Named Entity Recognition

LING 570

Fei Xia

Week 10: 11/30/09

Outline

- What is NER? Why NER?

- Common approach

- J&M Ch 22.1

What is NER?

- Task: Locate named entities in (usually) unstructured text

- Entities of interest include:
  - Person names
  - Location
  - Organization
  - Dates, times (relative and absolute)
  - Numbers
  - …

An example

- Microsoft released Windows Vista in 2007.

- <ORG>Microsoft</ORG> released <PRODUCT>Windows Vista</PRODUCT> in <YEAR>2007</YEAR>

- NE tags are often application-specific.

Why NER?

- Machine Translation:
  - E.g., translation of numbers, personal names
  - Ex1: 123,456,789 => 1,2345,6789

    thirty thousand => 三 (two) 万 (10-thousand)
  - Ex2: 李 ➔ Li, Lee, …

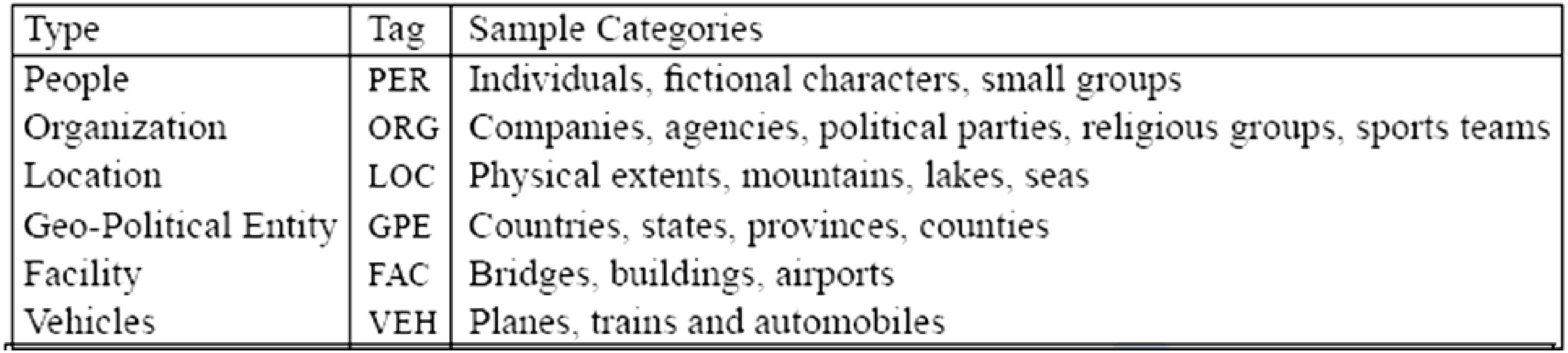
- IE:
  - Microsoft released Windows Vista in 2007.

  ➔ Company: Microsoft
  Product: Windows Vista
  Time: 2007

- IR

- Text-to-speech synthesis: 345 6789

Common NE categories



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

Ambiguity

- If all goes well, MATSUSHITA AND ROBERT BOSCH will … : person, or company
- Washington announced …: Location-for-organization
- Boston Power and Light …: one entity or two
- JFK: Person, Airport, Street

Evaluation

- Precision
- Recall
- F-score

Resources for NER

- Name lists:
  - Who-is-who lists: Famous people names
  - U.S. Securities and Exchange Commission - list of company names
  - Gazetteers: list of place names

- Tools:
  - LingPipe (on Patas)
  - OAK

Common methods:

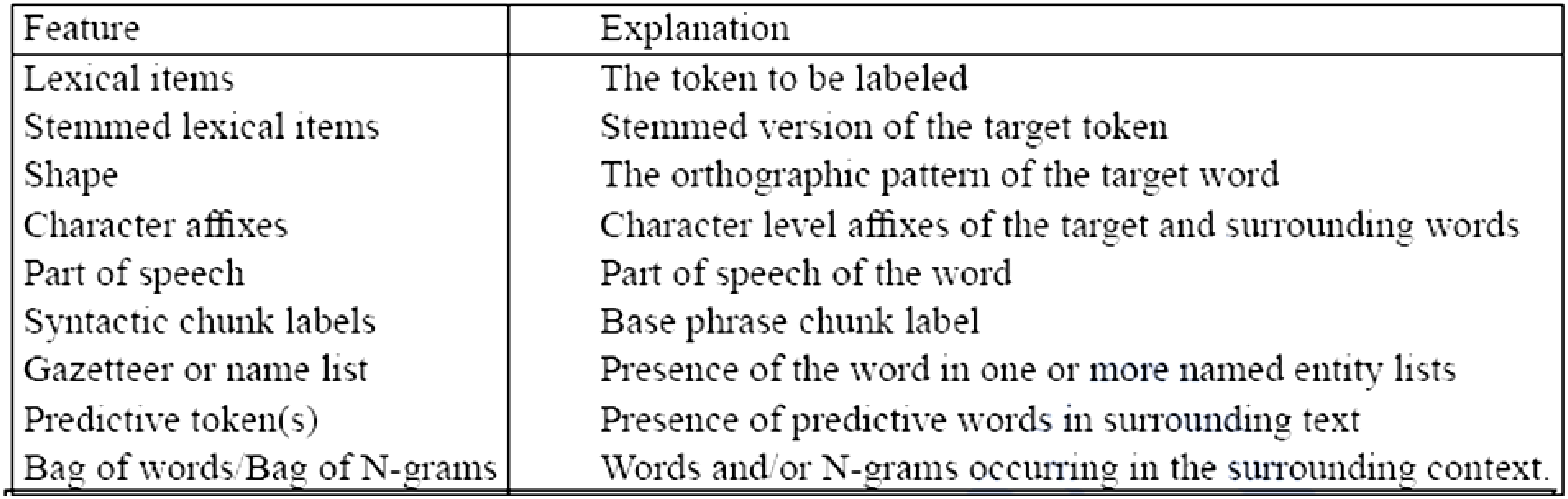
- Rule-based: regex patterns
  - Numbers:
  - Date: 07/08/06 (mm/dd/yy, dd/mm/yy, yy/mm/dd)
  - Money, etc.

- ML approaches: as sequence labeling
  - Proper names
  - Organization
  - Product
  - …

- Hybrid approach

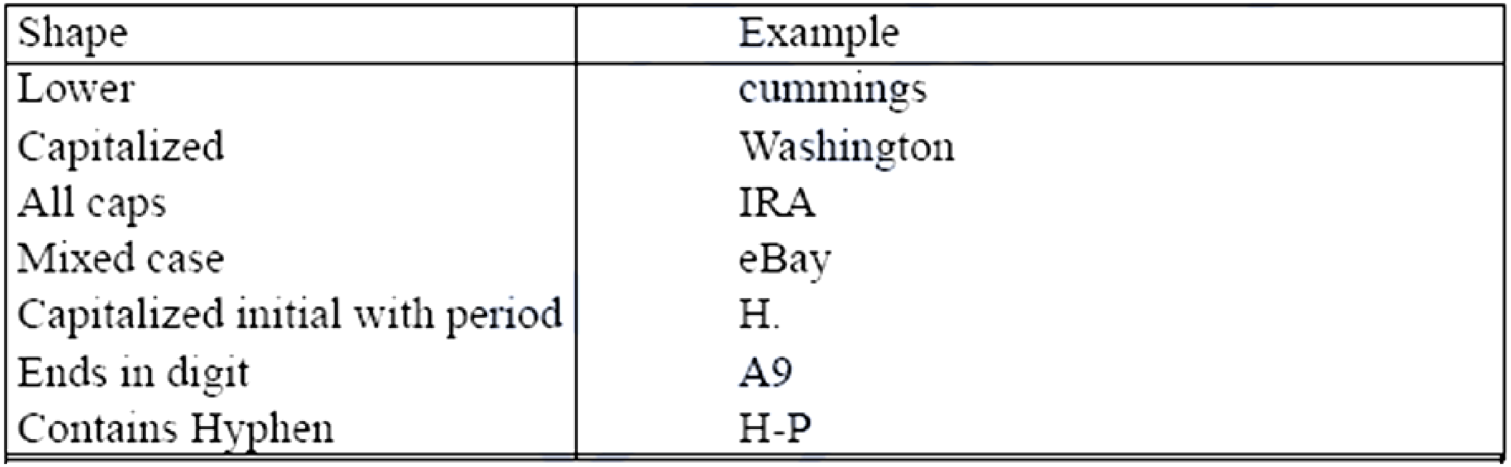
NER as sequence labeling problem

Commonly used features



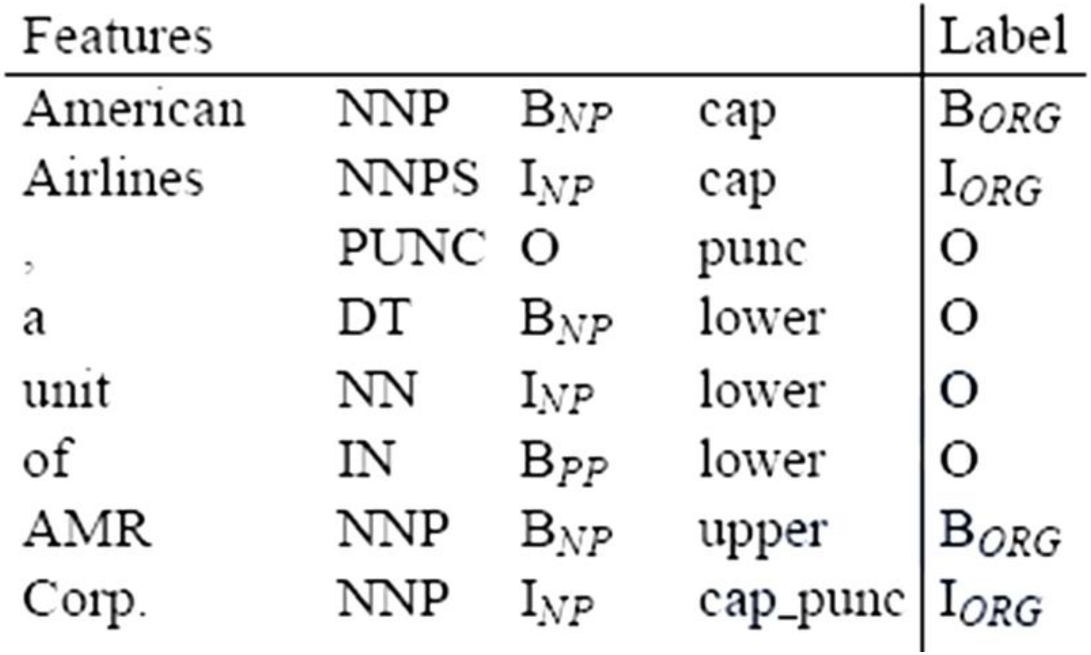
Feature	Explanation
Lexical items	The token to be labeled
Stemmed lexical items	Stemmed version of the target token
Shape	The orthographic pattern of the target word
Character affixes	Character level affixes of the target and surrounding words
Part of speech	Part of speech of the word
Syntactic chunk labels	Base phrase chunk label
Gazetteer or name list	Presence of the word in one or more named entity lists
Predictive token(s)	Presence of predictive words in surrounding text
Bag of words/Bag of N-grams	Words and/or N-grams occurring in the surrounding context.

Shape features



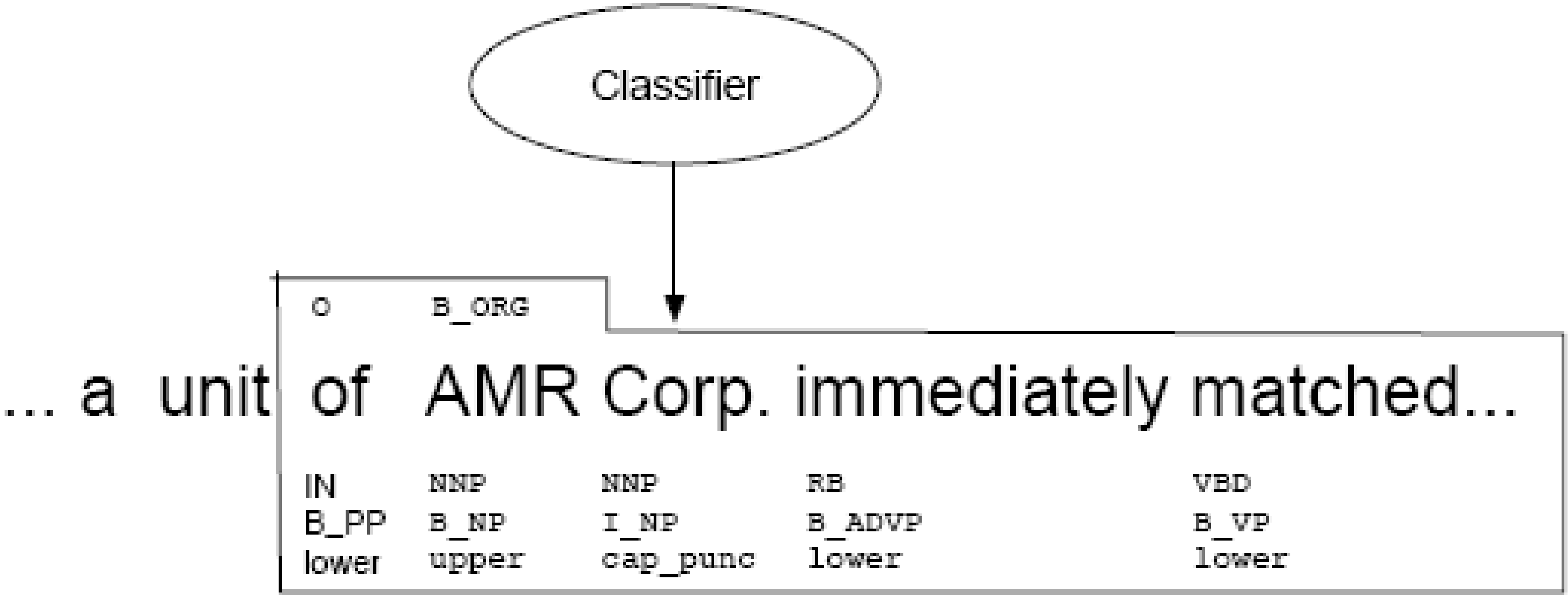
Shape	Example
Lower	cummings
Capitalized	Washington
All caps	IRA
Mixed case	eBay
Capitalized initial with period	H.
Ends in digit	A9
Contains Hyphen	H-P

An example



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}

Sequence labeling problem



Hybrid approaches

- Use both Regex patterns and supervised learning.

- Multiple passes:
  - First, apply sure rules that are high precision but low recall.

  - Then employ more error-prone statistical methods that take the output of the first pass into account