Discriminative Training
- Statistical model training involves maximizing some objective function
- For an HMM, we use maximum likelihood training
  - Maximize the probability of the training set
  - Reduction in errors is the true objective of learning
  - Another option is to try to directly optimize error rate or some other closely related objective
  - Consider not just truth, but also other candidates

Perceptron
- One approach that has been around since late 60s is the perceptron
- Basic idea:
  - Find the best scoring analysis (e.g. POS tag sequence)
  - Make its score lower, by penalizing its features
  - Make the score of the truth better, by rewarding its features
  - Go onto the next example

Formal Definition of Perceptron Algorithm
Formally, perceptron approach assumes:
- Training examples \((x_i, y_i)\) for \(i = 1 \ldots N\) where \(x_i\) is the input and \(y_i\) is the true output.
  - e.g. \((w_1, \ldots, w_n, t_1, \ldots, t_k)\) where \(t_1 \ldots t_k\) is the true tag sequence
- A function \(\text{GEN}\) which enumerates a set of candidates \(\text{GEN}(x)\) for an input \(x\).
  - e.g., run the tagger over input word sequence \(x\), to output tag-sequence candidates
- A representation \(\Phi\) mapping each \((x, y)\) \(\in X \times Y\) to a \(d\)-dimensional feature vector \(\Phi(x, y) \in \mathbb{R}^d\).
  - e.g., a vector of weights, one for each feature in \(\Phi\)

Perceptron Algorithm
- **Inputs:** Training examples \((x_i, y_i)\)
- **Initialization:** Set \(\alpha = 0\)
- **Algorithm:**
  - For \(t = 1 \ldots T, \ i = 1 \ldots N\)
    - Calculate \(z_i = \text{argmax}_{z \in \text{GEN}(x_i)} \Phi(x_i, z) \cdot \alpha\)
    - If \((z_i \neq y_i)\) then \(\alpha = \alpha + \Phi(x_i, y_i) - \Phi(x_i, z_i)\)
- **Output:** Parameters \(\alpha\)
Perceptron: Notes

- Because this technique is optimizing (sequence) error rate, it does not involve a normalization factor.
- Thus, it will overtrain,
  - i.e. it will do very well on the training set, but not so well on new data, like unsmoothed maximum likelihood.
  - Techniques exist for controlling overtraining, such as regularization, voting, and averaging.
  - Perceptron models outperform maximum likelihood–optimized models on a range of tasks.
  - POS-tagging, NP-chunking.

Conditional Random Fields (CRFs)

- The perceptron algorithm only pays attention to best-scoring (argmax) path.
- What if there were two top analyses, very close in score?
  - Should penalize features on both.
  - How do we allocate the penalty?
- CRFs are a way to do this, by optimizing the conditional log-likelihood of the truth.

CRF objective function

- Choose $\alpha$ to maximize the conditional log-likelihood of the training data:
  $$\text{LL}(\alpha) = \sum_{t=1}^{N} \log p_\theta(y_t|x_t) = \sum_{t=1}^{N} [\Phi(x_t, y_t) \cdot \alpha - \log Z(x_t, \alpha)]$$
- Use a zero-mean Gaussian prior on the parameters resulting in the regularized objective function:
  $$\text{LL}_R(\alpha) = \sum_{t=1}^{N} [\Phi(x_t, y_t) \cdot \alpha - \log Z(x_t, \alpha)] - \frac{||\alpha||^2}{2\sigma^2}$$
- Where the value $\sigma$ is typically estimated on heldout data.

Formal Definition of CRFs

- Define a conditional distribution over the members of $\text{GEN}(x)$ for a given input $x$:
  $$p_\theta(y|x) = \frac{1}{Z(x, \alpha)} \exp (\Phi(x, y) \cdot \alpha)$$
- Where
  $$Z(x, \alpha) = \sum_{y \in \text{GEN}(x)} \exp (\Phi(x, y) \cdot \alpha)$$
- (Can be calculated with forward-backward algorithm!)

CRF Optimization

- The objective function is convex and there is a globally optimal solution.
- Can use general numerical optimization techniques to find the global optimum.
  - e.g. for a language modeling project we used a general limited memory variable metric method to optimize $LL_R$ from a publicly available software library.
  - The optimizer needs the function value and the derivative (or gradient).

Agenda

- Homework
- Supervised Learning – Discriminative Training
  - Perceptron
  - CRFs
  - Features
- Midterm Review
Derivative of $LL_R$: Refresher

Remember the chain rule:
\[
\frac{df(g(x))}{dx} = \frac{df}{dg} \cdot \frac{dg}{dx}
\]
Also remember derivative of (natural) log:
\[
\frac{d \log(x)}{dx} = \frac{1}{x}
\]
And don’t forget the derivative of exp:
\[
\frac{d \exp(ax)}{dx} = a \exp(ax)
\]

Perceptron vs CRFs

- **Training time**
  - More expensive (calculating derivative) for CRFs...
  - ...but can be parallelized

- **Performance**
  - In Sha & Pereira, perceptron performance not statistically significantly different from CRF with same feature set

Features ($\Phi$)

- Good feature sets matter a lot
- These discriminative methods allow for easy use of many features
  - Unlike HMM based methods
- Examples of feature sets

Agenda

- **Homework**
- **Supervised Learning – Discriminative Training**
  - Perceptron
  - CRFs
  - Features
- **Midterm Review**

Features for Shallow Parsing

<table>
<thead>
<tr>
<th>$c_{i-1}(y_i)$</th>
<th>$c(y_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p = y_i$</td>
<td>$p + 1$</td>
</tr>
<tr>
<td>$p - 1$</td>
<td>$p - 1$</td>
</tr>
</tbody>
</table>

$y_i$ is the class of $w_i$
$t_i$ is the POS-tag of $w_i$
$\gamma_i = c_{i-1}, c_i$
e.g. Ill or Ko, but never Ol
$c(y_i) = c_i$

Sha & Pereira, 2003
Features for Tagging, & OOVs

Ratnaparkhi, 1993

<table>
<thead>
<tr>
<th>Conditions</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>$w_i$ is not rare</td>
<td>$w_i = X$ &amp; $l_i = 1$</td>
</tr>
<tr>
<td>$w_i$ in case</td>
<td>$X$ is prefix of $w_i$, $</td>
</tr>
<tr>
<td></td>
<td>$X$ is suffix of $w_i$, $</td>
</tr>
<tr>
<td></td>
<td>$w_i$ contains number &amp; $l_i = 1$</td>
</tr>
<tr>
<td></td>
<td>$w_i$ contains uppercase character &amp; $l_i = 1$</td>
</tr>
<tr>
<td></td>
<td>$w_i$ contains hyphen &amp; $l_i = 1$</td>
</tr>
<tr>
<td>$\forall w_i$</td>
<td>$l_{i-1} = X$ &amp; $l_i = 1$</td>
</tr>
<tr>
<td></td>
<td>$l_{i-1} = XY$ &amp; $l_i = 1$</td>
</tr>
<tr>
<td></td>
<td>$l_{i-1} = X$ &amp; $l_i = 1$</td>
</tr>
<tr>
<td></td>
<td>$l_{i-1} = X$ &amp; $l_i = 1$</td>
</tr>
<tr>
<td></td>
<td>$l_{i-1} = X$ &amp; $l_i = 1$</td>
</tr>
</tbody>
</table>

Table 1: Features on the current history $h_i$

Instantiated Features

<table>
<thead>
<tr>
<th>Word:</th>
<th>the stories about well-behaved communities and developers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tag:</td>
<td>DT NNS IN JJ NN CC NNS</td>
</tr>
<tr>
<td>Position:</td>
<td>1 2 3 4 5 6 7</td>
</tr>
</tbody>
</table>

| $w_{i-1}$ = short | $l_{i-1} = 10$ |
| $w_{i-2}$ = stories | $l_{i-2} = 10$ |
| $w_{i-3}$ = the | $l_{i-3} = 10$ |
| $w_{i-4}$ = well-behaved | $l_{i-4} = 10$ |
| $w_{i-5}$ = communities | $l_{i-5} = 10$ |
| $w_{i-6}$ = IN | $l_{i-6} = 10$ |
| $w_{i-7}$ = NNS IN | $l_{i-7} = 10$ |
| prob($w_i$|wo) | $l_{i-8} = 10$ |
| prob($w_i$|wo) | $l_{i-9} = 10$ |
| prob($w_i$|wo) | $l_{i-10} = 10$ |
| $w_i$ contains hyphen | $l_i = 12$ |

Table 2: Sample Data

<table>
<thead>
<tr>
<th>Word:</th>
<th>the stories about well-behaved communities and developers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tag:</td>
<td>DT NNS IN JJ NN CC NNS</td>
</tr>
<tr>
<td>Position:</td>
<td>1 2 3 4 5 6 7</td>
</tr>
</tbody>
</table>

| $w_{i-1}$ = short | $l_{i-1} = 10$ |
| $w_{i-2}$ = stories | $l_{i-2} = 10$ |
| $w_{i-3}$ = the | $l_{i-3} = 10$ |
| $w_{i-4}$ = well-behaved | $l_{i-4} = 10$ |
| $w_{i-5}$ = communities | $l_{i-5} = 10$ |
| $w_{i-6}$ = IN | $l_{i-6} = 10$ |
| $w_{i-7}$ = NNS IN | $l_{i-7} = 10$ |
| prob($w_i$|wo) | $l_{i-8} = 10$ |
| prob($w_i$|wo) | $l_{i-9} = 10$ |
| prob($w_i$|wo) | $l_{i-10} = 10$ |
| $w_i$ contains hyphen | $l_i = 12$ |

Table 3: Features Generated From $h_i$ (for tagging short) from Table 2

Agenda

- Homework
- Supervised Learning – Discriminative Training
  - Perceptron
  - CRFs
  - Features
- Midterm Review

Midterm Topics

- Sequences and n-grams
- FSAs, FSTs
  - Construction
  - Composition
- Smoothing
  - Algorithms
  - Interpolation, Backoff
- HMMs
  - Tagging
  - Viterbi
  - Forward-Backward

Midterm Format

- Some short answer questions
- Some basic numerical computation
- Questions from the homeworks
- No programming

Ground rules:
- Work completely independently – no communication of any kind
- No communication with the TA or instructor
- Open book, open note. Not open internet, except for web pages explicitly linked from the class webpage.
- Turn in a hard copy on Tuesday October 25 (or earlier, to Kristy)
Agenda: Summary

- Supervised Learning – Discriminative Training
  - Perceptron
  - CRFs
  - Features
- Midterm Review