Lecture 18
Information Extraction

CS 6320

Overview

- Information extraction – turns unstructured information buried in texts into structured data
- Extract proper nouns – “named entity recognition”
- Reference resolution –
  - named entity mentions
  - Pronoun references
- Relation Detection and classification
- Event detection and classification
- Temporal analysis
- Template filling

Sample text

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

- **Identify named entities**
- **Identify relations**

Template Filling

Example template for “airfare raise”

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

@ 2020 Dan I Moldovan, Human Language Technology Research Institute, UTD

Some Named Entity Types

Type	Tag	Sample Categories
People	PER	Individuals, fictional characters, small groups
Organization	ORG	Companies, agencies, political parties, religious groups, sports teams
Location	LOC	Physical extents, mountains, lakes, seas
Geo-Political Entity	GPE	Countries, states, provinces, counties
Facility	FAC	Bridges, buildings, airports
Vehicles	VEH	Planes, trains, and automobiles

Examples of Named Entity Types

Type	Example
People	*Turing* is often considered to be the father of modern computer science.
Organization	The *IPCC* said it is likely that future tropical cyclones will become more intense.
Location	The *Mt. Sanitas* loop hike begins at the base of *Sunshine Canyon*.
Geo-Political Entity	*Palo Alto* is looking at raising the fees for parking in the University Avenue district.
Facility	Drivers were advised to consider either the *Tappan Zee Bridge* or the *Lincoln Tunnel*.
Vehicles	The updated *Mini Cooper* retains its charm and agility.

Categorical Ambiguities

Name	Possible Categories
Washington	Person, Location, Political Entity, Organization, Facility
Downing St.	Location, Organization
IRA	Person, Organization, Monetary Instrument
Louis Vuitton	Person, Organization, Commercial Product

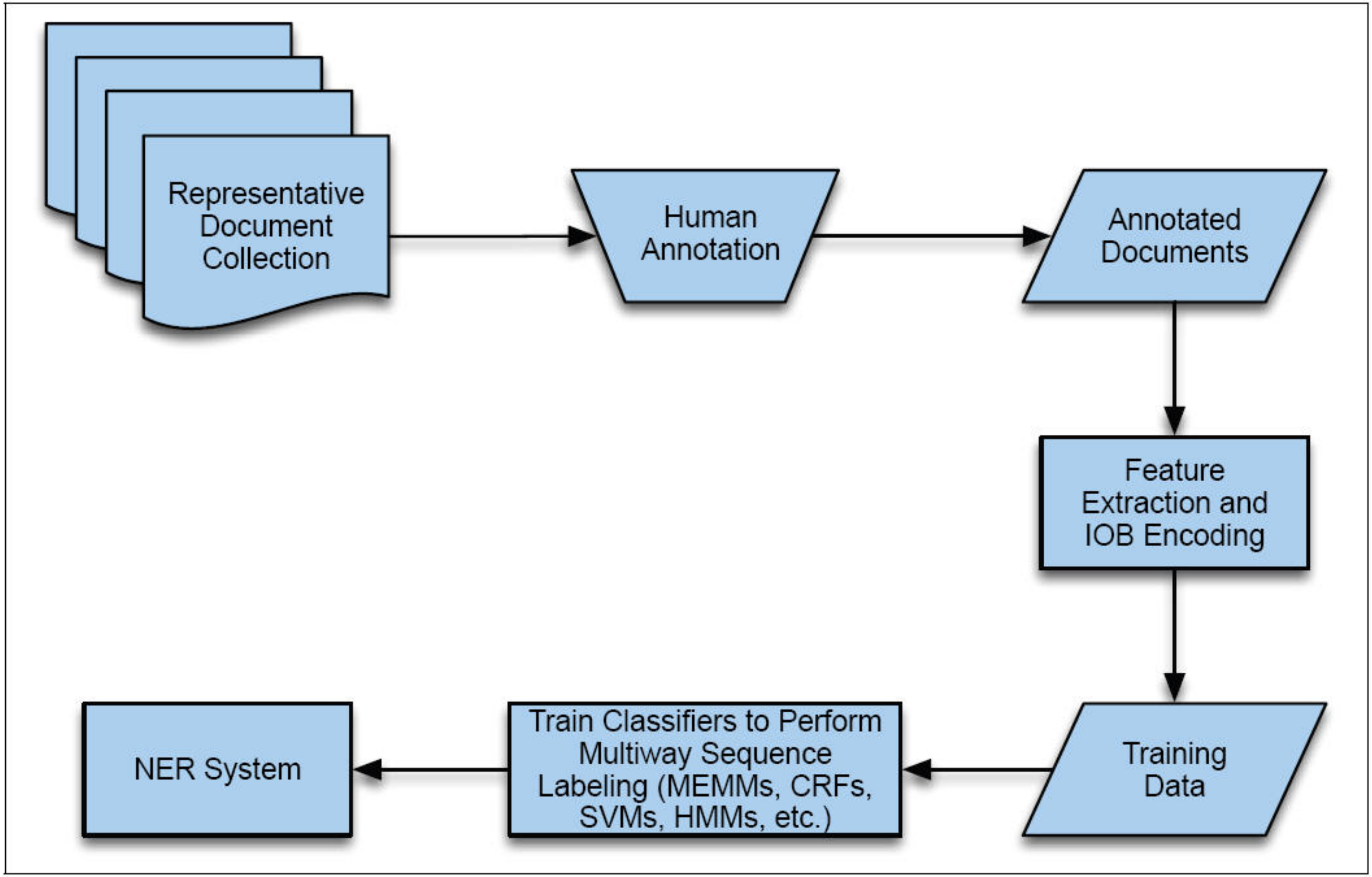
@ 2020 Dan I Moldovan, Human Language Technology Research Institute, UTD

Categorical Ambiguity

[$_{PERS}$ Washington] was born into slavery on the farm of James Burroughs.
[$_{ORG}$ Washington] went up 2 games to 1 in the four-game series.
Blair arrived in [$_{LOC}$ Washington] for what may well be his last state visit.
In June, [$_{GPE}$ Washington] passed a primary seatbelt law.
The [$_{FAC}$ Washington] had proved to be a leaky ship, every passage I made...

@ 2020 Dan I Moldovan, Human Language Technology Research Institute, UTD

Statistical Seq. Labeling



Features used in Training NER

Gazetteers – lists of place names

- www.geonames.com
- www.census.gov

Feature	Explanation
Lexical items	The token to be labeled
Stemmed lexical items	Stemmed version of the target token
Shape	The orthographic pattern of the target word
Character affixes	Character-level affixes of the target and surrounding words
Part of speech	Part of speech of the word
Syntactic chunk labels	Base-phrase chunk label
Gazetteer or name list	Presence of the word in one or more named entity lists
Predictive token(s)	Presence of predictive words in surrounding text
Bag of words/Bag of N-grams	Words and/or *N*-grams occurring in the surrounding context

@ 2020 Dan I Moldovan, Human Language Technology Research Institute, UTD

Selected Shape Features

Shape	Example
Lower	cummings
Capitalized	Washington
All caps	IRA
Mixed case	eBay
Capitalized character with period	H.
Ends in digit	A9
Contains hyphen	H-P

Boundary Detection

Words	IOB Label	IO Label
American	B-ORG	I-ORG
Airlines	I-ORG	I-ORG
,	O	O
a	O	O
unit	O	O
of	O	O
AMR	B-ORG	I-ORG
Corp.	I-ORG	I-ORG
,	O	O
immediately	O	O
matched	O	O
the	O	O
move	O	O
,	O	O
spokesman	O	O
Tim	B-PER	I-PER
Wagner	I-PER	I-PER
said	O	O
.	O	O

Figure 17.4 Named entity tagging as a sequence model, showing IOB and IO encodings.

BIO Encoding for NER

Word	POS	Chunk	Short shape	Label
American	NNP	B-NP	Xx	B-ORG
Airlines	NNPS	I-NP	Xx	I-ORG
,	,	O	,	O
a	DT	B-NP	x	O
unit	NN	I-NP	x	O
of	IN	B-PP	x	O
AMR	NNP	B-NP	X	B-ORG
Corp.	NNP	I-NP	Xx.	I-ORG
,	,	O	,	O
immediately	RB	B-ADVP	x	O
matched	VBD	B-VP	x	O
the	DT	B-NP	x	O
move	NN	I-NP	x	O
,	,	O	,	O
spokesman	NN	B-NP	x	O
Tim	NNP	I-NP	Xx	B-PER
Wagner	NNP	I-NP	Xx	I-PER
said	VBD	B-VP	x	O
.	,	O	.	O

Figure 17.6 Word-by-word feature encoding for NER.

@ 2020 Dan I Moldovan, Human Language Technology Research Institute, UTD

NER Classifier with Input Features

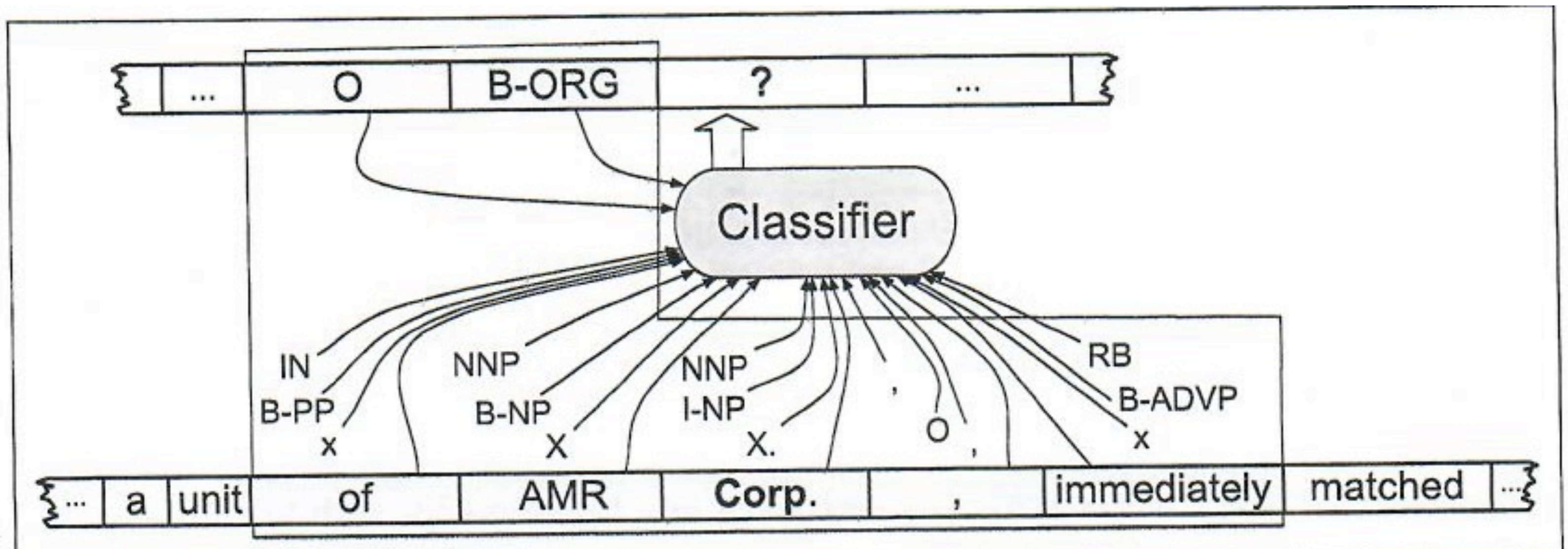


Figure 17.7 Named entity recognition as sequence labeling. The features available to the classifier during training and classification are those in the boxed area.

A Neural Approach to NER

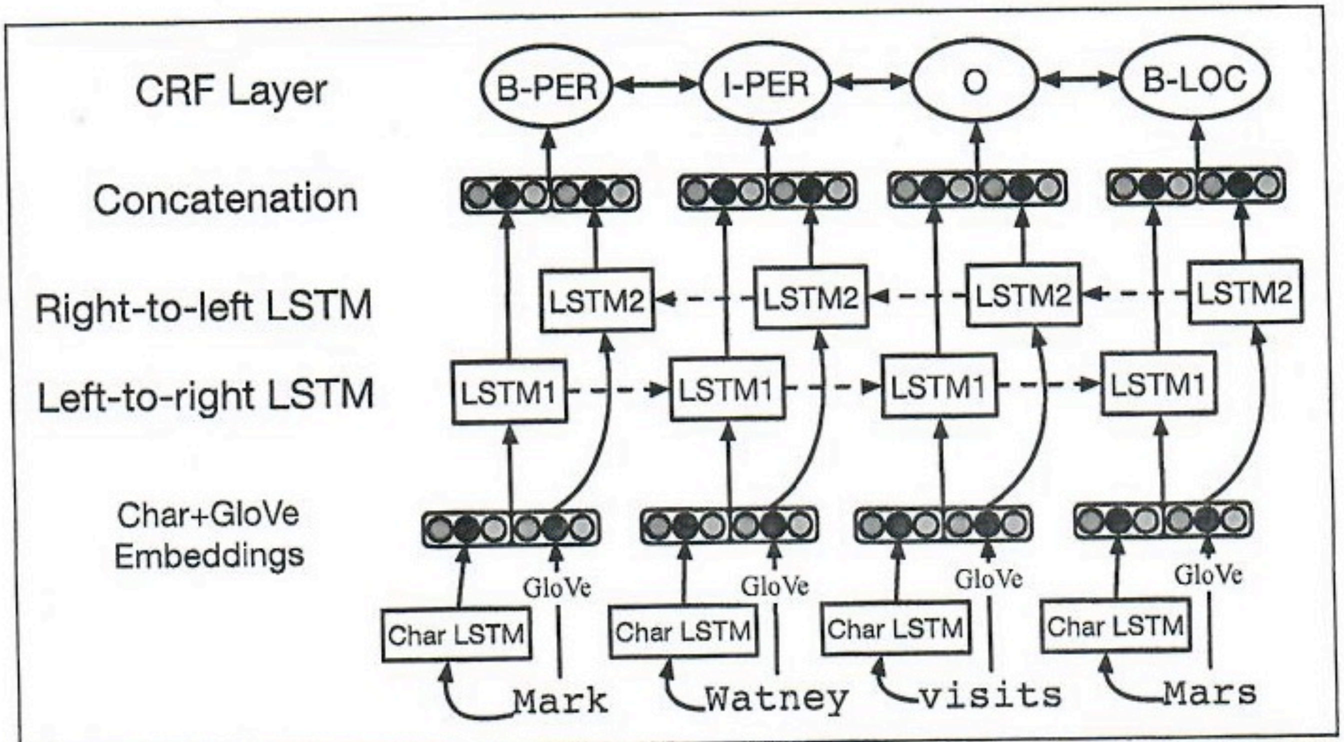


Figure 17.8 Putting it all together: character embeddings and words together a bi-LSTM sequence model. After (Lample et al., 2016)

Evaluation of N E R Systems

- Recall terms from Information retrieval
  - Recall = #correctly labeled / total # that should be labeled
  - Precision = # correctly labeled / total # labeled
- F- measure where β weights preferences
  - β=1 balanced
  - β>1 favors recall
  - β<1 favors precision

$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

NER Performance revisited

- NER performance revisited
  - Recall, Precision, F
  - High performance systems
    - » F ~ .92 for PERSONS and LOCATIONS and ~.84 for ORG
- Practical Rule-based NER
  - Make several passes on text, allowing the results of one pass to influence the next

@ 2020 Dan I Moldovan, Human Language Technology Research Institute, UTD

Relation Detection and classification

- **Consider Sample text:**
  - **Citing high fuel prices, [ORG United Airlines] said [TIME Friday] it has increased fares by [MONEY $6] per round trip on flights to some cities also served by lower-cost carriers. [ORG American Airlines], a unit of [ORG AMR Corp.], immediately matched the move, spokesman [PERSON Tim Wagner] said. [ORG United Airlines] an unit of [ORG UAL Corp.], said the increase took effect [TIME Thursday] and applies to most routes where it competes against discount carriers, such as [LOC Chicago] to [LOC Dallas] and [LOC Denver] to [LOC San Francisco].**
- **After identifying named entities what else can we extract?**
- **Relations**

@ 2020 Dan I Moldovan, Human Language Technology Research Institute, UTD

Example semantic relations

Relations		Examples	Types
Affiliations			
	Personal	*married to, mother of*	PER $\longrightarrow$ PER
	Organizational	*spokesman for, president of*	PER $\longrightarrow$ ORG
	Artifactual	*owns, invented, produces*	(PER $\mid$ ORG) $\longrightarrow$ ART
Geospatial			
	Proximity	*near, on outskirts*	LOC $\longrightarrow$ LOC
	Directional	*southeast of*	LOC $\longrightarrow$ LOC
Part-Of			
	Organizational	*a unit of, parent of*	ORG $\longrightarrow$ ORG
	Political	*annexed, acquired*	GPE $\longrightarrow$ GPE

@ 2020 Dan I Moldovan, Human Language Technology Research Institute, UTD

Sub-Relations for Association

1. Communication – lexical constraint
   a. COMMUNICATE
      i. WRITE_TO
         1. WRITTEN_COMMUNICATION
         2. ELECTRONIC_COMMUNICATION
      ii. TELEPHONE_TO
      iii. SPEAK_TO
      iv. COMMANDS_OR_CONTROLS(person)
      v. OTHER_COMM
      vi. RECRUITED
2. Meeting – lexical constraint
   a. MEET
3. Joint work – lexical constraint
   a. WORK_WITH
   b. SHARE_TASK_WITH
   c. IS_COWORKER_OF
4. . Economic/trade – usually lexical constraint
   a. SEND_TO
   b. RECEIVE_FROM
   c. SELL_TO
   d. PURCHASE_FROM
   e. TRANSFER
   f. IS_EMPLOYER_OF
5. Teacher/pupil – lexical constraint
   a. TEACH
6. Directly described association – lexical constraint
   a. IS_AFFILIATED_TO
7. Common group membership
   a. MEMBER_OF & MEMBER_OF – no constraint
   b. MEMBER_OF & COMMANDS_OR_CONTROLS(org)
   c. KINSHIP
8. Presence at shared location – no constraint
   a. SHARE_LOCATION
9. Common origin – no constraint
   a. SHARE_ORIGIN
10. Frequent travel to shared location – no constraint
    a. TRAVEL_TO

Example Extraction

Domain	$\mathcal{D} = \{a,b,c,d,e,f,g,h,i\}$
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	$Serves = \{\langle a,f\rangle, \langle a,g\rangle, \langle a,h\rangle, \langle a,i\rangle\}$

Supervised Learning Approaches to Relation Analysis

Algorithm two step process

1. Identify whether pair of named entities are related
2. Classifier is trained to label relations

function FINDRELATIONS(*words*) **returns** *relations*

  relations ← *nil*
  entities ← FINDENTITIES(*words*)
  forall entity pairs $\langle e1, e2 \rangle$ **in** *entities* **do**
    if RELATED?($e1, e2$)
      relations ← *relations*+CLASSIFYRELATION($e1, e2$)

Factors used in Classifying

Features of the named entities

- Named entity types of the two arguments
- Concatenation of the two entity types
- Headwords of the arguments
- Bag-of-words from each of the arguments

- Words in text
  - Bag-of-words and Bag-of-bigrams
  - Stemmed versions of the same
  - Distance between named entities (words / named entities)
- Syntactic structure
  - Parse related structures

Sample features Extracted

Entity-based features	
Entity$_1$ type	ORG
Entity$_1$ head	*airlines*
Entity$_2$ type	PERS
Entity$_2$ head	*Wagner*
Concatenated types	ORGPERS
Word-based features	
Between-entity bag of words	{ *a, unit, of, AMR, Inc., immediately, matched, the, move, spokesman* }
Word(s) before Entity$_1$	NONE
Word(s) after Entity$_2$	*said*
Syntactic features	
Constituent path	$NP \uparrow NP \uparrow S \uparrow S \downarrow NP$
Base syntactic chunk path	$NP \rightarrow NP \rightarrow PP \rightarrow NP \rightarrow VP \rightarrow NP \rightarrow NP$
Typed-dependency path	$Airlines \leftarrow_{subj} matched \leftarrow_{comp} said \rightarrow_{subj} Wagner$

Semantic Role Labeling

- SRL – is the task of finding semantic roles for each predicate.

- FrameNet

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

- PropBank

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

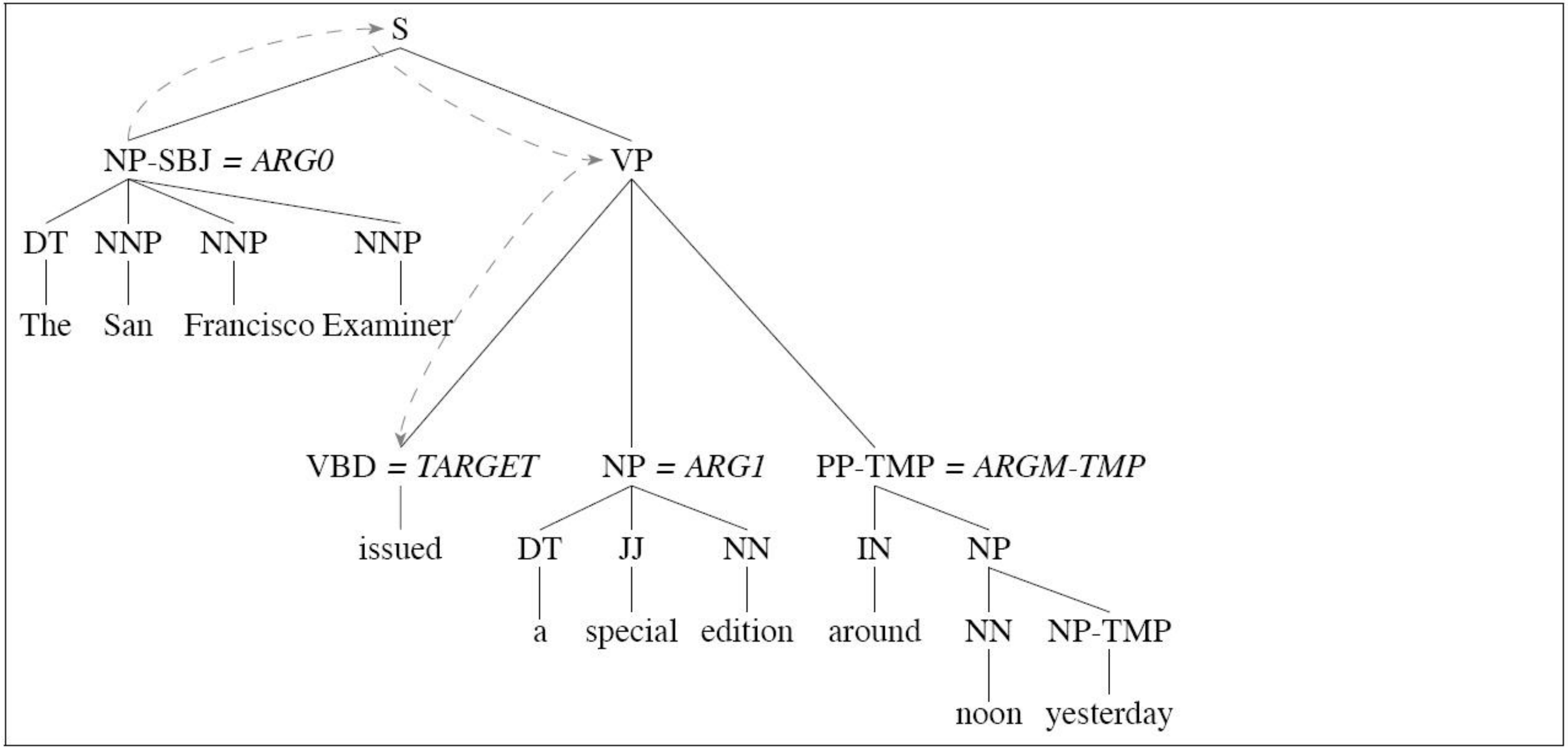
Semantic Role Labeling Algorithm

- **Need syntactic parser.**
- **Extract features.**
- **Classify node.**

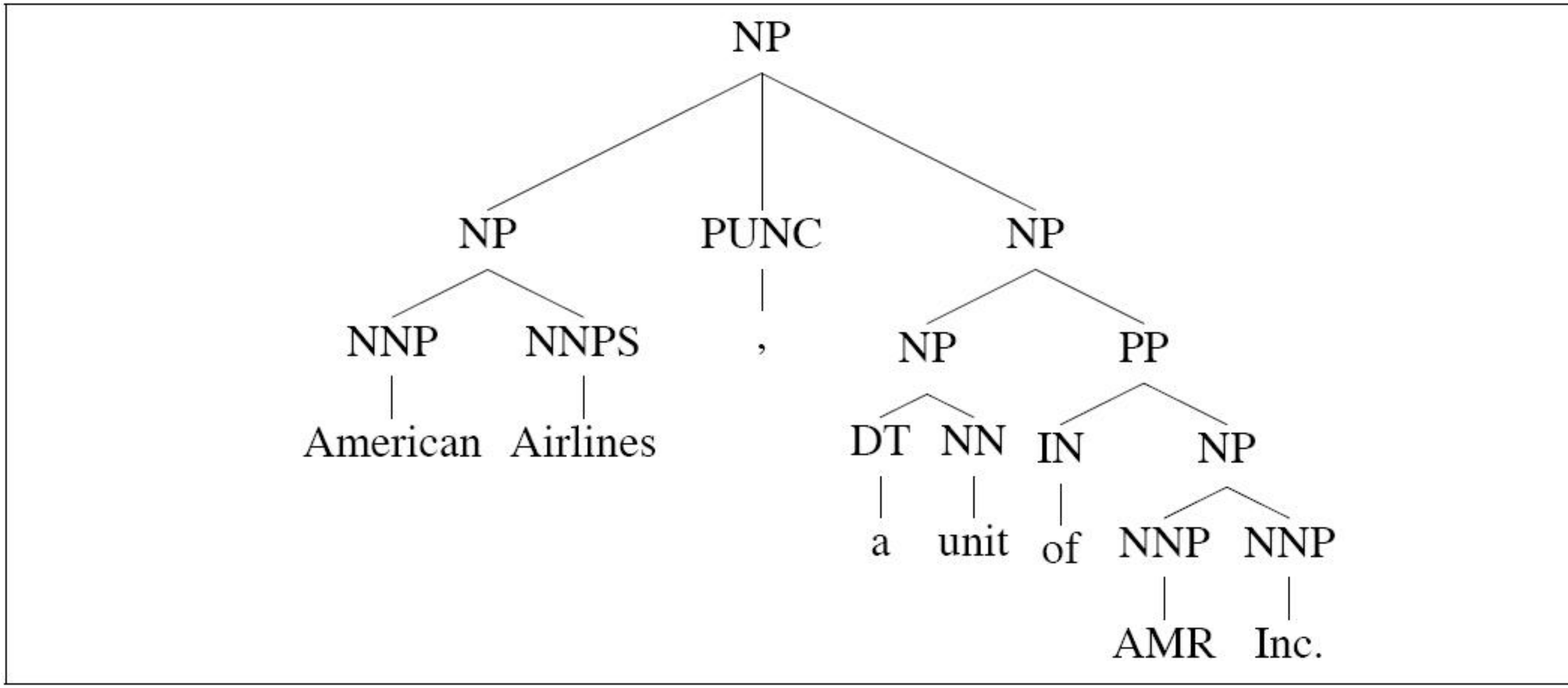
function SEMANTICROLELABEL(*words*) **returns** labeled tree

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

Semantic Role Labeling



a-part-of relation



Semantic Role Labeling-Features

- Governing Predicate.
- Phase type of constituent.
- Headword of constituent.
- Path in the parse tree from constituent to the predicate.
- Voice of the clause containing constituent.
- Binary respect to predicate (before or after).
- Sub categorization of predicate.

Bootstrapping Example "Has a hub at"

Consider the pattern

/ * has a hub at * /

Google search

Milwaukee-based Midwest has a hub at KCI

Delta has a hub at LaGuardia

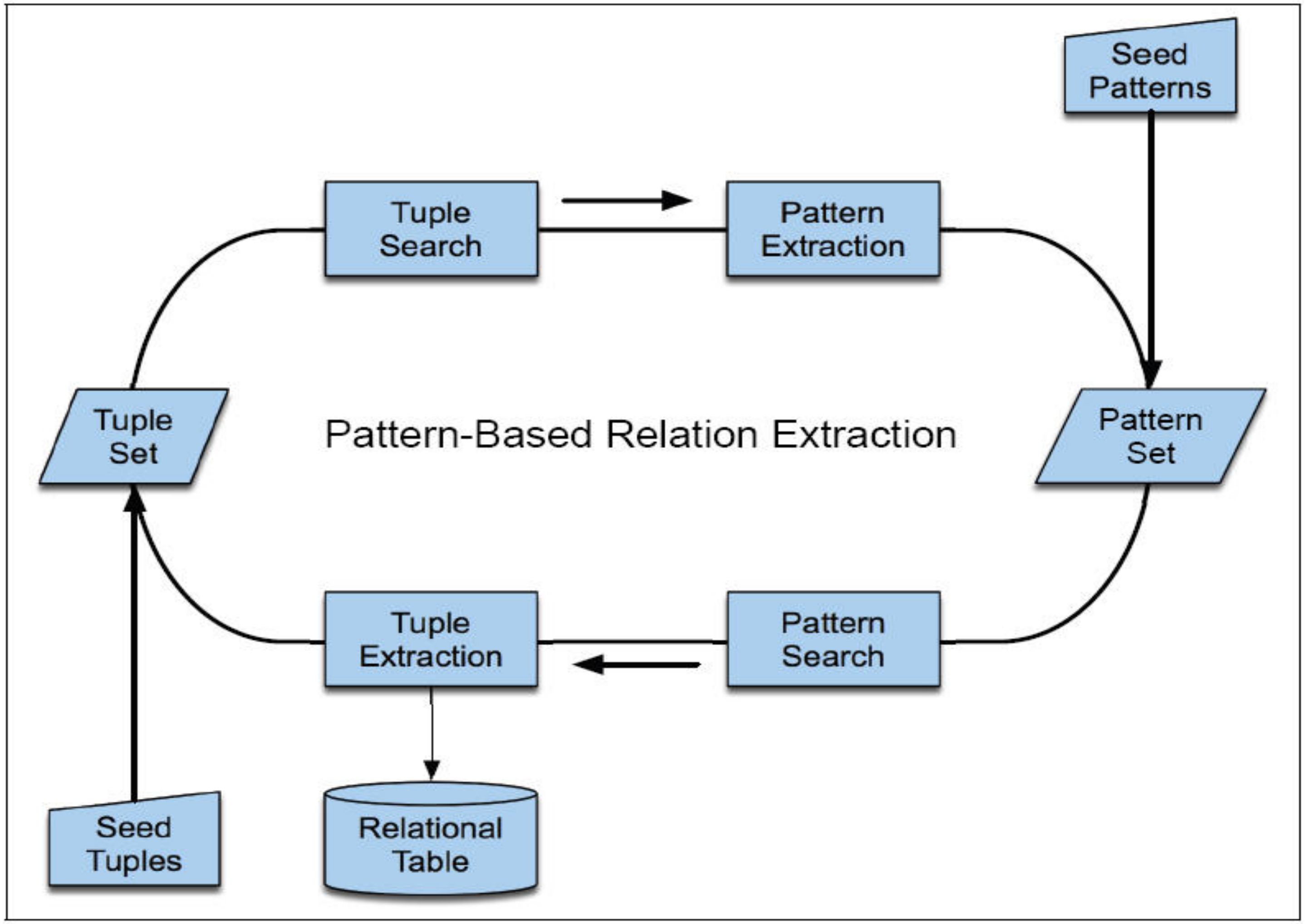
…

Two ways to fail

1. False positive: e.g. a star topology has a hub at its center
2. False negative? Just miss

No frill rival easyJet, which has established a hub at Liverpool

Bootstrapping Relation Extraction



Using Features to restrict patterns

Budget airline Ryanair, which uses Charleroi as a hub, scrapped all weekend flights ..

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

A spokesman at Charleroi, a main hub for Ryanair, estimated that …

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

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

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

Temporal and Durational Expressions

Absolute temporal expressions

Relative temporal expressions

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

@ 2020 Dan I Moldovan, Human Language Technology Research Institute, UTD

Temporal lexical triggers

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 Normalization

iSO 8601 - standard for encoding temporal values

YYYY-MM-DD

Sample ISO Patterns

Unit	Pattern	Sample Value
Fully specified dates	YYYY-MM-DD	1991-09-28
Weeks	YYYY-nnW	2007-27W
Weekends	PnWE	P1WE
24-hour clock times	HH:MM:SS	11:13:45
Dates and times	YYYY-MM-DDTHH:MM:SS	1991-09-28T11:00:00
Financial quarters	Qn	1999-Q3

Event Detection and Analysis

Event Detection and classification

[$_{EVENT}$ 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.

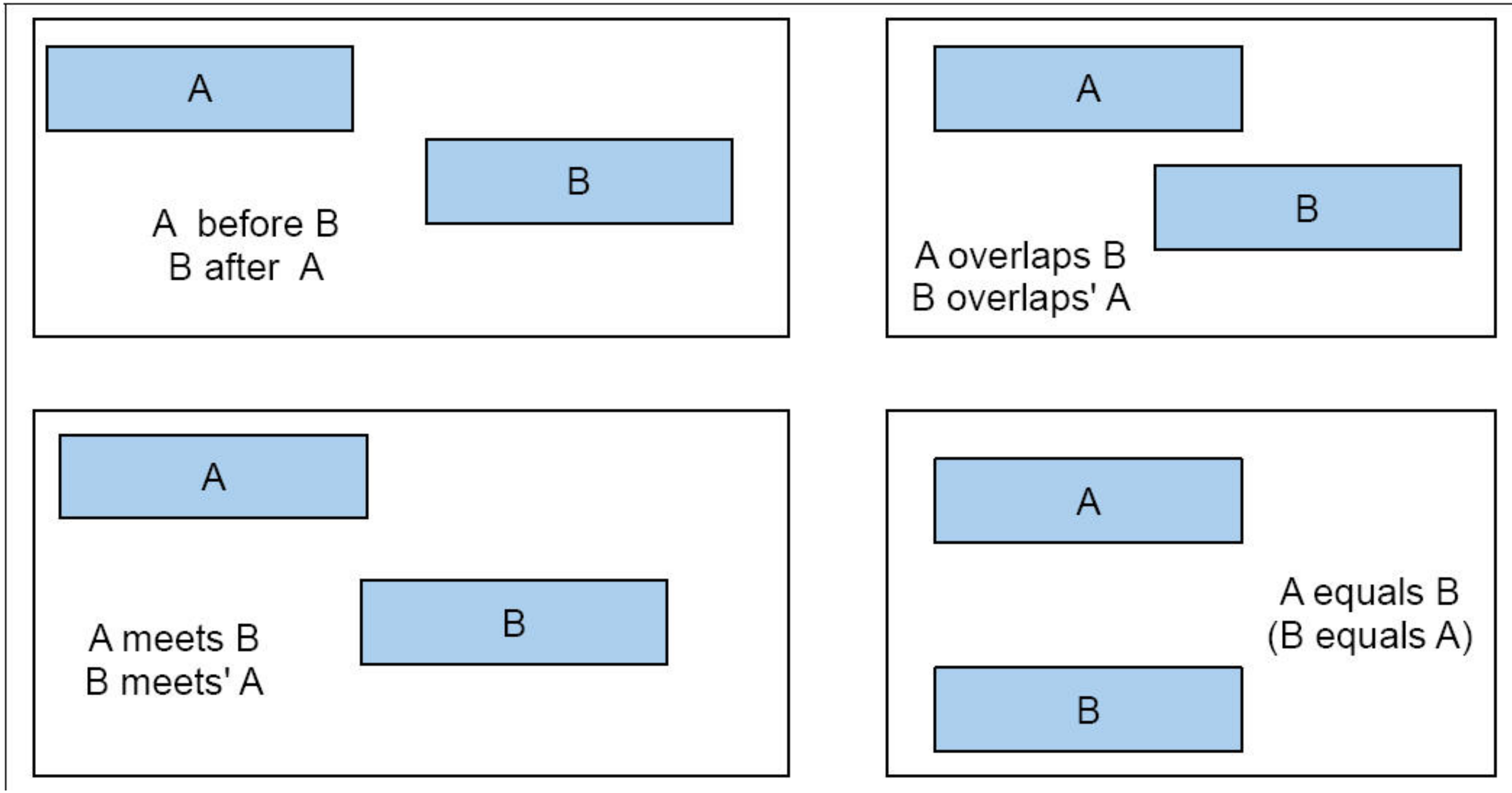
Features for Event Detection

Features used in rule-based and statistical techniques

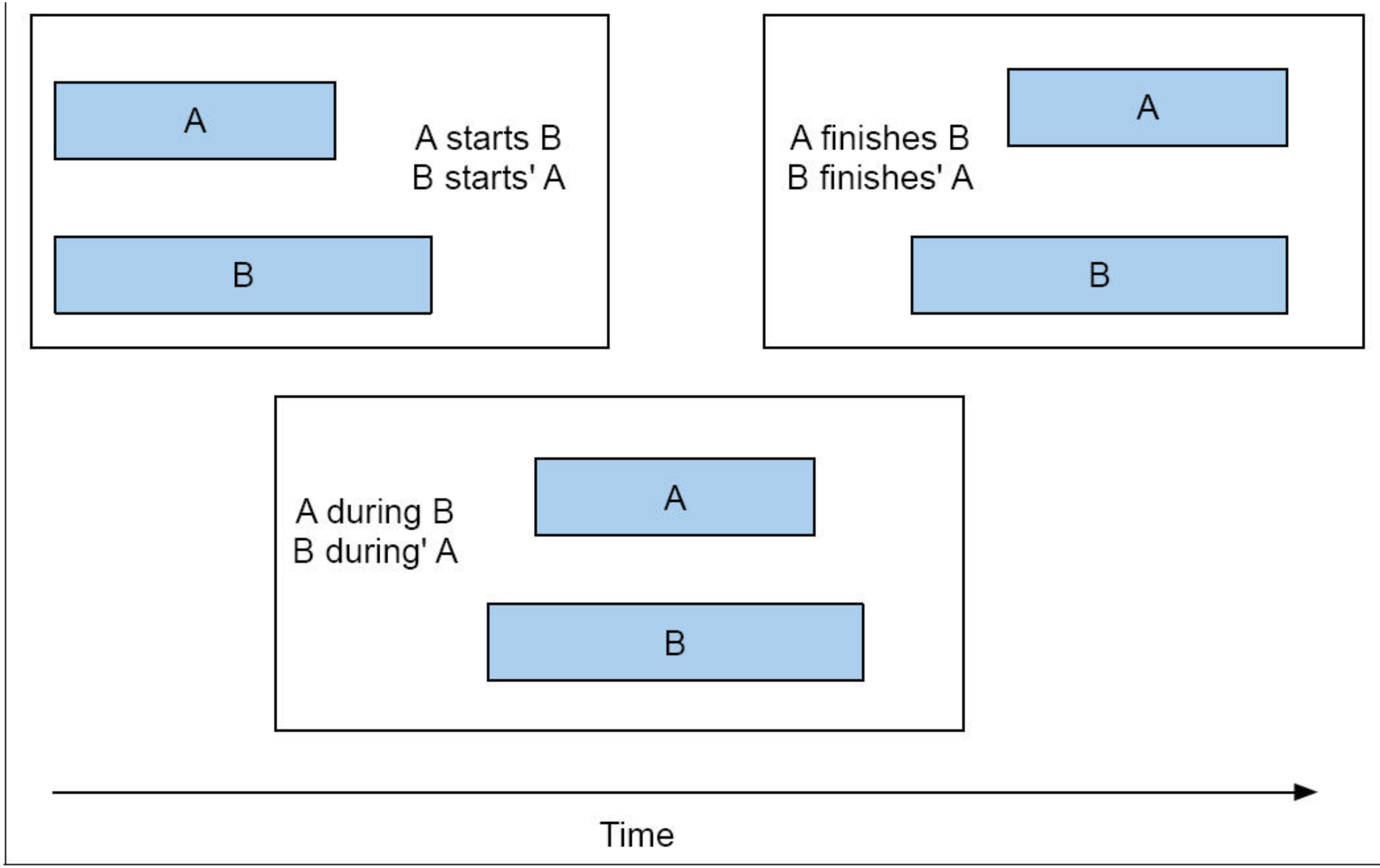
Feature	Explanation
Character affixes	Character-level prefixes and suffixes of target word
Nominalization suffix	Character level suffixes for nominalizations (e.g., *-tion*)
Part of speech	Part of speech of the target word
Light verb	Binary feature indicating that the target is governed by a light verb
Subject syntactic category	Syntactic category of the subject of the sentence
Morphological stem	Stemmed version of the target word
Verb root	Root form of the verb basis for a nominalization
WordNet hypernyms	Hypernym set for the target

@ 2020 Dan I Moldovan, Human Language Technology Research Institute, UTD

Allen’s 13 temporal Relations



continued



A
B
A starts B
B starts' A
A
B
A finishes B
B finishes' A
A during B
B during' A
A
B
Time

Example from Timebank Corpus

```
<TIMEX3 tid="t57" type="DATE" value="1989-10-26"  functionInDocument="CREATION_TIME">
10/26/89  </TIMEX3>

Delta Air Lines earnings <EVENT eid="e1" class="OCCURRENCE"> soared </EVENT> 33% to a
record in  <TIMEX3 tid="t58" type="DATE" value="1989-Q1" anchorTimeID="t57"> the
fiscal first quarter </TIMEX3>, <EVENT eid="e3"  class="OCCURRENCE">bucking</EVENT>
the industry trend toward <EVENT eid="e4" class="OCCURRENCE">declining</EVENT>
profits.
```