University of BRISTOL

5.1. Information Extraction

Edwin Simpson

Department of Computer Science,

University of Bristol, UK.

Information Extraction

Chapter 17, Speech and Language Processing, 3rd edition draft, Jurafsky & Martin (2021).

What is the relationship between airline announcements of fare increases and the behaviour of airline stocks the next day?

- Think about how you would answer this question.
- We can get hold of historical stock price data fairly easily from various web-based services.
- What other information do you need to relate the stock prices to the fare announcements? **Write down two pieces of information you need to know.**

Information Extraction

- **Did your answers match with any of these?**
- The names of airlines that have announced fare changes.
- How much the fare will be changing.
- Which routes are affected.
- When the change takes effect from.

Information Extraction

Chapter 17, Speech and Language Processing, 3rd edition draft, Jurafsky & Martin (2021).

- We can get all this information from a news article, for example:

"Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit of AMR Corp., immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL Corp., said the increase took effect Thursday and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Denver to San Francisco."

Information Extraction (IE)

- The information is all in the text, but it’s **unstructured**.
- IE extracts structured information from text:
  - Represents different pieces of information in a common format
  - E.g., a database table.

UNITED AIRLINES:	SPOKESPERSON	TIM WAGNER

FAIR RAISE ATTEMPT:	LEAD AIRLINE	UNITED AIRLINES
	AMOUNT	$6
	EFFECTIVE DATE	2006-10-26

Information Extraction (IE)

Chapter 17, Speech and Language Processing, 3rd edition draft, Jurafsky & Martin (2021).

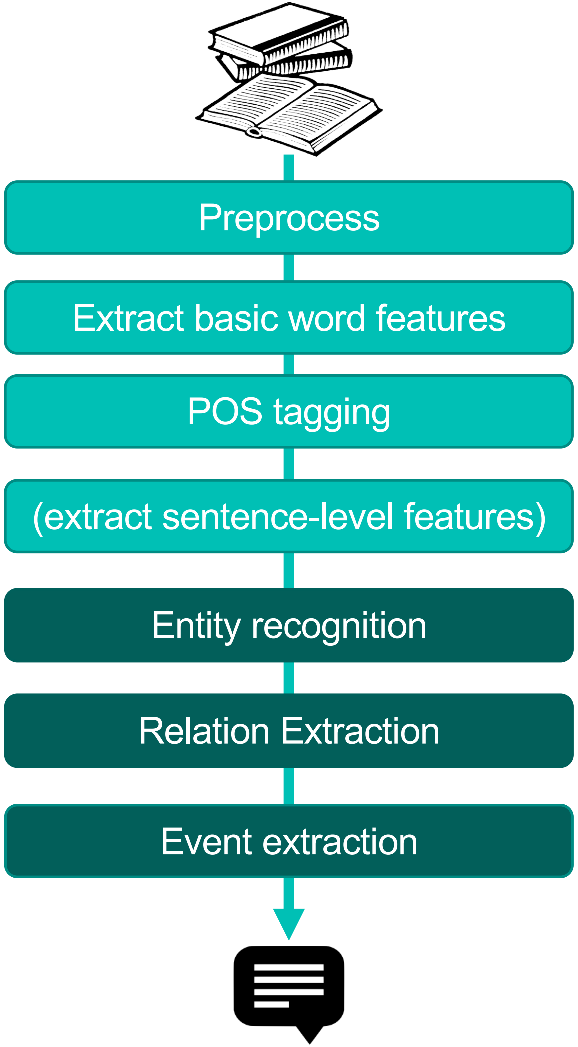
- IE involves several different steps:
  - Recognising entities (this week)
  - Finding semantic relations between entities (next week!)
  - Extracting information related to a single event (not in this unit):

UNITED AIRLINES:	SPOKESPERSON	TIM WAGNER

FAIR RAISE ATTEMPT:	LEAD AIRLINE	UNITED AIRLINES
	AMOUNT	$6
	EFFECTIVE DATE	2006-10-26

Information Extraction (IE)

- IE processes the features extracted at lower levels, such as word and syntax features.
- IE processes text at the semantic level to extract meaning.
- Its results are used in downstream tasks



Preprocess

Extract basic word features

POS tagging

(extract sentence-level features)

Entity recognition

Relation Extraction

Event extraction

Named Entities

Section 8.3, Speech and Language Processing, 3rd edition draft, Jurafsky & Martin (2021).

Named Entities

Section 8.3, Speech and Language Processing, 3rd edition draft, Jurafsky & Martin (2021).

- Look at the following excerpt from a news article again:
  1. Pause the video.
  2. Could you highlight the “entities”?
  3. How would you categorise the entities?

“Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit of AMR Corp., immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL Corp., said the increase took effect Thursday and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Denver to San Francisco.”

Named Entities

“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 Entities

- What is the definition of a named entity?

“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 Entities

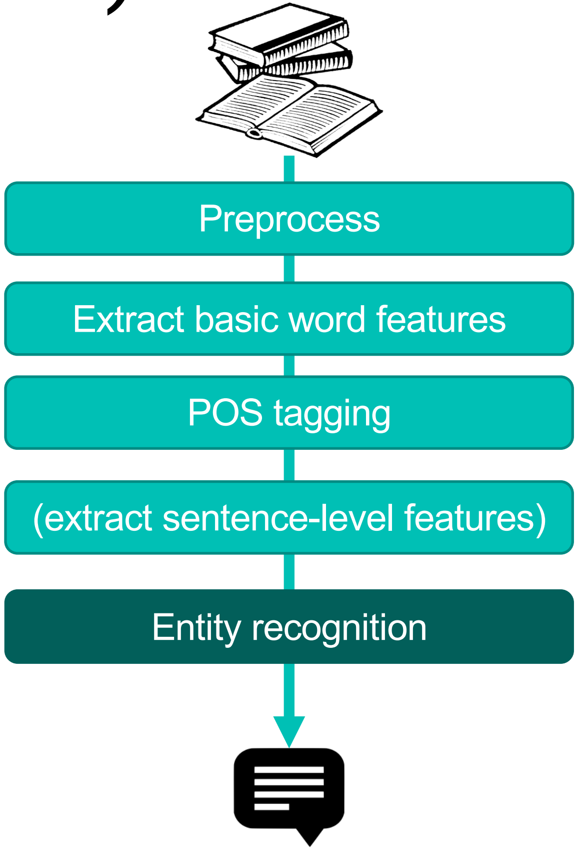
- What is the definition of a named entity?
- Anything that can be referred to by a proper name.
- Types can be domain-specific, e.g., PRODUCT in a shop review.

Table from https://www.nltk.org/book/ch07.html

Type	Examples
ORGANIZATION	Georgia-Pacific Corp., WHO
PERSON	Eddy Bonte, President Obama
LOCATION	Murray River, Mount Everest
DATE	June, 2008-06-29
TIME	two fifty a m, 1:30 p.m.
MONEY	175 million Canadian Dollars, GBP 10.40
PERCENT	twenty pct, 18.75 %
FACILITY	Washington Monument, Stonehenge
GPE	South East Asia, Midlothian

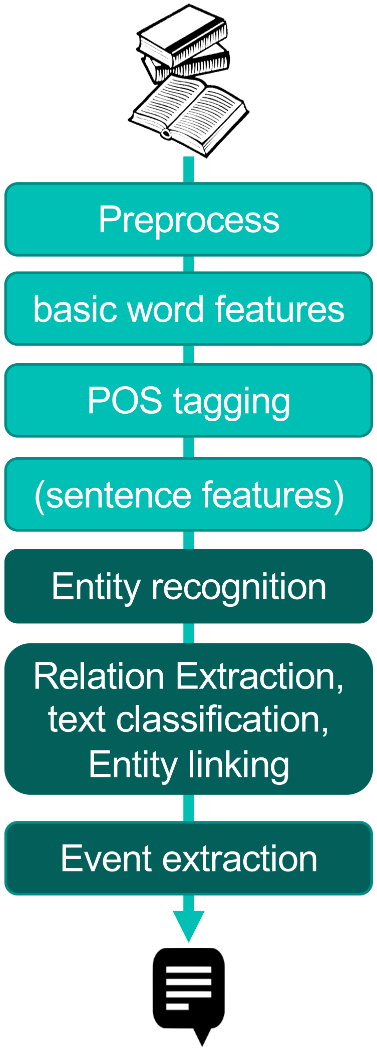
Named Entity Recognition (NER)

- Goal: label the entity spans in a text sequence.
- Named entities are useful information to extract of themselves, e.g.,:
  - To enable searching for particular entities.
  - To highlight important information in a piece of text.


Preprocess
Extract basic word features
POS tagging
(extract sentence-level features)
Entity recognition

Downstream Tasks

- A prerequisite for other IE steps like relation and event extraction.
- Tasks like sentiment analysis, where we need to identify which entity is the target of the author’s sentiment.
- Linking mentions of an entity to its Wikipedia page.
- In some text classification tasks, replace proper nouns with named entity tags to improve generalisation:
  - E.g. to avoid associating negative sentiment with a person’s name…
  - Use the PER tag instead so that the model has to process the context.



Preprocess

basic word features

POS tagging

(sentence features)

Entity recognition

Relation Extraction, text classification, Entity linking

Event extraction

Summary

- IE involves extracting structured information from text, e.g., to fill in entries in a database with facts about people, places and events.
- Named entity recognition (NER) is the task of identifying names of people, places, organisations, times and other things of interest in text.
- It provides useful information for information retrieval and various downstream tasks.