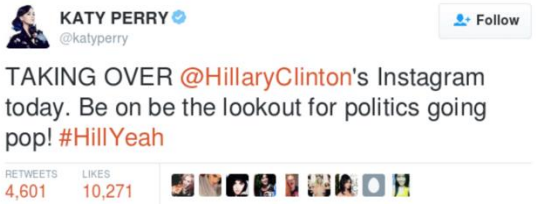
Named Entity Recognition

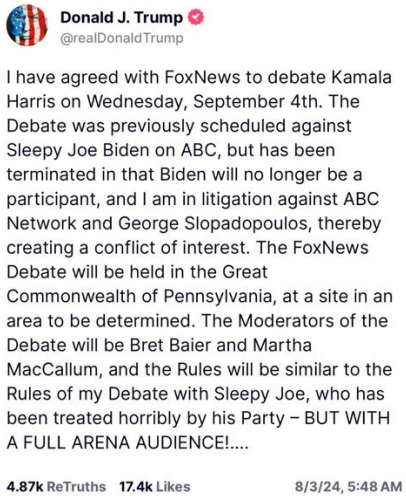
Information Extraction

Information Extraction

- Usually from unstructured or semi-structured data
- Examples
  - Online stories
  - News
  - Scientific papers
  - Resumes
- Entities
  - Who, when, where
- Build knowledge base

Named Entities





What named entities are mentioned?

Answer: Hillary, FoxNews, Kamala, Joe Biden, ABC, and so on

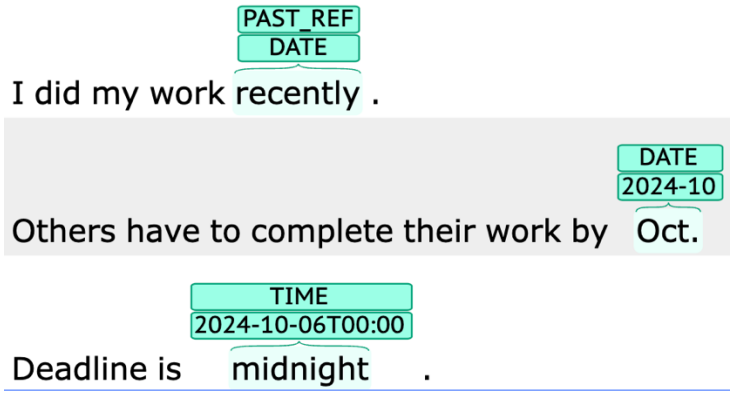
Named Entities

- Types:
  - People
  - Locations
  - Organizations
    - Teams, Newspapers, Companies
  - Geo-political entities
- Ambiguity:
  - London can be a person or a place (Adam London–English Cricketer)
- Useful for interfaces to databases, question answering, etc.

Times and Dates

Absolute expressions

Relative expressions (e.g., “last night”)



Event Extraction

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**].

Event Extraction

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

Dictionary Based Methods

- Segmentation
  - Which words belong to a named entity?
  - Brazilian football legend Pele's condition has improved, according to a Thursday evening statement from a Sao Paulo hospital.
- Classification
  - What type of named entity is it?
  - Use dictionary-based methods: gazetteers etc.

Named Entities

Type	Tag	Sample Categories	Example sentences
People	PER	people, characters	**Turing** is a giant of computer science.
Organization	ORG	companies, sports teams	The **IPCC** warned about the cyclone.
Location	LOC	regions, mountains, seas	The **Mt. Sanitas** loop is in **Sunshine Canyon**.
Geo-Political Entity	GPE	countries, states, provinces	**Palo Alto** is raising the fees for parking.

Let's try Dictionaries!

- Brazilian football legend Pele's condition has improved, according to a Thursday evening statement from a Sao Paulo hospital.
- There had been earlier concerns about Pele's health after Albert Einstein Hospital issued a release that said his condition was "unstable."
- Thursday night's release said Pele was relocated to the intensive care unit because a kidney dialysis machine he needed was in ICU.

Will Dictionary Work Here?

- [Brazilian] football legend [$_{PERSON}$ Pele]'s condition has improved, according to a [$_{TIME}$ Thursday evening] statement from a [$_{LOCATION}$ Sao Paulo] hospital.
- There had been earlier concerns about [Pele]'s health after [$_{ORG}$ Albert Einstein Hospital] issued a release that said his condition was "unstable."
- [$_{TIME}$ Thursday night]'s release said [Pele] was relocated] to the intensive care unit because a kidney dialysis machine he needed was in ICU.

Ambiguity

Name	Possible Categories
Washington	Person, Location, Political Entity, Organization, Vehicle
Downing St.	Location, Organization
IRA	Person, Organization, Monetary Instrument
Louis Vuitton	Person, Organization, Commercial Product

Figure 21.2 Common categorical ambiguities associated with various proper names.

[$_{PER}$ Washington] was born into slavery on the farm of James Burroughs.
[$_{ORG}$ Washington] went up 2 games to 1 in the four-game series.
Blair arrived in [$_{LOC}$ Washington] for what may well be his last state visit.
In June, [$_{GPE}$ Washington] passed a primary seatbelt law.
The [$_{VEH}$ Washington] had proved to be a leaky ship, every passage I made...

Figure 21.3 Examples of type ambiguities in the use of the name *Washington*.

NER Extraction Features

embeddings for w_i, embeddings for neighboring words
part of speech of w_i, part of speech of neighboring words
presence of w_i in a **gazetteer**
w_i contains a particular prefix (from all prefixes of length ≤ 4)
w_i contains a particular suffix (from all suffixes of length ≤ 4)
word shape of w_i, word shape of neighboring words
short word shape of w_i, short word shape of neighboring words
gazetteer features

Figure 17.15 Typical features for a feature-based NER system.

NER with Sequence Tagging

Sequence tagging is a common ML approach to NER.

Tokens are labeled as one of:

- `B`: Beginning of an entity
- `I`: Inside an entity
- `O`: Outside an entity

We train a Machine Learning model on a variety of text features to accomplish this. We'll see how to do this in the next session.

Word	Label	Tag
American	B	ORG
Airlines	I	ORG
a	O	–
unit	O	–
of	O	–
AMR	B	ORG
Corp.	I	ORG
immediately	O	–
matched	O	–
the	O	–
move	O	–
spokesman	O	–
Tim	B	PERS
Wagner	I	PERS
said	O	–

Word Shape

In English, the shape feature is one of the most predictive of entity names.

It is particularly useful for identifying businesses and products like Yahoo!, eBay, or iMac.

Shape is also a strong predictor of certain technical terms, such as gene names.

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

NER Extraction Features

$\text{prefix}(w_i) = \texttt{L}$
$\text{prefix}(w_i) = \texttt{L'}$
$\text{prefix}(w_i) = \texttt{L'O}$
$\text{prefix}(w_i) = \texttt{L'Oc}$
$\text{suffix}(w_i) = \texttt{tane}$
$\text{suffix}(w_i) = \texttt{ane}$
$\text{suffix}(w_i) = \texttt{ne}$
$\text{suffix}(w_i) = \texttt{e}$
$\text{word-shape}(w_i) = \texttt{X'Xxxxxxxx}$
$\text{short-word-shape}(w_i) = \texttt{X'Xx}$

Word: L'Occitane
or
Occitanie

Let’s Extract!

Jane Villanueva of United Airlines Holding discussed the Chicago route

Word shape, prefix/suffix of length 4, Gazetteer, BIO label

Feature Encoding in NER

Words	POS	Short shape	Gazetteer	BIO Label
Jane	NNP	Xx	0	B-PER
Villanueva	NNP	Xx	1	I-PER
of	IN	x	0	O
United	NNP	Xx	0	B-ORG
Airlines	NNP	Xx	0	I-ORG
Holding	NNP	Xx	0	I-ORG
discussed	VBD	x	0	O
the	DT	x	0	O
Chicago	NNP	Xx	1	B-LOC
route	NN	x	0	O
.	.	.	0	O

Figure 17.16 Some NER features for a sample sentence, assuming that Chicago and Villanueva are listed as locations in a gazetteer. We assume features only take on the values 0 or 1, so the first POS feature, for example, would be represented as $\mathbb{1}\{\text{POS} = \text{NNP}\}$.

NER as Sequence Labeling

- Many language problems can be cast as sequence labeling problems
  - POS – part of speech tagging
  - NER – named entity recognition
  - SRL – semantic role labeling
- Input
  - Sequence $w_1 w_2 w_3$
- Output
  - Labeled words
- Classification methods
  - Can use the categories of the previous tokens as features in classifying the next one
  - Direction matters

NER as Sequence Labeling

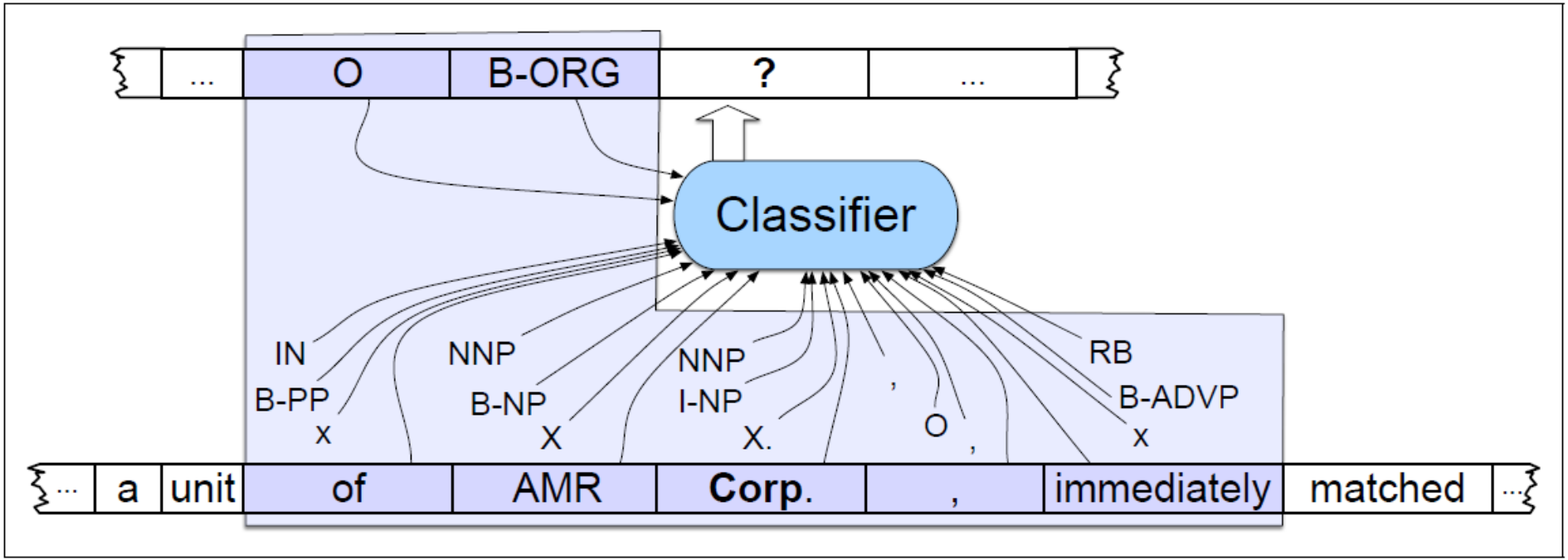


Figure 21.7 Named entity recognition as sequence labeling. The features available to the classifier during training and classification are those in the boxed area.

NER Demos

- http://nlp.stanford.edu:8080/ner/
- http://cogcomp.org/page/demo_view/ner
- http://demo.allennlp.org/named-entity-recognition

Other Examples

- Job announcements on online platforms
  - Location, title, starting date, qualifications, salary
- Seminar announcements
  - Time, title, location, speaker
- Medical papers
  - Drug, disease, gene/protein, cell line, species, substance

Filling the Templates

- Some fields get filled by text from the document
  - E.g., the names of people
- Others can be pre-defined values
  - E.g., successful/unsuccessful merger
- Some fields allow for multiple values

Evaluation metrics

- Precision, recall, F1

Acknowledgments

Lectures from Yale and Northeastern University