Lecture 4: Sequence Models I

Alan Ritter

(many slides from Greg Durrett, Dan Klein, Vivek Srikumar, Chris Manning, Yoav Artzi)
This Lecture

- Sequence modeling
- HMMs for POS tagging
- HMM parameter estimation
- Viterbi, forward-backward
Linguistic Structures

- Language is tree-structured
Linguistic Structures

- Language is tree-structured

*I ate the spaghetti with chopsticks*
Linguistic Structures

- Language is tree-structured

I ate the spaghetti with chopsticks    I ate the spaghetti with meatballs
Linguistic Structures

- Language is tree-structured

I ate the spaghetti with chopsticks  I ate the spaghetti with meatballs
Language is tree-structured

- I ate the spaghetti with chopsticks
- I ate the spaghetti with meatballs
Language is tree-structured

Understanding syntax fundamentally requires trees — the sentences have the same shallow analysis.
Linguistic Structures

- Language is tree-structured

- Understanding syntax fundamentally requires trees — the sentences have the same shallow analysis

```
I ate the spaghetti with chopsticks
```

```
I ate the spaghetti with meatballs
```
Linguistic Structures

- Language is sequentially structured: interpreted in an online way

Tanenhaus et al. (1995)
Language is sequentially structured: interpreted in an online way

Tanenhaus et al. (1995)
Linguistic Structures

- Language is sequentially structured: interpreted in an online way

Tanenhaus et al. (1995)
What tags are out there?

Ghana’s ambassador should have set up the big meeting in DC yesterday.
POS Tagging

Open class (lexical) words

Nouns
- Proper: IBM, Italy
- Common: cat / cats, snow

Verbs
- Main: see, registered

Adjectives
- yellow

Adverbs
- slowly

Numbers
- 122,312, one

Closed class (functional)

Determiners: the, some

Conjunctions: and, or

Pronouns: he, its

Auxiliary
- can, had

Prepositions: to, with

Particles: off, up

... more

... more

Slide credit: Dan Klein
POS Tagging
Fed raises interest rates 0.5 percent
Fed raises interest rates 0.5 percent

I hereby increase interest rates 0.5%
Fed raises interest rates 0.5 percent

I hereby increase interest rates 0.5%
Fed raises interest rates 0.5 percent

I hereby increase interest rates 0.5%
Fed raises interest rates 0.5 percent

I hereby increase interest rates 0.5%
Fed raises interest rates 0.5 percent

I hereby increase interest rates 0.5%
POS Tagging

Fed raises interest rates 0.5 percent

I hereby increase interest rates 0.5%
Fed raises interest rates 0.5 percent

I hereby increase interest rates 0.5%
Fed raises interest rates 0.5 percent

I hereby increase interest rates 0.5%

I’m 0.5% interested in the Fed’s raises!
Fed raises interest rates 0.5 percent

I hereby increase interest rates 0.5%

I’m 0.5% interested in the Fed’s raises!

Other paths are also plausible but even more semantically weird...
Fed raises interest rates 0.5 percent

Fed raises interest rates 0.5 percent

I hereby increase interest rates 0.5%

I’m 0.5% interested in the Fed’s raises!

Other paths are also plausible but even more semantically weird...

What governs the correct choice? Word + context
Fed raises interest rates 0.5 percent

I hereby increase interest rates 0.5%

I’m 0.5% interested in the Fed’s raises!

Other paths are also plausible but even more semantically weird...

What governs the correct choice? Word + context

- Word identity: most words have <=2 tags, many have one (percent, the)
- Context: nouns start sentences, nouns follow verbs, etc.
# POS Tagging

<table>
<thead>
<tr>
<th>CC</th>
<th>conjunction, coordinating</th>
<th>and both but either or</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD</td>
<td>numeral, cardinal</td>
<td>mid-1890 nine-thirty 0.5 one</td>
</tr>
<tr>
<td>DT</td>
<td>determiner</td>
<td>a all an every no that the</td>
</tr>
<tr>
<td>EX</td>
<td>existential there</td>
<td>there</td>
</tr>
<tr>
<td>FW</td>
<td>foreign word</td>
<td>gemeinschaft hund ich jeux</td>
</tr>
<tr>
<td>IN</td>
<td>preposition or conjunction, subordinating</td>
<td>among whether out on by if</td>
</tr>
<tr>
<td>JJ</td>
<td>adjective or numeral, ordinal</td>
<td>third ill-mannered regrettable</td>
</tr>
<tr>
<td>JJR</td>
<td>adjective, comparative</td>
<td>braver cheaper taller</td>
</tr>
<tr>
<td>JJS</td>
<td>adjective, superlative</td>
<td>bravest cheapest tallest</td>
</tr>
<tr>
<td>MD</td>
<td>modal auxiliary</td>
<td>can may might will would</td>
</tr>
<tr>
<td>NN</td>
<td>noun, common, singular or mass</td>
<td>cabbage thermostat investment subhumanity</td>
</tr>
<tr>
<td>NNP</td>
<td>noun, proper, singular</td>
<td>Molown Cougar Yvette Liverpool</td>
</tr>
<tr>
<td>NNPS</td>
<td>noun, proper, plural</td>
<td>Americans Materials States</td>
</tr>
<tr>
<td>NNS</td>
<td>noun, common, plural</td>
<td>undergraduates bric-a-brac averages</td>
</tr>
<tr>
<td>POS</td>
<td>genitive marker</td>
<td>'s</td>
</tr>
<tr>
<td>PRP</td>
<td>pronoun, personal</td>
<td>hers himself it we them</td>
</tr>
<tr>
<td>PRPS</td>
<td>pronoun, possessive</td>
<td>her his mine my our ours their thy your</td>
</tr>
<tr>
<td>RB</td>
<td>adverb</td>
<td>occasionally maddeningly adventurously</td>
</tr>
<tr>
<td>RBR</td>
<td>adverb, comparative</td>
<td>further gloomier heavier less-perfectly</td>
</tr>
<tr>
<td>RBS</td>
<td>adverb, superlative</td>
<td>best biggest nearest worst</td>
</tr>
<tr>
<td>RP</td>
<td>particle</td>
<td>aboard away back by on open through</td>
</tr>
<tr>
<td>TO</td>
<td>&quot;to&quot; as preposition or infinitive marker</td>
<td>to</td>
</tr>
<tr>
<td>UH</td>
<td>interjection</td>
<td>huh howdy uh whammo shucks heck</td>
</tr>
<tr>
<td>VB</td>
<td>verb, base form</td>
<td>ask bring fire see take</td>
</tr>
<tr>
<td>VBD</td>
<td>verb, past tense</td>
<td>pleaded swiped registered saw</td>
</tr>
<tr>
<td>VBG</td>
<td>verb, present participle or gerund</td>
<td>stirring focusing approaching erasing</td>
</tr>
<tr>
<td>VBN</td>
<td>verb, past participle</td>
<td>dilapidated imitated reunited unsettled</td>
</tr>
<tr>
<td>VBP</td>
<td>verb, present tense, not 3rd person singular</td>
<td>twist appear comprise mold postpone</td>
</tr>
<tr>
<td>VBZ</td>
<td>verb, present tense, 3rd person singular</td>
<td>bases reconstructs marks uses</td>
</tr>
<tr>
<td>WDT</td>
<td>WH-determiner</td>
<td>that what whatever which whichever</td>
</tr>
<tr>
<td>WP</td>
<td>WH-pronoun</td>
<td>that what whatever which who whom</td>
</tr>
<tr>
<td>WP$</td>
<td>WH-pronoun, possessive</td>
<td>whose</td>
</tr>
<tr>
<td>WRB</td>
<td>Wh-adverb</td>
<td>however whenever where why</td>
</tr>
</tbody>
</table>
What is this good for?
What is this good for?

- Text-to-speech: *record, lead*
What is this good for?

- Text-to-speech: *record, lead*
- Preprocessing step for syntactic parsers
What is this good for?

- Text-to-speech: *record, lead*
- Preprocessing step for syntactic parsers
- Domain-independent disambiguation for other tasks
What is this good for?

- Text-to-speech: *record, lead*
- Preprocessing step for syntactic parsers
- Domain-independent disambiguation for other tasks
- (Very) shallow information extraction
Sequence Models
Sequence Models

- Input $x = (x_1, \ldots, x_n)$  
  
  Output $y = (y_1, \ldots, y_n)$
Sequence Models

- Input $x = (x_1, \ldots, x_n)$  Output $y = (y_1, \ldots, y_n)$

- POS tagging: $x$ is a sequence of words, $y$ is a sequence of tags
Sequence Models

- Input \( x = (x_1, \ldots, x_n) \)  Output \( y = (y_1, \ldots, y_n) \)

- POS tagging: \( x \) is a sequence of words, \( y \) is a sequence of tags

- Today: generative models \( P(x, y) \); discriminative models next time
Hidden Markov Models

- Input $x = (x_1, \ldots, x_n)$  Output $y = (y_1, \ldots, y_n)$
Hidden Markov Models

- Input $x = (x_1, \ldots, x_n)$  
  Output $y = (y_1, \ldots, y_n)$

- Model the sequence of $y$ as a Markov process (dynamics model)
Hidden Markov Models

- Input $x = (x_1, ..., x_n)$  Output $y = (y_1, ..., y_n)$

- Model the sequence of $y$ as a Markov process (dynamics model)

- Markov property: future is conditionally independent of the past given the present
Hidden Markov Models

- Input \( x = (x_1, \ldots, x_n) \)  
  Output \( y = (y_1, \ldots, y_n) \)

- Model the sequence of \( y \) as a Markov process (dynamics model)

- Markov property: future is conditionally independent of the past given the present

\[
P(y_3|y_1, y_2) = P(y_3|y_2)
\]
Hidden Markov Models

- Input \( x = (x_1, \ldots, x_n) \)  
  Output \( y = (y_1, \ldots, y_n) \)

- Model the sequence of \( y \) as a Markov process (dynamics model)

- Markov property: future is conditionally independent of the past given the present

\[
P(y_3 | y_1, y_2) = P(y_3 | y_2)
\]

- Lots of mathematical theory about how Markov chains behave
Hidden Markov Models

- Input $x = (x_1, \ldots, x_n)$  Output $y = (y_1, \ldots, y_n)$

- Model the sequence of $y$ as a Markov process (dynamics model)

- Markov property: future is conditionally independent of the past given the present

\[ P(y_3|y_1, y_2) = P(y_3|y_2) \]

- Lots of mathematical theory about how Markov chains behave

- If $y$ are tags, this roughly corresponds to assuming that the next tag only depends on the current tag, not anything before
Hidden Markov Models

- Input $x = (x_1, \ldots, x_n)$  Output $y = (y_1, \ldots, y_n)$
Hidden Markov Models

- Input $x = (x_1, \ldots, x_n)$  
  Output $y = (y_1, \ldots, y_n)$

```
Fed raises ... percent
```

```
\begin{align*}
  y_1 & \rightarrow y_2 & \rightarrow & \ldots & \rightarrow & y_n \\
  x_1 & \rightarrow & x_2 & \rightarrow & \ldots & \rightarrow & x_n
\end{align*}
```
Hidden Markov Models

- Input $x = (x_1, \ldots, x_n)$
- Output $y = (y_1, \ldots, y_n)$
Hidden Markov Models

- Input $x = (x_1, \ldots, x_n)$  Output $y = (y_1, \ldots, y_n)$

$$P(y, x) = P(y_1) \prod_{i=2}^{n} P(y_i|y_{i-1}) \prod_{i=1}^{n} P(x_i|y_i)$$
Hidden Markov Models

- Input $\mathbf{x} = (x_1, \ldots, x_n)$  
  Output $\mathbf{y} = (y_1, \ldots, y_n)$

$$
\begin{align*}
\mathbf{y} & = (y_1, y_2, \ldots, y_n) \\
\mathbf{x} & = (x_1, x_2, \ldots, x_n)
\end{align*}
$$

$$
P(\mathbf{y}, \mathbf{x}) = P(y_1) \prod_{i=2}^{n} P(y_i | y_{i-1}) \prod_{i=1}^{n} P(x_i | y_i)
$$

Initial distribution
Hidden Markov Models

- Input $x = (x_1, \ldots, x_n)$  
  Output $y = (y_1, \ldots, y_n)$

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

- Initial distribution
- Transition probabilities
Hidden Markov Models

- Input $x = (x_1, \ldots, x_n)$  
  Output $y = (y_1, \ldots, y_n)$

$$P(y, x) = P(y_1) \prod_{i=2}^{n} P(y_i | y_{i-1}) \prod_{i=1}^{n} P(x_i | y_i)$$

- Initial distribution
- Transition probabilities
- Emission probabilities
Hidden Markov Models

- Input $x = (x_1, \ldots, x_n)$
- Output $y = (y_1, \ldots, y_n)$

- Observation $(x)$ depends only on current state $(y)$

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

- Initial distribution
- Transition probabilities
- Emission probabilities
**Hidden Markov Models**

- Input \( x = (x_1, \ldots, x_n) \)  
  Output \( y = (y_1, \ldots, y_n) \)

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

- Observation \( (x) \) depends only on current state \( (y) \)
- Multinomials: tag x tag transitions, tag x word emissions

![Diagram of Hidden Markov Model](image-url)
Hidden Markov Models

- Input $x = (x_1, ..., x_n)$  
  Output $y = (y_1, ..., y_n)$

- Observation $(x)$ depends only on current state $(y)$

- Multinomials: tag $x$ tag transitions, tag $x$ word emissions

- $P(x|y)$ is a distribution over all words in the vocabulary — not a distribution over features (but could be!)

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

- Initial distribution
- Transition probabilities
- Emission probabilities
Transitions in POS Tagging

- Dynamics model \( P(y_1) \prod_{i=2}^{n} P(y_i | y_{i-1}) \)

  - **VBD**: verb, past tense
  - **VB**: verb, base form
  - **VBN**: verb, past participle
  - **VBZ**: verb, 3rd person singular present
  - **VBP**: verb, present participle
  - **NNP**: proper noun, singular
  - **NNS**: noun, plural
  - **NN**: noun, singular or mass
  - **CD**: coordinating conjunction
  - **.**: period

*Fed raises interest rates 0.5 percent.*
Transitions in POS Tagging

- Dynamics model: \( P(y_1) \prod_{i=2}^{n} P(y_i | y_{i-1}) \)

- \( y_1 = \text{NNP} \) likely because start of sentence

- Fed raises interest rates 0.5 percent.

- **NNP** - proper noun, singular
- **VBZ** - verb, 3rd ps. sing. present
- **NN** - noun, singular or mass
Transitions in POS Tagging

- Dynamics model
  \[ P(y_1) \prod_{i=2}^{n} P(y_i | y_{i-1}) \]

- Fed raises interest rates 0.5 percent.

- \( P(y_1 = \text{NNP}) \) likely because start of sentence

- \( P(y_2 = \text{VBZ} | y_1 = \text{NNP}) \) likely because verb often follows noun

- **NNP** - proper noun, singular
- **VBZ** - verb, 3rd ps. sing. present
- **NN** - noun, singular or mass

Fed raises interest rates 0.5 percent.
Transitions in POS Tagging

- Dynamics model 
  \[ P(y_1) \prod_{i=2}^{n} P(y_i | y_{i-1}) \]

  - VBD - verb, past tense
  - VB - verb, base form
  - VBN - verb, past participle
  - VBZ - verb, 3rd person singular present
  - VBP - verb, present participle
  - NNP - proper noun, singular
  - NNS - noun, plural
  - NN - noun, singular or mass
  - CD - conjunction, coordinating

  Fed raises interest rates 0.5 percent.

- \( P(y_1 = \text{NNP}) \) likely because start of sentence

- \( P(y_2 = \text{VBZ}|y_1 = \text{NNP}) \) likely because verb often follows noun

- \( P(y_3 = \text{NN}|y_2 = \text{VBZ}) \) direct object follows verb, other verb rarely follows past tense verb (main verbs can follow modals though!)
Fed raises interest rates 0.5 percent.
Similar to Naive Bayes estimation: maximum likelihood solution = normalized counts (with smoothing) read off supervised data
Fed raises interest rates 0.5 percent.

Similar to Naive Bayes estimation: maximum likelihood solution = normalized counts (with smoothing) read off supervised data

\[ P(\text{tag} \mid \text{NN}) \]
Fed raises interest rates 0.5 percent.

Similar to Naive Bayes estimation: maximum likelihood solution = normalized counts (with smoothing) read off supervised data

\[ P(\text{tag} \mid \text{NN}) = (0.5, 0.5 \text{ NNS}) \]
Estimating Transitions

Fed raises interest rates 0.5 percent.

- Similar to Naive Bayes estimation: maximum likelihood solution = normalized counts (with smoothing) read off supervised data
- \( P(\text{tag} \mid \text{NN}) = (0.5, 0.5, \text{NNS}) \)
- How to smooth?
Estimating Transitions

Similar to Naive Bayes estimation: maximum likelihood solution = normalized counts (with smoothing) read off supervised data

P(tag | NN) = (0.5 ., 0.5 NNS)

How to smooth?

One method: smooth with unigram distribution over tags
Estimating Transitions

Fed raises interest rates 0.5 percent.

- Similar to Naive Bayes estimation: maximum likelihood solution = normalized counts (with smoothing) read off supervised data
  - \( P(\text{tag} \mid \text{NN}) = (0.5, 0.5 \text{ NNS}) \)
  - How to smooth?
  - One method: smooth with unigram distribution over tags
  \[
P(\text{tag}|\text{tag}_{-1}) = (1 - \lambda)\hat{P}(\text{tag}|\text{tag}_{-1}) + \lambda\hat{P}(\text{tag})
\]
Fed raises interest rates 0.5 percent.

- Similar to Naive Bayes estimation: maximum likelihood solution = normalized counts (with smoothing) read off supervised data
- \( P(\text{tag} \mid \text{NN}) = (0.5, 0.5 \text{ NNS}) \)
- How to smooth?
- One method: smooth with unigram distribution over tags

\[
P(\text{tag} \mid \text{tag}_{-1}) = (1 - \lambda)\hat{P}(\text{tag} \mid \text{tag}_{-1}) + \lambda\hat{P}(\text{tag})
\]

\( \hat{P} \) = empirical distribution (read off from data)
Fed raises interest rates 0.5 percent.
Emissions in POS Tagging

NNP VBZ NN NNS CD NN .
Fed raises interest rates 0.5 percent .

- Emissions $P(x \mid y)$ capture the distribution of words occurring with a given tag
Emissions in POS Tagging

- Emissions $P(x \mid y)$ capture the distribution of words occurring with a given tag
- $P(\text{word} \mid \text{NN}) = (0.05 \, \text{person}, 0.04 \, \text{official}, 0.03 \, \text{interest}, 0.03 \, \text{percent} \, ...)$

*Fed raises interest rates 0.5 percent.*
Emissions in POS Tagging

NNP VBZ NN NNS CD NN .

*Fed raises interest rates 0.5 percent .*

- Emissions $P(x \mid y)$ capture the distribution of words occurring with a given tag
- $P(\text{word} \mid \text{NN}) = (0.05 \text{ person}, 0.04 \text{ official}, 0.03 \text{ interest}, 0.03 \text{ percent} \ldots)$
- When you compute the posterior for a given word’s tags, the distribution favors tags that are more likely to generate that word
Emissions in POS Tagging

- Emissions $P(x \mid y)$ capture the distribution of words occurring with a given tag

  - $P(\text{word} \mid \text{NN}) = (0.05 \text{ person}, 0.04 \text{ official}, 0.03 \text{ interest}, 0.03 \text{ percent} \ldots)$

- When you compute the posterior for a given word’s tags, the distribution favors tags that are more likely to generate that word

- How should we smooth this?
Fed raises interest rates 0.5 percent
Fed raises interest rates 0.5 percent

- $P(\text{word} \mid \text{NN}) = (0.5 \text{ interest}, 0.5 \text{ percent})$ — hard to smooth!
Fed raises interest rates 0.5 percent

P(word | NN) = (0.5 interest, 0.5 percent) — hard to smooth!

Can interpolate with distribution looking at word shape
P(word shape | tag) (e.g., P(capitalized word of len >= 8 | tag))
Estimating Emissions

Fed raises interest rates 0.5 percent

- P(word | NN) = (0.5 interest, 0.5 percent) — hard to smooth!

- Can interpolate with distribution looking at word shape
  P(word shape | tag) (e.g., P(capitalized word of len >= 8 | tag))

- Alternative: use Bayes’ rule
Fed raises interest rates 0.5 percent

- $P(\text{word} \mid \text{NN}) = (0.5 \text{ interest}, 0.5 \text{ percent})$ — hard to smooth!

- Can interpolate with distribution looking at word shape
  $P(\text{word shape} \mid \text{tag})$ (e.g., $P(\text{capitalized word of len} \geq 8 \mid \text{tag})$)

- Alternative: use Bayes’ rule

\[
P(\text{word} \mid \text{tag}) = \frac{P(\text{tag} \mid \text{word})P(\text{word})}{P(\text{tag})}
\]
Estimating Emissions

Fed raises interest rates 0.5 percent

- $P(\text{word} \mid \text{NN}) = (0.5 \textit{interest}, 0.5 \textit{percent})$ — hard to smooth!

- Can interpolate with distribution looking at word shape
  $P(\text{word shape} \mid \text{tag})$ (e.g., $P(\text{capitalized word of len } \geq 8 \mid \text{tag})$)

- Alternative: use Bayes’ rule
  $$P(\text{word} \mid \text{tag}) = \frac{P(\text{tag} \mid \text{word})P(\text{word})}{P(\text{tag})}$$

  - Fancy techniques from language modeling, e.g. look at type fertility
    — $P(\text{tag} \mid \text{word})$ is flatter for some kinds of words than for others
Estimating Emissions

Fed raises interest rates 0.5 percent

- \( P(\text{word} \mid \text{NN}) = (0.5 \text{ interest, 0.5 percent}) \) — hard to smooth!

- Can interpolate with distribution looking at word shape
  \( P(\text{word shape} \mid \text{tag}) \) (e.g., \( P(\text{capitalized word of len } \geq 8 \mid \text{tag}) \))

- Alternative: use Bayes’ rule
  \[
  P(\text{word} \mid \text{tag}) = \frac{P(\text{tag} \mid \text{word}) P(\text{word})}{P(\text{tag})}
  \]
  
  - Fancy techniques from language modeling, e.g. look at type fertility
    — \( P(\text{tag} \mid \text{word}) \) is flatter for some kinds of words than for others

- \( P(\text{word} \mid \text{tag}) \) can be a log-linear model — we’ll see this in a few lectures
**Inference in HMMs**

- **Input** $\mathbf{x} = (x_1, ..., x_n)$
- **Output** $\mathbf{y} = (y_1, ..., y_n)$

$$P(\mathbf{y}, \mathbf{x}) = P(y_1) \prod_{i=2}^{n} P(y_i | y_{i-1}) \prod_{i=1}^{n} P(x_i | y_i)$$
Inference in HMMs

- Input \( x = (x_1, \ldots, x_n) \)  
  
- Output \( y = (y_1, \ldots, y_n) \)

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

- Inference problem: \( \arg\max_y P(y | x) = \arg\max_y \frac{P(y, x)}{P(x)} \)
**Inference in HMMs**

- **Input** \( \mathbf{x} = (x_1, \ldots, x_n) \)  
  **Output** \( \mathbf{y} = (y_1, \ldots, y_n) \)

- **Inference problem:** \( \arg\max_{\mathbf{y}} P(\mathbf{y} | \mathbf{x}) = \arg\max_{\mathbf{y}} \frac{P(\mathbf{y}, \mathbf{x})}{P(\mathbf{x})} \)

\[
P(\mathbf{y}, \mathbf{x}) = P(y_1) \prod_{i=2}^{n} P(y_i | y_{i-1}) \prod_{i=1}^{n} P(x_i | y_i)
\]
Inference in HMMs

- **Input** $x = (x_1, \ldots, x_n)$  
  - **Output** $y = (y_1, \ldots, y_n)$

$$P(y, x) = P(y_1) \prod_{i=2}^{n} P(y_i | y_{i-1}) \prod_{i=1}^{n} P(x_i | y_i)$$

- Inference problem: $\arg\max_y P(y | x) = \arg\max_y \frac{P(y, x)}{P(x)}$
- Exponentially many possible $y$ here!
Inference in HMMs

- **Input** \( \mathbf{x} = (x_1, \ldots, x_n) \) \hspace{1cm} **Output** \( \mathbf{y} = (y_1, \ldots, y_n) \)

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

- Inference problem: \( \arg\max_{\mathbf{y}} P(\mathbf{y} | \mathbf{x}) = \arg\max_{\mathbf{y}} \frac{P(\mathbf{y}, \mathbf{x})}{P(\mathbf{x})} \)

- Exponentially many possible \( \mathbf{y} \) here!

- Solution: dynamic programming (possible because of Markov structure!)
Inference in HMMs

- Input $x = (x_1, \ldots, x_n)$  
  Output $y = (y_1, \ldots, y_n)$

- Inference problem: $\arg\max_y P(y|x) = \arg\max_y \frac{P(y, x)}{P(x)}$

- Exponentially many possible $y$ here!

- Solution: dynamic programming (possible because of Markov structure!)

  - Many neural sequence models depend on entire previous tag sequence, need to use approximations like beam search

$$P(y, x) = P(y_1) \prod_{i=2}^{n} P(y_i|y_{i-1}) \prod_{i=1}^{n} P(x_i|y_i)$$
Viterbi Algorithm

\[ P(x_1, x_2, \cdots, x_n, y_1, y_2, \cdots y_n) = P(y_1) \prod_{i=1}^{n-1} P(y_{i+1}|y_i) \prod_{i=1}^{n} P(x_i|y_i) \]

\[
\max_{y_1, y_2, \cdots, y_n} P(y_n|y_{n-1})P(x_n|y_n) \cdots P(y_2|y_1)P(x_2|y_2)P(y_1)P(x_1|y_1)
\]
Viterbi Algorithm

\[
P(x_1, x_2, \cdots, x_n, y_1, y_2, \cdots y_n) = P(y_1) \prod_{i=1}^{n-1} P(y_{i+1}|y_i) \prod_{i=1}^{n} P(x_i|y_i)
\]

\[
\max_{y_1, y_2, \cdots, y_n} P(y_n|y_{n-1})P(x_n|y_n) \cdots P(y_2|y_1)P(x_2|y_2)P(y_1)P(x_1|y_1)
\]

\[
= \max_{y_2, y_3, \cdots, y_n} P(y_n|y_{n-1})P(x_n|y_n) \cdots \max_{y_1}P(y_2|y_1)P(x_2|y_2)P(y_1)P(x_1|y_1)
\]

The only terms that depend on \( y_1 \)

---

slide credit: Vivek Srikumar
Viterbi Algorithm

\[ P(x_1, x_2, \cdots, x_n, y_1, y_2, \cdots, y_n) = P(y_1) \prod_{i=1}^{n-1} P(y_{i+1} | y_i) \prod_{i=1}^{n} P(x_i | y_i) \]

\[
\max_{y_1, y_2, \cdots, y_n} P(y_n | y_{n-1}) P(x_n | y_n) \cdots P(y_2 | y_1) P(x_2 | y_2) P(y_1) P(x_1 | y_1)
\]

\[
= \max_{y_2, \cdots, y_n} P(y_n | y_{n-1}) P(x_n | y_n) \cdots \max_{y_1} P(y_2 | y_1) P(x_2 | y_2) P(y_1) P(x_1 | y_1)
\]

\[
= \max_{y_2, \cdots, y_n} P(y_n | y_{n-1}) P(x_n | y_n) \cdots \max_{y_1} P(y_2 | y_1) P(x_2 | y_2) \text{score}_1(y_1)
\]

Abstract away the score for all decisions till here into \( \text{score} \)

\[
\text{score}_1(s) = P(s) P(x_1 | s)
\]

slide credit: Vivek Srikumar
Viterbi Algorithm

\[ P(x_1, x_2, \ldots, x_n, y_1, y_2, \ldots y_n) = P(y_1) \prod_{i=1}^{n-1} P(y_{i+1} | y_i) \prod_{i=1}^{n} P(x_i | y_i) \]

\[
\begin{align*}
\max_{y_1, y_2, \ldots, y_n} P(y_n | y_{n-1}) P(x_n | y_n) & \cdot P(y_{n-1} | y_{n-2}) P(x_{n-1} | y_{n-1}) \cdots P(y_2 | y_1) P(x_2 | y_2) P(y_1 | y_1) P(x_1 | y_1) \\
= \max_{y_2, \ldots, y_n} P(y_n | y_{n-1}) P(x_n | y_n) & \cdot \max_{y_1} P(y_{n-1} | y_{n-2}) P(x_{n-1} | y_{n-1}) \cdots P(y_2 | y_1) P(x_2 | y_2) P(y_1 | y_1) P(x_1 | y_1) \\
= \max_{y_2, \ldots, y_n} P(y_n | y_{n-1}) P(x_n | y_n) & \cdot \max_{y_1} P(y_{n-1} | y_{n-2}) P(x_{n-1} | y_{n-1}) \cdots \max_{y_1} P(y_2 | y_1) P(x_2 | y_2) \text{score}_1(y_1)
\end{align*}
\]

Abstract away the score for all decisions till here into score

\[ \text{score}_1(s) = P(s)P(x_1 | s) \]

best (partial) score for a sequence ending in state \( s \)

slide credit: Vivek Srikumar
Viterbi Algorithm

\[ P(x_1, x_2, \cdots, x_n, y_1, y_2, \cdots y_n) = P(y_1) \prod_{i=1}^{n-1} P(y_{i+1} | y_i) \prod_{i=1}^{n} P(x_i | y_i) \]

\[
\max_{y_1, y_2, \cdots, y_n} P(y_n | y_{n-1}) P(x_n | y_n) \cdots P(y_2 | y_1) P(x_2 | y_2) P(y_1) P(x_1 | y_1)
\]

\[
= \max_{y_2, \cdots, y_n} P(y_n | y_{n-1}) P(x_n | y_n) \cdots \max_{y_1} P(y_2 | y_1) P(x_2 | y_2) P(y_1) P(x_1 | y_1)
\]

\[
= \max_{y_2, \cdots, y_n} P(y_n | y_{n-1}) P(x_n | y_n) \cdots \max_{y_2} P(y_3 | y_2) P(x_3 | y_3) \max_{y_1} P(y_2 | y_1) P(x_2 | y_2) \text{score}_1(y_1)
\]

Only terms that depend on \(y_2\)

slide credit: Vivek Srikumar
Viterbi Algorithm

\[ P(x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n) = P(y_1) \prod_{i=1}^{n-1} P(y_{i+1}|y_i) \prod_{i=1}^n P(x_i|y_i) \]

\[
\max_{y_1, y_2, \ldots, y_n} P(y_n|y_{n-1})P(x_n|y_n) \cdots P(y_2|y_1)P(x_2|y_2)P(x_1|y_1)
\]

\[
= \max_{y_2, \ldots, y_n} P(y_n|y_{n-1})P(x_n|y_n) \cdots \max_{y_1} P(y_2|y_1)P(x_2|y_2)P(x_1|y_1)
\]

\[
= \max_{y_2, \ldots, y_n} P(y_n|y_{n-1})P(x_n|y_n) \cdots \max_{y_1} P(y_3|y_2)P(x_3|y_3) \max_{y_1} P(y_2|y_1)P(x_2|y_2)
\]

\[
= \max_{y_3, \ldots, y_n} P(y_n|y_{n-1})P(x_n|y_n) \cdots \max_{y_2} P(y_3|y_2)P(x_3|y_3)\score_2(y_2)
\]

\[
\score_i(s) = \max_{y_{i-1}} P(s|y_{i-1})P(x_i|s)\score_{i-1}(y_{i-1})
\]

Abstract away the score for all decisions till here into \( \score \)
Viterbi Algorithm

slide credit: Dan Klein
“Think about” all possible immediate prior state values. Everything before that has already been accounted for by earlier stages.
Viterbi Algorithm

\[ P(x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n) = P(y_1) \prod_{i=1}^{n-1} P(y_{i+1} | y_i) \prod_{i=1}^{n} P(x_i | y_i) \]

\[
\max_{y_1, y_2, \ldots, y_n} P(y_n | y_{n-1}) P(x_n | y_n) \cdots P(y_2 | y_1) P(x_2 | y_2) P(y_1) P(x_1 | y_1)
\]

\[
= \max_{y_2, \ldots, y_n} P(y_n | y_{n-1}) P(x_n | y_n) \cdots \max_{y_1} P(y_2 | y_1) P(x_2 | y_2) \text{score}_1(y_1)
\]

\[
= \max_{y_2, \ldots, y_n} P(y_n | y_{n-1}) P(x_n | y_n) \cdots \max_{y_2} P(y_3 | y_2) P(x_3 | y_3) \max_{y_1} P(y_2 | y_1) P(x_2 | y_2) \text{score}_1(y_1)
\]

\[
= \max_{y_3, \ldots, y_n} P(y_n | y_{n-1}) P(x_n | y_n) \cdots \max_{y_2} P(y_3 | y_2) P(x_3 | y_3) \text{score}_2(y_2)
\]

\[\vdots\]

\[
= \max_{y_n} \text{score}_n(y_n)
\]

Abstract away the score for all decisions till here into score

slide credit: Vivek Srikumar
Viterbi Algorithm

\[
P(x_1, x_2, \cdots, x_n, y_1, y_2, \cdots, y_n) = P(y_1) \prod_{i=1}^{n-1} P(y_{i+1}|y_i) \prod_{i=1}^{n} P(x_i|y_i)
\]

\[
\max_{y_1, y_2, \cdots, y_n} P(y_n|y_{n-1})P(x_n|y_n) \cdots P(y_2|y_1)P(x_2|y_2)P(y_1)P(x_1|y_1)
\]

\[
= \max_{y_2, \cdots, y_n} P(y_n|y_{n-1})P(x_n|y_n) \cdots \max_{y_1} P(y_2|y_1)P(x_2|y_2)P(y_1)P(x_1|y_1)
\]

\[
= \max_{y_2, \cdots, y_n} P(y_n|y_{n-1})P(x_n|y_n) \cdots \max_{y_1} P(y_2|y_1)P(x_2|y_2)\text{score}_1(y_1)
\]

\[
= \max_{y_3, \cdots, y_n} P(y_n|y_{n-1})P(x_n|y_n) \cdots \max_{y_2} P(y_3|y_2)P(x_3|y_3)\max_{y_1} P(y_2|y_1)P(x_2|y_2)\text{score}_1(y_1)
\]

\[
= \max_{y_3, \cdots, y_n} P(y_n|y_{n-1})P(x_n|y_n) \cdots \max_{y_2} P(y_3|y_2)P(x_3|y_3)\text{score}_2(y_2)
\]

\[\vdots\]

\[
= \max_{y_n} \text{score}_n(y_n)
\]

\[
\text{score}_1(s) = P(s)P(x_1|s)
\]

\[
\text{score}_i(s) = \max_{y_{i-1}} P(s|y_{i-1})P(x_i|s)\text{score}_{i-1}(y_{i-1})
\]

slide credit: Vivek Srikumar
Viterbi Algorithm

1. Initial: For each state $s$, calculate
   \[ \text{score}_1(s) = P(s)P(x_1|s) = \pi_s B_{x_1,s} \]

2. Recurrence: For $i = 2$ to $n$, for every state $s$, calculate
   \[ \text{score}_i(s) = \max_{y_{i-1}} P(s|y_{i-1}) P(x_i|s) \text{score}_{i-1}(y_{i-1}) \]
   \[ = \max_{y_{i-1}} A_{y_{i-1},s} B_{s,x_i} \text{score}_{i-1}(y_{i-1}) \]

3. Final state: calculate
   \[ \max_y P(y, x|\pi, A, B) = \max_s \text{score}_n(s) \]

This only calculates the max. To get final answer (argmax),
- keep track of which state corresponds to the max at each step
- build the answer using these back pointers

\[ \pi: \text{Initial probabilities} \]
\[ A: \text{Transitions} \]
\[ B: \text{Emissions} \]
Forward-Backward Algorithm
In addition to finding the best path, we may want to compute marginal probabilities of paths $P(y_i = s | x)$.
In addition to finding the best path, we may want to compute marginal probabilities of paths $P(y_i = s | x)$

$$P(y_i = s | x) = \sum_{y_1, \ldots, y_i-1, y_{i+1}, \ldots, y_n} P(y | x)$$
In addition to finding the best path, we may want to compute marginal probabilities of paths $P(y_i = s|x)$

$$P(y_i = s|x) = \sum_{y_1, \ldots, y_i-1, y_{i+1}, \ldots, y_n} P(y|x)$$

What did Viterbi compute? $P(y_{\text{max}}|x) = \max_{y_1, \ldots, y_n} P(y|x)$
In addition to finding the best path, we may want to compute marginal probabilities of paths $P(y_i = s|x)$.

$$P(y_i = s|x) = \sum_{y_1,\ldots,y_{i-1},y_{i+1},\ldots,y_n} P(y|x)$$

What did Viterbi compute? $P(y_{\text{max}}|x) = \max_{y_1,\ldots,y_n} P(y|x)$

Can compute marginals with dynamic programming as well using an algorithm called forward-backward.
Forward-Backward Algorithm
Forward-Backward Algorithm

\[ P(y_3 = 2 | x) = \]
Forward-Backward Algorithm

\[ P(y_3 = 2|\mathbf{x}) = \]

\[ \frac{\text{sum of all paths through state 2 at time 3}}{\text{sum of all paths}} \]
$P(y_3 = 2|x) =$

sum of all paths through state 2 at time 3

sum of all paths
Forward-Backward Algorithm

\[ P(y_3 = 2 \mid x) = \]
\[ \frac{\text{sum of all paths through state 2 at time 3}}{\text{sum of all paths}} \]
Forward-Backward Algorithm

\[ P(y_3 = 2|x) = \]
\[
\frac{\text{sum of all paths through state 2 at time 3}}{\text{sum of all paths}}
\]
Forward-Backward Algorithm

\[ P(y_3 = 2|\mathbf{x}) = \frac{\text{sum of all paths through state 2 at time 3}}{\text{sum of all paths}} \]
Forward-Backward Algorithm

\[ P(y_3 = 2 | x) = \frac{\text{sum of all paths through state 2 at time 3}}{\text{sum of all paths}} \]

slide credit: Dan Klein
$P(y_3 = 2 | x) =$

\[
\text{sum of all paths through state 2 at time 3} \quad \frac{\text{sum of all paths}}{
\]

- Easiest and most flexible to do one pass to compute and one to compute

slide credit: Dan Klein
Forward-Backward Algorithm
Forward-Backward Algorithm

Initial:
Initial:

\[ \alpha_1(s) = P(s)P(x_1|s) \]
Forward-Backward Algorithm

Initial:
\[ \alpha_1(s) = P(s)P(x_1|s) \]

Recurrence:
Forward-Backward Algorithm

- Initial:
  \[
  \alpha_1(s) = P(s)P(x_1|s)
  \]

- Recurrence:
  \[
  \alpha_t(s_t) = \sum_{s_{t-1}} \alpha_{t-1}(s_{t-1})P(s_t|s_{t-1})P(x_t|s_t)
  \]
Forward-Backward Algorithm

- **Initial:**
  \[ \alpha_1(s) = P(s)P(x_1|s) \]

- **Recurrence:**
  \[ \alpha_t(s_t) = \sum_{s_{t-1}} \alpha_{t-1}(s_{t-1})P(s_t|s_{t-1})P(x_t|s_t) \]

- **Same as Viterbi but summing instead of maxing!**
Forward-Backward Algorithm

- Initial:
  \[ \alpha_1(s) = P(s)P(x_1|s) \]

- Recurrence:
  \[ \alpha_t(s_t) = \sum_{s_{t-1}} \alpha_{t-1}(s_{t-1})P(s_t|s_{t-1})P(x_t|s_t) \]

- Same as Viterbi but summing instead of maxing!

- These quantities get very small!
  Store everything as log probabilities
Forward-Backward Algorithm
Forward-Backward Algorithm

- Initial:
Forward-Backward Algorithm

- Initial:
  \[ \beta_n(s) = 1 \]
Forward-Backward Algorithm

- **Initial:**
  \[ \beta_n(s) = 1 \]

- **Recurrence:**
Forward-Backward Algorithm

- **Initial:**
  \[ \beta_n(s) = 1 \]

- **Recurrence:**
  \[ \beta_t(s_t) = \sum_{s_{t+1}} \beta_{t+1}(s_{t+1}) P(s_{t+1}|s_t) P(x_{t+1}|s_{t+1}) \]
Forward-Backward Algorithm

- **Initial:**
  \[ \beta_n(s) = 1 \]

- **Recurrence:**
  \[ \beta_t(s_t) = \sum_{s_{t+1}} \beta_{t+1}(s_{t+1}) P(s_{t+1}|s_t) P(x_{t+1}|s_{t+1}) \]

- **Big differences:** count emission for the next timestep (not current one)
Forward-Backward Algorithm

\[ \alpha_1(s) = P(s)P(x_1|s) \]
\[ \alpha_t(s_t) = \sum_{s_{t-1}} \alpha_{t-1}(s_{t-1})P(s_t|s_{t-1})P(x_t|s_t) \]
\[ \beta_n(s) = 1 \]
\[ \beta_t(s_t) = \sum_{s_{t+1}} \beta_{t+1}(s_{t+1})P(s_{t+1}|s_t)P(x_{t+1}|s_{t+1}) \]

- Big differences: count emission for the next timestep (not current one)
Forward-Backward Algorithm

\[ \alpha_1(s) = P(s)P(x_1|s) \]

\[ \alpha_t(s_t) = \sum_{s_{t-1}} \alpha_{t-1}(s_{t-1})P(s_t|s_{t-1})P(x_t|s_t) \]

\[ \beta_n(s) = 1 \]

\[ \beta_t(s_t) = \sum_{s_{t+1}} \beta_{t+1}(s_{t+1})P(s_{t+1}|s_t)P(x_{t+1}|s_{t+1}) \]
Forward-Backward Algorithm

\[ \alpha_1(s) = P(s)P(x_1 \mid s) \]

\[ \alpha_t(s_t) = \sum_{s_{t-1}} \alpha_{t-1}(s_{t-1})P(s_t \mid s_{t-1})P(x_t \mid s_t) \]

\[ \beta_n(s) = 1 \]

\[ \beta_t(s_t) = \sum_{s_{t+1}} \beta_{t+1}(s_{t+1})P(s_{t+1} \mid s_t)P(x_{t+1} \mid s_{t+1}) \]

\[ P(s_3 = 2 \mid x) = \frac{\alpha_3(2)\beta_3(2)}{\sum_i \alpha_3(i)\beta_3(i)} \]
Forward-Backward Algorithm

\[ \alpha_1(s) = P(s)P(x_1|s) \]

\[ \alpha_t(s_t) = \sum_{s_{t-1}} \alpha_{t-1}(s_{t-1})P(s_t|s_{t-1})P(x_t|s_t) \]

\[ \beta_n(s) = 1 \]

\[ \beta_t(s_t) = \sum_{s_{t+1}} \beta_{t+1}(s_{t+1})P(s_{t+1}|s_t)P(x_{t+1}|s_{t+1}) \]

\[ P(s_3 = 2|x) = \frac{\alpha_3(2)\beta_3(2)}{\sum_i \alpha_3(i)\beta_3(i)} \]

What is the denominator here?
Forward-Backward Algorithm

\[
\begin{align*}
\alpha_1(s) &= P(s)P(x_1|s) \\
\alpha_t(s_t) &= \sum_{s_{t-1}} \alpha_{t-1}(s_{t-1})P(s_t|s_{t-1})P(x_t|s_t) \\
\beta_n(s) &= 1 \\
\beta_t(s_t) &= \sum_{s_{t+1}} \beta_{t+1}(s_{t+1})P(s_{t+1}|s_t)P(x_{t+1}|s_{t+1}) \\
\end{align*}
\]

\[P(s_3 = 2|x) = \frac{\alpha_3(2)\beta_3(2)}{\sum_i \alpha_3(i)\beta_3(i)} = \]

- What is the denominator here? \( P(x) \)
HMM POS Tagging
HMM POS Tagging

- Baseline: assign each word its most frequent tag: ~90% accuracy
HMM POS Tagging

- Baseline: assign each word its most frequent tag: ~90% accuracy
- Trigram HMM: ~95% accuracy / 55% on unknown words
Fed raises interest rates 0.5 percent
Fed raises interest rates 0.5 percent

- Trigram model: $y_1 = (\langle S \rangle, \text{NNP})$, $y_2 = (\text{NNP}, \text{VBZ})$, ...
Trigram Taggers

Trigram model: \( y_1 = (<S>, \text{NNP}), \ y_2 = (\text{NNP}, \text{VBZ}), \ldots \)

\[ P((\text{VBZ, NN}) \mid (\text{NNP, VBZ})) \] — more context! Noun-verb-noun S-V-O

Fed raises interest rates 0.5 percent
Trigram Taggers

Trigram model: $y_1 = (<S>, \text{NNP}), y_2 = (\text{NNP}, \text{VBZ}), \ldots$

$P((\text{VBZ}, \text{NN}) \mid (\text{NNP}, \text{VBZ}))$ — more context! Noun-verb-noun S-V-O

Tradeoff between model capacity and data size — trigrams are a “sweet spot” for POS tagging

Fed raises interest rates 0.5 percent
HMM POS Tagging

- Baseline: assign each word its most frequent tag: ~90% accuracy
- Trigram HMM: ~95% accuracy / 55% on unknown words
HMM POS Tagging

- Baseline: assign each word its most frequent tag: ~90% accuracy
- Trigram HMM: ~95% accuracy / 55% on unknown words
- TnT tagger (Brants 1998, tuned HMM): 96.2% accuracy / 86.0% on unks
HMM POS Tagging

- Baseline: assign each word its most frequent tag: \(~90\%\) accuracy
- Trigram HMM: \(~95\%\) accuracy / \(55\%\) on unknown words
- TnT tagger (Brants 1998, tuned HMM): 96.2\% accuracy / 86.0\% on unks
- State-of-the-art (BiLSTM-CRFs): 97.5\% / 89\%+
### Errors

<table>
<thead>
<tr>
<th></th>
<th>JJ</th>
<th>NN</th>
<th>NNP</th>
<th>NNPS</th>
<th>RB</th>
<th>RP</th>
<th>IN</th>
<th>VB</th>
<th>VBD</th>
<th>VBN</th>
<th>VBP</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>JJ</td>
<td>0</td>
<td>177</td>
<td>56</td>
<td>0</td>
<td>61</td>
<td>2</td>
<td>5</td>
<td>10</td>
<td>15</td>
<td>108</td>
<td>0</td>
<td>488</td>
</tr>
<tr>
<td>NN</td>
<td>244</td>
<td>0</td>
<td>103</td>
<td>0</td>
<td>12</td>
<td>1</td>
<td>1</td>
<td>29</td>
<td>5</td>
<td>6</td>
<td>19</td>
<td>525</td>
</tr>
<tr>
<td>NNP</td>
<td>107</td>
<td>106</td>
<td>0</td>
<td>132</td>
<td>5</td>
<td>0</td>
<td>7</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>427</td>
</tr>
<tr>
<td>NNPS</td>
<td>1</td>
<td>0</td>
<td>110</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>142</td>
</tr>
<tr>
<td>RB</td>
<td>72</td>
<td>21</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>16</td>
<td>138</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>295</td>
</tr>
<tr>
<td>RP</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>39</td>
<td>65</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>104</td>
</tr>
<tr>
<td>IN</td>
<td>11</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>169</td>
<td>0</td>
<td>103</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>323</td>
</tr>
<tr>
<td>VB</td>
<td>17</td>
<td>64</td>
<td>9</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>7</td>
<td>85</td>
<td>189</td>
</tr>
<tr>
<td>VBD</td>
<td>10</td>
<td>5</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>143</td>
<td>2</td>
<td>166</td>
</tr>
<tr>
<td>VBN</td>
<td>101</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>108</td>
<td>0</td>
<td>1</td>
<td>221</td>
</tr>
<tr>
<td>VBP</td>
<td>5</td>
<td>34</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>49</td>
<td>6</td>
<td>3</td>
<td>104</td>
</tr>
<tr>
<td>Total</td>
<td>626</td>
<td>536</td>
<td>348</td>
<td>144</td>
<td>317</td>
<td>122</td>
<td>279</td>
<td>102</td>
<td>140</td>
<td>269</td>
<td>108</td>
<td>3651</td>
</tr>
</tbody>
</table>

Slide credit: Dan Klein / Toutanova + Manning (2000)
# Errors

<table>
<thead>
<tr>
<th>JJ</th>
<th>NN</th>
<th>NNP</th>
<th>NNPS</th>
<th>RB</th>
<th>RP</th>
<th>IN</th>
<th>VB</th>
<th>VBD</th>
<th>VBN</th>
<th>VBP</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>177</td>
<td>56</td>
<td>0</td>
<td>61</td>
<td>2</td>
<td>5</td>
<td>10</td>
<td>15</td>
<td>108</td>
<td>0</td>
<td>488</td>
</tr>
<tr>
<td>244</td>
<td>0</td>
<td>103</td>
<td>0</td>
<td>12</td>
<td>1</td>
<td>1</td>
<td>29</td>
<td>5</td>
<td>6</td>
<td>19</td>
<td>525</td>
</tr>
<tr>
<td>107</td>
<td>106</td>
<td>0</td>
<td>132</td>
<td>5</td>
<td>0</td>
<td>7</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>427</td>
</tr>
<tr>
<td>1</td>
<td>110</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>142</td>
</tr>
<tr>
<td>72</td>
<td>21</td>
<td>7</td>
<td>0</td>
<td>16</td>
<td>138</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>295</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>39</td>
<td>0</td>
<td>65</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>104</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>169</td>
<td>103</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>323</td>
</tr>
<tr>
<td>17</td>
<td>64</td>
<td>9</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>7</td>
<td>85</td>
<td>189</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>143</td>
<td>2</td>
<td>166</td>
</tr>
<tr>
<td>101</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>108</td>
<td>0</td>
<td>1</td>
<td>221</td>
</tr>
<tr>
<td>5</td>
<td>34</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>49</td>
<td>6</td>
<td>3</td>
<td>0</td>
<td>104</td>
</tr>
<tr>
<td>Total</td>
<td>626</td>
<td>536</td>
<td>348</td>
<td>144</td>
<td>317</td>
<td>122</td>
<td>279</td>
<td>102</td>
<td>140</td>
<td>269</td>
<td>108</td>
</tr>
</tbody>
</table>

 JJ/NN  NN

official knowledge
## Errors

<table>
<thead>
<tr>
<th>JJ</th>
<th>NN</th>
<th>NNP</th>
<th>NNPS</th>
<th>RB</th>
<th>RP</th>
<th>IN</th>
<th>VB</th>
<th>VBD</th>
<th>VBN</th>
<th>VBP</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>JJ</td>
<td>0</td>
<td>177</td>
<td>56</td>
<td>0</td>
<td>61</td>
<td>2</td>
<td>5</td>
<td>10</td>
<td>15</td>
<td>108</td>
<td>0</td>
</tr>
<tr>
<td>NN</td>
<td>244</td>
<td>0</td>
<td>103</td>
<td>0</td>
<td>12</td>
<td>1</td>
<td>1</td>
<td>29</td>
<td>5</td>
<td>6</td>
<td>19</td>
</tr>
<tr>
<td>NNP</td>
<td>107</td>
<td>106</td>
<td>0</td>
<td>132</td>
<td>5</td>
<td>0</td>
<td>7</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>NNPS</td>
<td>1</td>
<td>110</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>RB</td>
<td>72</td>
<td>21</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>16</td>
<td>138</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>RP</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>39</td>
<td>0</td>
<td>65</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IN</td>
<td>11</td>
<td>0</td>
<td>1</td>
<td>169</td>
<td>0</td>
<td>103</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>VB</td>
<td>17</td>
<td>64</td>
<td>9</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>7</td>
<td>85</td>
</tr>
<tr>
<td>VBD</td>
<td>10</td>
<td>5</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>143</td>
<td>2</td>
</tr>
<tr>
<td>VBN</td>
<td>101</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>108</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>VBP</td>
<td>5</td>
<td>34</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>49</td>
<td>6</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>626</td>
<td>536</td>
<td>348</td>
<td>144</td>
<td>317</td>
<td>122</td>
<td>279</td>
<td>102</td>
<td>140</td>
<td>269</td>
<td>108</td>
</tr>
</tbody>
</table>

**JJ/NN**

**NN**

**official knowledge**

**(NN NN: tax cut, art gallery, ...)**

---

Slide credit: Dan Klein / Toutanova + Manning (2000)
Errors

<table>
<thead>
<tr>
<th>JJ</th>
<th>NN</th>
<th>NNP</th>
<th>NNPS</th>
<th>RB</th>
<th>RP</th>
<th>IN</th>
<th>VB</th>
<th>VBD</th>
<th>VBN</th>
<th>VBP</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>JJ</td>
<td>0</td>
<td>177</td>
<td>56</td>
<td>0</td>
<td>61</td>
<td>2</td>
<td>5</td>
<td>10</td>
<td>15</td>
<td>108</td>
<td>0</td>
</tr>
<tr>
<td>NN</td>
<td>244</td>
<td>0</td>
<td>103</td>
<td>0</td>
<td>12</td>
<td>1</td>
<td>1</td>
<td>29</td>
<td>5</td>
<td>6</td>
<td>19</td>
</tr>
<tr>
<td>NNP</td>
<td>107</td>
<td>106</td>
<td>0</td>
<td>132</td>
<td>5</td>
<td>0</td>
<td>7</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>NNPS</td>
<td>1</td>
<td>0</td>
<td>110</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>RB</td>
<td>72</td>
<td>21</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>16</td>
<td>138</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>RP</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>39</td>
<td>0</td>
<td>65</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IN</td>
<td>11</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>169</td>
<td>103</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>VB</td>
<td>17</td>
<td>64</td>
<td>9</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>7</td>
<td>85</td>
</tr>
<tr>
<td>VBD</td>
<td>10</td>
<td>5</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>143</td>
<td>2</td>
</tr>
<tr>
<td>VBN</td>
<td>101</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>108</td>
<td>0</td>
<td>1</td>
<td>221</td>
</tr>
<tr>
<td>VBP</td>
<td>5</td>
<td>34</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>49</td>
<td>6</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>626</td>
<td>536</td>
<td>348</td>
<td>144</td>
<td>317</td>
<td>122</td>
<td>279</td>
<td>102</td>
<td>140</td>
<td>269</td>
<td>108</td>
</tr>
</tbody>
</table>

 JJ/NN   NN   VBD RP/IN DT NN
official knowledge made up the story

(NN NN: tax cut, art gallery, ...)
Errors

<table>
<thead>
<tr>
<th>JJ</th>
<th>NN</th>
<th>NNP</th>
<th>NNPS</th>
<th>RB</th>
<th>RP</th>
<th>IN</th>
<th>VB</th>
<th>VBD</th>
<th>VBN</th>
<th>VBP</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>177</td>
<td>56</td>
<td>0</td>
<td>61</td>
<td>2</td>
<td>5</td>
<td>10</td>
<td>15</td>
<td>108</td>
<td>0</td>
<td>488</td>
</tr>
<tr>
<td>244</td>
<td>0</td>
<td>103</td>
<td>0</td>
<td>12</td>
<td>1</td>
<td>1</td>
<td>29</td>
<td>5</td>
<td>6</td>
<td>19</td>
<td>525</td>
</tr>
<tr>
<td>107</td>
<td>106</td>
<td>0</td>
<td>132</td>
<td>5</td>
<td>0</td>
<td>7</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>427</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>110</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>142</td>
</tr>
<tr>
<td>72</td>
<td>21</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>16</td>
<td>138</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>295</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>39</td>
<td>0</td>
<td>65</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>104</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>169</td>
<td>103</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>323</td>
</tr>
<tr>
<td>17</td>
<td>64</td>
<td>9</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>7</td>
<td>85</td>
<td>189</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>143</td>
<td>2</td>
<td>166</td>
</tr>
<tr>
<td>101</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>108</td>
<td>0</td>
<td>1</td>
<td>221</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>34</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>49</td>
<td>6</td>
<td>3</td>
<td>0</td>
<td>104</td>
</tr>
<tr>
<td>626</td>
<td>536</td>
<td>348</td>
<td>144</td>
<td>317</td>
<td>122</td>
<td>279</td>
<td>102</td>
<td>140</td>
<td>269</td>
<td>108</td>
<td>3651</td>
</tr>
</tbody>
</table>

Slide credit: Dan Klein / Toutanova + Manning (2000)

- JJ/NN: official knowledge
- NN: tax cut, art gallery, ...
- VBD: made up
- RP/IN: the story
- DT: recently
- NN: sold
- VBP: shares

(NN NN: tax cut, art gallery, ...)
Remaining Errors

Manning 2011 “Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics?”
Remaining Errors

- Lexicon gap (word not seen with that tag in training) 4.5%

Manning 2011 “Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics?”
Remaining Errors

- Lexicon gap (word not seen with that tag in training) 4.5%
- Unknown word: 4.5%

Manning 2011 “Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics?”
Remaining Errors

- Lexicon gap (word not seen with that tag in training) 4.5%
- Unknown word: 4.5%
- Could get right: 16% (many of these involve parsing!)

Manning 2011 “Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics?”
Remaining Errors

- Lexicon gap (word not seen with that tag in training) 4.5%
- Unknown word: 4.5%
- Could get right: 16% (many of these involve parsing!)
- Difficult linguistics: 20%

Manning 2011 “Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics?”
Remaining Errors

- Lexicon gap (word not seen with that tag in training) 4.5%
- Unknown word: 4.5%
- Could get right: 16% (many of these involve parsing!)
- Difficult linguistics: 20%

VBD / VBP? (past or present?)

They set up absurd situations, detached from reality

Manning 2011 “Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics?”
Remaining Errors

- Lexicon gap (word not seen with that tag in training) 4.5%
- Unknown word: 4.5%
- Could get right: 16% (many of these involve parsing!)
- Difficult linguistics: 20%

  VBD / VBP? (past or present?)
  
  They set up absurd situations, detached from reality

- Underspecified / unclear, gold standard inconsistent / wrong: 58%

Manning 2011 “Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics?”
Remaining Errors

- Lexicon gap (word not seen with that tag in training) 4.5%
- Unknown word: 4.5%
- Could get right: 16% (many of these involve parsing!)
- Difficult linguistics: 20%

  VBD / VBP? (past or present?)
  
  They set up absurd situations, detached from reality

- Underspecified / unclear, gold standard inconsistent / wrong: 58%

  adjective or verbal participle? JJ / VBN?

  a $10 million fourth-quarter charge against discontinued operations

Manning 2011 “Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics?”
## Other Languages

<table>
<thead>
<tr>
<th>Language</th>
<th>Source</th>
<th># Tags</th>
<th>O/O</th>
<th>U/U</th>
<th>O/U</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>PADT/CoNLL07 (Hajič et al., 2004)</td>
<td>21</td>
<td>96.1</td>
<td>96.9</td>
<td>97.0</td>
</tr>
<tr>
<td>Basque</td>
<td>Basque3LB/CoNLL07 (Aduriz et al., 2003)</td>
<td>64</td>
<td>89.3</td>
<td>93.7</td>
<td>93.7</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>BTB/CoNLL06 (Simov et al., 2002)</td>
<td>54</td>
<td>95.7</td>
<td>97.5</td>
<td>97.8</td>
</tr>
<tr>
<td>Catalan</td>
<td>CESS-ECE/CoNLL07 (Martí et al., 2007)</td>
<td>54</td>
<td>98.5</td>
<td>98.2</td>
<td>98.8</td>
</tr>
<tr>
<td>Chinese</td>
<td>Penn ChineseTreebank 6.0 (Palmer et al., 2007)</td>
<td>34</td>
<td>91.7</td>
<td>93.4</td>
<td>94.1</td>
</tr>
<tr>
<td>Chinese</td>
<td>Sinica/CoNLL07 (Chen et al., 2003)</td>
<td>294</td>
<td>87.5</td>
<td>91.8</td>
<td>92.6</td>
</tr>
<tr>
<td>Czech</td>
<td>PDT/CoNLL07 (Böhmová et al., 2003)</td>
<td>63</td>
<td>99.1</td>
<td>99.1</td>
<td>99.1</td>
</tr>
<tr>
<td>Danish</td>
<td>DDT/CoNLL06 (Kromann et al., 2003)</td>
<td>25</td>
<td>96.2</td>
<td>96.4</td>
<td>96.9</td>
</tr>
<tr>
<td>Dutch</td>
<td>Alpino/CoNLL06 (Van der Beek et al., 2002)</td>
<td>12</td>
<td>93.0</td>
<td>95.0</td>
<td>95.0</td>
</tr>
<tr>
<td>English</td>
<td>PennTreebank (Marcus et al., 1993)</td>
<td>45</td>
<td>96.7</td>
<td>96.8</td>
<td>97.7</td>
</tr>
<tr>
<td>French</td>
<td>FrenchTreebank (Abeillé et al., 2003)</td>
<td>30</td>
<td>96.6</td>
<td>96.7</td>
<td>97.3</td>
</tr>
<tr>
<td>German</td>
<td>Tiger/CoNLL06 (Brants et al., 2002)</td>
<td>54</td>
<td>97.9</td>
<td>98.1</td>
<td>98.8</td>
</tr>
<tr>
<td>German</td>
<td>Negra (Skut et al., 1997)</td>
<td>54</td>
<td>96.9</td>
<td>97.9</td>
<td>98.6</td>
</tr>
<tr>
<td>Greek</td>
<td>GDT/CoNLL07 (Prokopidis et al., 2005)</td>
<td>38</td>
<td>97.2</td>
<td>97.5</td>
<td>97.8</td>
</tr>
<tr>
<td>Hungarian</td>
<td>Szeged/CoNLL07 (Csendes et al., 2005)</td>
<td>43</td>
<td>94.5</td>
<td>95.6</td>
<td>95.8</td>
</tr>
<tr>
<td>Italian</td>
<td>ISST/CoNLL07 (Montemagni et al., 2003)</td>
<td>28</td>
<td>94.9</td>
<td>95.8</td>
<td>95.8</td>
</tr>
<tr>
<td>Japanese</td>
<td>Verbmobil/CoNLL06 (Kawata and Bartels, 2000)</td>
<td>80</td>
<td>98.3</td>
<td>98.0</td>
<td>99.1</td>
</tr>
<tr>
<td>Japanese</td>
<td>Kyoto4.0 (Kurohashi and Nagao, 1997)</td>
<td>42</td>
<td>97.4</td>
<td>98.7</td>
<td>99.3</td>
</tr>
<tr>
<td>Korean</td>
<td>Sejong (<a href="http://www.sejong.or.kr">http://www.sejong.or.kr</a>)</td>
<td>187</td>
<td>96.5</td>
<td>97.5</td>
<td>98.4</td>
</tr>
<tr>
<td>Portuguese</td>
<td>Floresta Síntá(c)tica/CoNLL06 (Afonso et al., 2002)</td>
<td>22</td>
<td>96.9</td>
<td>96.8</td>
<td>97.4</td>
</tr>
<tr>
<td>Russian</td>
<td>SynTagRus-RNC (Boguslavsky et al., 2002)</td>
<td>11</td>
<td>96.8</td>
<td>96.8</td>
<td>96.8</td>
</tr>
<tr>
<td>Slovene</td>
<td>SDT/CoNLL06 (Džeroski et al., 2006)</td>
<td>29</td>
<td>94.7</td>
<td>94.6</td>
<td>95.3</td>
</tr>
<tr>
<td>Spanish</td>
<td>Ancora-Cast3LB/CoNLL06 (Civit and Martí, 2004)</td>
<td>47</td>
<td>96.3</td>
<td>96.3</td>
<td>96.9</td>
</tr>
<tr>
<td>Swedish</td>
<td>Talbanken05/CoNLL06 (Nivre et al., 2006)</td>
<td>41</td>
<td>93.6</td>
<td>94.7</td>
<td>95.1</td>
</tr>
<tr>
<td>Turkish</td>
<td>METU-Sabanci/CoNLL07 (Oflazer et al., 2003)</td>
<td>31</td>
<td>87.5</td>
<td>89.1</td>
<td>90.2</td>
</tr>
</tbody>
</table>

Petrov et al. 2012
## Other Languages

<table>
<thead>
<tr>
<th>Language</th>
<th>CRF+</th>
<th>CRF</th>
<th>BTS</th>
<th>BTS*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bulgarian</td>
<td>97.97</td>
<td>97.00</td>
<td>97.84</td>
<td>97.02</td>
</tr>
<tr>
<td>Czech</td>
<td>98.38</td>
<td>98.00</td>
<td>98.50</td>
<td>98.44</td>
</tr>
<tr>
<td>Danish</td>
<td>95.93</td>
<td>95.06</td>
<td>95.52</td>
<td>92.45</td>
</tr>
<tr>
<td>German</td>
<td>93.08</td>
<td>91.99</td>
<td>92.87</td>
<td>92.34</td>
</tr>
<tr>
<td>Greek</td>
<td>97.72</td>
<td>97.21</td>
<td>97.39</td>
<td>96.64</td>
</tr>
<tr>
<td>English</td>
<td>95.11</td>
<td>94.51</td>
<td>93.87</td>
<td>94.00</td>
</tr>
<tr>
<td>Spanish</td>
<td>96.08</td>
<td>95.03</td>
<td>95.80</td>
<td>95.26</td>
</tr>
<tr>
<td>Farsi</td>
<td>96.59</td>
<td>96.25</td>
<td>96.82</td>
<td>96.76</td>
</tr>
<tr>
<td>Finnish</td>
<td>94.34</td>
<td>92.82</td>
<td>95.48</td>
<td>96.05</td>
</tr>
<tr>
<td>French</td>
<td>96.00</td>
<td>95.93</td>
<td>95.75</td>
<td>95.17</td>
</tr>
<tr>
<td>Indonesian</td>
<td>92.84</td>
<td>92.71</td>
<td>92.85</td>
<td>91.03</td>
</tr>
<tr>
<td>Italian</td>
<td>97.70</td>
<td>97.61</td>
<td>97.56</td>
<td>97.40</td>
</tr>
<tr>
<td>Swedish</td>
<td>96.81</td>
<td>96.15</td>
<td>95.57</td>
<td>93.17</td>
</tr>
<tr>
<td><strong>AVERAGE</strong></td>
<td><strong>96.04</strong></td>
<td><strong>95.41</strong></td>
<td><strong>95.85</strong></td>
<td><strong>95.06</strong></td>
</tr>
</tbody>
</table>

---

**Byte-to-Span**

\[ \text{Óscar Romero was born in El Salvador.} \]

---

*Gillick et al. 2016*
Other Languages

<table>
<thead>
<tr>
<th>Language</th>
<th>CRF+</th>
<th>CRF</th>
<th>BTS</th>
<th>BTS*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bulgarian</td>
<td>97.97</td>
<td>97.00</td>
<td>97.84</td>
<td>97.02</td>
</tr>
<tr>
<td>Czech</td>
<td>98.38</td>
<td>98.00</td>
<td>98.50</td>
<td>98.44</td>
</tr>
<tr>
<td>Danish</td>
<td>95.93</td>
<td>95.06</td>
<td>95.52</td>
<td>92.45</td>
</tr>
<tr>
<td>German</td>
<td>93.08</td>
<td>91.99</td>
<td>92.87</td>
<td>92.34</td>
</tr>
<tr>
<td>Greek</td>
<td>97.72</td>
<td>97.21</td>
<td>97.39</td>
<td>96.64</td>
</tr>
<tr>
<td>English</td>
<td>95.11</td>
<td>94.51</td>
<td>93.87</td>
<td>94.00</td>
</tr>
<tr>
<td>Spanish</td>
<td>96.08</td>
<td>95.03</td>
<td>95.80</td>
<td>95.26</td>
</tr>
<tr>
<td>Farsi</td>
<td>96.59</td>
<td>96.25</td>
<td>96.82</td>
<td>96.76</td>
</tr>
<tr>
<td>Finnish</td>
<td>94.34</td>
<td>92.82</td>
<td>95.48</td>
<td>96.05</td>
</tr>
<tr>
<td>French</td>
<td>96.00</td>
<td>95.93</td>
<td>95.75</td>
<td>95.17</td>
</tr>
<tr>
<td>Indonesian</td>
<td>92.84</td>
<td>92.71</td>
<td>92.85</td>
<td>91.03</td>
</tr>
<tr>
<td>Italian</td>
<td>97.70</td>
<td>97.61</td>
<td>97.56</td>
<td>97.40</td>
</tr>
<tr>
<td>Swedish</td>
<td>96.81</td>
<td>96.15</td>
<td>95.57</td>
<td>93.17</td>
</tr>
<tr>
<td><strong>AVERAGE</strong></td>
<td><strong>96.04</strong></td>
<td><strong>95.41</strong></td>
<td><strong>95.85</strong></td>
<td><strong>95.06</strong></td>
</tr>
</tbody>
</table>

- Universal POS tagset (~12 tags), cross-lingual model works as well as tuned CRF using external resources

Gillick et al. 2016

Byte-to-Span

- Óscar Romero was born in El Salvador.

![Byte-to-Span Diagram](image)
Next Time
Next Time

- CRFs: feature-based discriminative models
Next Time

- CRFs: feature-based discriminative models
- Structured SVM for sequences
CRFs: feature-based discriminative models

Structured SVM for sequences

Named entity recognition