Named Entity Recognition

LING 570

Fei Xia

Outline

- What is NER? Why NER?
- Common approach
- J&M-ed3 Ch 17.1

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

- Apple released iPhone X in 2017.

- <ORG>Apple</ORG> released <PRODUCT>iPhone X</PRODUCT> in <YEAR>2017</YEAR>

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

NE tags are often application-specific:

- News: people, country, organization, dates, etc.
- Medical records: disease, medication, dosage, frequency, organism, etc.

Why NER?

- Machine Translation:
  - E.g., translation of numbers, personal names
  - Ex1: 123,456,789 => 1,2345,6789
    thirty thousand => 30000 => 3,0000 => 三 (three) 万 (10-thousand)
  - Ex2: 12/6/10 => 2010-12-6, 2010-6-12, 2012-6-10, …
  - Ex3: 李 ➔ Li, Lee

- IE:
  - Apple released iPhone X in 2017.

  ➔ Company: Apple
  Product: iPhone X
  Time: 2017

- IR: named entities focus of retrieval

- Text-to-speech synthesis: 911 (number vs. phone number), 9/11 (date vs. ratio)

Ambiguity

- If all goes well, MATSUSHITA AND ROBERT BOSCH will … : person, or company
- Washington chose …: state, city, country, person, univ, team, etc.
- Boston Power and Light …: one entity or two
- JFK: person, airport, street

Context & Ambiguity

[$_{PERS}$ 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...

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:
  - Stanford NLP package
  - 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.
- Machine learning via sequence labeling
  - Proper names
  - Organization
  - Product
  - …
- Hybrid approach

NER as sequence labeling problem

Use a classifier

- BIO scheme:
  - B-X: the 1st word in X
  - I-X: inside X
  - O: outside of any NE entities
- Any 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

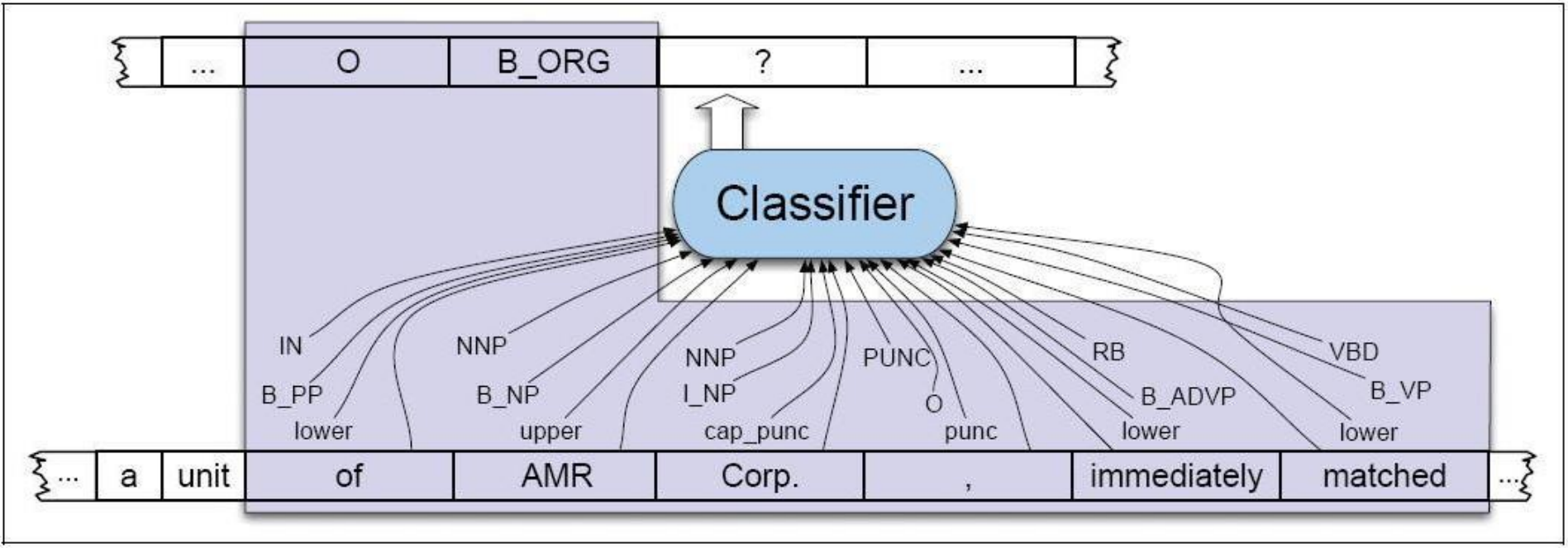
NER as Classification: Shape Features

- Shape types:
  - All lower case: e.g., company
  - Capitalized (first letter uppercase): e.g. Washington
  - all capitalized: e.g. WHO
  - mixed case: eBay
  - Capitalized with period: H.
  - Ends with 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}
,	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

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

Evaluation

- System: output of automatic tagging
- Gold Standard: true tags

- Precision: # correct chunks/# system chunks
- Recall: # correct chunks/# gold chunks
- F-measure: $F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P + R}$
- F_1 balances precision & recall

Evaluation

- Standard measures:
  - Precision, Recall, F-measure
  - Computed on entity types (Co-NLL evaluation)
- Classifiers vs evaluation measures
  - Classifiers optimize tag accuracy
    - Most common tag?
      - O – most tokens aren't NEs
  - Evaluation measures focuses on NE
- State-of-the-art:
  - Standard tasks: PER, LOC: 0.92; ORG: 0.84