Topic 10
Information Extraction

What counts as information extraction?

- Key phrase extraction
- Named entity recognition
- Named entity disambiguation and linking
- Relationship extraction

Where is IE used?

- Tagging content
- Chatbots
- Social media applications
- Data extraction from forms
- Event extraction
- Temporal information extraction
- Template filling

What are different IE tasks you can do with this document?

SAN FRANCISCO — Shortly after Apple used a new tax law last year to bring back most of the $252 billion it had held abroad, the company said it would buy back $100 billion of its stock.

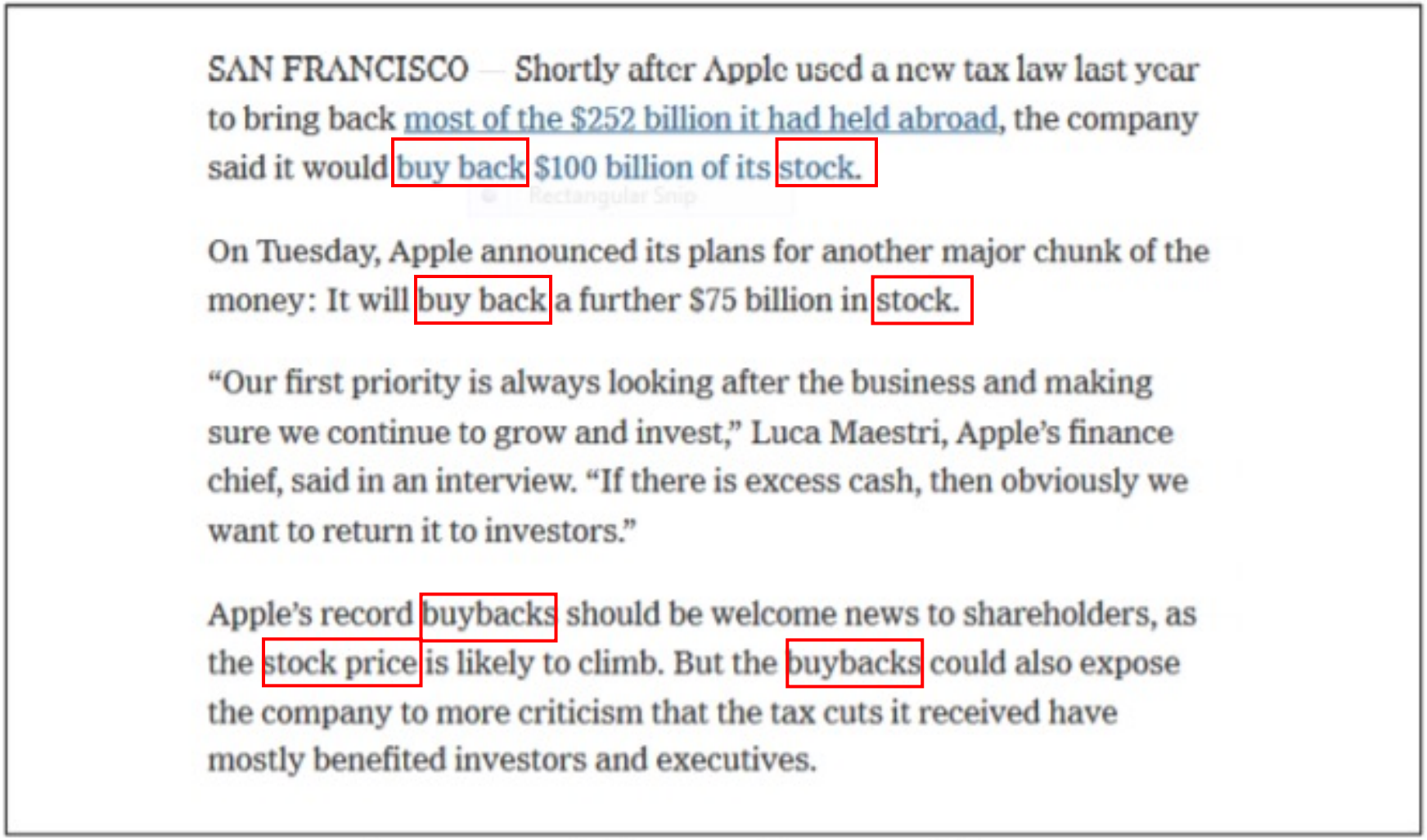
On Tuesday, Apple announced its plans for another major chunk of the money: It will buy back a further $75 billion in stock.

"Our first priority is always looking after the business and making sure we continue to grow and invest," Luca Maestri, Apple's finance chief, said in an interview. "If there is excess cash, then obviously we want to return it to investors."

Apple's record buybacks should be welcome news to shareholders, as the stock price is likely to climb. But the buybacks could also expose the company to more criticism that the tax cuts it received have mostly benefited investors and executives.

Figure 5-2. A New York Times article from April 30, 2019 [7]

What are different IE tasks you can do with this document?



SAN FRANCISCO — Shortly after Apple used a new tax law last year to bring back most of the $252 billion it had held abroad, the company said it would buy back $100 billion of its stock.

On Tuesday, Apple announced its plans for another major chunk of the money: It will buy back a further $75 billion in stock.

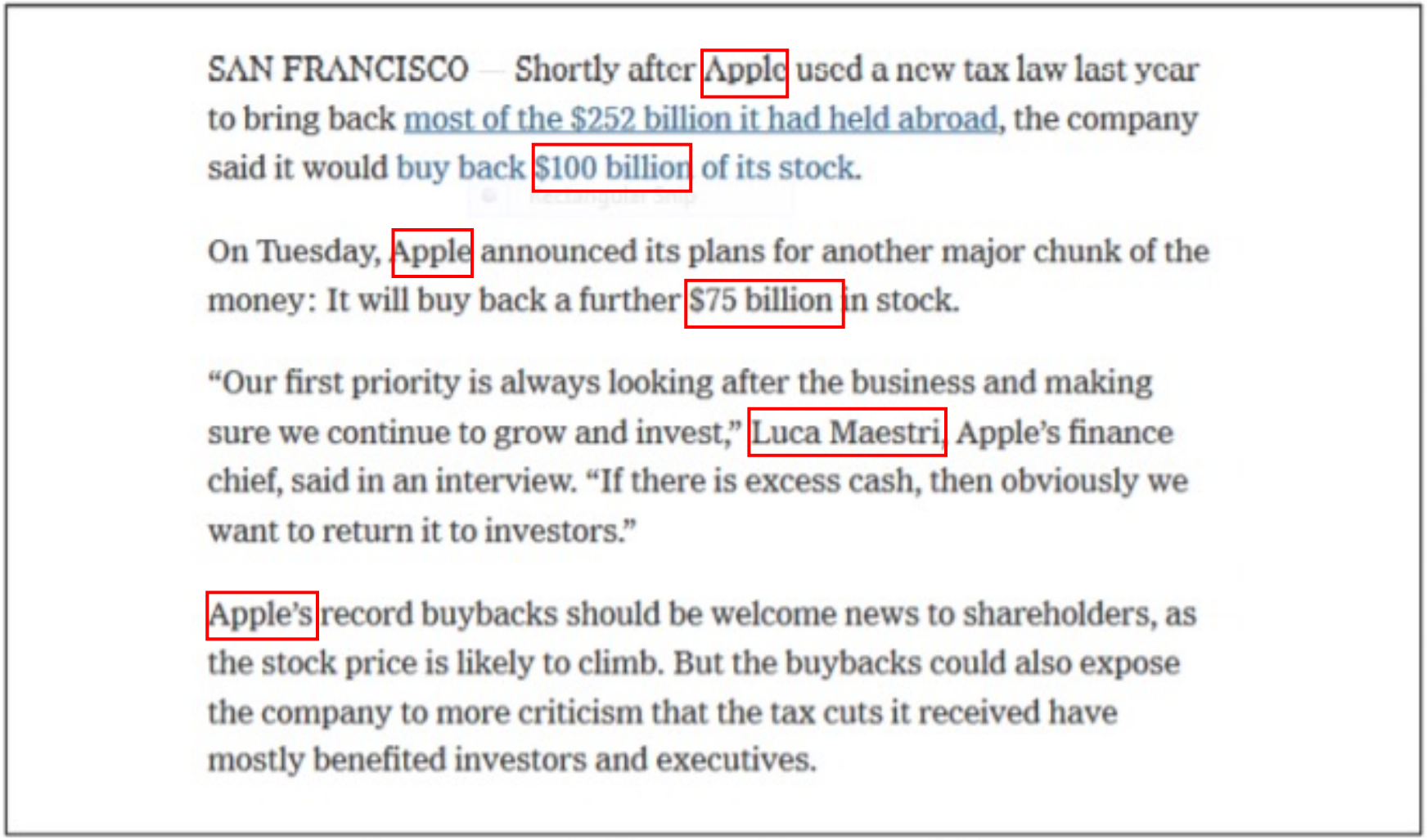
"Our first priority is always looking after the business and making sure we continue to grow and invest," Luca Maestri, Apple's finance chief, said in an interview. "If there is excess cash, then obviously we want to return it to investors."

Apple's record buybacks should be welcome news to shareholders, as the stock price is likely to climb. But the buybacks could also expose the company to more criticism that the tax cuts it received have mostly benefited investors and executives.

Figure 5-2. A New York Times article from April 30, 2019 [7]

Keyword or keyphrase extraction

What are different IE tasks you can do with this document?



SAN FRANCISCO — Shortly after Apple used a new tax law last year to bring back most of the $252 billion it had held abroad, the company said it would buy back $100 billion of its stock.

On Tuesday, Apple announced its plans for another major chunk of the money: It will buy back a further $75 billion in stock.

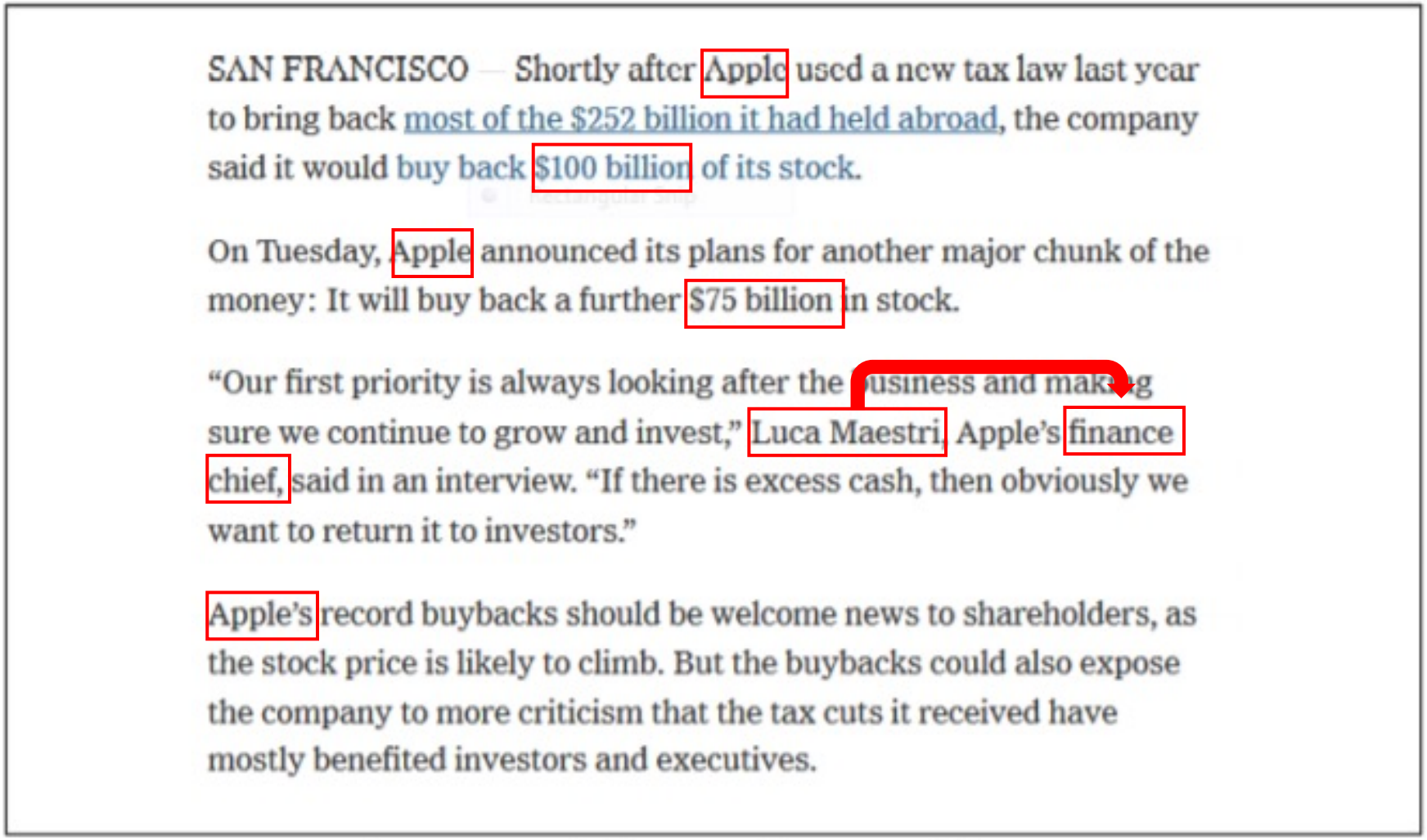
"Our first priority is always looking after the business and making sure we continue to grow and invest," Luca Maestri, Apple's finance chief, said in an interview. "If there is excess cash, then obviously we want to return it to investors."

Apple's record buybacks should be welcome news to shareholders, as the stock price is likely to climb. But the buybacks could also expose the company to more criticism that the tax cuts it received have mostly benefited investors and executives.

Figure 5-2. A New York Times article from April 30, 2019 [7]

Named Entity Recognition

What are different IE tasks you can do with this document?



SAN FRANCISCO — Shortly after Apple used a new tax law last year to bring back most of the $252 billion it had held abroad, the company said it would buy back $100 billion of its stock.

On Tuesday, Apple announced its plans for another major chunk of the money: It will buy back a further $75 billion in stock.

"Our first priority is always looking after the business and making sure we continue to grow and invest," Luca Maestri, Apple's finance chief, said in an interview. "If there is excess cash, then obviously we want to return it to investors."

Apple's record buybacks should be welcome news to shareholders, as the stock price is likely to climb. But the buybacks could also expose the company to more criticism that the tax cuts it received have mostly benefited investors and executives.

Figure 5-2. A New York Times article from April 30, 2019 [7]

Relationship extraction

IE sometimes requires more advanced techniques than simple classification

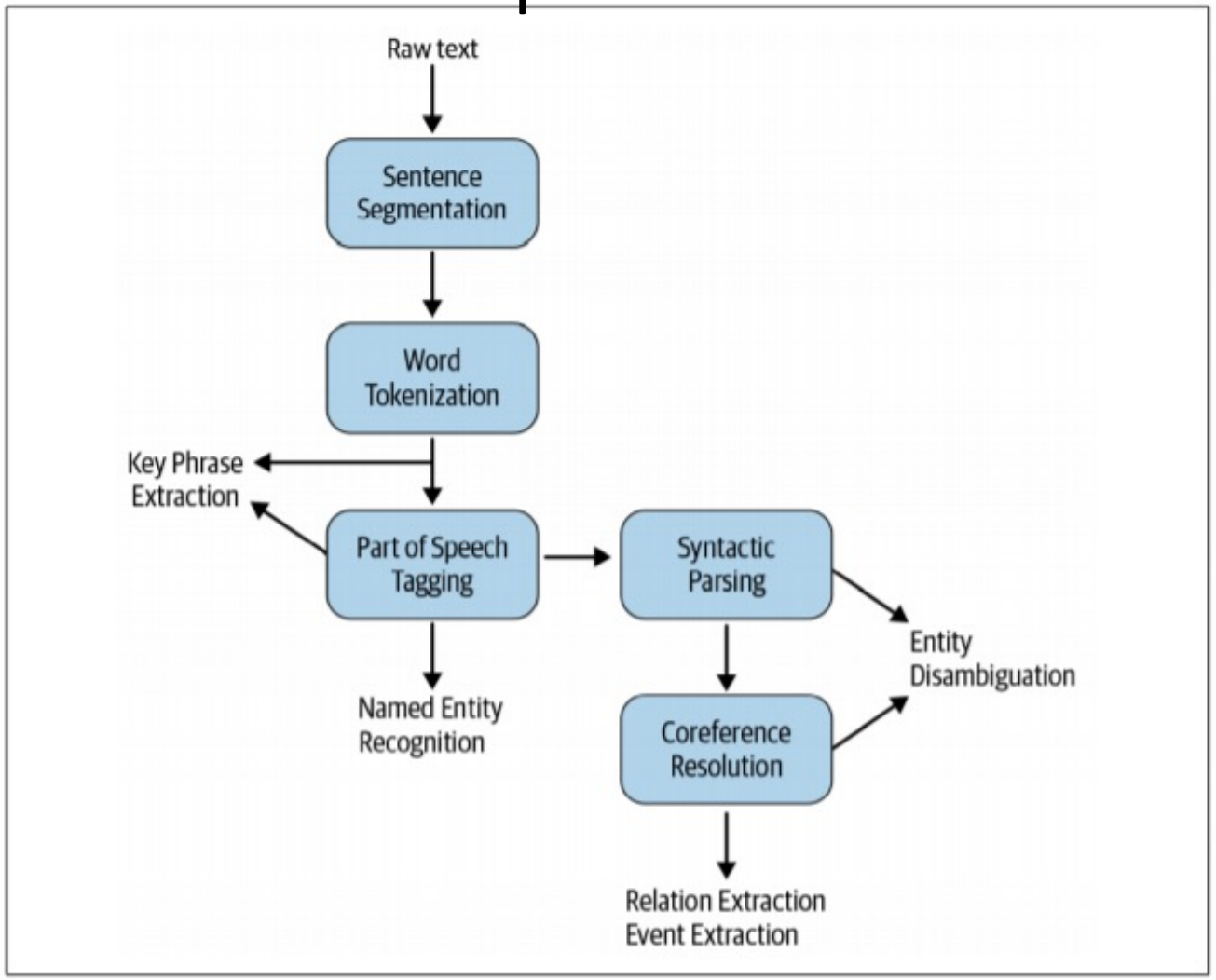


Figure 5-3. IE Pipeline illustrating NLP processing needed for some IE tasks

Keyphrase Extraction

- Extract nouns and noun phrases
- Textacy built on top of SpaCy
- TextRank

```
import textacy

text = "Apple is looking at buying U.K. startup for $1 billion. This
would be their largest acquisition ever."

doc = textacy.make_spacy_doc(text, lang='en_core_web_sm') #
key_phrases = textacy.extract.keyterms.most_important_terms(doc)

print(key_phrases)
```

Review: Named Entity Recognition (NER)

- Entity is the specific piece of information extracted
- Anything that can be attributed to a proper name can be considered a named entity

NER Systems

- BIO
- Can have a sequence labeling approach opposed to classification we covered in chapter 4
- **What does BOW and TF-IDF representation assume?**

```
Essex   B-ORG
,       O
however O
,       O
look    O
certain O
to      O
regain  O
their   O
top     O
spot    O
after   O
Nasser  B-PER
Hussain I-PER
and     O
Peter   B-PER
Such    I-PER
gave    O
them    O
a       O
firm    O
grip    O
on      O
their   O
match   O
against O
Yorkshire       B-ORG
at      O
Headingley      B-LOC
.       O
```

Figure 5-7. NER training data format example

Conditional Random Field, Lafferty et al. 2001

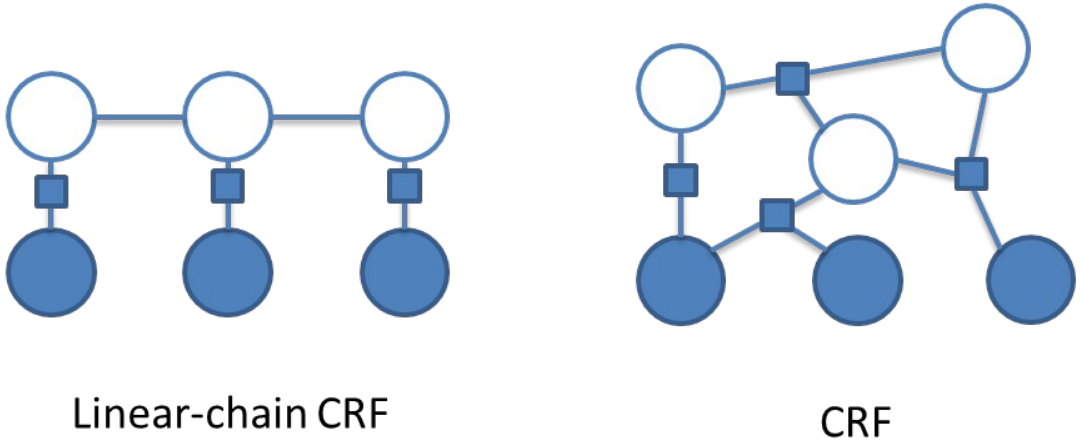
- A Conditional Random Field (CRF) is a probabilistic graphical model

Given:

  - some feature extractors (feature extractors need to output real numbers)
  - weights associated with the features (which are learned)
  - **previous labels**

Task: Predict the current label

- sklearn-crfsuite



Linear-chain CRF

CRF

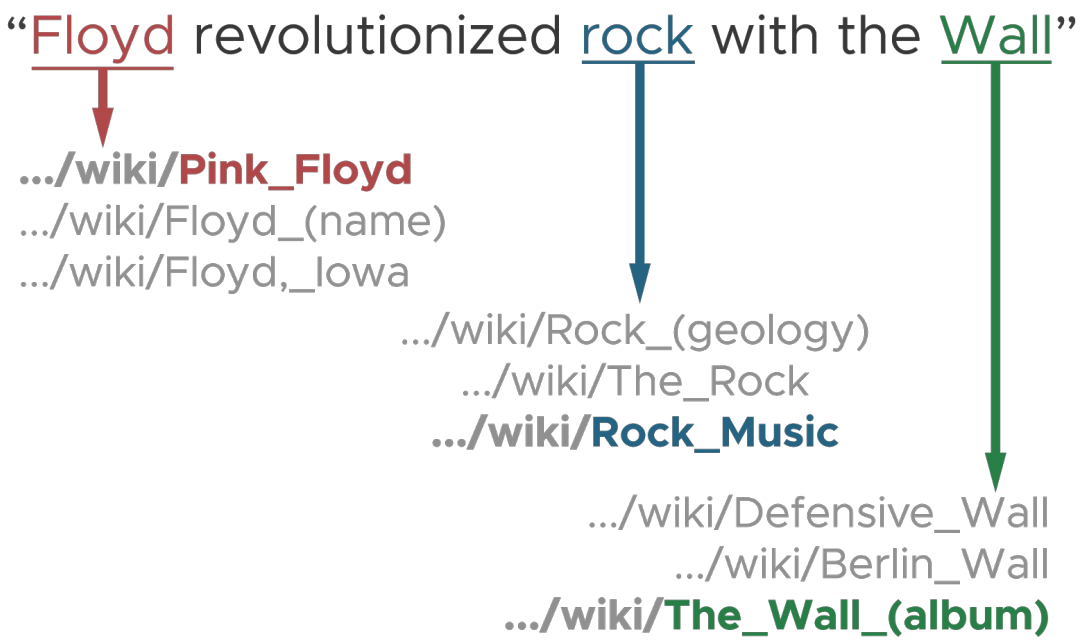
A CRF works by modeling the conditional probability distribution of the label sequence given the input sequence

$$p(\mathbf{y}|\mathbf{x}) = \underbrace{\frac{1}{Z(\mathbf{x})}}_{\text{Normalization}} \prod_{t=1}^{T} \exp \left\{ \sum_{k=1}^{K} \underbrace{\theta_k}_{\text{Weight}} \underbrace{f_k(y_t, y_{t-1}, \mathbf{x}_t)}_{\text{Feature}} \right\}$$

- t is the position of the datapoint we want to predict, y is the label, x is the feature vector

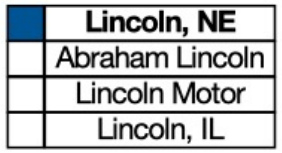
Named Entity Disambiguation (NED)

- Assigning a unique identity to entities in a text



Patterns

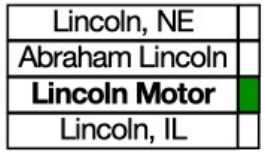
Entity Memorization
textual cues indicate a specific entity



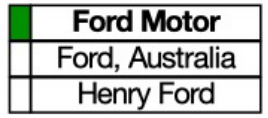
Where is Lincoln Nebraska?

co-occurrence of Lincoln, NE with "Nebraska"

Type Consistency
certain textual signals indicate that the types of entities are likely similar

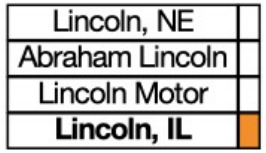


Is a Lincoln or Ford more expensive?

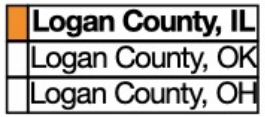


consistent "car" types with sequence keyword "or"

KG Relations
textual signals indicate that two entities have a relationship

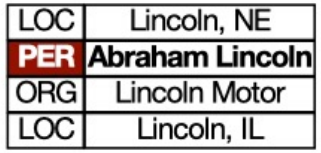


Where is Lincoln in Logan County?



"capital-of" relation with location keyword "in"

Type Affordance
textual cues indicate a specific type

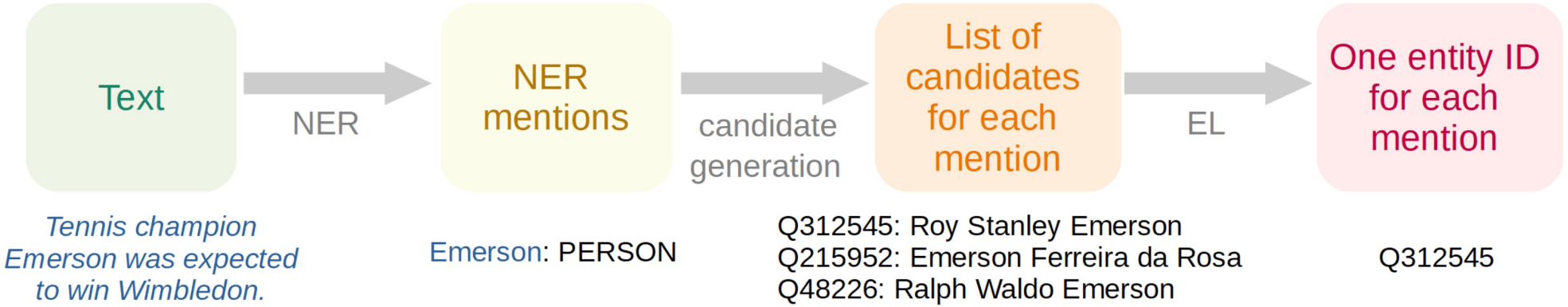


How tall is Lincoln?

"height" is associated with people types

Entity Linking (NER + NED)

- Linking entities in a text with a unique entity in a knowledge base



Relationship Extraction

- The next step, connecting entities in some type of relation
- Approaches
  - Rules like “PER [something] of ORG” -> is-a-part-of relation
  - Binary classifier (are two entities related?)
  - Multiclass classification (what kind of relation is it?)
  - Unsupervised

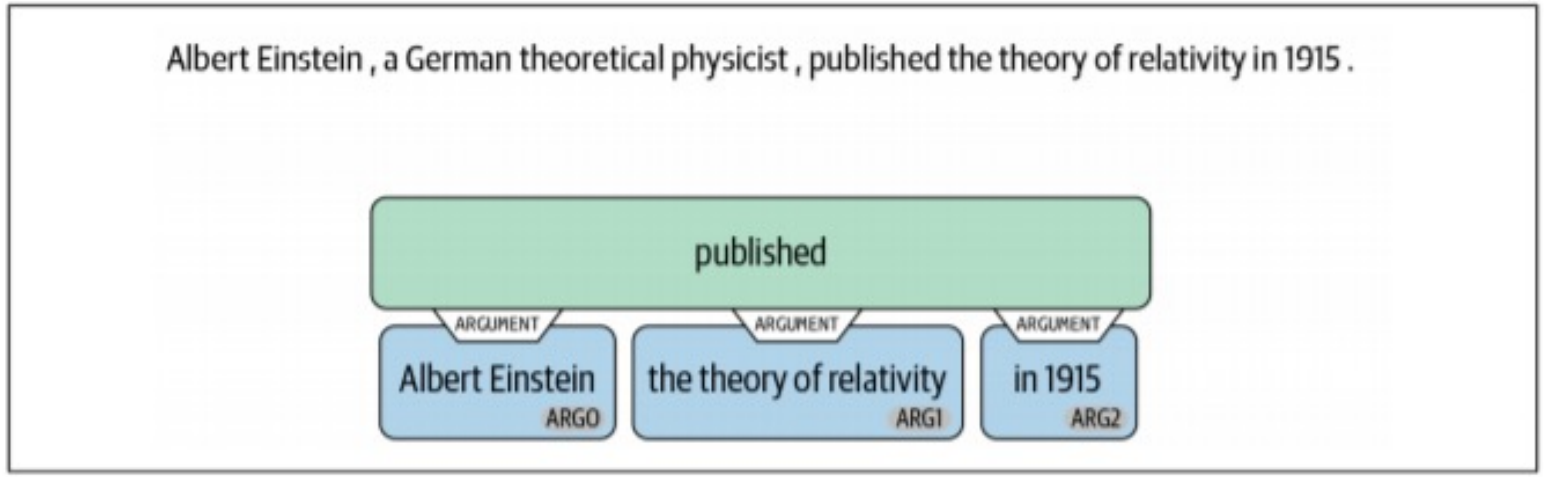


Figure 5-11. Open IE demo by AllenNLP

Other activities

- Temporal information extraction
- Event extraction
- Template filling
- Assertion status detection
- Semantic annotation
- etc.