Last class we worked through an example where we had a diary of the number of ice creams Jason Eisner ate each date and wanted to get the weather from that.

We had this Hidden Markov Model:

- Transition probabilities:
  - $p(1|C) = 0.7$, $p(2|C) = 0.2$, $p(3|C) = 0.1$
  - $p(1|H) = 0.1$, $p(2|H) = 0.2$, $p(3|H) = 0.7$

- State probabilities:
  - Start state: $0.1$, $0.1$, $0.8$
  - Hot state: $0.8$, $0.1$
  - Cold state: $0.5$, $0.1$
  - End state: $0.1$
Wrapping up the Ice Cream Example

- What is the likelihood that Jason ate 3 ice creams, followed by 2 ice creams, followed by 1 ice cream.

\[
\begin{align*}
\alpha_1(H) &= 0.35 \\
\alpha_1(C) &= 0.05 \\
\alpha_2(H) &= 0.057 \\
\alpha_2(C) &= 0.015 \\
\alpha_3(H) &= 0.00471 \\
\alpha_3(C) &= 0.01239 \\
\alpha_F(H) &= 0.00171
\end{align*}
\]
Wrapping up the Ice Cream Example

- What is the most probable sequence of temperatures given that Jason ate 3 ice creams, followed by 2 ice creams, followed by 1 ice cream. \( \nu_F(E) = 0.00084 \)

![Diagram of temperature sequence and probabilities]
- The Viterbi Algorithm is used to determine the most likely sequence of states given a sequence of observations.
- Given a sentence $w_0 w_1 w_2 \ldots w_n$, what is the most likely sequence of tags $T_0 T_1 T_2 \ldots T_n$?

\[
\nu_1(q) = a_{0q} b_q(o_1), \text{ for all states, } 1 \leq q \leq N \\
\nu_t(q) = \max_r \nu_{t-1}(r) a_{rq} b_q(o_t), \text{ for all states and all times } 1 \leq q \leq N, 1 \leq t \leq T \\
\nu_T(q_F) = \max_r \nu_{T-1}(r) a_r F
\]
The Viterbi Algorithm is used to determine the most likely sequence of states given a sequence of observations.

For each state $q$ from 1 to $N$

$$V[q,1] = P(q|<s>) \times P(w_1|q)$$

$$B[q,1] = 0$$

For each time $t$ from 2 to $T$

For each state $q$ from 1 to $N$

$$V[q,t] = \max [ V[r,t-1] \times P(q|r) \times P(w_t|q) ] \text{ for all states } r \text{ from } 1 \text{ to } N$$

$$B[t][q,t] = \arg\max [ V[r,t-1] \times P(q|r) \times P(w_t|q) ] \text{ for all states } r \text{ from } 1 \text{ to } N$$

score = $\max [ V[r,T] \times P(<\text{end}|r) ] \text{ for all states } r \text{ from } 1 \text{ to } N$

Backtrace_{END} = \arg\max [ V[r,T] \times P(<\text{end}|r) ] \text{ for all states } r \text{ from } 1 \text{ to } N$

Return Reverse(Backtrace Path, Following B[] until 0)
To use an HMM as a PoS tagger, we use the states to represent parts of speech, while the text is the sequence of observations.

An HMM for Part of Speech tagging has a lot more states, upwards of 40 usually.

- $p(\text{baby}|\text{NN}) = 0.5$
- $p(\text{dog}|\text{NN}) = 0.3$
- $p(\text{koala}|\text{NN}) = 0.2$
- $p(\text{a}|\text{H}) = 0.2$
- $p(\text{the}|\text{H}) = 0.7$
- $p(\text{an}|\text{H}) = 0.1$
The list of part of speech tags used to label the training data is called a **tagset**.

In theory, each corpora could define its own, but in practice there are a few common tagsets that most researchers use.

The Penn Treebank Tagset is one of the most commonly used:

- A detailed description of the labels and guidelines to annotators for manually tagging text is available [here](#).
- It has 36 tags, some of which are:

<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD</td>
<td>Cardinal number</td>
</tr>
<tr>
<td>DT</td>
<td>Determiner</td>
</tr>
<tr>
<td>JJ</td>
<td>Adjective</td>
</tr>
<tr>
<td>JJR</td>
<td>Comparative Adjective</td>
</tr>
<tr>
<td>NN</td>
<td>Singular Noun</td>
</tr>
<tr>
<td>NNS</td>
<td>Plural Noun</td>
</tr>
<tr>
<td>NNP</td>
<td>Singular Proper Noun</td>
</tr>
<tr>
<td>VB</td>
<td>Base Verb</td>
</tr>
<tr>
<td>VBZ</td>
<td>3rd Person Singular Present Verb</td>
</tr>
<tr>
<td>VBD</td>
<td>Past Tense Verb</td>
</tr>
</tbody>
</table>
Researchers from Google have proposed a Universal Tagset meant to work with all languages. Because it is universal, it is very coarse, only containing 12 tags. Mappings are provided from existed tagsets. The existing tagsets range in size from 11 tags for Russian to 294 for Mandarin (Sinica Treebank).

Some differentiations made in the Sinica treebank are:

- Between count and mass nouns
- Between intransitive, transitive, and ditransitive verbs
- Between active and stative verbs
- Different types of adverbs
  - Manner
  - Locative
  - Etc.
Annotated Corpora

- A Corpus that has additional information besides the text is known as an **annotated corpus**
  - They are very valuable for many tasks, including PoS tagging
- A common representation of words and their parts of speech is "word/PoS"
  - Implementation of Georgia's automobile title law was also recommended by the outgoing jury.
  - Each subject center library was chosen because of its demonstrated strength in a particular area, which headquarters could then build upon.
Tagging using HMM

- To get the part of speech tags from and HMM, given a sentence, we use the Viterbi algorithm.
- Our trellis will be much bigger in this case, with dimension $T \times N$, where $T$ is the number of tags in the tagset, and $N$ is the length of the sentence.
- Let's look at an example of tagging the sentence:

  That was a great talk.

- We will discuss how to get the probabilities in a few slides.
Rather than the state diagram, this example uses a matrix to hold the \( v \)'s

- Each cell \( i,t \) represents \( v_t(i) \)

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>at</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>bedz</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>dt</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>jj</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>nn</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
v_1(\text{at}) = P(\text{at} | <s>) \cdot P(\text{That} | \text{at})
\]
\[
v_1(\text{bedz}) = P(\text{bedz} | <s>) \cdot P(\text{That} | \text{bedz})
\]
\[
v_1(\text{dt}) = P(\text{dt} | <s>) \cdot P(\text{That} | \text{dt})
\]
\[
v_1(\text{jj}) = P(\text{jj} | <s>) \cdot P(\text{That} | \text{jj})
\]
\[
v_1(\text{nn}) = P(\text{nn} | <s>) \cdot P(\text{That} | \text{nn})
\]
\[
v_1(\text{.}) = P(\text{.} | <s>) \cdot P(\text{That} | \text{.})
\]

That was a great talk.
Rather than the state diagram, this example uses a matrix to hold the v’s.

- Each cell $i,t$ represents $v_t(i)$.

<table>
<thead>
<tr>
<th>at</th>
<th>0.0</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>bedz</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dt</td>
<td>6.8E-4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>jj</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nn</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>t</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>t = 1</td>
<td>That was a great talk.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t = 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t = 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t = 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t = 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t = 6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$v_1(\text{at}) = 0.1848 \times 0.0 = 0$
$v_1(\text{bedz}) = 0.000114 \times 0.0 = 0$
$v_1(\text{dt}) = 0.0363 \times 0.019 = 0.0006897$
$v_1(\text{jj}) = 0.0318 \times 0.0 = 0$
$v_1(\text{nn}) = 0.0571 \times 0.0 = 0$
$v_1(\text{.}) = 0.0016 \times 0.0 = 0$
Rather than the state diagram, this example uses a matrix to hold the v's

- Each cell $i, t$ represents $v_t(i)$

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>at</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bedz</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dt</td>
<td>6.8E-4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>jj</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nn</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>t</th>
<th>t = 1</th>
<th>t = 2</th>
<th>t = 3</th>
<th>t = 4</th>
<th>t = 5</th>
<th>t = 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>That</td>
<td>was</td>
<td>a</td>
<td>great</td>
<td>talk</td>
<td>.</td>
<td></td>
</tr>
</tbody>
</table>

$V_2(at) = \max \{ P(at | x) \cdot P(was | at) \cdot V_1(x) \}$

$V_2(bedz) = \max \{ P(at | x) \cdot P(was | at) \cdot V_1(x) \}$

$V_2(dt) = \max \{ P(at | x) \cdot P(was | at) \cdot V_1(x) \}$

$V_2(jj) = \max \{ P(at | x) \cdot P(was | at) \cdot V_1(x) \}$

$V_2(nn) = \max \{ P(at | x) \cdot P(was | at) \cdot V_1(x) \}$

$V_2(.) = \max \{ P(at | x) \cdot P(was | at) \cdot V_1(x) \}$
Rather than the state diagram, this example uses a matrix to hold the $v$'s

- Each cell $i,t$ represents $v_t(i)$

<table>
<thead>
<tr>
<th></th>
<th>t = 1</th>
<th>t = 2</th>
<th>t = 3</th>
<th>t = 4</th>
<th>t = 5</th>
<th>t = 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>at</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bedz</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dt</td>
<td>6.8 E-4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>jj</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nn</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$v_2(\text{at}) = \max [ \begin{align*}
& P(\text{at} | \text{at}) \cdot P(\text{was} | \text{at}) \cdot v_1(\text{at}), \\
& P(\text{at} | \text{bedz}) \cdot P(\text{was} | \text{bedz}) \cdot v_1(\text{bedz}), \\
& P(\text{at} | \text{dt}) \cdot P(\text{was} | \text{dt}) \cdot v_1(\text{dt}), \\
& P(\text{at} | \text{jj}) \cdot P(\text{was} | \text{jj}) \cdot v_1(\text{jj}), \\
& P(\text{at} | \text{nn}) \cdot P(\text{was} | \text{nn}) \cdot v_1(\text{nn}), \\
& P(\text{at} | .) \cdot P(\text{was} | .) \cdot v_1(.) 
\end{align*} ]$
Rather than the state diagram, this example uses a matrix to hold the \( v \)'s

- Each cell \( i,t \) represents \( v_t(i) \)

\[
\begin{array}{c|ccccccc}
\text{at} & 0.0 & & & & & & \\
\text{bedz} & 0.0 & & & & & & \\
\text{dt} & 6.8 \times 10^{-4} & & & & & & \\
\text{jj} & 0.0 & & & & & & \\
\text{nn} & 0.0 & & & & & & \\
. & 0.0 & & & & & & \\
\end{array}
\]

<table>
<thead>
<tr>
<th>t = 1</th>
<th>t = 2</th>
<th>t = 3</th>
<th>t = 4</th>
<th>t = 5</th>
<th>t = 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>That</td>
<td>was</td>
<td>a</td>
<td>great</td>
<td>talk</td>
<td>.</td>
</tr>
</tbody>
</table>

\[
v_2(\text{at}) = \max \left[ P(\text{at} | \text{at}) \cdot P(\text{was} | \text{at}) \cdot 0.0, P(\text{at} | \text{bedz}) \cdot P(\text{was} | \text{bedz}) \cdot 0.0, P(\text{at} | \text{dt}) \cdot P(\text{was} | \text{dt}) \cdot 0.0006897, P(\text{at} | \text{jj}) \cdot P(\text{was} | \text{jj}) \cdot 0.0, P(\text{at} | \text{nn}) \cdot P(\text{was} | \text{nn}) \cdot 0.0, P(\text{at} | .) \cdot P(\text{was} | .) \cdot 0.0, \right]
\]
Rather than the state diagram, this example uses a matrix to hold the v’s
  
  - Each cell \( i, t \) represents \( v_t(i) \)

\[
\begin{array}{c|c|c|c|c|c|c}
\text{at} & \text{bedz} & \text{dt} & \text{jj} & \text{nn} & \cdot \\
0.0 & 0.0 & 6.8 \times 10^{-4} & 0.0 & 0.0 & 0.0 \\
\hline
\end{array}
\]

<table>
<thead>
<tr>
<th>( t )</th>
<th>( t = 1 )</th>
<th>( t = 2 )</th>
<th>( t = 3 )</th>
<th>( t = 4 )</th>
<th>( t = 5 )</th>
<th>( t = 6 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>\text{That}</td>
<td>\text{was}</td>
<td>\text{a}</td>
<td>\text{great}</td>
<td>\text{talk}</td>
<td>\ldots</td>
<td></td>
</tr>
</tbody>
</table>

\[
v_2(\text{at}) = \max \left[ P(\text{at} | \text{dt}) \times P(\text{was} | \text{dt}) \times 0.0006897, 0.0 \right]
\]
Rather than the state diagram, this example uses a matrix to hold the $v$'s.

- Each cell $i, t$ represents $v_t (i)$

\[
\begin{array}{cccccc}
 & t = 1 & t = 2 & t = 3 & t = 4 & t = 5 & t = 6 \\
\text{at} & 0.0 & 0.0 & & & & \\
\text{bedz} & 0.0 & & & & & \\
\text{dt} & 6.8 \times 10^{-4} & & & & & \\
\text{jj} & 0.0 & & & & & \\
\text{nn} & 0.0 & & & & & \\
. & 0.0 & & & & & \\
\end{array}
\]

$\frac{d}{dt} v_2 (at) = \max \left[ 0.0024 \times 0.0 \times 0.006897 , \right.$

$0.0$

$\left. \right]$

$\frac{d}{dt} v_2 (at) = 0.0$

That was a great talk.
Rather than the state diagram, this example uses a matrix to hold the $v$'s

- Each cell $i,t$ represents $v_t(i)$

```
\[
\begin{array}{cccccc}
\text{at} & \text{bedz} & \text{dt} & \text{jj} & \text{nn} & \text{.} \\
0.0 & 0.0 & 6.8 \times 10^{-4} & 0.0 & 0.0 & 0.0 \\
\end{array}
\]
```

<table>
<thead>
<tr>
<th></th>
<th>$t=1$</th>
<th>$t=2$</th>
<th>$t=3$</th>
<th>$t=4$</th>
<th>$t=5$</th>
<th>$t=6$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>That</strong></td>
<td>was</td>
<td>a</td>
<td><strong>great</strong></td>
<td>talk</td>
<td>.</td>
<td></td>
</tr>
</tbody>
</table>

$v_2(\text{bedz}) = \max[
\begin{align*}
P(\text{bedz} | \text{at}) \times P(\text{was} | \text{at}) \times v_1(\text{at}), \\
P(\text{bedz} | \text{bedz}) \times P(\text{was} | \text{bedz}) \times v_1(\text{bedz}), \\
P(\text{bedz} | \text{dt}) \times P(\text{was} | \text{dt}) \times v_1(\text{dt}), \\
P(\text{bedz} | \text{jj}) \times P(\text{was} | \text{jj}) \times v_1(\text{jj}), \\
P(\text{bedz} | \text{nn}) \times P(\text{was} | \text{nn}) \times v_1(\text{nn}), \\
P(\text{bedz} | .) \times P(\text{was} | .) \times v_1(.),
\end{align*}
]
Rather than the state diagram, this example uses a matrix to hold the v’s

- Each cell i,t represents $v_t(i)$

<table>
<thead>
<tr>
<th></th>
<th>t = 1</th>
<th>t = 2</th>
<th>t = 3</th>
<th>t = 4</th>
<th>t = 5</th>
<th>t = 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>at</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bedz</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dt</td>
<td>6.8E-4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>jj</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nn</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

That was a great talk.

$v_2(\text{bedz}) = \max [P(\text{bedz} | \text{dt}) * P(\text{was} | \text{dt}) * v_1(\text{dt}), 0]$. 
Rather than the state diagram, this example uses a matrix to hold the v’s
- Each cell $i,t$ represents $v_t(i)$

<table>
<thead>
<tr>
<th></th>
<th>t = 1</th>
<th>t = 2</th>
<th>t = 3</th>
<th>t = 4</th>
<th>t = 5</th>
<th>t = 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>at</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bedz</td>
<td>0.0</td>
<td>1.6E-5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dt</td>
<td>6.8E-4</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>jj</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nn</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

That was a great talk.

$$v_2(\text{bedz}) = \max \left[ 0.0229 \times 0.9991 \times 0.0006897, 0 \right]$$

$$v_2(\text{bedz}) = 0.00001577991$$
We also have to update the BT matrix

<table>
<thead>
<tr>
<th></th>
<th>at</th>
<th>bedz</th>
<th>dt</th>
<th>jj</th>
<th>nn</th>
<th>.</th>
</tr>
</thead>
<tbody>
<tr>
<td>at</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bedz</td>
<td>0.0</td>
<td>1.6</td>
<td>6.8</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>dt</td>
<td></td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>jj</td>
<td></td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nn</td>
<td></td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*In reality we would store a numerical index into a tag lookup*
Your Turn

\[ v_t(q) = \max_r v_{t-1}(r) a_{rq} b_q(o_t) \]

<table>
<thead>
<tr>
<th></th>
<th>at</th>
<th>bedz</th>
<th>dt</th>
<th>jj</th>
<th>nn</th>
</tr>
</thead>
<tbody>
<tr>
<td>t = 1</td>
<td>0.0</td>
<td>0.0</td>
<td>6.8E-4</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>t = 2</td>
<td>0.0</td>
<td>1.6E-5</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>t = 3</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>t = 4</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>t = 5</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>t = 6</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

That was a great talk.

\[
P(a | X) = \begin{array}{cccccc}
0.2158 & 0 & 0 & 0 & 1.09E-5 & 0 \\
\end{array}
\]

\[
P(X | \text{bedz}) = \begin{array}{cccccc}
0.1432 & 0 & 0.0041 & 0.0881 & 0.0183 & 0.0025 \\
\end{array}
\]
Getting the probabilities

- The procedure is very similar to building a language model
- But now rather than just sliding a window across the sentence, we have to slide a window across the tags, and a “vertical” window along the word tag pairs
- Use the simple MLE of counts to get probabilities

\[
P(w|t) = \frac{C(w, t)}{C(t)} \quad P(t_{n-1}t_n) = \frac{C(t_{n-1}t_n)}{C(t_{n-1})}
\]
Some Issues to deal with

- **Should we worry about 0 probabilities?**
  - For the tag bigrams, it is a closed set, so even with a relatively small sample size we can be confident we’ve seen almost all possible combinations of tags.

- **What about for words given a tag?**
  - This is a much more likely scenario.
  - Rather than use any smoothing, let’s set our vocabulary ahead of time.

- **Set the vocab to have at most N words in it**
  - Leave the N most frequent words as is.
  - Replace all other words with a special unknown symbol (UNK is a common one).

- **This might result in the sentence**
  - Masala/NP had/HVD chicken/NN today/NR ./.
  - UNK/NP had/HVD chicken/NN today/NR ./.

- **If we use this UNKed corpus to train our we get the probability of UNKNOWN | TAG empirically**
Fancy ways to handle UNKNOWN

- We can calculate UNKNOWN | TAG now, but it turns out there are other correlations to take advantage of.
- Words that start with a capital letter tend to be unproportionally labeled as some sort of noun.
  - To get this count number of words start with a capital labeled with NN, etc.
  - Also count number of words that start with a lower case labeled with tag X.
- Word endings also correlate decently with parts of speech
  - Count the occurrence of the last letter of a word and a label
  - Get fancy and run a morphological analyzer on the word and count the occurrence of the last morpheme and a label.
Evaluation

- To evaluate a tagging method, run the tagger on an unseen corpus for which you know the correct labels for.
- Compare the predicted labels against the correct labels and count the number of labels correctly predicted.
  - \[
    \frac{\text{#Correct}}{\text{Total # of Labels (Sentence Length)}} = \text{Accuracy}
  \]

To/to round/vb out/rp the/at blockading/vbg force/nn ,/, submarines/nns would/md be/be needed/vbn

[to, vb, rp, at, vbg, nn, ‘’, nns, md, be, vbn, .]

[to, vb, rp, at, jj, nn, ‘’, nns, md, be, vbn, .]

Accuracy = \(\frac{11}{12} = 91.666666\)
Evaluation

- Accuracy is a good way to compare two taggers and quickly know which one is “better”.
- Because this is a labeling task, we can get much more informative statistics.
- The simplest is the per class accuracy
  - Out of everything that should have been tagged X, how many were correctly tagged X?
  - This tells us if we are having errors disproportionately with a certain tag
- Taking this further, we can compute a confusion matrix of size TAG x TAG
  - What the the percent of times we incorrectly tagged a word as X, when the correct TAG as actually Y.
  - This tells use if we are confusing two specific tags frequently