Information Extraction

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

specifying the extraction task for a new domain

Today

Specifying the Extraction Task

- **Define the domain**
- **Slots/components in the output template**
  - String fill?
  - Set fill?
  - Normalization?
  - One/multiple fills?
  - Cross-referencing with other slots?
- **Develop manual annotation instructions**

IE system: natural disasters

Disaster Type: earthquake

- location: *Afghanistan*
- date: *today*
- magnitude: *6.9*
- epicenter: *a remote part of the country*
- damage:
    - human-effect:
        - victim: *Thousands of people*
        - number: *Thousands*
        - outcome: dead
    - physical-effect:
        - object: *entire villages*
        - outcome: damaged

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

Changes in Management

Evergreen Information said Barry Nelsen, who had a heart-bypass operation last week, resigned as president and chief executive. The board formally accepted the resignation of Thomas Casey, its former chairman, who stepped down effective Feb. 2.

Martin Bell was named president, CEO, and chairman. Mr. Bell -- who has been chief financial officer since the fall -- also got voting control of 970,000 shares held by the Evergreen Partnership, a vehicle for the company's three co-founders, including Mr. Nelsen.

Excluding these shares, Evergreen Information has more than two million shares or exercisable warrants outstanding, according to a spokeswoman.

The computer products and services concern has cut its staff to fewer than 10 employees from about 35, and has deferred and reduced managers' salaries. In a press release, it said it believes the company is still viable.

Evergreen Information said Barry Nelsen, who had a heart-bypass operation last week, resigned as president and chief executive. The board formally accepted the resignation of Thomas Casey, its former chairman, who stepped down effective Feb. 2.

Martin Bell was named president, CEO, and chairman. Mr. Bell -- who has been chief financial officer since the fall -- also got voting control of 970,000 shares held by the Evergreen Partnership, a vehicle for the company's three co-founders, including Mr. Nelsen.

Excluding these shares, Evergreen Information has more than two million shares or exercisable warrants outstanding, according to a spokeswoman.

The computer products and services concern has cut its staff to fewer than 10 employees from about 35, and has deferred and reduced managers' salaries. In a press release, it said it believes the company is still viable.

```
<TEMPLATE-9303020074-1> :=
   DOC_NR: "9303020074"
   CONTENT: <SUCCESSION_EVENT-9303020074-1>
          <SUCCESSION_EVENT-9303020074-2>
          <SUCCESSION_EVENT-9303020074-3>
          <SUCCESSION_EVENT-9303020074-4>
<SUCCESSION_EVENT-9303020074-1> :=
   SUCCESSION_ORG: <ORGANIZATION-9303020074-1>
   POST: "president"
   IN_AND_OUT: <IN_AND_OUT-9303020074-1>
            <IN_AND_OUT-9303020074-2>
   VACANCY_REASON: REASSIGNMENT
   COMMENT: "Nelson out, Bell in as pres of Evergreen Info"
        / "This event could be collapsed with SUCCESSION_EVENT-2"
```

```
<SUCCESSION_EVENT-9303020074-2> :=
   SUCCESSION_ORG:  <ORGANIZATION-9303020074-1>
   POST: "chief executive" / "CEO"
   IN_AND_OUT: <IN_AND_OUT-9303020074-3>
            <IN_AND_OUT-9303020074-4>
   VACANCY_REASON:  REASSIGNMENT
   COMMENT: "Nelson out, Bell in as CEO of Evergreen Info"
<SUCCESSION_EVENT-9303020074-3> :=
   SUCCESSION_ORG:  <ORGANIZATION-9303020074-1>
   POST: "chairman"
   IN_AND_OUT: <IN_AND_OUT-9303020074-5>
            <IN_AND_OUT-9303020074-6>
   VACANCY_REASON:  REASSIGNMENT
   COMMENT: "Casey out, Bell in as chmn of Evergreen Info"
<SUCCESSION_EVENT-9303020074-4> :=
   SUCCESSION_ORG:  <ORGANIZATION-9303020074-1>
   POST: "chief financial officer"
   IN_AND_OUT: <IN_AND_OUT-9303020074-7>
   VACANCY_REASON:  OTH_UNK
   COMMENT: "Bell in as CFO at Evergreen Info 'since the fall'"
```

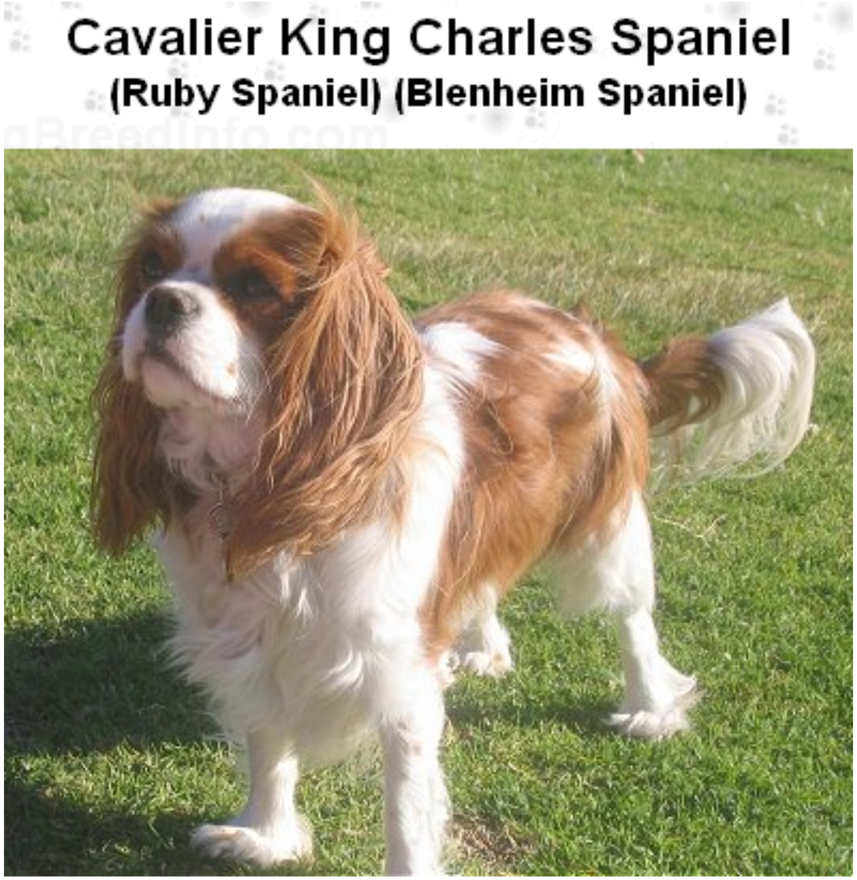
```
<IN_AND_OUT-9303020074-1> :=
   IO_PERSON: <PERSON-9303020074-1>
   NEW_STATUS: OUT
   ON_THE_JOB: UNCLEAR
   COMMENT: "Nelson out as pres"
        / "ON_THE_JOB: 'resign' (headline), 'resigned'"
<IN_AND_OUT-9303020074-2> :=
   IO_PERSON: <PERSON-9303020074-3>
   NEW_STATUS: IN
   ON_THE_JOB: UNCLEAR
   OTHER_ORG: <ORGANIZATION-9303020074-1>
   REL_OTHER_ORG: SAME_ORG
   COMMENT: "Bell in as pres -- was already CFO at same org"
        / "ON_THE_JOB: 'was named'"
<IN_AND_OUT-9303020074-3> :=
   IO_PERSON: <PERSON-9303020074-1>
   NEW_STATUS: OUT
   ON_THE_JOB: UNCLEAR
   COMMENT: "Nelson out as CEO"
        / "This obj identical to IN_AND_OUT-1"
```

```
<IN_AND_OUT-9303020074-4> :=
   IO_PERSON: <PERSON-9303020074-3>
   NEW_STATUS: IN
   ON_THE_JOB: UNCLEAR
   OTHER_ORG: <ORGANIZATION-9303020074-1>
   REL_OTHER_ORG: SAME_ORG
   COMMENT: "Bell in as CEO"
         / "This obj identical to IN_AND_OUT-2"
<IN_AND_OUT-9303020074-5> :=
   IO_PERSON: <PERSON-9303020074-2>
   NEW_STATUS: OUT
   ON_THE_JOB: NO
   COMMENT: "Casey out"
         / "ON_THE_JOB: 'stepped down effective Feb. 2'"
<IN_AND_OUT-9303020074-6> :=
   IO_PERSON: <PERSON-9303020074-3>
   NEW_STATUS: IN
   ON_THE_JOB: UNCLEAR
   OTHER_ORG: <ORGANIZATION-9303020074-1>
   REL_OTHER_ORG: SAME_ORG
   COMMENT: "Bell in as chmn"
         / "This obj identical to IN_AND_OUT-2"
```

```
<IN_AND_OUT-9303020074-7> :=
    IO_PERSON: <PERSON-9303020074-3>
    NEW_STATUS: IN
    ON_THE_JOB: YES
    COMMENT: "Bell in"
           / "ON_THE_JOB: has been CFO 'since the fall'"
<ORGANIZATION-9303020074-1> :=
    ORG_NAME: "Evergreen Information Technologies Inc."
    ORG_ALIAS: "Evergreen Information Technologies"
               "Evergreen"
               "Evergreen Information"
    ORG_DESCRIPTOR: "The computer products and services concern"
    ORG_TYPE: COMPANY
    ORG_LOCALE: McLean CITY
    ORG_COUNTRY: United States
```

```
<PERSON-9303020074-1> :=
   PER_NAME: "Barry Nelsen"
   PER_ALIAS: "Nelsen"
   PER_TITLE: "Mr."
<PERSON-9303020074-2> :=
   PER_NAME: "Thomas Casey"
<PERSON-9303020074-3> :=
   PER_NAME: "Martin Bell"
   PER_ALIAS: "Bell"
   PER_TITLE: "Mr."
```

IE: dogs


Cavalier King Charles Spaniel
(Ruby Spaniel) (Blenheim Spaniel)
gBreedInfo.com

Height: 12-13 inches (30-33 cm.)
Weight: 10-18 pounds (5-8 kg.)

Prone to syringomyelia, hereditary eye disease, dislocating kneecaps (patella), back troubles, ear infections, early onset of deafness or hearing trouble. Sometime's hip dysplasia. Don't over feed. This breed tends to gain weight easily. Some lines are genetically disposed early onset to a serious heart problem, which sometimes causes early death. When selecting one of these dogs, it is extremely important to check the medical history of several previous generations.

Cavalier King Charles Spaniels are good for apartment life. They are moderately active indoors and a small yard will be sufficient. The Cavalier does not do well in very warm conditions.

Cavalier King Charles Spaniels need a daily walk. Play will take care of a lot of their exercise needs, however, as with all breeds, play will not fulfill their primal instinct to walk. Dogs who do not get to go on daily walks are more likely to display behavior problems. They will also enjoy a good romp in a safe open area off lead, such as a large fenced in yard.

Instead…ended up with Marseille

…a purebred mutt

- American Eskimo Dog

- + Cocker Spaniel

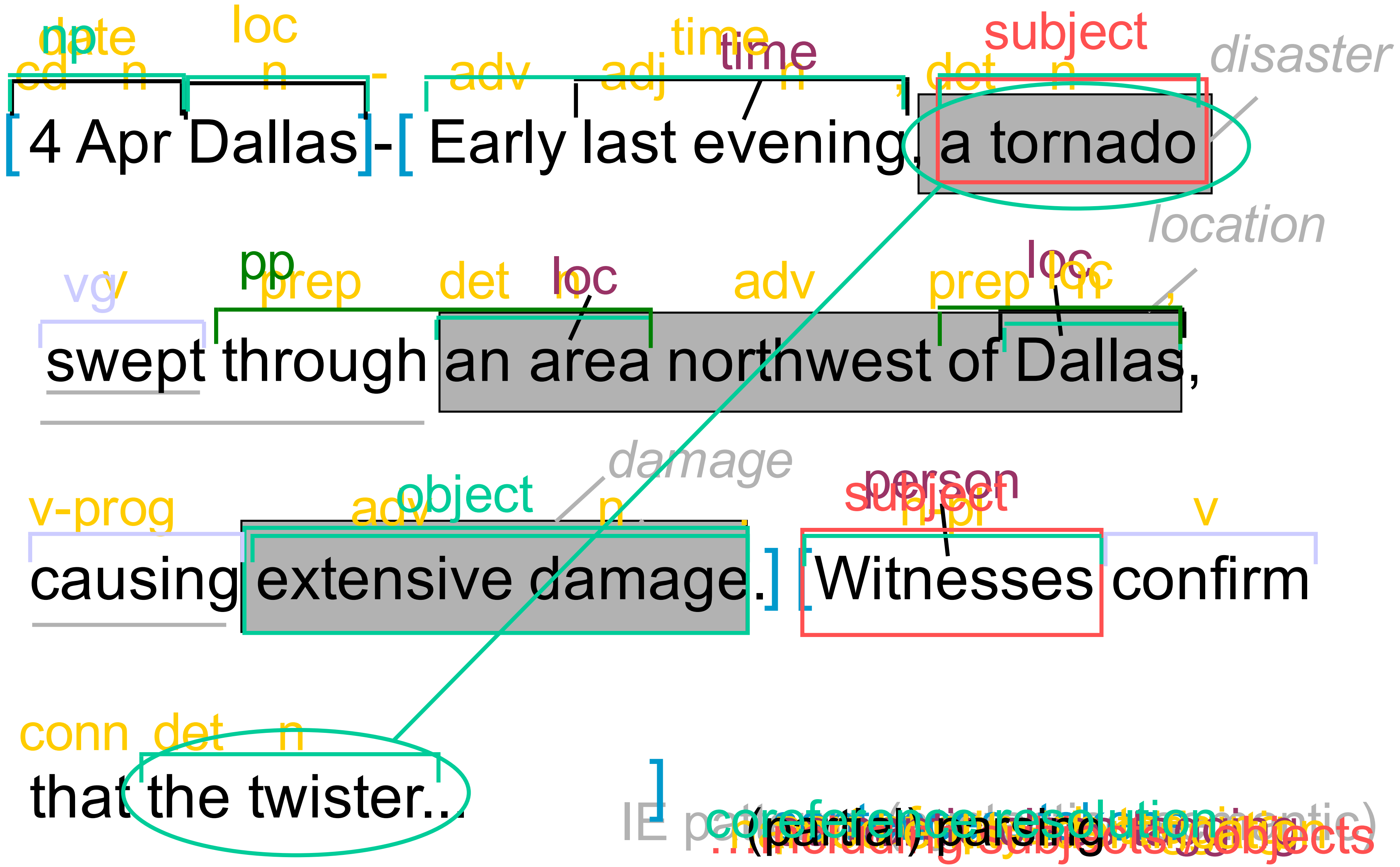
Information Extraction

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

specifying the extraction task
for a new domain

Today

IE system components


date
np
loc
cd
n
n
-
adv
adj
time
time
n
,
det
n
subject
disaster
[4 Apr Dallas] - [Early last evening, a tornado
location
vg
v
pp
prep
det
loc
n
adv
prep
loc
loc
n
,
swept through an area northwest of Dallas,
damage
v-prog
adv
object
n
.
person
subject
n-pl
v
causing extensive damage.] [Witnesses confirm
conn
det
n
that the twister...
]
IE pattern (partial, semantic)
coreference resolution
(partial) parsing
...including subjects & objects

Issues…

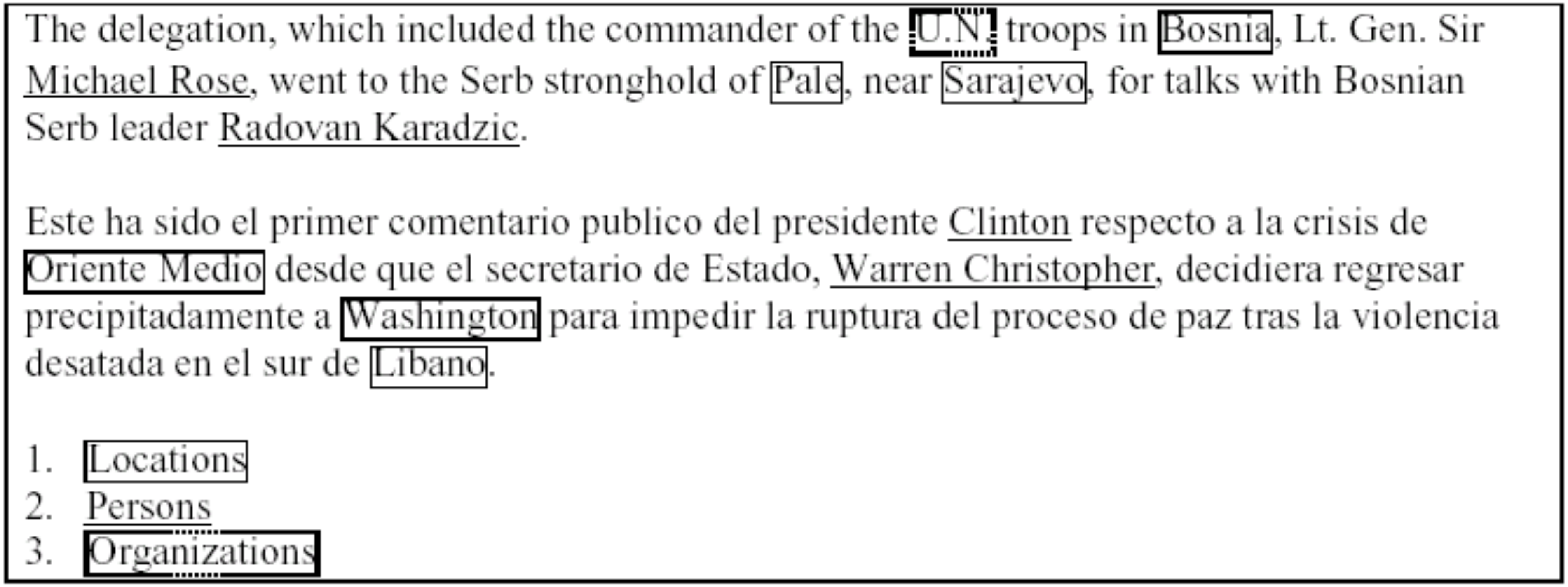
- tension between **domain-independent** and **domain-dependent** language processing
  - treating task in a domain-independent way allows the use of general IR/NLP techniques and tools
  - treating task in a domain-dependent way allows for tailoring of techniques for better performance
- IE is generally 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 months/system for IE from unstructured text
  - requires the expertise of computational linguists

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.

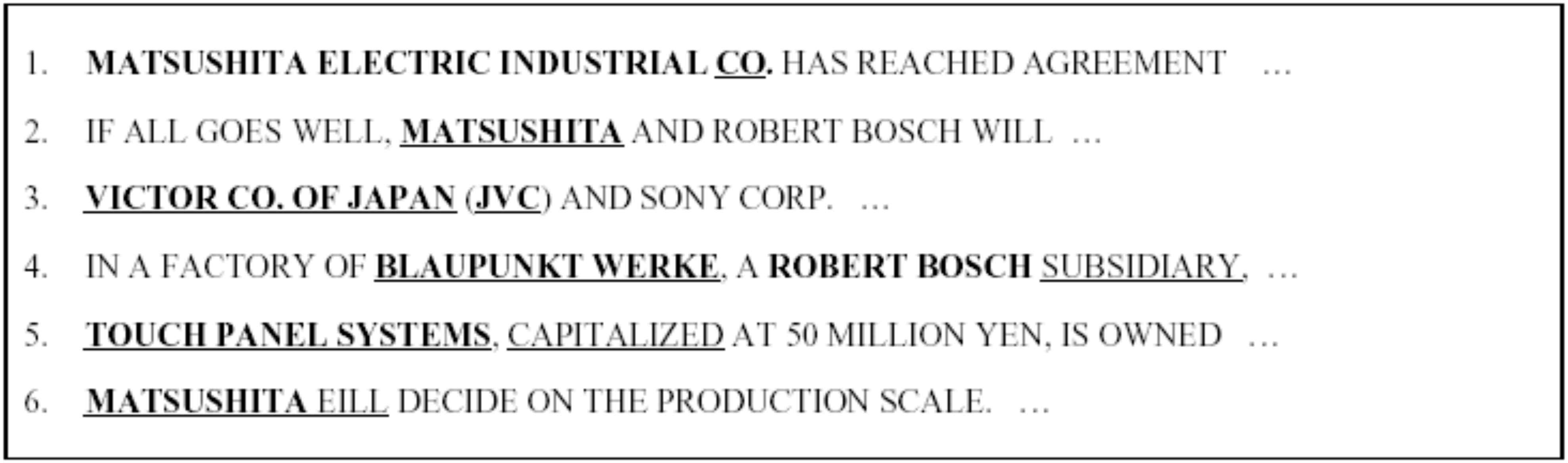
1. Locations
2. Persons
3. Organizations

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

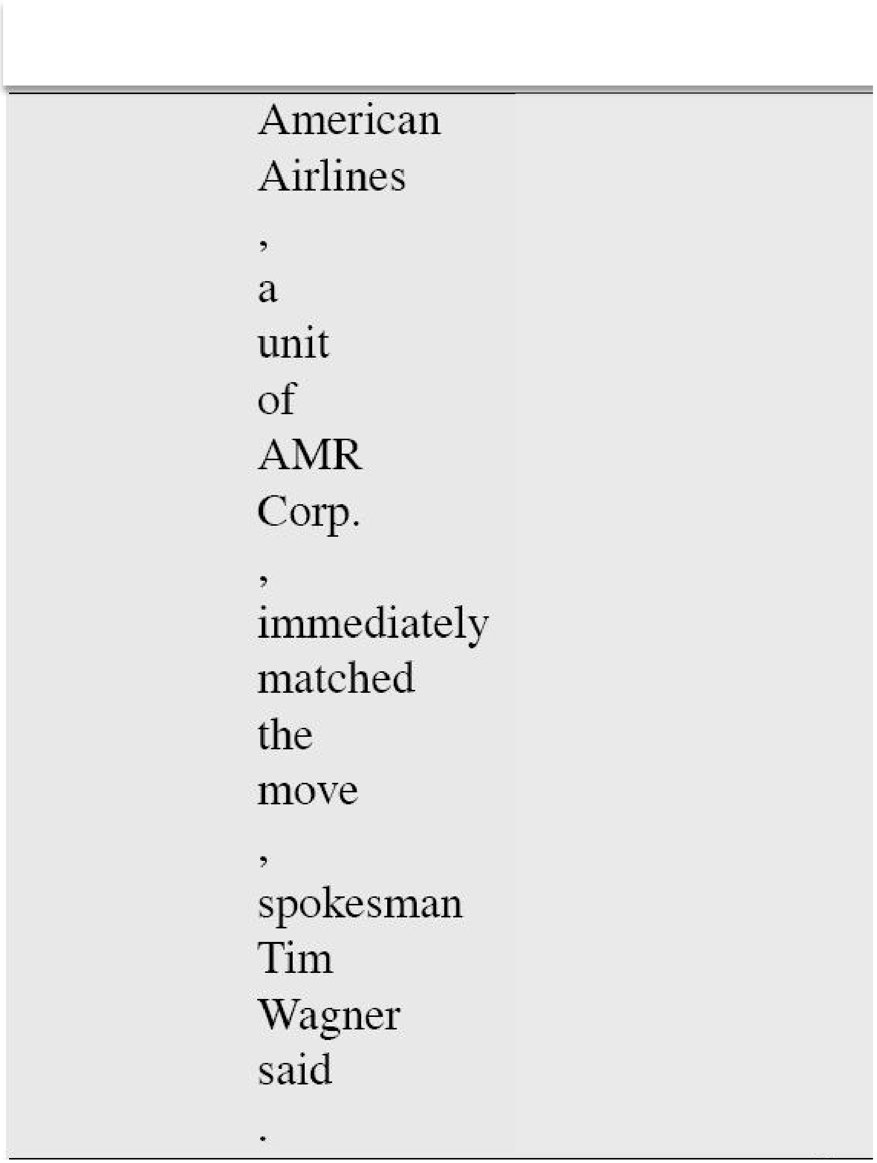


1. **MATSUSHITA ELECTRIC INDUSTRIAL CO.** HAS REACHED AGREEMENT …
2. IF ALL GOES WELL, **MATSUSHITA** AND ROBERT BOSCH WILL …
3. **VICTOR CO. OF JAPAN** (**JVC**) AND SONY CORP. …
4. IN A FACTORY OF **BLAUPUNKT WERKE**, A **ROBERT BOSCH** SUBSIDIARY, …
5. **TOUCH PANEL SYSTEMS**, CAPITALIZED AT 50 MILLION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

Clickers

- **How many named entities?**
  A. 0
  B. 1
  C. 2
  D. 3
  E. >3



American
Airlines
,
a
unit
of
AMR
Corp.
,
immediately
matched
the
move
,
spokesman
Tim
Wagner
said
.

ML approaches for NER?

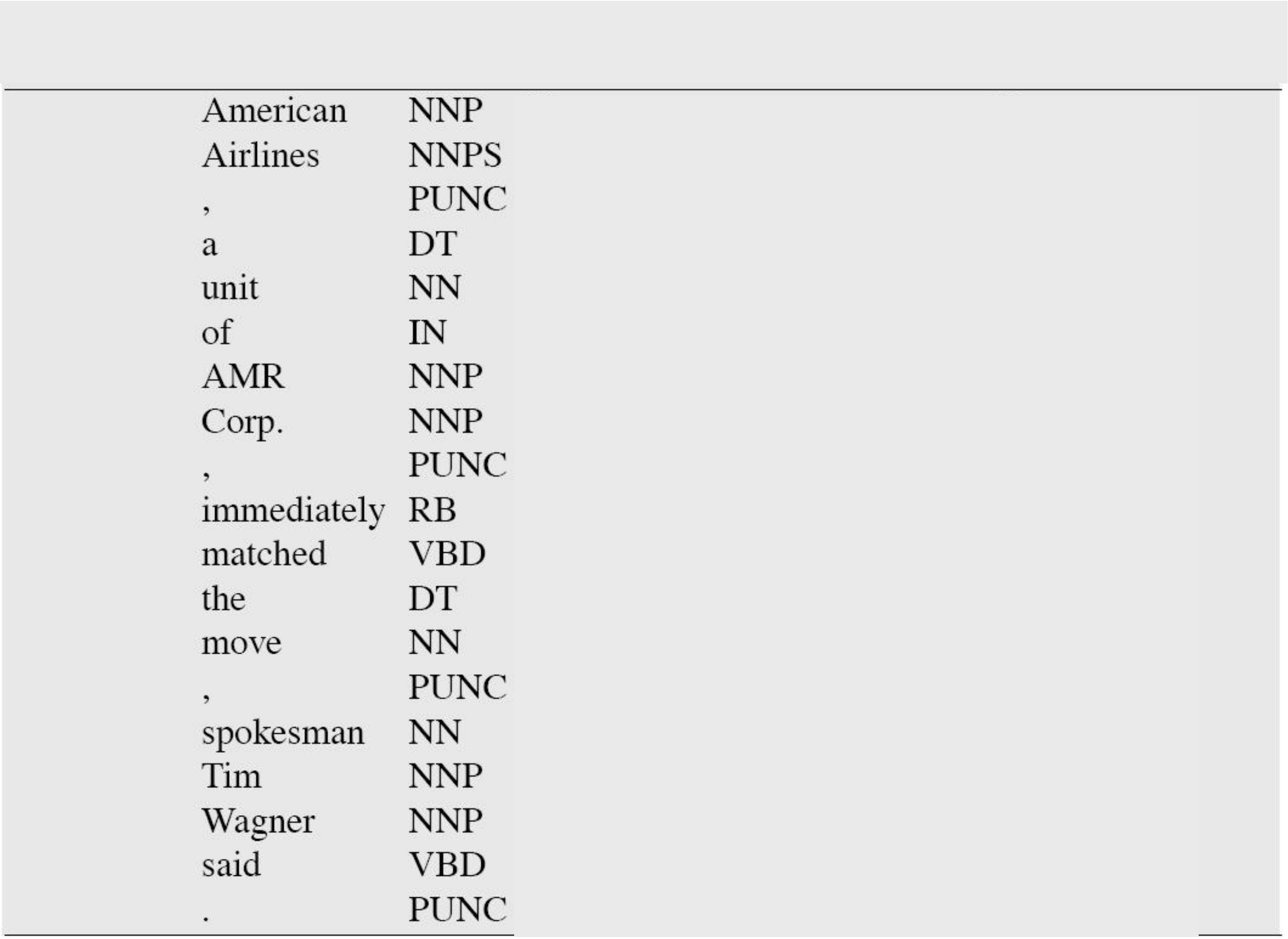
A. Classification (e.g. NB, SVMs)

B. Sequence tagging (e.g. HMMs)

C. Both A and B

D. Neither A nor B

HMMs for NE detection



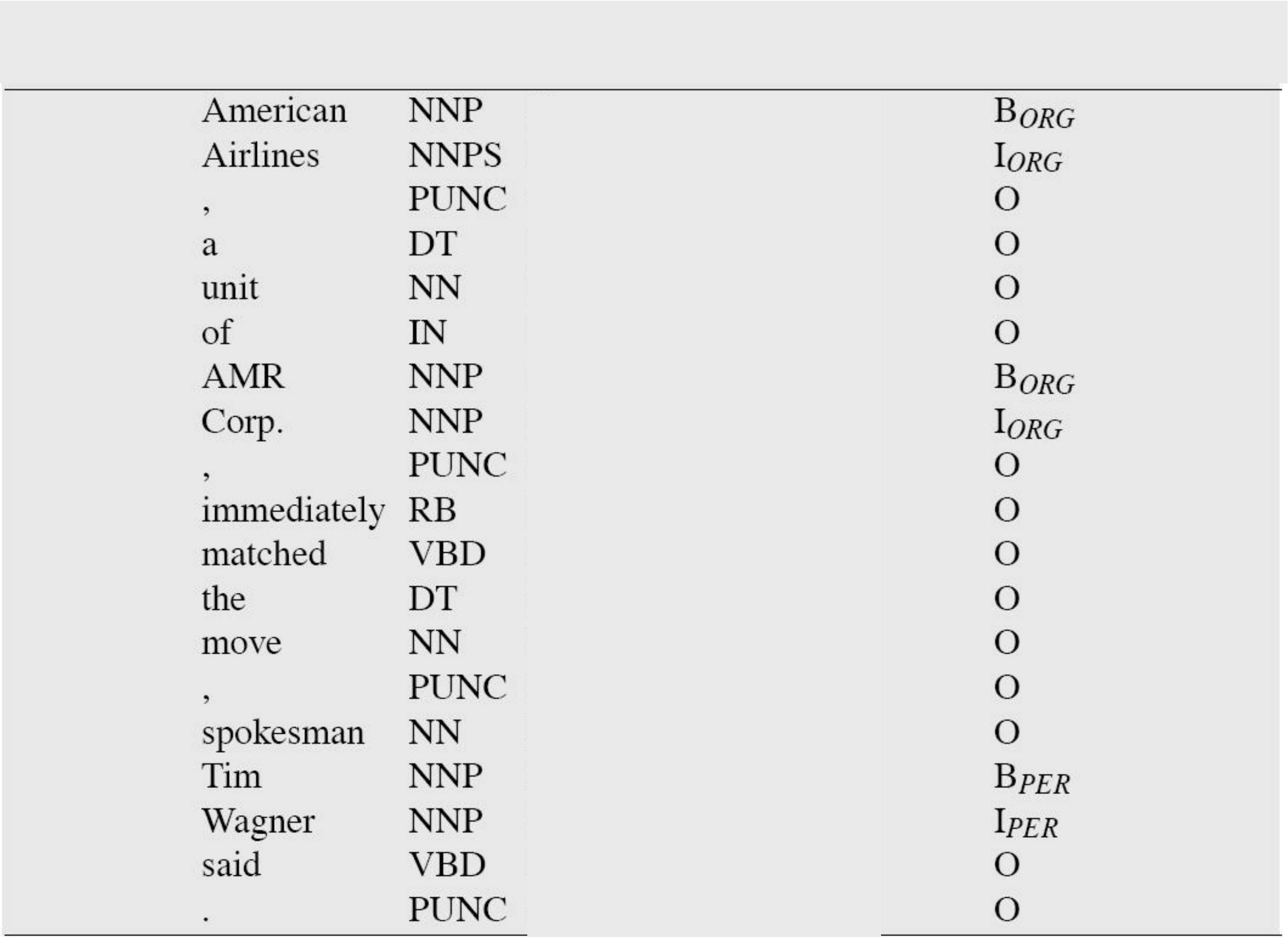
American	NNP
Airlines	NNPS
,	PUNC
a	DT
unit	NN
of	IN
AMR	NNP
Corp.	NNP
,	PUNC
immediately	RB
matched	VBD
the	DT
move	NN
,	PUNC
spokesman	NN
Tim	NNP
Wagner	NNP
said	VBD
.	PUNC

Figure, copyright J&M 2nd ed

Tag Set for NER

- **IOB tags**
  - B-xxx
    - First (i.e. Beginning) token in a NE of type xxx
  - I-xxx
    - Inside an entity of type xxx
  - O
    - Outside all NEs

HMMs for NE detection

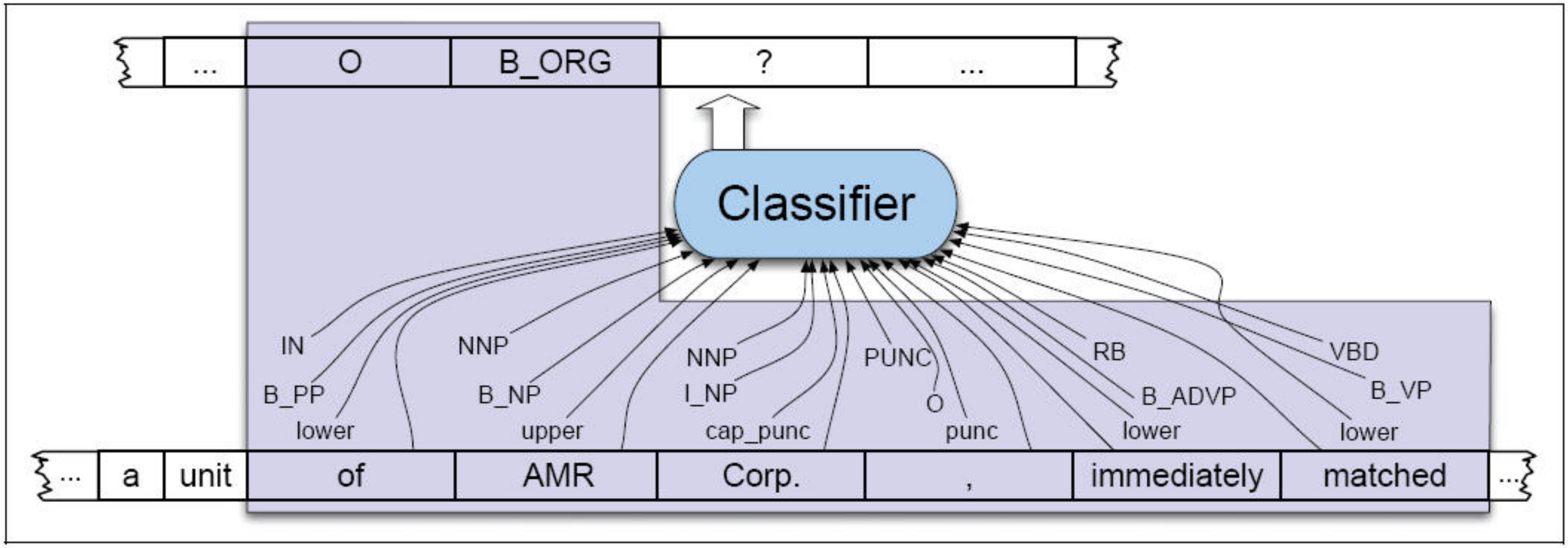


American	NNP	B_{ORG}
Airlines	NNPS	I_{ORG}
,	PUNC	O
a	DT	O
unit	NN	O
of	IN	O
AMR	NNP	B_{ORG}
Corp.	NNP	I_{ORG}
,	PUNC	O
immediately	RB	O
matched	VBD	O
the	DT	O
move	NN	O
,	PUNC	O
spokesman	NN	O
Tim	NNP	B_{PER}
Wagner	NNP	I_{PER}
said	VBD	O
.	PUNC	O

Figure, copyright J&M 2nd ed

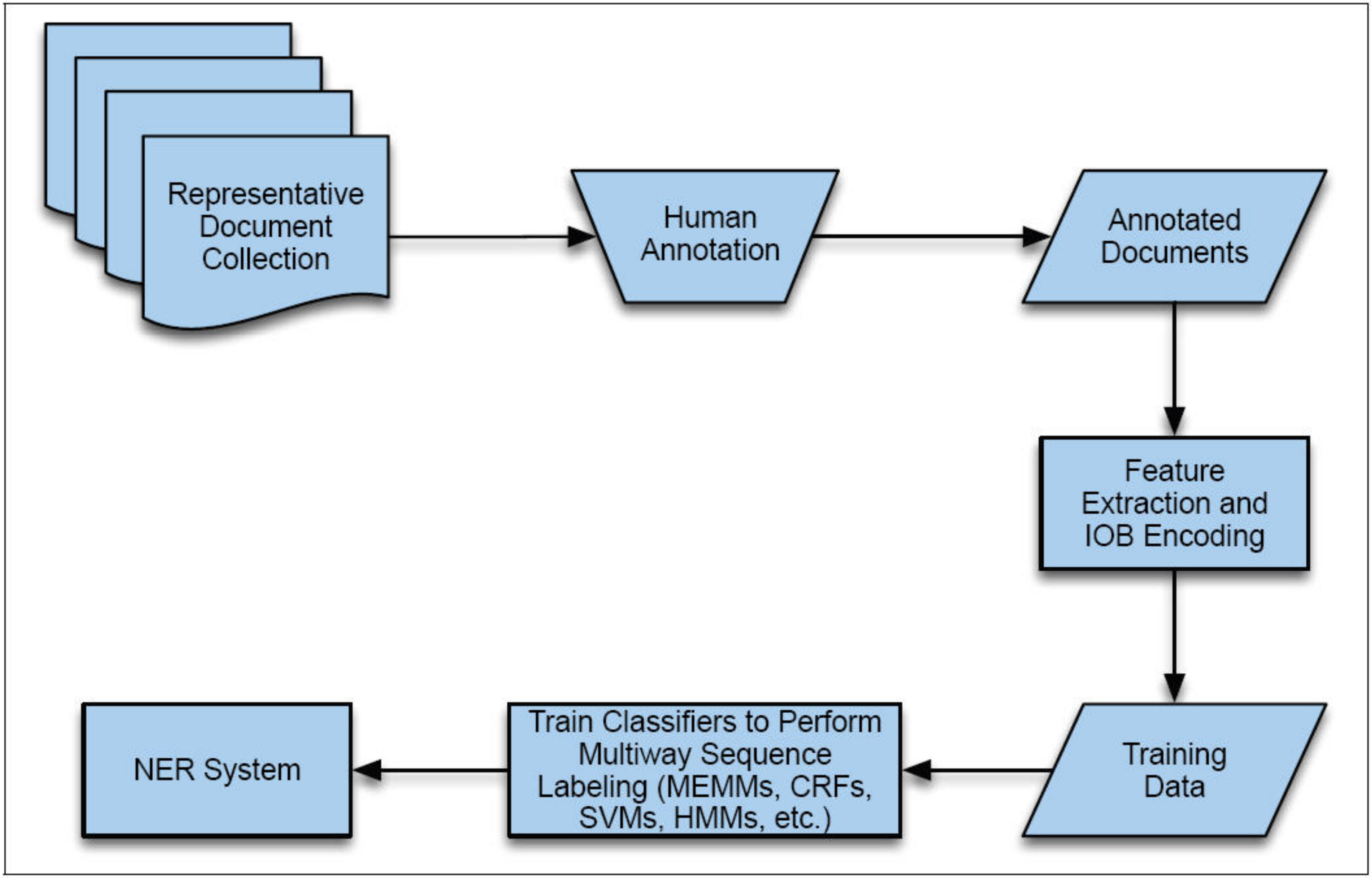
Window-based Classification

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



Figure, copyright J&M 2nd ed

End-to-end process


Representative Document Collection
Human Annotation
Annotated Documents
Feature Extraction and IOB Encoding
Training Data
Train Classifiers to Perform Multiway Sequence Labeling (MEMMs, CRFs, SVMs, HMMs, etc.)
NER System

NE Results Using HMM’s

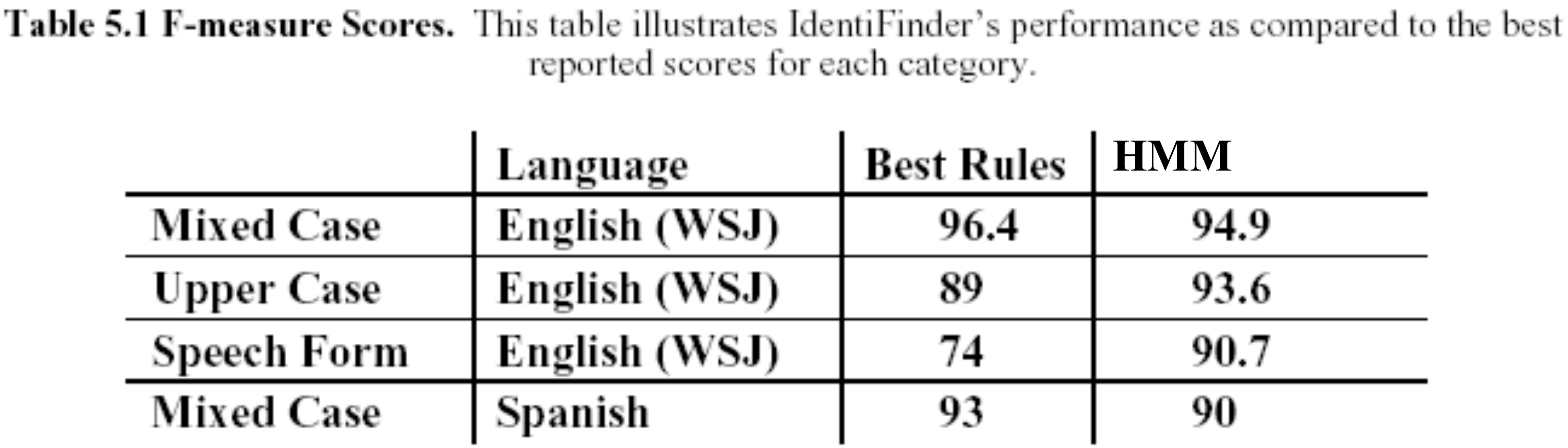


Table 5.1 F-measure Scores. This table illustrates IdentiFinder’s performance as compared to the best reported scores for each category.

	Language	Best Rules	HMM
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