Natural Language Processing
Sequence labeling: CRFs

Sofia Serrano
sofias6@cs.washington.edu

Credit to Xiaochuang Han for slides

Announcements

- A2 is due a week from tonight
- Extra office hours next week– we'll update the course google calendar (also embedded on the course website below the calendar table) with those extra office hours by Sunday evening (and send out an Ed announcement once that's done)
- A1 grades are out!
  - We'll be taking A1 regrade requests through the end of Thursday 2/16

A couple of closing words about Viterbi (because there's always something)

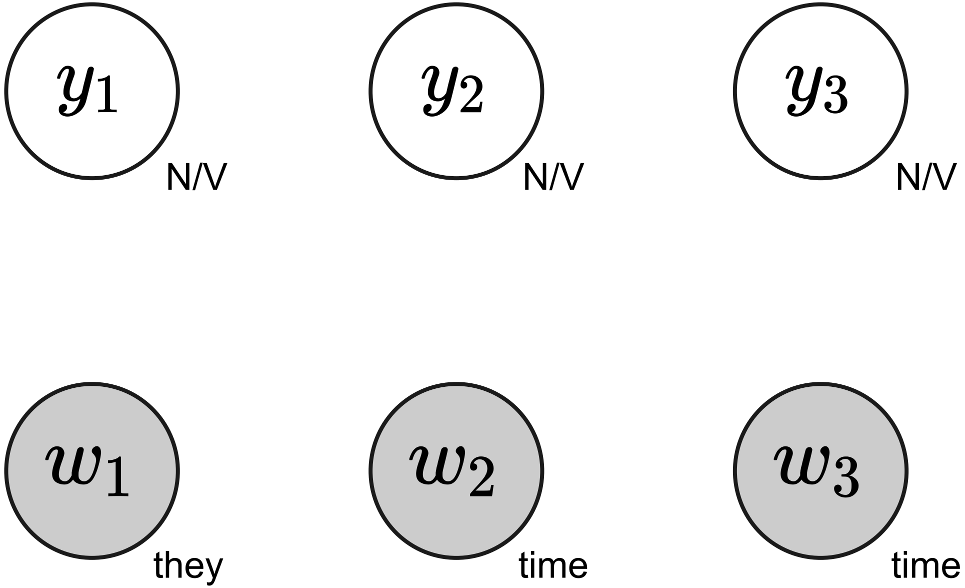
- Multiplying together a bunch of probabilities → we want to do our calculations in log space instead!
- We can think about doing Viterbi either by filling in a table, or by recursively filling out column through column– *as long as we don't accidentally throw out old columns' backpointers by doing so*
- Make sure your table of transition probabilities and your table of observation likelihoods are all estimated *before* you start decoding for any particular input. (They're tables, but they're separate from your Viterbi dynamic-programming table!)

Conditional Random Fields

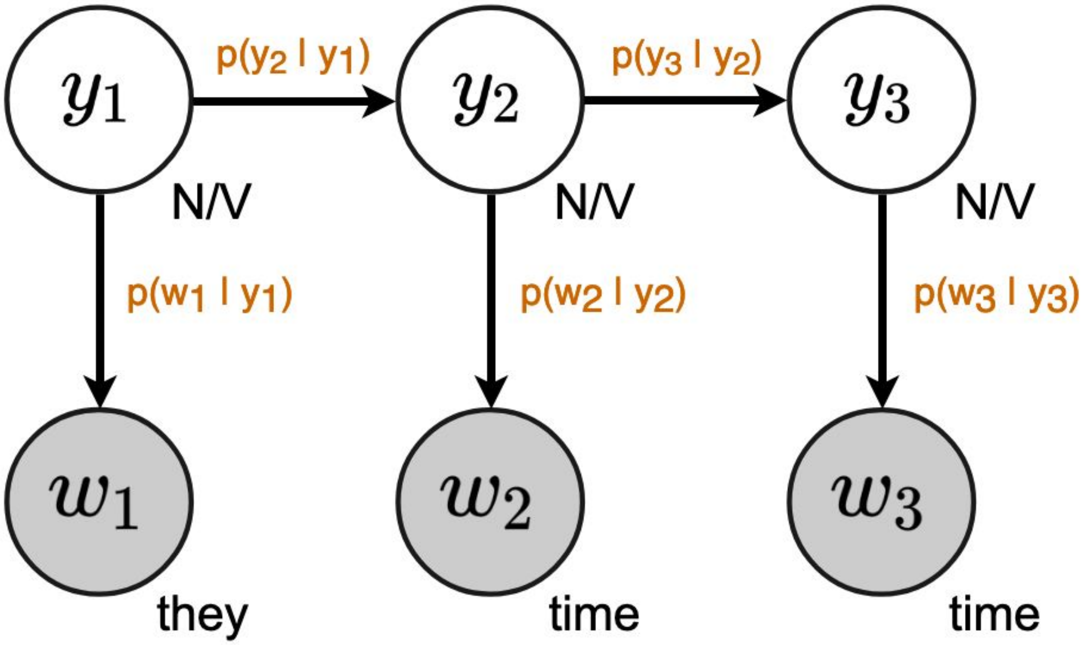
A POS-tagging example

POS?	*POS?*	*POS?*
they	time	time

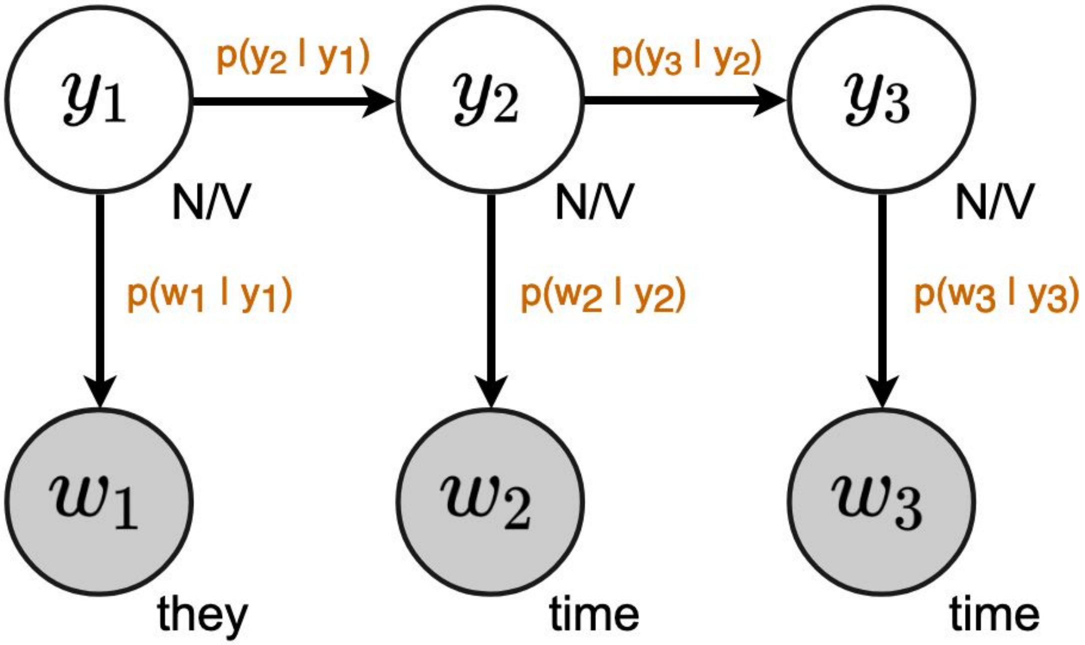
A POS-tagging example



Our POS-tagging example as an HMM



Our POS-tagging example as an HMM



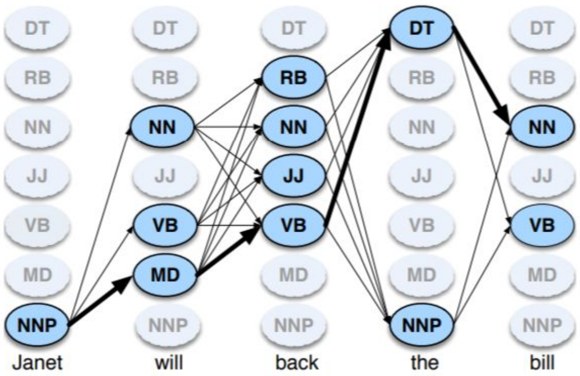
Score of specific sequence of states/emissions: The product of all the orange numbers

Review: Parameters of our HMM model

We learned these parameters using count-based estimation.

	NNP	MD	VB	JJ	NN	RB	DT
<s>	0.2767	0.0006	0.0031	0.0453	0.0449	0.0510	0.2026
NNP	0.3777	0.0110	0.0009	0.0084	0.0584	0.0090	0.0025
MD	0.0008	0.0002	0.7968	0.0005	0.0008	0.1698	0.0041
VB	0.0322	0.0005	0.0050	0.0837	0.0615	0.0514	0.2231
JJ	0.0366	0.0004	0.0001	0.0733	0.4509	0.0036	0.0036
NN	0.0096	0.0176	0.0014	0.0086	0.1216	0.0177	0.0068
RB	0.0068	0.0102	0.1011	0.1012	0.0120	0.0728	0.0479
DT	0.1147	0.0021	0.0002	0.2157	0.4744	0.0102	0.0017

	Janet	will	back	the	bill
NNP	0.000032	0	0	0.000048	0
MD	0	0.308431	0	0	0
VB	0	0.000028	0.000672	0	0.000028
JJ	0	0	0.000340	0	0
NN	0	0.000200	0.000223	0	0.002337
RB	0	0	0.010446	0	0
DT	0	0	0	0.506099	0

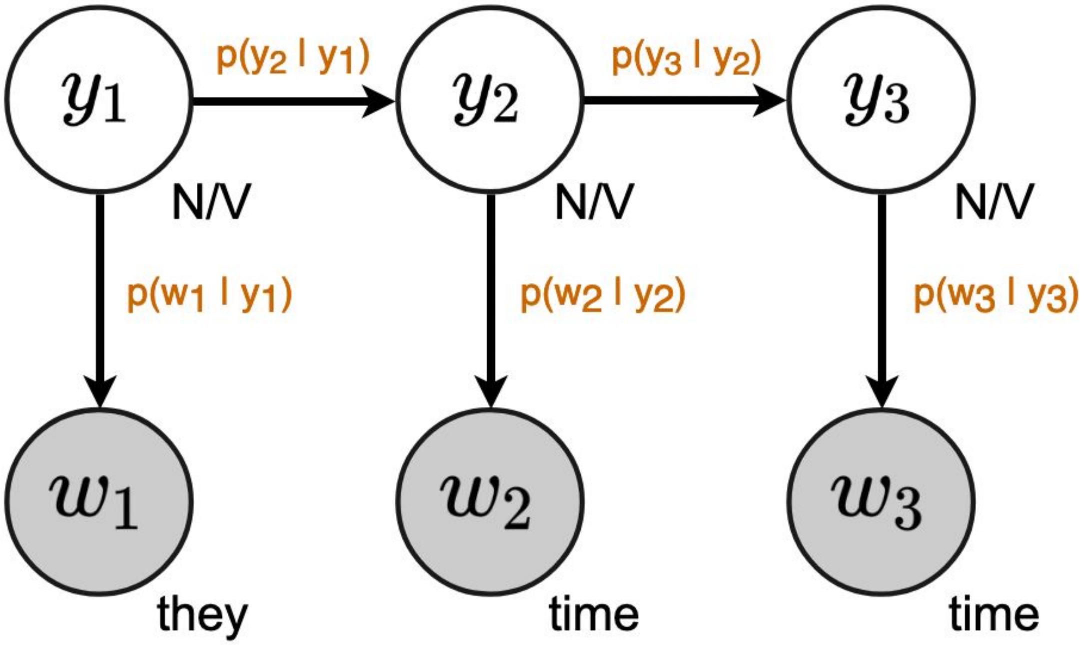


A

$$v_t(j) = \max_{i=1}^{N} v_{t-1}(i)\, a_{ij}\, b_j(o_t)$$

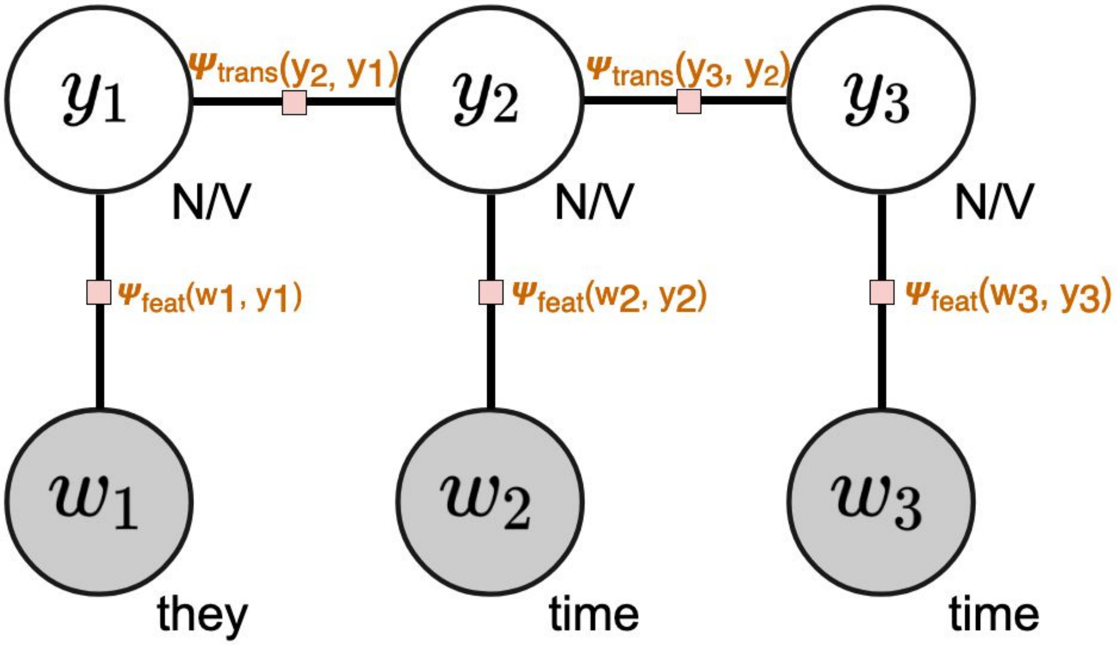
B

Our POS-tagging example as an HMM



Score of specific sequence of states/emissions: The product of all the orange numbers

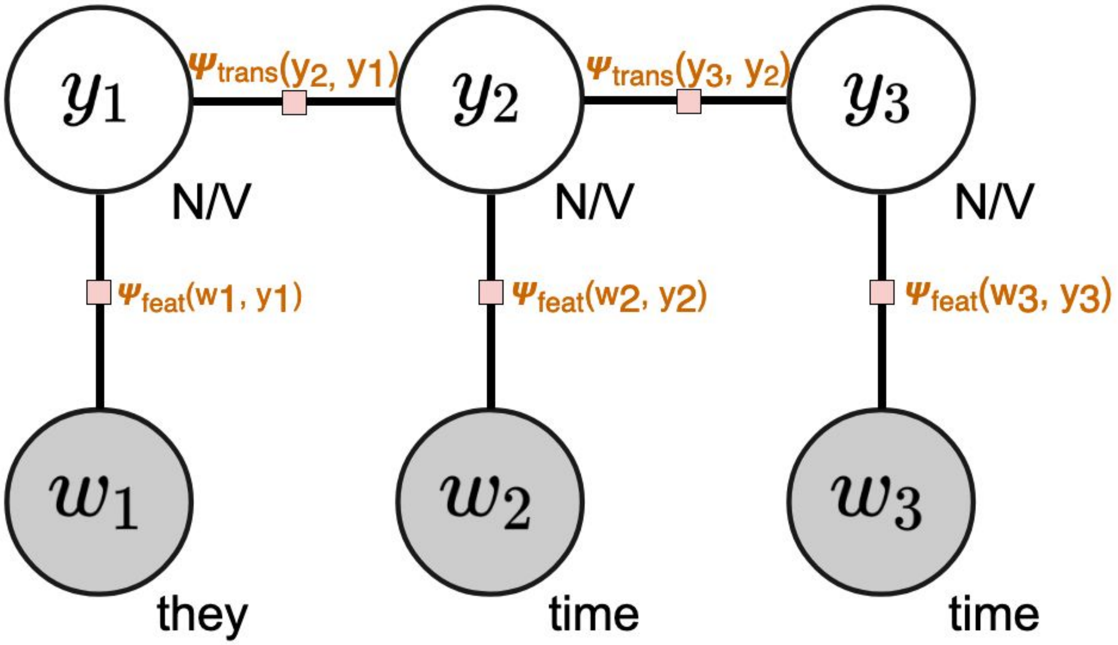
Switching to a CRF: p(A | B) → Ψ(A, B)



Score of specific sequence of states/emissions: The product of all the orange numbers

Switching to a CRF: p(A | B) → **Ψ**(A, B)

Key change: we allow ***Ψ*** functions to output *any positive number*

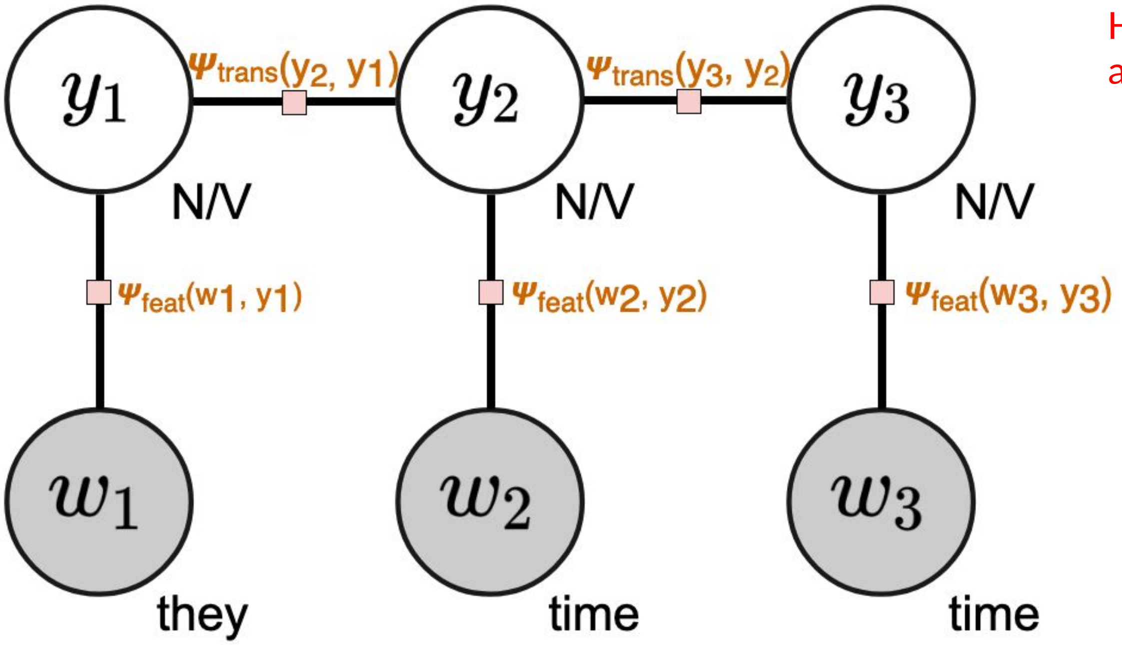


Score of specific sequence of states/emissions: The product of all the orange numbers

Switching to a CRF: p(A | B) → **Ψ**(A, B)

Key change: we allow **Ψ** functions to output *any positive number*

How do we do this? Exponentiate as the final part of each **Ψ**



Score of specific sequence of states/emissions: The product of all the orange numbers

Where have we seen this shift from parameters representing probabilities to learned weights before?

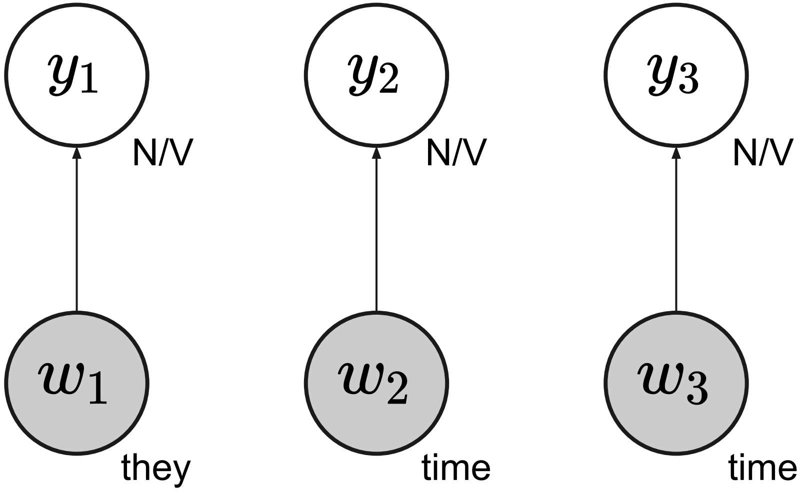
In moving from Naive Bayes to Logistic Regression!

In switching from an HMM to a CRF, we've moved from a generative sequence labeling model to a discriminative one.

How do we learn the parameters for the **potentials** (those $\boldsymbol{\Psi}$ functions)?

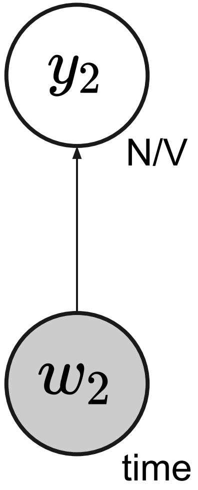
Follow our strategy that we used for learning a logistic regression model!

Logistic regression

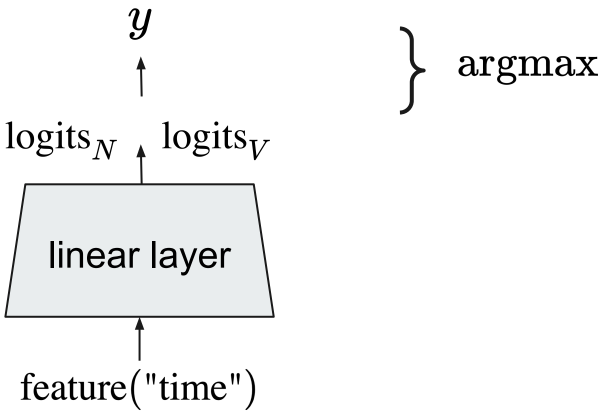


Predict each individual tag with logistic regression

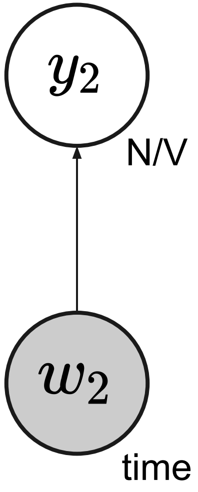
Logistic regression



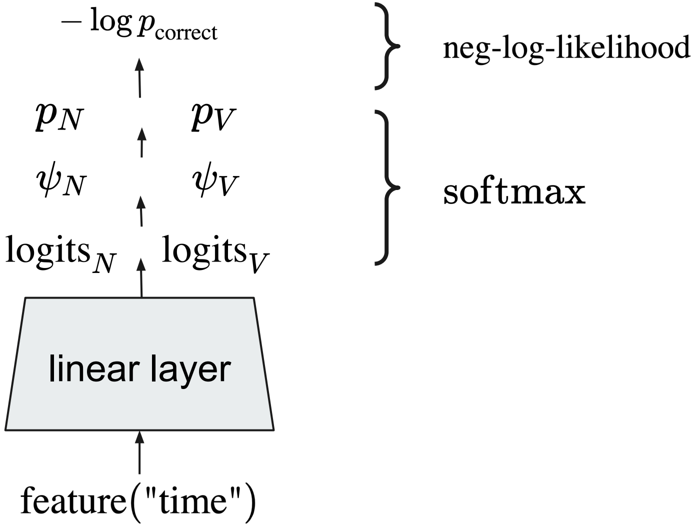
Inference



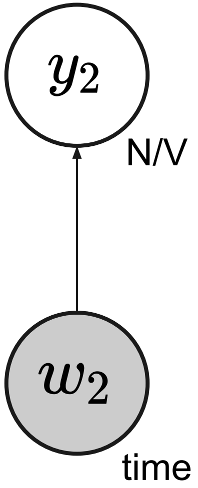
Logistic regression



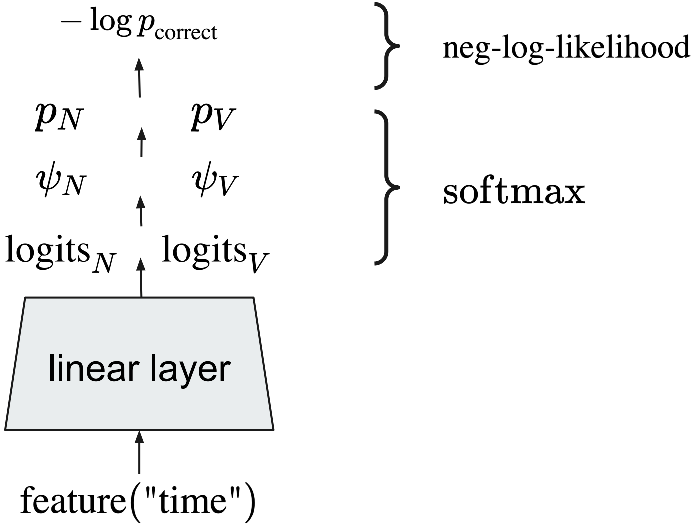
Training



Logistic regression



Training

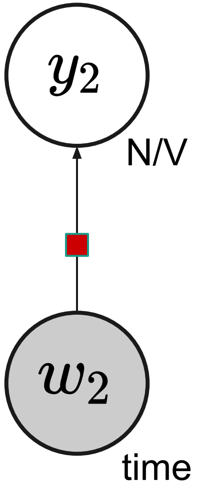


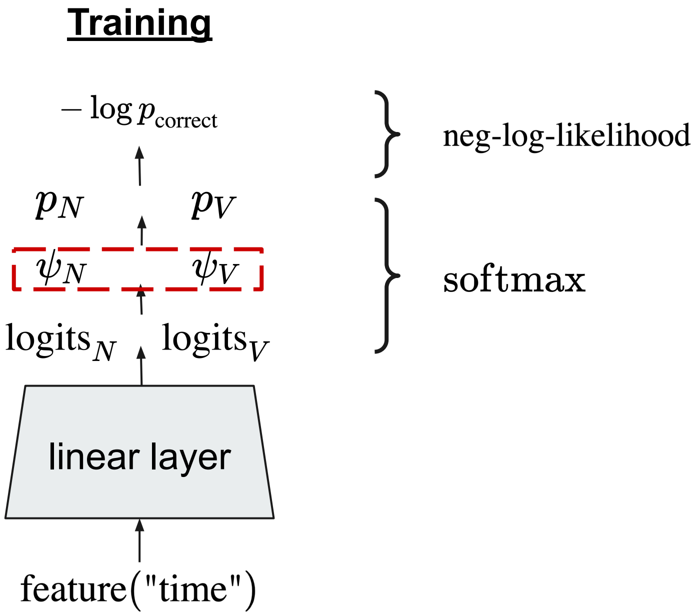
$$\psi_N = \exp(\text{logits}_N)$$

$p_N \propto \psi_N$ or more precisely, $p_N = \frac{\psi_N}{\psi_N + \psi_V}$

softmax

Logistic regression

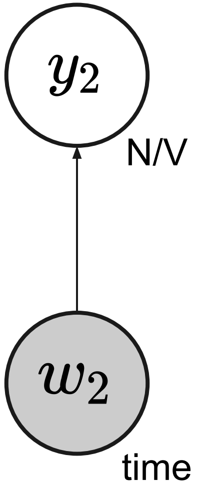




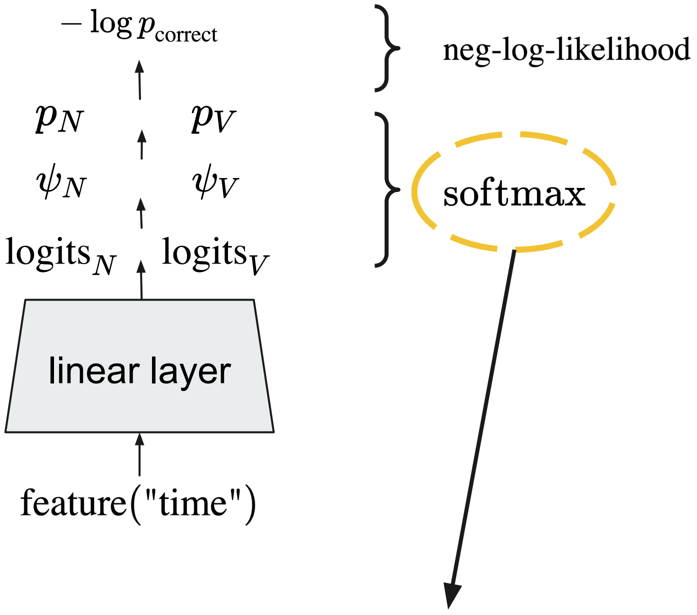
$$\psi_N = \exp(\text{logits}_N)$$

$p_N \propto \psi_N$ or more precisely, $p_N = \frac{\psi_N}{\psi_N + \psi_V}$

Logistic regression

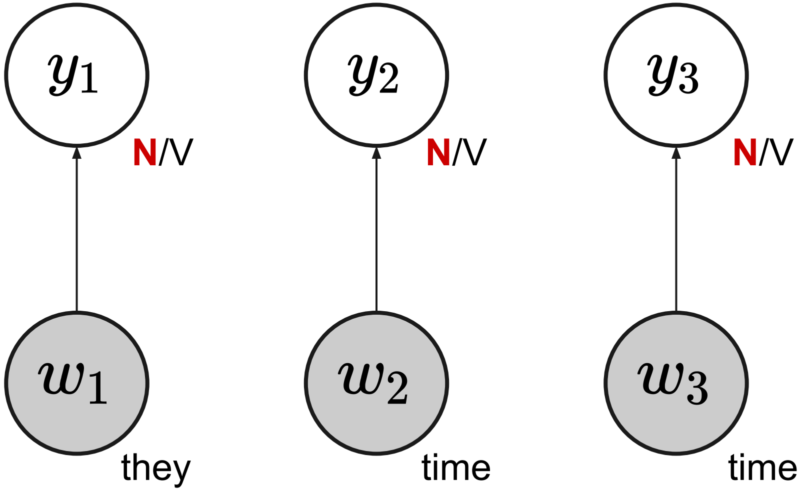


Training



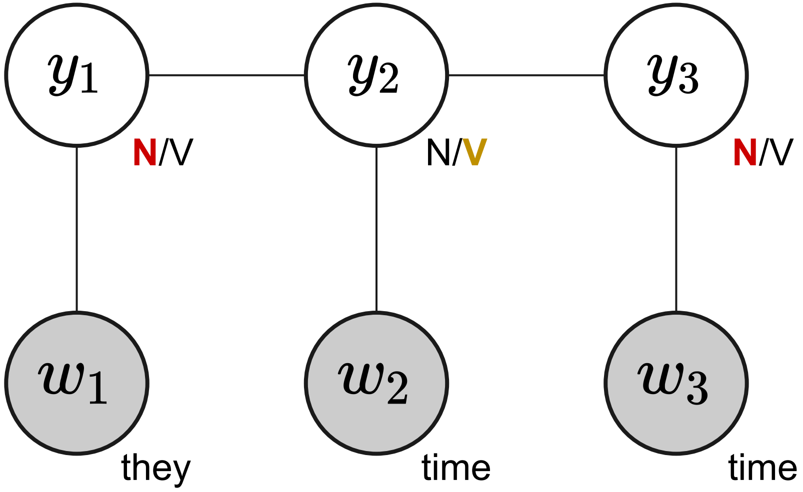
normalization is important but difficult in the sequence setup

Logistic regression



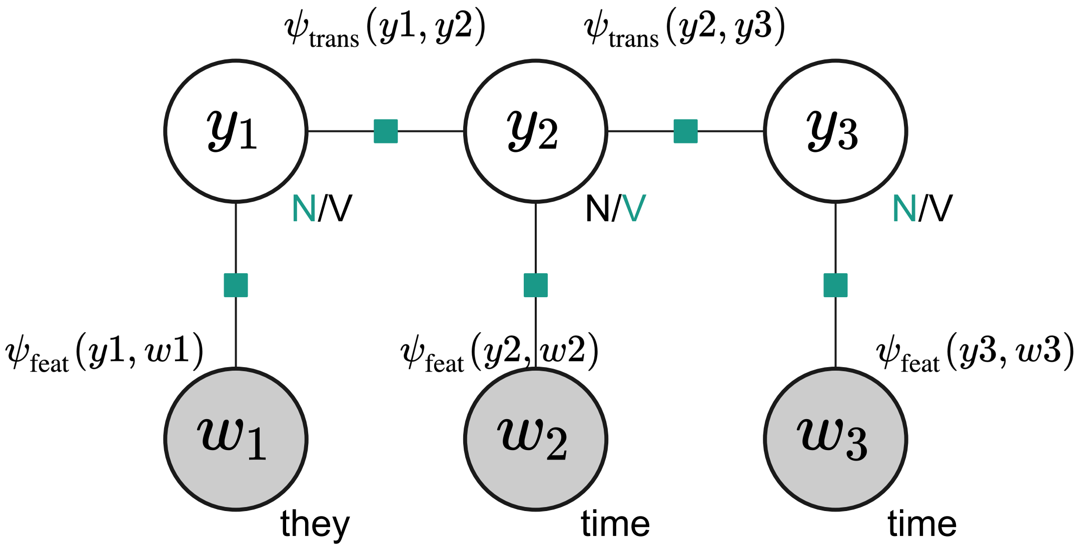
Predicting each individual tag with logistic regression is suboptimal

Conditional random fields



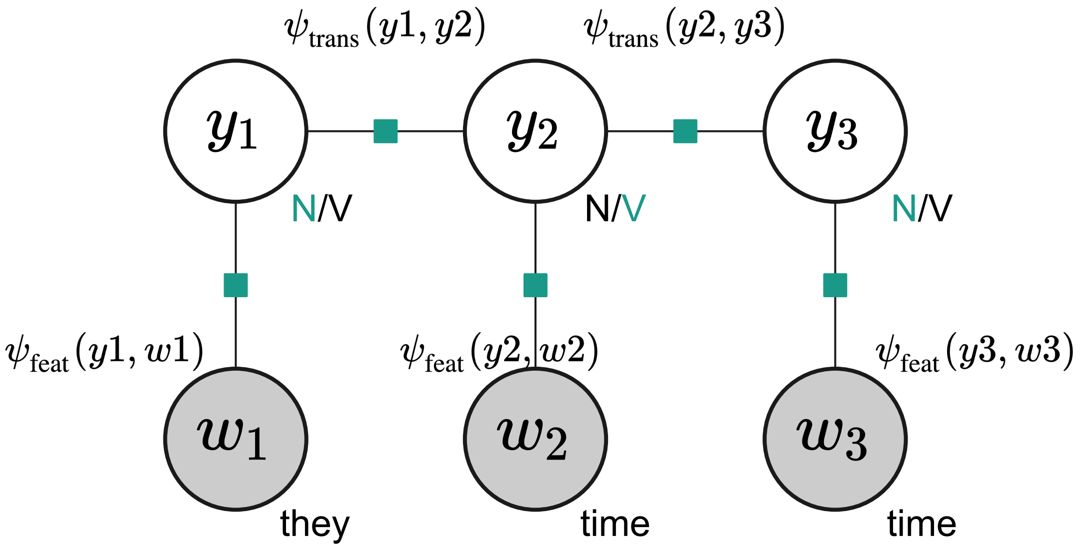
Incorporate structures between the labels

Conditional random fields



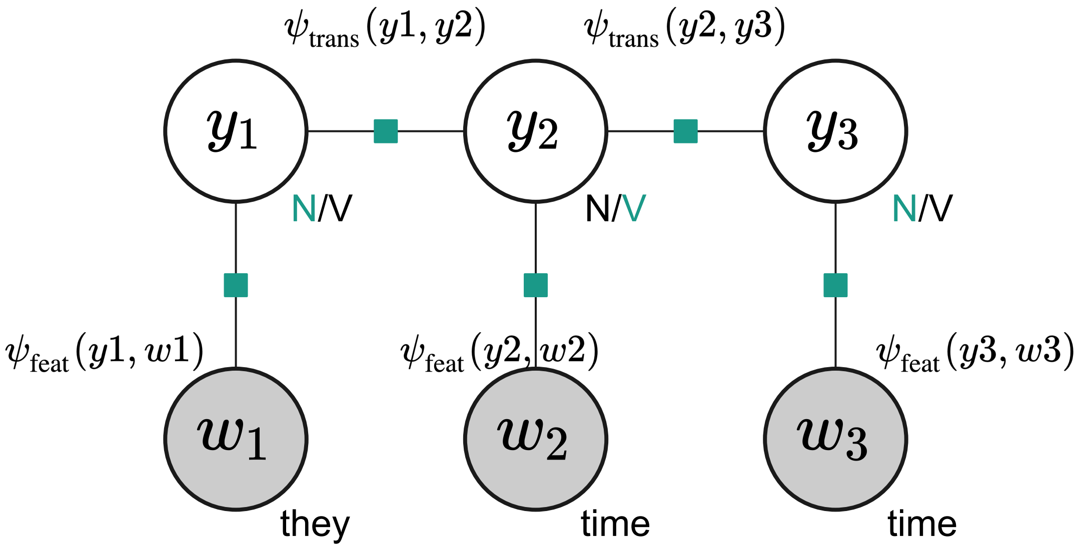
We define a series of scores ψ

Conditional random fields



These scores are similar to their counterparts in logistic regression: (0, +inf)

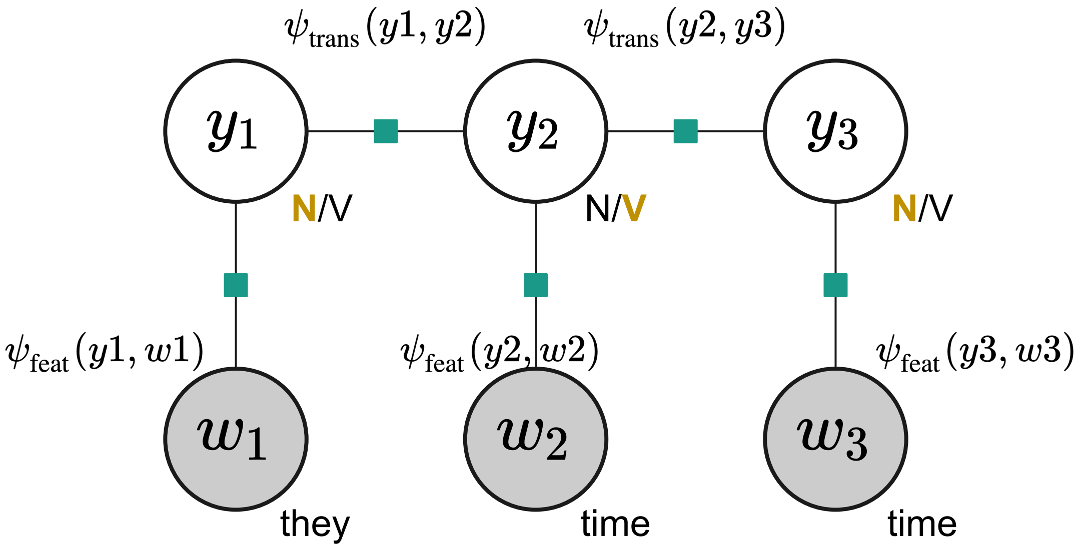
Conditional random fields



Again like in LR, these scores come from models with learnable parameters. In the homework:

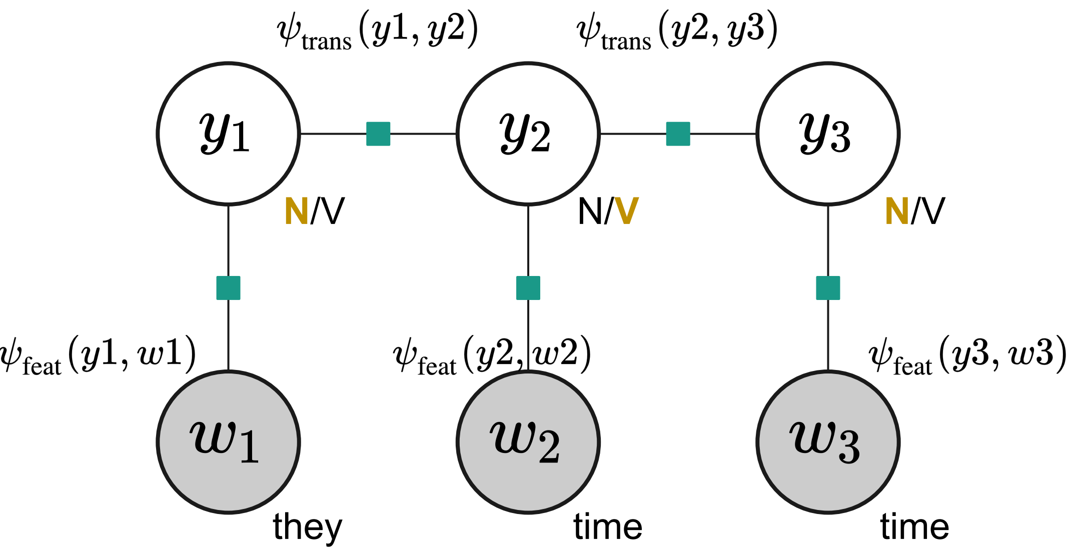
- ψ_{feat} is parameterized by a bidirectional LSTM
- ψ_{trans} is parameterized by a simple lookup table

Conditional random fields



The goal of training a CRF is to obtain the gold label sequence, and optimize the model parameters to maximize that sequence's probability.

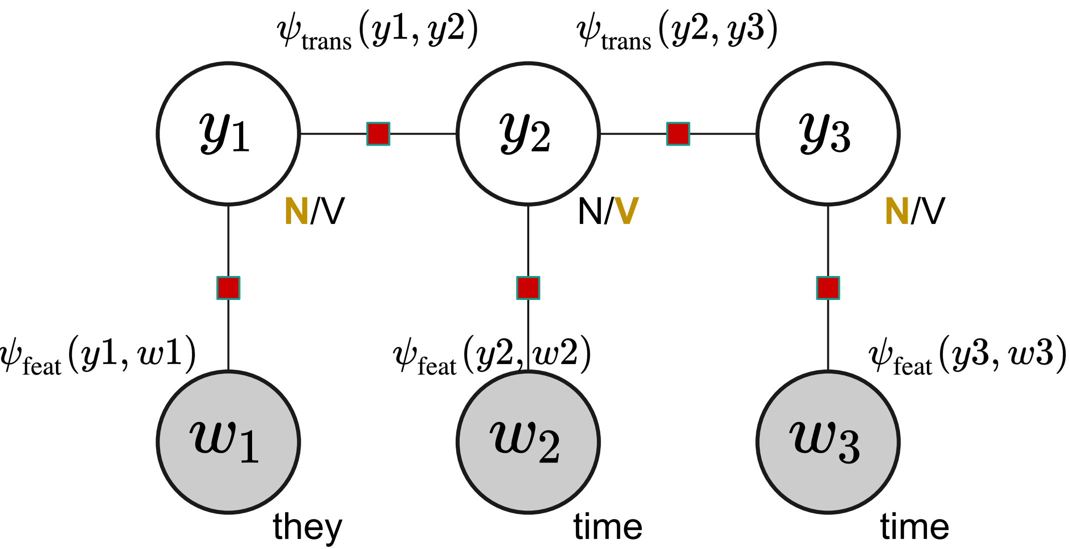
Conditional random fields



$$\psi(\cdot) \in (0, +\inf)$$

How to define the $\underbrace{\text{label sequence probability}}_{p_\theta(\boldsymbol{y} \mid \boldsymbol{w})}$?

Conditional random fields

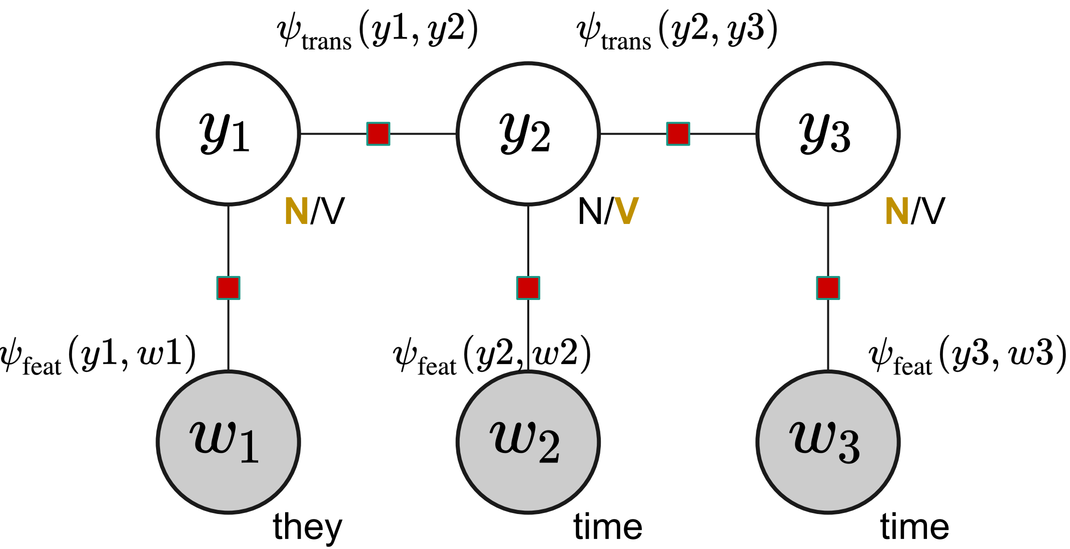


$$\psi(\cdot) \in (0, +\inf)$$

How to define the $\underbrace{\text{label sequence probability}}_{p_\theta(\boldsymbol{y} \mid \boldsymbol{w})}$?

$$p_\theta(\boldsymbol{y} \mid \boldsymbol{w}) \propto \prod_{n=1}^{N} \psi_{\text{feat}}(y_n, w_n)\psi_{\text{trans}}(y_n, y_{n+1})$$

Conditional random fields



$$\psi(\cdot) \in (0, +\inf)$$

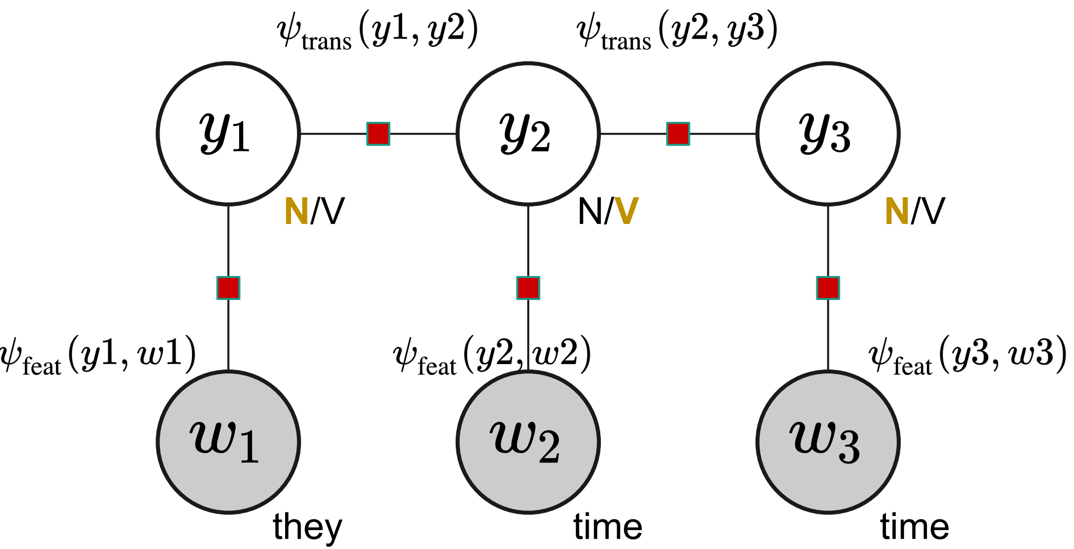
How to define the label sequence probability?

$$\underbrace{\text{label sequence probability}}_{p_\theta(\boldsymbol{y} \mid \boldsymbol{w})}$$

$$p_\theta(\boldsymbol{y} \mid \boldsymbol{w}) \propto \prod_{n=1}^{N} \psi_{\text{feat}}(y_n, w_n)\psi_{\text{trans}}(y_n, y_{n+1})$$

$$= \frac{\prod_{n=1}^{N} \psi_{\text{feat}}(y_n, w_n)\psi_{\text{trans}}(y_n, y_{n+1})}{\sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{w})} \prod_{n=1}^{N} \psi_{\text{feat}}(y'_n, w_n)\psi_{\text{trans}}(y'_n, y'_{n+1})}$$

Conditional random fields



$$\psi(\cdot) \in (0, +\inf)$$

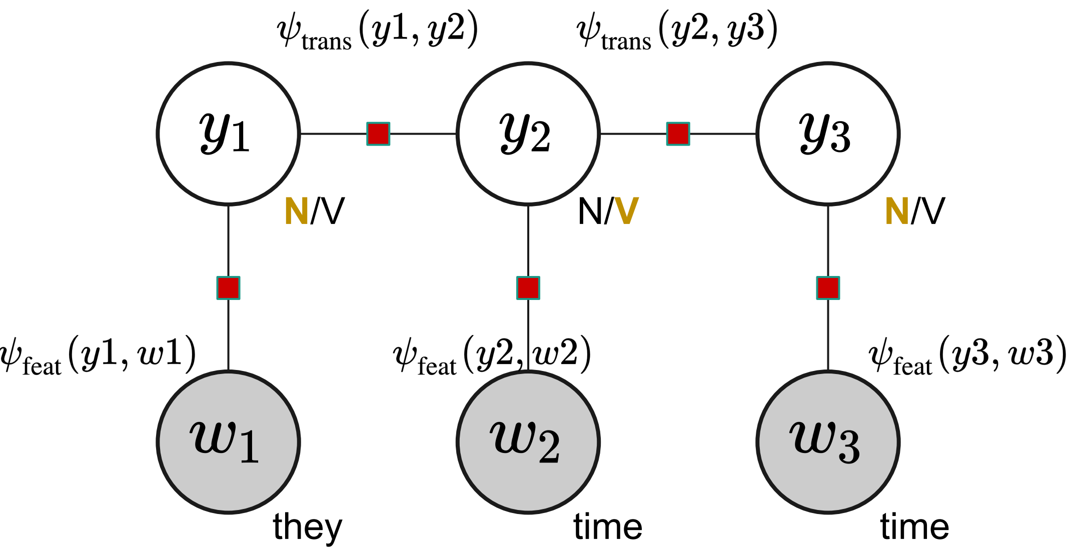
How to define the $\underbrace{\text{label sequence probability}}_{p_\theta(\boldsymbol{y} \mid \boldsymbol{w})}$?

$$p_\theta(\boldsymbol{y} \mid \boldsymbol{w}) \propto \prod_{n=1}^{N} \psi_{\text{feat}}(y_n, w_n)\psi_{\text{trans}}(y_n, y_{n+1})$$

$$= \frac{\prod_{n=1}^{N} \psi_{\text{feat}}(y_n, w_n)\psi_{\text{trans}}(y_n, y_{n+1})}{\sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{w})} \prod_{n=1}^{N} \psi_{\text{feat}}(y'_n, w_n)\psi_{\text{trans}}(y'_n, y'_{n+1})}$$

We are essentially doing softmax here, but the denominator is difficult to calculate!

Conditional random fields



$$\psi(\cdot) \in (0, +\inf)$$

How to define the label sequence probability?

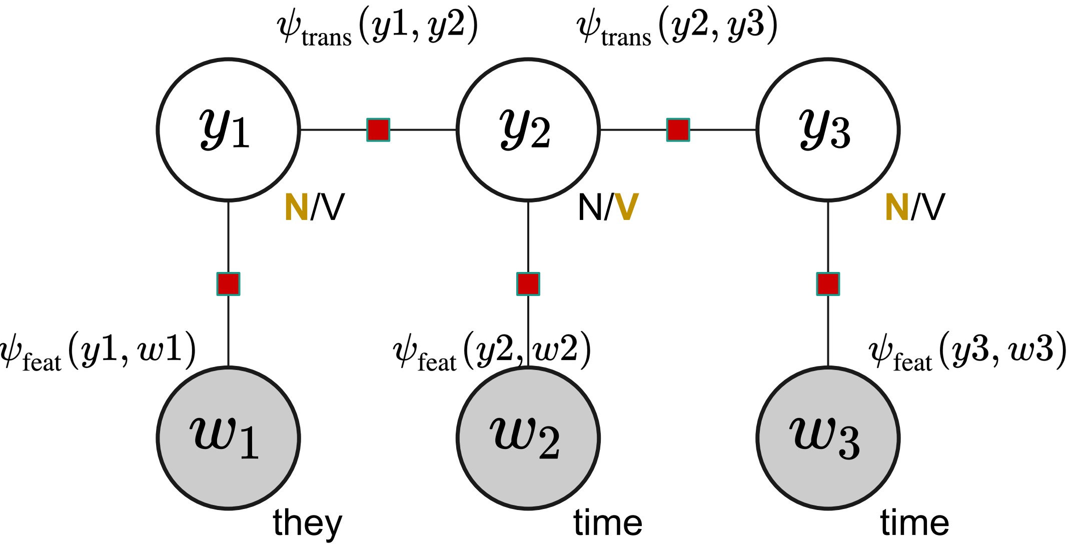
$$\underbrace{\text{label sequence probability}}_{p_\theta(\boldsymbol{y} \mid \boldsymbol{w})}$$

$$p_\theta(\boldsymbol{y} \mid \boldsymbol{w}) \propto \prod_{n=1}^{N} \psi_{\text{feat}}(y_n, w_n)\psi_{\text{trans}}(y_n, y_{n+1})$$

$$= \frac{\prod_{n=1}^{N} \psi_{\text{feat}}(y_n, w_n)\psi_{\text{trans}}(y_n, y_{n+1})}{\sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{w})} \prod_{n=1}^{N} \psi_{\text{feat}}(y'_n, w_n)\psi_{\text{trans}}(y'_n, y'_{n+1})}$$

We introduce the forward algorithm (a.k.a. sum-product algorithm) to obtain the denominator (a.k.a. partition function).

Conditional random fields



$$\psi(\cdot) \in (0, +\inf)$$

How to define the label sequence probability?

$$\underbrace{\text{the label sequence probability?}}_{p_\theta(\boldsymbol{y} \mid \boldsymbol{w})}$$

$$p_\theta(\boldsymbol{y} \mid \boldsymbol{w}) \propto \prod_{n=1}^{N} \psi_{\text{feat}}(y_n, w_n)\psi_{\text{trans}}(y_n, y_{n+1})$$

HW 6.2

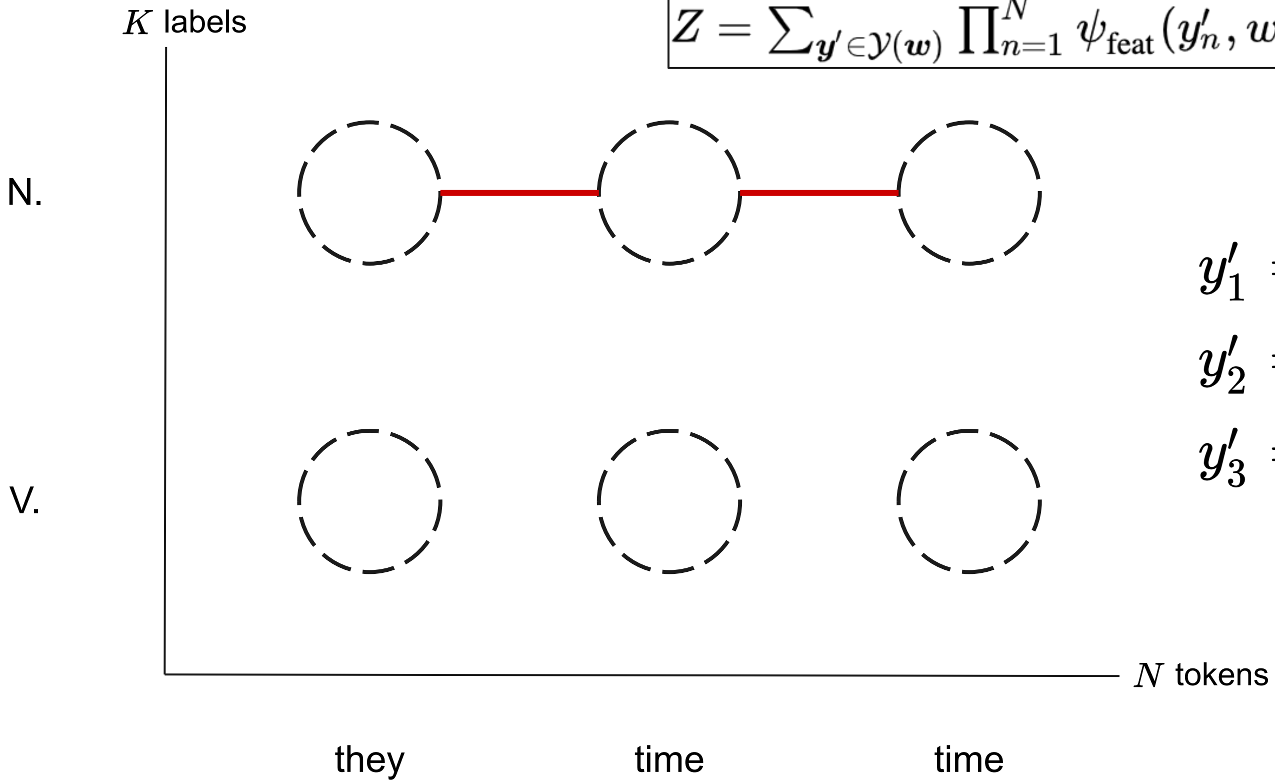
$$= \frac{\prod_{n=1}^{N} \psi_{\text{feat}}(y_n, w_n)\psi_{\text{trans}}(y_n, y_{n+1})}{\sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{w})} \prod_{n=1}^{N} \psi_{\text{feat}}(y'_n, w_n)\psi_{\text{trans}}(y'_n, y'_{n+1})}$$

HW 6.1

We introduce the forward algorithm (a.k.a. sum-product algorithm) to obtain the denominator (a.k.a. partition function).

Conditional random fields

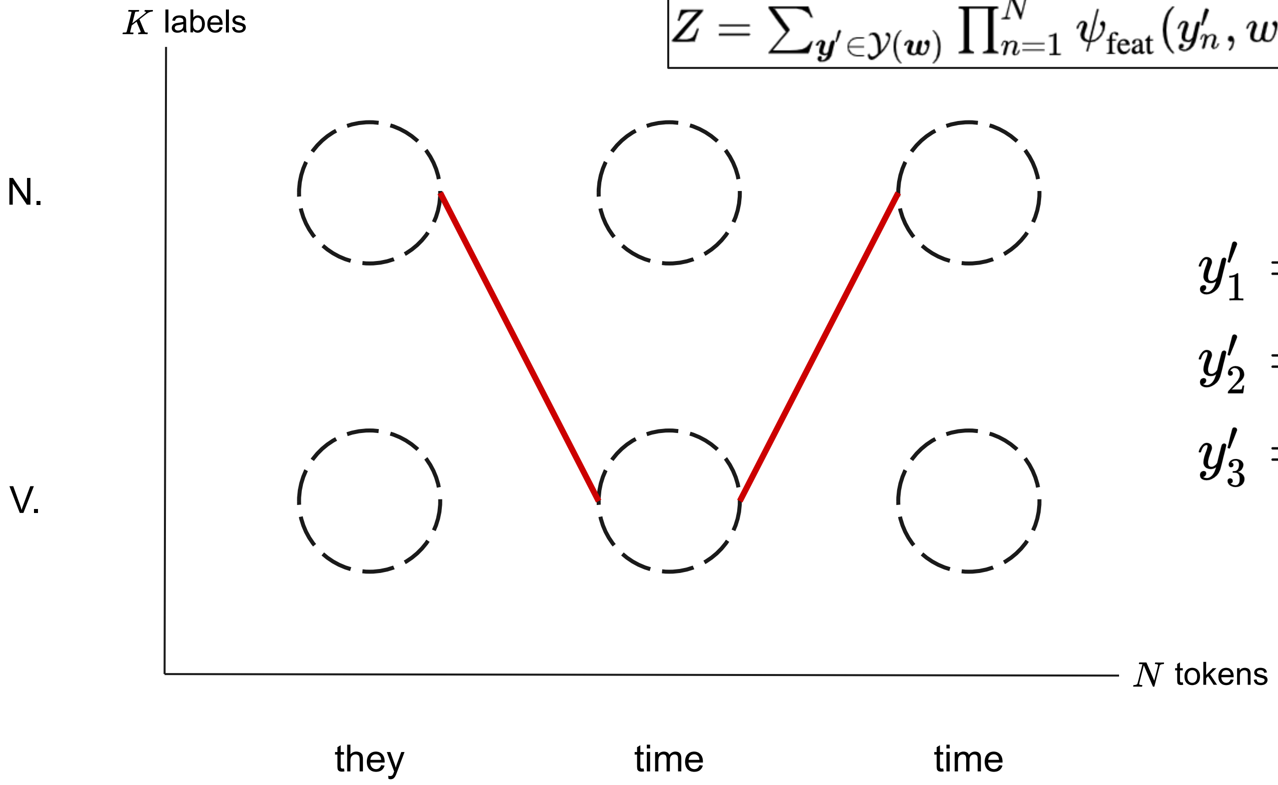
$$Z = \sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{w})} \prod_{n=1}^{N} \psi_{\text{feat}}(y'_n, w_n)\psi_{\text{trans}}(y'_n, y'_{n+1})$$



$$y'_1 = \text{NOUN}$$
$$y'_2 = \text{NOUN}$$
$$y'_3 = \text{NOUN}$$

Conditional random fields

$$Z = \sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{w})} \prod_{n=1}^{N} \psi_{\text{feat}}(y'_n, w_n)\psi_{\text{trans}}(y'_n, y'_{n+1})$$



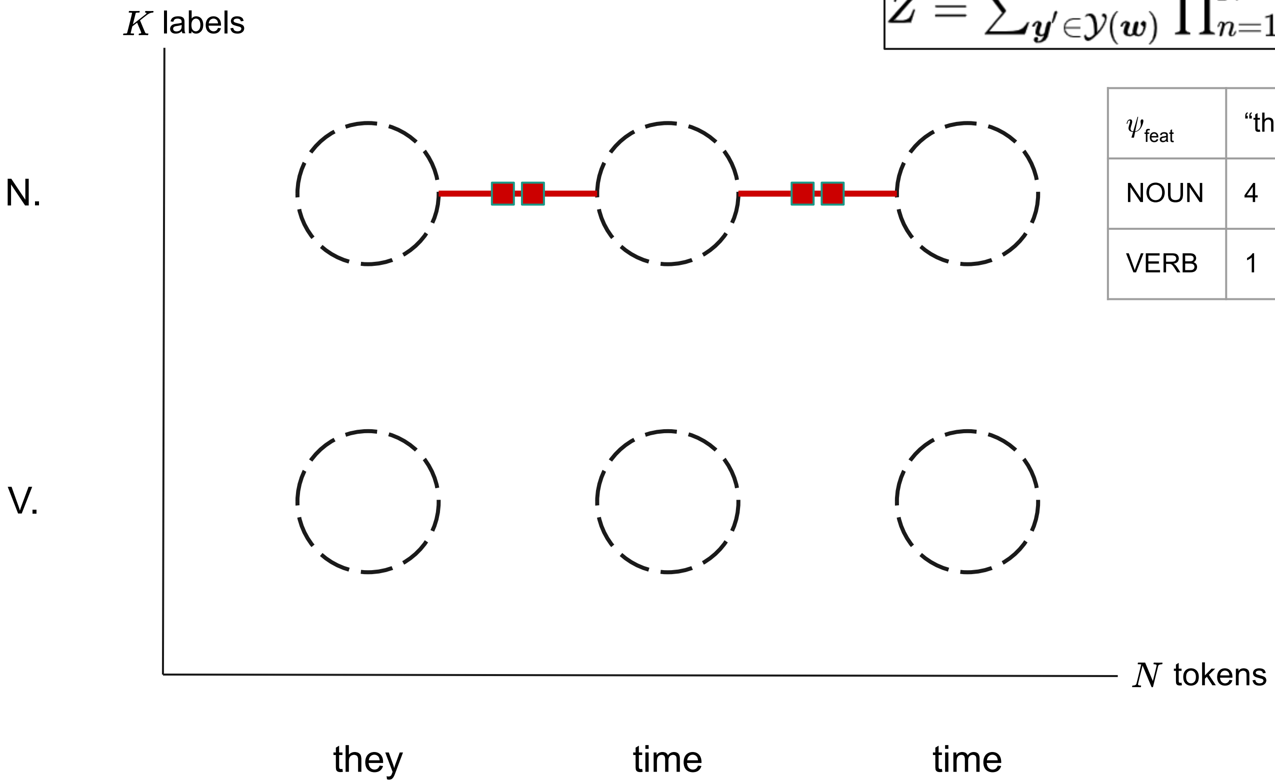
$$y'_1 = \text{NOUN}$$
$$y'_2 = \text{VERB}$$
$$y'_3 = \text{NOUN}$$

Conditional random fields

$$Z = \sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{w})} \prod_{n=1}^{N} \psi_{\text{feat}}(y'_n, w_n)\psi_{\text{trans}}(y'_n, y'_{n+1})$$

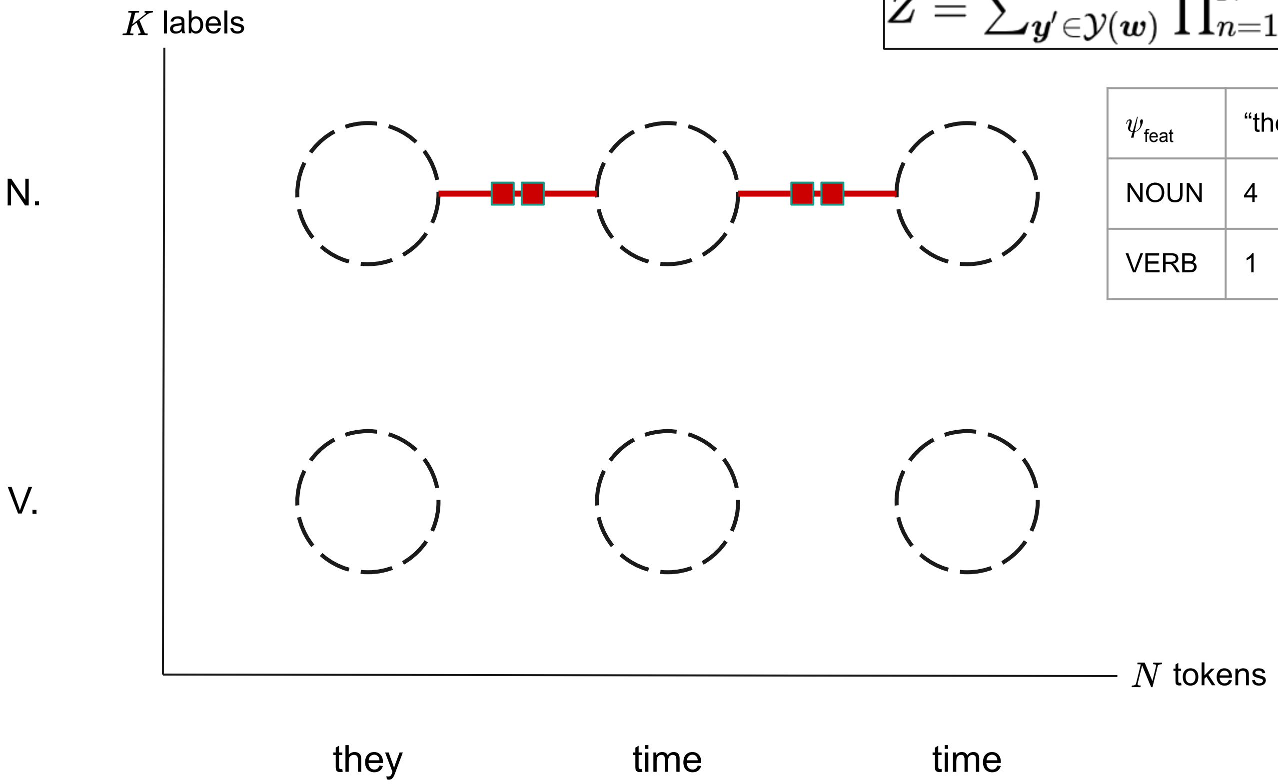
ψ_{feat}	“they”	“time”
NOUN	4	3
VERB	1	1

ψ_{trans}	NOUN	VERB
NOUN	2	3
VERB	3	1



Conditional random fields

$$Z = \sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{w})} \prod_{n=1}^{N} \psi_{\text{feat}}(y'_n, w_n)\psi_{\text{trans}}(y'_n, y'_{n+1})$$



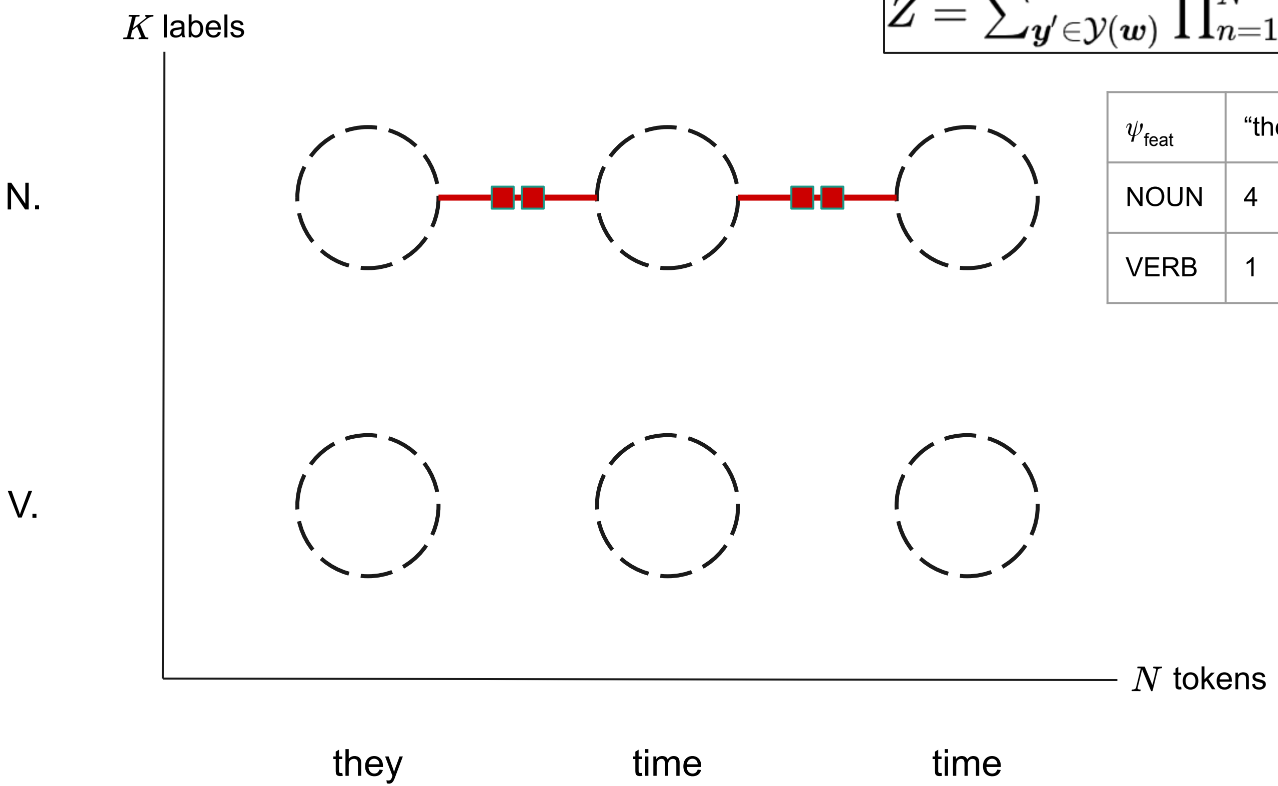
ψ_{feat}	"they"	"time"
NOUN	4	3
VERB	1	1

ψ_{trans}	NOUN	VERB
NOUN	2	3
VERB	3	1

$$\prod \psi(\cdot) = 4 \times 2 \times 3 \times 2 \times 3$$

Conditional random fields

$$Z = \sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{w})} \prod_{n=1}^{N} \psi_{\text{feat}}(y'_n, w_n)\psi_{\text{trans}}(y'_n, y'_{n+1})$$



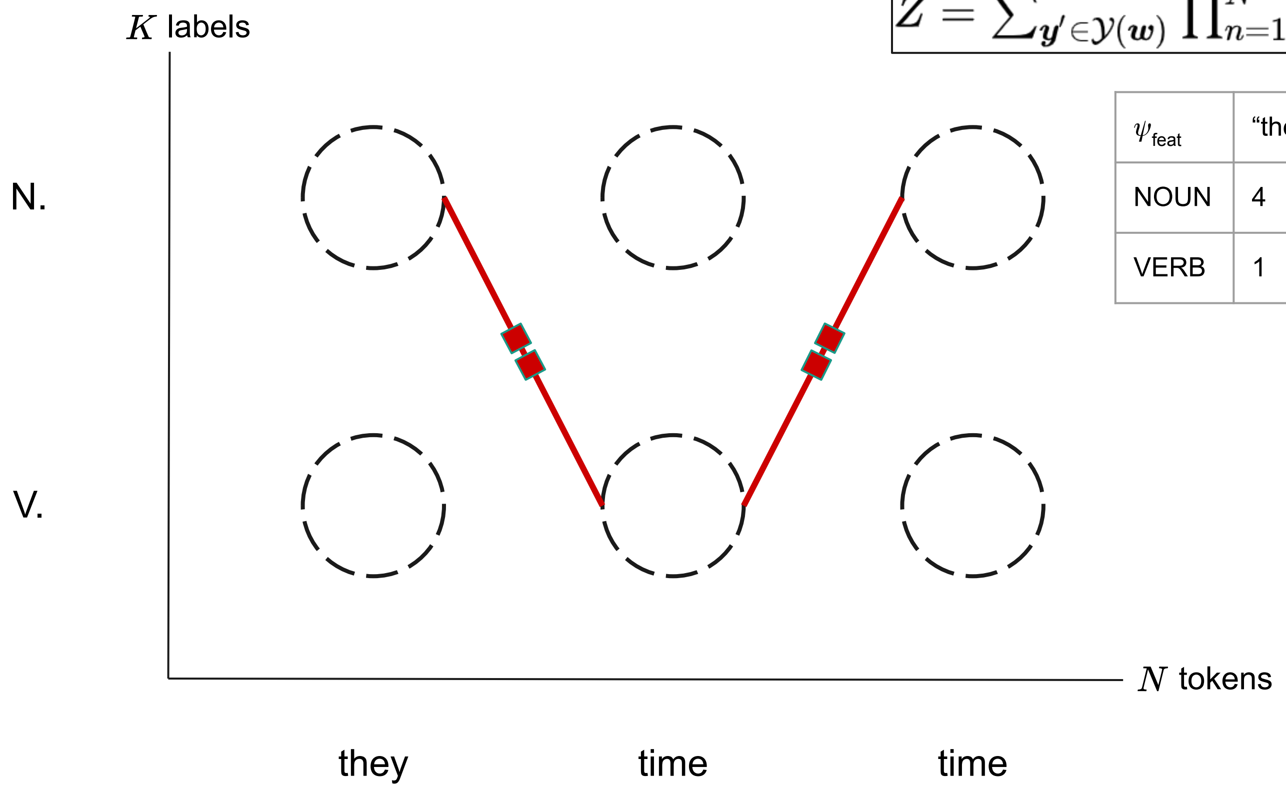
ψ_{feat}	"they"	"time"
NOUN	4	3
VERB	1	1

ψ_{trans}	NOUN	VERB
NOUN	2	3
VERB	3	1

$$\prod \psi(\cdot) = 4 \times 2 \times 3 \times 2 \times 3$$

Conditional random fields

$$Z = \sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{w})} \prod_{n=1}^{N} \psi_{\text{feat}}(y'_n, w_n)\psi_{\text{trans}}(y'_n, y'_{n+1})$$

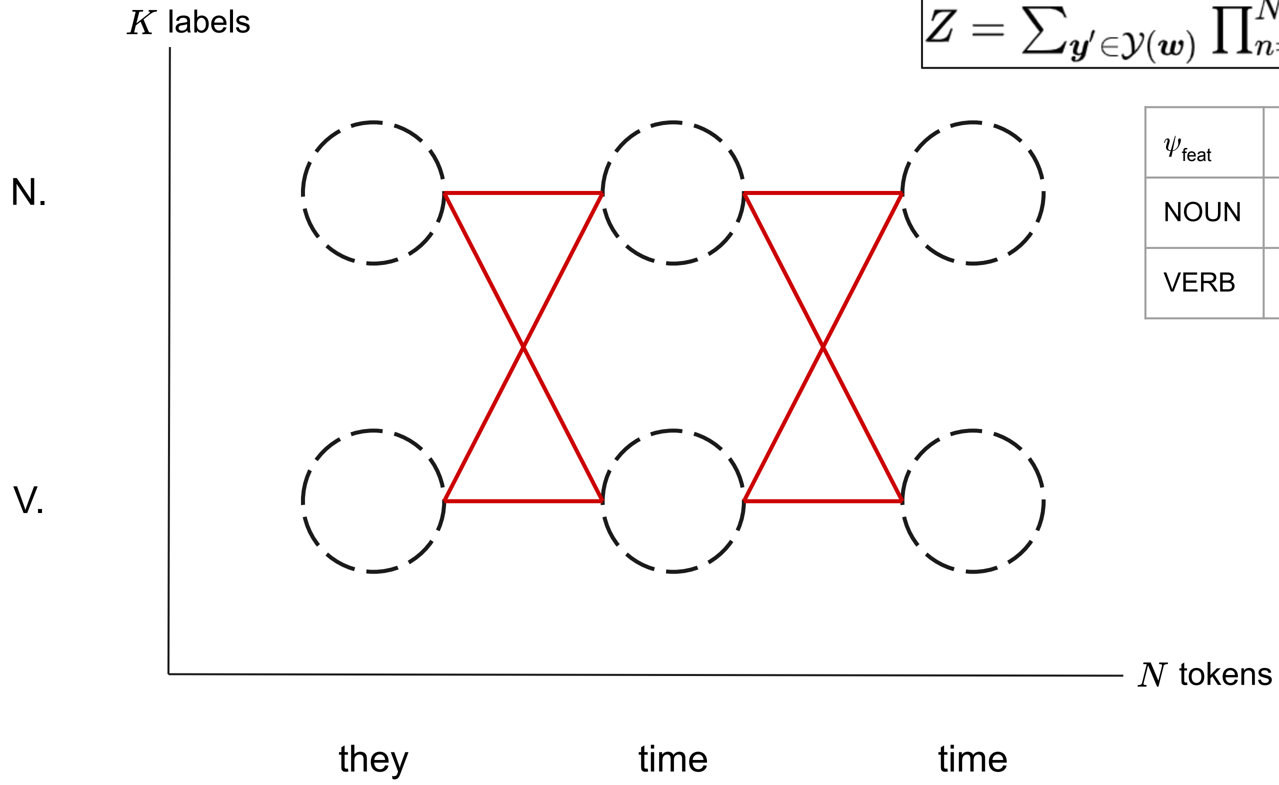


ψ_{feat}	"they"	"time"
NOUN	4	3
VERB	1	1

ψ_{trans}	NOUN	VERB
NOUN	2	3
VERB	3	1

$$\prod \psi(\cdot) = 4 \times 3 \times 1 \times 3 \times 3$$

Conditional random fields



$$Z = \sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{w})} \prod_{n=1}^{N} \psi_{\text{feat}}(y'_n, w_n) \psi_{\text{trans}}(y'_n, y'_{n+1})$$

ψ_{feat}	"they"	"time"
NOUN	4	3
VERB	1	1

ψ_{trans}	NOUN	VERB
NOUN	2	3
VERB	3	1

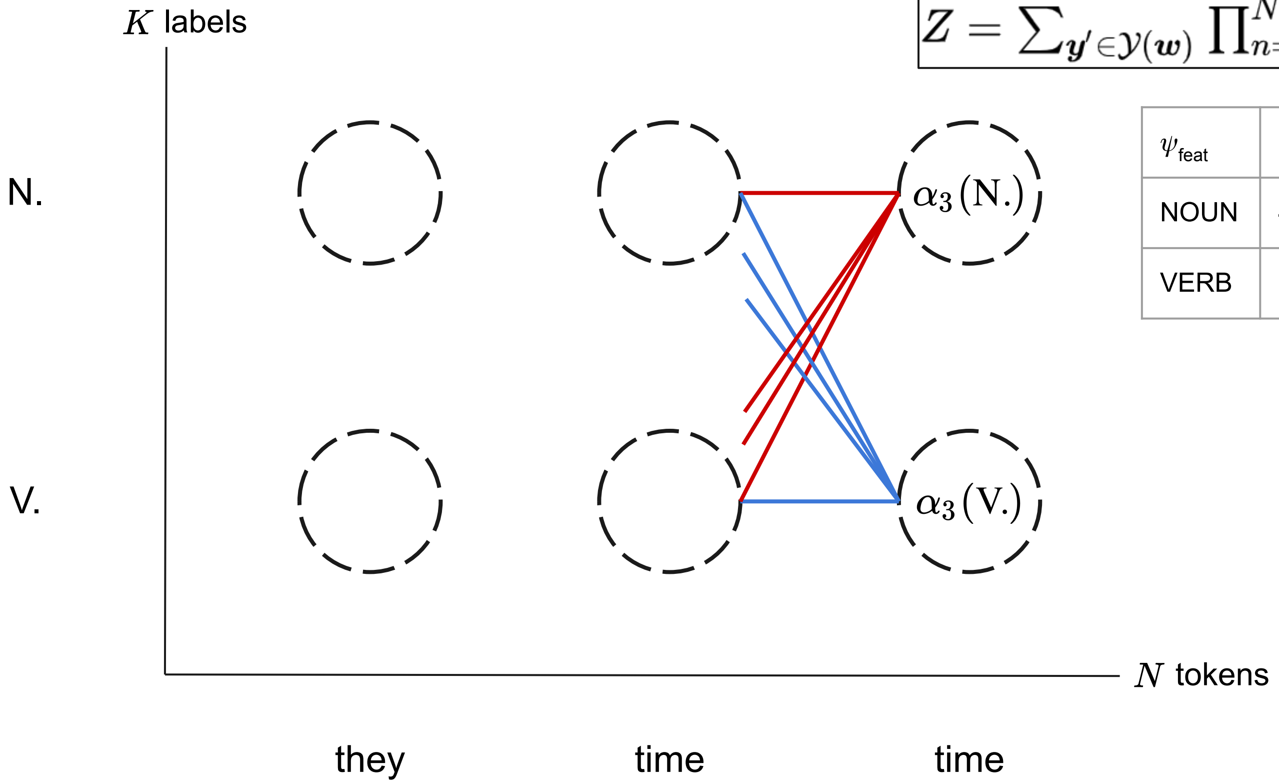
$$\prod \psi(\cdot) = 4 \times 2 \times 3 \times 2 \times 3$$

⋮

(we want their sum)

Conditional random fields

$$\boxed{Z = \sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{w})} \prod_{n=1}^{N} \psi_{\text{feat}}(y'_n, w_n)\psi_{\text{trans}}(y'_n, y'_{n+1})}$$



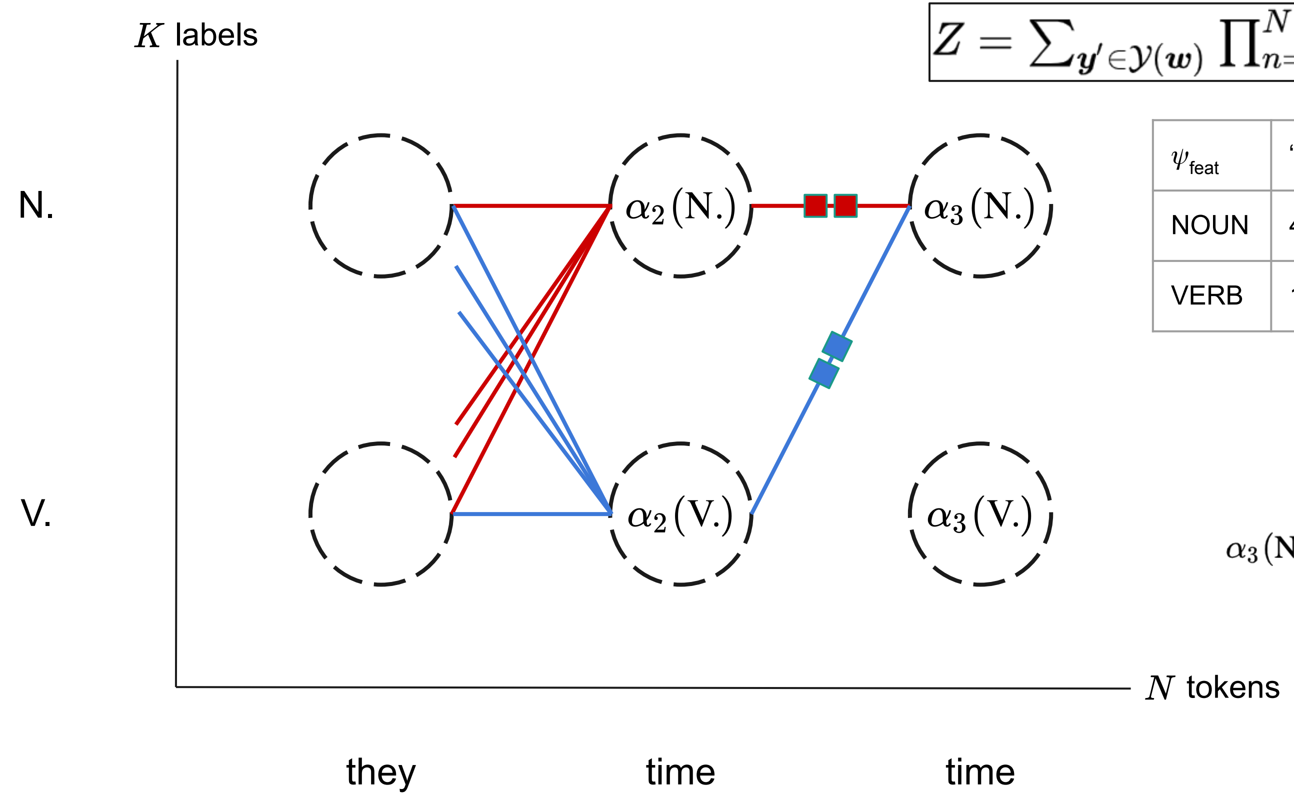
ψ_{feat}	"they"	"time"
NOUN	4	3
VERB	1	1

ψ_{trans}	NOUN	VERB
NOUN	2	3
VERB	3	1

Here the alpha variable denotes the sum of all paths ***ending*** at the cell.

$$Z = \alpha_3(\text{N.}) + \alpha_3(\text{N.})$$

Conditional random fields



$$Z = \sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{w})} \prod_{n=1}^{N} \psi_{\text{feat}}(y'_n, w_n)\psi_{\text{trans}}(y'_n, y'_{n+1})$$

ψ_{feat}	"they"	"time"
NOUN	4	3
VERB	1	1

ψ_{trans}	NOUN	VERB
NOUN	2	3
VERB	3	1

We can compute each of the alpha variable recursively.

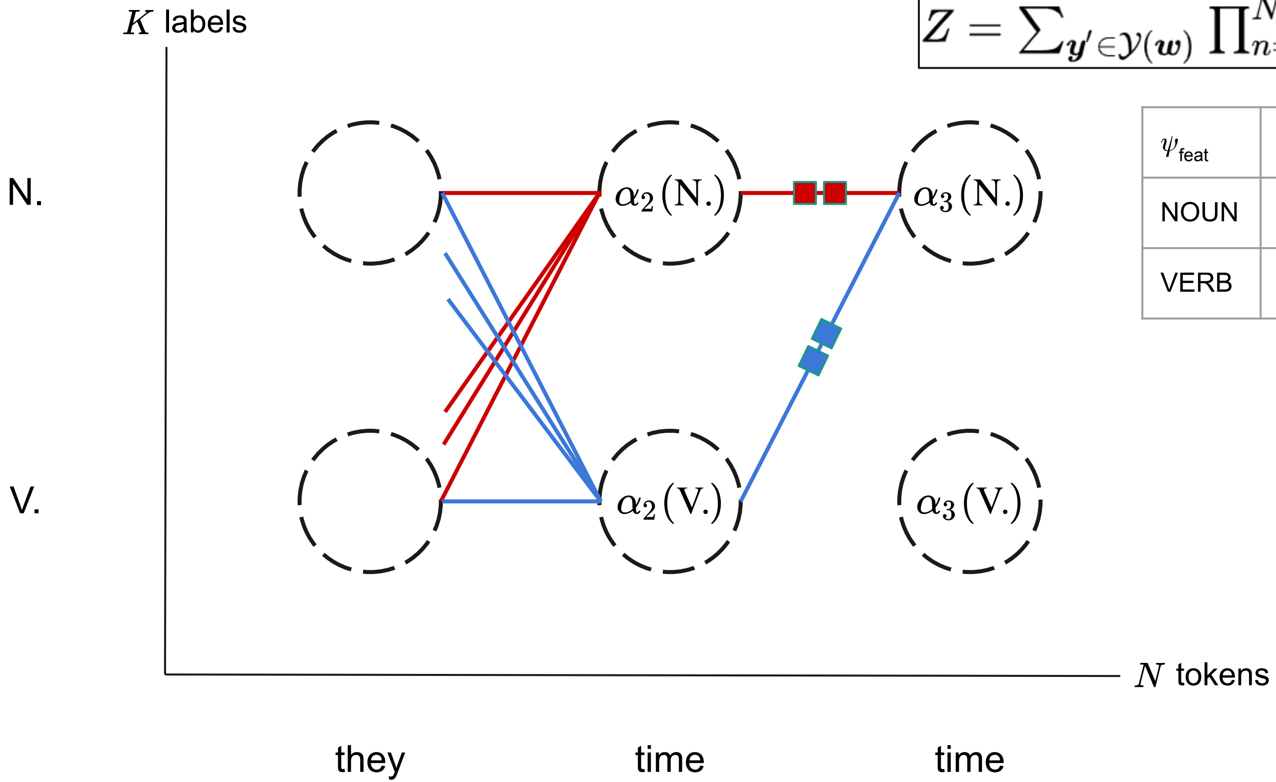
$$\alpha_3(\text{N.}) = \alpha_2(\text{N.})\psi_{\text{trans}}(\text{N.}, \text{N.})\psi_{\text{feat}}(\text{N.}, \text{"time"}) + \alpha_2(\text{V.})\psi_{\text{trans}}(\text{V.}, \text{N.})\psi_{\text{feat}}(\text{N.}, \text{"time"})$$

Conditional random fields

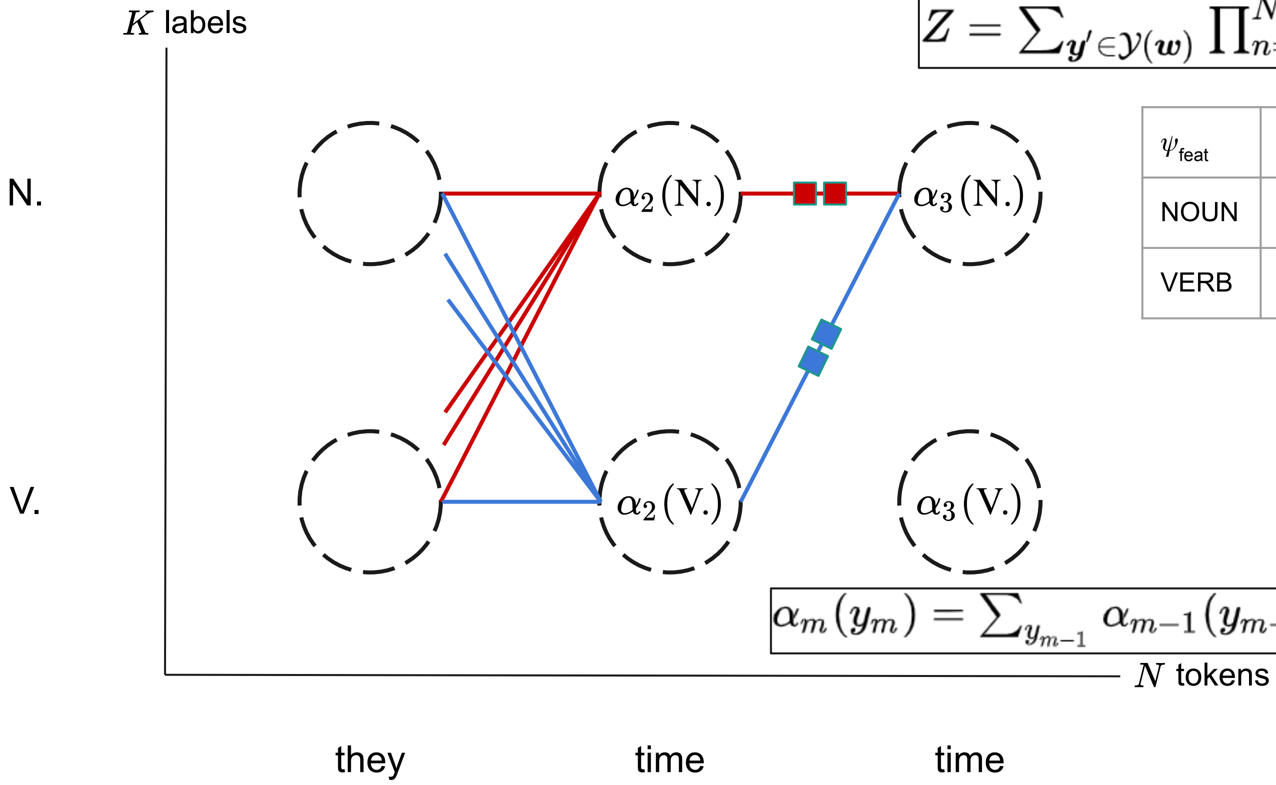
$$Z = \sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{w})} \prod_{n=1}^{N} \psi_{\text{feat}}(y'_n, w_n)\psi_{\text{trans}}(y'_n, y'_{n+1})$$

ψ_{feat}	"they"	"time"
NOUN	4	3
VERB	1	1

ψ_{trans}	NOUN	VERB
NOUN	2	3
VERB	3	1



Conditional random fields



$$Z = \sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{w})} \prod_{n=1}^{N} \psi_{\text{feat}}(y'_n, w_n)\psi_{\text{trans}}(y'_n, y'_{n+1})$$

ψ_{feat}	"they"	"time"
NOUN	4	3
VERB	1	1

ψ_{trans}	NOUN	VERB
NOUN	2	3
VERB	3	1

$$\alpha_m(y_m) = \sum_{y_{m-1}} \alpha_{m-1}(y_{m-1})\psi_{\text{trans}}(y_{m-1}, y_m)\psi_{\text{feat}}(y_m, w_m)$$

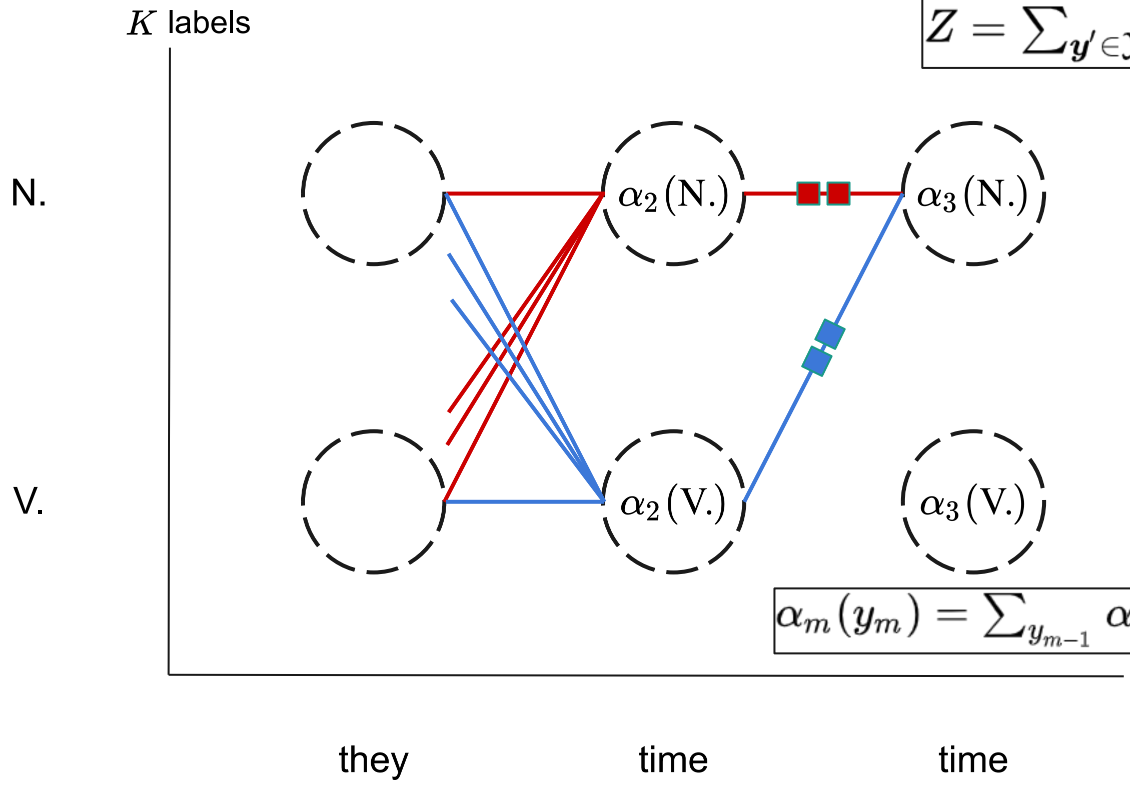
for $m = 1, 2, \ldots, N$

Conditional random fields

$$Z = \sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{w})} \prod_{n=1}^{N} \psi_{\text{feat}}(y'_n, w_n)\psi_{\text{trans}}(y'_n, y'_{n+1})$$

ψ_{feat}	"they"	"time"
NOUN	4	3
VERB	1	1

ψ_{trans}	NOUN	VERB
NOUN	2	3
VERB	3	1

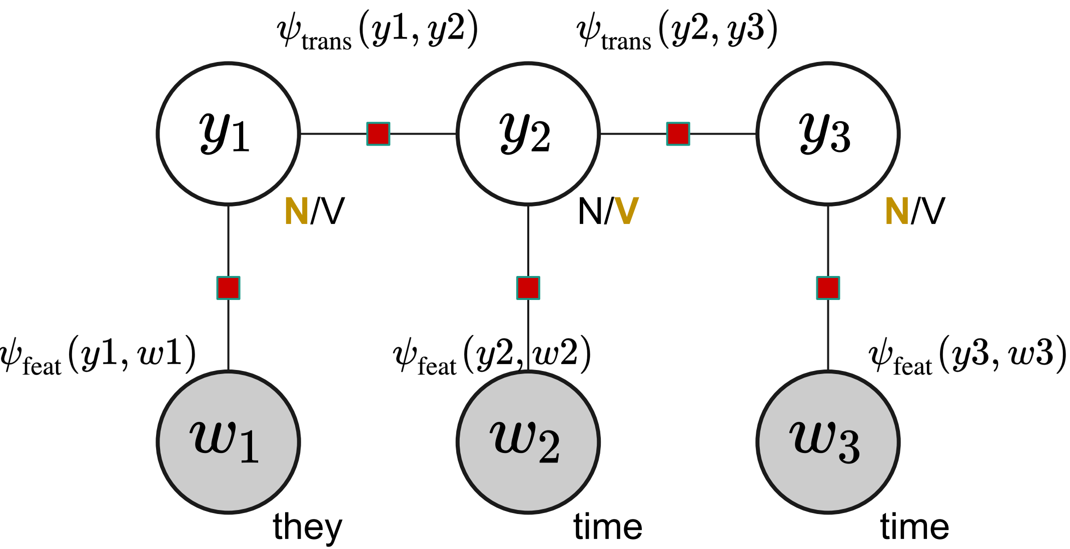


Time complexity?
Space complexity?

$$\alpha_m(y_m) = \sum_{y_{m-1}} \alpha_{m-1}(y_{m-1})\psi_{\text{trans}}(y_{m-1}, y_m)\psi_{\text{feat}}(y_m, w_m)$$

for $m = 1, 2, \ldots, N$

Conditional random fields



$$\psi(\cdot) \in (0, +\inf)$$

How to define the label sequence probability?

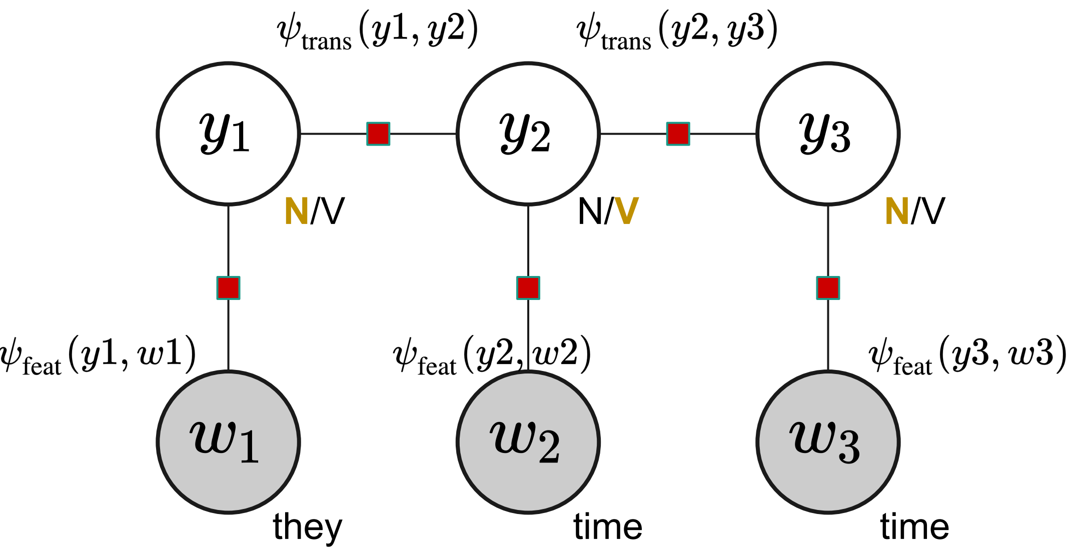
$$\underbrace{\text{label sequence probability?}}_{p_\theta(\boldsymbol{y} \mid \boldsymbol{w})}$$

$$p_\theta(\boldsymbol{y} \mid \boldsymbol{w}) \propto \prod_{n=1}^{N} \psi_{\text{feat}}(y_n, w_n)\psi_{\text{trans}}(y_n, y_{n+1})$$

$$= \frac{\prod_{n=1}^{N} \psi_{\text{feat}}(y_n, w_n)\psi_{\text{trans}}(y_n, y_{n+1})}{\sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{w})} \prod_{n=1}^{N} \psi_{\text{feat}}(y'_n, w_n)\psi_{\text{trans}}(y'_n, y'_{n+1})}$$

$$= \text{something computable!!!}$$

Conditional random fields



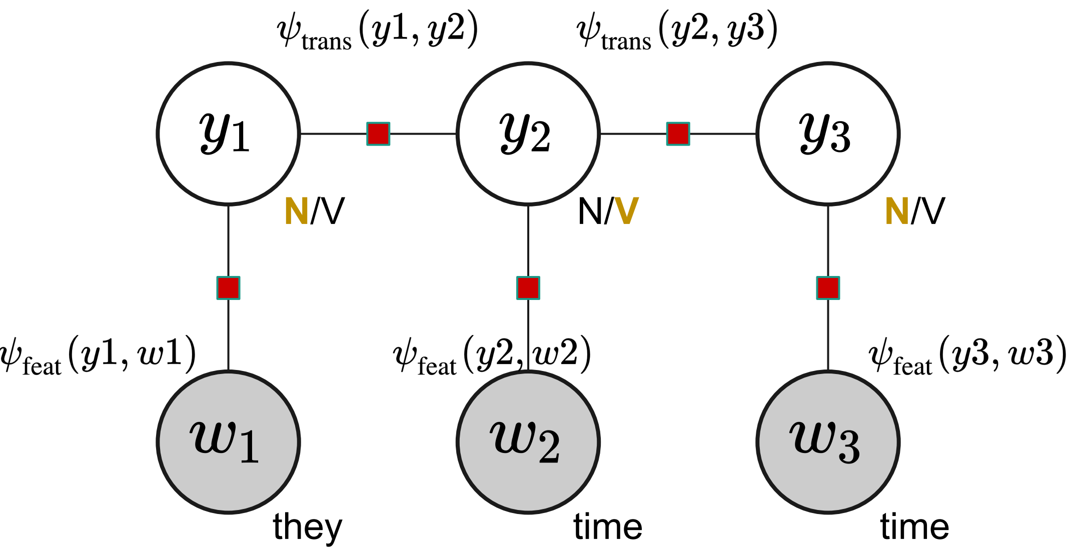
$$\psi(\cdot) \in (0, +\inf)$$

How to maximize the $\underbrace{\text{gold sequence probability}}_{p_\theta(\boldsymbol{y} \mid \boldsymbol{w})}$?

$$\theta \leftarrow \theta - \eta \nabla_\theta (-\log p_\theta(\boldsymbol{y} \mid \boldsymbol{w}))$$

Gradient descent, or any of your favorite optimizers :)

Conditional random fields



$$\psi(\cdot) \in (0, +\inf)$$

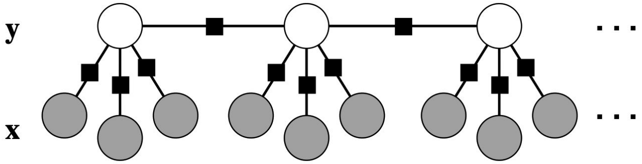
How to do inference on test-time inputs with the learned model?

The Viterbi algorithm, a.k.a. max-product algorithm
(This part is the same as HMMs)

Further details can be found in Chapter 7 of the Eisenstein textbook here.

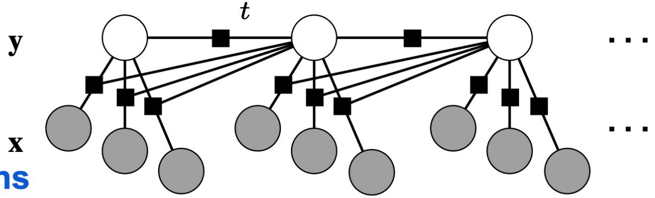
Other types of CRF

Direct Extension of HMM



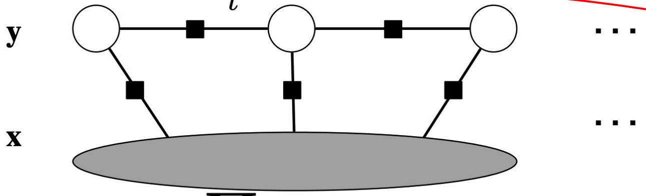
$$p(y \mid x) \propto \prod_t \psi_t(y_t, x_t)\psi_{t,t+1}(y_t, y_{t+1})$$

State Transitions depend on Observations



$$p(y \mid x) \propto \prod_t \psi_t(y_t, x_t)\psi_{t,t+1}(y_t, y_{t+1}, x_t)$$

Arbitrary Non-Local Features



$$p(y \mid x) \propto \prod_t \psi_t(y_t, x)\psi_{t,t+1}(y_t, y_{t+1}, x)$$

source