What is “Text”? 

Language by itself is complex.

The Web further complicates language.

- Tons of data on the web
- A lot of it is text
- In many languages
- In many genres
NLP for fun and profit

- Making NLP more accessible
  - Provide APIs for common NLP tasks
  ```
  var text = document.get(...);
  var entities = agent.markNE(text);
  ```
- Big $$$
- Backend to intelligent processing of text

Desideratum: Multilinguality

- Except for feature extraction, systems should be language agnostic

In this lecture

- Understand how to solve and ace in NLP tasks
- Learn general methodology or approaches
- End-to-End development using an example task
- Overview of (un)common NLP tasks

Case study: Named Entity Recognition

- Demo: http://viewer.opencalais.com

- How do we build something like this?
- How do we find out well we are doing?
- How can we improve?
Case study: Named Entity Recognition

- Collect data to learn from
  - Sentences with words marked as PER, ORG, LOC, NONE
- How do we get this data?

Pay the experts

Wisdom of the crowds

Getting the data: Annotation

- Time consuming
- Costs $$$
- Need for quality control
  - Inter-annotator agreement
  - Kappa score (Krippendorf, 1980)
- Smarter ways to annotate
  - Get fewer annotations: Active Learning
  - Rationales (Zaidan, Eisner & Piatko, 2007)
Our recipe ...

- 1. Formalize some insights
- 2. Study the formalism mathematically
- 3. Develop & implement algorithms
- 4. Test on real data

NER: Designing features

Need to segment sentences text format

Tokens in sequences

Preprocessing

- Not as trivial as you think
- Original text itself might be in an ugly HTML
- Clean eval!

NER: Designing features

Only  
France  
and  
Great  
Britain  
backed  
Fischler  
' s  
proposal

NER: Designing features

Only  
France  
and  
Great  
Britain  
backed  
Fischler  
' s  
proposal

NER: Designing features

Only  
France  
and  
Great  
Britain  
backed  
Fischler  
' s  
proposal

NER: Designing features

Only  
France  
and  
Great  
Britain  
backed  
Fischler  
' s  
proposal
### NER: Designing features

These are extracted during preprocessing!

<table>
<thead>
<tr>
<th>Only</th>
<th>POS=RB</th>
<th>IS_CAPITALIZED</th>
<th>PREV_WO__NONE_</th>
</tr>
</thead>
<tbody>
<tr>
<td>France</td>
<td>POS=NNP</td>
<td>IS_CAPITALIZED</td>
<td>PREV_WO_only</td>
</tr>
<tr>
<td>and</td>
<td>POS=CC</td>
<td>...</td>
<td>PREV_WO=france</td>
</tr>
<tr>
<td>Great</td>
<td>POS=NNP</td>
<td>IS_CAPITALIZED</td>
<td>PREV_WO=and</td>
</tr>
<tr>
<td>Britain</td>
<td>POS=NNP</td>
<td>IS_CAPITALIZED</td>
<td>PREV_WO=great</td>
</tr>
<tr>
<td>backed</td>
<td>POS=VBD</td>
<td>...</td>
<td>PREV_WO=britain</td>
</tr>
<tr>
<td>Fischer</td>
<td>POS=NNP</td>
<td>IS_CAPITALIZED</td>
<td>PREV_WO=backed</td>
</tr>
</tbody>
</table>

---

### NER: Designing features

- Can you think of other features?

- **WORD**
- **PREV_WO**
- **NEXT_WO**
- **PREV_BIGRAM**
- **NEXT_BIGRAM**
- **POS**
- **PREV_POS**
- **NEXT_POS**
- **PREV_POS_BIGRAM**
- **NEXT_POS_BIGRAM**
- **IN_LEXICON_PER**
- **IN_LEXICON_LOC**
- **IN_LEXICON_ORG**
- **IS_CAPITALIZED**
- **HAS_DIGITS**
- **IS_HYPHENATED**
- **IS_ALLCAPS**
- **RARE_WO**
- **USEFUL_UNIGRAM_PER**
- **USEFUL_BIGRAM_PER**
- **USEFUL_UNIGRAM_LOC**
- **USEFUL_BIGRAM_LOC**
- **USEFUL_UNIGRAM_ORG**
- **USEFUL_BIGRAM_ORG**
- **USEFUL_SUFFIX_PER**
- **USEFUL_SUFFIX_LOC**
- **USEFUL_SUFFIX_ORG**
- **WORD**
- **PREV_WO**
- **NEXT_WO**
- **PREV_BIGRAM**
- **NEXT_BIGRAM**
- **POS**
- **PREV_POS**
- **NEXT_POS**
- **PREV_POS_BIGRAM**
- **NEXT_POS_BIGRAM**
- **IN_LEXICON_PER**
- **IN_LEXICON_LOC**
- **IN_LEXICON_ORG**
- **IS_CAPITALIZED**
- **USEFUL_SUFFIX_ORG**
NER: How else can we improve?

- Unlabeled data!

With more unlabeled data
instance 1: ... headquartered in (Washington State) ... 
instance 2: ... (Mr. Washington)”, the vice president of ... 
instance 3: ... headquartered in (Kazakhstan) ... 
instance 4: ... flew to (Kazakhstan) ... 
test: ... (Robert Jordan), a partner at Steptoe & Johnson ...
test: ... flew to (China) ...

example from Jerry Zhu

NER : Challenges

- Domain transfer
  - WSJ  NYT
  - WSJ  Blogs ??
  - WSJ  Twitter ??!
- Tough nut: Organizations
- Non textual data?

Entity Extraction is a Boring Solved Problem – or is it? (Vilain, Su and Lubar, 2007)

NER: Related application

- Extracting real estate information from Criagslist Ads

Our oversized one, two and three bedroom apartment homes with floor plans featuring 1 and 2 baths offer space unlike any competition. Relax and enjoy the views from your own private balcony or patio, or feel free to entertain, with plenty of space in your large living room, dining area and eat in kitchens. The lovely pool and sun deck make summer fun a splash. Our location makes commuting a breeze – Near MTA bus lines, the Metro station, major shopping areas, and for the little ones, an elementary school is right next door.

NER: Related Application

- BioNLP: Annotation of chemical entities

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>CM</td>
<td>chemical compound</td>
<td>chloric acid</td>
</tr>
<tr>
<td>RN</td>
<td>chemical reaction</td>
<td>1,3-dimethylstilbesterol</td>
</tr>
<tr>
<td>CJ</td>
<td>chemical adjective</td>
<td>quinone</td>
</tr>
<tr>
<td>ASI</td>
<td>enzyme</td>
<td>ornithine</td>
</tr>
<tr>
<td>CPR</td>
<td>chemical prefix</td>
<td>1,3-</td>
</tr>
</tbody>
</table>

Corbet, Batchelor & Teufel, 2007

Shared Tasks: NLP in practice

- Shared Task
  - Everybody works on a (mostly) common dataset
  - Evaluation measures are defined
  - Participants get ranked on the evaluation measures
  - Advance the state of the art
  - Set benchmarks
- Tasks involve common hard problems or new interesting problems