Topics for Today

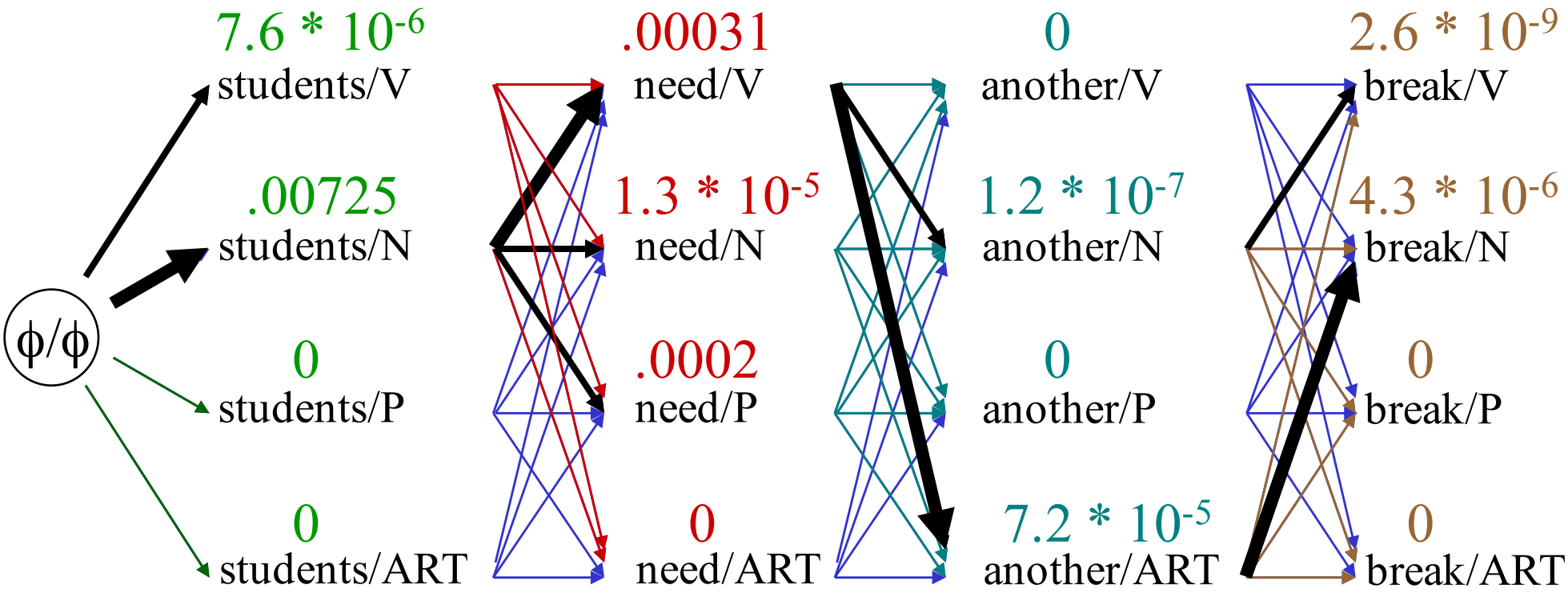
- **Finish up HMMs for POS tagging**
- **Information extraction**
- **Named entity detection**

Hidden Markov Models

$Q = q_1 q_2 \ldots q_N$	a set of N **states**
$A = a_{11} a_{12} \ldots a_{n1} \ldots a_{nn}$	a **transition probability matrix** A, each a_{ij} representing the probability of moving from state i to state j, s.t. $\sum_{j=1}^{n} a_{ij} = 1 \quad \forall i$
$O = o_1 o_2 \ldots o_T$	a sequence of T **observations**, each one drawn from a vocabulary $V = v_1, v_2, ..., v_V$
$B = b_i(o_t)$	a sequence of **observation likelihoods**, also called **emission probabilities**, each expressing the probability of an observation o_t being generated from a state i
q_0, q_F	a special **start state** and **end (final) state** that are not associated with observations, together with transition probabilities $a_{01} a_{02} \ldots a_{0n}$ out of the start state and $a_{1F} a_{2F} \ldots a_{nF}$ into the end state

Figure, copyright J&M 2nd ed

Viterbi Algorithm



Results

- Effective if probability estmates are computed from a large corpus
- Effective if corpus is of the same style as the input to be classified
- Consistently achieve accuracies of 97% or better using trigram model
- Cuts error rate in half vs. naive algorithm (90% accuracy rate)
- Can be smoothed using backoff or interpolation or discounting...

Information Extraction

- **Introduction**
  - Task definition
  - Evaluation
  - IE system architecture
- **Named entity detection**
- **Relation extraction**

Information extraction

individual documents

Information Extraction System

Who: _____
What: _____
Where: _____
When: _____
How: _____

Who: _____
What: _____
Where: _____
When: _____
How: _____

IE system: natural disasters

Disaster Type: earthquake
- •location: *Afghanistan*
- •date: *today*
- •magnitude: *6.9*
- •magnitude-confidence: high
- •epicenter: *a remote part of the country*
- •damage:
  - •human-effect:
    - •victim: *Thousands of people*
    - •number: *Thousands*
    - •outcome: dead
    - •confidence: medium
    - •confidence-marker: *feared*
  - •physical-effect:
    - •object: *entire villages*
    - •outcome: damaged
    - •confidence: medium
    - •confidence-marker: *Details now hard to come by / reports say*

Thousands of people are feared dead following... (voice-over) ...a powerful earthquake that hit Afghanistan today. The quake registered 6.9 on the Richter scale, centered in a remote part of the country. (on camera) Details now hard to come by, but reports say entire villages were buried by the quake.

Document no.: ABC19980530.1830.0342
Date/time: 05/30/1998 18:35:42.49

IE system: terrorism

SAN SALVADOR, 15 JAN 90 (ACAN-EFE) -- [TEXT] ARMANDO CALDERON SOL, PRESIDENT OF THE NATIONALIST REPUBLICAN ALLIANCE (ARENA), THE RULING SALVADORAN PARTY, TODAY CALLED FOR AN INVESTIGATION INTO ANY POSSIBLE CONNECTION BETWEEN THE **MILITARY PERSONNEL IMPLICATED IN THE ASSASSINATION OF JESUIT PRIESTS**.

"IT IS SOMETHING SO HORRENDOUS, SO MONSTROUS, THAT WE MUST INVESTIGATE THE **POSSIBILITY THAT THE FMLN (FARABUNDO MARTI NATIONAL LIBERATION FRONT) STAGED THIS ASSASSINATION** TO DISCREDIT THE GOVERNMENT," CALDERON SOL SAID.

SALVADORAN PRESIDENT ALFREDO CRISTIANI **IMPLICATED FOUR OFFICERS, INCLUDING ONE COLONEL, AND FIVE MEMBERS OF THE ARMED FORCES IN THE ASSASSINATION OF SIX JESUIT PRIESTS AND TWO WOMEN ON 16 NOVEMBER AT THE CENTRAL AMERICAN UNIVERSITY.**

IE system: output

1. DATE	- 15 JAN 90
2. LOCATION	EL SALVADOR: CENTRAL AMERICAN UNIVERSITY
3. TYPE	MURDER
4. STAGE OF EXECUTION	ACCOMPLISHED
5. INCIDENT CATEGORY	TERRORIST ACT
6. PERP: INDIVIDUAL ID	"FOUR OFFICERS" "ONE COLONEL" "FIVE MEMBERS OF THE ARMED FORCES"
7. PERP: ORGANIZATION ID	"ARMED FORCES", "FMLN"
8. PERP: CONFIDENCE	REPORTED AS FACT
9. HUM TGT: DESCRIPTION	"JESUIT PRIESTS" "WOMEN"
10. HUM TGT: TYPE	CIVILIAN: "JESUIT PRIESTS" CIVILIAN: "WOMEN"
11. HUM TGT: NUMBER	6: "JESUIT PRIESTS" 2: "WOMEN"
12. EFFECT OF INCIDENT	DEATH: "JESUIT PRIESTS" DEATH: "WOMEN"

Output Template

1. DATE	10 NOV 88
2. LOCATION	CHILE: SANTIAGO (CITY)
3. TYPE	MURDER
4. STAGE OF EXECUTION	ACCOMPLISHED
5. INCIDENT CATEGORY	TERRORIST ACT
6. PERP: INDIVIDUAL ID	"THEY"
7. PERP: CONFIDENCE	REPORTED AS FACT
9. HUM TGT: DESCRIPTION	"BIRDS"
10. HUM TGT: TYPE	CIVILIAN: "BIRDS"
11. HUM TGT: NUMBER	2: "BIRDS"
12. EFFECT OF INCIDENT	DEATH: "BIRDS"
13. INSTRUMENT	STONE

IE Example: Input Text

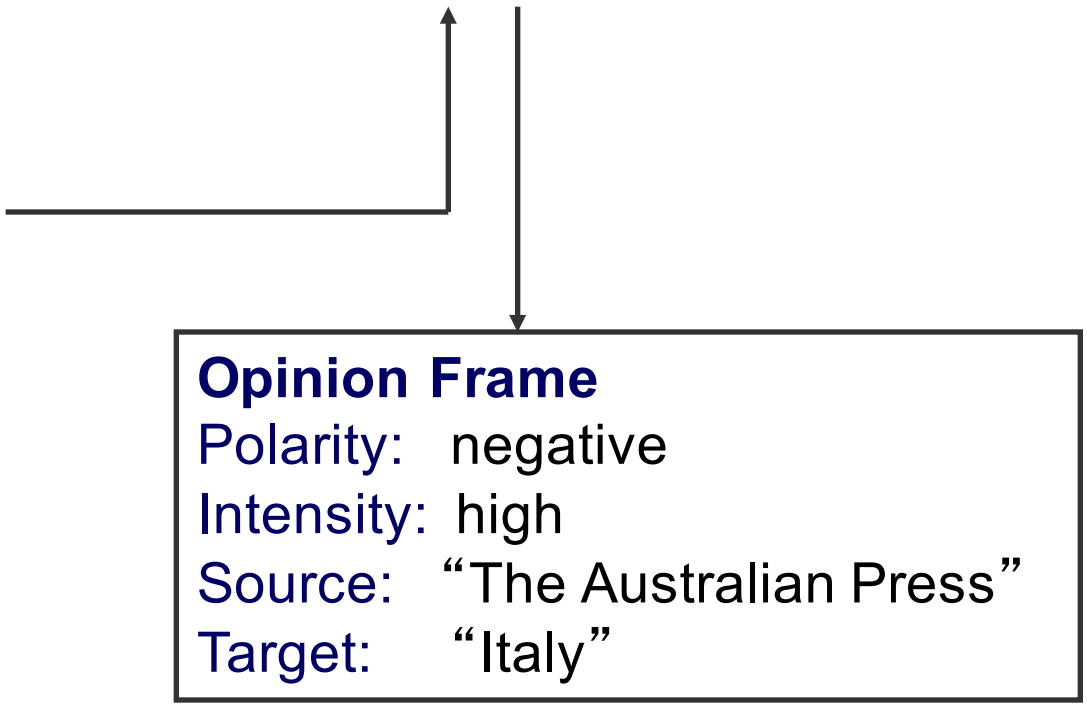
SANTIAGO, 10 NOV 88 (QUE PASA) -- [TEXT] [CONTINUED]

… THE PLENUM OF THE SOCIALIST PARTY [PS]-ALMEYDA WAS, OF COURSE, THE MOST EAGERLY ANTICIPATED…THEY AMBITIOUSLY FELT THAT THIS WAS THE OPPORTUNITY TO REMOVE SOME STRATEGIC OBSTACLES, SORT OF LIKE **KILLING TWO BIRDS WITH ONE STONE**: REGISTRATION AND THE SOUGHT-AFTER SOCIALIST UNITY...

Fine-grained Opinion Extraction

"The Australian Press **launched a bitter attack** on Italy"

- **Five components**
  - Opinion trigger
  - Polarity
    - positive
    - negative
    - neutral
  - Strength/intensity
    - low..extreme
  - Source (opinion holder)
  - Target (topic)



Information extraction (IE)

- Identify specific pieces of information (data) in a unstructured or semi-structured textual document.
- 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
  - Subjective language

Template slot types

- Slots in template typically filled by a **subSTRING** from the document.
- Some slots may have a **fixed SET of pre-specified possible fillers** that may not occur in the text itself.
  - Terrorist act: THREATENED, ATTEMPTED, ACCOMPLISHED.
  - Job type: CLERICAL, SERVICE, CUSTODIAL, etc.
  - Disaster type: EARTHQUAKE, TORNADO, etc.
- Some slots may allow **multiple fillers**.
  - E.g. victims
- Some domains may allow **multiple extracted templates per document**.
  - Multiple terrorist events in a single story

Evaluating IE systems

- Evaluate system performance vs. independent, manually-annotated test data not used during system development.
- Compute average value of metrics adapted from IR:
  - **Recall** = *# correct extractions / # extractions in gold standard*
  - **Precision** = *# correct extractions / # extractions by system*
  - **F-Measure** = Harmonic mean of recall and precision

State of the art

Unrestricted text:
65-70% R; 70-80% P

Semi-structured text:
90+% R/P

MUC
[1991-94]

ACE
[1991-94]

- terrorist activities
- business joint ventures
- microelectronic chip fabrication
- changes in corporate management
- natural disasters
- summarize medical patient records
- create job-listing databases from newsgroups
- bioinformatics

Information Extraction

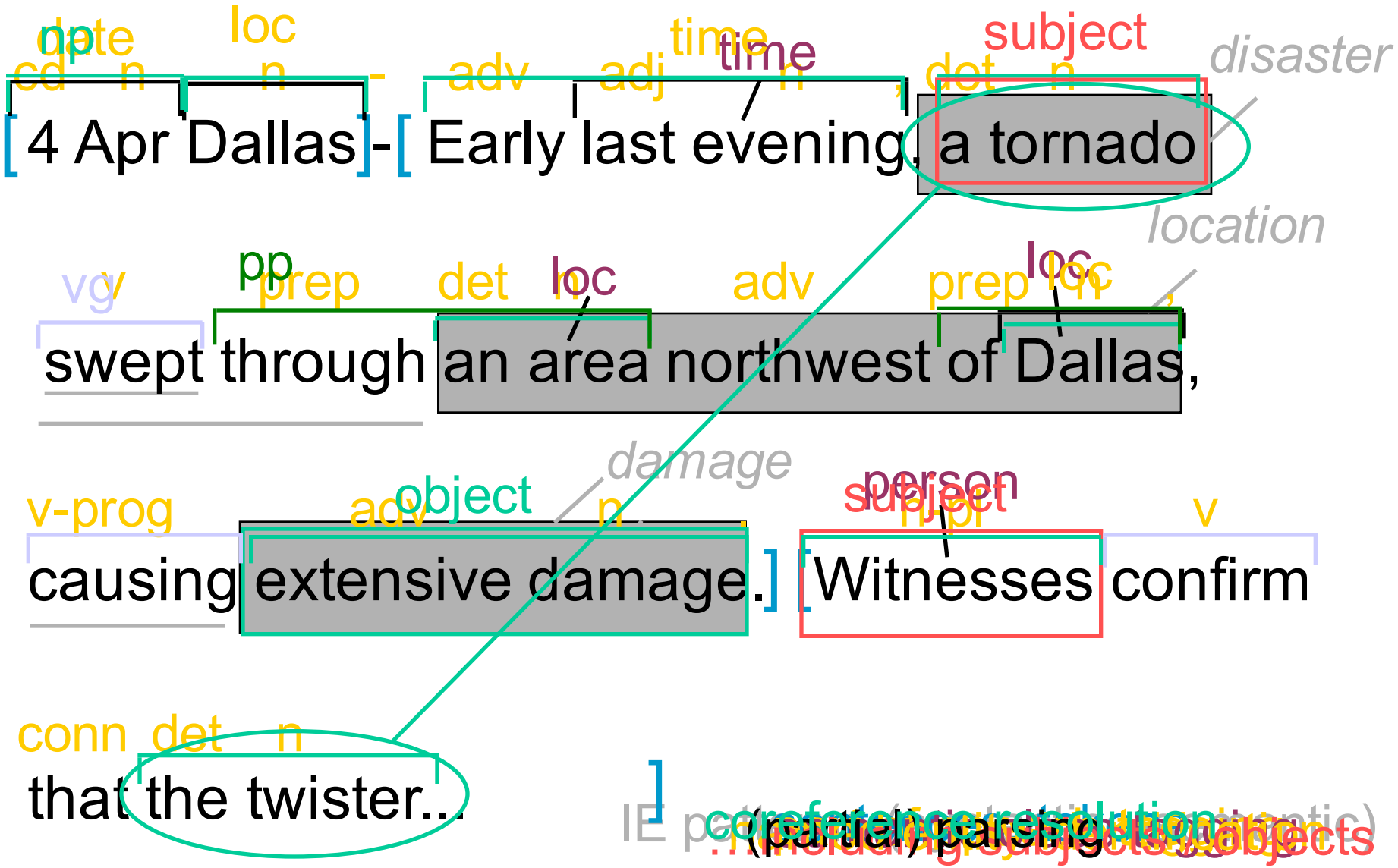
- **Introduction**
  - Task definition
  - Evaluation
  - IE system architecture
- **Named entity detection**
- **Relation extraction**

Natural disasters example

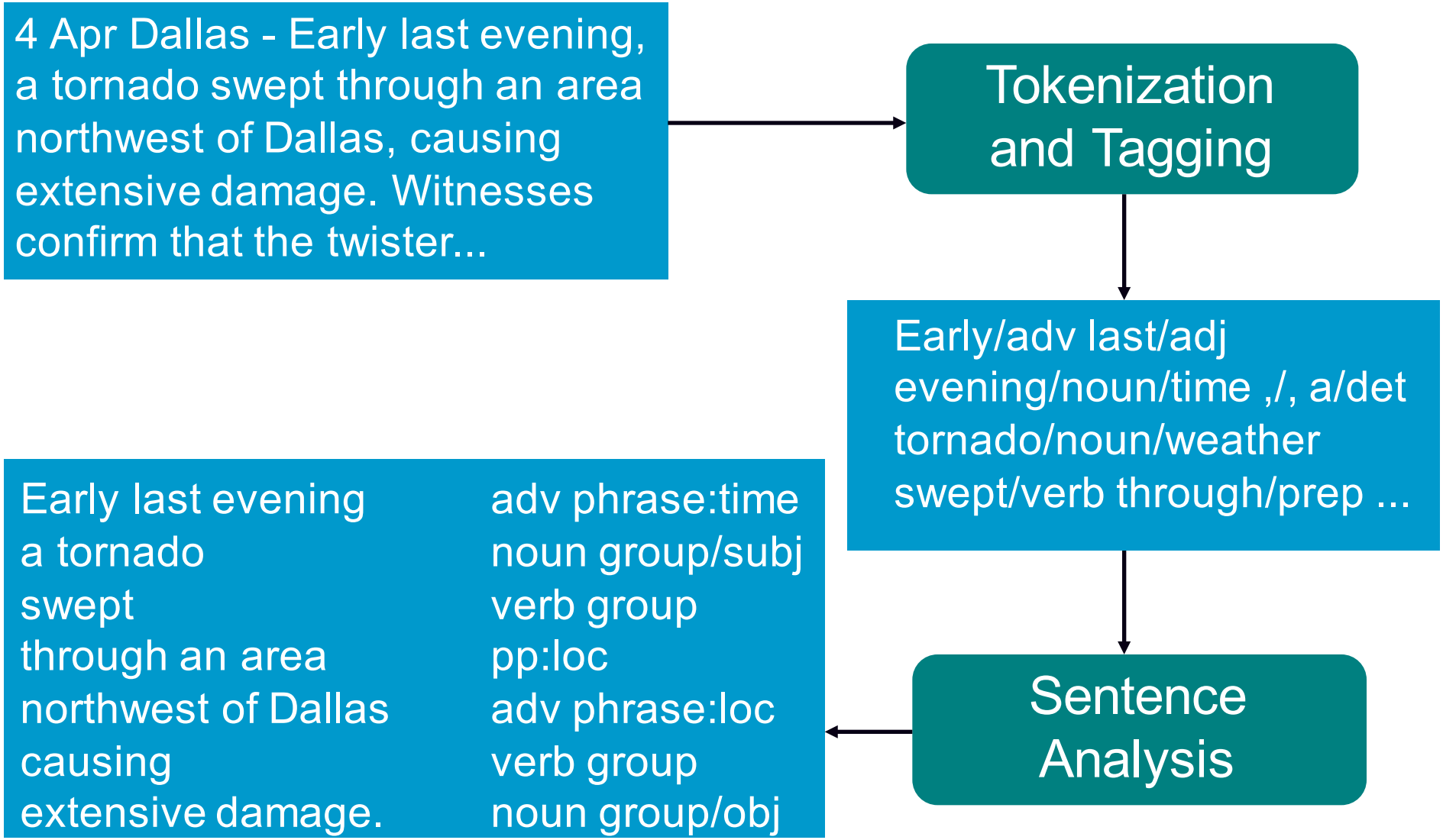
4 Apr Dallas - Early last evening, a tornado swept through an area northwest of Dallas, causing extensive damage. Witnesses confirm that the twister occurred without warning at approximately 7:15pm and destroyed two mobile homes. The Texaco station, at 102 Main Street, Farmers Branch, TX, was also severely damaged, but no injuries were reported. Total property damages are estimated to be $350,000.

Event:	tornado
Date:	4/3/2008
Time:	19:15
Location:	Farmers Branch: "northwest of Dallas": TX: USA
Damage:	"mobile homes" (2) "Texaco station" (1)
Estimated Losses:	$350,000
Injuries:	none

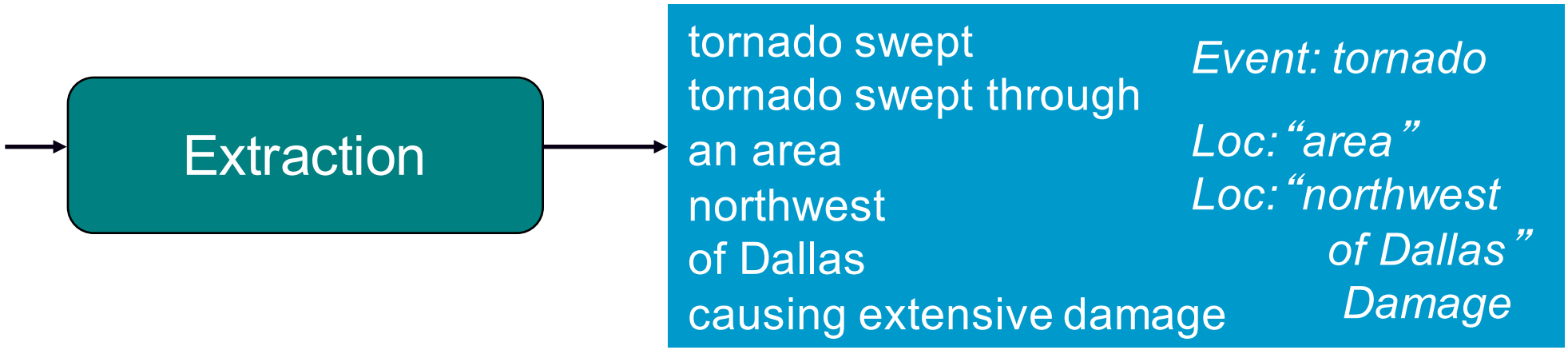
IE system components



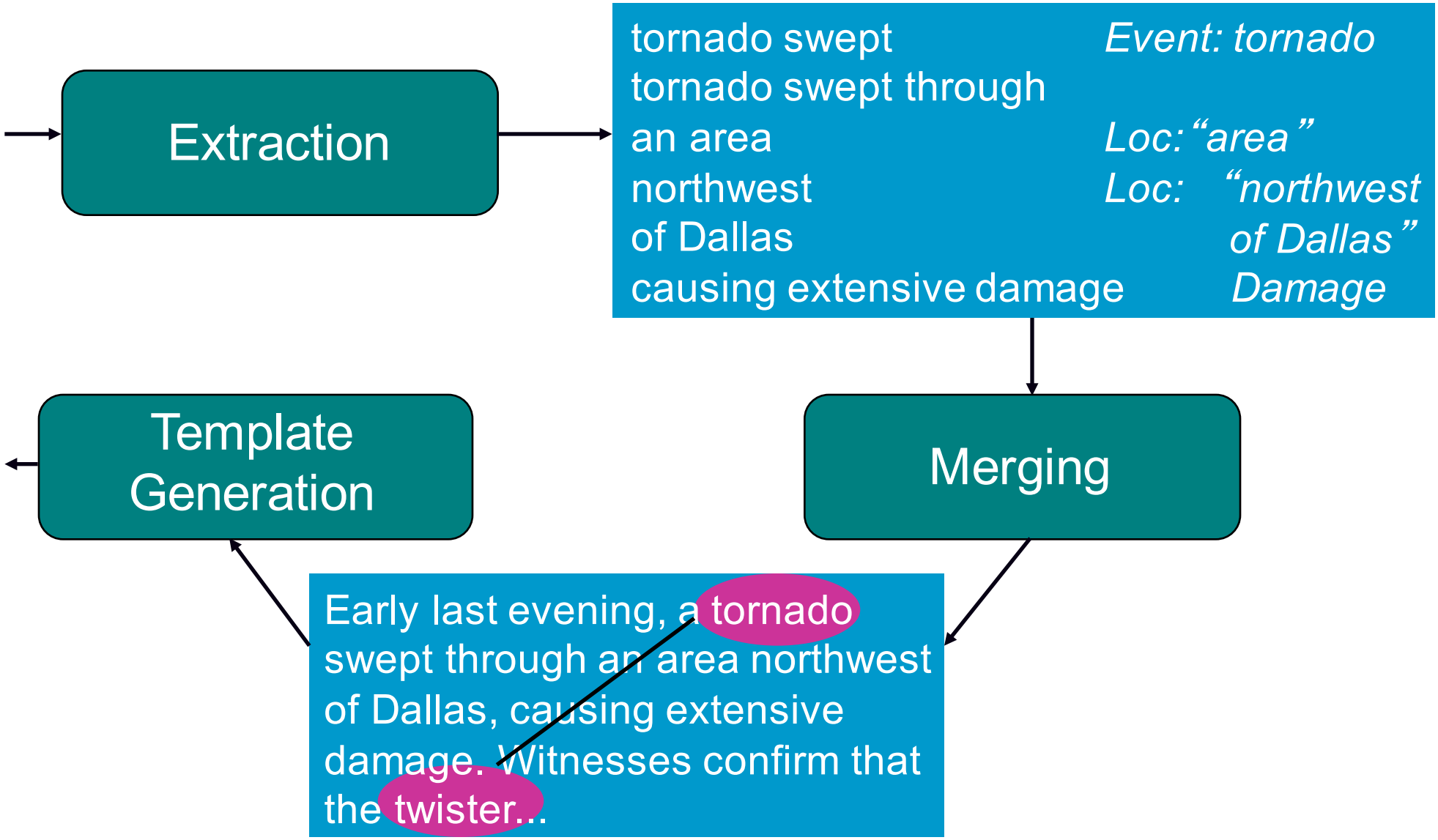
Pre-processing



Learning


Extraction
tornado swept
tornado swept through
an area
northwest
of Dallas
causing extensive damage
Event: tornado
Loc: "area"
Loc: "northwest of Dallas"
Damage

Post-processing



Issues…

- tension between **domain-independent** and **domain-dependent** language processing
  - treating task in a domain-independent way allows the use of general IR/NLP techniques and tools
  - treating task in a domain-dependent way allows for tailoring of techniques for better performance
- IE is generally handled as **domain-specific text understanding**
  - key system components need to be re-built for each new domain
  - difficult and time-consuming to build if constructed manually
    - Initially, ~6-12 months/system for IE from unstructured text
  - requires the expertise of computational linguists

Machine learning methods

- **acquire linguistic knowledge** by applying statistical and symbolic learning methods; derive training examples from the texts themselves
- **automate** the construction of each IE system component
- improve **robustness** of final systems while maintaining (or at least approaching) the accuracies of handcrafted systems

Information Extraction

- **Introduction**
  - Task definition
  - Evaluation
  - IE system architecture
- **Named entity detection**
- **Relation extraction**

NE Identification

- **Identify all named locations, named persons, named organizations, dates, times, monetary amounts, and percentages.**

The delegation, which included the commander of the U.N. troops in Bosnia, Lt. Gen. Sir Michael Rose, went to the Serb stronghold of Pale, near Sarajevo, for talks with Bosnian Serb leader Radovan Karadzic.

Este ha sido el primer comentario publico del presidente Clinton respecto a la crisis de Oriente Medio desde que el secretario de Estado, Warren Christopher, decidiera regresar precipitadamente a Washington para impedir la ruptura del proceso de paz tras la violencia desatada en el sur de Libano.

1. Locations
2. Persons
3. Organizations

Figure 1.1 Examples. Examples of correct labels for English text and for Spanish text.

Guidelines need to be specified

- ***The Wall Street Journal*** : artifact or organization?
- ***White House*** : organization or location?
- Is a street name a location?
- Should ***yesterday*** and ***last Tuesday*** be labeled as dates?
- Is ***mid-morning*** a time?

Examples

1. **MATSUSHITA ELECTRIC INDUSTRIAL CO.** HAS REACHED AGREEMENT …
2. IF ALL GOES WELL, **MATSUSHITA** AND ROBERT BOSCH WILL …
3. **VICTOR CO. OF JAPAN** (**JVC**) AND SONY CORP. …
4. IN A FACTORY OF **BLAUPUNKT WERKE**, A **ROBERT BOSCH** SUBSIDIARY, …
5. **TOUCH PANEL SYSTEMS**, CAPITALIZED AT 50 MILLION YEN, IS OWNED …
6. **MATSUSHITA** EILL DECIDE ON THE PRODUCTION SCALE. …

Figure 2.1 English Examples. Finding names ranges from the easy to the challenging. Company names are in boldface. It is crucial for any name-finder to deal with the underlined text.

ML approaches for NER?

HMMs for NE detection

American	NNP
Airlines	NNPS
,	PUNC
a	DT
unit	NN
of	IN
AMR	NNP
Corp.	NNP
,	PUNC
immediately	RB
matched	VBD
the	DT
move	NN
,	PUNC
spokesman	NN
Tim	NNP
Wagner	NNP
said	VBD
.	PUNC

Figure, copyright J&M 2nd ed

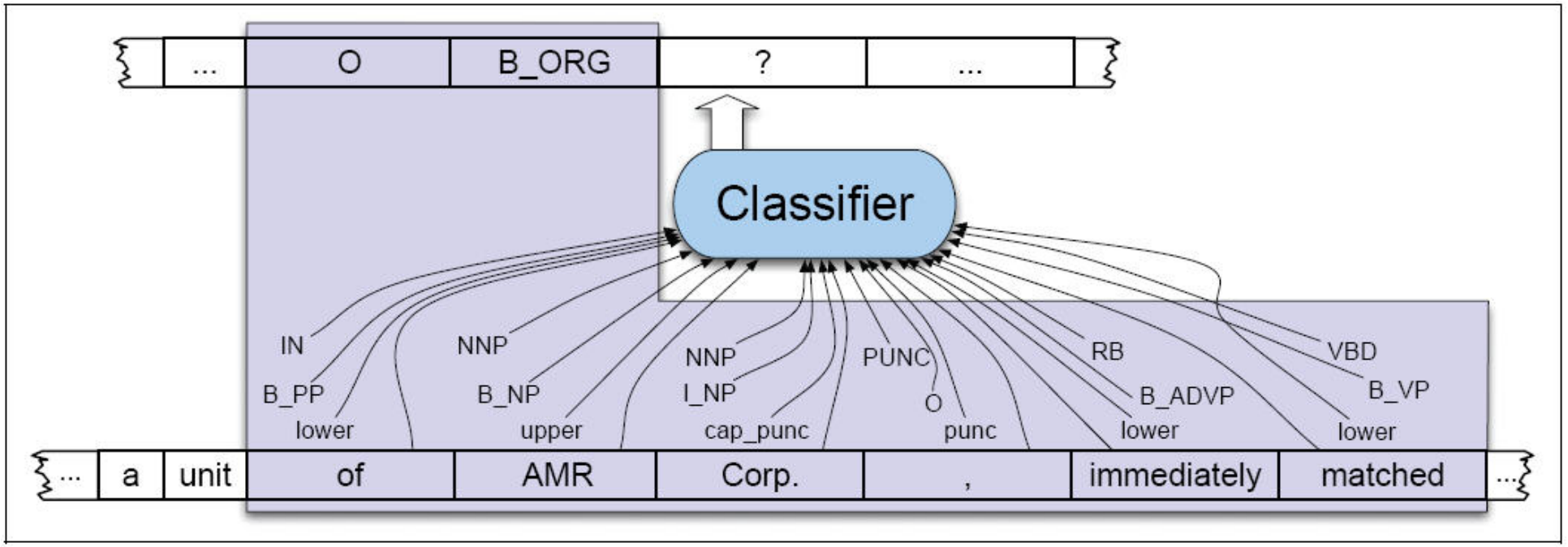
HMMs for NE detection

American	NNP	B_{ORG}
Airlines	NNPS	I_{ORG}
,	PUNC	O
a	DT	O
unit	NN	O
of	IN	O
AMR	NNP	B_{ORG}
Corp.	NNP	I_{ORG}
,	PUNC	O
immediately	RB	O
matched	VBD	O
the	DT	O
move	NN	O
,	PUNC	O
spokesman	NN	O
Tim	NNP	B_{PER}
Wagner	NNP	I_{PER}
said	VBD	O
.	PUNC	O

Figure, copyright J&M 2nd ed

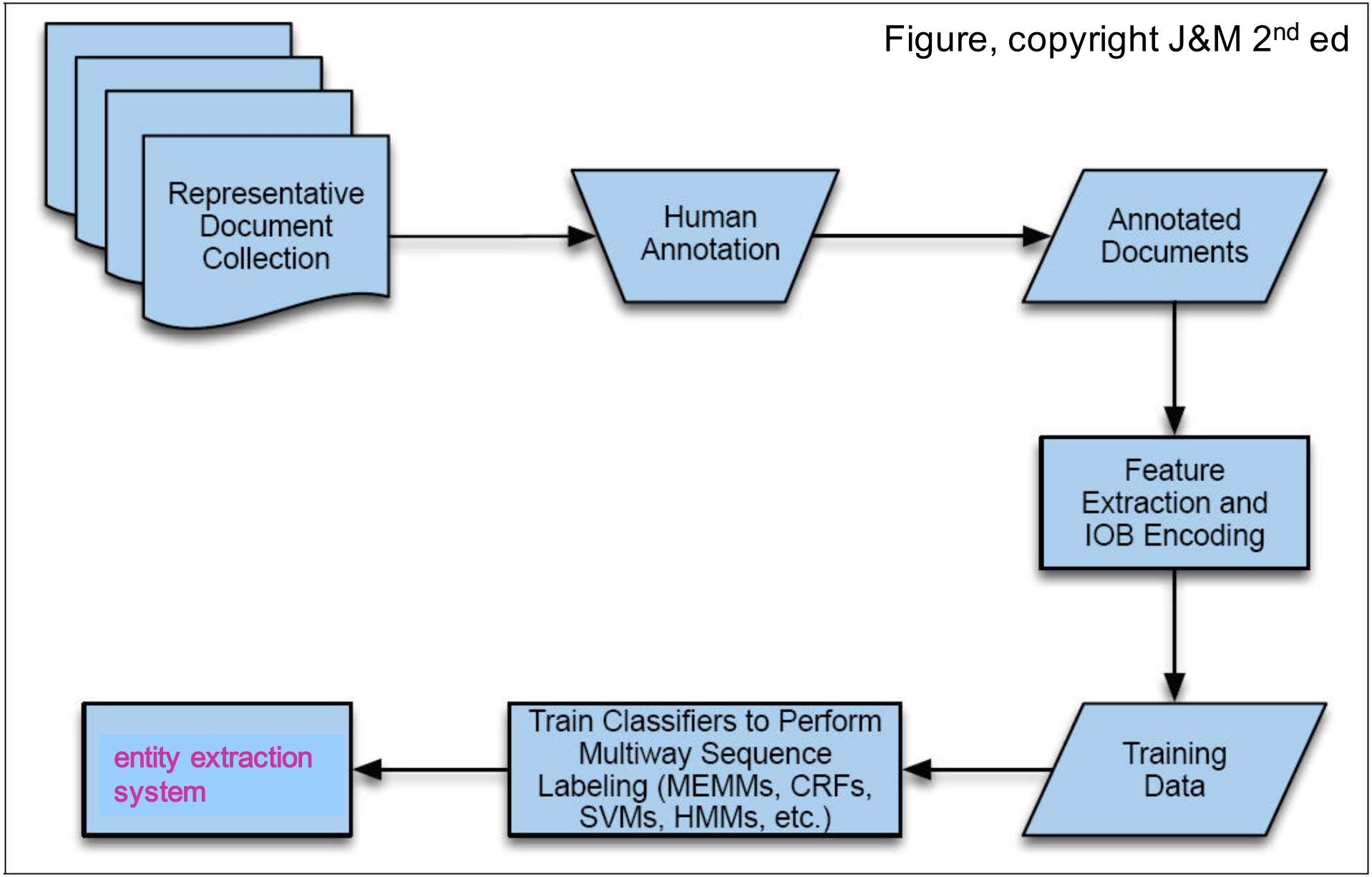
Window-based Classification

- **Fixed-size moving window**
- **Classify the target token as one of IOB**



Figure, copyright J&M 2nd ed

End-to-end process



NE Results Using HMM's

Table 5.1 F-measure Scores. This table illustrates IdentiFinder's performance as compared to the best reported scores for each category.

	Language	Best Rules	IdentiFinder
Mixed Case	English (WSJ)	96.4	94.9
Upper Case	English (WSJ)	89	93.6
Speech Form	English (WSJ)	74	90.7
Mixed Case	Spanish	93	90