Named Entity Recognition (NER)

(Natural Language Processing)

Randil Pushpananda
rpn@ucsc.cmb.ac.lk

Named Entity Recognition (NER)

- Find and classify all the named entities in a text.
- What is a named entity?
  - A mention of an entity using its name.
    - Kansas Jayhawks
  - This is a subset of the possible mentions...
    - Kansas, Jayhawks, the team, it, they
- Find means identify the exact span of the mention
- Classify means determine the category of the entity being referred to

Proper Nouns	Common Nouns
Havaianas	slippers
Coconut	tree
Jolibee	fastfood
Acer	computer
Robert	man/boy
Maria	woman/girl

Example

- The category can be generic like Organization, Person, Location, Time, etc., or a custom category depending on the use case such as Healthcare Terms, Programming Language, etc.
- For example, an NER model detects “football“ as an entity in a paragraph and classifies it into the category of sports.

Example

- The decision by the independent MP **Andrew Wilkie** to withdraw his support for the minority **Labor** government sounded dramatic but it should not further threaten its stability. When, after the **2010** election, **Wilkie**, **Rob Oakeshott**, **Tony Windsor** and the **Greens** agreed to support **Labor**, they gave just two guarantees: confidence and supply.

Example

- The decision by the independent MP **Andrew Wilkie** to withdraw his support for the minority **Labor** government sounded dramatic but it should not further threaten its stability. When, after the **2010** election, **Wilkie**, **Rob Oakeshott**, **Tony Windsor** and the **Greens** agreed to support **Labor**, they gave just two guarantees: confidence and supply.

Person, **Date**, **Organization**

NE Types

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.

If your task is to find out

‘where’,

‘what’,

‘who’,

‘when’ from a sentence,

NER is the solution.

Ambiguity in NE

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

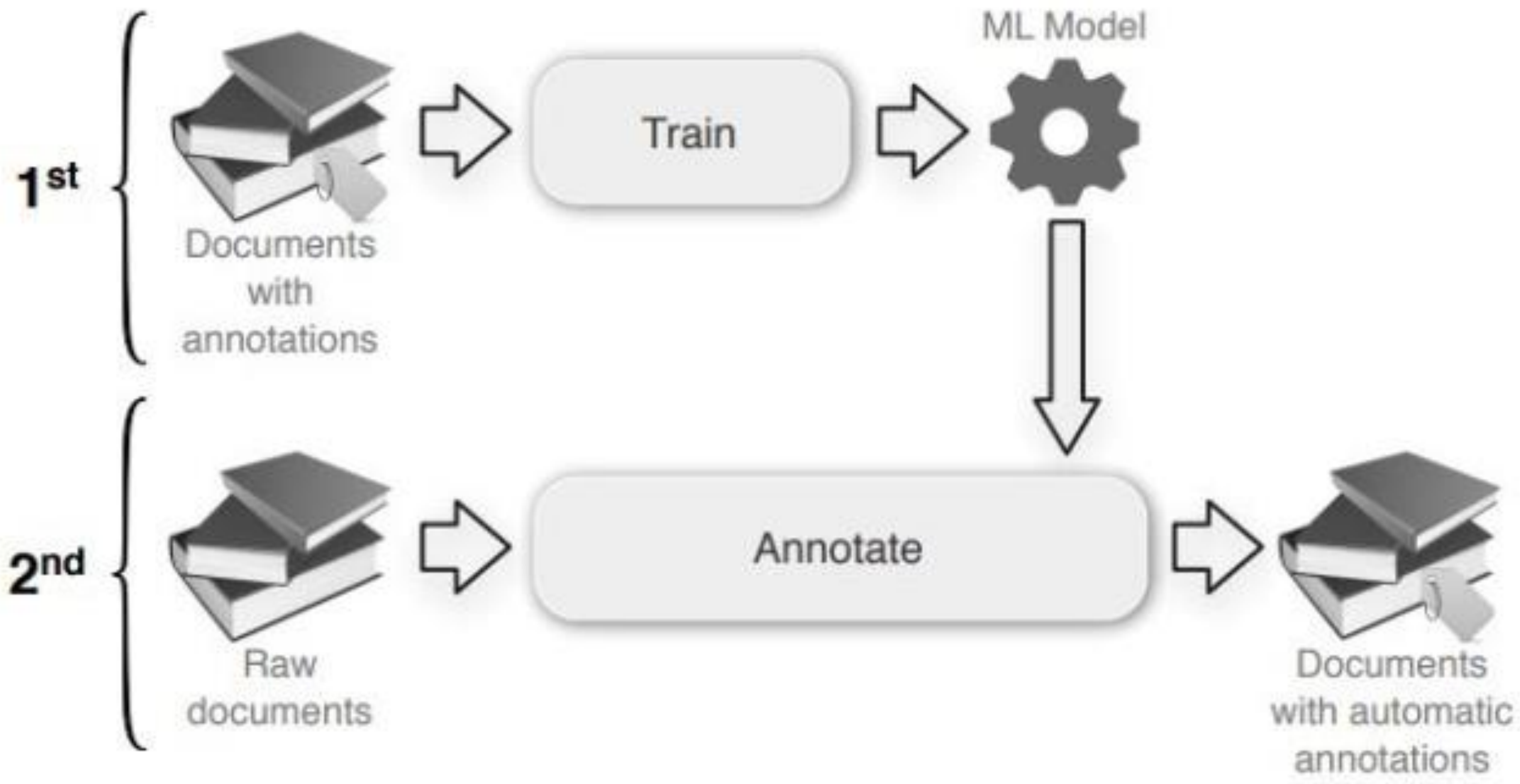
[$_{PERS}$ Washington] was born into slavery on the farm of James Burroughs.
[$_{ORG}$ Washington] went up 2 games to 1 in the four-game series.
Blair arrived in [$_{LOC}$ Washington] for what may well be his last state visit.
In June, [$_{GPE}$ Washington] passed a primary seatbelt law.
The [$_{FAC}$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

Named Entity Visualizer

NER Approaches

- As with partial parsing and chunking there are two basic approaches (and hybrids)
  - Rule-based (regular expressions)
    - Lists of names
    - Patterns to match things that look like names
    - Patterns to match the environments that classes of names tend to occur in.
  - ML-based approaches
    - Get annotated training data
    - Extract features
    - Train systems to replicate the annotation

NER Pipeline



NER Features

- Features may include
  - the word
  - POS tag
  - IOB tag
    - Inside
    - Outside
    - Beginning
  - the shape of the word

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

Evaluation of NER

- The 2-by-2 contingency table

	Correct	Not correct
Selected	True-Positive	False-Positive
Not Selected	False-Negative	True-Negative

Evaluation of NER

- The 2-by-2 contingency table

	Correct	Not correct
Selected	True-Positive	False-Positive
Not Selected	False-Negative	True-Negative

Selected and Not selected (Predicted Values)
Correct and not correct (Actual Values)

Evaluation of NER

	Correct	Not correct
Selected	True-Positive	False-Positive
Not Selected	False-Negative	True-Negative

Precision: % of selected items that are correct = tp / (tp + fp)

- From all the classes we have predicted as positive, how many are actually positive.
- Precision should be high as possible.

Evaluation of NER

	Correct	Not correct
Selected	True-Positive	False-Positive
Not Selected	False-Negative	True-Negative

Recall: % of correct items that are selected = tp / (tp + fn)

- From all the positive classes, how many we predicted correctly.
- Recall should be high as possible.

Precision & Recall

Actual Values

True False

Predicted Values

Positive Negative

<table>
<tr><th>y</th><th>y pred</th><th>output for threshold 0.6</th><th>Recall</th><th>Precision</th><th>Accuracy</th></tr>
<tr><td>0</td><td>0.5</td><td>0</td><td rowspan="7">1/2</td><td rowspan="7">2/3</td><td rowspan="7">4/7</td></tr>
<tr><td>1</td><td>0.9</td><td>1</td></tr>
<tr><td>0</td><td>0.7</td><td>1</td></tr>
<tr><td>1</td><td>0.7</td><td>1</td></tr>
<tr><td>1</td><td>0.3</td><td>0</td></tr>
<tr><td>0</td><td>0.4</td><td>0</td></tr>
<tr><td>1</td><td>0.5</td><td>0</td></tr>
</table>

$$\textbf{\textit{Precision}} = \frac{TP}{TP + FP}$$

$$\textbf{\textit{Recall}} = \frac{TP}{TP + FN}$$

`Accuracy = TP + TN / Total`

Evaluation of NER

- A combined measure: F
  - A combined measure that assesses the P/R tradeoff is F measure (weighted harmonic mean):

$$F = \frac{1}{a\frac{1}{P} + (1 - a)\frac{1}{R}} = \frac{(b^2 + 1)PR}{b^2 P + R}$$

  - People usually use balanced F_1 measure
    - i.e., with $\beta = 1$ (that is, $\alpha = ½$), then

$$F_1 = 2PR/(P+R)$$

$$\textbf{\textit{F - measure}} = \frac{2*\text{Recall}*\text{Precision}}{\text{Recall} + \text{Precision}}$$

Quiz

$F_1 = 2PR/(P+R)$

- $P = 40\%$ $R = 40\%$ $F_1 =$
- $P = 75\%$ $R = 25\%$ $F_1 =$

Quiz

$F_1 = 2PR/(P+R)$

- $P = 40\%$ $R = 40\%$ $F_1 = 40.0\%$
- $P = 75\%$ $R = 25\%$ $F_1 = 37.5\%$