Named Entity Recognition

Natural Language Processing
CSCI 6350
Dr. Sal Barbosa

Named Entity Recognition

- Specific type of information extraction in which the goal is to extract formal names of particular types of entities such as people, places, organizations, etc.
- Usually a preprocessing step for subsequent task-specific IE, or other tasks such as question answering.

2

from slides by Raymond Mooney, University of Austin

Named Entity Recognition Example

U.S. Supreme Court quashes 'illegal' Guantanamo trials

Military trials arranged by the Bush administration for detainees at Guantanamo Bay are illegal, the United States Supreme Court ruled Thursday. The court found that the trials — known as military commissions — for people detained on suspicion of terrorist activity abroad do not conform to any act of Congress. The justices also rejected the government's argument that the Geneva Conventions regarding prisoners of war do not apply to those held at Guantanamo Bay. Writing for the 5-3 majority, Justice Stephen Breyer said the White House had overstepped its powers under the U.S. Constitution. "Congress has not issued the executive a blank cheque," Breyer wrote.

President George W. Bush said he takes the ruling very seriously and would find a way to both respect the court's findings and protect the American people.

3

from slides by Raymond Mooney, University of Austin

Named Entity Recognition Example

people **places** **organizations**

U.S. Supreme Court quashes 'illegal' Guantanamo trials

Military trials arranged by the Bush administration for detainees at Guantanamo Bay are illegal, the United States Supreme Court ruled Thursday. The court found that the trials — known as military commissions — for people detained on suspicion of terrorist activity abroad do not conform to any act of Congress. The justices also rejected the government's argument that the Geneva Conventions regarding prisoners of war do not apply to those held at Guantanamo Bay. Writing for the 5-3 majority, Justice Stephen Breyer said the White House had overstepped its powers under the U.S. Constitution. "Congress has not issued the executive a blank cheque," Breyer wrote.

President George W. Bush said he takes the ruling very seriously and would find a way to both respect the court's findings and protect the American people.

4

from slides by Raymond Mooney, University of Austin

Named Entity Types and Examples

Type	Tag	Sample Categories
People	PER	Individuals, fictional characters, small groups
Organization	ORG	Companies, agencies, political parties, religious groups, sports teams
Location	LOC	Physical extents, mountains, lakes, seas
Geo-Political Entity	GPE	Countries, states, provinces, counties
Facility	FAC	Bridges, buildings, airports
Vehicles	VEH	Planes, trains, and automobiles

Type	Example
People	*Turing* is often considered to be the father of modern computer science.
Organization	The *IPCC* said it is likely that future tropical cyclones will become more intense.
Location	The *Mt. Sanitas* loop hike begins at the base of *Sunshine Canyon*.
Geo-Political Entity	*Palo Alto* is looking at raising the fees for parking in the University Avenue district.
Facility	Drivers were advised to consider either the *Tappan Zee Bridge* or the *Lincoln Tunnel*.
Vehicles	The updated *Mini Cooper* retains its charm and agility.

Ambiguous Named Entity Types

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

[$_{PERS}$ 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 [$_{FAC}$ Washington] had proved to be a leaky ship, every passage I made...

Features Used in Training NER

identity of w_i, identity of neighboring words
embeddings for w_i, embeddings for neighboring words
part of speech of w_i, part of speech of neighboring words
base-phrase syntactic chunk label of w_i and 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)
w_i is all upper case
word shape of w_i, word shape of neighboring words
short word shape of w_i, short word shape of neighboring words
presence of hyphen

Shape Features

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

Feature Encoding for NER

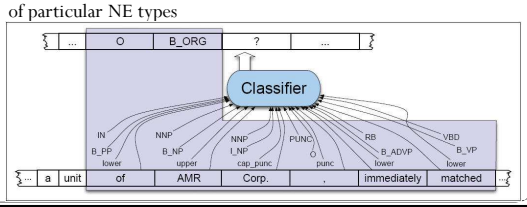
	Features				Label
	American	NNP	B_{NP}	cap	B_{ORG}
	Airlines	NNPS	I_{NP}	cap	I_{ORG}
	,	PUNC	O	punc	O
	a	DT	B_{NP}	lower	O
	unit	NN	I_{NP}	lower	O
	of	IN	B_{PP}	lower	O
	AMR	NNP	B_{NP}	upper	B_{ORG}
	Corp.	NNP	I_{NP}	cap_punc	I_{ORG}
	,	PUNC	O	punc	O
	immediately	RB	B_{ADVP}	lower	O
	matched	VBD	B_{VP}	lower	O
	the	DT	B_{NP}	lower	O
	move	NN	I_{NP}	lower	O
	,	PUNC	O	punc	O
	spokesman	NN	B_{NP}	lower	O
	Tim	NNP	I_{NP}	cap	B_{PER}
	Wagner	NNP	I_{NP}	cap	I_{PER}
	said	VBD	B_{VP}	lower	O
	.	PUNC	O	punc	O

9

from Speech and Language Processing by Jurafsky and Martin

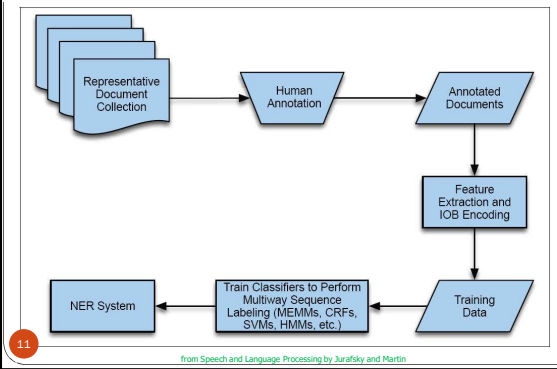
NER as Sequence Labeling Task

- The standard algorithm for named entity recognition is as a word-by-word sequence labeling task
- Assigned tags capture both the boundary and the NE type
- Sequence classifiers like MEMM/CRF, bi-LSTM, or transformers are trained to label the tokens with tags that indicate the presence of particular NE types



10

Statistical Sequence Labeling



11

from Speech and Language Processing by Jurafsky and Martin

Code Demo

12