**Announcements**

- Reading Report #8 due now
- Assignment #3
  - Due: next Friday
  - Through Learning Suite

**Objectives**

- Gain further insight into MEMMs
- Understand another application: Named Entity Tagging
- Think about other sequence labeling problems
- Understand the label bias problem in local conditional models
- Time Permitting:
  - Reinforce the distinction between joint (generative) and conditional models
  - Distinguish between discriminative training and conditional models

**Feature Templates**

**Emission:**
- Feature: `<w0=future, t0=JJ>`
- Feature template: `<w0, t0>`

```java
public class CurrentWordExtractor implements FeatureTemplate<LocalContext<String, String>> {
    […]
    public String getName() {
        return "CURWORD";
    }
    […]
    public void extract(LocalContext<String, String> inputDatum, Counter<String> outputVector) {
        outputVector.incrementCount(getName() + ":" + inputDatum.getCurrentWord());
    }
}
```

**Edges or Transitions:**
- Feature: `<t0=DT, t1=JJ>`
- Feature template: `<t0, t1>`
- Higher order: `<t0, t1, t2>`

```java
public class PreviousTagExtractor implements FeatureTemplate<LocalContext<String, String>> {
    […]
    public String getName() {
        return "PREVIOUS_TAG";
    }
    […]
    public void extract(LocalContext<String, String> inputDatum, Counter<String> outputVector) {
        outputVector.incrementCount(getName() + ":" + inputDatum.getPreviousTag());
    }
}
```

**We discussed others:**
- Feature: `<w0 ends in "ly", t0=RB>`
- Feature template: `<2-letter suffix(w0), t0>`

**Also possible:**
- Mixed feature templates: `<t1, w0, t0>`
MEMMs

Local Model:

Mathematically:

\[ P_{ME}(t_1|f(w, l), t_{i-1}, t_{i-2}) \]

Example local context (at training time):

\[ \langle \text{Fed} \rangle \text{ NNP VBZ NN NNS} \]

Example feature vector:

CURWORD:interest, PREVWORD:raises, NEXTWORD:rates, POS:3, 2CHARSUFFIX:st, PREVTAG:VBZ PREVPREVTAG:NNP, ...

Decoding

- Decoding with MaxEnt (or any CMM) taggers:
  - Like decoding in HMMs
  - Same techniques: Viterbi, beam search, posterior decoding

- Viterbi algorithm (HMMs):
  \[ \delta_i(s) = \max_x P(w_i|s)P(s') \cdot \delta_{i-1}(s') \]

- Viterbi algorithm (MEMMs/CMMs):
  \[ \delta_i(s) = \max_x P_{ME}(s'|f(w, l), w, l) \cdot \delta_{i-1}(s') \]

- Beam search is effective! Why?

Tagging Workflow

New Problem

- Named Entity Recognition

Named Entity Recognition

Which features are dynamic?

Feature Weights:

<table>
<thead>
<tr>
<th>Feature Type</th>
<th>Feature</th>
<th>8-PERS</th>
<th>8-LOC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Previous word</td>
<td>at</td>
<td>-0.73</td>
<td>0.94</td>
</tr>
<tr>
<td>Current word</td>
<td>Grace</td>
<td>0.03</td>
<td>0.09</td>
</tr>
<tr>
<td>Beginning bigram</td>
<td>&lt;G</td>
<td>0.45</td>
<td>-0.04</td>
</tr>
<tr>
<td>Current POS tag</td>
<td>NNP</td>
<td>0.47</td>
<td>0.48</td>
</tr>
<tr>
<td>Prev and cur tags</td>
<td>IN NNP</td>
<td>-0.70</td>
<td>0.14</td>
</tr>
<tr>
<td>Previous state</td>
<td>Other</td>
<td>-0.70</td>
<td>-0.92</td>
</tr>
<tr>
<td>Current signature</td>
<td>Xx</td>
<td>0.80</td>
<td>0.48</td>
</tr>
<tr>
<td>Prev state, cur sig</td>
<td>O-Xx</td>
<td>0.88</td>
<td>0.37</td>
</tr>
<tr>
<td>Prev-cur-next sig</td>
<td>O-Xx-Xx</td>
<td>-0.66</td>
<td>0.37</td>
</tr>
<tr>
<td>P. state - p-cur sig</td>
<td>O-Xx</td>
<td>-0.20</td>
<td>0.92</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>-0.58</td>
<td>2.68</td>
</tr>
</tbody>
</table>
Other Applications
- Word breaking
- Sentence breaking
- Phrasal “chunking” / shallow parsing
- Information extraction
- Dialog act mark-up
- ...

Label Bias Problem
- An issue with conditional local models worth knowing about:
  - “Label bias”
  - Maxent taggers’ local scores can be near one without having both good “transitions” and “emissions”
  - This means that often evidence doesn’t flow properly in the MEMM (as a sequence model)

Label Bias Problem
- Search space:
- Think: probabilistic garden path

CRF Taggers
- Fix: normalize globally!
- Do not decompose training into independent local regions
  - Con: Can be deadly slow to train – require repeated inference on training set
  - Pro: Avoid label bias
- Higher-powered discriminatively trained sequence models:
  - Conditional Random Fields (CRFs)
  - Voted perceptrons
  - Maximum Margin Markov Networks (M3Ns)
- Differences tend not to be too important for POS tagging, at least for English WSJ
  - Why?

CRFs are Slow to Train!
- POS: 45 labels, 1,000,000 words = over a week
- NER: 11 labels, 200,000 words = 2 hours
- Using Mallet in the NLP Lab (2008) on new hardware

Reminder: Conditional vs. Joint
- Joint model, conditional query:
  \[ \hat{\ell} = \arg \max_{\ell} \hat{P}(\ell | w) = \arg \max_{\ell} \hat{P}(\ell | w) \]
  - In practice, we restrict the search over possible tags for \( w \): \( \hat{P}(w) \)
  \[ \hat{\ell} = \arg \max_{\ell \in \mathcal{L}} \hat{P}(\ell | w) \]
- Conditional model (with discriminative training):
  \[ \hat{\ell} = \arg \max_{\ell} \hat{P}_{\text{ME}}(\ell | w) \]
Conditional vs. Joint

\[ \hat{P}(L | W) \quad \hat{P}(L, W) \]

Joint vs. Conditional Pairs

We will explore “factor graphs” / undirected graphical models next time.

Next Time

- Undirected Graphical Models
- Markov Random Fields