Introduction to NLP

232.

Information Extraction

Information Extraction

- Usually from unstructured (or semi-structured) data
- Examples
  - News stories
  - Scientific papers
  - Resumes
- Entities
  - Who did what, when, where, why
- Build knowledge base (KBP Task)

Named Entities

- Types:
  - People
  - Locations
  - Organizations
    - Teams, Newspapers, Companies
  - Geo-political entities
- Ambiguity:
  - London can be a person, city, country (by metonymy) etc.
- Useful for interfaces to databases, question answering, etc.

Named Entities

Type	**Tag**	**Sample Categories**	**Example sentences**
People	PER	people, characters	**Turing** is a giant of computer science.
Organization	ORG	companies, sports teams	The **IPCC** warned about the cyclone.
Location	LOC	regions, mountains, seas	The **Mt. Sanitas** loop is in **Sunshine Canyon**.
Geo-Political Entity	GPE	countries, states, provinces	**Palo Alto** is raising the fees for parking.
Facility	FAC	bridges, buildings, airports	Consider the **Tappan Zee Bridge**.
Vehicles	VEH	planes, trains, automobiles	It was a classic **Ford Falcon**.

Figure 21.1 A list of generic named entity types with the kinds of entities they refer to.

Named Entity Recognition (NER)

- Segmentation
  - Which words belong to a named entity?
  - Brazilian football legend Pele's condition has improved, according to a Thursday evening statement from a Sao Paulo hospital.
- Classification
  - What type of named entity is it?
  - Use gazetteers, spelling, adjacent words, etc.
  - Brazilian football legend [$_{\text{PERSON}}$ Pele]'s condition has improved, according to a [$_{\text{TIME}}$ Thursday evening] statement from a [$_{\text{LOCATION}}$ Sao Paulo] hospital.

Times and Events

- Times
  - Absolute expressions
  - Relative expressions (e.g., “last night”)
- Events
  - E.g., a plane went past the end of the runway

NER, Time, and Event extraction

- Brazilian football legend [$_{\text{PERSON}}$ Pele]'s condition has improved, according to a [$_{\text{TIME}}$ Thursday evening] statement from a [$_{\text{LOCATION}}$ Sao Paulo] hospital.
- There had been earlier concerns about Pele's health after [$_{\text{ORG}}$ Albert Einstein Hospital] issued a release that said his condition was "unstable.“
- [$_{\text{TIME}}$ Thursday night]'s release said [$_{\text{EVENT}}$ Pele was relocated] to the intensive care unit because a kidney dialysis machine he needed was in ICU.

Event Extraction

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**].

Event Extraction

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

Named Entity Recognition (NER)

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

Figure 21.2 Common categorical ambiguities associated with various proper names.

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

Figure 21.3 Examples of type ambiguities in the use of the name *Washington*.

Sample Input for NER

```
( (S
    (NP-SBJ-1
      (NP (NNP Rudolph) (NNP Agnew) )
      (, ,)
      (UCP
        (ADJP
          (NP (CD 55) (NNS years) )
          (JJ old) )
        (CC and)
        (NP
          (NP (JJ former) (NN chairman) )
          (PP (IN of)
            (NP (NNP Consolidated) (NNP Gold) (NNP Fields) (NNP PLC) ))))
      (, ,) )
    (VP (VBD was)
      (VP (VBN named)
        (S
          (NP-SBJ (-NONE- *-1) )
          (NP-PRD
            (NP (DT a) (JJ nonexecutive) (NN director) )
            (PP (IN of)
              (NP (DT this) (JJ British) (JJ industrial) (NN conglomerate) ))))))
    (. .) ))
```

Sample Output for NER (IOB format)

file_id	sent_id	word_id	iob_inner	pos	word
0002	**1**	**0**	**B-PER**	**NNP**	**Rudolph**
0002	**1**	**1**	**I-PER**	**NNP**	**Agnew**
0002	1	2	O	COMMA	COMMA
0002	1	3	B-NP	CD	55
0002	1	4	I-NP	NNS	years
0002	1	5	B-ADJP	JJ	old
0002	1	6	O	CC	and
0002	1	7	B-NP	JJ	former
0002	1	8	I-NP	NN	chairman
0002	1	9	B-PP	IN	of
0002	**1**	**10**	**B-ORG**	**NNP**	**Consolidated**
0002	**1**	**11**	**I-ORG**	**NNP**	**Gold**
0002	**1**	**12**	**I-ORG**	**NNP**	**Fields**
0002	**1**	**13**	**I-ORG**	**NNP**	**PLC**
0002	1	14	O	COMMA	COMMA
0002	1	15	B-VP	VBD	was
0002	1	16	I-VP	VBN	named
0002	1	17	B-NP	DT	a
0002	1	18	I-NP	JJ	nonexecutive
0002	1	19	I-NP	NN	director
0002	1	20	B-PP	IN	of
0002	1	21	B-NP	DT	this
0002	1	22	I-NP	JJ	British
0002	1	23	I-NP	JJ	industrial
0002	1	24	I-NP	NN	conglomerate
0002	1	25	O	.	.

NER Demos

- http://nlp.stanford.edu:8080/ner/
- http://cogcomp.org/page/demo_view/ner
- http://demo.allennlp.org/named-entity-recognition

NER Extraction Features

identity of w_i
identity of neighboring words
part of speech of w_i
part of speech of neighboring words
base-phrase syntactic chunk label of w_i and neighboring words
presence of w_i in a **gazetteer**
w_i contains a particular prefix (from all prefixes of length ≤ 4)
w_i contains a particular suffix (from all suffixes of length ≤ 4)
w_i is all upper case
word shape of w_i
word shape of neighboring words
short word shape of w_i
short word shape of neighboring words
presence of hyphen

Figure 21.5 Features commonly used in training named entity recognition systems.

NER Extraction Features

$$\text{prefix}(w_i) = \texttt{L}$$
$$\text{prefix}(w_i) = \texttt{L'}$$
$$\text{prefix}(w_i) = \texttt{L'O}$$
$$\text{prefix}(w_i) = \texttt{L'Oc}$$
$$\text{suffix}(w_i) = \texttt{tane}$$
$$\text{suffix}(w_i) = \texttt{ane}$$
$$\text{suffix}(w_i) = \texttt{ne}$$
$$\text{suffix}(w_i) = \texttt{e}$$
$$\text{word-shape}(w_i) = \texttt{X'Xxxxxxxx}$$
$$\text{short-word-shape}(w_i) = \texttt{X'Xx}$$

Feature Encoding in 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 21.6 Word-by-word feature encoding for NER.

NER as Sequence Labeling

- Many NLP problems can be cast as sequence labeling problems
  - POS – part of speech tagging
  - NER – named entity recognition
  - SRL – semantic role labeling
- Input
  - Sequence $w_1 w_2 \ldots w_n$
- Output
  - Labeled words
- Classification methods
  - Can use the categories of the previous tokens as features in classifying the next one
  - Direction matters

NER as Sequence Labeling

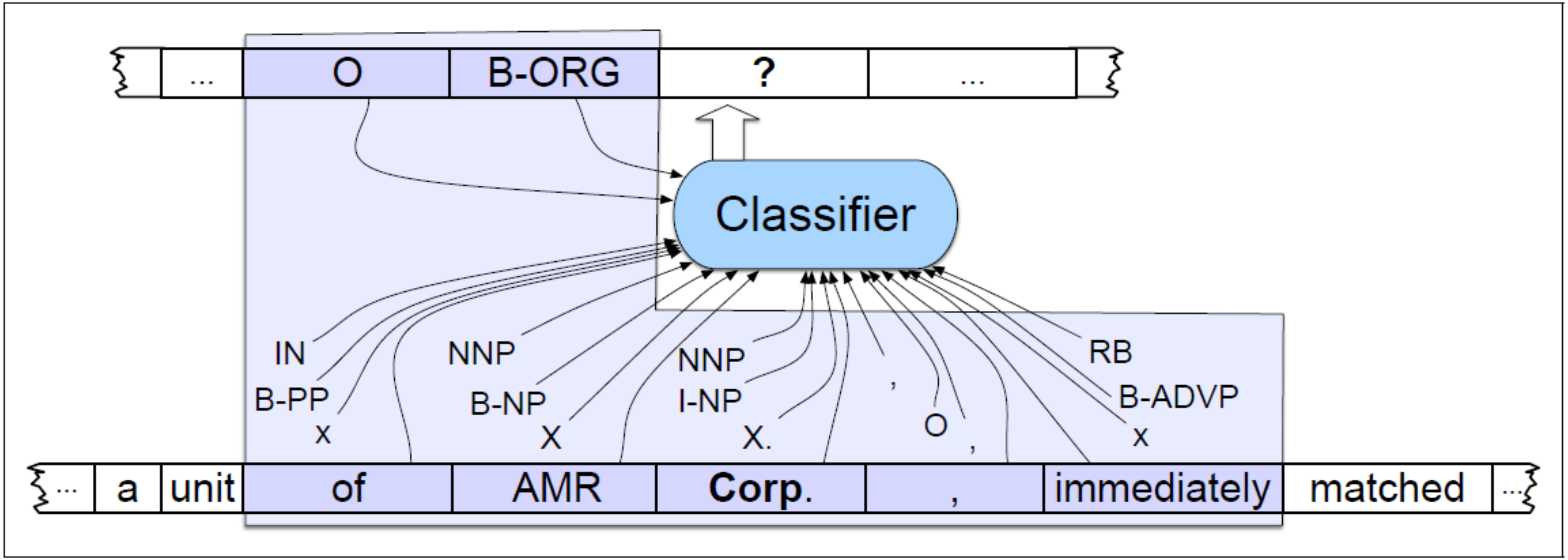

Figure 21.7 Named entity recognition as sequence labeling. The features available to the classifier during training and classification are those in the boxed area.

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

Figure 21.17 Examples of absolute, relational and durational temporal expressions.

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*

Figure 21.18 Examples of temporal lexical triggers.

TempEx Example

```
# yesterday/today/tomorrow
$string =~ s/(($OT+(early|earlier|later?)$CT+\s+)?(($OT+the$CT+\s+)?$OT+day$CT+\s+
$OT+(before|after)$CT+\s+)?$OT+$TERelDayExpr$CT+(\s+$OT+(morning|afternoon|
evening|night)$CT+)?)/<TIMEX2 TYPE=\"DATE\">$1<\/TIMEX2>/gio;

$string =~ s/($OT+\w+$CT+\s+)
<TIMEX2 TYPE=\"DATE\"[^>]*>($OT+(Today|Tonight)$CT+)<\/TIMEX2>/$1$2/gso;

# this/that (morning/afternoon/evening/night)
$string =~ s/(($OT+(early|earlier|later?)$CT+\s+)?$OT+(this|that|every|the$CT+\s+
$OT+(next|previous|following))$CT+\s*$OT+(morning|afternoon|evening|night)
$CT+(\s+$OT+thereafter$CT+)?)/<TIMEX2 TYPE=\"DATE\">$1<\/TIMEX2>/gosi;
```

Figure 21.19 Fragment of Perl code from MITRE's TempEx temporal tagging system.

TimeML

```
<TIMEX3 id=''t1'' type="DATE" value="2007-07-02" functionInDocument="CREATION_TIME">
July 2, 2007 </TIMEX3>  A fare increase initiated <TIMEX3 id="t2" type="DATE"
value="2007-W26"  anchorTimeID="t1">last week</TIMEX3> by UAL Corp's United  Airlines
was matched by competitors over <TIMEX3 id="t3" type="DURATION" value="P1WE"
anchorTimeID="t1"> the weekend </TIMEX3>, marking the second successful fare increase
in <TIMEX3 id="t4" type="DURATION" value="P2W" anchorTimeID="t1"> two weeks </TIMEX3>.
```

Figure 21.21 TimeML markup including normalized values for temporal expressions.

TimeBank

```
<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.
```

Figure 21.25 Example from the TimeBank corpus.

The Message Understanding Conference (MUC)

- Slot Filling

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.

MUC Example

Tie-up-1		**Activity-1**:	
RELATIONSHIP	tie-up	COMPANY	Bridgestone Sports Taiwan Co.
ENTITIES	Bridgestone Sports Co.	PRODUCT	iron and "metal wood" clubs
	a local concern	START DATE	DURING: January 1990
	a Japanese trading house		
JOINT VENTURE	Bridgestone Sports Taiwan Co.		
ACTIVITY	Activity-1		
AMOUNT	NT$20000000		

Figure 21.26 The templates produced by FASTUS given the input text on page 25.

Biomedical example

- Gene labeling
- Sentence:
  - [$_{GENE}$ BRCA1] and [$_{GENE}$ BRCA2] are human genes that produce tumor suppressor proteins

Other Examples

- Job announcements
  - Location, title, starting date, qualifications, salary
- Seminar announcements
  - Time, title, location, speaker
- Medical papers
  - Drug, disease, gene/protein, cell line, species, substance

Filling the Templates

- Some fields get filled by text from the document
  - E.g., the names of people
- Others can be pre-defined values
  - E.g., successful/unsuccessful merger
- Some fields allow for multiple values

Evaluating Template-Based NER

- For each test document
  - Number of correct template extractions
  - Number of slot/value pairs extracted
  - Number of extracted slot/value pairs that are correct

Introduction to NLP

233.

Relation Extraction

Relation Extraction

- Person-person
  - ParentOf, MarriedTo, Manages
- Person-organization
  - WorksFor
- Organization-organization
  - IsPartOf
- Organization-location
  - IsHeadquarteredAt

Relation Extraction

- Core NLP task
  - Used for building knowledge bases, question answering
- Input
  - **Mazda North American Operations** *is headquartered in* **Irvine, Calif.**, and oversees the sales, marketing, parts and customer service support of Mazda vehicles in the United States and Mexico through nearly 700 dealers.
- Output (predicate)
  - IsHeadquarteredIn (Mazda North American Operations, Irvine)

Relation extraction

- Using patterns
  - Regular expressions
  - Gazetteers
- Supervised learning
- Semi-supervised learning
  - Using seeds

Relation Extraction

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**].

The ACE Evaluation

- Newspaper data
- Entities:
  - Person, Organization, Facility, Location, Geopolitical Entity
- Relations:
  - Role, Part, Located, Near, Social

The ACE Evaluation

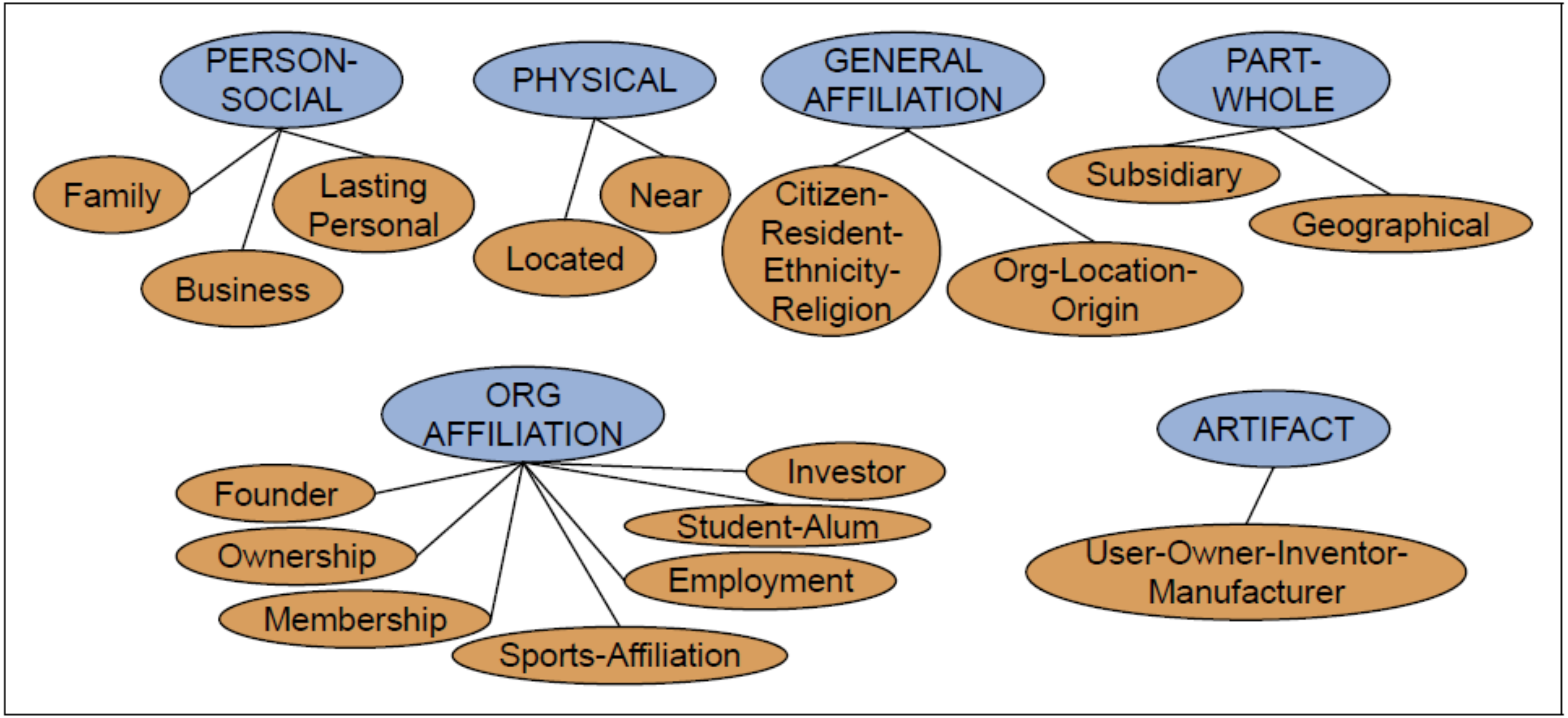

Figure 21.8 The 17 relations used in the ACE relation extraction task.

Semantic Relations

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**...

Figure 21.9 Semantic relations with examples and the named entity types they involve.

Extracting IS-A Relations

- Hearst's patterns
  - X and other Y
  - X or other Y
  - Y such as X
  - Y, including X
  - Y, especially X
- Example
  - Evolutionary relationships between the platypus and other mammals

Hypernym Extraction (Hearst)

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

Figure 21.11 Hand-built lexico-syntactic patterns for finding hypernyms, using {} to mark optionality (Hearst, 1992a, 1998).

Supervised Relation Extraction

- Look for sentences that have two entities that we know are part of the target relation
- Look at the other words in the sentence, especially the ones between the two entities
- Use a classifier to determine whether the relation exists

Semi-supervised Relation Extraction

- Start with some seeds, e.g.,
  - **Beethoven** *was born* in December **1770** in Bonn
- Look for other sentences with the same words
- Look for expressions that appear nearby
- Look for other sentences with the same expressions

Bootstrapping

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 *seeds*
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

Figure 21.14 Bootstrapping from seed entity pairs to learn relations.

Bootstrapping

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

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

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

Bootstrapping

```
/ [ORG], which uses [LOC] as a hub /
/ [ORG]'s hub at [LOC] /
/ [LOC] a main hub for [ORG] /
```

Evaluating Relation Extraction

- Precision P
  - correctly extracted relations/all extracted relations
- Recall R
  - correctly extracted relations/all existing relations
- F1 measure
  - F1 = 2PR/(P+R)
- If there is no annotated data
  - only measure precision