Natural Language Processing (CSEP 517): Sequence Models

Noah Smith

University of Washington
nasmith@cs.washington.edu

April 17, 2017

To-Do List

- Online quiz: due Sunday
- Read: Collins (2011), which has somewhat different notation; Jurafsky and Martin (2016a,b,c)
- A2 due April 23 (Sunday)

Linguistic Analysis: Overview

Every linguistic analyzer is comprised of:

1. Theoretical motivation from linguistics and/or the text domain
2. An algorithm that maps $\mathcal{V}^\dagger$ to some output space $\mathcal{Y}$.
3. An implementation of the algorithm
   - Once upon a time: rule systems and crafted rules
   - Most common now: supervised learning from annotated data
   - Frontier: less supervision (semi-, un-, reinforcement, distant, ...)

Sequence Labeling

After text classification ($\mathcal{V}^\dagger \rightarrow \mathcal{L}$), the next simplest type of output is a **sequence labeling**.

$$\langle x_1, x_2, \ldots, x_\ell \rangle \mapsto \langle y_1, y_2, \ldots, y_\ell \rangle$$
$$\boldsymbol{x} \mapsto \boldsymbol{y}$$

Every word gets a label in $\mathcal{L}$.
Example problems:

- part-of-speech tagging (Church, 1988)
- spelling correction (Kernighan et al., 1990)
- word alignment (Vogel et al., 1996)
- named-entity recognition (Bikel et al., 1999)
- compression (Conroy and O'Leary, 2001)

The Simplest Sequence Labeler: "Local" Classifier

Define features of a labeled word in context: $\boldsymbol{\phi}(\boldsymbol{x}, i, y)$.

Train a classifier, e.g.,

$$\begin{aligned}\hat{y}_i &= \operatorname*{argmax}_{y \in \mathcal{L}} s(\boldsymbol{x}, i, y) \\ &\stackrel{\text{linear}}{=} \operatorname*{argmax}_{y \in \mathcal{L}} \mathbf{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}, i, y)\end{aligned}$$

Decide the label for each word independently.

The Simplest Sequence Labeler: "Local" Classifier

Define features of a labeled word in context: $\boldsymbol{\phi}(\boldsymbol{x}, i, y)$.

Train a classifier, e.g.,

$$\hat{y}_i = \operatorname*{argmax}_{y \in \mathcal{L}} s(\boldsymbol{x}, i, y)$$
$$\overset{\text{linear}}{=} \operatorname*{argmax}_{y \in \mathcal{L}} \mathbf{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}, i, y)$$

Decide the label for each word independently.

Sometimes this works!

The Simplest Sequence Labeler: "Local" Classifier

Define features of a labeled word in context: $\boldsymbol{\phi}(\boldsymbol{x}, i, y)$.

Train a classifier, e.g.,

$$\hat{y}_i = \operatorname*{argmax}_{y \in \mathcal{L}} s(\boldsymbol{x}, i, y)$$
$$\stackrel{\text{linear}}{=} \operatorname*{argmax}_{y \in \mathcal{L}} \mathbf{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}, i, y)$$

Decide the label for each word independently.

Sometimes this works!

We can do better when there are predictable relationships between Y_i and Y_{i+1}.

Generative Sequence Labeling: Hidden Markov Models

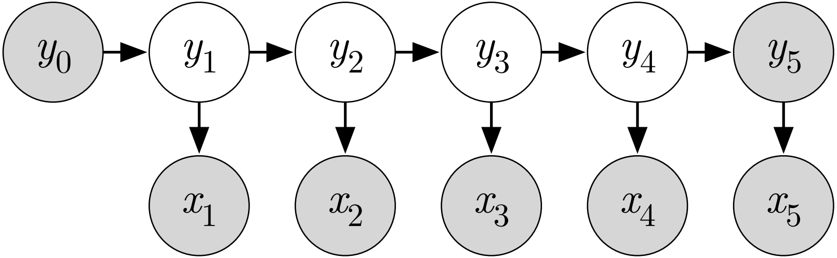
$$p(\boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1}^{\ell+1} p(x_i \mid y_i) \cdot p(y_i \mid y_{i-1})$$

For each state/label $y \in \mathcal{L}$:

- $p(X_i \mid Y_i = y)$ is the "emission" distribution for y
- $p(Y_i \mid Y_{i-1} = y)$ is called the "transition" distribution for y

Assume Y_0 is always a start state and $Y_{\ell+1}$ is always a stop state; $x_{\ell+1}$ is always the stop symbol.

Graphical Representation of Hidden Markov Models



Note: handling of beginning and end of sequence is a bit different than before. Last x is known since $p($🛑 $\mid$ 🛑$) = 1$.

Structured vs. Not

Each of these has an advantage over the other:

- The HMM lets the different labels "interact."
- The local classifier makes all of $\boldsymbol{x}$ available for every decision.

Prediction with HMMs

The classical HMM tells us to choose:

$$\operatorname*{argmax}_{\boldsymbol{y} \in \mathcal{L}^{\ell+1}} \prod_{i=1}^{\ell+1} p(x_i, \mid y_i) \cdot p(y_i \mid y_{i-1})$$

How to optimize over $|\mathcal{L}|^{\ell}$ choices without explicit enumeration?

Prediction with HMMs

The classical HMM tells us to choose:

$$\operatorname*{argmax}_{\boldsymbol{y} \in \mathcal{L}^{\ell+1}} \prod_{i=1}^{\ell+1} p(x_i, \mid y_i) \cdot p(y_i \mid y_{i-1})$$

How to optimize over $|\mathcal{L}|^\ell$ choices without explicit enumeration?

Key: exploit the conditional independence assumptions:

$$Y_i \perp \boldsymbol{Y}_{1:i-2} \mid Y_{i-1}$$
$$Y_i \perp \boldsymbol{Y}_{i+2:\ell} \mid Y_{i+1}$$

Part-of-Speech Tagging Example

	I	suspect	the	present	forecast	is	pessimistic	.
noun	•	•	•	•	•	•		
adj.		•		•	•		•	
adv.				•				
verb		•		•	•	•		
num.	•							
det.			•					
punc.								•

With this very simple tag set, $7^8 = 5.7$ million labelings.
(Even restricting to the possibilities above, 288 labelings.)

Two Obvious Solutions

Brute force: Enumerate all solutions, score them, pick the best.

Greedy: Pick each $\hat{y}_i$ according to:

$$\hat{y}_i = \operatorname*{argmax}_{y \in \mathcal{L}} p(y \mid \hat{y}_{i-1}) \cdot p(x_i \mid y)$$

What's wrong with these?

Two Obvious Solutions

Brute force: Enumerate all solutions, score them, pick the best.

Greedy: Pick each $\hat{y}_i$ according to:

$$\hat{y}_i = \operatorname*{argmax}_{y \in \mathcal{L}} p(y \mid \hat{y}_{i-1}) \cdot p(x_i \mid y)$$

What's wrong with these?

Consider:
"the old dog the footsteps of the young" (credit: Julia Hirschberg)
"the horse raced past the barn fell"

Conditional Independence

We can get an exact solution in polynomial time!

$$Y_i \perp \boldsymbol{Y}_{1:i-2} \mid Y_{i-1}$$
$$Y_i \perp \boldsymbol{Y}_{i+2:\ell} \mid Y_{i+1}$$

Given the adjacent labels to Y_i, others do not matter.

Let's start at the last position, ℓ . . .

High-Level View of Viterbi

- The decision about Y_ℓ is a function of $y_{\ell-1}$, x_ℓ, and nothing else!

$$
\begin{aligned}
p(Y_\ell = y \mid \boldsymbol{x}, \boldsymbol{y}_{1:(\ell-1)}) &= p\left(Y_\ell = y \,\middle|\, \begin{array}{l} X_\ell = x_\ell, \\ Y_{\ell-1} = y_{\ell-1}, \\ Y_{\ell+1} = {\color{red}\bigcirc} \end{array} \right) \\
&= \frac{p(Y_\ell = y, X_\ell = x_\ell, Y_{\ell-1} = y_{\ell-1}, Y_{\ell+1} = {\color{red}\bigcirc})}{p(X_\ell = x_\ell, Y_{\ell-1} = y_{\ell-1}, Y_{\ell+1} = {\color{red}\bigcirc})} \\
&\propto p({\color{red}\bigcirc} \mid y) \cdot p(x_\ell \mid y) \cdot p(y \mid y_{\ell-1})
\end{aligned}
$$

High-Level View of Viterbi

- The decision about Y_ℓ is a function of $y_{\ell-1}$, x_ℓ, and nothing else!

$$
\begin{aligned}
p(Y_\ell = y \mid \boldsymbol{x}, \boldsymbol{y}_{1:(\ell-1)}) &= p\left(Y_\ell = y \;\middle|\; \begin{array}{l} X_\ell = x_\ell, \\ Y_{\ell-1} = y_{\ell-1}, \\ Y_{\ell+1} = \bigcirc \end{array} \right) \\
&= \frac{p(Y_\ell = y, X_\ell = x_\ell, Y_{\ell-1} = y_{\ell-1}, Y_{\ell+1} = \bigcirc)}{p(X_\ell = x_\ell, Y_{\ell-1} = y_{\ell-1}, Y_{\ell+1} = \bigcirc)} \\
&\propto p(\bigcirc \mid y) \cdot p(x_\ell \mid y) \cdot p(y \mid y_{\ell-1})
\end{aligned}
$$

- If, for each value of $y_{\ell-1}$, we knew the best $\boldsymbol{y}_{1:(\ell-1)}$, then picking y_ℓ would be easy.

High-Level View of Viterbi

- The decision about Y_ℓ is a function of $y_{\ell-1}$, x_ℓ, and nothing else!

$$
\begin{aligned}
p(Y_\ell = y \mid \boldsymbol{x}, \boldsymbol{y}_{1:(\ell-1)}) &= p\left(Y_\ell = y \;\middle|\; \begin{array}{l} X_\ell = x_\ell, \\ Y_{\ell-1} = y_{\ell-1}, \\ Y_{\ell+1} = \bigcirc \end{array} \right) \\
&= \frac{p(Y_\ell = y, X_\ell = x_\ell, Y_{\ell-1} = y_{\ell-1}, Y_{\ell+1} = \bigcirc)}{p(X_\ell = x_\ell, Y_{\ell-1} = y_{\ell-1}, Y_{\ell+1} = \bigcirc)} \\
&\propto p(\bigcirc \mid y) \cdot p(x_\ell \mid y) \cdot p(y \mid y_{\ell-1})
\end{aligned}
$$

- If, for each value of $y_{\ell-1}$, we knew the best $\boldsymbol{y}_{1:(\ell-1)}$, then picking y_ℓ would be easy.
- Idea: for each position i, calculate the score of the best label prefix $\boldsymbol{y}_{1:i}$ ending in each possible value for Y_i.

High-Level View of Viterbi

- The decision about Y_ℓ is a function of $y_{\ell-1}$, x_ℓ, and nothing else!

$$
\begin{aligned}
p(Y_\ell = y \mid \boldsymbol{x}, \boldsymbol{y}_{1:(\ell-1)}) &= p\left(Y_\ell = y \;\middle|\; \begin{array}{l} X_\ell = x_\ell, \\ Y_{\ell-1} = y_{\ell-1}, \\ Y_{\ell+1} = \bigcirc \end{array} \right) \\
&= \frac{p(Y_\ell = y, X_\ell = x_\ell, Y_{\ell-1} = y_{\ell-1}, Y_{\ell+1} = \bigcirc)}{p(X_\ell = x_\ell, Y_{\ell-1} = y_{\ell-1}, Y_{\ell+1} = \bigcirc)} \\
&\propto p(\bigcirc \mid y) \cdot p(x_\ell \mid y) \cdot p(y \mid y_{\ell-1})
\end{aligned}
$$

- If, for each value of $y_{\ell-1}$, we knew the best $\boldsymbol{y}_{1:(\ell-1)}$, then picking y_ℓ would be easy.
- Idea: for each position i, calculate the score of the best label prefix $\boldsymbol{y}_{1:i}$ ending in each possible value for Y_i.
- With a little bookkeeping, we can then trace backwards and recover the best label sequence.

Chart Data Structure

	x_1	x_2	$\ldots$	x_ℓ
y				
y'				
$\vdots$				
y^{last}				

Recurrence

First, think about the *score* of the best sequence.

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

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

Recurrence

First, think about the *score* of the best sequence.

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

$$s_\ell(y) = p(\; \mid y) \cdot p(x_\ell \mid y) \cdot \max_{y' \in \mathcal{L}} p(y \mid y') \cdot \boxed{s_{\ell-1}(y')}$$

$$s_{\ell-1}(y) = p(x_{\ell-1} \mid y) \cdot \max_{y' \in \mathcal{L}} p(y \mid y') \cdot \boxed{s_{\ell-2}(y')}$$

Recurrence

First, think about the *score* of the best sequence.

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

$$
\begin{aligned}
s_\ell(y) &= p(\;\mid y) \cdot p(x_\ell \mid y) \cdot \max_{y' \in \mathcal{L}} p(y \mid y') \cdot \boxed{s_{\ell-1}(y')} \\
s_{\ell-1}(y) &= p(x_{\ell-1} \mid y) \cdot \max_{y' \in \mathcal{L}} p(y \mid y') \cdot \boxed{s_{\ell-2}(y')} \\
s_{\ell-2}(y) &= p(x_{\ell-2} \mid y) \cdot \max_{y' \in \mathcal{L}} p(y \mid y') \cdot \boxed{s_{\ell-3}(y')}
\end{aligned}
$$

Recurrence

First, think about the *score* of the best sequence.

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

$$s_\ell(y) = p(\octagon \mid y) \cdot p(x_\ell \mid y) \cdot \max_{y' \in \mathcal{L}} p(y \mid y') \cdot \boxed{s_{\ell-1}(y')}$$

$$s_{\ell-1}(y) = p(x_{\ell-1} \mid y) \cdot \max_{y' \in \mathcal{L}} p(y \mid y') \cdot \boxed{s_{\ell-2}(y')}$$

$$s_{\ell-2}(y) = p(x_{\ell-2} \mid y) \cdot \max_{y' \in \mathcal{L}} p(y \mid y') \cdot \boxed{s_{\ell-3}(y')}$$

$$\vdots$$

$$s_i(y) = p(x_i \mid y) \cdot \max_{y' \in \mathcal{L}} p(y \mid y') \cdot \boxed{s_{i-1}(y')}$$

Recurrence

First, think about the *score* of the best sequence.

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

$$
\begin{aligned}
s_\ell(y) &= p(\;\mid y) \cdot p(x_\ell \mid y) \cdot \max_{y' \in \mathcal{L}} p(y \mid y') \cdot \boxed{s_{\ell-1}(y')} \\
s_{\ell-1}(y) &= p(x_{\ell-1} \mid y) \cdot \max_{y' \in \mathcal{L}} p(y \mid y') \cdot \boxed{s_{\ell-2}(y')} \\
s_{\ell-2}(y) &= p(x_{\ell-2} \mid y) \cdot \max_{y' \in \mathcal{L}} p(y \mid y') \cdot \boxed{s_{\ell-3}(y')} \\
&\vdots \\
s_i(y) &= p(x_i \mid y) \cdot \max_{y' \in \mathcal{L}} p(y \mid y') \cdot \boxed{s_{i-1}(y')} \\
&\vdots \\
s_1(y) &= p(x_1 \mid y) \cdot p(y \mid y_0)
\end{aligned}
$$

Viterbi Procedure (Part I: Prefix Scores)

	x_1	x_2	$\dots$	x_ℓ
y				
y'				
$\vdots$				
y^{last}				

Viterbi Procedure (Part I: Prefix Scores)

	x_1	x_2	$\ldots$	x_ℓ
y	$s_1(y)$			
y'	$s_1(y')$			
$\vdots$				
y^{last}	$s_1(y^{last})$			

$$s_1(y) = p(x_1 \mid y) \cdot p(y \mid y_0)$$

Viterbi Procedure (Part I: Prefix Scores)

	x_1	x_2	$\dots$	x_ℓ
y	$s_1(y)$	$s_2(y)$		
y'	$s_1(y')$	$s_2(y')$		
$\vdots$				
y^{last}	$s_1(y^{last})$	$s_2(y^{last})$		

$$s_i(y) = p(x_i \mid y) \cdot \max_{y' \in \mathcal{L}} p(y \mid y') \cdot \boxed{s_{i-1}(y')}$$

Viterbi Procedure (Part I: Prefix Scores)

	x_1	x_2	$\ldots$	x_ℓ
y	$s_1(y)$	$s_2(y)$		$s_\ell(y)$
y'	$s_1(y')$	$s_2(y')$		$s_\ell(y')$
$\vdots$				
y^{last}	$s_1(y^{last})$	$s_2(y^{last})$		$s_\ell(y^{last})$

$$s_\ell(y) = p(\quad \mid y) \cdot p(x_\ell \mid y) \cdot \max_{y' \in \mathcal{L}} p(y \mid y') \cdot \boxed{s_{\ell-1}(y')}$$

Claim: $\max_{y \in \mathcal{L}} s_\ell(y) = \max_{\boldsymbol{y} \in \mathcal{L}^{\ell+1}} p(\boldsymbol{x}, \boldsymbol{y})$

Claim: $\max\limits_{y \in \mathcal{L}} s_\ell(y) = \max\limits_{\boldsymbol{y} \in \mathcal{L}^{\ell+1}} p(\boldsymbol{x}, \boldsymbol{y})$

$$\max_{y \in \mathcal{L}} s_\ell(y) = \max_{y \in \mathcal{L}} p(\;\mid y) \cdot p(x_\ell \mid y) \cdot \max_{y' \in \mathcal{L}} p(y \mid y') \cdot \boxed{s_{\ell-1}(y')}$$

Claim: $\max_{y \in \mathcal{L}} s_\ell(y) = \max_{\boldsymbol{y} \in \mathcal{L}^{\ell+1}} p(\boldsymbol{x}, \boldsymbol{y})$

$$\begin{aligned}\max_{y \in \mathcal{L}} s_\ell(y) &= \max_{y \in \mathcal{L}} p(\bigcirc \mid y) \cdot p(x_\ell \mid y) \cdot \max_{y' \in \mathcal{L}} p(y \mid y') \cdot \boxed{s_{\ell-1}(y')} \\ &= \max_{y \in \mathcal{L}} p(\bigcirc \mid y) \cdot p(x_\ell \mid y) \cdot \max_{y' \in \mathcal{L}} p(y \mid y') \cdot \boxed{p(x_{\ell-1} \mid y') \cdot \max_{y'' \in \mathcal{L}} p(y' \mid y'') \cdot \boxed{s_{\ell-2}(y'')}}\end{aligned}$$

Claim: $\max\limits_{y \in \mathcal{L}} s_\ell(y) = \max\limits_{\boldsymbol{y} \in \mathcal{L}^{\ell+1}} p(\boldsymbol{x}, \boldsymbol{y})$

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

$$= \max_{y \in \mathcal{L}} p(\bigcirc \mid y) \cdot p(x_\ell \mid y) \cdot \max_{y' \in \mathcal{L}} p(y \mid y') \cdot \boxed{p(x_{\ell-1} \mid y') \cdot \max_{y'' \in \mathcal{L}} p(y' \mid y'') \cdot \boxed{s_{\ell-2}(y'')}}$$

$$= \max_{y \in \mathcal{L}} p(\bigcirc \mid y) \cdot p(x_\ell \mid y) \cdot \max_{y' \in \mathcal{L}} p(y \mid y') \cdot$$

$$\boxed{p(x_{\ell-1} \mid y') \cdot \max_{y'' \in \mathcal{L}} p(y' \mid y'') \cdot \boxed{p(x_{\ell-2} \mid y'') \cdot \max_{y''' \in \mathcal{L}} p(y'' \mid y''') \cdot \boxed{s_{\ell-3}(y''')}}}$$

Claim: $\max\limits_{y \in \mathcal{L}} s_\ell(y) = \max\limits_{\boldsymbol{y} \in \mathcal{L}^{\ell+1}} p(\boldsymbol{x}, \boldsymbol{y})$

$$
\begin{aligned}
\max_{y \in \mathcal{L}} s_\ell(y) &= \max_{y \in \mathcal{L}} p(\bigcirc \mid y) \cdot p(x_\ell \mid y) \cdot \max_{y' \in \mathcal{L}} p(y \mid y') \cdot \boxed{s_{\ell-1}(y')} \\
&= \max_{y \in \mathcal{L}} p(\bigcirc \mid y) \cdot p(x_\ell \mid y) \cdot \max_{y' \in \mathcal{L}} p(y \mid y') \cdot \boxed{p(x_{\ell-1} \mid y') \cdot \max_{y'' \in \mathcal{L}} p(y' \mid y'') \cdot \boxed{s_{\ell-2}(y'')}} \\
&= \max_{y \in \mathcal{L}} p(\bigcirc \mid y) \cdot p(x_\ell \mid y) \cdot \max_{y' \in \mathcal{L}} p(y \mid y') \cdot \\
&\quad \boxed{p(x_{\ell-1} \mid y') \cdot \max_{y'' \in \mathcal{L}} p(y' \mid y'') \cdot \boxed{p(x_{\ell-2} \mid y'') \cdot \max_{y''' \in \mathcal{L}} p(y'' \mid y''') \cdot \boxed{s_{\ell-3}(y''')}}} \\
&= \max_{\boldsymbol{y} \in \mathcal{L}^{\ell+1}} p(\bigcirc \mid y_\ell) \cdot p(x_\ell \mid y_\ell) \cdot p(y_\ell \mid y_{\ell-1}) \cdot p(x_{\ell-1} \mid y_{\ell-1}) \cdot p(y_{\ell-1} \mid y_{\ell-2}) \cdot \\
&\quad p(x_{\ell-2} \mid y_{\ell-2}) \cdots p(x_1 \mid y_1) \cdot p(y_1 \mid y_0)
\end{aligned}
$$

Claim: $\max_{y \in \mathcal{L}} s_\ell(y) = \max_{\boldsymbol{y} \in \mathcal{L}^{\ell+1}} p(\boldsymbol{x}, \boldsymbol{y})$

$$\max_{y \in \mathcal{L}} s_\ell(y) = \max_{y \in \mathcal{L}} p(\quad \mid y) \cdot p(x_\ell \mid y) \cdot \max_{y' \in \mathcal{L}} p(y \mid y') \cdot \boxed{s_{\ell-1}(y')}$$

$$= \max_{y \in \mathcal{L}} p(\quad \mid y) \cdot p(x_\ell \mid y) \cdot \max_{y' \in \mathcal{L}} p(y \mid y') \cdot \boxed{p(x_{\ell-1} \mid y') \cdot \max_{y'' \in \mathcal{L}} p(y' \mid y'') \cdot \boxed{s_{\ell-2}(y'')}}$$

$$= \max_{y \in \mathcal{L}} p(\quad \mid y) \cdot p(x_\ell \mid y) \cdot \max_{y' \in \mathcal{L}} p(y \mid y') \cdot$$

$$\boxed{p(x_{\ell-1} \mid y') \cdot \max_{y'' \in \mathcal{L}} p(y' \mid y'') \cdot \boxed{p(x_{\ell-2} \mid y'') \cdot \max_{y''' \in \mathcal{L}} p(y'' \mid y''') \cdot \boxed{s_{\ell-3}(y''')}}}$$

$$= \max_{\boldsymbol{y} \in \mathcal{L}^{\ell+1}} p(\quad \mid y_\ell) \cdot p(x_\ell \mid y_\ell) \cdot p(y_\ell \mid y_{\ell-1}) \cdot p(x_{\ell-1} \mid y_{\ell-1}) \cdot p(y_{\ell-1} \mid y_{\ell-2}) \cdot$$

$$p(x_{\ell-2} \mid y_{\ell-2}) \cdots p(x_1 \mid y_1) \cdot p(y_1 \mid y_0)$$

$$= \max_{\boldsymbol{y} \in \mathcal{L}^{\ell+1}} \prod_{i=1}^{\ell+1} p(x_i \mid y_i) \cdot p(y_i \mid y_{i-1})$$

High-Level View of Viterbi

- The decision about Y_ℓ is a function of $y_{\ell-1}$, x_ℓ, and nothing else!

$$
\begin{aligned}
p(Y_\ell = y \mid \boldsymbol{x}, \boldsymbol{y}_{1:(\ell-1)}) &= p\left(Y_\ell = y \;\middle|\; \begin{array}{l} X_\ell = x_\ell, \\ Y_{\ell-1} = y_{\ell-1}, \\ Y_{\ell+1} = \bigcirc \end{array}\right) \\
&= \frac{p(Y_\ell = y, X_\ell = x_\ell, Y_{\ell-1} = y_{\ell-1}, Y_{\ell+1} = \bigcirc)}{p(X_\ell = x_\ell, Y_{\ell-1} = y_{\ell-1}, Y_{\ell+1} = \bigcirc)} \\
&\propto p(\bigcirc \mid y) \cdot p(x_\ell \mid y) \cdot p(y \mid y_{\ell-1})
\end{aligned}
$$

- If, for each value of $y_{\ell-1}$, we knew the best $\boldsymbol{y}_{1:(\ell-1)}$, then picking y_ℓ would be easy.
- Idea: for each position i, calculate the score of the best label prefix $\boldsymbol{y}_{1:i}$ ending in each possible value for Y_i.
- With a little bookkeeping, we can then trace backwards and recover the best label sequence.

Viterbi Procedure (Part I: Prefix Scores and Backpointers)

	x_1	x_2	$\ldots$	x_ℓ
y				
y'				
$\vdots$				
y^{last}				

Viterbi Procedure (Part I: Prefix Scores and Backpointers)

	x_1	x_2	$\ldots$	x_ℓ
y	$s_1(y)$ $b_1(y)$			
y'	$s_1(y')$ $b_1(y')$			
$\vdots$				
y^{last}	$s_1(y^{last})$ $b_1(y^{last})$			

$$s_1(y) = p(x_1 \mid y) \cdot p(y \mid y_0)$$
$$b_1(y) = y_0$$

Viterbi Procedure (Part I: Prefix Scores and Backpointers)

	x_1	x_2	$\ldots$	x_ℓ
y	$s_1(y)$ $b_1(y)$	$s_2(y)$ $b_2(y)$		
y'	$s_1(y')$ $b_1(y')$	$s_2(y')$ $b_2(y')$		
$\vdots$				
y^{last}	$s_1(y^{last})$ $b_1(y^{last})$	$s_2(y^{last})$ $b_2(y^{last})$		

$$s_i(y) = p(x_i \mid y) \cdot \max_{y' \in \mathcal{L}} p(y \mid y') \cdot \boxed{s_{i-1}(y')}$$

$$b_i(y) = \operatorname*{argmax}_{y' \in \mathcal{L}} p(y \mid y') \cdot s_{i-1}(y')$$

Viterbi Procedure (Part I: Prefix Scores and Backpointers)

	x_1	x_2	$\ldots$	x_ℓ
y	$s_1(y)$ $b_1(y)$	$s_2(y)$ $b_2(y)$		$s_\ell(y)$ $b_\ell(y)$
y'	$s_1(y')$ $b_1(y')$	$s_2(y')$ $b_2(y')$		$s_\ell(y')$ $b_\ell(y')$
$\vdots$				
y^{last}	$s_1(y^{last})$ $b_1(y^{last})$	$s_2(y^{last})$ $b_2(y^{last})$		$s_\ell(y^{last})$ $b_\ell(y^{last})$

$$s_\ell(y) = p(\quad \mid y) \cdot p(x_\ell \mid y) \cdot \max_{y' \in \mathcal{L}} p(y \mid y') \cdot \boxed{s_{\ell-1}(y')}$$

$$b_\ell(y) = \operatorname*{argmax}_{y' \in \mathcal{L}} p(y \mid y') \cdot s_{\ell-1}(y')$$

Full Viterbi Procedure

Input: $\boldsymbol{x}$, $p(X_i \mid Y_i)$, $p(Y_{i+1} \mid Y_i)$

Output: $\hat{\boldsymbol{y}}$

1. For $i \in \langle 1, \ldots, \ell \rangle$:
   - Solve for $s_i(*)$ and $b_i(*)$.
     - Special base case for $i = 1$ to handle start state y_0 (no max)
     - General recurrence for $i \in \langle 2, \ldots, \ell - 1 \rangle$
     - Special case for $i = \ell$ to handle stopping probability
2. $\hat{y}_\ell \leftarrow \operatorname*{argmax}_{y \in \mathcal{L}} s_\ell(y)$
3. For $i \in \langle \ell, \ldots, 1 \rangle$:
   - $\hat{y}_{i-1} \leftarrow b(y_i)$

Viterbi Asymptotics

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

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

	x_1	x_2	$\ldots$	x_ℓ
y				
y'				
$\vdots$				
y^{last}				

Generalizing Viterbi

- Instead of HMM parameters, we can "featurize" or "neuralize."

Generalizing Viterbi

- Instead of HMM parameters, we can "featurize" or "neuralize."
  Define features of adjacent labeled words in context: $\boldsymbol{\phi}(\boldsymbol{x}, i, y, y')$
  "Structured" classifer/predictor:

$$\hat{\boldsymbol{y}} = \operatorname*{argmax}_{\boldsymbol{y} \in \mathcal{L}^{\ell+1}} \sum_{i=1}^{\ell+1} \mathbf{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}, i, y_i, y_{i-1})$$

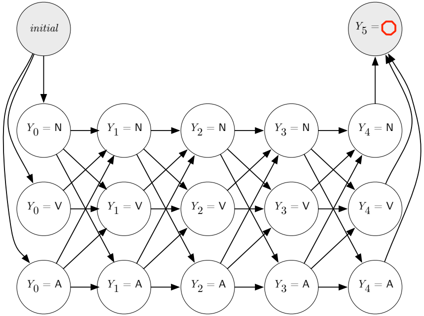
$$\stackrel{\text{HMM}}{=} \operatorname*{argmax}_{\boldsymbol{y} \in \mathcal{L}^{\ell+1}} \sum_{i=1}^{\ell+1} \log p(x_i \mid y_i) + \log p(y_i \mid y_{i-1})$$

Generalizing Viterbi

- Instead of HMM parameters, we can "featurize" or "neuralize."
- Viterbi instantiates an general algorithm called **max-product variable elimination**, for inference along a chain of variables with pairwise "links."
  HMMs are the simplest example of a **structured predictor**: a collection of classifiers whose decisions depend on each other.

Generalizing Viterbi

- Instead of HMM parameters, we can "featurize" or "neuralize."
- Viterbi instantiates an general algorithm called **max-product variable elimination**, for inference along a chain of variables with pairwise "links."
  HMMs are the simplest example of a **structured predictor**: a collection of classifiers whose decisions depend on each other.
- Viterbi solves a special case of the "best path" problem.



Generalizing Viterbi

- Instead of HMM parameters, we can "featurize" or "neuralize."
- Viterbi instantiates an general algorithm called **max-product variable elimination**, for inference along a chain of variables with pairwise "links."
  HMMs are the simplest example of a **structured predictor**: a collection of classifiers whose decisions depend on each other.
- Viterbi solves a special case of the "best path" problem.
- Higher-order dependencies among $\boldsymbol{Y}$ are also possible.

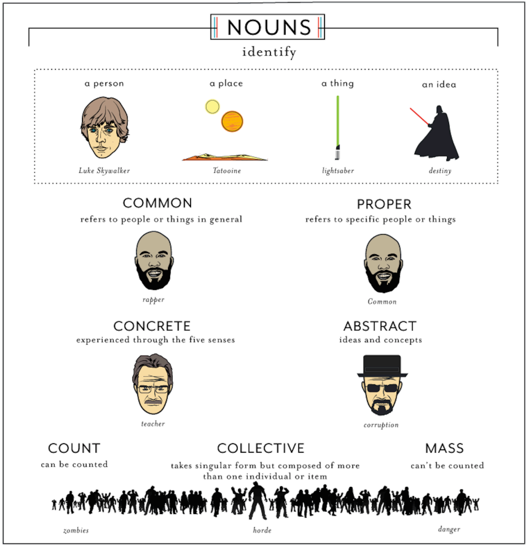
$$s_i(y, y') = \max_{y'' \in \mathcal{L}} p(x_i \mid y) \cdot p(y \mid y', y'') \cdot s_{i-1}(y', y'')$$

Applications of Sequence Models

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

Parts of Speech

http://mentalfloss.com/article/65608/master-particulars-grammar-pop-culture-primer



Parts of Speech

- "Open classes": Nouns, verbs, adjectives, adverbs, numbers
- "Closed classes":
  - Modal verbs
  - Prepositions (*on*, *to*)
  - Particles (*off*, *up*)
  - Determiners (*the*, *some*)
  - Pronouns (*she*, *they*)
  - Conjunctions (*and*, *or*)

Parts of Speech in English: Decisions

Granularity decisions regarding:

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

Some linguistic reasoning required:

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

Interactions with tokenization:

- Punctuation
- Compounds (*Mark'll*, *someone's*, *gonna*)

Penn Treebank: 45 tags, ∼40 pages of guidelines (Marcus et al., 1993)

Parts of Speech in English: Decisions

Granularity decisions regarding:

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

Some linguistic reasoning required:

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

Interactions with tokenization:

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

Penn Treebank: 45 tags, ∼40 pages of guidelines (Marcus et al., 1993)
TweetNLP: 20 tags, 7 pages of guidelines (Gimpel et al., 2011)

Example: Part-of-Speech Tagging

ikr smh he asked fir yo last name

so he can add u on fb lololol

Example: Part-of-Speech Tagging

I know, right	shake my head			for	your		
ikr	smh	he	asked	fir	yo	last	name

				you		Facebook	laugh out loud
so	he	can	add	u	on	fb	lololol

Example: Part-of-Speech Tagging

I know, right	shake my head			for	your		
ikr	smh	he	asked	fir	yo	last	name
!	G	O	V	P	D	A	N
interjection	acronym	pronoun	verb	prep.	det.	adj.	noun

				you		Facebook	laugh out loud
so	he	can	add	u	on	fb	lololol
P	O	V	V	O	P	^	!
preposition						proper noun	

Why POS?

- Text-to-speech: *record*, *lead*, *protest*
- Lemmatization: *saw*/V $\rightarrow$ *see*; *saw*/N $\rightarrow$ *saw*
- Quick-and-dirty multiword expressions: (Adjective | Noun)* Noun (Justeson and Katz, 1995)
- Preprocessing for harder disambiguation problems:
  - *The Georgia branch had taken* **on** *loan commitments ...*
  - *The average of interbank* **offered** *rates plummeted ...*

A Simple POS Tagger

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

A Simple POS Tagger

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

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

A Simple POS Tagger

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

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

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

A Simple POS Tagger

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

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

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

All datasets have some errors; estimated upper bound for Penn Treebank is 98%.

Supervised Training of Hidden Markov Models

Given: annotated sequences $\langle\langle \boldsymbol{x}_1, \boldsymbol{y}_1, \rangle, \ldots, \langle \boldsymbol{x}_n, \boldsymbol{y}_n \rangle\rangle$

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

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

- $\boldsymbol{\theta}_{*|y}$ is the "emission" distribution, estimating $p(x \mid y)$ for each $x \in \mathcal{V}$
- $\gamma_{*|y}$ is called the "transition" distribution, estimating $p(y' \mid y)$ for each $y' \in \mathcal{L}$

Supervised Training of Hidden Markov Models

Given: annotated sequences $\langle\langle \boldsymbol{x}_1, \boldsymbol{y}_1, \rangle, \ldots, \langle \boldsymbol{x}_n, \boldsymbol{y}_n \rangle\rangle$

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

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

- $\boldsymbol{\theta}_{*|y}$ is the "emission" distribution, estimating $p(x \mid y)$ for each $x \in \mathcal{V}$
- $\boldsymbol{\gamma}_{*|y}$ is called the "transition" distribution, estimating $p(y' \mid y)$ for each $y' \in \mathcal{L}$

Maximum likelihood estimate: count and normalize!

Back to POS

TnT, a trigram HMM tagger with smoothing: 96.7% (Brants, 2000)

Back to POS

TnT, a trigram HMM tagger with smoothing: 96.7% (Brants, 2000)

State of the art: ∼97.5% (Toutanova et al., 2003); uses a feature-based model with:

- capitalization features
- spelling features
- name lists ("gazetteers")
- context words
- hand-crafted patterns

Back to POS

TnT, a trigram HMM tagger with smoothing: 96.7% (Brants, 2000)

State of the art: ~97.5% (Toutanova et al., 2003); uses a feature-based model with:

- capitalization features
- spelling features
- name lists ("gazetteers")
- context words
- hand-crafted patterns

There might be very recent improvements to this.

Other Labels

Parts of speech are a minimal *syntactic* representation.

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

Supersenses

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

Supersenses

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

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

- E.g., from a dictionary

Supersenses

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

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

- E.g., from a dictionary

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

- WordNet (Fellbaum, 1998) is a fascinating resource in its own right! See `http://wordnetweb.princeton.edu/perl/webwn` to get an idea.

Supersenses

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

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

- E.g., from a dictionary

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

- WordNet (Fellbaum, 1998) is a fascinating resource in its own right! See `http://wordnetweb.princeton.edu/perl/webwn` to get an idea.

This represents a coarsening of the annotations in the Semcor corpus (Miller et al., 1993).

Example: *box*'s Thirteen Synonym Sets, Eight Supersenses

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

Example: *box*'s Thirteen Synonym Sets, Eight Supersenses

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

Supersense Tagging Example

Clara	Harris	,	one	of	the	guests	in	the
	N.PERSON					N.PERSON		

box	,	stood	up	and	demanded
N.ARTIFACT			V.MOTION		V.COMMUNICATION

water	.
N.SUBSTANCE	

Ciaramita and Altun's Approach

Features at each position in the sentence:

- word
- "first sense" from WordNet (also conjoined with word)
- POS, coarse POS
- shape (case, punctuation symbols, etc.)
- previous label

All of these fit into "$\boldsymbol{\phi}(\boldsymbol{x}, i, y, y')$."

Featurizing HMMs

Log-probability score of $\boldsymbol{y}$ (given $\boldsymbol{x}$) decomposes into a sum of local scores:

$$\mathrm{score}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{\ell+1} \overbrace{\left(\log p(x_i \mid y_i) + \log p(y_i \mid y_{i-1})\right)}^{\text{local score at position } i} \tag{1}$$

Featurized HMM:

$$\mathrm{score}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{\ell+1} \overbrace{\left(\mathbf{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}, i, y_i, y_{i-1})\right)}^{\text{local score at position } i} \tag{2}$$

$$= \mathbf{w} \cdot \underbrace{\sum_{i=1}^{\ell+1} \boldsymbol{\phi}(\boldsymbol{x}, i, y_i, y_{i-1})}_{\text{global features, } \boldsymbol{\Phi}(\boldsymbol{x}, \boldsymbol{y})} \tag{3}$$

What Changes?

Algorithmically, not much!

Viterbi recurrence before (using log math):

$$s_1(y) = \log p(x_1 \mid y) + \log p(y \mid y_0)$$

$$s_i(y) = \log p(x_i \mid y) + \max_{y' \in \mathcal{L}} \log p(y \mid y') + \boxed{s_{i-1}(y')}$$

$$s_\ell(y) = \log p(\quad \mid y) + \log p(x_\ell \mid y) + \max_{y' \in \mathcal{L}} \log p(y \mid y') + \boxed{s_{\ell-1}(y')}$$

After:

$$s_1(y) = \mathbf{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}, 1, y, y_0)$$

$$s_i(y) = \max_{y' \in \mathcal{L}} \mathbf{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}, i, y, y') + \boxed{s_{i-1}(y')}$$

$$s_\ell(y) = \max_{y' \in \mathcal{L}} \mathbf{w} \cdot \left(\boldsymbol{\phi}(\boldsymbol{x}, \ell, y, y') + \boldsymbol{\phi}(\boldsymbol{x}, \ell+1, \quad , y)\right) + \boxed{s_{\ell-1}(y')}$$

Supervised Training of Sequence Models (Discriminative)

Given: annotated sequences $\langle\langle \boldsymbol{x}_1, \boldsymbol{y}_1, \rangle, \ldots, \langle \boldsymbol{x}_n, \boldsymbol{y}_n \rangle\rangle$

Assume:

$$\begin{aligned}
\text{predict}(\boldsymbol{x}) &= \operatorname*{argmax}_{\boldsymbol{y} \in \mathcal{L}^{\ell+1}} \text{score}(\boldsymbol{x}, \boldsymbol{y}) \\
&= \operatorname*{argmax}_{\boldsymbol{y} \in \mathcal{L}^{\ell+1}} \sum_{i=1}^{\ell+1} \mathbf{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}, i, y_i, y_{i-1}) \\
&= \operatorname*{argmax}_{\boldsymbol{y} \in \mathcal{L}^{\ell+1}} \mathbf{w} \cdot \sum_{i=1}^{\ell+1} \boldsymbol{\phi}(\boldsymbol{x}, i, y_i, y_{i-1}) \\
&= \operatorname*{argmax}_{\boldsymbol{y} \in \mathcal{L}^{\ell+1}} \mathbf{w} \cdot \boldsymbol{\Phi}(\boldsymbol{x}, \boldsymbol{y})
\end{aligned}$$

Estimate: $\mathbf{w}$

Perceptron

Perceptron algorithm for **classification**:

- For $t \in \{1, \ldots, T\}$:
  - Pick i_t uniformly at random from $\{1, \ldots, n\}$.
  - $\hat{\ell}_{i_t} \leftarrow \operatorname*{argmax}_{\ell \in \mathcal{L}} \mathbf{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}_{i_t}, \ell)$
  - $\mathbf{w} \leftarrow \mathbf{w} - \alpha \left(\boldsymbol{\phi}(\boldsymbol{x}_{i_t}, \hat{\ell}_{i_t}) - \boldsymbol{\phi}(\boldsymbol{x}_{i_t}, \ell_{i_t}) \right)$

Structured Perceptron

Collins (2002)

Perceptron algorithm for ~~classification~~ **structured prediction**:

- For $t \in \{1, \ldots, T\}$:
  - Pick i_t uniformly at random from $\{1, \ldots, n\}$.
  - $\hat{\boldsymbol{y}}_{i_t} \leftarrow \operatorname*{argmax}_{\boldsymbol{y} \in \mathcal{L}^{\ell+1}} \mathbf{w} \cdot \boldsymbol{\Phi}(\boldsymbol{x}_{i_t}, \boldsymbol{y})$
  - $\mathbf{w} \leftarrow \mathbf{w} - \alpha \left(\boldsymbol{\Phi}(\boldsymbol{x}_{i_t}, \hat{\boldsymbol{y}}_{i_t}) - \boldsymbol{\Phi}(\boldsymbol{x}_{i_t}, \boldsymbol{y}_{i_t}) \right)$

This can be viewed as stochastic subgradient descent on the *structured* hinge loss:

$$\sum_{i=1}^{n} \underbrace{\max_{\boldsymbol{y} \in \mathcal{L}^{\ell_i+1}} \mathbf{w} \cdot \boldsymbol{\Phi}(\boldsymbol{x}_i, \boldsymbol{y})}_{\text{fear}} - \underbrace{\mathbf{w} \cdot \boldsymbol{\Phi}(\boldsymbol{x}_i, \boldsymbol{y}_i)}_{\text{hope}}$$

Back to Supersenses

Clara	Harris	,	one	of	the	guests	in	the
	N.PERSON					N.PERSON		

box	,	stood	up	and	demanded
N.ARTIFACT			V.MOTION		V.COMMUNICATION

water	.
N.SUBSTANCE	

Shouldn't *Clara Harris* and *stood up* be respectively "grouped"?

Segmentations

Segmentation:

- Input: $\boldsymbol{x} = \langle x_1, x_2, \ldots, x_\ell \rangle$
- Output:

$$\left\langle \begin{array}{l} \boldsymbol{x}_{1:\ell_1}, \\ \boldsymbol{x}_{(1+\ell_1):(\ell_1+\ell_2)}, \\ \boldsymbol{x}_{(1+\ell_1+\ell_2):(\ell_1+\ell_2+\ell_3)}, \ldots, \\ \boldsymbol{x}_{(1+\sum_{i=1}^{m-1}\ell_i):\sum_{i=1}^{m}\ell_i} \end{array} \right\rangle \tag{4}$$

where $\ell = \sum_{i=1}^{m} \ell_i$.

Application: word segmentation for writing systems without whitespace.

Segmentations

Segmentation:

- Input: $\boldsymbol{x} = \langle x_1, x_2, \ldots, x_\ell \rangle$
- Output:

$$\left\langle \begin{array}{l} \boldsymbol{x}_{1:\ell_1}, \\ \boldsymbol{x}_{(1+\ell_1):(\ell_1+\ell_2)}, \\ \boldsymbol{x}_{(1+\ell_1+\ell_2):(\ell_1+\ell_2+\ell_3)}, \ldots, \\ \boldsymbol{x}_{(1+\sum_{i=1}^{m-1} \ell_i):\sum_{i=1}^{m} \ell_i} \end{array} \right\rangle \tag{4}$$

where $\ell = \sum_{i=1}^{m} \ell_i$.

Application: word segmentation for writing systems without whitespace.

With arbitrarily long segments, this does not look like a job for $\boldsymbol{\phi}(\boldsymbol{x}, i, y, y')$!

Segmentation as Sequence Labeling

Ramshaw and Marcus (1995)

Two labels: B ("beginning of new segment"), I ("inside segment")

- $\ell_1 = 4, \ell_2 = 3, \ell_3 = 1, \ell_4 = 2 \longrightarrow \langle \text{B, I, I, I, B, I, I, B, B, I} \rangle$

Three labels: B, I, O ("outside segment")

Five labels: B, I, O, E ("end of segment"), S ("singleton")

Segmentation as Sequence Labeling

Ramshaw and Marcus (1995)

Two labels: B ("beginning of new segment"), I ("inside segment")

- $\ell_1 = 4, \ell_2 = 3, \ell_3 = 1, \ell_4 = 2 \longrightarrow \langle$B, I, I, I, B, I, I, B, B, I$\rangle$

Three labels: B, I, O ("outside segment")

Five labels: B, I, O, E ("end of segment"), S ("singleton")

Bonus: combine these with a label to get *labeled* segmentation!

Named Entity Recognition as Segmentation and Labeling

An older and narrower subset of supersenses used in information extraction:

- person,
- location,
- organization,
- geopolitical entity,
- . . . and perhaps domain-specific additions.

Named Entity Recognition

With $\underbrace{\text{Commander Chris Ferguson}}_{\text{person}}$ at the helm ,

$\underbrace{\text{Atlantis}}_{\text{spacecraft}}$ touched down at $\underbrace{\text{Kennedy Space Center}}_{\text{location}}$.

Named Entity Recognition

With $\underbrace{\text{Commander Chris Ferguson}}_{\text{person}}$ at the helm ,

O B I I O O O O

$\underbrace{\text{Atlantis}}_{\text{spacecraft}}$ touched down at $\underbrace{\text{Kennedy Space Center}}_{\text{location}}$.

B O O O B I I O

Named Entity Recognition: Evaluation

	1	2	3	4	5	6	7	8	9
$\boldsymbol{x}$ =	Britain	sent	warships	across	the	English	Channel	Monday	to
$\boldsymbol{y}$ =	B	O	O	O	O	B	I	B	O
$\boldsymbol{y}'$ =	O	O	O	O	O	B	I	B	O

10	11	12	13	14	15	16	17	18	19
rescue	Britons	stranded	by	Eyjafjallajökull	's	volcanic	ash	cloud	.
O	B	O	O	B	O	O	O	O	O
O	B	O	O	B	O	O	O	O	O

Segmentation Evaluation

Typically: precision, recall, and F_1.

Multiword Expressions

Schneider et al. (2014b)

- **MW compounds:** *red tape*, *motion picture*, *daddy longlegs*, *Bayes net*, *hot air balloon*, *skinny dip*, *trash talk*
- **verb-particle:** *pick up*, *dry out*, *take over*, *cut short*
- **verb-preposition:** *refer to*, *depend on*, *look for*, *prevent from*
- **verb-noun(-preposition):** *pay attention (to)*, *go bananas*, *lose it*, *break a leg*, *make the most of*
- **support verb:** *make decisions*, *take breaks*, *take pictures*, *have fun*, *perform surgery*
- **other phrasal verb:** *put up with*, *miss out (on)*, *get rid of*, *look forward to*, *run amok*, *cry foul*, *add insult to injury*, *make off with*
- **PP modifier:** *above board*, *beyond the pale*, *under the weather,at all*, *from time to time*, *in the nick of time*
- **coordinated phrase:** *cut and dry*, *more or less*, *up and leave*
- **conjunction/connective:** *as well as*, *let alone*, *in spite of*, *on the face of it/on its face*
- **semi-fixed VP:** *smack* <one>*'s lips*, *pick up where* <one> *left off*, *go over* <thing> *with a fine-tooth(ed) comb*, *take* <one>*'s time*, *draw* <oneself> *up to* <one>*'s full height*
- **fixed phrase:** *easy as pie*, *scared to death*, *go to hell in a handbasket*, *bring home the bacon*, *leave of absence*, *sense of humor*
- **phatic:** *You're welcome. Me neither!*
- **proverb:** *Beggars can't be choosers. The early bird gets the worm. To each his own. One man's* <thing$_1$> *is another man's* <thing$_2$>.

Sequence Labeling with Nesting

Schneider et al. (2014a)

he	was	willing	to	budge$_1$	a$_2$	little$_2$	on$_1$	the	price
O	O	O	O	B	b	ī	Ī	O	O

which	means^4	a$_3^4$	lot$_3^4$	to^4	me^4	.
O	B	Ĩ	Ī	Ĩ	Ĩ	O

Strong (subscript) vs. weak (superscript) MWEs.

One level of nesting, plus strong/weak distinction, can be handled with an eight-tag scheme.

Back to Syntax

Base noun phrase chunking:

[He]$_{NP}$ reckons [the current account deficit]$_{NP}$ will narrow to [only $ 1.8 billion]$_{NP}$ in [September]$_{NP}$

(What is a base noun phrase?)

"Chunking" used generically includes base verb and prepositional phrases, too.

Sequence labeling with BIO tags and features can be applied to this problem (Sha and Pereira, 2003).

Remarks

Sequence models are extremely useful:

- syntax: part-of-speech tags, base noun phrase chunking
- semantics: supersense tags, named entity recognition, multiword expressions

All of these are called "shallow" methods (why?).

Remarks

Sequence models are extremely useful:

- syntax: part-of-speech tags, base noun phrase chunking
- semantics: supersense tags, named entity recognition, multiword expressions

All of these are called "shallow" methods (why?).

Issues to be aware of:

- Supervised data for these problems is not cheap.
- Performance always suffers when you test on a different style, genre, dialect, etc. than you trained on.
- Runtime depends on the size of $\mathcal{L}$ and the number of consecutive labels that features can depend on.

References I

Daniel M. Bikel, Richard Schwartz, and Ralph M. Weischedel. An algorithm that learns what's in a name. *Machine learning*, 34(1–3):211–231, 1999. URL http://people.csail.mit.edu/mcollins/6864/slides/bikel.pdf.

Thorsten Brants. TnT – a statistical part-of-speech tagger. In *Proc. of ANLP*, 2000.

Kenneth W. Church. A stochastic parts program and noun phrase parser for unrestricted text. In *Proc. of ANLP*, 1988.

Massimiliano Ciaramita and Yasemin Altun. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proc. of EMNLP*, 2006.

Massimiliano Ciaramita and Mark Johnson. Supersense tagging of unknown nouns in WordNet. In *Proc. of EMNLP*, 2003.

Michael Collins. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In *Proc. of EMNLP*, 2002.

Michael Collins. Tagging with hidden Markov models, 2011. URL http://www.cs.columbia.edu/~mcollins/courses/nlp2011/notes/hmms.pdf.

John M. Conroy and Dianne P. O'Leary. Text summarization via hidden Markov models. In *Proc. of SIGIR*, 2001.

Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

References II

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proc. of ACL*, 2011.

Daniel Jurafsky and James H. Martin. Hidden Markov models (draft chapter), 2016a. URL `https://web.stanford.edu/~jurafsky/slp3/9.pdf`.

Daniel Jurafsky and James H. Martin. Information extraction (draft chapter), 2016b. URL `https://web.stanford.edu/~jurafsky/slp3/21.pdf`.

Daniel Jurafsky and James H. Martin. Part-of-speech tagging (draft chapter), 2016c. URL `https://web.stanford.edu/~jurafsky/slp3/10.pdf`.

John S. Justeson and Slava M. Katz. Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1:9–27, 1995.

Mark D. Kernighan, Kenneth W. Church, and William A. Gale. A spelling correction program based on a noisy channel model. In *Proc. of COLING*, 1990.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics*, 19(2):313–330, 1993.

G. A. Miller, C. Leacock, T. Randee, and R. Bunker. A semantic concordance. In *Proc. of HLT*, 1993.

Lance A Ramshaw and Mitchell P. Marcus. Text chunking using transformation-based learning, 1995. URL `http://arxiv.org/pdf/cmp-lg/9505040.pdf`.

References III

Nathan Schneider and Noah A. Smith. A corpus and model integrating multiword expressions and supersenses. In *Proc. of NAACL*, 2015.

Nathan Schneider, Emily Danchik, Chris Dyer, and Noah A. Smith. Discriminative lexical semantic segmentation with gaps: Running the MWE gamut. *Transactions of the Association for Computational Linguistics*, 2:193–206, April 2014a.

Nathan Schneider, Spencer Onuffer, Nora Kazour, Emily Danchik, Michael T. Mordowanec, Henrietta Conrad, and Noah A. Smith. Comprehensive annotation of multiword expressions in a social web corpus. In *Proc. of LREC*, 2014b.

Fei Sha and Fernando Pereira. Shallow parsing with conditional random fields. In *Proc. of NAACL*, 2003.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proc. of NAACL*, 2003.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. HMM-based word alignment in statistical translation. In *Proc. of COLING*, 1996.