High-Level View of Viterbi

- The decision about $Y_\ell$ is a function of $y_{\ell-1}$, $x$, and nothing else!
High-Level View of Viterbi

- The decision about $Y_\ell$ is a function of $y_{\ell-1}$, $x$, and nothing else!
- If, for each value of $y_{\ell-1}$, we knew the best $y_{1:(\ell-1)}$, then picking $y_\ell$ (and $y_{\ell-1}$) would be easy.
High-Level View of Viterbi

- The decision about $Y_\ell$ is a function of $y_{\ell-1}$, $x$, and nothing else!
- If, for each value of $y_{\ell-1}$, we knew the best $y_{1:(\ell-1)}$, then picking $y_\ell$ (and $y_{\ell-1}$) would be easy.
- Idea: for each position $i$, calculate the score of the best label prefix $y_{1:i}$ ending in each possible value for $Y_i$. 

High-Level View of Viterbi

- The decision about $Y_\ell$ is a function of $y_{\ell-1}$, $x$, and nothing else!
- If, for each value of $y_{\ell-1}$, we knew the best $y_{1:(&-1)}$, then picking $y_\ell$ (and $y_{\ell-1}$) would be easy.
- Idea: for each position $i$, calculate the score of the best label prefix $y_{1:i}$ ending in each possible value for $Y_i$.
- With a little bookkeeping, we can then trace backwards and recover the best label sequence.
First, think about the score of the best sequence.

Let $s_i(y)$ be the score of the best label sequence for $x_{1:i}$ that ends in $y$. It is defined recursively:

$$s_\ell(y) = \gamma_{\bigcirc \mid y} \cdot \theta_{x_{\ell} \mid y} \cdot \max_{y' \in \mathcal{L}} \gamma_{y \mid y'} \cdot s_{\ell-1}(y')$$
Recurrence

First, think about the score of the best sequence.

Let \( s_i(y) \) be the score of the best label sequence for \( x_{1:i} \) that ends in \( y \). It is defined recursively:

\[
\begin{align*}
    s_\ell(y) &= \gamma_{\bigcirc |y} \cdot \theta_{x_\ell|y} \cdot \max_{y' \in \mathcal{L}} \gamma_{y|y'} \cdot s_{\ell-1}(y') \\
    s_{\ell-1}(y) &= \theta_{x_\ell|y} \cdot \max_{y' \in \mathcal{L}} \gamma_{y|y'} \cdot s_{\ell-2}(y')
\end{align*}
\]
Recurrence

First, think about the score of the best sequence.

Let $s_i(y)$ be the score of the best label sequence for $x_{1:i}$ that ends in $y$. It is defined recursively:

$$s_\ell(y) = \gamma_{|y} \cdot \theta_{x_\ell|y} \cdot \max_{y' \in L} \gamma_{y|y'} \cdot s_{\ell-1}(y')$$

$$s_{\ell-1}(y) = \theta_{x_\ell|y} \cdot \max_{y' \in L} \gamma_{y|y'} \cdot s_{\ell-2}(y')$$

$$s_{\ell-2}(y) = \theta_{x_\ell|y} \cdot \max_{y' \in L} \gamma_{y|y'} \cdot s_{\ell-3}(y')$$
Recurrence

First, think about the score of the best sequence.

Let $s_i(y)$ be the score of the best label sequence for $x_{1:i}$ that ends in $y$. It is defined recursively:

\[
\begin{align*}
  s_{\ell}(y) &= \gamma \bigcirc | y \bigcdot \theta_{x_{\ell}|y} \cdot \max_{y' \in \mathcal{L}} \gamma_{y|y'} \cdot s_{\ell-1}(y') \\
  s_{\ell-1}(y) &= \theta_{x_{\ell}|y} \cdot \max_{y' \in \mathcal{L}} \gamma_{y|y'} \cdot s_{\ell-2}(y') \\
  s_{\ell-2}(y) &= \theta_{x_{\ell}|y} \cdot \max_{y' \in \mathcal{L}} \gamma_{y|y'} \cdot s_{\ell-3}(y') \\
  &\vdots \\
  s_i(y) &= \theta_{x_i|y} \cdot \max_{y' \in \mathcal{L}} \gamma_{y|y'} \cdot s_{i-1}(y')
\end{align*}
\]
Recurrence

First, think about the score of the best sequence. Let \( s_i(y) \) be the score of the best label sequence for \( x_{1:i} \) that ends in \( y \). It is defined recursively:

\[
\begin{align*}
s_\ell(y) &= \gamma_{\ell \mid y} \cdot \theta_{x_\ell \mid y} \cdot \max_{y' \in L} \gamma_{y \mid y'} \cdot s_{\ell - 1}(y') \\
s_{\ell - 1}(y) &= \theta_{x_{\ell - 1} \mid y} \cdot \max_{y' \in L} \gamma_{y \mid y'} \cdot s_{\ell - 2}(y') \\
s_{\ell - 2}(y) &= \theta_{x_{\ell - 2} \mid y} \cdot \max_{y' \in L} \gamma_{y \mid y'} \cdot s_{\ell - 3}(y') \\
&\vdots \\
s_i(y) &= \theta_{x_i \mid y} \cdot \max_{y' \in L} \gamma_{y \mid y'} \cdot s_{i - 1}(y') \\
&\vdots \\
s_1(y) &= \theta_{x_1 \mid y} \cdot \max_{y' \in L} \gamma_{y \mid y'} \cdot \pi_{y'}
\end{align*}
\]
Viterbi Procedure (Part I: Prefix Scores)

<table>
<thead>
<tr>
<th></th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>...</th>
<th>$x_\ell$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$y'$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$y^{\text{last}}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Viterbi Procedure (Part I: Prefix Scores)

\[
\begin{array}{c|cccc}
  & x_1 & x_2 & \ldots & x_\ell \\
\hline
y & s_1(y) & & & \\
\hline
y' & s_1(y') & & & \\
\hline
\vdots & & & & \\
\hline
y_{last} & s_1(y_{last}) & & & \\
\end{array}
\]

\[
s_1(y) = \theta_{x_1|y} \cdot \max_{y'|y' \in \mathcal{L}} \gamma_{y'|y'} \cdot \pi_{y'}
\]
Viterbi Procedure (Part I: Prefix Scores)

\[
\begin{array}{cccc}
 x_1 & x_2 & \ldots & x_\ell \\
 y & s_1(y) & s_2(y) & \ldots & s_\ell(y) \\
y' & s_1(y') & s_2(y') & \ldots & s_\ell(y') \\
\vdots & & & & \\
y^{last} & s_1(y^{last}) & s_2(y^{last}) & \ldots & s_\ell(y^{last}) \\
\end{array}
\]

\[
s_i(y) = \theta_{x_i|y} \cdot \max_{y' \in \mathcal{L}} \gamma_{y|y'} \cdot s_{i-1}(y')
\]
Viterbi Procedure (Part I: Prefix Scores)

<table>
<thead>
<tr>
<th></th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$\ldots$</th>
<th>$x_\ell$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>$s_1(y)$</td>
<td>$s_2(y)$</td>
<td>$s_\ell(y)$</td>
<td></td>
</tr>
<tr>
<td>$y'$</td>
<td>$s_1(y')$</td>
<td>$s_2(y')$</td>
<td>$s_\ell(y')$</td>
<td></td>
</tr>
<tr>
<td>$\vdots$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$y_{last}$</td>
<td>$s_1(y_{last})$</td>
<td>$s_2(y_{last})$</td>
<td>$s_\ell(y_{last})$</td>
<td></td>
</tr>
</tbody>
</table>

$$s_\ell(y) = \gamma_{\bigcirc |y} \cdot \theta_{x_\ell | y} \cdot \max_{y' \in \mathcal{L}} \gamma_{y | y'} \cdot s_{\ell - 1}(y')$$
Claim: $\max_{y \in \mathcal{L}} s_\ell(y) = \max_{y \in \mathcal{L}^{\ell+1}} p(x, y)$
Claim: \( \max_{y \in \mathcal{L}} s_{\ell}(y) = \max_{y \in \mathcal{L}^{\ell+1}} p(\mathbf{x}, y) \)
Claim: \( \max_{y \in \mathcal{L}} s_{\ell}(y) = \max_{y \in \mathcal{L}^{\ell+1}} p(x, y) \)

\[
\max_{y \in \mathcal{L}} s_{\ell}(y) = \max_{y \in \mathcal{L}} \gamma_{x|y} \cdot \theta_{x_{\ell}|y} \cdot \max_{y' \in \mathcal{L}} \gamma_{y|y'} \cdot s_{\ell-1}(y')
\]

\[
= \max_{y \in \mathcal{L}} \gamma_{x|y} \cdot \theta_{x_{\ell-1}|y'} \cdot \max_{y'' \in \mathcal{L}} \gamma_{y'|y''} \cdot \max_{y' \in \mathcal{L}} \gamma_{y|y'} \cdot s_{\ell-2}(y'')
\]
Claim: \( \max_{y \in \mathcal{L}} s_{\ell}(y) = \max_{y \in \mathcal{L}^{\ell+1}} p(\bm{x}, y) \)
Claim: $\max_{y \in \mathcal{L}} s_\ell(y) = \max_{y \in \mathcal{L}^{\ell+1}} p(x, y)$

\[
\max_{y \in \mathcal{L}} s_\ell(y) = \max_{y \in \mathcal{L}} \gamma_{y_0} \cdot \theta_{y_1} \cdot \max_{y' \in \mathcal{L}} \gamma_{y'y'} \cdot s_{\ell-1}(y')
\]

\[
= \max_{y \in \mathcal{L}} \gamma_{y_0} \cdot \theta_{y_1} \cdot \max_{y' \in \mathcal{L}} \gamma_{y'y'} \cdot \theta_{y_1} \cdot \max_{y'' \in \mathcal{L}} \gamma_{y'y''} \cdot s_{\ell-2}(y'')
\]

\[
= \max_{y \in \mathcal{L}} \gamma_{y_0} \cdot \theta_{y_1} \cdot \max_{y' \in \mathcal{L}} \gamma_{y'y'} \cdot \theta_{y_1} \cdot \max_{y'' \in \mathcal{L}} \gamma_{y'y''} \cdot s_{\ell-3}(y''')
\]

\[
= \max_{y \in \mathcal{L}^{\ell+1}} \gamma_{y_\ell} \cdot \theta_{y_\ell} \cdot \gamma_{y_{\ell-1}} \cdot \theta_{y_1} \cdot \max_{y_{\ell-1}} \gamma_{y_{\ell-1}} \cdot \gamma_{y_{\ell-2}} \cdot \theta_{y_{\ell-2}} \cdot \cdots \cdot \theta_{y_1} \cdot \gamma_{y_0} \cdot \pi_{y_0}
\]
Claim: \( \max_{y \in L} s_\ell(y) = \max_{y \in L^{\ell+1}} p(x, y) \)

\[
\max_{y \in L} s_\ell(y) = \max_{y \in L} \gamma_0 |y| \cdot \theta_{x_\ell | y} \cdot \max_{y' \in L} \gamma_{y|y'} \cdot \max_{y'' \in L} s_{\ell-1}(y')
\]

\[
= \max_{y \in L} \gamma_0 |y| \cdot \theta_{x_\ell | y} \cdot \max_{y' \in L} \gamma_{y|y'} \cdot \theta_{x_{\ell-1} | y'} \cdot \max_{y'' \in L} \gamma_{y'|y''} \cdot \max_{y''' \in L} s_{\ell-2}(y'')
\]

\[
= \max_{y \in L} \gamma_0 |y| \cdot \theta_{x_\ell | y} \cdot \max_{y' \in L} \gamma_{y|y'} \cdot \theta_{x_{\ell-1} | y'} \cdot \max_{y'' \in L} \gamma_{y'|y''} \cdot \max_{y''' \in L} s_{\ell-3}(y''')
\]

\[
= \max_{y \in L^{\ell+1}} \gamma_0 |y_\ell| \cdot \theta_{x_\ell | y_\ell} \cdot \gamma_{y_\ell | y_{\ell-1}} \cdot \theta_{x_{\ell-1} | y_{\ell-1}} \cdot \gamma_{y_{\ell-1} | y_{\ell-2}} \cdot \theta_{x_{\ell-2} | y_{\ell-2}} \cdots \theta_{x_1 | y_1} \cdot \gamma_{y_1 | y_0} \cdot \pi_{y_0}
\]

\[
= \max_{y \in L^{\ell+1}} \pi_{y_0} \prod_{i=1}^{\ell+1} \theta_{x_i | y_i} \cdot \gamma_{y_i | y_{i-1}}
\]
High-Level View of Viterbi

- The decision about $Y_\ell$ is a function of $y_{\ell-1}$, $x$, and nothing else!
- If, for each value of $y_{\ell-1}$, we knew the best $y_{1:(\ell-1)}$, then picking $y_\ell$ (and $y_{\ell-1}$) would be easy.
- Idea: for each position $i$, calculate the score of the best label prefix $y_{1:i}$ ending in each possible value for $Y_i$.
- With a little bookkeeping, we can then trace backwards and recover the best label sequence.
Viterbi Procedure (Part I: Prefix Scores and Backpointers)

<table>
<thead>
<tr>
<th></th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>\ldots</th>
<th>$x_\ell$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$y'$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\vdots$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$y_{\text{last}}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Viterbi Procedure (Part I: Prefix Scores and Backpointers)

<table>
<thead>
<tr>
<th></th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$\ldots$</th>
<th>$x_\ell$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>$s_1(y)$</td>
<td>$b_1(y)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$y'$</td>
<td>$s_1(y')$</td>
<td>$b_1(y')$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\vdots$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$y_{last}$</td>
<td>$s_1(y_{last})$</td>
<td>$b_1(y_{last})$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$s_1(y) = \theta_{x_1|y} \cdot \max_{y'|y' \in \mathcal{L}} \gamma_{y|y'} \cdot \pi_{y'}$

$b_1(y) = \arg\max_{y'|y' \in \mathcal{L}} \gamma_{y|y'} \cdot \pi_{y'}$
Viterbi Procedure (Part I: Prefix Scores and Backpointers)

<table>
<thead>
<tr>
<th></th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>...</th>
<th>$x_\ell$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>$s_1(y)$</td>
<td>$s_2(y)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$b_1(y)$</td>
<td>$b_2(y)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$y'$</td>
<td>$s_1(y')$</td>
<td>$s_2(y')$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$b_1(y')$</td>
<td>$b_2(y')$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>$y^{last}$</td>
<td>$s_1(y^{last})$</td>
<td>$s_2(y^{last})$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$b_1(y^{last})$</td>
<td>$b_2(y^{last})$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$$s_i(y) = \theta_{x_i|y} \cdot \max_{y'|y' \in \mathcal{L}} \gamma_{y|y'} \cdot s_{i-1}(y')$$

$$b_i(y) = \arg\max_{y' \in \mathcal{L}} \gamma_{y|y'} \cdot s_{i-1}(y')$$
Viterbi Procedure (Part I: Prefix Scores and Backpointers)

<table>
<thead>
<tr>
<th></th>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>( \ldots )</th>
<th>( x_\ell )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y )</td>
<td>( s_1(y) )</td>
<td>( s_2(y) )</td>
<td>( \ldots )</td>
<td>( s_\ell(y) )</td>
</tr>
<tr>
<td></td>
<td>( b_1(y) )</td>
<td>( b_2(y) )</td>
<td></td>
<td>( b_\ell(y) )</td>
</tr>
<tr>
<td>( y' )</td>
<td>( s_1(y') )</td>
<td>( s_2(y') )</td>
<td>( \ldots )</td>
<td>( s_\ell(y') )</td>
</tr>
<tr>
<td></td>
<td>( b_1(y') )</td>
<td>( b_2(y') )</td>
<td></td>
<td>( b_\ell(y') )</td>
</tr>
<tr>
<td>( \vdots )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( y_{\text{last}} )</td>
<td>( s_1(y_{\text{last}}) )</td>
<td>( s_2(y_{\text{last}}) )</td>
<td>( \ldots )</td>
<td>( s_\ell(y_{\text{last}}) )</td>
</tr>
<tr>
<td></td>
<td>( b_1(y_{\text{last}}) )</td>
<td>( b_2(y_{\text{last}}) )</td>
<td></td>
<td>( b_\ell(y_{\text{last}}) )</td>
</tr>
</tbody>
</table>

\[
s_\ell(y) = \gamma \big| y \big) \cdot \theta_{x_\ell | y} \cdot \max_{y' \in \mathcal{L}} \gamma_{y | y'} \cdot \mathbb{1}_{s_{\ell-1}(y')}
\]

\[
b_\ell(y) = \arg \max_{y' \in \mathcal{L}} \gamma_{y | y'} \cdot s_{\ell-1}(y')
\]
Full Viterbi Procedure

Input: $x, \theta, \gamma, \pi$

Output: $\hat{y}$

1. For $i \in \langle 1, \ldots, \ell \rangle$:
   ▶ Solve for $s_i(*)$ and $b_i(*)$.
     ▶ Special base case for $i = 1$ to handle $\pi$
     ▶ General recurrence for $i \in \langle 2, \ldots, \ell - 1 \rangle$
     ▶ Special case for $i = \ell$ to handle stopping probability

2. $\hat{y}_\ell \leftarrow \arg\max_{y \in \mathcal{L}} s_\ell(y)$

3. For $i \in \langle \ell, \ldots, 1 \rangle$:
   ▶ $\hat{y}_{i-1} \leftarrow b(y_i)$
Full Viterbi Procedure

Input: $x, \theta, \gamma, \pi$
Output: $\hat{y}$

1. For $i \in \langle 1, \ldots, \ell \rangle$:
   ▶ Solve for $s_i(*)$ and $b_i(*)$.
   ▶ Special base case for $i = 1$ to handle $\pi$ (base case)
   ▶ General recurrence for $i \in \langle 2, \ldots, \ell - 1 \rangle$
     \[
     s_i(y) = \theta_{x_i|y} \cdot \max_{y' \in \mathcal{L}} \gamma_{y|y'} \cdot s_{i-1}(y')
     \]
     \[
     b_i(y) = \arg\max_{y' \in \mathcal{L}} \gamma_{y|y'} \cdot s_{i-1}(y')
     \]
   ▶ Special case for $i = \ell$ to handle stopping probability

2. $\hat{y}_{\ell} \leftarrow \arg\max_{y \in \mathcal{L}} s_{\ell}(y)$

3. For $i \in \langle \ell, \ldots, 1 \rangle$:
   ▶ $\hat{y}_{i-1} \leftarrow b(y_i)$
Viterbi Asymptotics

Space: $O(|\mathcal{L}|\ell)$

Runtime: $O(|\mathcal{L}|^2\ell)$
Generalizing Viterbi

- Instead of HMM parameters, we can use the featurized variant.

\[ s_i(y) = \max_{y' \in \mathcal{L}} \exp (w \cdot \phi(x, i, y, y')) \cdot s_{i-1}(y') \]

More features may increase runtime, but asymptotic dependence on \( \ell \) and \( |\mathcal{L}| \) is the same.

- For this case and for the HMM case, taking logarithms is a good idea.
- Note that dependence on entirety of \( x \) doesn’t affect asymptotics.
Generalizing Viterbi

- Instead of HMM parameters, we can use the featurized variant.
- Viterbi instantiates an general algorithm called **max-product variable elimination** for inference along a chain of variables with pairwise links.
  - Applicable to Bayesian networks and Markov networks.
Generalizing Viterbi

- Instead of HMM parameters, we can use the featurized variant.
- Viterbi instantiates an general algorithm called max-product variable elimination for inference along a chain of variables with pairwise links.
- Viterbi solves a special case of the “best path” problem.
Generalizing Viterbi

- Instead of HMM parameters, we can use the featurized variant.
- Viterbi instantiates an general algorithm called **max-product variable elimination** for inference along a chain of variables with pairwise links.
- Viterbi solves a special case of the “best path” problem.
- Higher-order dependencies among $Y$ are also possible.

$$s_i(y, y') = \max_{y'' \in \mathcal{L}} \exp \left( w \cdot \phi(x, i, y, y', y'') \right) \cdot s_{i-1}(y', y'')$$
Generalizing Viterbi

- Instead of HMM parameters, we can use the featurized variant.
- Viterbi instantiates an general algorithm called max-product variable elimination for inference along a chain of variables with pairwise links.
- Viterbi solves a special case of the “best path” problem.
- Higher-order dependencies among $Y$ are also possible.
- Dynamic programming algorithms.
Generalizing Viterbi

- Instead of HMM parameters, we can use the featurized variant.
- Viterbi instantiates an general algorithm called **max-product variable elimination** for inference along a chain of variables with pairwise links.
- Viterbi solves a special case of the “best path” problem.
- Higher-order dependencies among $Y$ are also possible.
- Dynamic programming algorithms.
- Weighted finite-state analysis.
Applications of Sequence Models

- part-of-speech tagging (Church, 1988)
- supersense tagging (Ciaramita and Altun, 2006)
- named-entity recognition (Bikel et al., 1999)
- multiword expressions (Schneider and Smith, 2015)
- base noun phrase chunking (Sha and Pereira, 2003)

Along the way, we’ll briefly mention two ways to learn sequence models.
Parts of Speech

- “Open classes”: Nouns, verbs, adjectives, adverbs, numbers
- “Closed classes”:
  - Modal verbs
  - Prepositions (on, to)
  - Particles (off, up)
  - Determiners (the, some)
  - Pronouns (she, they)
  - Conjunctions (and, or)
Parts of Speech in English: Decisions

Granularity decisions regarding:
- verb tenses, participles
- plural/singular for verbs, nouns
- proper nouns
- comparative, superlative adjectives and adverbs

Some linguistic reasoning required:
- Existential *there*
- Infinitive marker *to*
- *wh* words (pronouns, adverbs, determiners, possessive *whose*)

Interactions with tokenization:
- Punctuation
- Compounds (*Mark’ll, someone’s, gonna*)

Penn Treebank: 45 tags, ~40 pages of guidelines (Marcus et al., 1993)
Parts of Speech in English: Decisions

Granularity decisions regarding:

- verb tenses, participles
- plural/singular for verbs, nouns
- proper nouns
- comparative, superlative adjectives and adverbs

Some linguistic reasoning required:

- Existential *there*
- Infinitive marker *to*
- *wh* words (pronouns, adverbs, determiners, possessive *whose*)

Interactions with tokenization:

- Punctuation
- Compounds (*Mark’ll, someone’s, gonna*)
- Social media: hashtag, at-mention, discourse marker (*RT*), URL, emoticon, abbreviations, interjections, acronyms

Penn Treebank: 45 tags, ~40 pages of guidelines (Marcus et al., 1993)
TweetNLP: 20 tags, 7 pages of guidelines (Gimpel et al., 2011)
Example: Part-of-Speech Tagging

ikr  smh  he  asked  fir  yo  last  name

so  he  can  add  u  on  fb  lololol
Example: Part-of-Speech Tagging

I know, right shake my head for your

ikr smh he asked fir yo last name

so he can add u on fb lololol
Example: Part-of-Speech Tagging

I know, right shake my head for your

ikr smh he asked for yo last name!

interjection acronym pronoun verb prep. det. adj. noun

so he can add u on fb lololol

you Facebook laugh out loud

P O V V O P ∧ !

preposition proper noun
Why POS?

- Text-to-speech: *record, lead, protest*
- Lemmatization: *saw/V → see; saw/N → saw*
- Preprocessing for harder disambiguation problems:
  - *The Georgia branch had taken on loan commitments . . .*
  - *The average of interbank offered rates plummeted . . .*
A Simple POS Tagger

Define a map $\mathcal{V} \rightarrow \mathcal{L}$. 
A Simple POS Tagger

Define a map \( V \rightarrow L \).

How to pick the single POS for each word? E.g., raises, Fed, . . .
A Simple POS Tagger

Define a map $V \rightarrow L$.

How to pick the single POS for each word? E.g., raises, Fed, ... 

Penn Treebank: most frequent tag rule gives 90.3%, 93.7% if you're clever about handling unknown words.
A Simple POS Tagger

Define a map $\mathcal{V} \rightarrow \mathcal{L}$.

How to pick the single POS for each word? E.g., raises, Fed, . . .

Penn Treebank: most frequent tag rule gives 90.3%, 93.7% if you’re clever about handling unknown words.

All datasets have some errors; estimated upper bound for Penn Treebank is 98%.
Supervised Training of Hidden Markov Models

Given: annotated sequences \( \langle x_1, y_1, \rangle, \ldots, \langle x_n, y_n \rangle \)

\[
p(x, y) = \pi_{y_0} \prod_{i=1}^{\ell+1} \theta_{x_i|y_i} \cdot \gamma_{y_i|y_{i-1}}
\]

Parameters: for each state/label \( y \in L \):

- \( \pi \) is the “start” distribution
- \( \theta_{*|y} \) is the “emission” distribution
- \( \gamma_{*|y} \) is called the “transition” distribution
Supervised Training of Hidden Markov Models

Given: annotated sequences \( \langle \langle x_1, y_1, \rangle, \ldots, \langle x_n, y_n \rangle \rangle \)

\[
p(x, y) = \pi_{y_0} \prod_{i=1}^{\ell+1} \theta_{x_i | y_i} \cdot \gamma_{y_i | y_{i-1}}
\]

Parameters: for each state/label \( y \in \mathcal{L} \):

- \( \pi \) is the “start” distribution
- \( \theta_{*|y} \) is the “emission” distribution
- \( \gamma_{*|y} \) is called the “transition” distribution

Maximum likelihood estimate: count and normalize!
TnT, a trigram HMM tagger with smoothing: 96.7% (Brants, 2000)
TnT, a trigram HMM tagger with smoothing: 96.7% (Brants, 2000)

State of the art: \(~97.5\%\) (Toutanova et al., 2003); uses a feature-based model with:
- capitalization features
- spelling features
- name lists ("gazetteers")
- context words
- hand-crafted patterns
Other Labels

Parts of speech are a minimal *syntactic* representation.

Sequence labeling can get you a lightweight *semantic* representation, too.
Supersenses

A problem with a long history: word-sense disambiguation.
A problem with a long history: word-sense disambiguation.

Classical approaches assumed you had a list of ambiguous words and their senses.

- E.g., from a dictionary
Supersenses

A problem with a long history: word-sense disambiguation.

Classical approaches assumed you had a list of ambiguous words and their senses.

- E.g., from a dictionary

Ciaramita and Johnson (2003) and Ciaramita and Altun (2006) used a lexicon called WordNet to define 41 semantic classes for words.

- WordNet (Fellbaum, 1998) is a fascinating resource in its own right! See http://wordnetweb.princeton.edu/perl/webwn to get an idea.
A problem with a long history: word-sense disambiguation.

Classical approaches assumed you had a list of ambiguous words and their senses.
  ▶ E.g., from a dictionary

Ciaramita and Johnson (2003) and Ciaramita and Altun (2006) used a lexicon called WordNet to define 41 semantic classes for words.
  ▶ WordNet (Fellbaum, 1998) is a fascinating resource in its own right! See http://wordnetweb.princeton.edu/perl/webwn to get an idea.

This represents a coarsening of the annotations in the Semcor corpus (Miller et al., 1993).
Example: *box's Thirteen Synonym Sets, Eight Supersenses*

1. box: a (usually rectangular) container; may have a lid. “he rummaged through a box of spare parts”
2. box/loge: private area in a theater or grandstand where a small group can watch the performance. “the royal box was empty”
3. box/boxful: the quantity contained in a box. “he gave her a box of chocolates”
4. corner/box: a predicament from which a skillful or graceful escape is impossible. “his lying got him into a tight corner”
5. box: a rectangular drawing. “the flowchart contained many boxes”
6. box/boxwood: evergreen shrubs or small trees
7. box: any one of several designated areas on a ball field where the batter or catcher or coaches are positioned. “the umpire warned the batter to stay in the batter’s box”
8. box/box seat: the driver’s seat on a coach. “an armed guard sat in the box with the driver”
9. box: separate partitioned area in a public place for a few people. “the sentry stayed in his box to avoid the cold”
10. box: a blow with the hand (usually on the ear). “I gave him a good box on the ear”
11. box/package: put into a box. “box the gift, please”
12. box: hit with the fist. “I’ll box your ears!”
Example: *box's Thirteen Synonym Sets, Eight Supersenses*

1. box: a (usually rectangular) container; may have a lid. “he rummaged through a box of spare parts” \(\sim\) **N.ARTIFACT**
2. box/loge: private area in a theater or grandstand where a small group can watch the performance. “the royal box was empty” \(\sim\) **N.ARTIFACT**
3. box/boxful: the quantity contained in a box. “he gave her a box of chocolates” \(\sim\) **N.QUANTITY**
4. corner/box: a predicament from which a skillful or graceful escape is impossible. “his lying got him into a tight corner” \(\sim\) **N.STATE**
5. box: a rectangular drawing. “the flowchart contained many boxes” \(\sim\) **N.SHAPE**
6. box/boxwood: evergreen shrubs or small trees \(\sim\) **N.PLANT**
7. box: any one of several designated areas on a ball field where the batter or catcher or coaches are positioned. “the umpire warned the batter to stay in the batter’s box” \(\sim\) **N.ARTIFACT**
8. box/box seat: the driver’s seat on a coach. “an armed guard sat in the box with the driver” \(\sim\) **N.ARTIFACT**
9. box: separate partitioned area in a public place for a few people. “the sentry stayed in his box to avoid the cold” \(\sim\) **N.ARTIFACT**
10. box: a blow with the hand (usually on the ear). “I gave him a good box on the ear” \(\sim\) **N.ACT**
11. box/package: put into a box. “box the gift, please” \(\sim\) **V.CONTACT**
12. box: hit with the fist. “I’ll box your ears!” \(\sim\) **V.CONTACT**
13. box: engage in a boxing match. \(\sim\) **V.COMPETITION**
References I


