Entity Recognition; Unsupervised Methods; Social Media.

Leon Derczynski

AU 2016
Named Entity Recognition

- Named Entities (NEs) are named objects
- Usually an identifier for a single example of a thing
  - e.g. Novo Nordisk is a single example of a company
- Related to Kripke's idea of “fixed designator”
  - Something that always uniquely identifies
- The types can be subjective!
  - Organisation, location, person
  - Time, TV show, sports team, building
Named Entity Recognition

• NER provides a foundation for complex systems

• Once we have found NEs we can:
  – Find co-referring mentions: “Dr Smith”, “John Smith”, “John”, “he”
  – Find relations: “He (Dr Smith) is CEO of XY”
  – Link to Ontologies/Knowledge Bases: “Athens, Georgia” vs “Athens, Greece” (disambiguation)
NER pipeline

- **Pre-processing**
  - Tokenisation: characters → tokens/words
  - Sentence splitting
  - POS tagging: nouns, verbs, adjectives, ...
  - Morphological analysis: word root/lemma

- **Find entity mentions:** Persons, Locations, ...

- **Type classification** (Person or Location or ..?)

- **Coreferencing:** (different) mentions of the same entity
Example of IE

John lives in London. He works there for Polar Bear Design.
John lives in London. He works there for Polar Bear Design.
Basic NE Recognition

John lives in London. He works there for Polar Bear Design.
John lives in London. He works there for Polar Bear Design.
John lives in London. He works there for Polar Bear Design.
John lives in London. He works there for Polar Bear Design.
John lives in London. He works there for Polar Bear Design.
NER design

• Four part model:
  – What feature representation to use for tokens;
  – Which inference algorithm to use;
  – How to capture non-local dependencies;
  – How to incorporate external knowledge.

• Roth and Ratinov 2009, “Design challenges and misconceptions in named entity recognition”
Representation and labeling

• Token feature representation options:
  - Token itself
  - Previous and following token
  - Word shape, to model capitalisation
  - Part of speech tag
  - Lexical features (e.g. character n-grams) to help with OOV terms

This means, taking sequences of letters to give hints about NE types

E.g. in Danish, if word ends
  - “-ssen”
  - “-strup”
  - Others?
## Representation and labeling

### Labelling scheme:

<table>
<thead>
<tr>
<th>Tag</th>
<th>Entity</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Facebook</td>
<td>B-company</td>
</tr>
<tr>
<td>O</td>
<td>Job-Hunting</td>
<td>O</td>
</tr>
<tr>
<td>O</td>
<td>App</td>
<td>O</td>
</tr>
<tr>
<td>I</td>
<td>BranchOut</td>
<td>B-product</td>
</tr>
<tr>
<td>O</td>
<td>Raises</td>
<td>O</td>
</tr>
<tr>
<td>O</td>
<td>$6</td>
<td>O</td>
</tr>
<tr>
<td>O</td>
<td>Million</td>
<td>O</td>
</tr>
<tr>
<td>O</td>
<td>From</td>
<td>O</td>
</tr>
<tr>
<td>I</td>
<td>Accel</td>
<td>B-company</td>
</tr>
<tr>
<td>O</td>
<td>And</td>
<td>O</td>
</tr>
<tr>
<td>I</td>
<td>Super</td>
<td>B-company</td>
</tr>
<tr>
<td>I</td>
<td>Angels</td>
<td>I-company</td>
</tr>
</tbody>
</table>

- BIO (Begin, In, Out) separates adjacent entities
- CRF with BIO popular
- SVM-U with IO can give better performance
NER inference algorithms

- As with part of speech tagging, sequence labelling can work well (e.g. CRF)
  - Assumes well-formed sentences and lots of training data
  - If this is inappropriate, then local context in token features can compensate
NER inference algorithms

- SVM-U: “uneven” (Li 2009)
- Adjust margins between supporting examples and decision hyperplane to reflect class balance

What's the class balance in NER?
- Is it even between in/out, or does one dominate?
Dependencies & external knowledge

- Typically, only the first mention of an entity is referred to in full:
  - Manchester United are great. They're my favourite football team. Man U forever!

- Using only local features will lead to missed entities.

- How can we incorporate external knowledge for NER?
  - Useful to tell us when unusual/unexpected words are an entity: “Szeged” “White House”
Dependencies & external knowledge

• Unlabelled text
  - NEs found in distributionally similar contexts
  - Deviation from typical context indicates special use
  - Labelled LDA can produce phrase lists given an entity type (Ramage 2009, Ritter 2011)
Phrase list extraction

• Extraction with topic models:
  – Techniques like this work on “filling the gap”
  – First, we find the gaps!
    • Did you see Donald Trump on Fox News last night?
    • Did you see Silvio Berlusconi on Fox News last night?
    • Did you see Donald Trump on TV2 last night?
  – Two slots can be extracted: one person, one TV channel
    • see _ _ on
    • on _ * last night
Phrase list extraction

• How do we find the contexts?
  - Automatically!
  - How likely is a context to indicate a phrase?
  - $P(\text{entity} \mid w_{i-1}) P(\text{entity} \mid w_{i+1}) > P(w_{i-1}) P(w_{i+1})$

• This gives some gaps for finding possible entities
  - Facility: voodoo lounge, grand ballroom, crash mansion, sullivan hall, memorial union, rogers arena, rockwood music hall, amway center, el mocambo, madison square, bridgestone arena, cat club, le poisson rouge, bryant park, mandalay bay, broadway bar, ritz carlton, mgm grand, olympia theatre, consol energy center
  - TV show: pretty little, american skins, nof, order svu, greys, kktny, rhobh, parks & recreation, parks & rec, dawson ’s creek, big fat gypsy weddings, big fat gypsy wedding, winter wipeout, jersey shores, idiot abroad, royle, jerseyshore, mr . sunshine, hawaii five-0, new jersey shore
Gazetteers

- Lists of words/phrases of a certain type
- Can be constructed manually or automatically

- Gaz. completeness gives P/R tradeoff
- Won't catch terms not seem in gazetteer, which makes domain adaptation tough
Gazetteers

• Having identified a match, we need to handle it
  – Feature representation?
  – What about this:
    • “Brøndby were great tonight!”
    • Location or organisation?
    • Metonymy
  – Scope is not always clear:
    • New York Times Square
    • Greedy matches might help!
NER: summary

- **Representation:**
  - in-vs-out?
  - BIO?

- **Inference:**
  - CRF?
  - SVM?

- **Dependencies**
  - Capturing both “Manchester United” and “Man U”

- **World knowledge**
  - Using resources like gazetteers
Unsupervised Learning

● What if we don't have labels?
● We can still discover something from the data

(thanks to Poggio, Ullman, Harari, Zysman @ MIT)
Most learning is unsupervised!

- “Most of human and animal learning is unsupervised learning.
- If intelligence was a cake, unsupervised learning would be the cake, supervised learning would be the icing on the cake, and reinforcement learning would be the cherry on the cake.
- We know how to make the icing and the cherry, but we don't know how to make the cake.”
Clustering

- The organization of unlabeled data into similarity groups called clusters.
  - A cluster is a collection of data items which are “similar” between them
  - And “dissimilar” to data items in other clusters.
Historic application of clustering

- John Snow, a London physician, plotted location of cholera deaths during 1850s outbreak
- Locations clustered near polluted wells
- Exposed problem – and solution!
What do we need to do clustering?

- Proximity measure
  - How close to items are, e.g. $s(x_i, x_k)$
  - Or – how different items are; distance $d(x_i, x_k)$
- Way to evaluate whole clustering
- Algorithm to compute clustering
Distance (dissimilarity) measures

• Euclidean distance
  – E.g. as-the-crow-flies

• Manhattan distance
  – “city block” route
Cluster evaluation (a hard problem)

• Intra-cluster cohesion (compactness):
  – Cohesion measures how near the data points in a cluster are to the cluster centroid.
  – Sum of squared error (SSE) is a commonly used measure.

• This tries to keep clusters relatively close to a centre
Clustering
How many clusters?

• Difficult question to answer!
  – How many parts of speech are in the data?
  – What kinds of named entities are there?
  – Which different writing styles exist?

• Often a good place to compare with theory

• Humans seem to do a pretty good job of this
  – Stereotypes
  – No requirement for strict boundaries
Hierarchy

• Up to now, considered “flat” clustering
• For some data, hierarchical clustering makes sense
Example: biological taxonomy
Types of hierarchical clustering

- **Divisive (top down) clustering**
  - Starts with all data points in one cluster, the root, then:
    - Splits the root into a set of child clusters. Each child cluster is recursively divided further
    - stops when only singleton clusters of individual data points remain, i.e., each cluster with only a single point

- Start from a single cluster, and break it up
Types of hierarchical clustering

• Agglomerative (bottom up) clustering
  – The hierarchy is built from the bottom level by
    • merging the most similar (or nearest) pair of clusters
    • stopping when all the data points are merged into a single cluster (i.e., the root cluster).

• Start with pieces, and group them in order
Divisive hierarchical clustering

- Any “flat” method that produces fixed number of clusters is OK.
- If c=2:
Agglomerative hierarchical clustering

- Initialise with each example in its own cluster
- While there is more than one cluster:
  - Find the two nearest clusters
  - Merge them