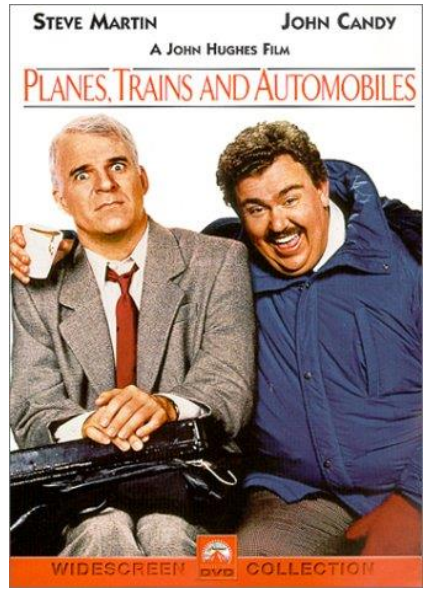
Named Entity Recognition
Chapter 22

Named Entity

- Anything that can be referred to with a proper name

  - People (PER): Individuals, fiction characters, …
  - Organization (ORG): Companies, Agencies, …
  - Location (LOC): Physical extents, mountain ranges, seas, ….
  - Geo-political entity (GPE): Countries, states, …
  - Facility (FAC): Bridges, airports, buildings
  - Vehicles (VEH): Planes, trains and automobiles



Goal of NER

Turing is often considered to be the father of modern computer science.

NER: identifying ***Turing*** is a Person (PER)

Generic NER

- Focuses on: person, location, and organization

- Specialized applications:
  - Weapons
  - **De-identification**
  - Drug-drug interactions
  - Nano-particle characteristics
  - Works of art
  - Proteins
  - Genes

Extended NER

- The notion of NER is commonly extended to include things that are not entities per se
  - Temporal expressions
    - dates, times, named events
  - Numerical expressions
    - Measurements, prices, counts

Example of Annotated NER Text

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, 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].

Text Contains

- 12 mentions
  - 4 organizations (ORG)
  - 4 locations (LOC)
  - 2 times (TIME)
  - 1 person (PER)
  - 1 money (MONEY)

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, 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].

Ambiguity in NER

- Two types of ambiguity
  - The mention can refer to different entities of the same type
    - JFK could refer to the former president or his son
  - The mention can refer to more than one entity type
    - JFK could be a person (PER) or an airport (LOC)

Sequence problems

- Many problems in NLP have data which is a sequence of characters, words, phrases, lines, or sentences …
- We can think of our task as one of labeling each item

VBG	NN	IN	DT	NN	IN	NN
Chasing	opportunity	in	an	age	of	upheaval

POS tagging

PERS	O	O	O	ORG	ORG
Murdoch	discusses	future	of	News	Corp.

Named entity recognition

B	B	I	I	B	I	B	I	B	B
而	相	对	于	这	些	品	牌	的	价

Word segmentation

Text segmen-tation

Encoding classes for sequence labeling

	IO encoding	IOB encoding
Fred	PER	B-PER
showed	O	O
Sue	PER	B-PER
Mengqiu	PER	B-PER
Huang	PER	I-PER
‘s	O	O
new	O	O
painting	O	O

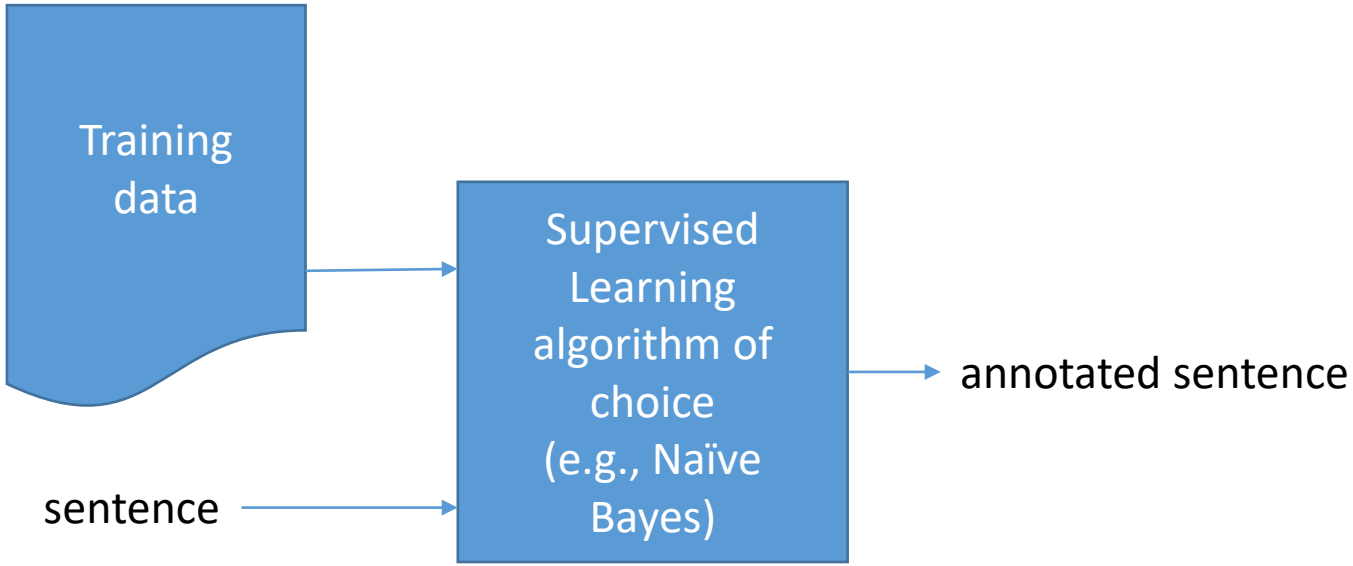
NER as Sequence Labeling

- **Standard approach** to NER:
  - Word-by-word labeling task
  - **Classifiers** are trained to label the tokens in a text with tags that indicate the presence of a particular kind of name entity

[ORG American Airlines], a unit of [ORG AMR Corp.], immediately matched the move, spokesman [PER Time Wagner] said.

Words	Label
American	Borg
Airlines	Iorg
,	O
a	O
unit	O
of	O
AMR	Borg
Corp.	Iorg
,	O
immediately	O
matched	O
the	O
move	O
,	O
spokesman	O
Tim	Bper
Wagner	Iper
said	O
.	O

Supervised Learning NER System



Christopher Manning

The ML sequence model approach to NER

Training

1. Collect a set of representative training documents
2. Label each token for its entity class or other (O)
3. Design feature extractors appropriate to the text and classes
4. Train a sequence classifier to predict the labels from the data

Testing

1. Receive a set of testing documents
2. Run sequence model inference to label each token
3. Appropriately output the recognized entities

Stanford Slide Added by Dr. Liao

Supervised Learning Algorithms

- Commonly used:
  - Sequence prediction algorithms:
    - CRFs
      - Conditional Random Field (CRF) model -> 1st-order linear-chain Markov
      - Eg., **Stanford NER** (Developed in Java)
    - HMMs

$$Estimating\colon\ P(y_i|x_{i-k}\ldots x_{i+l}, y_{i-m}\ldots y_{i-1})$$

where:

$$X = (x_1, \ldots, x_N)\ \ is\ an\ \boldsymbol{input}\ sequence\ (your\ sentence)$$
$$Y = (y_1, \ldots, y_N)\ is\ the\ \boldsymbol{output}\ sequence\ (NER\ tags)$$

Contents Added by Dr. Liao

Features

Feature	Explanation
Lexical items	The token to be labeled
Stemmed Lexical Item	The stem of the token to be labeled
Shape	Orthographic pattern of the target word
Character affixes	Character-level affixes of the target and surrounding words
Syntactic chunk label	Base-phrase chunk label
Gazetteer	Presence of word in one or more named entity lists
Predictive token(s)	Presence of predictive words in surrounding text
Bag of word / Bag of n-grams	Words and/or n-grams of the surrounding context

Shape Features

Shape Feature	Example
Lower	cummings
Capitalized	Washington
All caps	IRA
Mixed case	eBay
Capitalized character with period	H.
Ends in digit	A9
Contains hyphen	H-P

Predictive Words

Predictive Feature	Entity
Mr.	Person
Rev.	Person
MD	Person
Inc.	Organization
Corp.	Organization

Gazeteers

- Where do these Gazeteers come from:
  - Previously:
    - Census data
    - Lists of companies
  - Now: Wikipedia
    - Artwork: novels, books, paintings, operas, plays
    - Named Objects: aircraft tanks, rifles, weapons
    - Events: playoffs, championships, races

Available NER Systems

- Apache OpenNLP
  - https://opennlp.apache.org
- NameFinder module (OpenNLP NER)
  - https://opennlp.apache.org/docs/1.5.3/manual/opennlp.html#tools.namefind
- Stanford NER
  - https://nlp.stanford.edu/software/CRF-NER.html
- UIUC NET
  - http://cogcomp.org/page/demo_view/ner

Next Up

- Rest of today:
  - Applications of NER: De-identification

- Coming up:
  - Information Retrieval (read Chapter 23)