Information Extraction

Chapter 20-20.3 J&M 3rd edition

IP disclosure: content on these slides was adapted from Ray Mooney and Ellen Riloff

Today's Lecture

- Information extraction
- Named Entity Recognition
- Relation Extraction
- Temporal Expression Processing

Information Extraction (IE)

- Identify specific pieces of information (data) in an unstructured or semi-structured textual document or speech transcription.
- Transform unstructured information in a corpus of documents or web pages into a structured database.
- Applied to different types of text:
  - Newspaper articles
  - Web pages
  - Scientific articles
  - Newsgroup messages
  - Classified ads
  - Medical notes

MUC

- DARPA funded significant efforts in IE in the early to mid 1990's.
- Message Understanding Conference (MUC) was an annual event/competition where results were presented.
- Focused on extracting information from news articles:
  - Terrorist events
  - Industrial joint ventures
  - Company management changes
- Information extraction of particular interest to the intelligence community (CIA, NSA).
- Established standard evaluation methodology using development (training) and test data and metrics: precision, recall, F-measure.

Named Entity (NE) Recognition

- Specific type of information extraction in which the goal is to extract proper names of particular types of entities such as people, places, organizations, etc.
- Usually a preprocessing step for subsequent task-specific IE, or other tasks such as question answering.
- NEs are application specific

Named Entity Recognition Example

U.S. Supreme Court quashes 'illegal' Guantanamo trials

Military trials arranged by the Bush administration for detainees at Guantanamo Bay are illegal, the United States Supreme Court ruled Thursday. The court found that the trials — known as military commissions — for people detained on suspicion of terrorist activity abroad do not conform to any act of Congress. The justices also rejected the government's argument that the Geneva Conventions regarding prisoners of war do not apply to those held at Guantanamo Bay. Writing for the 5-3 majority, Justice Stephen Breyer said the White House had overstepped its powers under the U.S. Constitution. "Congress has not issued the executive a blank cheque," Breyer wrote.

President George W. Bush said he takes the ruling very seriously and would find a way to both respect the court's findings and protect the American people.

Named Entity Recognition Example

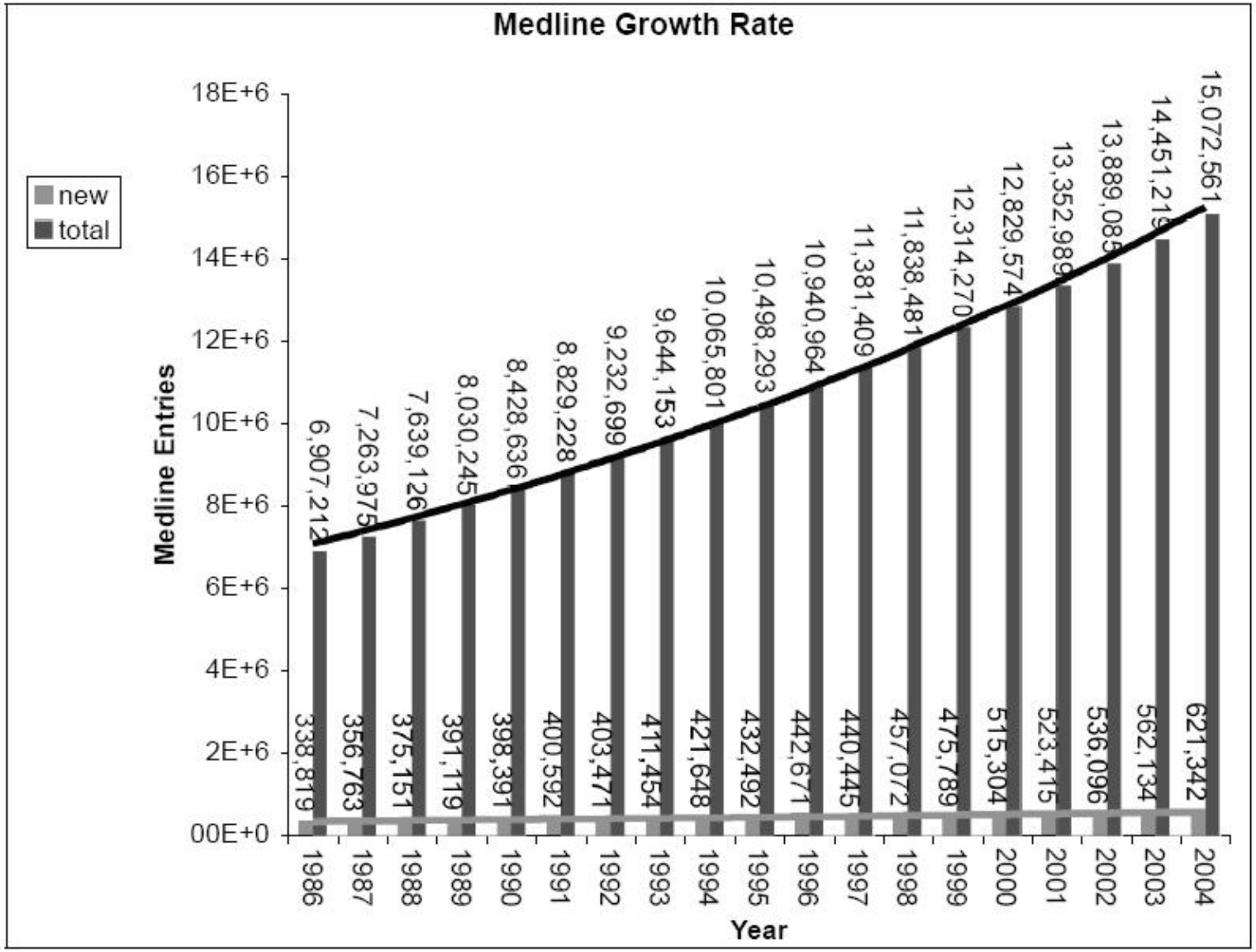
people places organizations

U.S. Supreme Court quashes 'illegal' Guantanamo trials

Military trials arranged by the Bush administration for detainees at Guantanamo Bay are illegal, the United States Supreme Court ruled Thursday. The court found that the trials — known as military commissions — for people detained on suspicion of terrorist activity abroad do not conform to any act of Congress. The justices also rejected the government's argument that the Geneva Conventions regarding prisoners of war do not apply to those held at Guantanamo Bay. Writing for the 5-3 majority, Justice Stephen Breyer said the White House had overstepped its powers under the U.S. Constitution. "Congress has not issued the executive a blank cheque," Breyer wrote.

President George W. Bush said he takes the ruling very seriously and would find a way to both respect the court's findings and protect the American people.

Biomedical Information Extraction



Medline Corpus

TI - Two potentially oncogenic cyclins, cyclin A and cyclin D1, share common properties of subunit configuration, tyrosine phosphorylation and physical association with the Rb protein

AB - Originally identified as a 'mitotic cyclin', cyclin A exhibits properties of growth factor sensitivity, susceptibility to viral subversion and association with a tumor-suppressor protein, properties which are indicative of an S-phase-promoting factor (SPF) as well as a candidate proto-oncogene …

Moreover, cyclin D1 was found to be phosphorylated on tyrosine residues in vivo and, like cyclin A, was readily phosphorylated by pp60c-src in vitro.

In synchronized human osteosarcoma cells, cyclin D1 is induced in early G1 and becomes associated with p9Ckshs1, a Cdk-binding subunit.

Immunoprecipitation experiments with human osteosarcoma cells and Ewing's sarcoma cells demonstrated that cyclin D1 is associated with both p34cdc2 and p33cdk2, and that cyclin D1 immune complexes exhibit appreciable histone H1 kinase activity …

Medline Corpus: Named Entity Recognition (Proteins)

TI - Two potentially oncogenic cyclins, cyclin A and cyclin D1, share common properties of subunit configuration, tyrosine phosphorylation and physical association with the Rb protein

AB - Originally identified as a 'mitotic cyclin', cyclin A exhibits properties of growth factor sensitivity, susceptibility to viral subversion and association with a tumor-suppressor protein, properties which are indicative of an S-phase-promoting factor (SPF) as well as a candidate proto-oncogene …

Moreover, cyclin D1 was found to be phosphorylated on tyrosine residues in vivo and, like cyclin A, was readily phosphorylated by pp60c-src in vitro.

In synchronized human osteosarcoma cells, cyclin D1 is induced in early G1 and becomes associated with p9Ckshs1, a Cdk-binding subunit.

Immunoprecipitation experiments with human osteosarcoma cells and Ewing's sarcoma cells demonstrated that cyclin D1 is associated with both p34cdc2 and p33cdk2, and that cyclin D1 immune complexes exhibit appreciable histone H1 kinase activity …

Amazon Book Description

```
....
</td></tr>
</table>
<b class="sans">The Age of Spiritual Machines : When Computers Exceed Human Intelligence</b><br>
<font face=verdana,arial,helvetica size=-1>
by <a href="/exec/obidos/search-handle-url/index=books&field-author=
        Kurzweil%2C%20Ray/002-6235079-4593641">
Ray Kurzweil</a><br>
</font>
<br>
<a href="http://images.amazon.com/images/P/0140282025.01.LZZZZZZZ.jpg">
<img src="http://images.amazon.com/images/P/0140282025.01.MZZZZZZZ.gif" width=90
   height=140 align=left border=0></a>
<font face=verdana,arial,helvetica size=-1>
<span class="small">
<span class="small">
<b>List Price:</b> <span class=listprice>$14.95</span><br>
<b>Our Price: <font color=#990000>$11.96</font></b><br>
<b>You Save:</b> <font color=#990000><b>$2.99 </b>
(20%)</font><br>
</span>
<p> <br>…
```

Extracted Book Template

Title: The Age of Spiritual Machines :
When Computers Exceed Human Intelligence
Author: Ray Kurzweil
List-Price: $14.95
Price: $11.96
:
:

Other Applications

- Job postings
- Job resumes
- Seminar announcements
- Company information from the web
- Continuing education course info from the web
- University information from the web
- Apartment rental ads
- Molecular biology information from MEDLINE

How Difficult is NER?

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

[$_{PERS}$ 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...

IE as Sequence Labeling

- Can extract features describing each token in the text.
- Can apply a sliding window classifier using various classification algorithms.
- Can apply probabilistic sequence models:
  - HMM
  - CRF (Conditional Random Fields)

Sequence Labeling for NER

Words	Label
American	B_{ORG}
Airlines	I_{ORG}
,	O
a	O
unit	O
of	O
AMR	B_{ORG}
Corp.	I_{ORG}
,	O
immediately	O
matched	O
the	O
move	O
,	O
spokesman	O
Tim	B_{PERS}
Wagner	I_{PERS}
said	O
.	O

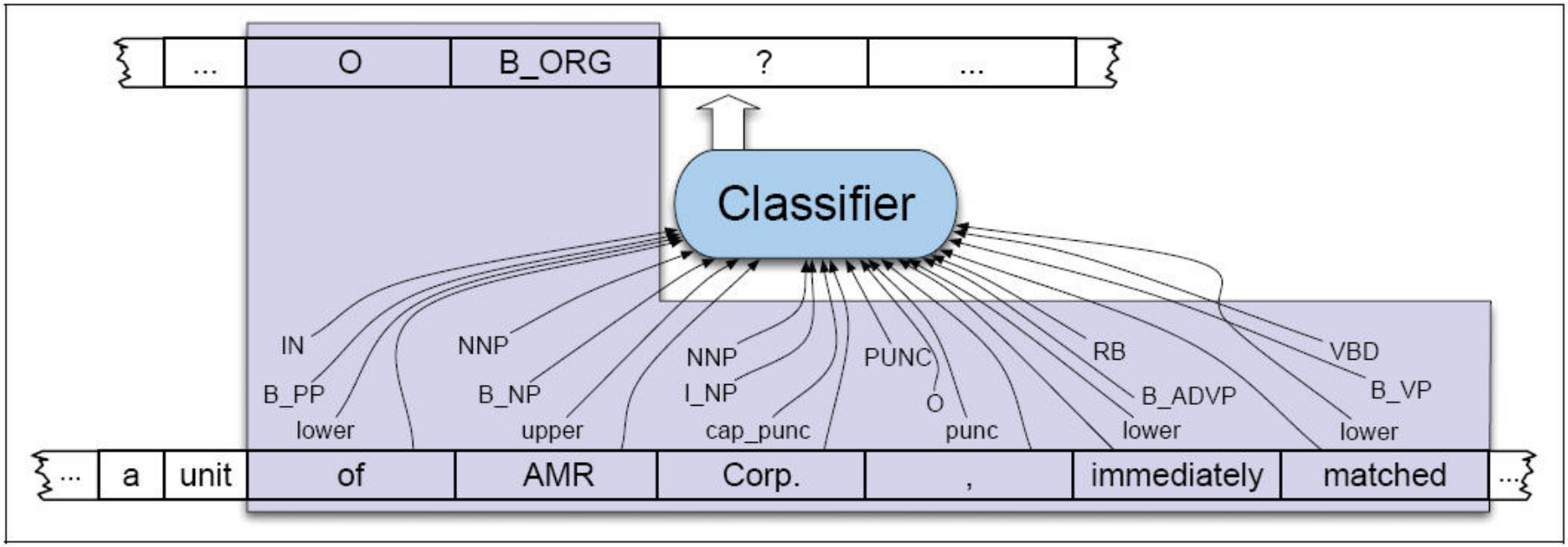
Typical features

- Lexical items
- Shape features
- Gazetteers
- Stemmed lexical item
- POS
- Trigger words

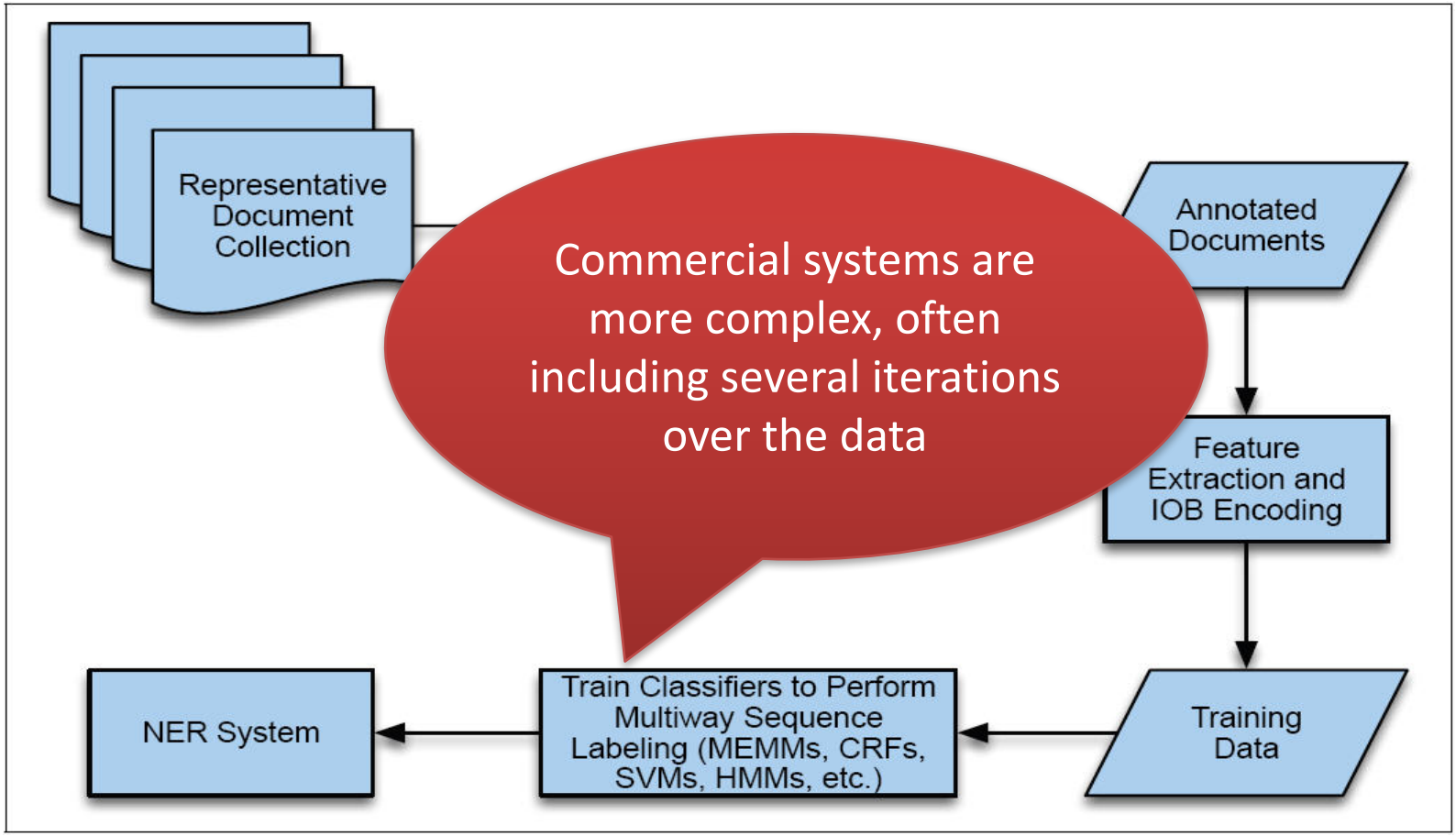
Sequence Labeling for NER

Features				Label
American	NNP	B_{NP}	cap	B_{ORG}
Airlines	NNPS	I_{NP}	cap	I_{ORG}
,	PUNC	O	punc	O
a	DT	B_{NP}	lower	O
unit	NN	I_{NP}	lower	O
of	IN	B_{PP}	lower	O
AMR	NNP	B_{NP}	upper	B_{ORG}
Corp.	NNP	I_{NP}	cap_punc	I_{ORG}
,	PUNC	O	punc	O
immediately	RB	B_{ADVP}	lower	O
matched	VBD	B_{VP}	lower	O
the	DT	B_{NP}	lower	O
move	NN	I_{NP}	lower	O
,	PUNC	O	punc	O
spokesman	NN	B_{NP}	lower	O
Tim	NNP	I_{NP}	cap	B_{PER}
Wagner	NNP	I_{NP}	cap	I_{PER}
said	VBD	B_{VP}	lower	O
.	PUNC	O	punc	O

Sequence Labeling for NER



Sequence Labeling for NER



Evaluating IE Accuracy

- Always evaluate performance on independent, manually-annotated test data not used during system development.
- Measure for each test document:
  - Total number of NEs in the solution template: *N*
  - Total number of NEs extracted by the system: *E*
  - Number of extracted NEs that are correct (i.e. in the gold standard): *C*
- Compute average value of metrics adapted from IR:
  - Recall = *C*/*N*
  - Precision = *C*/*E*
  - F-Measure = Harmonic mean of recall and precision

    2 x Precision x Recall / (Precision + Recall)

Relation Extraction

Relation Extraction

- Once entities are recognized, identify specific relations between entities
  - Employed-by
  - Located-at
  - Part-of
- Example:
  - Michael Dell is the CEO of Dell Computer Corporation and lives in Austin Texas.

Medline Corpus: Relation Extraction Protein Interactions

TI - Two potentially oncogenic cyclins, cyclin A and cyclin D1, share common properties of subunit configuration, tyrosine phosphorylation and physical association with the Rb protein

AB - Originally identified as a 'mitotic cyclin', cyclin A exhibits properties of growth factor sensitivity, susceptibility to viral subversion and association with a tumor-suppressor protein, properties which are indicative of an S-phase-promoting factor (SPF) as well as a candidate proto-oncogene …

Moreover, cyclin D1 was found to be phosphorylated on tyrosine residues in vivo and, like cyclin A, was readily phosphorylated by pp60c-src in vitro.

In synchronized human osteosarcoma cells, cyclin D1 is induced in early G1 and becomes associated with p9Ckshs1, a Cdk-binding subunit.

Immunoprecipitation experiments with human osteosarcoma cells and Ewing's sarcoma cells demonstrated that cyclin D1 is associated with both p34cdc2 and p33cdk2, and that cyclin D1 immune complexes exhibit appreciable histone H1 kinase activity …

Relation Extraction

- The goal is to identify a set of ordered tuples over elements of a domain.

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

Supervised Learning for Relation Extraction

```
function FINDRELATIONS(words) returns relations

    relations ← nil
    entities ← FINDENTITIES(words)
    forall entity pairs ⟨e1, e2⟩ in entities do
        if RELATED?(e1, e2)
            relations ← relations+CLASSIFYRELATION(e1, e2)
```

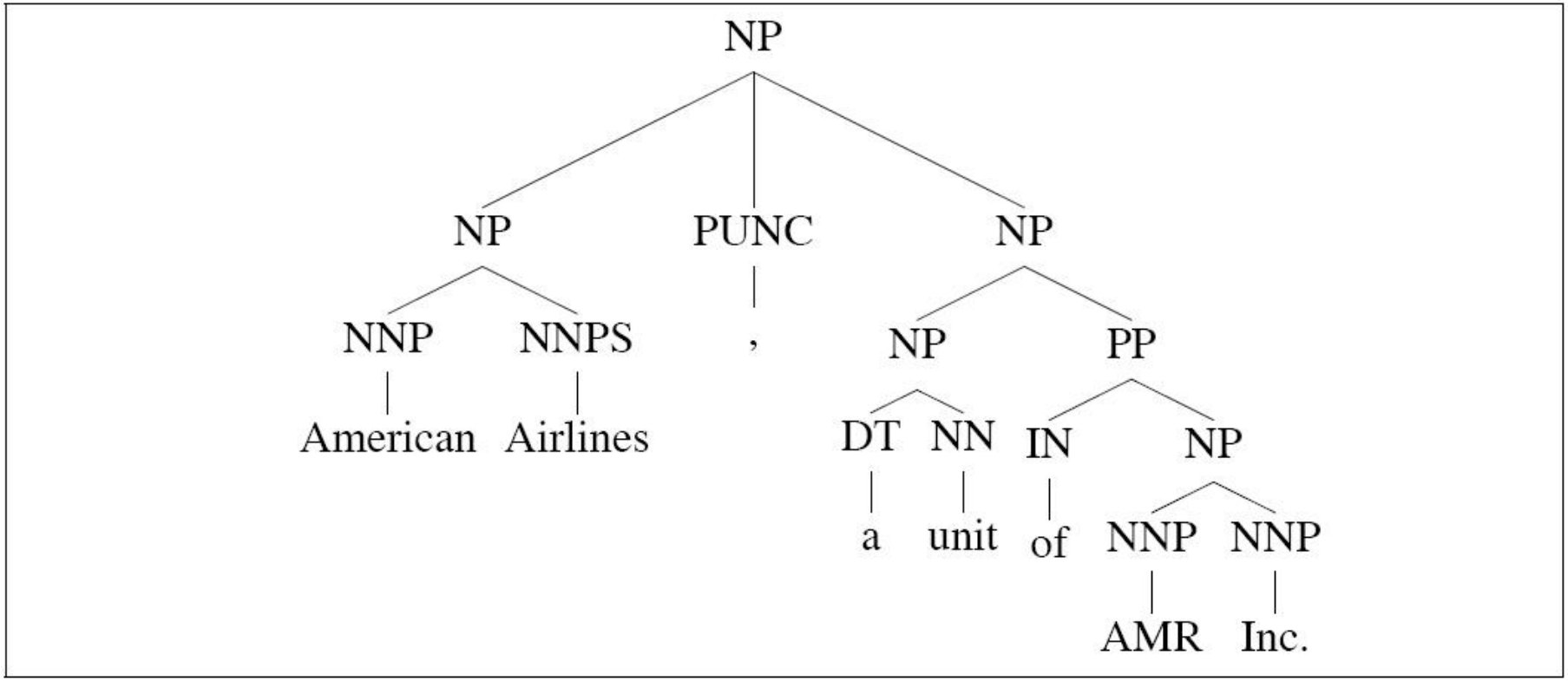
Supervised Learning for Relation Extraction

- Features commonly used include:
  - NE types
  - Bag of words for each argument
  - Headwords of the arguments
  - Bag of words and bigrams between entities
    - Stemmed versions
  - Words and stems in the immediate context
  - Presence of particular constructions
  - Chunk based-phrase paths
  - Constituent-tree paths

Supervised Learning for Relation Extraction

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	$Airlines \leftarrow_{subj} matched \leftarrow_{comp} said \rightarrow_{subj} Wagner$

Syntactic Features for Supervised Learning for Relation Extraction



Example of an appositive construction

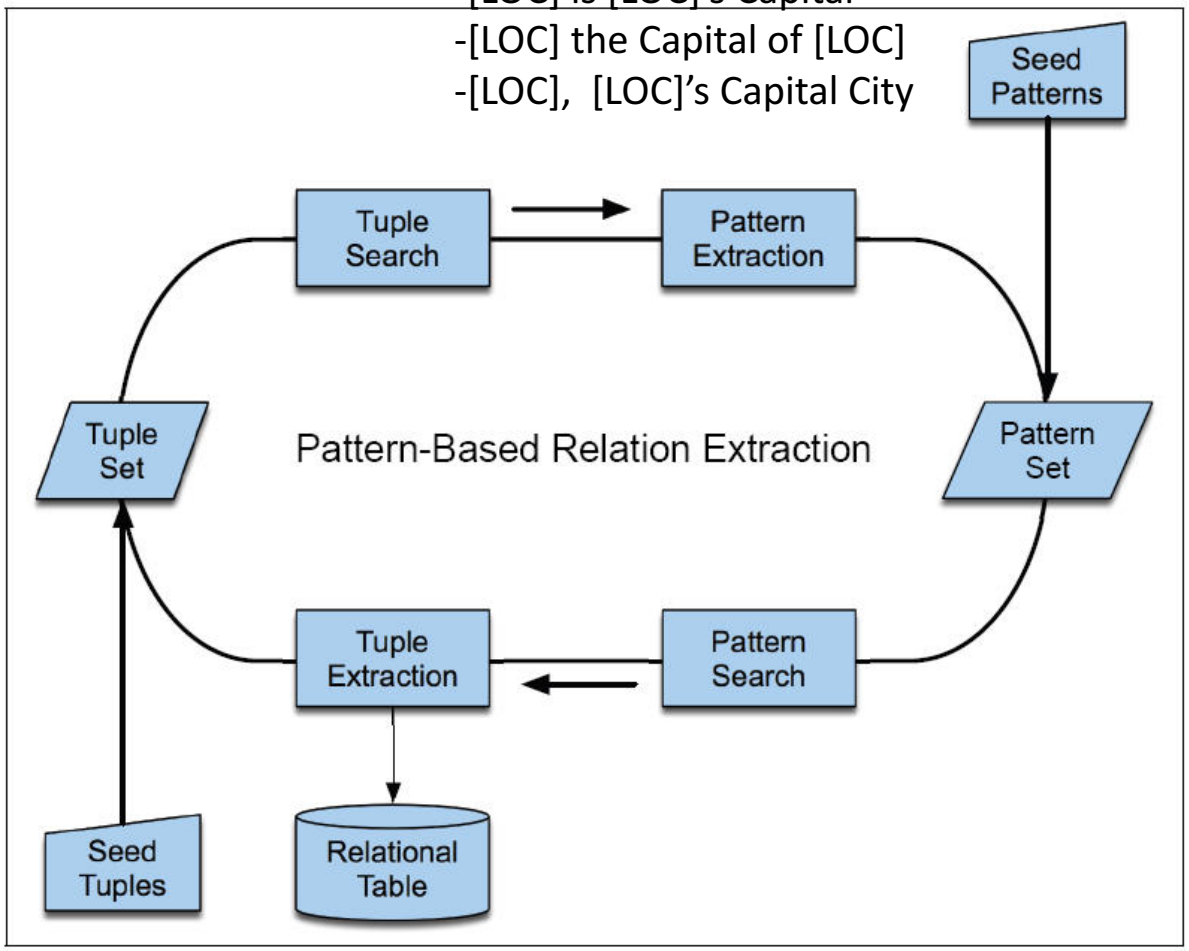
Lightly Supervised Relation Extraction

- We can use regular expressions to mine the web for relations:
  - * *is the Capital of* *
- Will return the following examples:
  - Cairo *is the Capital of* Egypt
  - Madrid *is the Capital of* Spain
- But also:
  - New York City *is the Capital of* the World
  - Harlem *is the Capital of* every Ghetto in Town
- Are we missing something?

Lightly Supervised Relation Extraction

-Cairo is Egypt's Capital
-Cairo the Capital of Egypt
-Cairo, Egypt's Capital City

-[LOC] is [LOC]'s Capital
-[LOC] the Capital of [LOC]
-[LOC], [LOC]'s Capital City



Cairo, Capital, Egypt

Lightly Supervised Relation Extraction

We need to address the following problems:

- Representation of the search patterns
- Assessing accuracy and coverage of discovered patterns
  - (Riloff and Jones, 1999)
- Assessing reliability of discovered tuples

Evaluation of Relation Extraction Systems

Can focus on:

1. Measuring relation extraction (can use labeled or unlabeled recall, precision and f-measure)
2. Measuring how much the system can discover (humans analyze the output of the system and compute accuracy, no recall)
3. Use external sources such as gazetteers to measure recall

Temporal Expression Recognition

Temporal Expression Recognition

Types of 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

Temporal Expression Recognition

Examples of 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 Expression Recognition

Annotation scheme

A fare increase initiated <TIMEX3>last week </TIMEX3> by UAL Corp's United Airlines was matched by competitors over <TIMEX3>the weekend </TIMEX3>, marking the second successful fare increase in <TIMEX3>two weeks </TIMEX3>.

Temporal Expression Recognition

Approaches:

- Rule-based systems
- Sequence labeling systems
- Constituent based classification

Rule-based Systems for Temporal Expression Recognition

```
# 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;
```

Sequence Labeling Systems for Temporal Expression Recognition

Use the same I,O,B scheme to train a classification algorithm:

A fare increase initiated last week by UAL Corp's

O O O O B I O O O

Constituent Based Approaches

- Give the segmentation task to the syntactic parser
- Train binary classifiers on chunks

Evaluation of Temporal Expression Recognition

- Standard measures: precision, recall, and f-measure
- Best systems are at ~.87 (labeled p, and r)

Temporal Processing

- Ambiguities from trigger words:
  - I was listening to *Manic Monday* on the radio
- After extraction of temporal expressions we need to do normalization.

Information Extraction Issues

- Better active learning methods
- Integrating entity and relation extraction
- Semi-supervised IE
- Adaptation and transfer to new tasks
- Mining extracted data to find cross-document regularities.