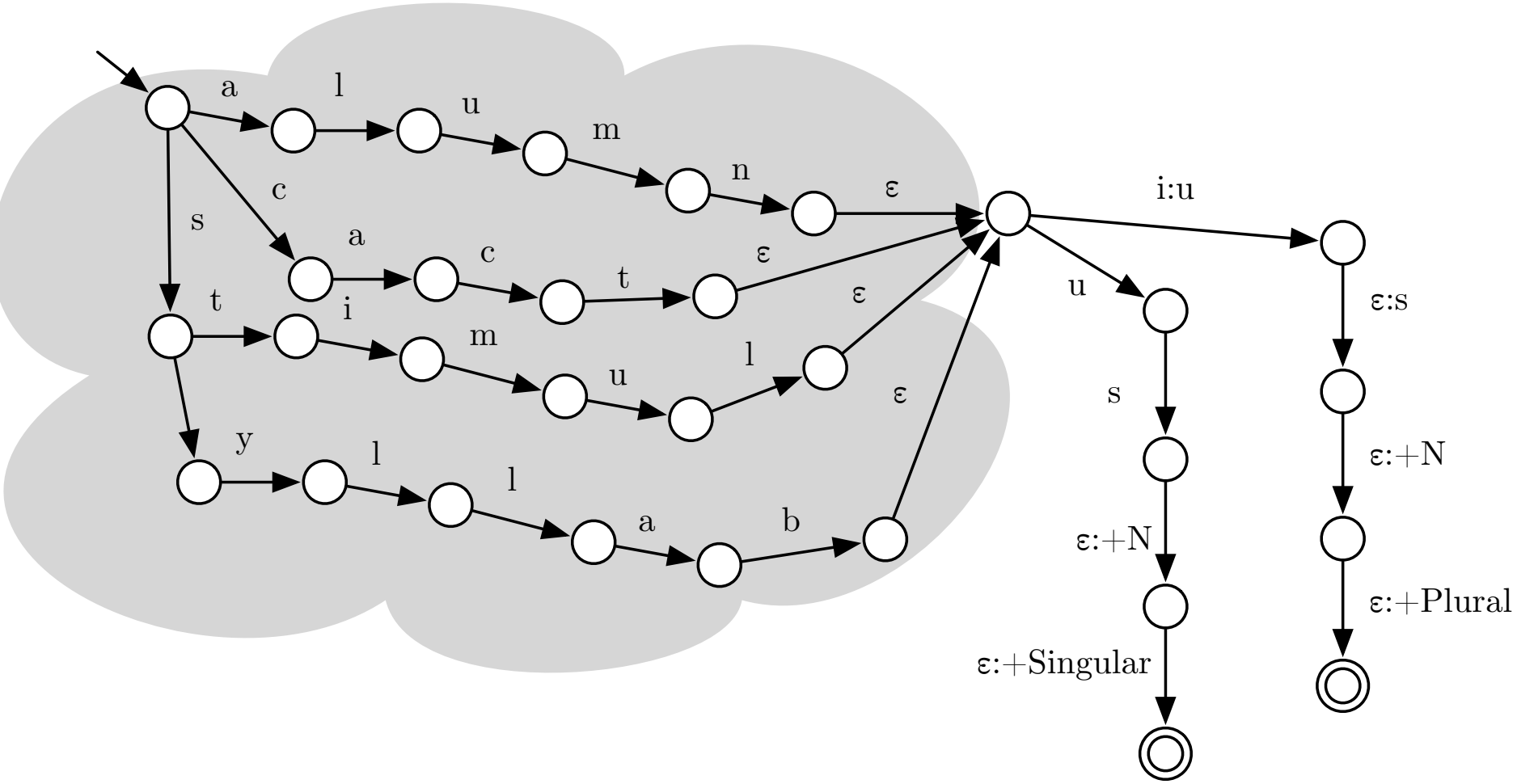
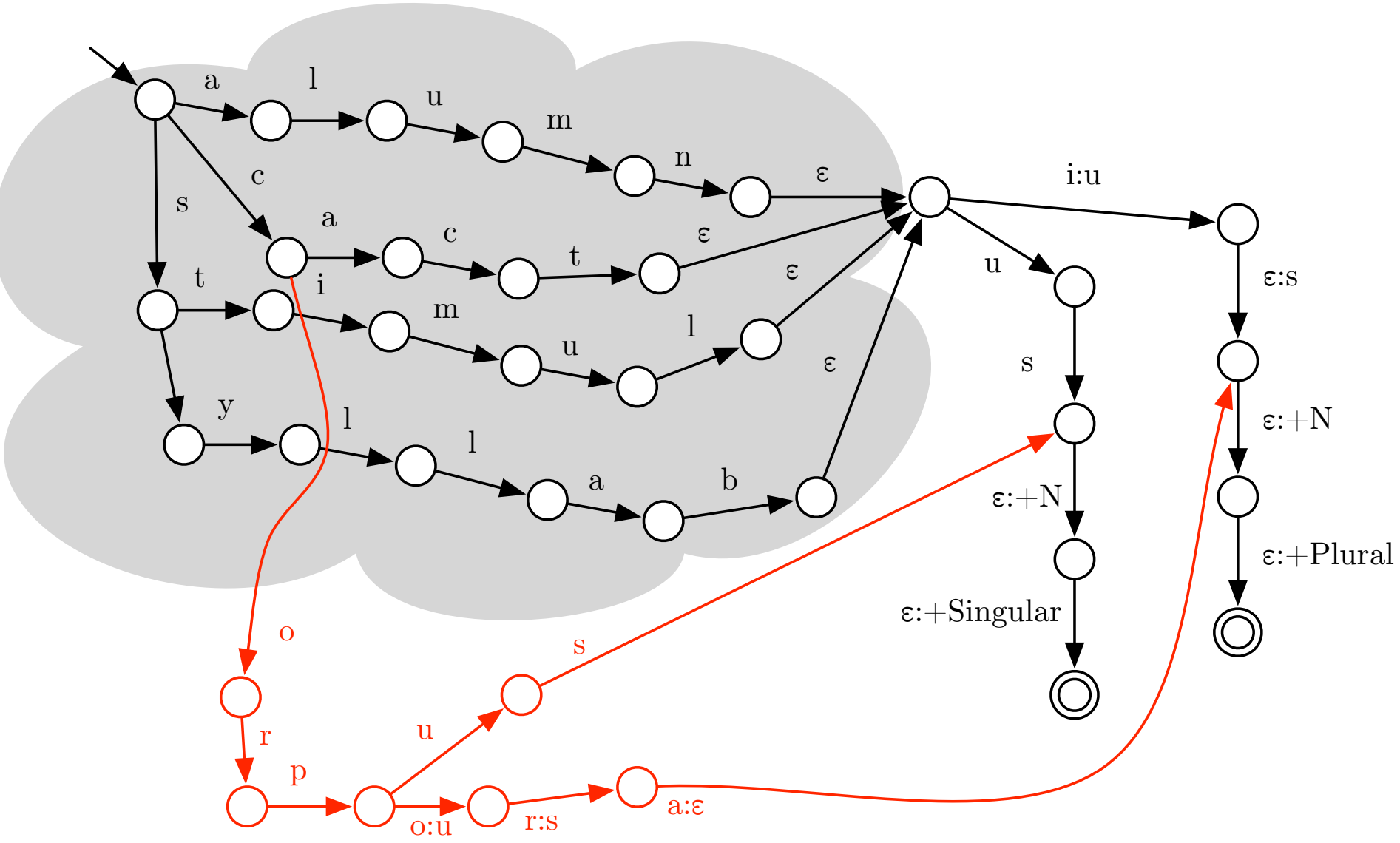
Natural Language Processing

Lecture 5: Applications of NLP


a
l
u
m
n
ε
s
c
a
c
t
ε
t
i
m
u
l
ε
y
l
l
a
b
ε
i:u
u
s
ε:+N
ε:+Singular
ε:s
ε:+N
ε:+Plural


a
l
u
m
n
ε
s
c
a
c
t
ε
t
i
m
u
l
ε
ε
y
l
l
a
b
i:u
u
s
ε:+N
ε:+Singular
ε:s
ε:+N
ε:+Plural
o
r
p
u
s
o:u
r:s
a:ε

Information Extraction (From 10,000 Feet)

- **Input**: text, empty relational database
- **Output**: populated relational database

```
Senator John Edwards is to
drop out of the race to
become the Democratic party's
presidential candidate after
consistently trailing in
third place.
In the latest primary, held
in Florida yesterday, Edwards
gained only 14% of the vote,
with Hillary Clinton polling
50% and Barack Obama on 33%.
A reported 1.5m voters turned
out to vote.
```

State	Party	Candidate	Fraction
FL	D	Edwards	0.14
FL	D	Clinton	0.50
FL	D	Obama	0.33

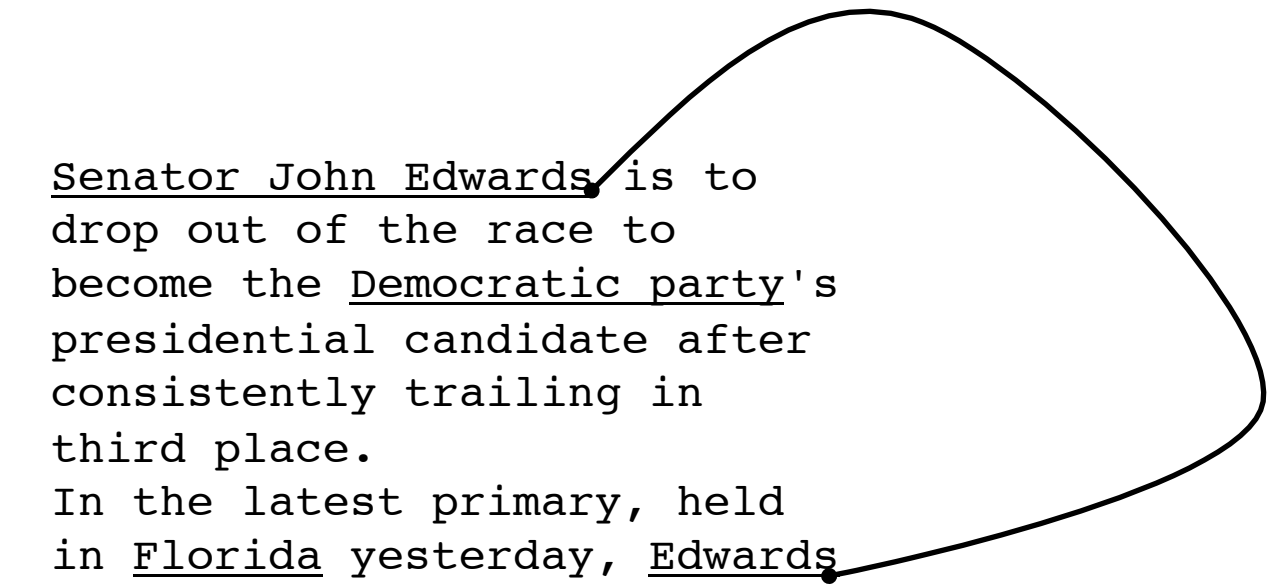
Named Entity Recognition

Senator John Edwards is to drop out of the race to become the Democratic party's presidential candidate after consistently trailing in third place.
In the latest primary, held in Florida yesterday, Edwards gained only 14% of the vote, with Hillary Clinton polling 50% and Barack Obama on 33%.
A reported 1.5m voters turned out to vote.

Reference Resolution

Senator John Edwards is to drop out of the race to become the Democratic party's presidential candidate after consistently trailing in third place.
In the latest primary, held in Florida yesterday, Edwards gained only 14% of the vote, with Hillary Clinton polling 50% and Barack Obama on 33%.
A reported 1.5m voters turned out to vote.

Coreference Resolution



Senator John Edwards is to drop out of the race to become the Democratic party's presidential candidate after consistently trailing in third place.
In the latest primary, held in Florida yesterday, Edwards gained only 14% of the vote, with Hillary Clinton polling 50% and Barack Obama on 33%.
A reported 1.5m voters turned out to vote.

Relation Detection

Senator John Edwards is to drop out of the race to become the Democratic party's presidential candidate after consistently trailing in third place.
In the latest primary, held in Florida yesterday, Edwards gained only 14% of the vote, with Hillary Clinton polling 50% and Barack Obama on 33%.
A reported 1.5m voters turned out to vote.

member_of	
John Edwards	Democratic party
Hillary Clinton	Democratic party
Barack Obama	Democratic party

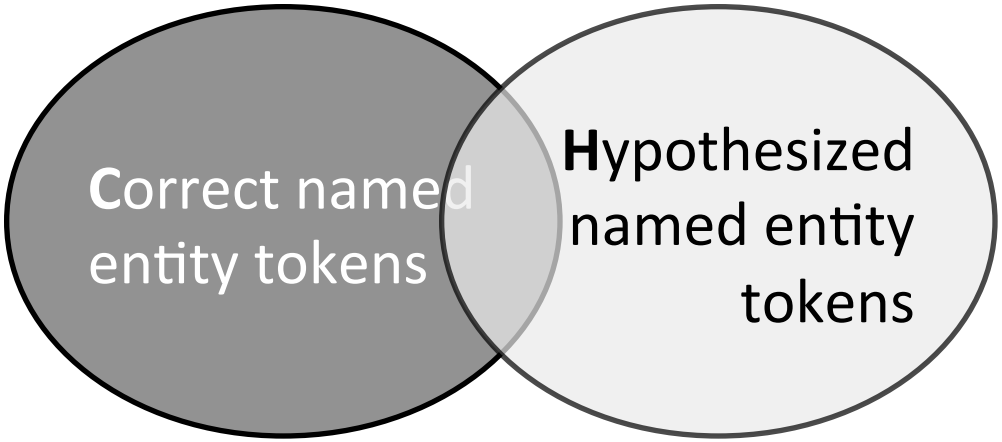
NER

Encoding the NER Problem

With that, Edwards' campaign will end the way it began 13 months ago -- with the candidate pitching in to rebuild lives in a city still ravaged by Hurricane Katrina. Edwards embraced New Orleans as a glaring symbol of what he described as a Washington that didn't hear the cries of the downtrodden.

...	...
O	ravaged
O	by
B-NAT	Hurricane
I-NAT	Katrina
O	.
B-PER	Edwards
O	embraced
B-LOC	New
I-LOC	Orleans
O	as
O	a
O	glaring
O	symbol
O	of
O	what
O	he
O	described
O	as
O	a
B-GPE	Washington
O	that
O	didn't
O	hear
...	...

Evaluating an NER System



$$\text{recall} = \frac{|C \cap H|}{|C|}$$

$$\text{precision} = \frac{|C \cap H|}{|H|}$$

NER System Building Process

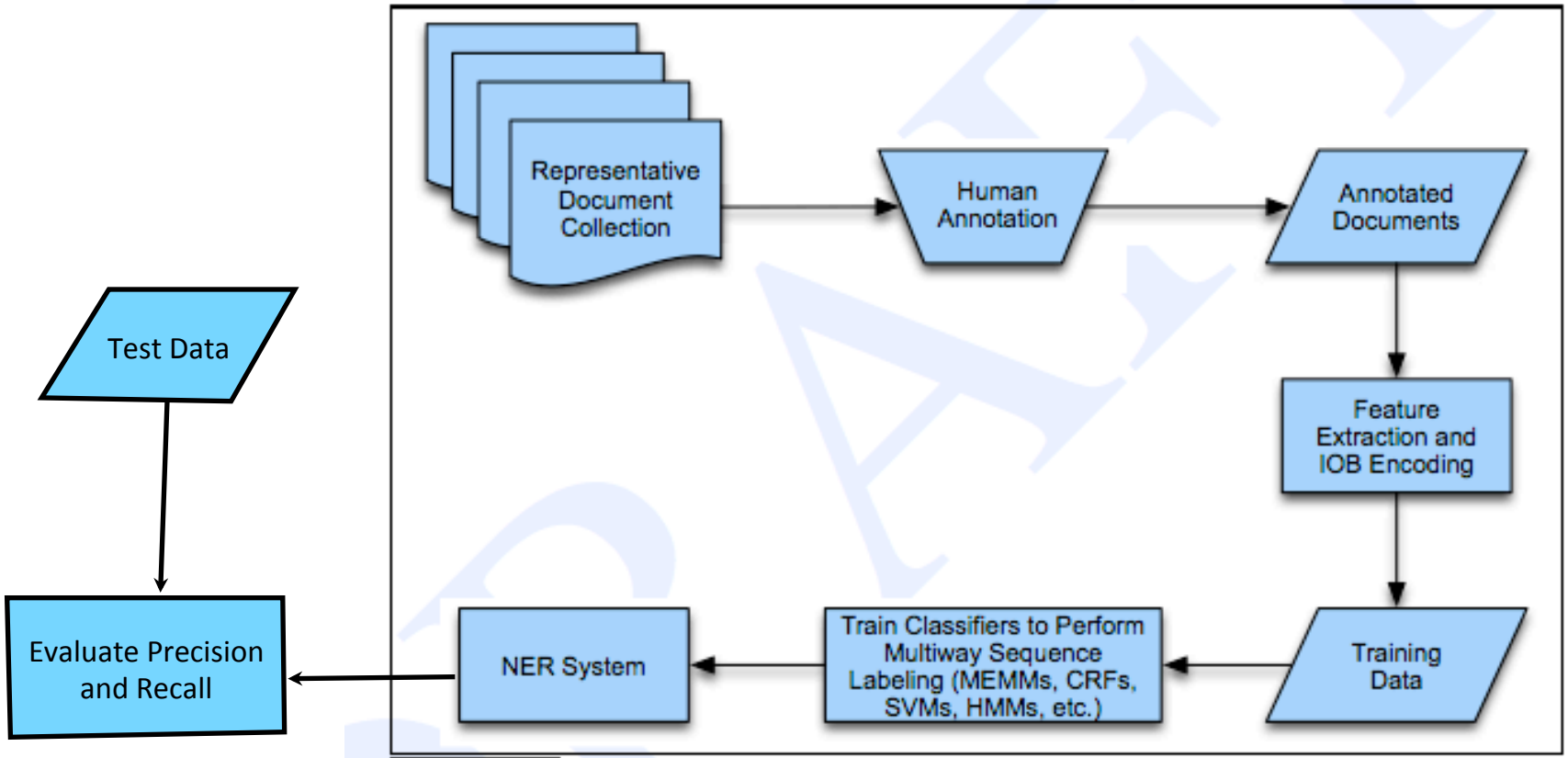


Figure 22.10 Basic steps in the statistical sequence labeling approach to creating a named entity recognition system.

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.

Bootstrapping Relations

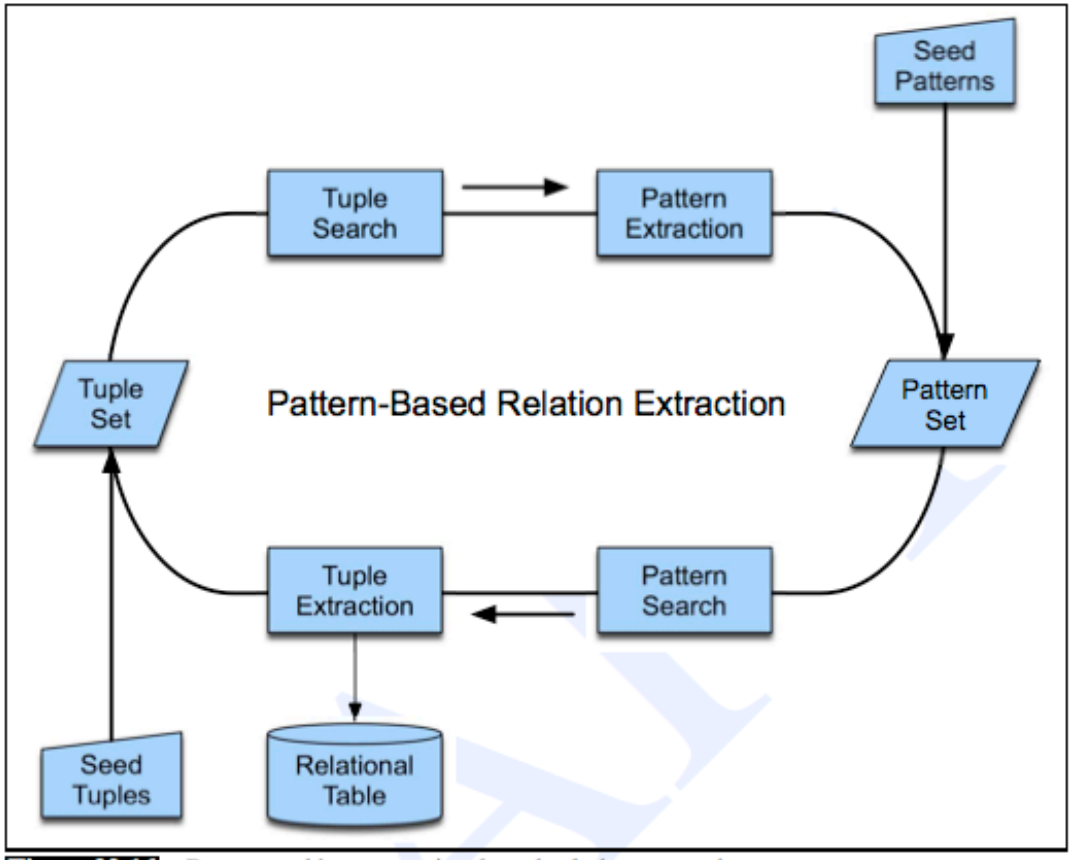


Figure 22.16 Pattern and bootstrapping-based relation extraction.

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.

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

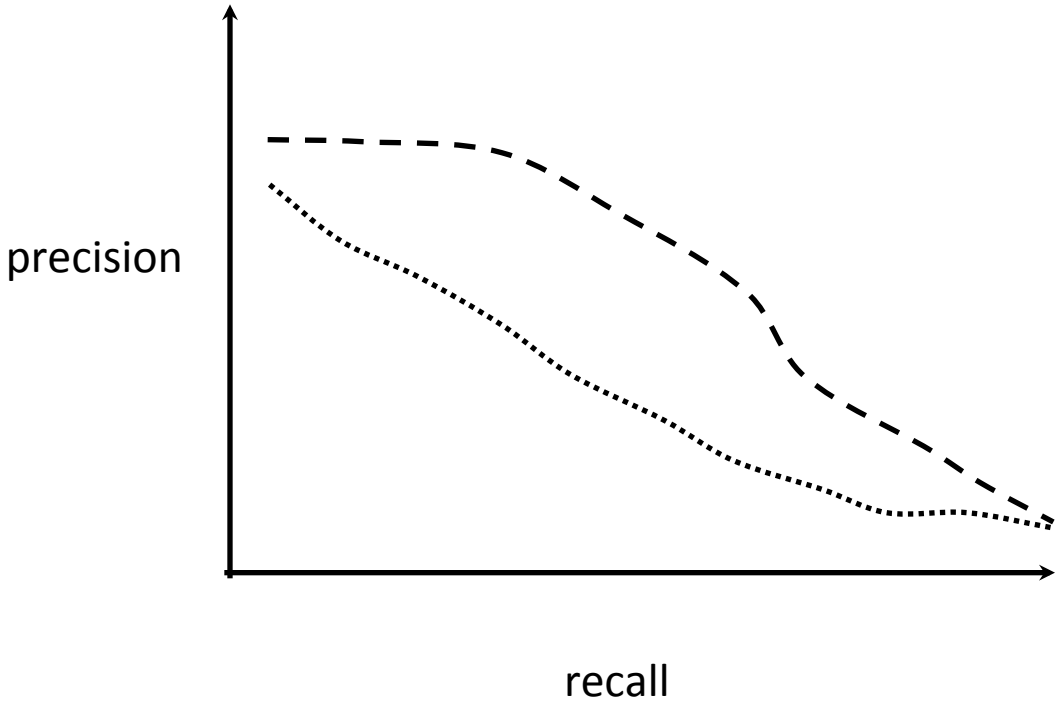
- Similarity of vectors ⇒ relevance:

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

- Twists: **tf-idf**

$$x[j] = \text{count}(w_j) \times \log \frac{\#\text{ docs}}{\#\text{ docs with } w_j}$$

Evaluating Information Retrieval



Question Answering

Question Answering Architecture

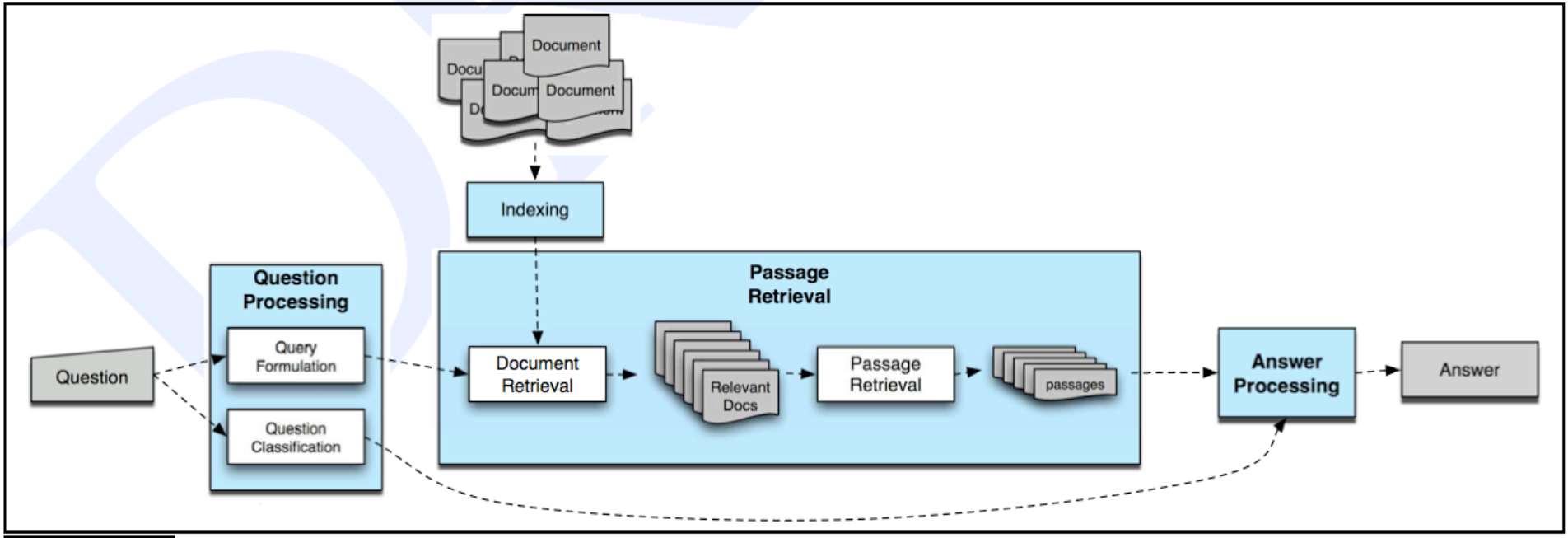


Figure 23.8 The 3 stages of a generic question answering system: question processing, passage retrieval, and answer processing..

SLP p. 793

Your Project (Perhaps)

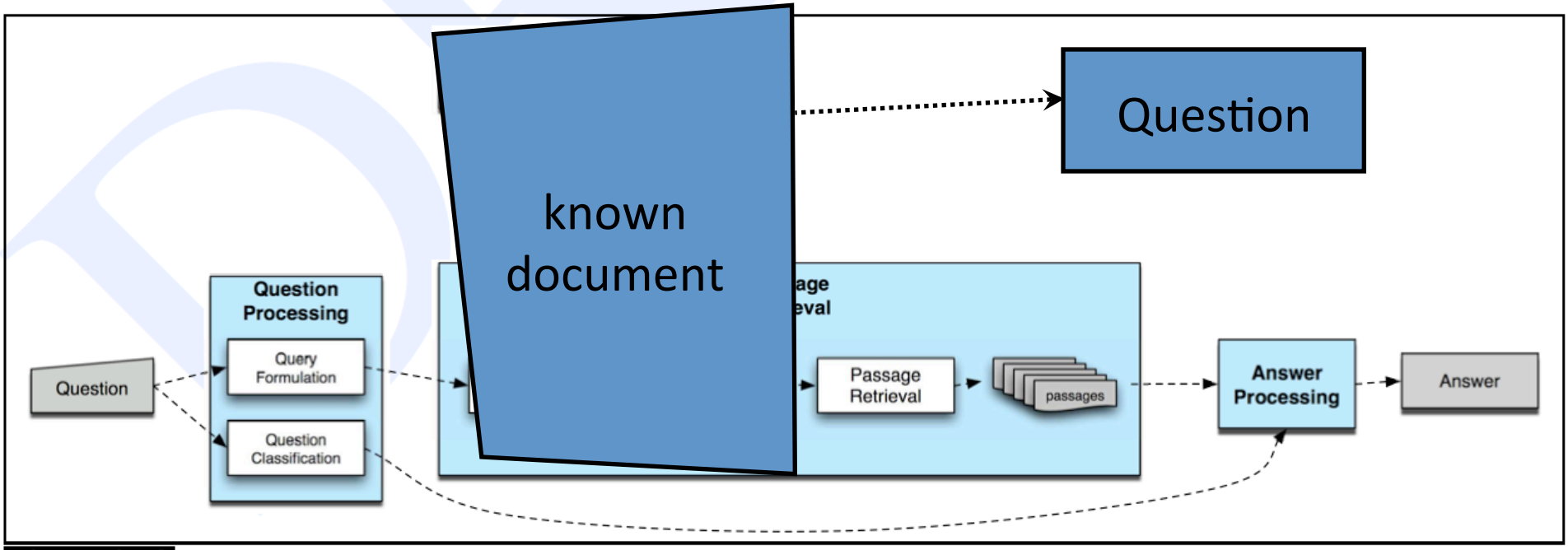


Figure 23.8 The 3 stages of a generic question answering system: question processing, passage retrieval, and answer processing..

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