# Natural Language Processing

Lecture 2:  Applications of NLP

# Information Extraction (from 10,000 feet)

WASHINGTON/SELMA, Ala. (Reuters) - Democratic U.S. presidential front-runner Bernie Sanders raised $46.5 million in February, his campaign said on Sunday, and will launch new television ad buys in nine states with primaries later this month after this week's Super Tuesday contests. Joe Biden's campaign reported raising $5 million the day of the South Carolina primary. His February haul was $18 million, spokesman Michael Gwin said. Meanwhile, rival Elizabeth Warren, who struggled to a fifth-place finish in South Carolina, raised more than $29 million in February, her campaign manager Roger Lau said in a memo to supporters on Sunday.

| Candidate | Party | Feb Fundraising Total |
|-----------|-------|-----------------------|
| Sanders | D | $46,500,000 |
| Biden | D | $18,000,000 |
| Warren | D | $29,000,000 |

# Named Entity Recognition

**Elizabeth Warren**, the liberal firebrand who emerged as a top Democratic contender for the **White House** on the strength of an anti-corruption platform backed by a dizzying array of policy proposals, ended her campaign on Thursday. A former bankruptcy law professor who forged a national reputation as a scourge of Wall Street even before entering politics, **Warren** had banked on a strong showing on Super Tuesday after a string of disappointing finishes in the early states. But she trailed far behind front-runners **Bernie Sanders** and **Joe Biden**, placing third in her home state of Massachusetts, which she continues to represent in the U.S. Senate.

- Label certain kinds of proper nouns:
  - Personal names
  - Organizations
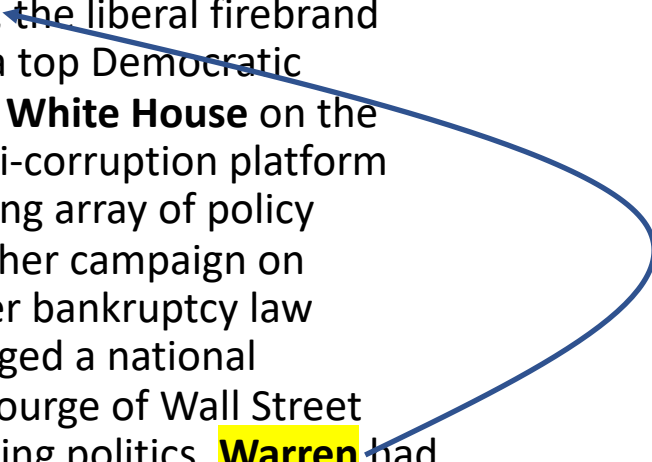  - Geopolitical entities
  - Locations
  - etc.

# Reference resolution

**Elizabeth Warren**, the liberal firebrand who emerged as a top Democratic contender for the **White House** on the strength of an anti-corruption platform backed by a dizzying array of policy proposals, ended her campaign on Thursday. A former bankruptcy law professor who forged a national reputation as a scourge of Wall Street even before entering politics, **Warren** had banked on a strong showing on Super Tuesday after a string of disappointing finishes in the early states. But she trailed far behind front-runners **Bernie Sanders** and **Joe Biden**, placing third in her home state of Massachusetts, which she continues to represent in the U.S. Senate.

# Coreference resolution

**Elizabeth Warren**, the liberal firebrand who emerged as a top Democratic contender for the **White House** on the strength of an anti-corruption platform backed by a dizzying array of policy proposals, ended her campaign on Thursday. A former bankruptcy law professor who forged a national reputation as a scourge of Wall Street even before entering politics, **Warren** had banked on a strong showing on Super Tuesday after a string of disappointing finishes in the early states. But she trailed far behind front-runners **Bernie Sanders** and **Joe Biden**, placing third in her home state of Massachusetts, which she continues to represent in the U.S. Senate.

# Relation detection

WASHINGTON/SELMA, Ala. (Reuters) - Democratic U.S. presidential front-runner Bernie Sanders raised $46.5 million in February, his campaign said on Sunday, and will launch new television ad buys in nine states with primaries later this month after this week's Super Tuesday contests. Joe Biden's campaign reported raising $5 million the day of the South Carolina primary. His February haul was $18 million, spokesman Michael Gwin said. Meanwhile, rival Elizabeth Warren, who struggled to a fifth-place finish in South Carolina, raised more than $29 million in February, her campaign manager Roger Lau said in a memo to supporters on Sunday.

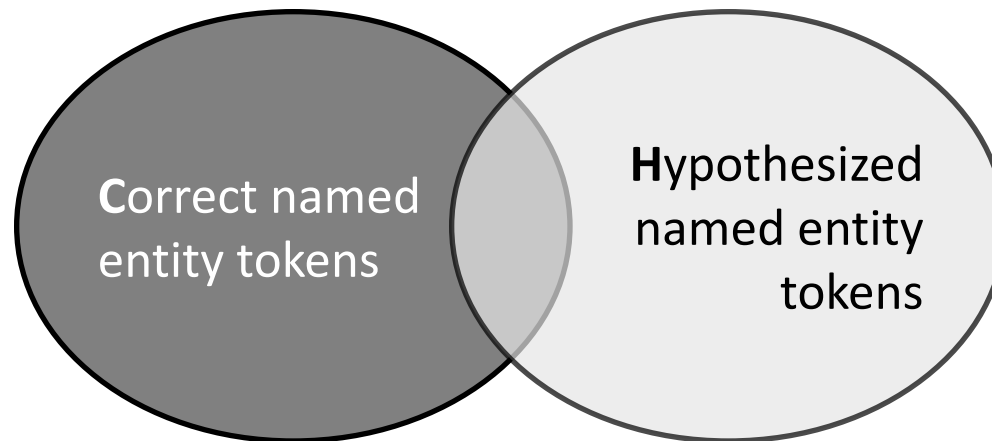| Candidate | member_of |
|---|---|
| Bernie Sanders | Democratic Party |
| Joe Biden | Democratic Party |
| Elizabeth Warren | Democratic Party |

# NER

# Encoding the NER Problem

President Donald Trump met with local leaders and federal responders shortly after landing at an Air Force base in Carolina, Puerto Rico, for what was supposed to be a briefing on the situation on the island. Instead, Trump turned it into an opportunity to congratulate himself and the federal government's response to the disaster…. He downplayed throughout his remarks how dire things are in Puerto Rico, where more than half of the people don't have power, running water, or cellphone service two weeks after Hurricane Maria, a Category 4 storm, tore through the island.

| | |
|---|---|
| ... | ... |
| B-PER | President |
| I-PER | Donald |
| I-PER | Trump |
| O | met |
| O | with |
| O | local |
| O | leaders |
| O | and |
| O | federal |
| O | responders |
| O | shortly |
| O | after |
| O | landing |
| O | at |
| O | an |
| B-ORG | Air |
| I-ORG | Force |
| O | base |
| O | in |
| B-LOC | Carolina |
| O | , |
| B-LOC | Puerto |
| I-LOC | Rico |
| ... | ... |

# Some Named Entity Types

- Different annotation schemes for NER use different types
- Common types include
  - PER—person
  - ORG—Organization
  - LOC—Location
  - GPE—Geopolitical Entity
  - FAC—Facility
  - NAT—Natural phenomenon
- These are only tagged when they are proper names

# Evaluating an NER System



$$\text{recall} = \frac{|C \cap H|}{|C|} \qquad\qquad \text{precision} = \frac{|C \cap H|}{|H|}$$
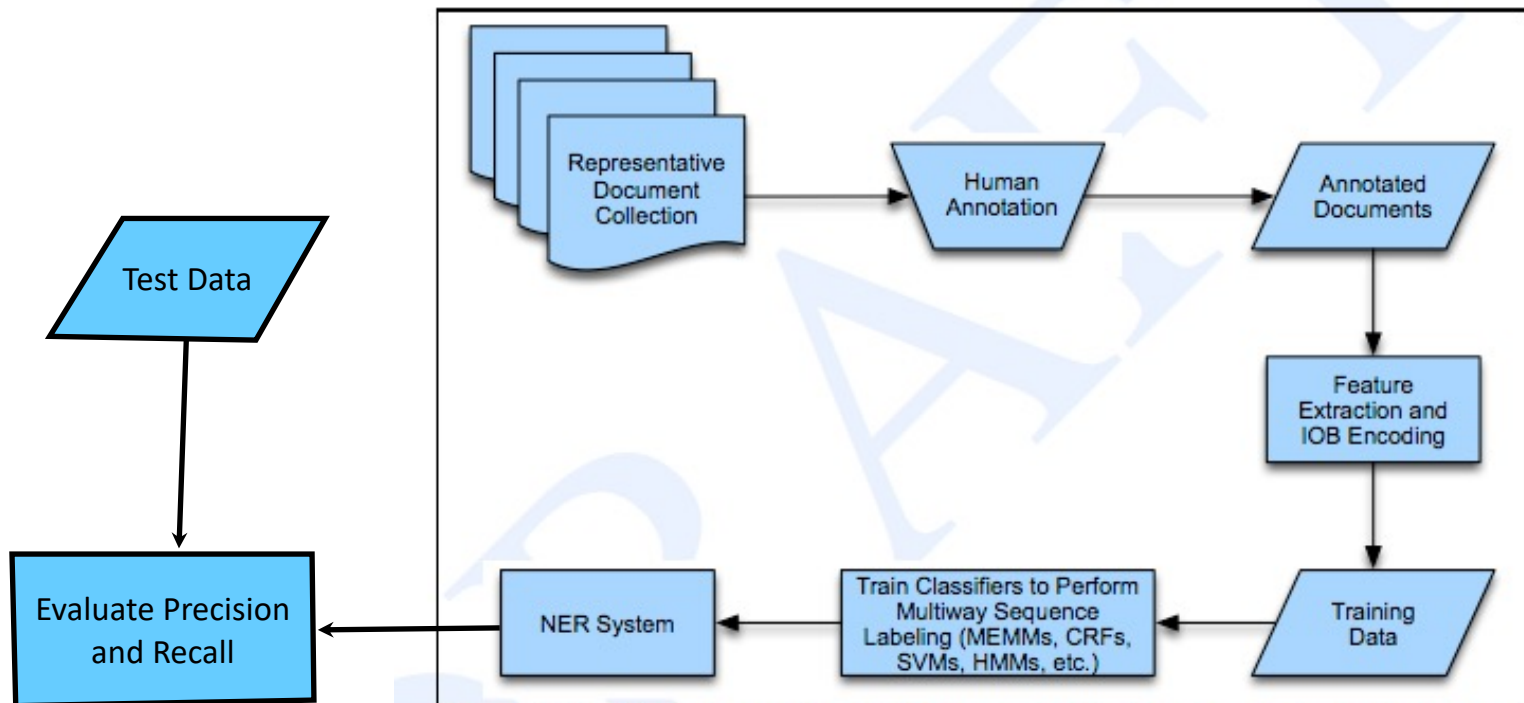
# NER System Building Process



Figure 22.10 Basic steps in the statistical sequence labeling approach to creating a named entity recognition system.

# Relation Extraction

# Examples of Relations

| 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 |

**Figure 22.11** Semantic relations with examples and the named entity types they involve.

# Seeding Tuples

Examples
  Brad is married to Angelina.
  Bill is married to Hillary.
  Hillary is married to Bill.
  Hillary is the wife of Bill.
Seeds
    X is married to Y
    X is the wife of Y
Find all X,Y where these templates exist
Extract the X,Ys, find all other X ... Y
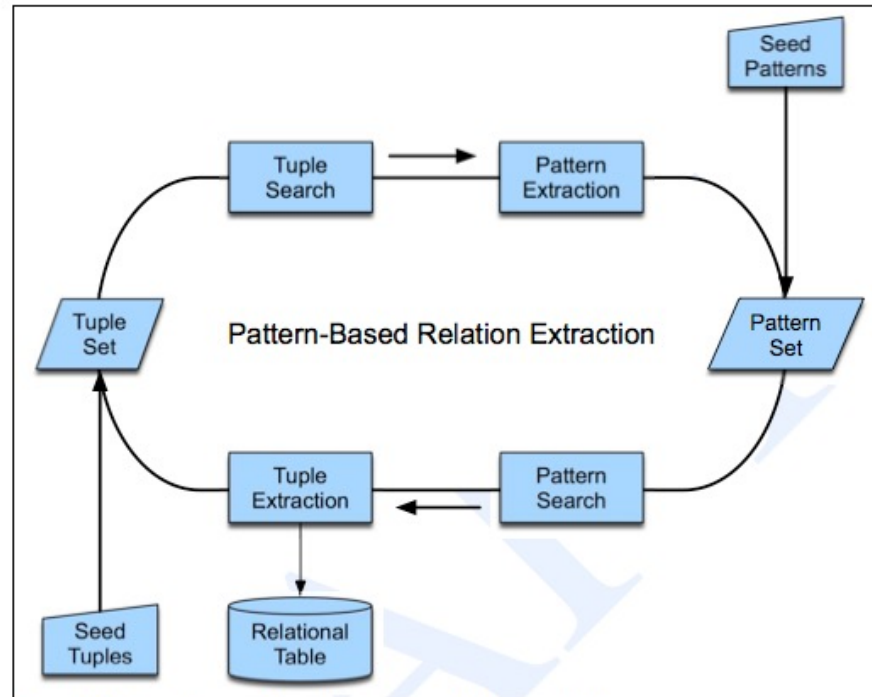
# Bootstrapping Relations



**Figure 22.16** Pattern and bootstrapping-based relation extraction.

# Information Retrieval

# The Vector Space Model

- Each document is a $|\Sigma|$-dimensional vector $d_i$:

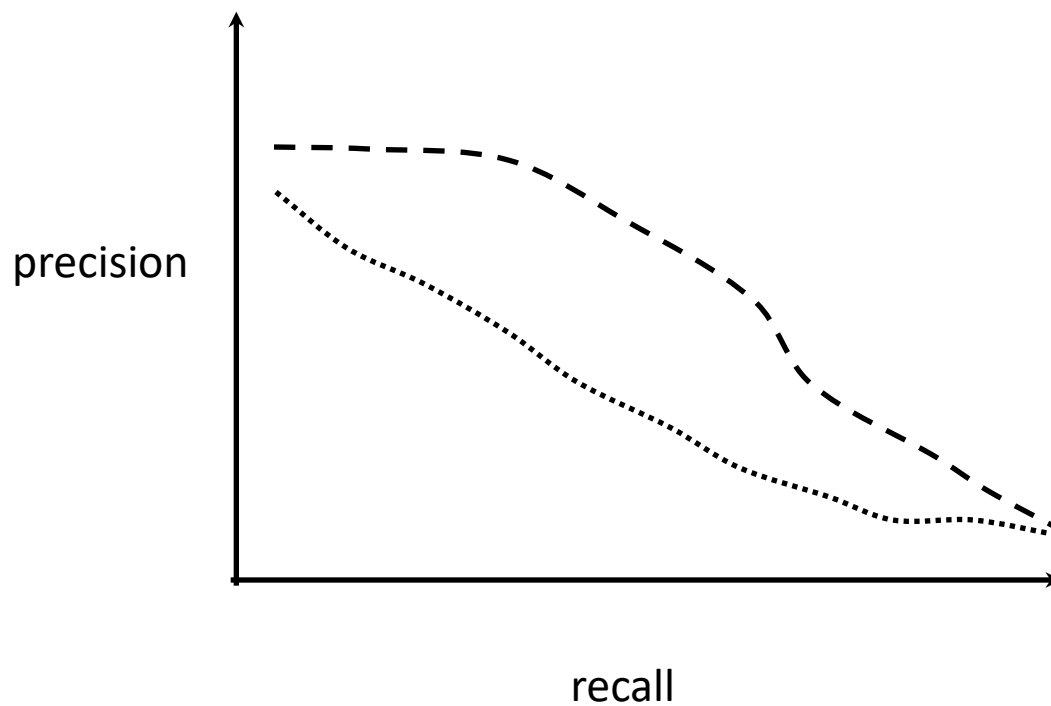$$d_i[j] = \text{count of } w_j \text{ in document } i$$

- Given a query q, represent it in the same way.

$$q[j] = \text{count of } w_j \text{ in query}$$

- Similarity of vectors $\Rightarrow$ relevance:

$$\text{cosine\_similarity}(d_i, q) = \frac{\sum_j d_i[j] \times q[j]}{\sqrt{\sum_j d_i[j]^2} \times \sqrt{\sum_j q[j]^2}}$$

- Twists: **tf-idf** $\quad x[j] = \text{count}(w_j) \times \log \dfrac{\# \text{ docs}}{\# \text{ docs with } w_j}$

# Evaluating Information Retrieval

# Question Answering
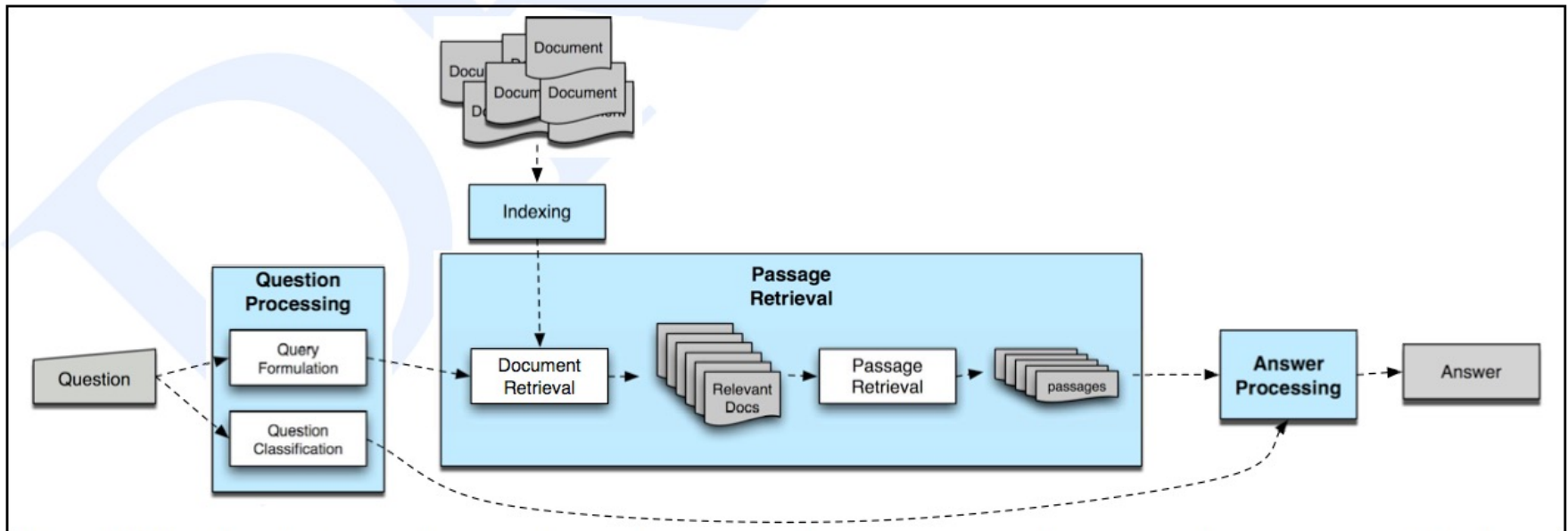
# Question Answering Architecture



**Figure 23.8** The 3 stages of a generic question answering system: question processing, passage retrieval, and answer processing..
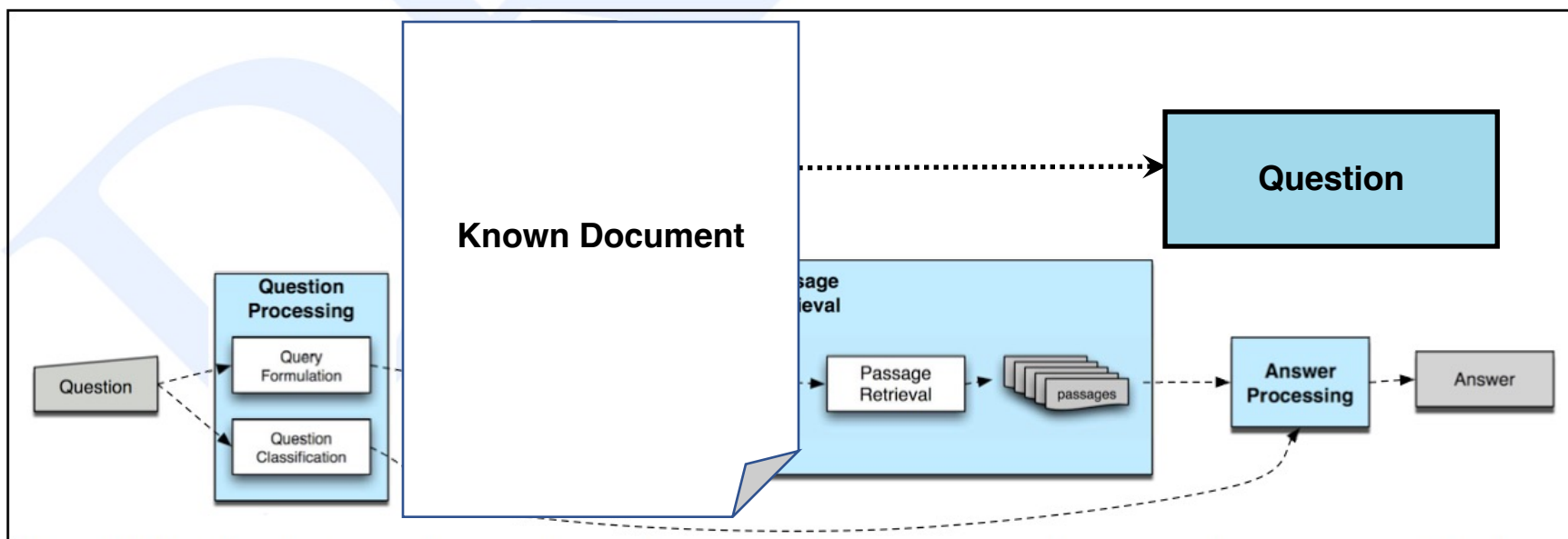
*SLP* p. 793

# Your Project (Perhaps)



**Figure 23.8** The 3 stages of a generic question answering system: question processing, passage retrieval, and answer processing..

*SLP* p. 793

# Evaluating QA

$$\text{mean reciprocal rank} = \frac{1}{T} \sum_{i=1}^{T} \frac{1}{\text{rank of first correct answer to question } i}$$

# Some General Tools

- Supervised classification

- Feature vector representations

- Bootstrapping

- Evaluation:
  - Precision and recall (and their curves)
  - Mean reciprocal rank