



Applied Natural Language Processing

Info 256

Lecture 19: Named entity recognition (April 4, 2019)

David Bamman, UC Berkeley

Named entity recognition

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 [PER Tim Wagner] said. [ORG United], a 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].

Named entity recognition

[tim cook]PER is the ceo of [apple]ORG

- Identifying spans of text that correspond to typed entities that are proper names.

Named entity recognition

| 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 Golden Gate Bridge . |
| Vehicles | VEH | planes, trains, automobiles | It was a classic Ford Falcon . |

Figure 17.1 A list of generic named entity types with the kinds of entities they refer to.

ACE NER categories (+weapon)

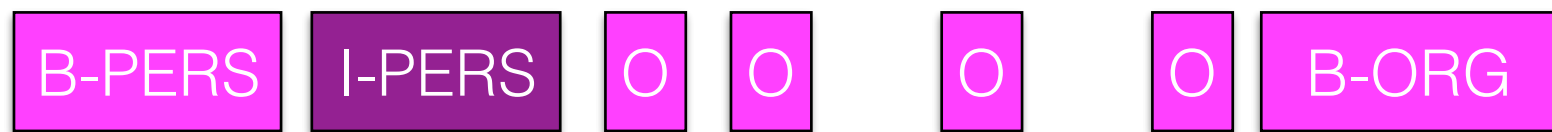
Named entity recognition

- GENIA corpus of MEDLINE abstracts (biomedical)

We have shown that [interleukin-1]_{PROTEIN} ([IL-1]_{PROTEIN}) and [IL-2]_{PROTEIN} control [IL-2 receptor alpha (IL-2R alpha) gene]_{DNA} transcription in [CD4-CD8- murine T lymphocyte precursors]_{CELL LINE}

| |
|-----------|
| protein |
| cell line |
| cell type |
| DNA |
| RNA |

BIO notation



tim cook is the ceo of apple

- **B**eginning of entity
- **I**nside entity
- **O**utside entity

[tim cook]_{PER} is the ceo of [apple]_{ORG}

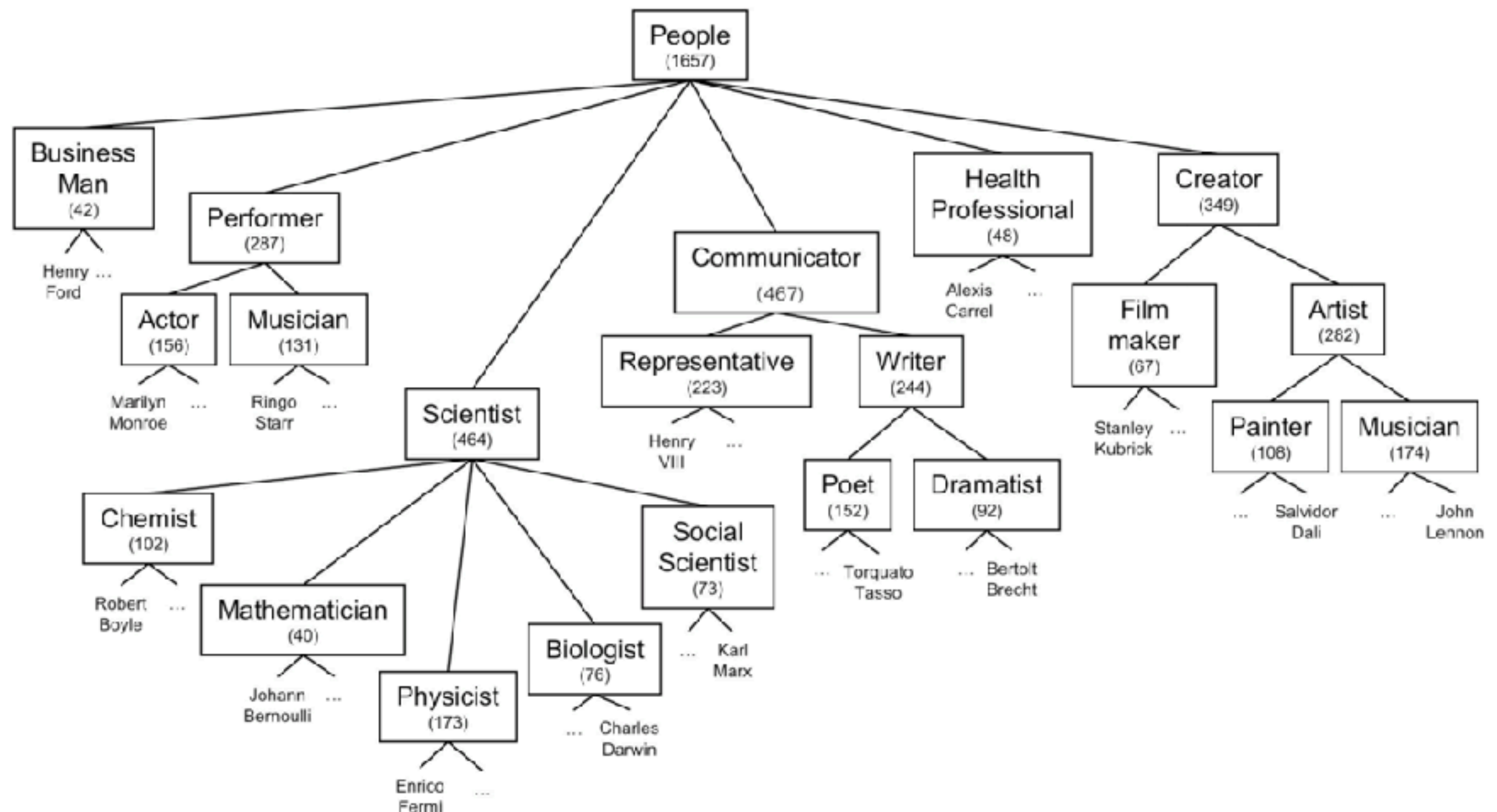
Named entity recognition

B-PER

B-PER

After he saw Harry Tom went to the store

Fine-grained NER



Fine-grained NER

WordNet Search - 3.1

- [WordNet home page](#) - [Glossary](#) - [Help](#)

Word to search for:

Display Options:

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations
Display options for sense: (gloss) "an example sentence"

Noun

- [S: \(n\) Brecht](#), **Bertolt Brecht** (German dramatist and poet who developed a style of epic theater (1898–1956))
 - [instance](#)
 - [S: \(n\) dramatist](#), [playwright](#) (someone who writes plays)
 - [S: \(n\) poet](#) (a writer of poems (the term is usually reserved for writers of good poetry))

Entity recognition

| | |
|--------------|---|
| Person | ... named after [the daughter of a Mattel co-founder] ... |
| Organization | [The Russian navy] said the submarine was equipped with 24 missiles |
| Location | Fresh snow across [the upper Midwest] on Monday, closing schools |
| GPE | The [Russian] navy said the submarine was equipped with 24 missiles |
| Facility | Fresh snow across the upper Midwest on Monday, closing [schools] |
| Vehicle | The Russian navy said [the submarine] was equipped with 24 missiles |
| Weapon | The Russian navy said the submarine was equipped with [24 missiles] |

ACE entity categories

<https://www ldc.upenn.edu/sites/www ldc.upenn.edu/files/english-entities-guidelines-v6.6.pdf>

Named entity recognition

- Most **named** entity recognition datasets have flat structure (i.e., non-hierarchical labels).
 - ✓ [The University of California]**ORG**
 - ✗ [The University of [California]**GPE**]**ORG**
- Mostly fine for **named** entities, but more problematic for general entities:

[[John]**PER**'s mother]**PER** said ...

Nested NER

| | | | | | | | |
|-------|-------|-------|----------|-------|-------|--------|------------|
| named | after | the | daughter | of | a | Mattel | co-founder |
| B-ORG | | | | | | | |
| | | | | | B-PER | I-PER | I-PER |
| B-PER | | I-PER | I-PER | I-PER | I-PER | I-PER | I-PER |

Sequence labeling

$$x = \{x_1, \dots, x_n\}$$

$$y = \{y_1, \dots, y_n\}$$

- For a set of inputs x with n sequential time steps, one corresponding label y_i for each x_i
- Model correlations in the labels y .

Maximum Entropy Markov Model (MEMM)

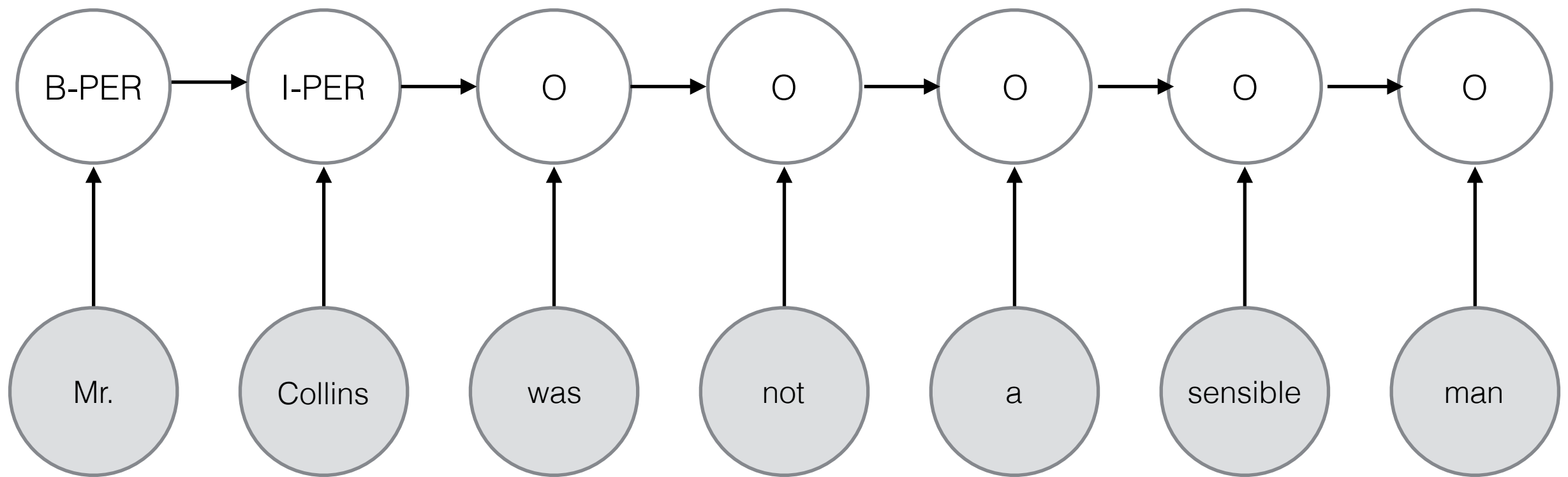
General maxent form

$$\arg \max_y P(y \mid x, \beta)$$

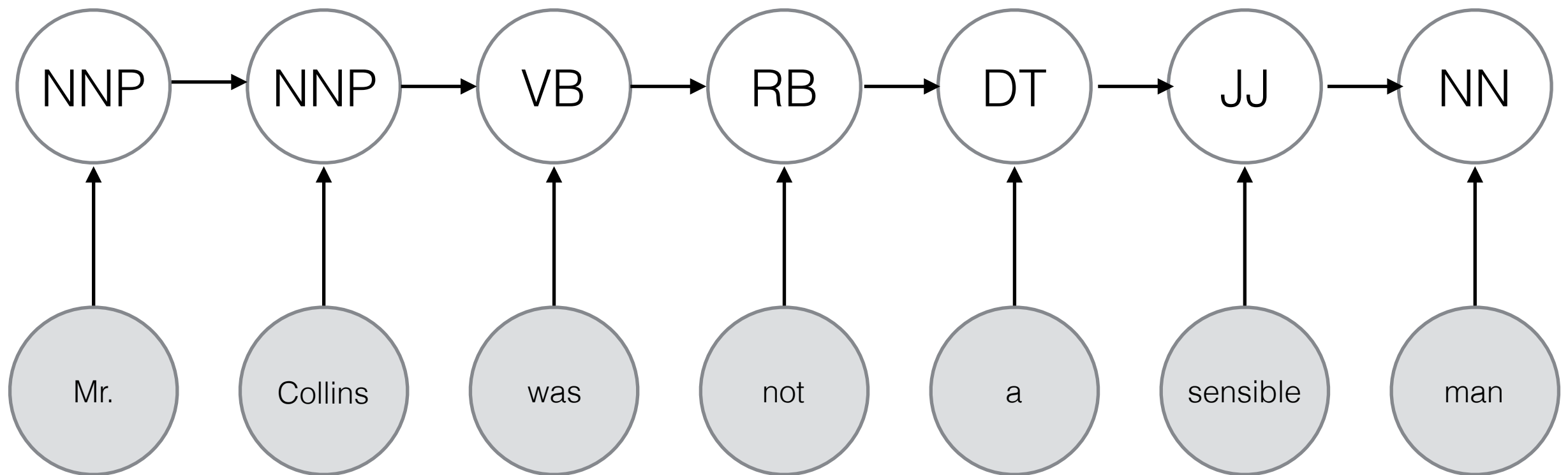
Maxent with Markov
assumption: Maximum Entropy
Markov Model

$$\arg \max_y \prod_{i=1}^n P(y_i \mid y_{i-1}, x)$$

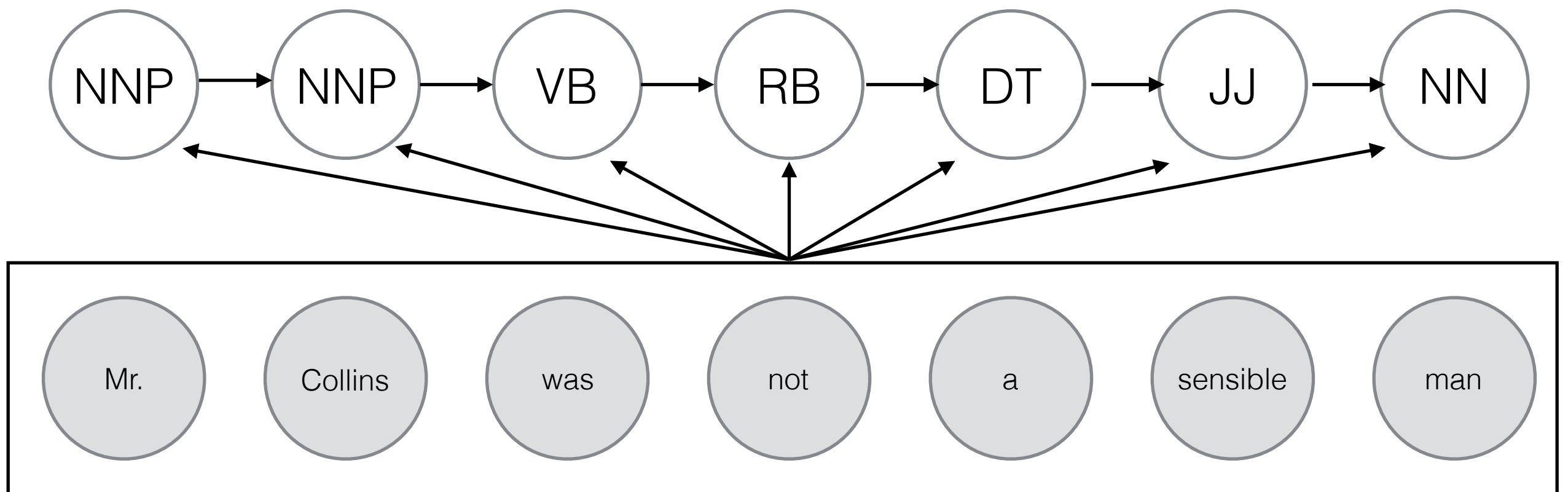
MEMM



MEMM

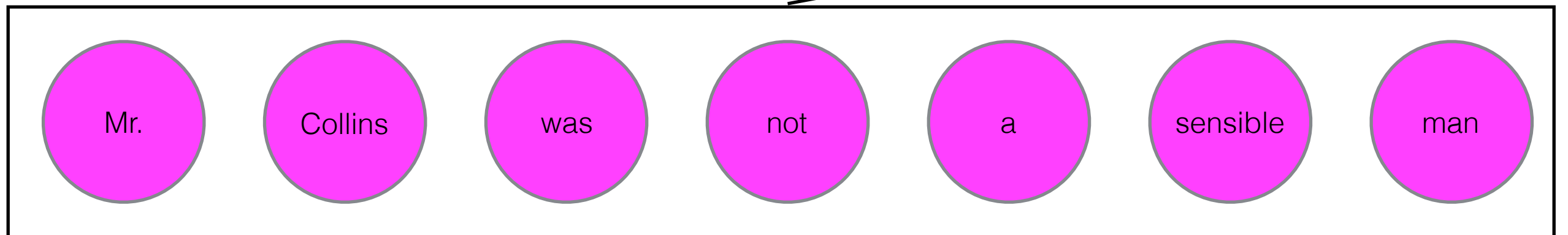
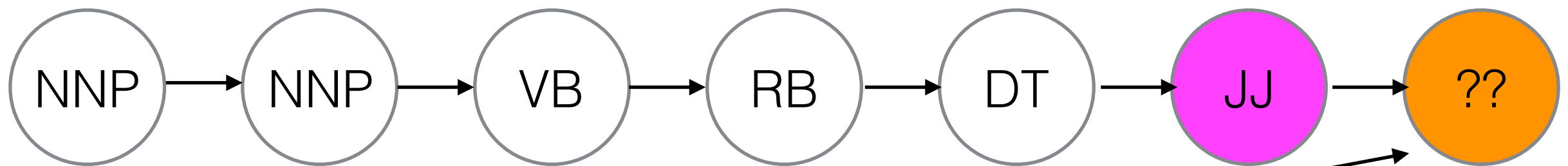


MEMM



MEMMs condition on the *entire* input

MEMM



Features

$$f(y_i, y_{i-1}; x_1, \dots, x_n)$$

Features are scoped over
the previous predicted
tag and the entire
observed input

| feature | example |
|-------------------------------|---------|
| $x_i = \text{man}$ | 1 |
| $y_{i-1} = \text{JJ}$ | 1 |
| $i=n$ (last word of sentence) | 1 |
| x_i ends in -ly | 0 |

NER sequence labeling

identity of w_i , identity of neighboring words
embeddings for w_i , embeddings for 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 17.5 Typical features for a feature-based NER system.

Gazetteers

- List of place names; more generally, list of names of some typed category
- GeoNames (GEO), US SSN (PER), Getty Thesaurus of Geographic Placenames, Getty Thesaurus of Art and Architecture

Bun Crannich
Dromore West
Dromore
Youghal Harbour
Youghal Bay
Youghal
Eochail
Yellow River
Yellow Furze
Woodville
Wood View
Woodtown House
Woodstown
Woodstock House
Woodsgift House
Woodrooff House
Woodpark
Woodmount
Wood Lodge
Woodlawn Station
Woodlawn
Woodlands Station
Woodhouse
Wood Hill
Woodfort
Woodford River
Woodford
Woodfield House
Woodenbridge Junction Station
Woodenbridge
Woodbrook House
Woodbrook
Woodbine Hill
Wingfield House
Windy Harbour
Windy Gap

Training

$$\prod_{i=1}^n P(y_i \mid y_{i-1}, x, \beta)$$

For all training data, we want probability of the true label y_i conditioned on the previous true label y_{i-1} to be high.

This is simply multiclass logistic regression

Decoding

- With logistic regression, our prediction is simply the $\text{argmax } y$:

$$P(y \mid x, \beta)$$

- With an MEMM, we know the true y_{i-1} during training but we never of course know it at test time

$$P(y_i \mid y_{i-1}, x, \beta)$$

Decoding

- Greedy: proceed left to right, committing to the best tag for each time step (given the sequence seen so far)

Fruit flies like a banana

| | | | | |
|----|----|----|----|----|
| NN | VB | IN | DT | NN |
|----|----|----|----|----|

Decoding

DT NN VBD IN DT NN ???

The horse raced past the barn fell

Decoding

DT NN VBD IN DT NN ???

The horse raced past the barn fell

DT NN VBN IN DT NN VBD

Information later on in the sentence can influence the best tags earlier on.

All paths

| | | | | | | | |
|-------|--|--|--|--|--|--|--|
| END | | | | | | | |
| DT | | | | | | | |
| NNP | | | | | | | |
| VB | | | | | | | |
| NN | | | | | | | |
| MD | | | | | | | |
| START | | | | | | | |

^

Janet

will

back

the

bill

\$

Ideally, what we want is to calculate the joint probability of **each path** and pick the one with the highest probability. But for N time steps and K labels, number of possible paths = K^N

5 word sentence with 45 Penn Treebank tags

$45^5 = 184,528,125$ different paths

$45^{20} = 1.16e33$ different paths

Viterbi algorithm

- Basic idea: if an optimal path through a sequence uses **label L at time T**, then it must have used an optimal path to get to label L at time T
- We can discard all non-optimal paths up to label L at time T

Evaluation

- We evaluate NER with precision/recall/F1 over typed chunks.

Evaluation

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---------------|-------|-------|----|-----|-------|----|-------|
| | tim | cook | is | the | CEO | of | Apple |
| <i>gold</i> | B-PER | I-PER | O | O | O | O | B-ORG |
| <i>system</i> | B-PER | O | O | O | B-PER | O | B-ORG |

<start, end, type>

| | |
|-----------|-----|
| Precision | 1/3 |
| Recall | 1/2 |

gold

<1,2,PER>
<7,7,ORG>

system

<1,1,PER>
<5,5,PER>
<7,7,ORG>

Activity

- `12.ner/ToponymResolution.ipynb`