Announcements

- A2 due today
- A3 out tomorrow
- Midterm: list of topics next week. Covers content up to March 7
  - CRFs will NOT be on the midterm, a couple other topics too
Today

- Conditional random fields
- Named entity recognition
- Syntax and constituency parsing
CRFs and NER
### Named Entity Recognition

<table>
<thead>
<tr>
<th>B-PER</th>
<th>I-PER</th>
<th>O</th>
<th>O</th>
<th>O</th>
<th>B-LOC</th>
<th>O</th>
<th>O</th>
<th>O</th>
<th>B-ORG</th>
<th>O</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barack Obama</td>
<td>will travel to</td>
<td>Hangzhou</td>
<td>today</td>
<td>for the</td>
<td>G20</td>
<td>meeting</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Frame as a sequence problem with a BIO tagset: begin, inside, outside
- Why might an HMM not do so well here?
  - Lots of O’s, so tags aren’t as informative about context
  - Need sub-word features on unknown words
- CRFs are discriminative models that will solve these problems
HMMs are expressible as Bayes nets (factor graphs)

\[ P(y, x) = P(y_1)P(x_1|y_1)P(y_2|y_1)P(x_2|y_2) \ldots \]

Locally normalized model: each factor is a probability distribution that normalizes
Conditional Random Fields

- **HMMs**: \( P(y, x) = P(y_1)P(x_1|y_1)P(y_2|y_1)P(x_2|y_2) \ldots \)

- **CRFs**: discriminative models with the following globally-normalized form:
  \[
P(y|x) = \frac{\prod_k \exp(\phi_k(x, y))}{\sum_{y'} \prod_k \exp(\phi_k(x, y'))}
  \]
  any real-valued scoring function of its arguments
  normalizer \( Z \)

- **Naive Bayes**: logistic regression :: HMMs : CRFs
  local vs. global normalization <-> generative vs. discriminative

- **How do we max over \( y \)?** Requires considering an exponential number of sequences in general
Sequential CRFs

- **HMMs:** \( P(y, x) = P(y_1)P(x_1|y_1)P(y_2|y_1)P(x_2|y_2) \ldots \)

- **CRFs:**

\[
P(y|x) \propto \prod_k \exp(\phi_k(x, y))
\]

\[
P(y|x) \propto \exp(\phi_o(y_1)) \prod_{i=2}^n \exp(\phi_t(y_{i-1}, y_i)) \prod_{i=1}^n \exp(\phi_e(x_i, y_i))
\]
Sequential CRFs

\[
P(y|x) \propto \exp(\phi_o(y_1)) \prod_{i=2}^{n} \exp(\phi_t(y_{i-1}, y_i)) \prod_{i=1}^{n} \exp(\phi_e(x_i, y_i))
\]

- We condition on \(x\), so every factor can depend on all of \(x\)
- \(y\) can’t depend arbitrarily on \(x\) in a generative model

- token index — lets us look at current word
Don’t include initial distribution, can bake into other factors

Sequential CRFs:

\[
P(y|x) = \frac{1}{Z} \prod_{i=2}^{n} \exp(\phi_t(y_{i-1}, y_i)) \prod_{i=1}^{n} \exp(\phi_e(y_i, i, x))
\]
Feature Functions

\[ P(y|x) = \frac{1}{Z} \prod_{i=2}^{n} \exp(\phi_{t}(y_{i-1}, y_{i})) \prod_{i=1}^{n} \exp(\phi_{e}(y_{i}, i, x)) \]

- Phis can be almost anything! Here we use linear functions of sparse features

\[ \phi_{e}(y_{i}, i, x) = w^{\top} f_{e}(y_{i}, i, x) \quad \phi_{t}(y_{i-1}, y_{i}) = w^{\top} f_{t}(y_{i-1}, y_{i}) \]

\[ P(y|x) \propto \exp w^{\top} \left[ \sum_{i=2}^{n} f_{t}(y_{i-1}, y_{i}) + \sum_{i=1}^{n} f_{e}(y_{i}, i, x) \right] \]

- Looks like our single weight vector multiclass logistic regression model
Basic Features for NER

\[ P(y|x) \propto \exp w^\top \left[ \sum_{i=2}^{n} f_t(y_{i-1}, y_i) + \sum_{i=1}^{n} f_e(y_i, i, x) \right] \]

Transitions: \( f_t(y_{i-1}, y_i) = \text{Ind}[y_{i-1} \& y_i] = \text{Ind}[O \rightarrow \text{B-LOC}] \)

Emissions: \( f_e(y_6, 6, x) = \text{Ind}[\text{B-LOC} \& \text{Current word} = \text{Hangzhou}] \)
\( \text{Ind}[\text{B-LOC} \& \text{Prev word} = \text{to}] \)

Barack Obama will travel to \textcolor{orange}{\text{Hangzhou}} today for the G20 meeting.
CRFs Outline

- Model: \[ P(y|x) = \frac{1}{Z} \prod_{i=2}^{n} \exp(\phi_t(y_{i-1}, y_i)) \prod_{i=1}^{n} \exp(\phi_e(y_i, i, x)) \]

\[ P(y|x) \propto \exp w^\top \left[ \sum_{i=2}^{n} f_t(y_{i-1}, y_i) + \sum_{i=1}^{n} f_e(y_i, i, x) \right] \]

- Inference: \ \text{argmax} \ P(y|x) \ \text{from Viterbi}

- Learning: requires running sum-product Viterbi to compute posterior probabilities \( P(y \mid x) \) at each step \( i \)
Features for NER

- Word features (can use in HMM)
  - Capitalization
  - Word shape
  - Prefixes/suffixes
  - Lexical indicators
- Context features (can’t use in HMM!)
  - Words before/after
  - Tags before/after
- Word clusters
- Gazetteers

Leicestershire

Boston

Apple released a new version...

According to the New York Times...
Evaluating NER

- Prediction of all Os still gets 66% accuracy on this example!
- What we really want to know: how many named entity chunk predictions did we get right?
  - Precision: of the ones we predicted, how many are right?
  - Recall: of the gold named entities, how many did we find?
  - F-measure: harmonic mean of these two
NER

- CRF with lexical features can get around 85 F1 on this problem
- Other pieces of information that many systems capture
- World knowledge:

  The delegation met the president at the airport, Tanjug said.

---

Tanjug

From Wikipedia, the free encyclopedia

Tanjug (/ˈtadjʊɡ/) (Serbian Cyrillic: Танжур) is a Serbian state news agency based in Belgrade.[2]
Nonlocal Features

The news agency **Tanjug** reported on the outcome of the meeting.

The delegation met the president at the airport, **Tanjug** said.

- More complex factor graph structures can let you capture this, or just decode sentences in order and use features on previous sentences.

Finkel and Manning (2008), Ratinov and Roth (2009)
# How well do NER systems do?

<table>
<thead>
<tr>
<th>System</th>
<th>Resources Used</th>
<th>$F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ LBJ-NER</td>
<td>Wikipedia, Nonlocal Features, Word-class Model</td>
<td>90.80</td>
</tr>
<tr>
<td>- (Suzuki and Isozaki, 2008)</td>
<td>Semi-supervised on 1G-word unlabeled data</td>
<td>89.92</td>
</tr>
<tr>
<td>- (Ando and Zhang, 2005)</td>
<td>Semi-supervised on 27M-word unlabeled data</td>
<td>89.31</td>
</tr>
<tr>
<td>- (Kazama and Torisawa, 2007a)</td>
<td>Wikipedia</td>
<td>88.02</td>
</tr>
<tr>
<td>- (Krishnan and Manning, 2006)</td>
<td>Non-local Features</td>
<td>87.24</td>
</tr>
<tr>
<td>- (Kazama and Torisawa, 2007b)</td>
<td>Non-local Features</td>
<td>87.17</td>
</tr>
<tr>
<td>+ (Finkel et al., 2005)</td>
<td>Non-local Features</td>
<td>86.86</td>
</tr>
</tbody>
</table>

Lample et al. (2016)

<table>
<thead>
<tr>
<th>System</th>
<th>$F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM-CRF (no char)</td>
<td>90.20</td>
</tr>
<tr>
<td>LSTM-CRF</td>
<td><strong>90.94</strong></td>
</tr>
<tr>
<td>S-LSTM (no char)</td>
<td>87.96</td>
</tr>
<tr>
<td>S-LSTM</td>
<td>90.33</td>
</tr>
</tbody>
</table>

BiLSTM-CRF + ELMo

Peters et al. (2018)

<table>
<thead>
<tr>
<th>System</th>
<th>$F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM-CRF</td>
<td><strong>90.94</strong></td>
</tr>
<tr>
<td>S-LSTM</td>
<td>90.33</td>
</tr>
<tr>
<td>BiLSTM-CRF + ELMo</td>
<td><strong>92.2</strong></td>
</tr>
</tbody>
</table>
Takeaways

- CRFs are structured feature-based models
- Efficient to do inference and learning using dynamic programs
- Looks like logistic regression, but requires more effort to implement
Constituency Parsing
Syntax

- Study of word order and how words form sentences

- Why do we care about syntax?
  - Multiple interpretations of words (noun or verb? Fed raises... example)
  - Recognize verb-argument structures (who is doing what to whom?)
  - Higher level of abstraction beyond words: some languages are SVO, some are VSO, some are SOV, parsing can canonicalize
Constituency Parsing

- Tree-structured syntactic analyses of sentences
- Common things: noun phrases, verb phrases, prepositional phrases
- Bottom layer is POS tags
- Examples will be in English. Constituency makes sense for a lot of languages but not all
She told me that she would never amount to anything.
Constituency Parsing

Examples
Challenges

- PP attachment

If we do not annotate, these trees differ only in one rule:

- $\text{VP} \rightarrow \text{VP} PP$
- $\text{NP} \rightarrow \text{NP} PP$

Parse will go one way or the other, regardless of words.

Lexicalization allows us to be sensitive to specific words.

same parse as “the cake with some icing”
Challenges

- NP internal structure: tags + depth of analysis
Constituency

- How do we know what the constituents are?

- Constituency tests:
  - Substitution by proform (e.g., pronoun)
  - Clefting (*It was with a spoon that...*)
  - Answer ellipsis (What did they eat? *the cake*)
    (How? *with a spoon*)

- Sometimes constituency is not clear, e.g., coordination: *she went to and bought food at the store*
Context-Free Grammars, CKY
1. The pace of the first few lectures (naive Bayes, logistic regression, perceptron, etc.) was [too fast/too slow/just right]

2. The pace of the last few lectures (tagging, Viterbi, parsing) was [too fast/too slow/just right]

3. The homeworks overall are [too hard/too easy/just right]

4. I would prefer A3 be due on [Friday March 8 / Monday March 11] (midterm is on Thursday, March 14)

5. Other comments (likes/dislikes)