrm{si} \, X$		X starts Y
$X \, \mathrm{d} \, Y$ $Y \, \mathrm{di} \, X$		X during Y
$X \, \mathrm{f} \, Y$ $Y \, \mathrm{fi} \, X$		X finishes Y
$X = Y$		X is equal to Y

This is a simplification of Allen's interval algebra

37

Subordination relations

- Event-Event

- (counter-) <u>Factive</u>: introduce entailments/presupp of X
  - John forgot that X
  - John regrets that X
- (negative) <u>Evidential</u>: reporting, perception
  - John said that X
  - John denied that X
- <u>Modal</u>: reference to possible world
  - Mary wanted John to X
  - John promised to X