Named Entity Recognition

LING 570

Fei Xia
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


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

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

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: 夏 ➔ Xia, summer
    飞 ➔ Fei, fly
    李 ➔ Li, Lee

• IE:
    ➔ Company: Microsoft
    Product: Windows Vista
    Time: 2007

• 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:
  – 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
Commonly used features

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

<table>
<thead>
<tr>
<th>Features</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>American</td>
<td>B\textsubscript{ORG}</td>
</tr>
<tr>
<td>Airlines</td>
<td>I\textsubscript{ORG}</td>
</tr>
<tr>
<td>,</td>
<td>O</td>
</tr>
<tr>
<td>a</td>
<td>O</td>
</tr>
<tr>
<td>unit</td>
<td>O</td>
</tr>
<tr>
<td>of</td>
<td>O</td>
</tr>
<tr>
<td>AMR</td>
<td>B\textsubscript{ORG}</td>
</tr>
<tr>
<td>Corp.</td>
<td>I\textsubscript{ORG}</td>
</tr>
<tr>
<td>,</td>
<td>O</td>
</tr>
<tr>
<td>immediately</td>
<td>O</td>
</tr>
<tr>
<td>matched</td>
<td>O</td>
</tr>
<tr>
<td>the</td>
<td>O</td>
</tr>
<tr>
<td>move</td>
<td>O</td>
</tr>
<tr>
<td>,</td>
<td>O</td>
</tr>
<tr>
<td>spokesman</td>
<td>O</td>
</tr>
<tr>
<td>Tim</td>
<td>B\textsubscript{PER}</td>
</tr>
<tr>
<td>Wagner</td>
<td>I\textsubscript{PER}</td>
</tr>
<tr>
<td>said</td>
<td>O</td>
</tr>
<tr>
<td>.</td>
<td>O</td>
</tr>
</tbody>
</table>
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: \[ F = \frac{(1 + \frac{2}{PR})}{\frac{2}{PR} + 1} \]
• \( F_1 \) 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