Sequence Labelling

Dr. Uzair Ahmad

Program

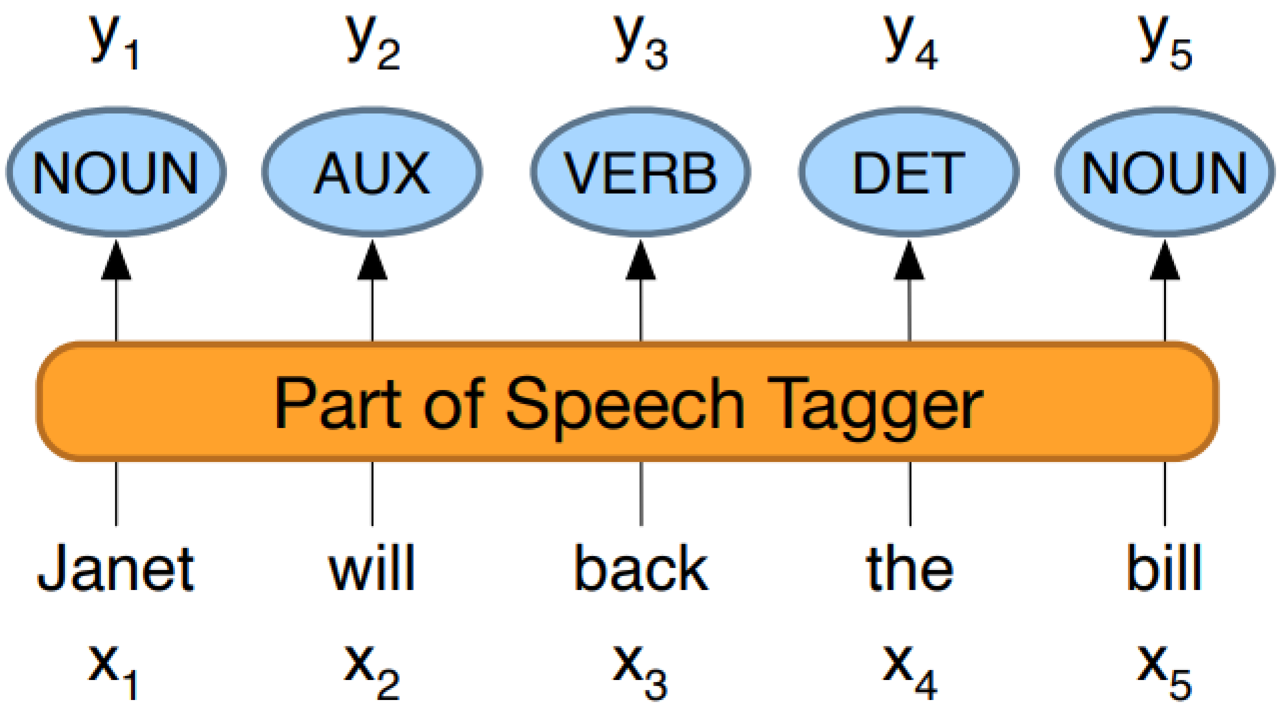
- Sequence labeling:
    - Part-of-speech tagging
    - NER
- Learning approaches
    - Classification approach
    - POS as structure prediction
        - Hidden Markov Models
        - The Viterbi algorithm
    - Discriminative approaches
        - RNNs and POS tagging

POS Tags

Tag	Description	Example	Tag	Description	Example	Tag	Description	Example
CC	coord. conj.	*and, but, or*	NNP	proper noun, sing.	*IBM*	TO	“to”	*to*
CD	cardinal number	*one, two*	NNPS	proper noun, plu.	*Carolinas*	UH	interjection	*ah, oops*
DT	determiner	*a, the*	NNS	noun, plural	*llamas*	VB	verb base	*eat*
EX	existential ‘there’	*there*	PDT	predeterminer	*all, both*	VBD	verb past tense	*ate*
FW	foreign word	*mea culpa*	POS	possessive ending	*’s*	VBG	verb gerund	*eating*
IN	preposition/ subordin-conj	*of, in, by*	PRP	personal pronoun	*I, you, he*	VBN	verb past participle	*eaten*
JJ	adjective	*yellow*	PRP$	possess. pronoun	*your, one’s*	VBP	verb non-3sg-pr	*eat*
JJR	comparative adj	*bigger*	RB	adverb	*quickly*	VBZ	verb 3sg pres	*eats*
JJS	superlative adj	*wildest*	RBR	comparative adv	*faster*	WDT	wh-determ.	*which, that*
LS	list item marker	*1, 2, One*	RBS	superlatv. adv	*fastest*	WP	wh-pronoun	*what, who*
MD	modal	*can, should*	RP	particle	*up, off*	WP$	wh-possess.	*whose*
NN	sing or mass noun	*llama*	SYM	symbol	*+,%, &*	WRB	wh-adverb	*how, where*

Penn Treebank part-of-speech tags.

POS Tagging Task



SLP: P157

Sequence labeling

- Why difficult ?
  - They can fish. → (N N V) (N V N)
  - Can of fish → (N P N)

Sequence labeling

	They	can	fish
Possible assignments	V	V	V
	V	V	N
	V	N	N
	V	N	V
	N	N	N
	N	V	N
	N	N	V
	N	V	V

Classification approach

- $\boldsymbol{w} = (w_1, w_2, w_3, \ldots, w_M)$
- Feature function
  - $f((w, m),\ y)$

$$\boldsymbol{f}((\boldsymbol{w} = \textit{they can fish}, m = 1), \mathrm{N}) = (\textit{they}, \mathrm{N})$$
$$\boldsymbol{f}((\boldsymbol{w} = \textit{they can fish}, m = 2), \mathrm{V}) = (\textit{can}, \mathrm{V})$$
$$\boldsymbol{f}((\boldsymbol{w} = \textit{they can fish}, m = 3), \mathrm{V}) = (\textit{fish}, \mathrm{V}).$$

Classification approach

- Grammatical ambiguity
  - They can fish. → (N N V) (N V N)
  - The can of fish → (D N P N)
- The tagger must rely on context

Classification approach

- Context and Grammatical ambiguity

$$
\begin{aligned}
\boldsymbol{f}((\boldsymbol{w} = \textit{they can fish}, 1), \mathrm{N}) = \{&(w_m = \textit{they}, y_m = \mathrm{N}),\\
&(w_{m-1} = \square, y_m = \mathrm{N}),\\
&(w_{m+1} = \textit{can}, y_m = \mathrm{N})\}\\
\boldsymbol{f}((\boldsymbol{w} = \textit{they can fish}, 2), \mathrm{V}) = \{&(w_m = \textit{can}, y_m = \mathrm{V}),\\
&(w_{m-1} = \textit{they}, y_m = \mathrm{V}),\\
&(w_{m+1} = \textit{fish}, y_m = \mathrm{V})\}\\
\boldsymbol{f}((\boldsymbol{w} = \textit{they can fish}, 3), \mathrm{V}) = \{&(w_m = \textit{fish}, y_m = \mathrm{V}),\\
&(w_{m-1} = \textit{can}, y_m = \mathrm{V}),\\
&(w_{m+1} = \blacksquare, y_m = \mathrm{V})\}.
\end{aligned}
$$

Classification approach

	1	2	3	4	5
W	The	old	man	the	boat.
POS	Det	Adj	N	Det	N
POS	Det	N	V	Det	N

Structure prediction

- Model the joint probability distribution $\mathrm{P}(w, y)$
  - often using probabilistic graphical models.

  - Set of tokens $\boldsymbol{w} = (w_1, w_2, w_3, \dots, w_M)$
  - Set of possible tags $\Upsilon(\mathrm{w}) = \Upsilon^M = (y_1, y_2,\ y_3, \dots,\ y_M)$
    - $\Upsilon = \{\mathrm{N}, \mathrm{V}, \mathrm{D}, \dots\}$

$$\hat{y} = \underset{y \in \Upsilon(\mathrm{w})}{\operatorname{argmax}} \Psi(w, y)$$

Structure prediction

- Restricting the scoring function

$$\Psi(w, y) = \sum_{m=1}^{M+1} \psi(w, y_m, y_{m-1}, m)$$

$$\psi(w_{1:M}, y_m, y_{m-1}, m) = \theta \cdot f(w, y_m, y_{m-1}, m)$$

Structure prediction

$$\boldsymbol{f}(\boldsymbol{w} = \textit{they can fish}, \boldsymbol{y} = \mathrm{N\ V\ V}\ ) = \sum_{m=1}^{M+1} \boldsymbol{f}(\boldsymbol{w}, y_m, y_{m-1}, m)$$

$$\begin{aligned}
&\boldsymbol{f}(\boldsymbol{w}, \mathrm{N}, \Diamond, 1) && (w_m = \textit{they}, y_m = \mathrm{N}) + (y_m = \mathrm{N}, y_{m-1} = \Diamond) \\
+\ &\boldsymbol{f}(\boldsymbol{w}, \mathrm{V}, \mathrm{N}, 2) && (w_m = \textit{can}, y_m = \mathrm{V}) + (y_m = \mathrm{V}, y_{m-1} = \mathrm{N}) \\
+\ &\boldsymbol{f}(\boldsymbol{w}, \mathrm{V}, \mathrm{V}, 3) && (w_m = \textit{fish}, y_m = \mathrm{V}) + (y_m = \mathrm{V}, y_{m-1} = \mathrm{V}) \\
+\ &\boldsymbol{f}(\boldsymbol{w}, \blacklozenge, \mathrm{V}, 4) && (y_m = \blacklozenge, y_{m-1} = \mathrm{V})
\end{aligned}$$

Structure prediction

- Inference by restricted scoring function

$$\Psi(w, y) = \sum_{m=1}^{M+1} \psi(w, y_m, y_{m-1}, m)$$

$$= \underset{\boldsymbol{y}_{1:M}}{\operatorname{argmax}} \sum_{m=1}^{M+1} \psi(\boldsymbol{w}, y_m, y_{m-1}, m)$$

$$= \underset{\boldsymbol{y}_{1:M}}{\operatorname{argmax}} \sum_{m=1}^{M+1} s_m(y_m, y_{m-1}),$$

Hidden Markov Models

- Probabilistic estimation of scores $s_m(\mathbf{y}, \boldsymbol{y}')$
- Naïve bayes classifier
  - $p(y|x) \propto p(x, y)$
- Probabilistic sequence labeling
  - $p(y|w) \propto p(y, w)$

$$\Psi(w, y) = \sum_{m=1}^{M+1} \psi(w, y_m, y_{m-1}, m)$$

Hidden Markov Models

- Probabilistic sequence labeling
  - $\mathrm{p}(y|w) \propto \mathrm{p}(y, w)$

each token depends **only** on its tag

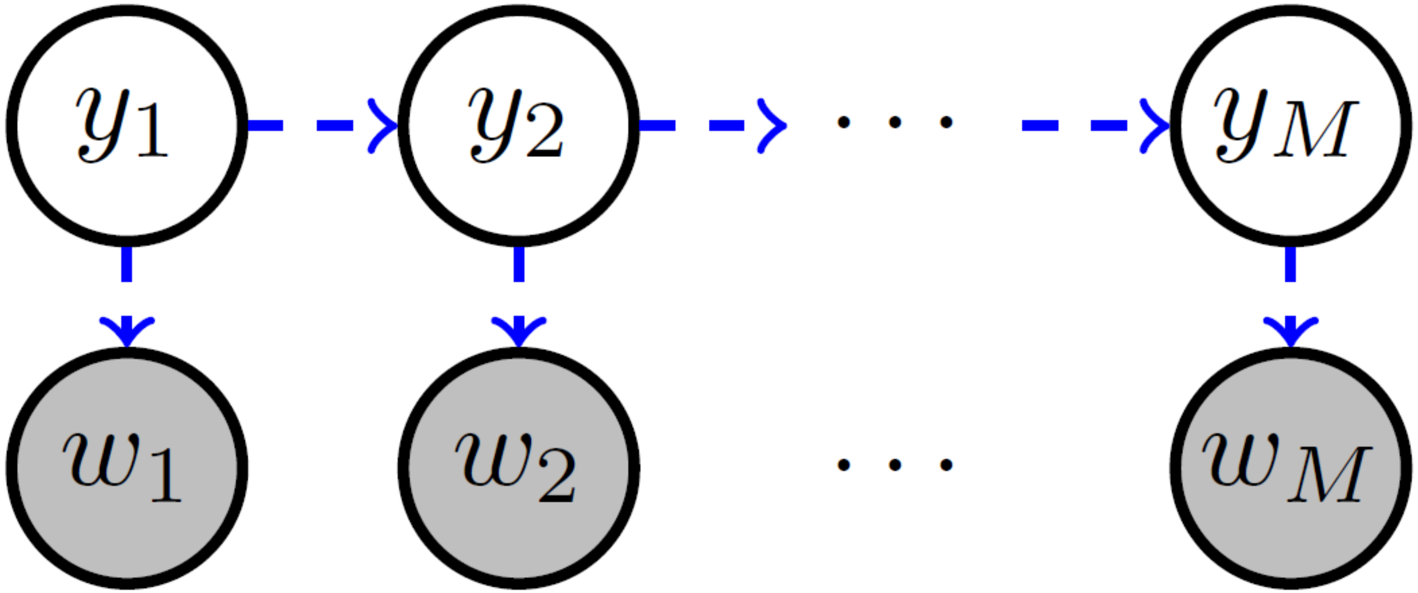
each tag depends **only** on its predecessor

$$\mathrm{p}(y|w) = \prod_{m=1}^{M} \mathrm{p}(w_m \, | y_m)$$

$$\mathrm{p}(y) = \prod_{m=1}^{M} \mathrm{p}(y_m \, | y_{m-1})$$

Hidden Markov Models

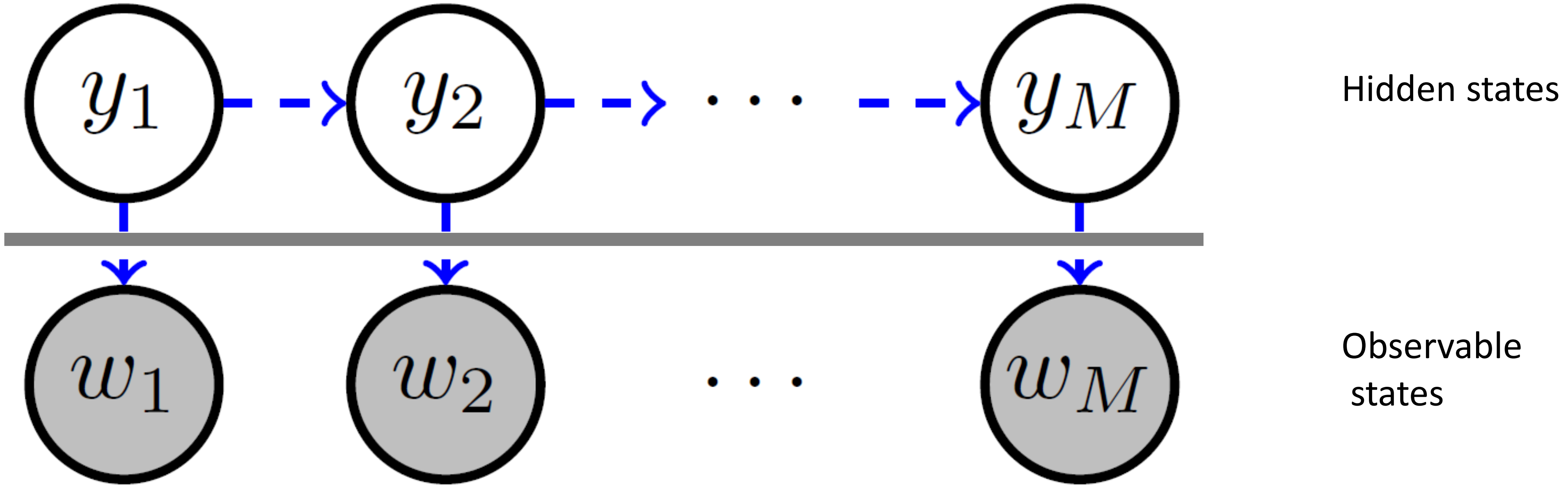
$$p(y) = \prod_{m=1}^{M} \mathrm{p}(y_m \mid y_{m-1})$$



$$p(y|w) = \prod_{m=1}^{M} \mathrm{p}(w_m \mid y_m)$$

Hidden Markov Models

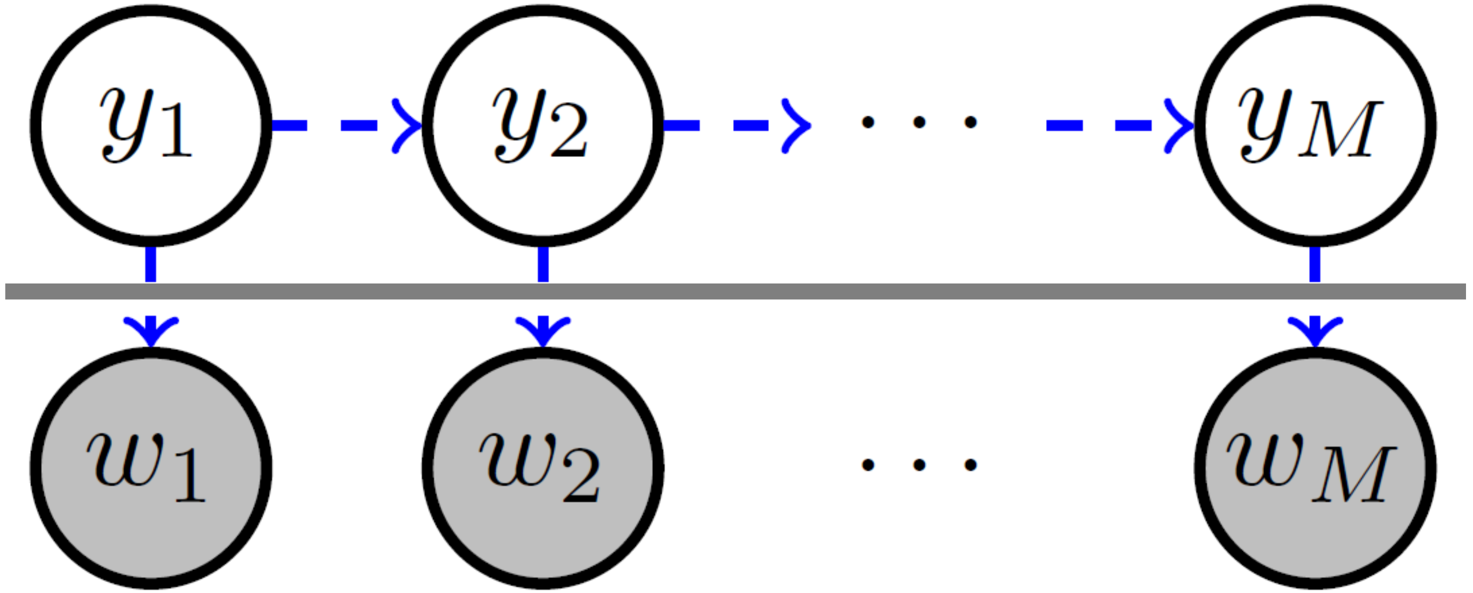
$$p(y) = \prod_{m=1}^{M} \mathrm{p}(y_m \,|y_{m-1})$$ → Transition probabilities



$$p(y|w) = \prod_{m=1}^{M} \mathrm{p}(w_m \,|y_m)$$ → Emission probabilities

Hidden Markov Models

$$p_t(y) = \prod_{m=1}^{M} \mathrm{p}(y_m \mid y_{m-1}; \lambda)$$ → Transition probabilities



Hidden states

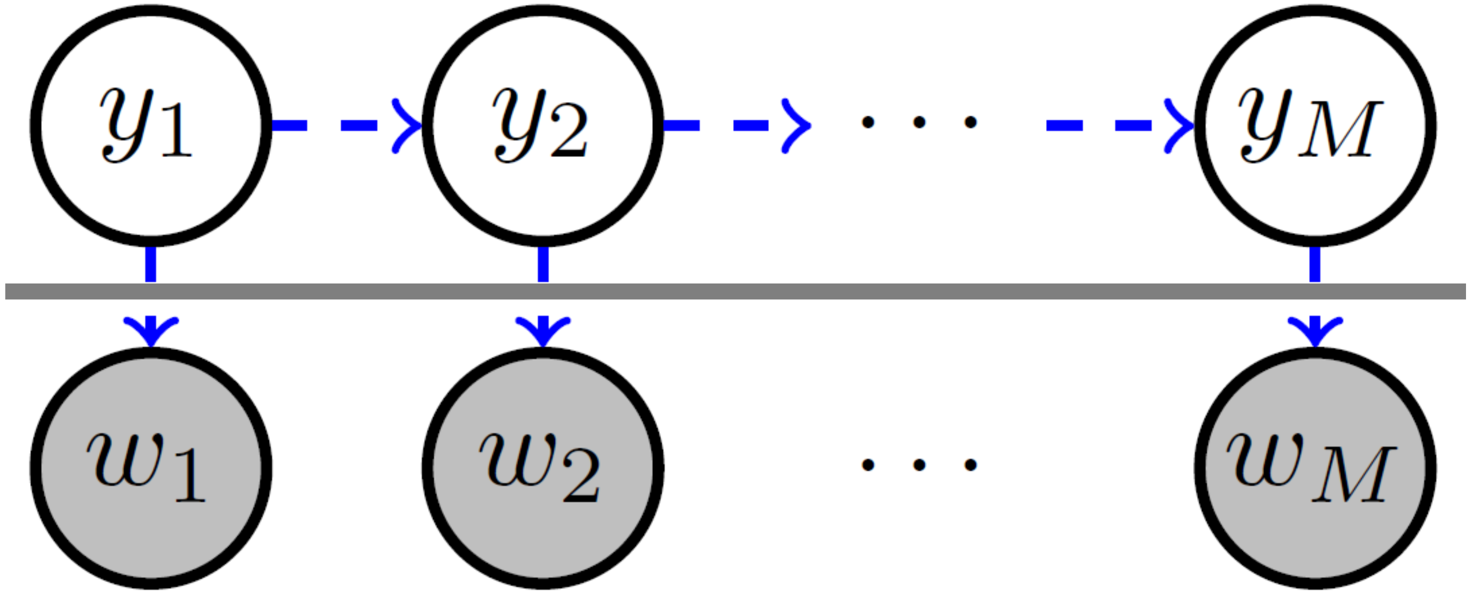
Observable states

$$p_e(y|w) = \prod_{m=1}^{M} \mathrm{p}(w_m \mid y_m; \phi)$$ → Emission probabilities

Hidden Markov Models

$$p_t(y) = \prod_{m=1}^{M} \mathrm{p}(y_m \,|y_{m-1}; \lambda)$$ → Transition probabilities

$$\lambda_{k,k'} \triangleq \Pr(Y_m = k' \mid Y_{m-1} = k) = \frac{\text{count}(Y_m = k', Y_{m-1} = k)}{\text{count}(Y_{m-1} = k)}$$



Hidden states

Observable symbols

$$p_e(y|w) = \prod_{m=1}^{M} \mathrm{p}(w_m \,|y_m; \phi)$$ → Emission probabilities

$$\phi_{k,i} \triangleq \Pr(W_m = i \mid Y_m = k) = \frac{\text{count}(W_m = i, Y_m = k)}{\text{count}(Y_m = k)}$$

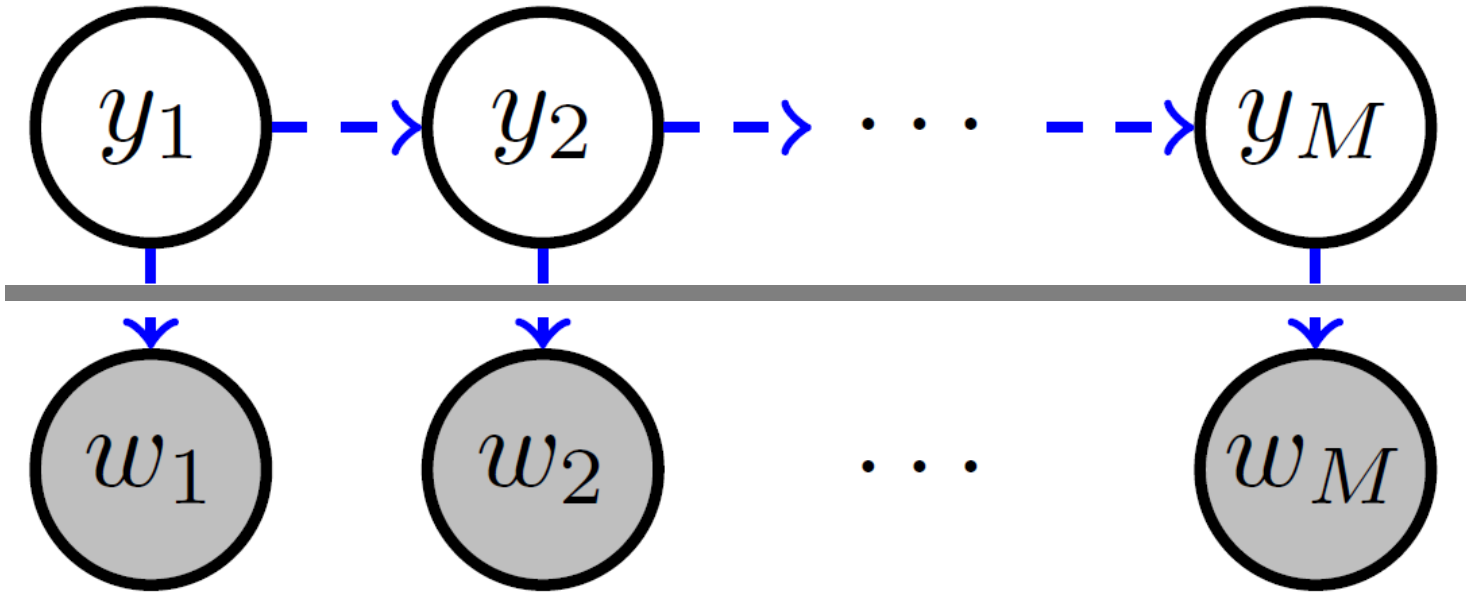
Hidden Markov Models

Inference

$$\hat{y} = \underset{y}{\operatorname{argmax}}\, \mathrm{p}(y|w) \approx \max_{y} \mathrm{p}(y, w)$$

$$\mathrm{p}(y, w) = \mathrm{p}(y \mid w) \times \mathrm{p}(w) \propto \mathrm{p}(y \mid w)$$

$$\mathrm{p}(y, w) = \mathrm{p}(w \mid y) \times \mathrm{p}(y) \propto \mathrm{p}(w \mid y)$$



Hidden states

Observable states

Hidden Markov Models

Inference

$$\boxed{\hat{y} = \underset{y}{\operatorname{argmax}} \log\left(\mathrm{p}(y|w)\right) \approx \max_{y} \log\left(\mathrm{p}(y,w)\right)}$$

$$\log\left(\mathrm{p}(y,w)\right) = \log\left(\mathrm{p}(\mathrm{y})\right) \times \log \mathrm{p}(w|y)$$

$$\log\left(\mathrm{p}(y,w)\right) = \sum_{m=1}^{M} \log \mathrm{p}_{Y}(y_m \mid y_{m-1}) + \log \mathrm{p}_{W|Y}(w_m \mid y_m)$$

$$\log\left(\mathrm{p}(y,w)\right) = \sum_{m=1}^{M} \log \lambda_{y_m, y_{m-1}} + \log \phi_{y_m, w_m}$$

$$\boxed{s_m(y_m, y_{m-1})}$$

Viterbi Algorithm

$$
\begin{aligned}
v_m(y_m) &\triangleq \max_{\boldsymbol{y}_{1:m-1}} \sum_{n=1}^{m} s_n(y_n, y_{n-1}) \\
&= \max_{y_{m-1}} s_m(y_m, y_{m-1}) + \max_{\boldsymbol{y}_{1:m-2}} \sum_{n=1}^{m-1} s_n(y_n, y_{n-1}) \\
&= \max_{y_{m-1}} s_m(y_m, y_{m-1}) + v_{m-1}(y_{m-1}).
\end{aligned}
$$

Viterbi Algorithm

for $k \in \{0, \dots K\}$ **do** **Initialize**

$$v_1(k) = s_1(k, \Diamond)$$

for $m \in \{2, \dots, M\}$ **do** **Recursion**

for $k \in \{0, \dots, K\}$ **do**

$$v_m(k) = \max_{k'} s_m(k, k') + v_{m-1}(k')$$

$$b_m(k) = \operatorname{argmax}_{k'} s_m(k, k') + v_{m-1}(k')$$

$$y_M = \operatorname{argmax}_k s_{M+1}(\blacklozenge, k) + v_M(k)$$ **Termination**

for $m \in \{M-1, \dots 1\}$ **do** **Backtracking**

$$y_m = b_m(y_{m+1})$$

return $y_{1:M}$

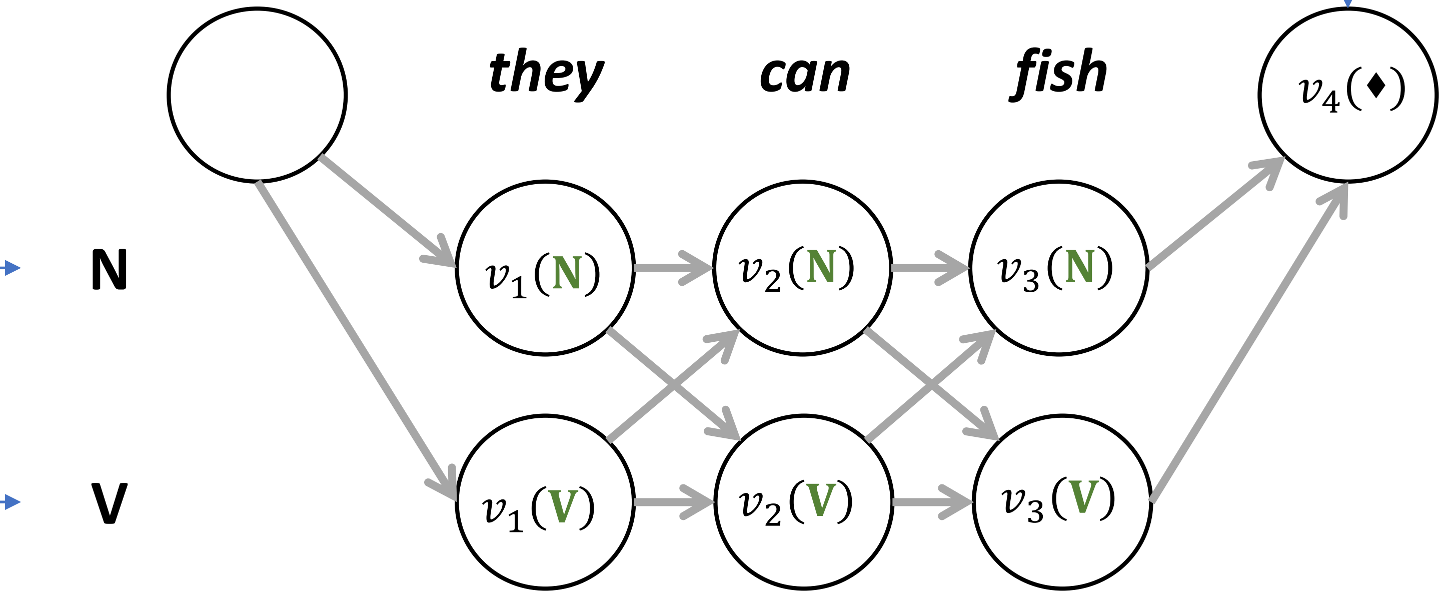
Viterbi Algorithm

	They	can	fish
Possible assignments	V	V	V
	V	V	N
	V	N	N
	V	N	V
	N	N	N
	N	V	N
	N	N	V
	N	V	V

Viterbi Algorithm

for $m \in \{2, \ldots, M\}$ **do**
  for $k \in \{0, \ldots, K\}$ **do**
    $v_m(k) = \max_{k'} s_m(k, k') + v_{m-1}(k')$
    $b_m(k) = \text{argmax}_{k'} s_m(k, k') + v_{m-1}(k')$

for $k \in \{0, \ldots K\}$ **do**
  $v_1(k) = s_1(k, \Diamond)$



Viterbi Algorithm

M

K

	they	*can*	*fish*
N	-2	-3	-3
V	-10	-1	-3

Emissions

K - to

K'- from

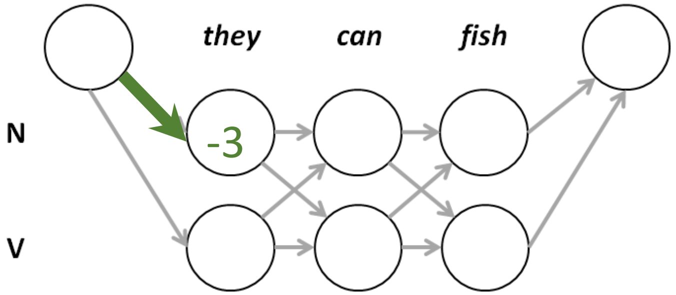
	N	V	$\blacklozenge$
$\lozenge$	-1	-2	$-\infty$
N	-3	-1	-1
V	-1	-3	-1

Transitions

$$\boxed{\begin{array}{l}\textbf{for } k \in \{0, \dots K\} \textbf{ do} \\ \quad v_1(k) = s_1(k, \Diamond)\end{array}}$$

$$v_1(\mathbf{N}) = s_1(\mathbf{N}, \Diamond) = -2 - 1 = -3$$

$$b_1(\mathbf{N}) = y_{m-1} = \Diamond$$

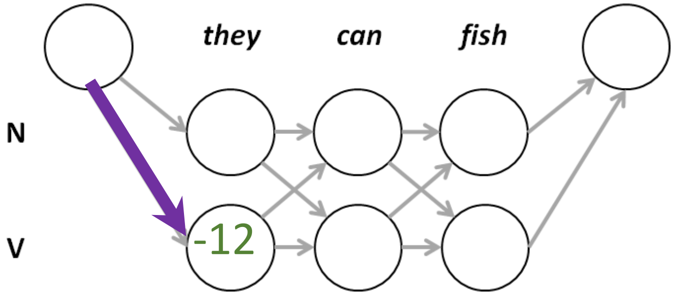


	they	*can*	*fish*
N	−2	−3	−3
V	−10	−1	−3

	N	V	◆
◊	−1	−2	−∞
N	−3	−1	−1
V	−1	−3	−1

$$v_1(\mathbf{V}) = s_1(\mathbf{V}, \Diamond) = -10 - 2 = -12$$

$$b_1(\mathbf{V}) = y_{m-1} = \Diamond$$



	they	*can*	*fish*
N	−2	−3	−3
V	−10	−1	−3

	N	V	◆
◊	−1	−2	−∞
N	−3	−1	−1
V	−1	−3	−1

$$
\begin{aligned}
&\textbf{for } m \in \{2, \ldots, M\} \textbf{ do}\\
&\quad \textbf{for } k \in \{0, \ldots, K\} \textbf{ do}\\
&\qquad v_m(k) = \max_{k'} s_m(k, k') + v_{m-1}(k')\\
&\qquad b_m(k) = \operatorname{argmax}_{k'} s_m(k, k') + v_{m-1}(k')
\end{aligned}
$$

$$v_2(\mathbf{N}) = \max(v_1(\mathbf{N}) + s_2(\mathbf{N},\mathbf{N}), v_1(\mathbf{V}) + s_2(\mathbf{N},\mathbf{V}))$$

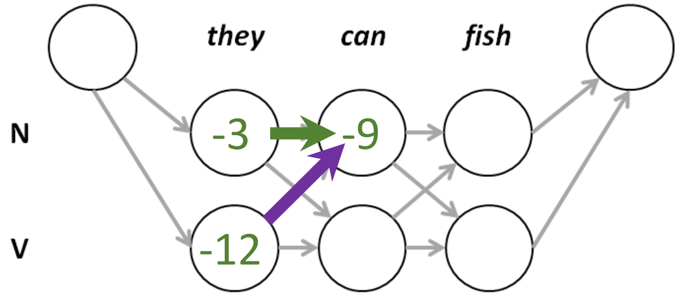
$$v_2(\mathbf{N}) = \max(-3 + s_2(\mathbf{N},\mathbf{N}), -12 + s_2(\mathbf{N},\mathbf{V}))$$

$$s_2(\mathbf{N},\mathbf{N}) = -3 - 3$$

$$s_2(\mathbf{N},\mathbf{V}) = -3 - 1$$

$$v_2(\mathbf{N}) = \max(-3-3-3, -12-3-1) = -9$$

$$b_2(\mathbf{N}) = y_{m-1} = \mathbf{N}$$



	they	*can*	*fish*
N	-2	-3	-3
V	-10	-1	-3

K - to

K'- from

	N	V	◆
◇	-1	-2	$-\infty$
N	-3	-1	-1
V	-1	-3	-1

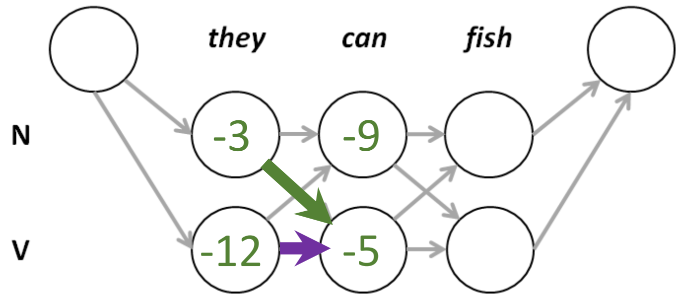
for $m \in \{2, \ldots, M\}$ **do**
  for $k \in \{0, \ldots, K\}$ **do**
    $v_m(k) = \max_{k'} s_m(k, k') + v_{m-1}(k')$
    $b_m(k) = \mathrm{argmax}_{k'} s_m(k, k') + v_{m-1}(k')$

$$v_2(\mathbf{V}) = \max(v_1(\mathbf{N}) + s_2(\mathbf{V}, \mathbf{N}), v_1(\mathbf{V}) + s_2(\mathbf{V}, \mathbf{V}))$$

$$v_2(\mathbf{V}) = \max(-3 + \underbrace{s_2(\mathbf{V}, \mathbf{N})}_{-1-1}, -12 + \underbrace{s_2(\mathbf{V}, \mathbf{V})}_{-1-3})$$

$$v_2(\mathbf{V}) = \max(-3-1-1, -12-1-3) = -5$$

$$b_2(\mathbf{V}) = y_{m-1} = \mathbf{N}$$



	they	*can*	*fish*
N	-2	-3	-3
V	-10	-1	-3

K - to

K'- from

	N	V	♦
◇	-1	-2	$-\infty$
N	-3	-1	-1
V	-1	-3	-1

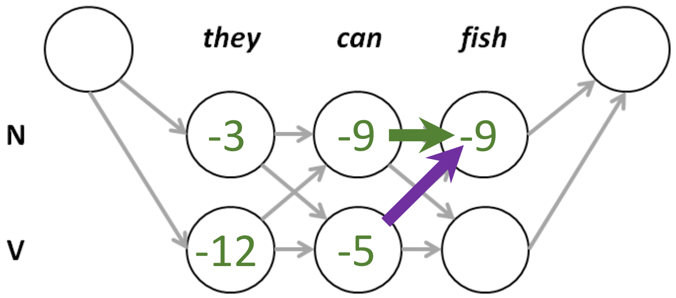
for $m \in \{2, \ldots, M\}$ **do**
  for $k \in \{0, \ldots, K\}$ **do**
    $v_m(k) = \max_{k'} s_m(k, k') + v_{m-1}(k')$
    $b_m(k) = \text{argmax}_{k'} s_m(k, k') + v_{m-1}(k')$

$$v_3(\mathbf{N}) = \max(v_2(\mathbf{N}) + s_3(\mathbf{N}, \mathbf{N}), v_2(\mathbf{V}) + s_3(\mathbf{N}, \mathbf{V}))$$

$$v_3(\mathbf{N}) = \max(-9 + \underbrace{s_3(\mathbf{N}, \mathbf{N})}_{-3-3}, -5 + \underbrace{s_3(\mathbf{N}, \mathbf{V})}_{-3-1})$$

$$v_3(\mathbf{N}) = \max(-9-3-3, -5-3-1) = -9$$

$$b_3(\mathbf{N}) = y_{m-1} = \mathbf{V}$$



	they	*can*	*fish*
N	-2	-3	-3
V	-10	-1	-3

K - to

K'- from

	N	V	♦
◇	-1	-2	$-\infty$
N	-3	-1	-1
V	-1	-3	-1

for $m \in \{2, \ldots, M\}$ **do**
  for $k \in \{0, \ldots, K\}$ **do**
    $v_m(k) = \max_{k'} s_m(k, k') + v_{m-1}(k')$
    $b_m(k) = \operatorname{argmax}_{k'} s_m(k, k') + v_{m-1}(k')$

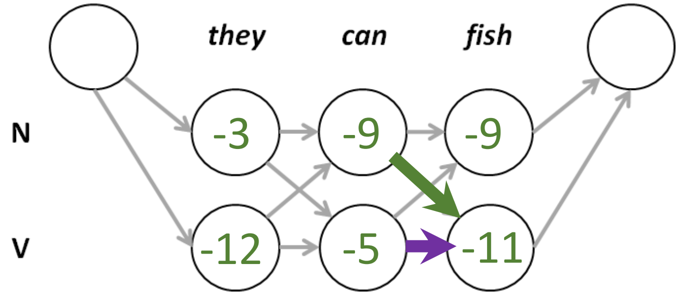
$$v_3(\mathbf{V}) = \max(v_2(\mathbf{N}) + s_3(\mathbf{V}, \mathbf{N}), v_2(\mathbf{V}) + s_3(\mathbf{V}, \mathbf{V}))$$

$$v_3(\mathbf{V}) = \max(-9 + s_3(\mathbf{V}, \mathbf{N}), -5 + s_3(\mathbf{V}, \mathbf{V}))$$

$$-3-1 \qquad\qquad -3-3$$

$$v_3(\mathbf{V}) = \max(-9-3-1, -5-3-3) = -11$$

$$b_3(\mathbf{V}) = y_{m-1} = \mathbf{V}$$



	they	*can*	*fish*
N	-2	-3	-3
V	-10	-1	-3

K - to

K'- from

	N	V	◆
◇	-1	-2	$-\infty$
N	-3	-1	-1
V	-1	-3	-1

Termination

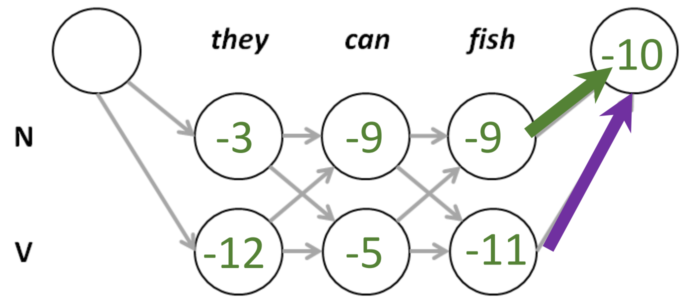
$$\boxed{y_M = \text{argmax}_k \, s_{M+1}(\blacklozenge, k) + v_M(k)}$$

$$v_4(\blacklozenge) = \max(v_3(\mathbf{N}) + s_4(\blacklozenge, \mathbf{N}), v_3(\mathbf{V}) + s_4(\blacklozenge, \mathbf{V}))$$

$$v_4(\blacklozenge) = \max(-9 + \underbrace{s_4(\blacklozenge, \mathbf{N})}_{0-1}, \, -11 + \underbrace{s_4(\blacklozenge, \mathbf{V})}_{0-1})$$

$$v_4(\blacklozenge) = \max(-9 + 0-1, \, -11+0-1) = -10$$

$$b_4(\blacklozenge) = y_{m-1} = N$$



	they	*can*	*fish*
N	-2	-3	-3
V	-10	-1	-3

K - to

K'- from

	N	V	$\blacklozenge$
$\lozenge$	-1	-2	$-\infty$
N	-3	-1	-1
V	-1	-3	-1

Backtracking

for $m \in \{M-1, \ldots 1\}$ **do**

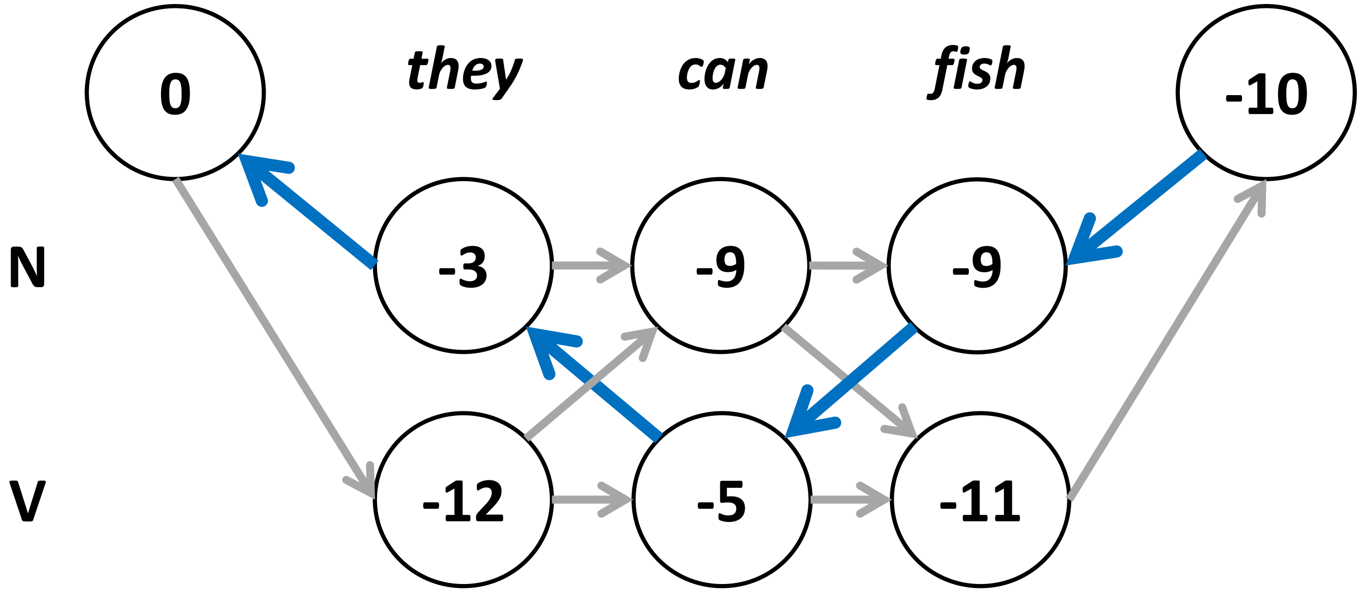
$$y_m = b_m(y_{m+1})$$

$y_0 = b_1(\mathbf{N}) = \Diamond$

$y_1 = b_2(\mathbf{N}) = \mathbf{N}$

$y_2 = b_3(\mathbf{V}) = \mathbf{V}$

$y_3 = b_4(\blacklozenge) = N$



Viterbi Algorithm

- Restricting scoring function to local parts
  - Only pairs of adjacent tags
  - Akin to a bigram language model (over tags)
- Higher-order features ?

$$\Psi(w, y) = \sum_{m=1}^{M+2} \psi(w, y_m, y_{m-1}, y_{m-2}, m)$$

Discriminative approaches

- Structured Perceptron
  - Increase weights of correct classifications
  - Decrease weights of incorrect classification
- Structured SVM
  - Push class boundary away from training instances
- Conditional Random Fields
  - conditional probabilistic model for sequence labeling;
  - built on the logistic regression classifier

References

- Natural Language Processing | J. Eisenstein