Sequence Models I

Wei Xu

(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
Language is tree-structured

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

- Language is sequentially structured: interpreted in an online way

Tanenhaus et al. (1995)
POS Tagging

- 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
- Auxiliary: can, had
- Prepositions: to with
- Particles: off up

Closed class (functional)

- Determiners: the, some
- Conjunctions: and, or
- Pronouns: he, its

Slide credit: Dan Klein
<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>JJ$</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>Motown 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 maddeeningly 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 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>
### POS Tagging

<table>
<thead>
<tr>
<th>Fed raises interest rates 0.5 percent</th>
<th>Fed raises interest rates 0.5 percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>VBD VBN VBP VBZ NNP NNS NN</td>
<td>VBD VBN VBZ VBP VBZ NNP NNS CD NN</td>
</tr>
</tbody>
</table>

**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.
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

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

- POS tagging: $\mathbf{x}$ is a sequence of words, $\mathbf{y}$ is a sequence of tags

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

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

- Model the sequence of $y$ as a Markov process

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

Fed raises percent
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
- \(P(x|y)\) is a distribution over all words in the vocabulary — not a distribution over features (but could be!)
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
- \( P(y_3 = \text{NN}|y_2 = \text{VBZ}) \) direct object follows verb, other verb rarely follows past tense verb (main verbs can follow modals though!)

**Tagging Examples:**
- VBD
- VB
- VBN
- VBZ
- VBP
- VB
- NNP
- NNS
- NN
- NNS
- CD
- NN

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

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

\[\hat{P} = \text{empirical distribution (read off from data)}\]
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?
Estimating Emissions

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

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

- 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
Viterbi Algorithm

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

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

Transition probabilities  Emission probabilities  Initial probability

slide credit: Vivek Srikumar
Viterbi Algorithm

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

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

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

The only terms that depend on \(y_1\)

slide credit: Vivek Srikumar
Viterbi Algorithm

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

\[
\max_{y_1, y_2, \cdots, y_n} P(y_n|y_{n-1})P(x_n|y_n) \cdots P(y_2|y_1)P(x_2|y_2)P(y_1)P(x_1|y_1) \\
= \max_{y_1, y_2, \cdots, y_n} P(y_n|y_{n-1})P(x_n|y_n) \cdots \max_{y_1} P(y_2|y_1)P(x_2|y_2)P(y_1)P(x_1|y_1) \\
= \max_{y_1, y_2, \cdots, y_n} P(y_n|y_{n-1})P(x_n|y_n) \cdots \max_{y_1} P(y_2|y_1)P(x_2|y_2)\text{score}_1(y_1)
\]

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

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

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

slide credit: Vivek Srikumar
Viterbi Algorithm

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

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

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

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

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

Only terms that depend on \(y_2\)

slide credit: Vivek Srikumar
Viterbi Algorithm

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

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

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

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

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

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

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

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

slide credit: Vivek Srikumar
Viterbi Algorithm

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

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

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

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

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

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

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

\[
\vdots
\]

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

Abstract away the score for all decisions till here into \text{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
   \[
   \text{score}_i(s) = \max_{y_{i-1}} P(s|y_{i-1})P(x_i|s)\text{score}_{i-1}(y_{i-1})
   \]
   \[
   = \max_{y_{i-1}} A_{y_{i-1},s} B_{s,x_i}\text{score}_{i-1}(y_{i-1})
   \]
   \[
   \Pi: \text{Initial probabilities}
   \]
   \[
   A: \text{Transitions}
   \]
   \[
   B: \text{Emissions}
   \]

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 \textit{(argmax)},

- keep track of which state corresponds to the max at each step
- build the answer using these back pointers

slide credit: Vivek Srikumar
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

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

\[
\frac{\text{sum of all paths through state 2 at time 3}}{\text{sum of all paths}}
\]
Forward-Backward Algorithm

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

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

- **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

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

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

\[ \beta_n(s) = 1 \]

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

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

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

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

Slide credit: Dan Klein
Trigram Taggers

- **NNP**  **VBZ**  **NN**  **NNS**  **CD**  **NN**
  Fed raises interest rates 0.5 percent

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

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

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

Slide credit: Dan Klein
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>39</td>
<td>0</td>
<td>65</td>
<td>0</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>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>

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

NN NN: tax cut, art gallery, ...

official knowledge made up the story recently sold shares
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?”
<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>FrencheTreebank (Abeilé 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 Sintáctica/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-Sabancı/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
Next Time

- CRFs: feature-based discriminative models
- Structured SVM for sequences
- Named entity recognition