TTIC 31210:
Advanced Natural Language Processing

Kevin Gimpel

Spring 2019

Lecture 7:
Structured Prediction 1

Roadmap

- intro (1 lecture)
- deep learning for NLP (5 lectures)
- **structured prediction (4 lectures)**
  - **introducing/formalizing structured prediction, categories of structures**
  - **inference: dynamic programming, greedy algorithms, beam search**
  - **inference with non-local features**
  - **learning in structured prediction**
- generative models, latent variables, unsupervised learning, variational autoencoders (2 lectures)
- Bayesian methods in NLP (2 lectures)
- Bayesian nonparametrics in NLP (2 lectures)
- review & other topics (1 lecture)

Assignments

- we will briefly go over Assignment 1 today
- Assignment 2 was posted last week, due May 1st
- reminder: for those graduating this quarter, Assignment 5 is optional

What is Structured Prediction?

Classifiers

- a function from inputs ***x*** to outputs ***y***
- one simple type of classifier:
  - for any input ***x***, assign a score to each output ***y***, parameterized by parameters ***w***:

$$\text{score}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta})$$

  - classify by choosing highest-scoring output:

$$\text{classify}(\boldsymbol{x}, \boldsymbol{\theta}) = \underset{\boldsymbol{y}}{\text{argmax}} \ \text{score}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta})$$

Notation

$\mathbf{u}$ = a vector

u_i = entry *i* in the vector

$\mathbf{W}$ = a matrix

w_{ij} = entry (*i*,*j*) in the matrix

$\boldsymbol{x}$ = a structured object

x_i = item *i* in the structured object

Modeling, Inference, Learning

inference: solve argmax

modeling: define score function

$$\text{classify}(\boldsymbol{x}, \boldsymbol{\theta}) = \underset{\boldsymbol{y}}{\text{argmax}} \ \text{score}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta})$$

learning: choose $\boldsymbol{\theta}$

Applications of our Classifier Framework

task	input (x)	output (y)	output space ($\mathcal{L}$)	size of $\mathcal{L}$
text classification	a sentence	gold standard label for $\boldsymbol{x}$	pre-defined, small label set (e.g., {positive, negative})	2-10
word sense disambiguation	instance of a particular word (e.g., *bass*) with its context	gold standard word sense of target word	pre-defined sense inventory from WordNet for *bass*	2-30
learning skip-gram word embeddings	instance of a word in a corpus	a word in the context of x in a corpus	vocabulary	$\|V\|$
part-of-speech tagging	a sentence	gold standard part-of-speech tags for $\boldsymbol{x}$	all possible part-of-speech tag sequences with same length as $\boldsymbol{x}$	$\|P\|^{\|\boldsymbol{x}\|}$

Applications of our Classifier Framework

task	input (x)	output (y)	output space ($\mathcal{L}$)	size of $\mathcal{L}$
text classification	a sentence	gold standard label for x	pre-defined, small label set (e.g., {positive, negative})	2-10
word sense disambiguation	instance of a particular word (e.g., *bass* ... its cont...	gold standard word sense of ...	pre-defined sense inventory from ...	2-30
learning skip-gram word embeddings	instance ... word in a c...			
part-of-speech tagging	a sentence	gold standard part-of-speech tags for x	all possible part-of-speech tag sequences with same length as x	$\lvert P \rvert^{\lvert x \rvert}$

exponential in size of input!
"structured prediction"

Applications of Classifier Framework (continued)

task	input (*x*)	output (*y*)	output space ($\mathcal{L}$)	size of $\mathcal{L}$
named entity recognition	a sentence	gold standard named entity labels for ***x*** (BIO tags)	all possible BIO label sequences with same length as ***x***	$\lvert P \rvert^{\lvert \boldsymbol{x} \rvert}$
constituency parsing	a sentence	gold standard constituent parse (labeled bracketing) of ***x***	all possible labeled bracketings of ***x***	exponential in length of ***x*** (Catalan number)
dependency parsing	a sentence	gold standard dependency parse (labeled directed spanning tree) of ***x***	all possible labeled directed spanning trees of ***x***	exponential in length of ***x***
machine translation	a sentence	a translation of ***x***	all possible translations of ***x***	potentially infinite

Modeling, Inference, Learning

inference: solve argmax

modeling: define score function

$$\mathrm{classify}(\boldsymbol{x}, \boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{y}} \ \mathrm{score}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta})$$

learning: choose $\boldsymbol{\theta}$

Working definition of structured prediction:

size of output space is exponential in size of input

or is unbounded (e.g., machine translation)

(we can't just enumerate all possible outputs)

What is Structured Prediction?

- however, just because the output is a structured object does not necessarily mean we are doing “structured prediction”
- we can model many structured output spaces with traditional “local” or “unstructured” predictors
- today we will aim to make this more formal
- in short, we may be predicting structures but we might not necessarily be using a “structured predictor”

Example NLP Tasks

- we’ll go through some examples of NLP tasks that involve predicting output structures

Sequence Labeling

(e.g., Part-of-Speech Tagging)

determiner	verb (past)	prep.	proper noun	proper noun	poss.	adj.	noun
Some	questioned	if	Tim	Cook	’s	first	product

modal	verb	det.	adjective	noun	prep.	proper noun	punc.
would	be	a	breakaway	hit	for	Apple	.

Unlabeled Segmentations
(Chinese Word Segmentation)

- some languages are written without whitespace
- task: insert spaces to form "words"
  - 莎拉波娃现在居住在美国东南部的佛罗里达。
  - 莎拉波娃 现在 居住 在 美国 东南部 的 佛罗里达
  - Sharapova now lives in US southeastern Florida

Labeled Segmentations

(Named Entity Recognition)

Some questioned if Tim Cook's first product would be a breakaway hit for Apple.

PERSON (Tim Cook) ORGANIZATION (Apple)

Labeled Segmentations

(Entity Linking)

Some questioned if Tim Cook's first product would be a breakaway hit for Apple.

Labeled Segmentation as Sequence Labeling

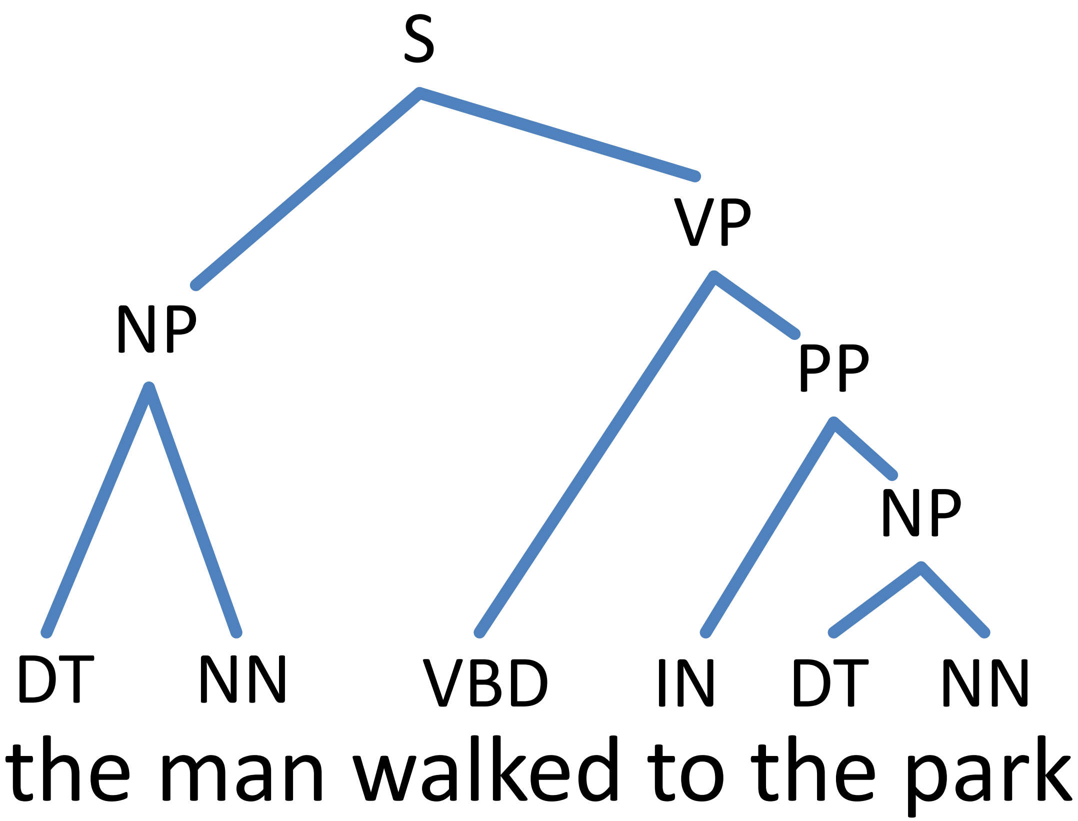
O	O	O	B-PERSON	I-PERSON	O	O	O
Some	questioned	if	Tim	Cook	’s	first	product

O	O	O	O	O	O	B-ORGANIZATION	O
would	be	a	breakaway	hit	for	Apple	.

B = “begin”
I = “inside”
O = “outside”

Trees
(Constituency Parsing)

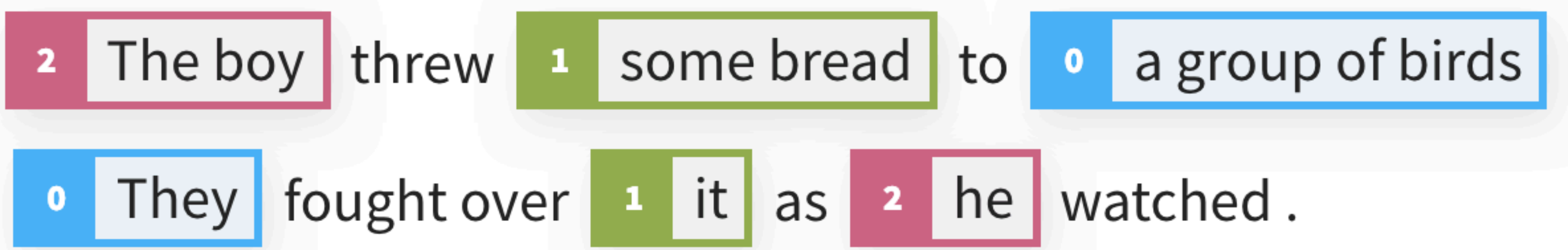
(S (NP the man) (VP walked (PP to (NP the park))))



Key:
S = sentence
NP = noun phrase
VP = verb phrase
PP = prepositional phrase
DT = determiner
NN = noun
VBD = verb (past tense)
IN = preposition

Unlabeled Segmentation + Clustering
(Coreference Resolution)

The boy threw some bread to a group of birds . They fought over it as he watched .



Generation

- there are many language generation tasks that involve predicting a phrase, sentence, document, or some other textual sequence

Answers
(Question Answering)

amazon alexa

“Alexa, who was President when Barack Obama was nine?”

“Alexa, how’s my commute?

“Alexa, what’s the weather?”

“Alexa, did the 49ers win?”

Sentences
(Machine Translation)

Summaries
(Summarization)

The Apple Watch has drawbacks. There are other smartwatches that offer more capabilities.

Structured Prediction

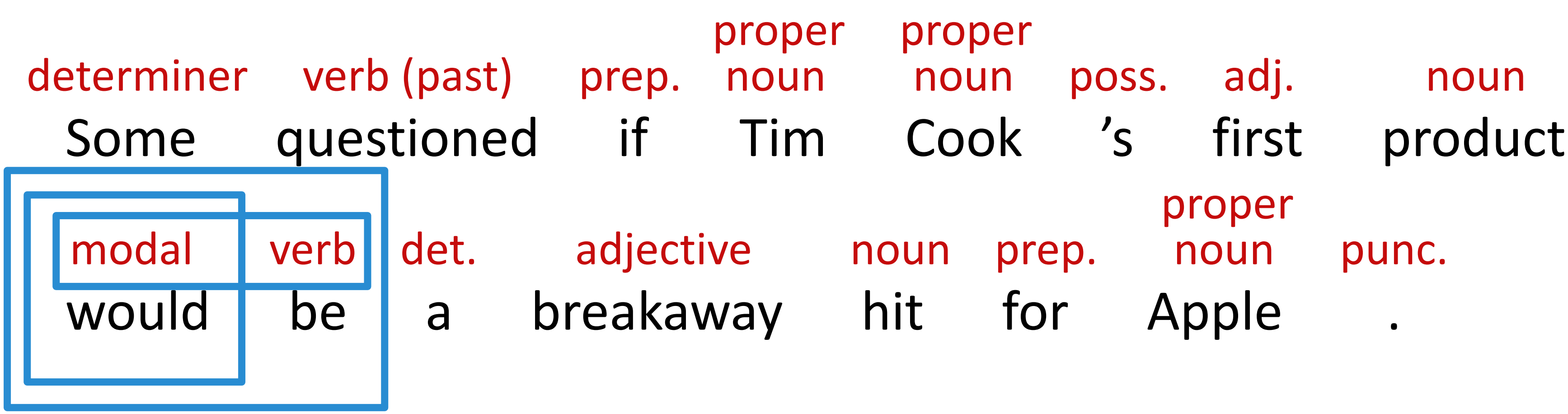
- what is and is not structured prediction?
- we use the term "structured prediction" when:
  - we use a structured score function or a structured loss function
- a structured score/loss function does not decompose across "minimal parts" of output
- to apply this definition we need to define "parts" and "minimal parts"

- parts:
  - each “part” is a subcomponent of entire input/output pair
  - e.g., a single word and its associated POS tag for POS tagging
  - or a sequence of two words and their POS tags
  - or a sequence of two POS tags

determiner	verb (past)	prep.	proper noun	proper noun	poss.	adj.	noun
Some	questioned	if	Tim	Cook	’s	first	product

modal	verb	det.	adjective	noun	prep.	proper noun	punc.
would	be	a	breakaway	hit	for	Apple	.

- parts:
  - each “part” is a subcomponent of entire input/output pair
  - “parts function” = decomposition of input/output pair into a set of parts
  - parts functions defined for score/loss function, rather than for task (many parts functions possible for a task)
  - parts may overlap



- parts:
  - each "part" is a subcomponent of entire input/output pair
  - "parts function" = decomposition of input/output pair into a set of parts
  - parts functions defined for score/loss function, rather than for task (many parts functions possible for a task)
  - parts may overlap

- minimal parts:
  - smallest possible parts for the task
  - minimal parts function defined for task (structured output space), not for structured score/loss function
  - minimal parts are non-overlapping

determiner	verb (past)	prep.	proper noun	proper noun	poss.	adj.	noun
Some	questioned	if	Tim	Cook	’s	first	product

modal	verb	det.	adjective	noun	prep.	proper noun	punc.
would	be	a	breakaway	hit	for	Apple	.

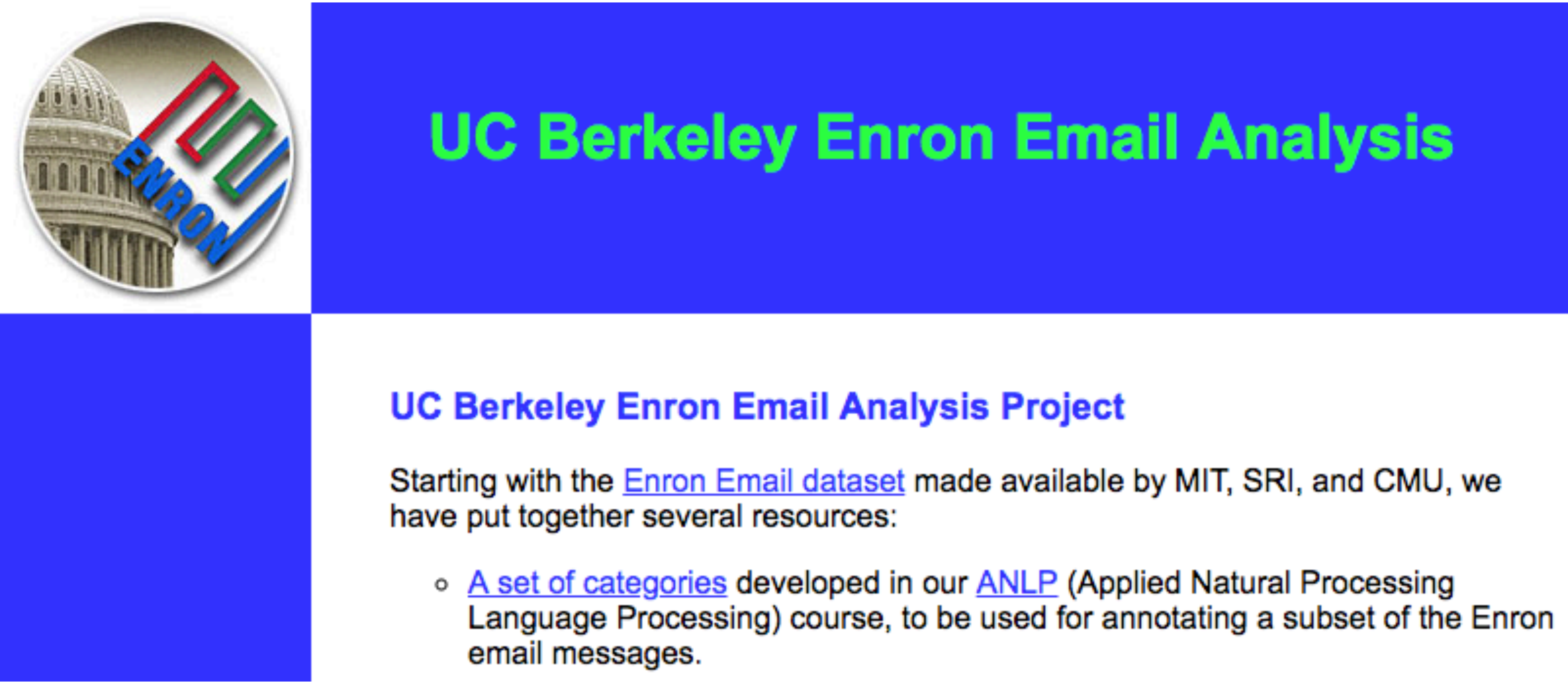
- minimal parts:
  - smallest possible parts for the task
  - minimal parts function defined for task (structured output space), not for structured score/loss function
  - minimal parts are non-overlapping

Categories of Structured Prediction Problems

- multi-label classification:
  - each input can be labeled with multiple labels
  - e.g., document classification where each document can have multiple labels

Categories of Structured Prediction Problems

- multi-label classification in NLP:



UC Berkeley Enron Email Analysis

UC Berkeley Enron Email Analysis Project

Starting with the Enron Email dataset made available by MIT, SRI, and CMU, we have put together several resources:

- A set of categories developed in our ANLP (Applied Natural Processing Language Processing) course, to be used for annotating a subset of the Enron email messages.

http://bailando.sims.berkeley.edu/enron_email.html

1 Coarse genre
1.1 Company Business, Strategy, etc. (elaborate in Section 3 [Topics])
1.2 Purely Personal
1.3 Personal but in professional context (e.g., it was good working with you)
1.4 Logistic Arrangements (meeting scheduling, technical support, etc)
…
4 Emotional tone (if not neutral)
4.1 jubilation
4.2 hope / anticipation
4.3 humor
4.4 camaraderie
4.5 admiration
4.6 gratitude
4.7 friendship / affection
…

http://bailando.sims.berkeley.edu/enron_email.html

Categories of Structured Prediction Problems

- multi-label classification:
  - each input can be labeled with multiple labels
  - if there are *N* possible labels, output space has size ___? **(Q1 on handout)**

Categories of Structured Prediction Problems

- multi-label classification:
  - each input can be labeled with multiple labels
  - if there are N possible labels, output space has size 2^N
  - what are the minimal parts? **(Q2 on handout)**

- parts:
  - each “part” is a subcomponent of entire input/output pair
  - “parts function” = decomposition of input/output pair into a set of parts
  - parts functions defined for score/loss function, rather than for task (many parts functions possible for a task)
  - parts may overlap

- minimal parts:
  - smallest possible parts for the task
  - minimal parts function defined for task (structured output space), not for structured score/loss function
  - minimal parts are non-overlapping

Categories of Structured Prediction Problems

- multi-label classification:
  - each input can be labeled with multiple labels
  - if there are N possible labels, output space has size 2^N
  - what are the minimal parts? individual labels

$$\text{mp}(\boldsymbol{y}) = \{y_1, ..., y_N\}$$
$$\text{where each } y_i \in \{0, 1\}$$

  - the mp(***y***) function defines the set of minimal parts of the structured output ***y***

Multi-Label Classification

– minimal parts: $\mathrm{mp}(\boldsymbol{y}) = \{y_1, ..., y_N\}$

where each $y_i \in \{0, 1\}$

– if score & loss functions factor across minimal parts, then we are not doing structured prediction

- e.g., we could build *N* binary classifiers, one for each label, and use them to independently predict each label for each input
- this would not be considered structured prediction

Parts and Score Functions

– let's define a "parts" function to characterize structured score/loss functions

$$\text{parts}(\boldsymbol{x}, \boldsymbol{y})$$

– where our score function is then defined:

$$\text{score}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}) = \sum_{\langle \boldsymbol{x}_r, \boldsymbol{y}_r \rangle \in \text{parts}(\boldsymbol{x}, \boldsymbol{y})} \text{score}_{\text{part}}(\boldsymbol{x}_r, \boldsymbol{y}_r, \boldsymbol{\theta})$$

– score function decomposes additively across parts

– each part is a subcomponent of input/output pair

Multi-Label Classification

– minimal parts: $\mathrm{mp}(\boldsymbol{y}) = \{y_1, ..., y_N\}$
where each $y_i \in \{0, 1\}$

– a parts function that uses the same decomposition as the minimal parts:

$$\mathrm{parts}(\boldsymbol{x}, \boldsymbol{y}) = \{\langle \boldsymbol{x}, y_1 \rangle, \ldots, \langle \boldsymbol{x}, y_N \rangle\}$$

– if we use this parts function, we will not be doing structured prediction

Multi-Label Classification

– minimal parts: $\mathrm{mp}(\boldsymbol{y}) = \{y_1, ..., y_N\}$

where each $y_i \in \{0, 1\}$

– a parts function that uses the same decomposition as the minimal parts:

$$\mathrm{parts}(\boldsymbol{x}, \boldsymbol{y}) = \{\langle \boldsymbol{x}, y_1 \rangle, \ldots, \langle \boldsymbol{x}, y_N \rangle\}$$

– parts function that does not decompose like the minimal parts? **(Q3 on handout)**

Multi-Label Classification

– minimal parts: $\mathrm{mp}(\boldsymbol{y}) = \{y_1, ..., y_N\}$ where each $y_i \in \{0, 1\}$

– a parts function that uses the same decomposition as the minimal parts:

$$\mathrm{parts}(\boldsymbol{x}, \boldsymbol{y}) = \{\langle \boldsymbol{x}, y_1 \rangle, \ldots, \langle \boldsymbol{x}, y_N \rangle\}$$

– parts function that does not decompose like the minimal parts?

$$\mathrm{parts}_1(\boldsymbol{x}, \boldsymbol{y}) = \{\langle \boldsymbol{x}, y_1 \rangle, \ldots, \langle \boldsymbol{x}, y_N \rangle\}$$

$$\mathrm{parts}_2(\boldsymbol{x}, \boldsymbol{y}) = \{\langle \boldsymbol{x}, \{y_i, y_j\} \rangle : 1 \leq i < j \leq N\}$$

$$\mathrm{parts}(\boldsymbol{x}, \boldsymbol{y}) = \mathrm{parts}_1(\boldsymbol{x}, \boldsymbol{y}) \cup \mathrm{parts}_2(\boldsymbol{x}, \boldsymbol{y})$$

this parts function uses parts for individual labels as well as all pairs of labels

other possibilities:

parts for all label triples, a part for the full set of labels, etc.

– parts function that does not decompose like the minimal parts?

$$\text{parts}_1(\boldsymbol{x}, \boldsymbol{y}) = \{\langle \boldsymbol{x}, y_1 \rangle, \ldots, \langle \boldsymbol{x}, y_N \rangle\}$$

$$\text{parts}_2(\boldsymbol{x}, \boldsymbol{y}) = \{\langle \boldsymbol{x}, \{y_i, y_j\} \rangle : 1 \leq i < j \leq N\}$$

$$\text{parts}(\boldsymbol{x}, \boldsymbol{y}) = \text{parts}_1(\boldsymbol{x}, \boldsymbol{y}) \cup \text{parts}_2(\boldsymbol{x}, \boldsymbol{y})$$

Categories of Structured Prediction Problems

- multi-label classification
- sequence labeling:
  - input is a sequence of length T
  - output is a sequence of length T
  - each position in output sequence is one of N labels
  - output space has size ___? **(Q4 on handout)**

Categories of Structured Prediction Problems

- multi-label classification
- sequence labeling:
  - input is a sequence of length T
  - output is a sequence of length T
  - each position in output sequence is one of N labels
  - output space has size N^T

Sequence Labeling

- input is a sequence of length T, output is a sequence of length T
- each position in output sequence is one of N labels
- minimal parts? **(Q5 on handout)**

Sequence Labeling

- input is a sequence of length *T*, output is a sequence of length *T*
- each position in output sequence is one of *N* labels
- minimal parts?

individual labels in output sequence

$$\mathrm{mp}(\boldsymbol{y}) = \{y_1, \ldots, y_T\}$$
$$\text{where each } y_i \in \{1, \ldots, N\}$$

Sequence Labeling

- minimal parts: $\text{mp}(\boldsymbol{y}) = \{y_1, \ldots, y_T\}$
  where each $y_i \in \{1, \ldots, N\}$

- parts function that uses same decomposition as minimal parts:
  $$\text{parts}(\boldsymbol{x}, \boldsymbol{y}) = \{\langle \boldsymbol{x}, y_1 \rangle, \ldots, \langle \boldsymbol{x}, y_T \rangle\}$$

- parts function that does not decompose like minimal parts? **(Q6 on handout)**

Sequence Labeling

- minimal parts: $\text{mp}(\boldsymbol{y}) = \{y_1, \ldots, y_T\}$
  where each $y_i \in \{1, \ldots, N\}$

- parts function that uses same decomposition as minimal parts:

$$\text{parts}(\boldsymbol{x}, \boldsymbol{y}) = \{\langle \boldsymbol{x}, y_1 \rangle, \ldots, \langle \boldsymbol{x}, y_T \rangle\}$$

- parts function that does not decompose like minimal parts?

$$\text{parts}_1(\boldsymbol{x}, \boldsymbol{y}) = \{\langle \boldsymbol{x}, y_1 \rangle, \ldots, \langle \boldsymbol{x}, y_T \rangle\}$$

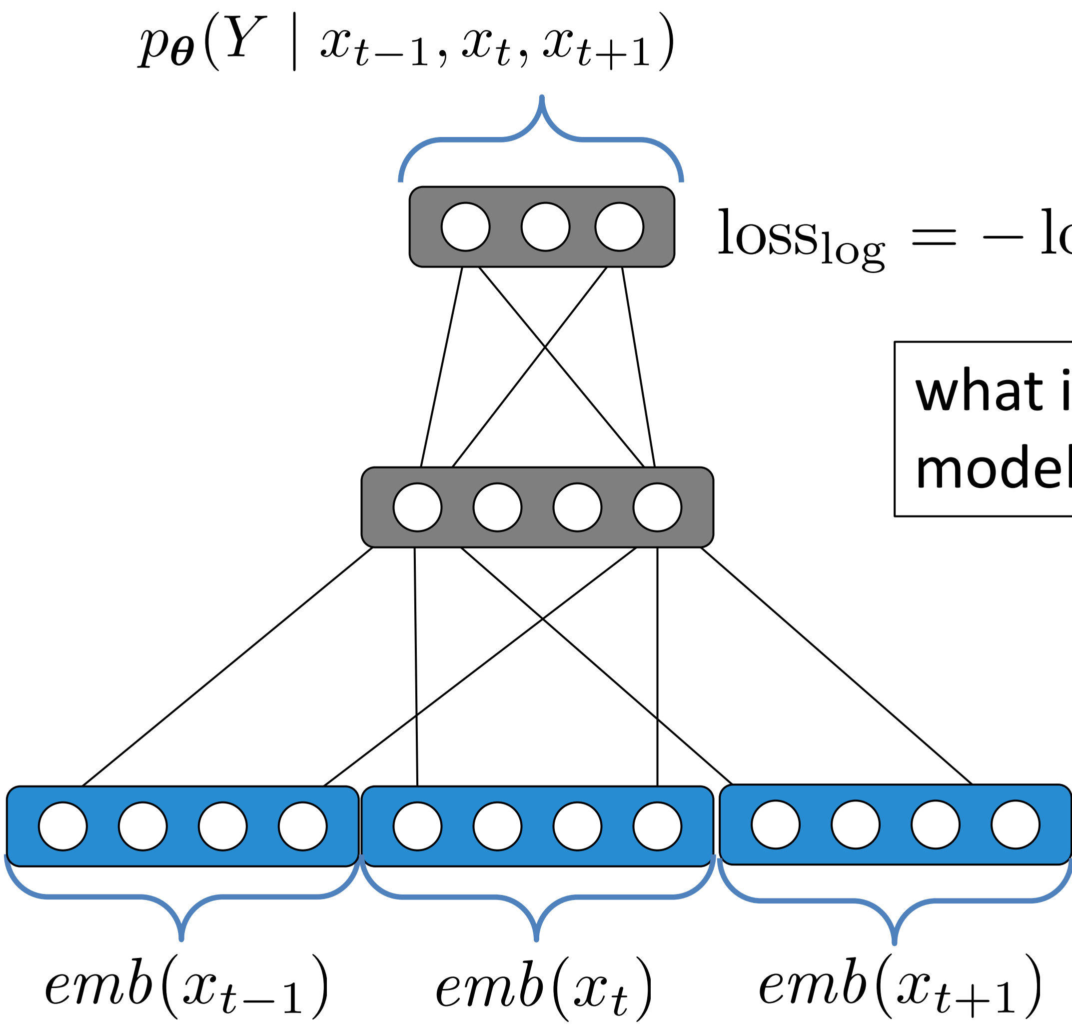
$$\text{parts}_2(\boldsymbol{x}, \boldsymbol{y}) = \{\langle \boldsymbol{x}, \langle y_i, y_{i+1} \rangle \rangle : 1 \leq i < T\}$$

$$\text{parts}(\boldsymbol{x}, \boldsymbol{y}) = \text{parts}_1(\boldsymbol{x}, \boldsymbol{y}) \cup \text{parts}_2(\boldsymbol{x}, \boldsymbol{y})$$

Examples of Models for Sequence Labeling

- consider a feed-forward neural network for POS tagging
- input is a word along with 1 word to either side of it
- output is predicted tag for center word
- training loss: log loss of correct tag at each position, summed over positions in sentence

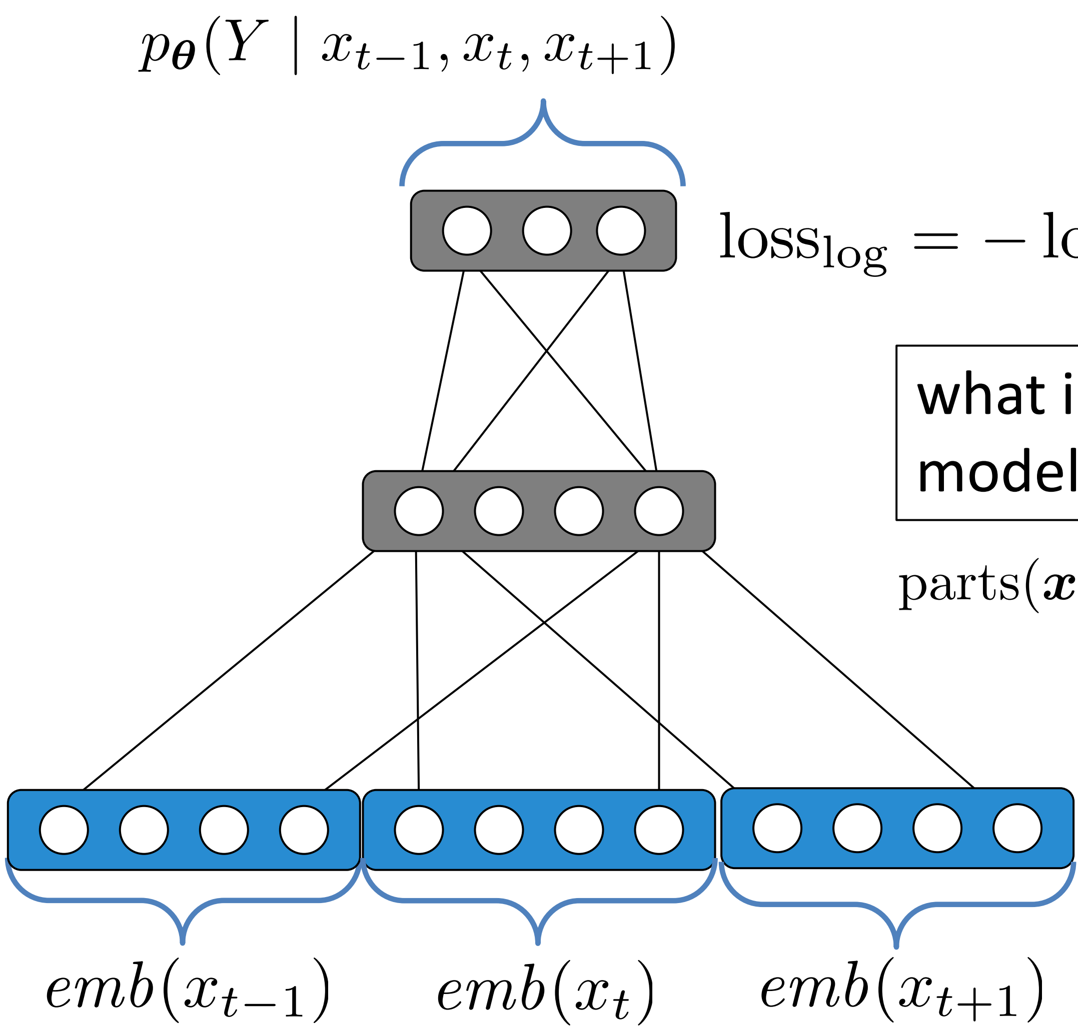
Feed-Forward POS Tagger



$$\text{loss}_{\log} = -\log p_{\boldsymbol{\theta}}(Y = y_t \mid x_{t-1}, x_t, x_{t+1})$$

what is the parts function for this model & loss? **(Q7 on handout)**

Feed-Forward POS Tagger

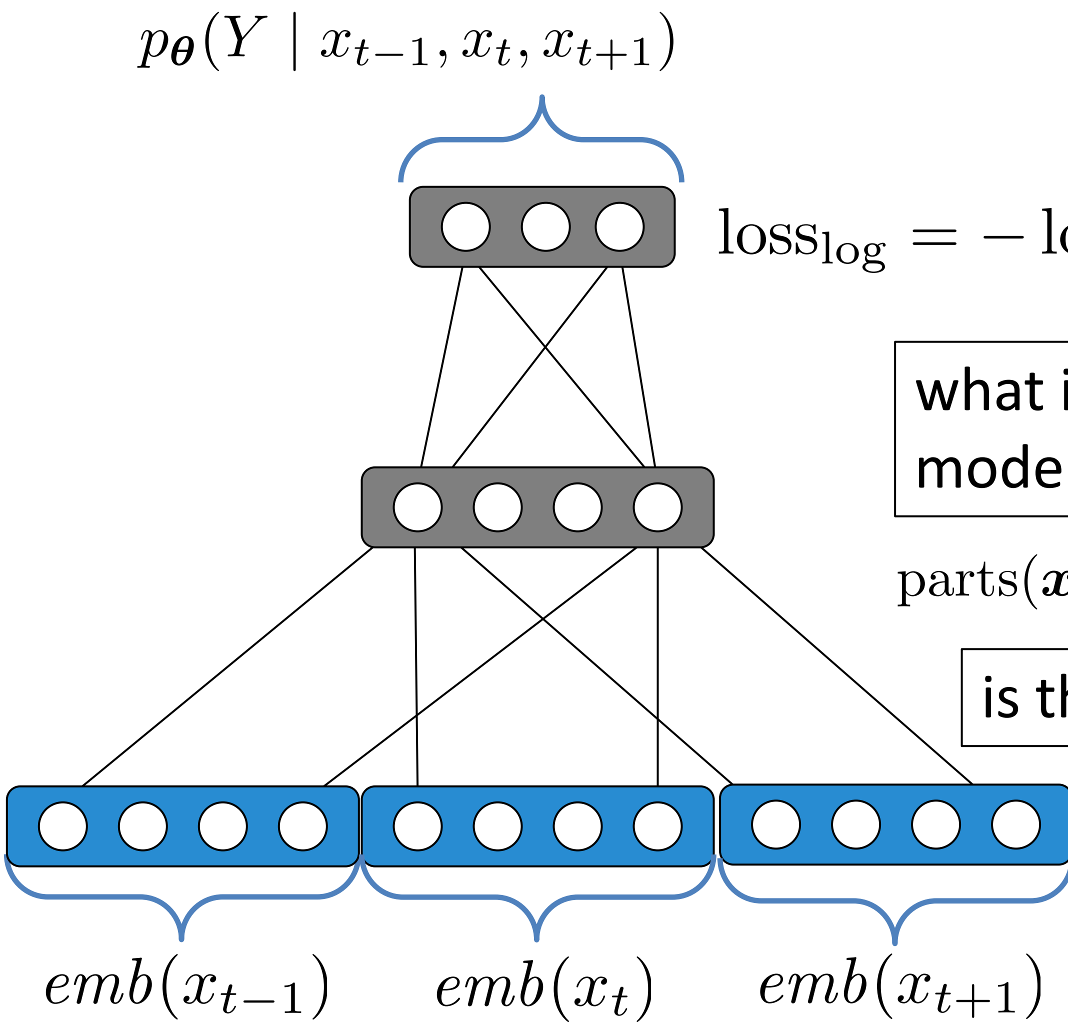


$$\text{loss}_{\log} = -\log p_{\boldsymbol{\theta}}(Y = y_t \mid x_{t-1}, x_t, x_{t+1})$$

what is the parts function for this model & loss?

$$\text{parts}(\boldsymbol{x}, \boldsymbol{y}) = \{\langle \boldsymbol{x}_{t-1:t+1}, y_t \rangle : 1 \leq t \leq T\}$$

Feed-Forward POS Tagger



$$\text{loss}_{\log} = -\log p_{\boldsymbol{\theta}}(Y = y_t \mid x_{t-1}, x_t, x_{t+1})$$

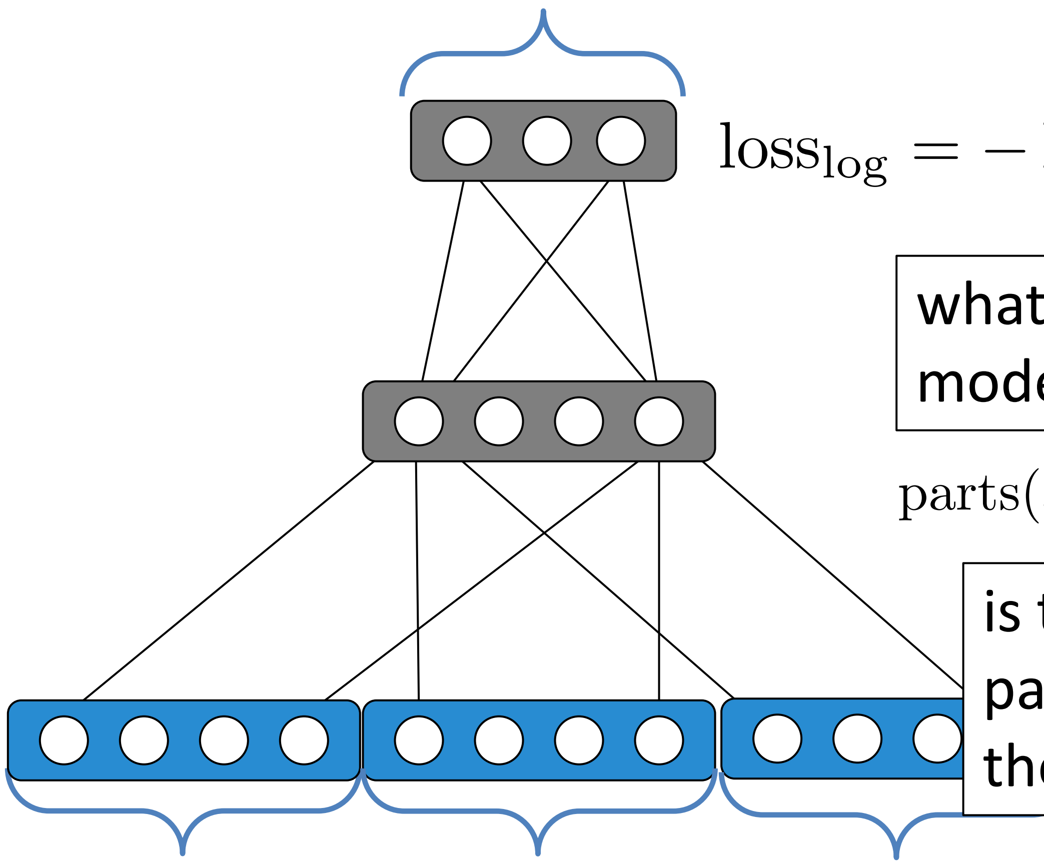
what is the parts function for this model & loss?

$$\text{parts}(\boldsymbol{x}, \boldsymbol{y}) = \{\langle \boldsymbol{x}_{t-1:t+1}, y_t \rangle : 1 \leq t \leq T\}$$

is this structured prediction?

Feed-Forward POS Tagger

$p_{\boldsymbol{\theta}}(Y \mid x_{t-1}, x_t, x_{t+1})$



$emb(x_{t-1})$ $emb(x_t)$ $emb(x_{t+1})$

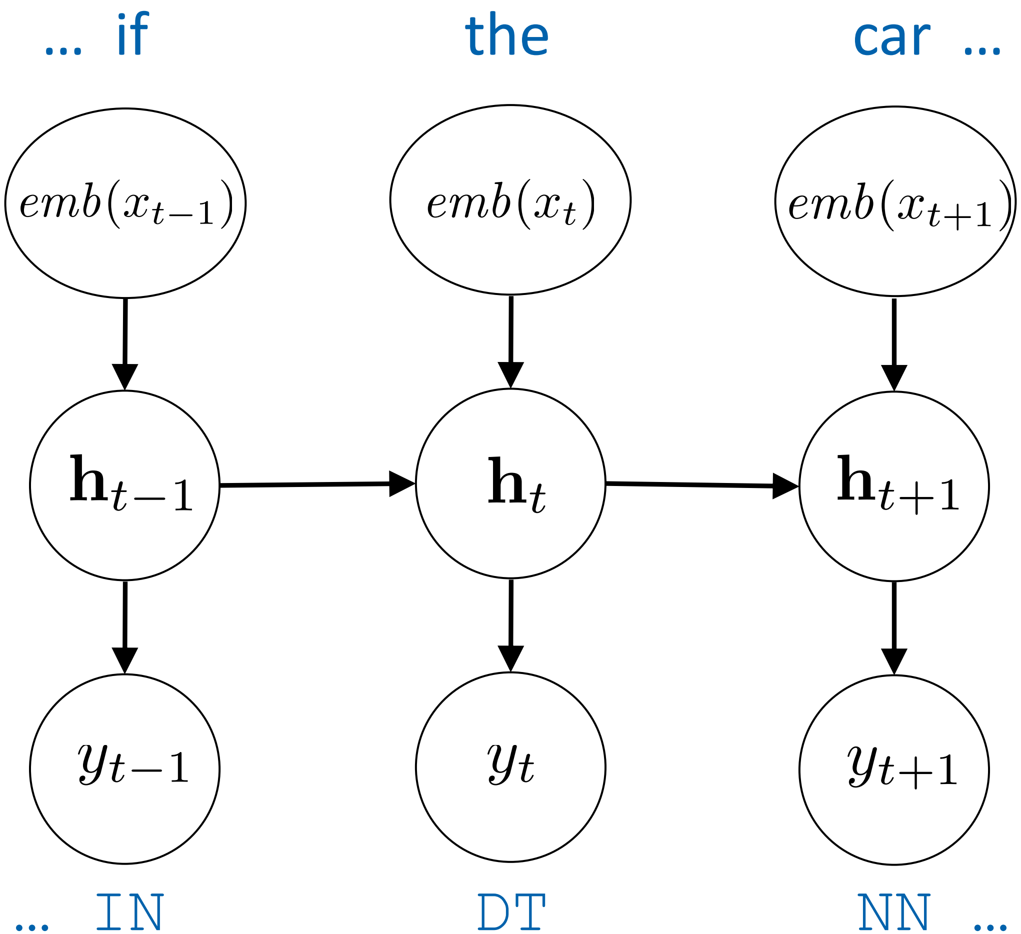
$$\text{loss}_{\log} = -\log p_{\boldsymbol{\theta}}(Y = y_t \mid x_{t-1}, x_t, x_{t+1})$$

what is the parts function for this model & loss?

$$\text{parts}(\boldsymbol{x}, \boldsymbol{y}) = \{\langle \boldsymbol{x}_{t-1:t+1}, y_t \rangle : 1 \leq t \leq T\}$$

is this structured prediction? no. parts functions decomposes like the minimal parts (in terms of y)

Forward RNN for Part-of-Speech Tagging



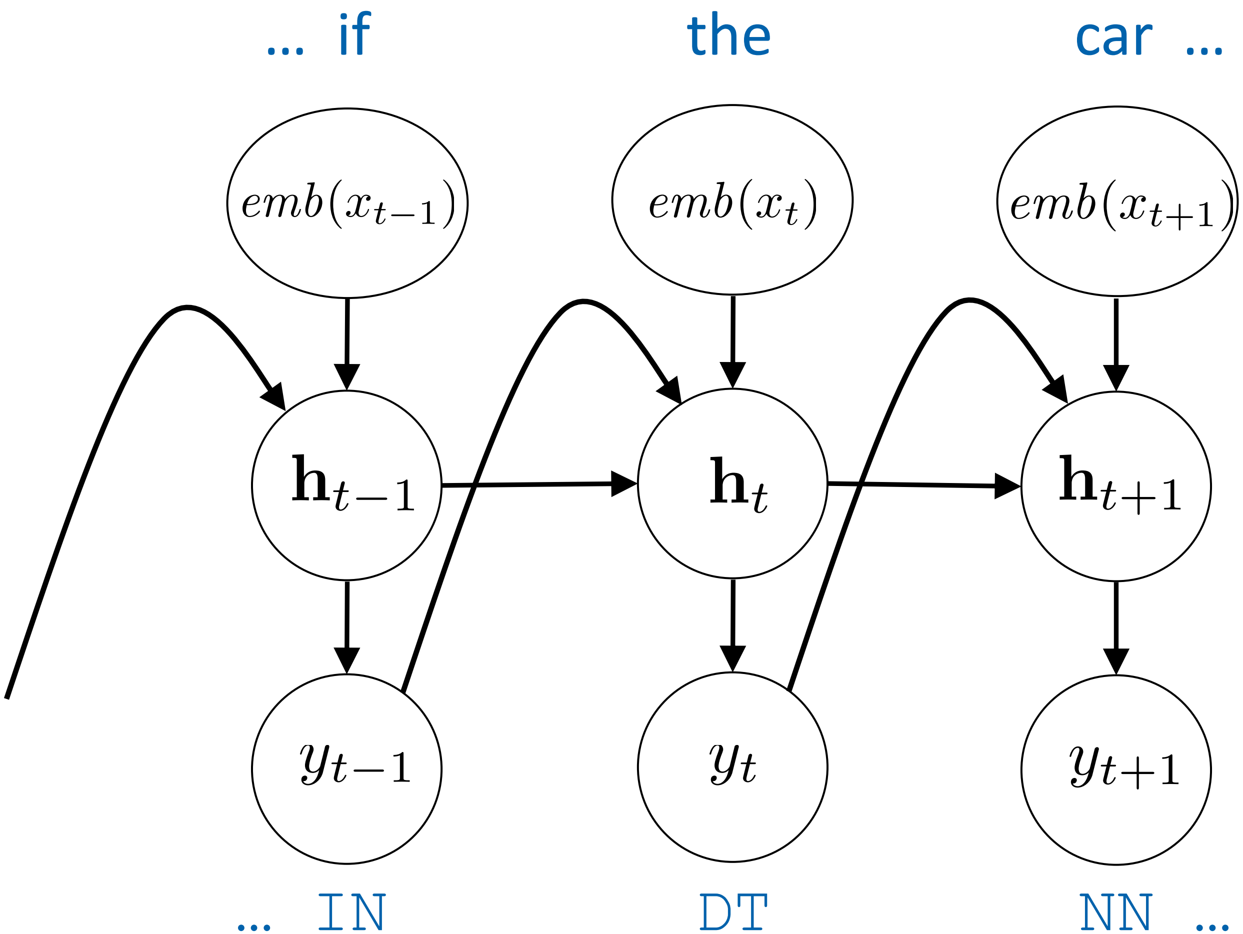
Forward RNN Tagger

- training loss: log loss of correct tag at each position, summed over positions in sentence
- is this structured prediction?
  - no
- parts function decomposes like minimal parts:

$$\text{parts}(\boldsymbol{x}, \boldsymbol{y}) = \{\langle \boldsymbol{x}_{1:t}, y_t \rangle : 1 \leq t \leq T\}$$

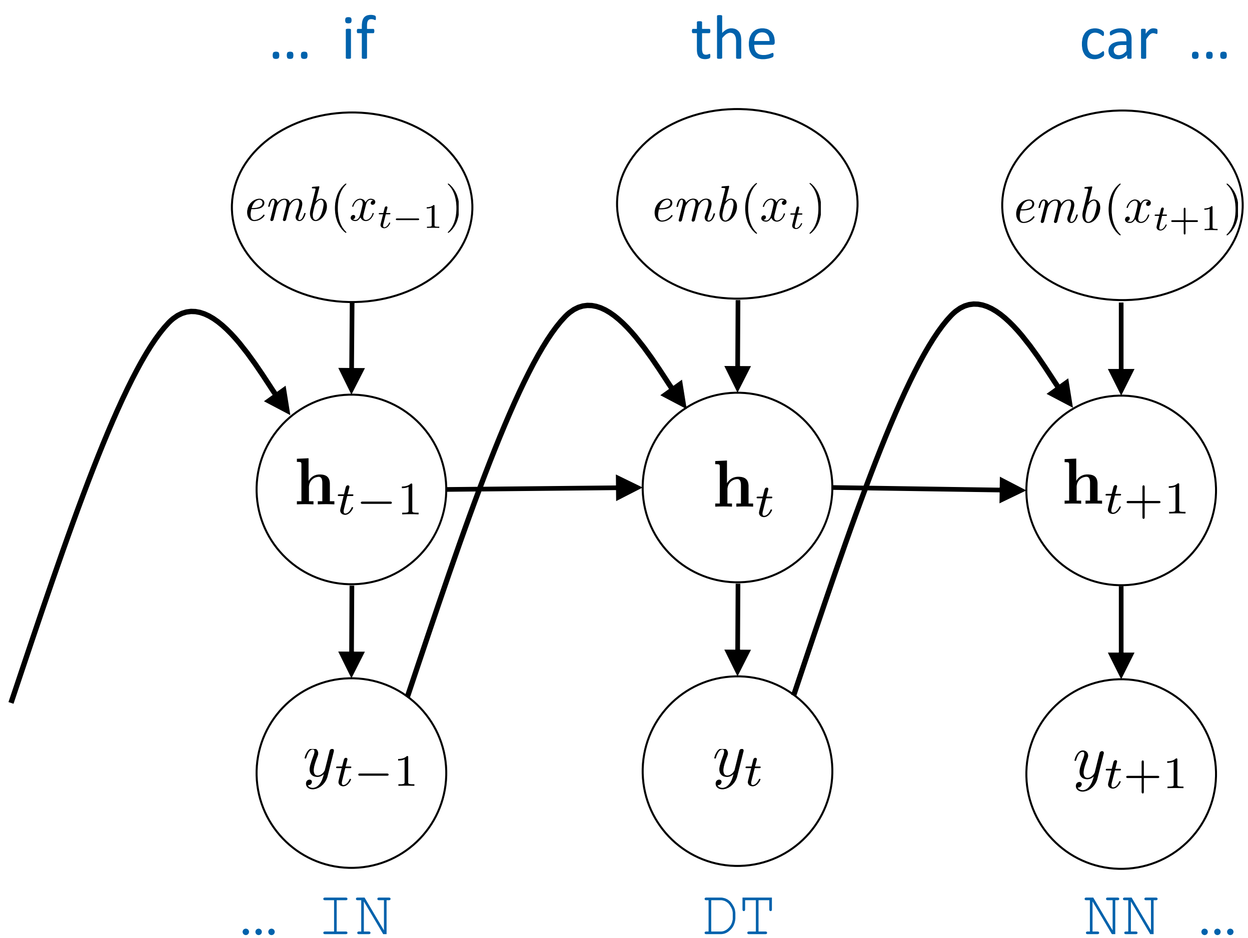
- there are many effective and popular approaches to sequence labeling that do not fit our definition of a “structured predictor”

Forward RNN for Part-of-Speech Tagging with Previous Label



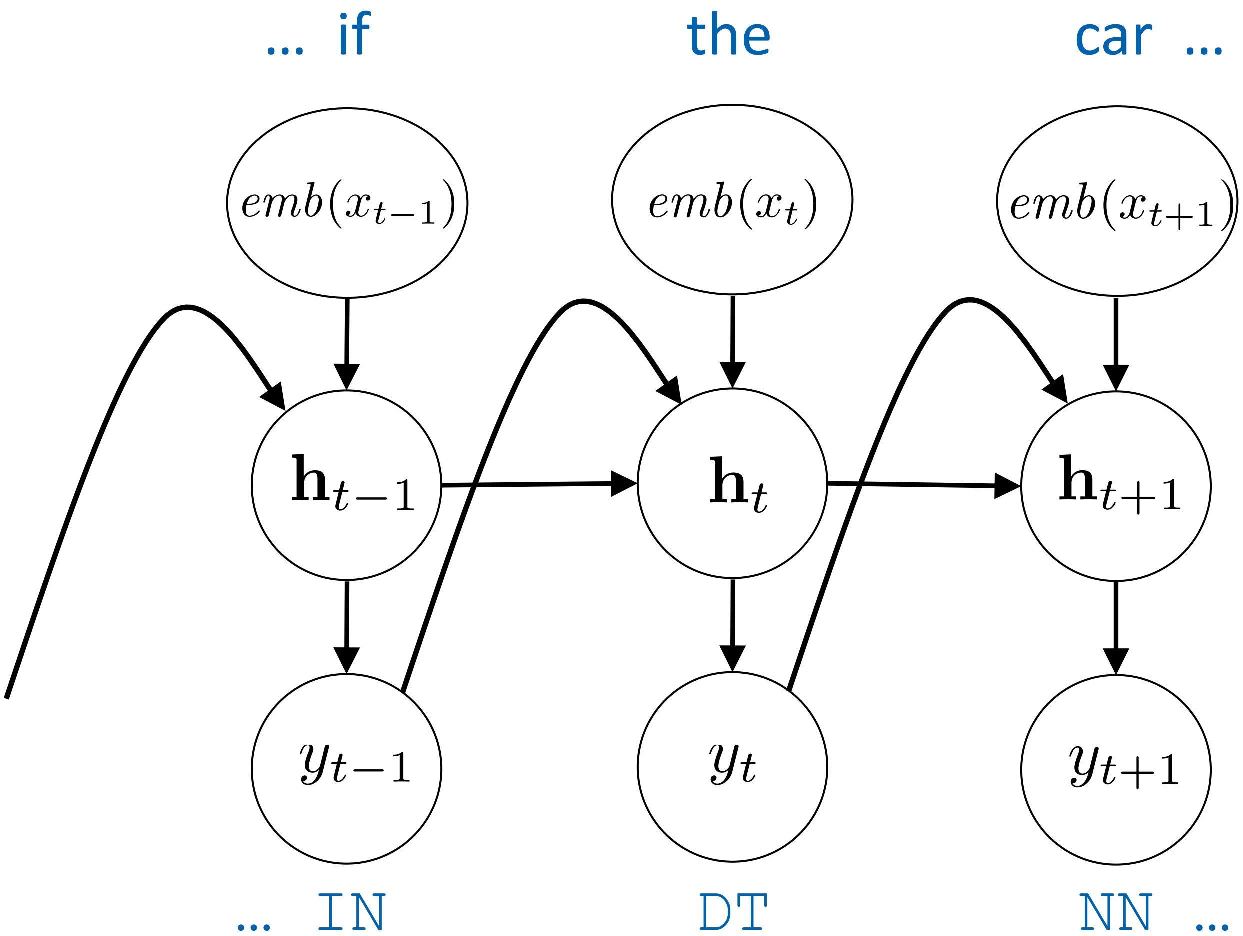
this model uses the previous *y* to compute a hidden vector

Forward RNN for Part-of-Speech Tagging with Previous Label



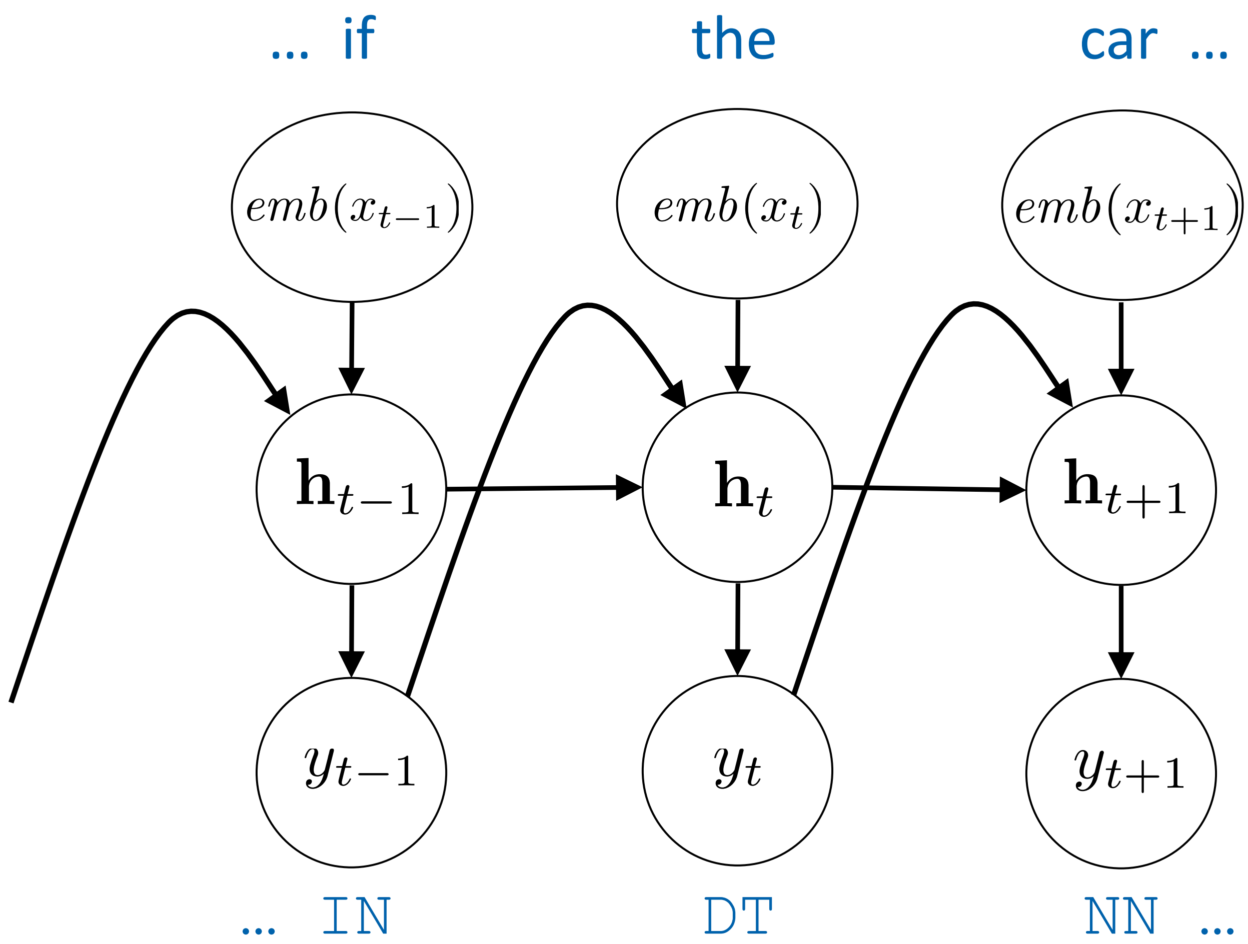
hidden vector used to compute probability distribution over tags at each position: $p_{\boldsymbol{\theta}}(Y_t \mid \boldsymbol{x}_{1:t}, \boldsymbol{y}_{1:t-1})$

Forward RNN for Part-of-Speech Tagging with Previous Label



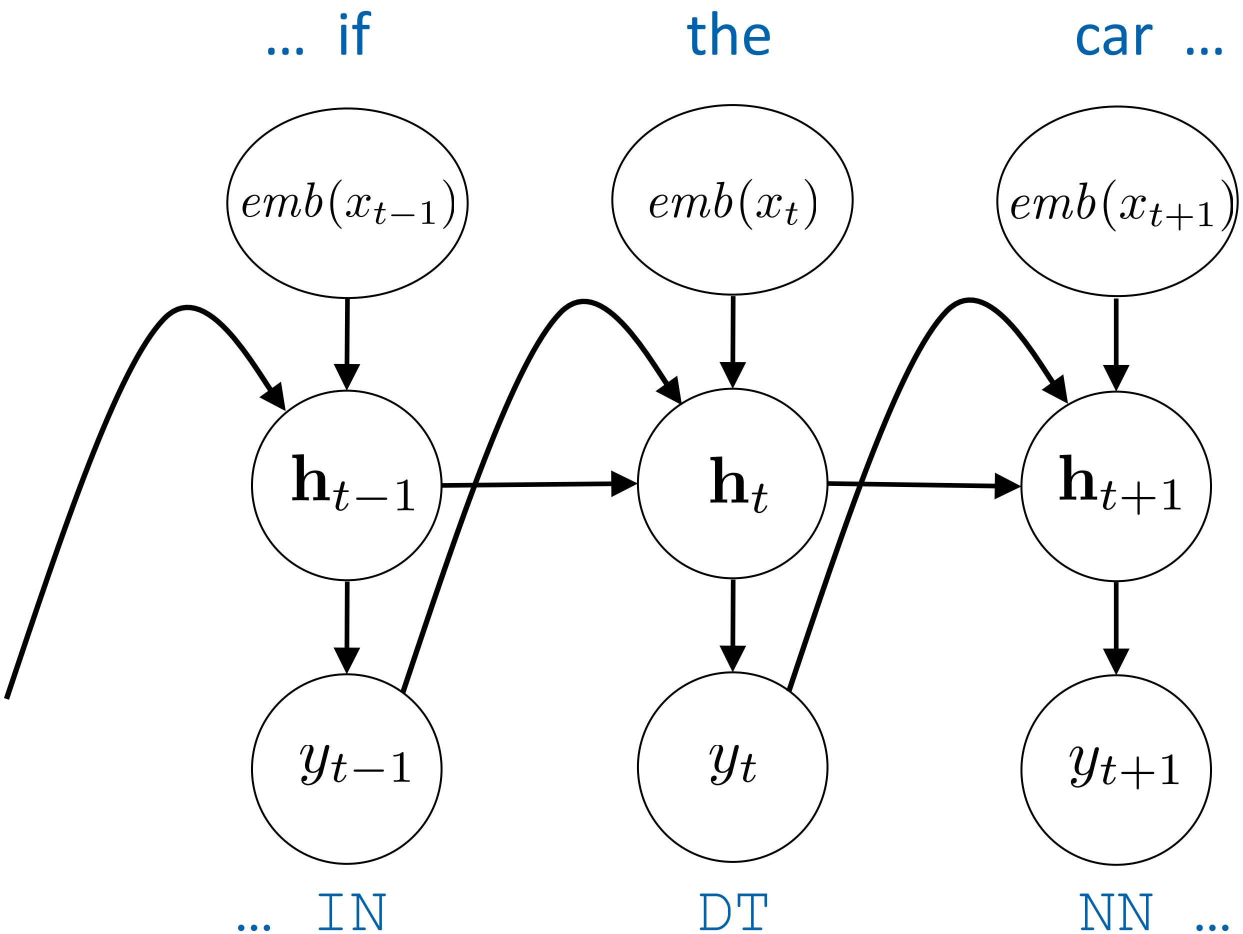
loss: $-\sum_t \log p_{\boldsymbol{\theta}}(Y_t = y_t \mid \boldsymbol{x}_{1:t}, \boldsymbol{y}_{1:t-1})$

Forward RNN for Part-of-Speech Tagging with Previous Label



what is the parts function for this model & loss?

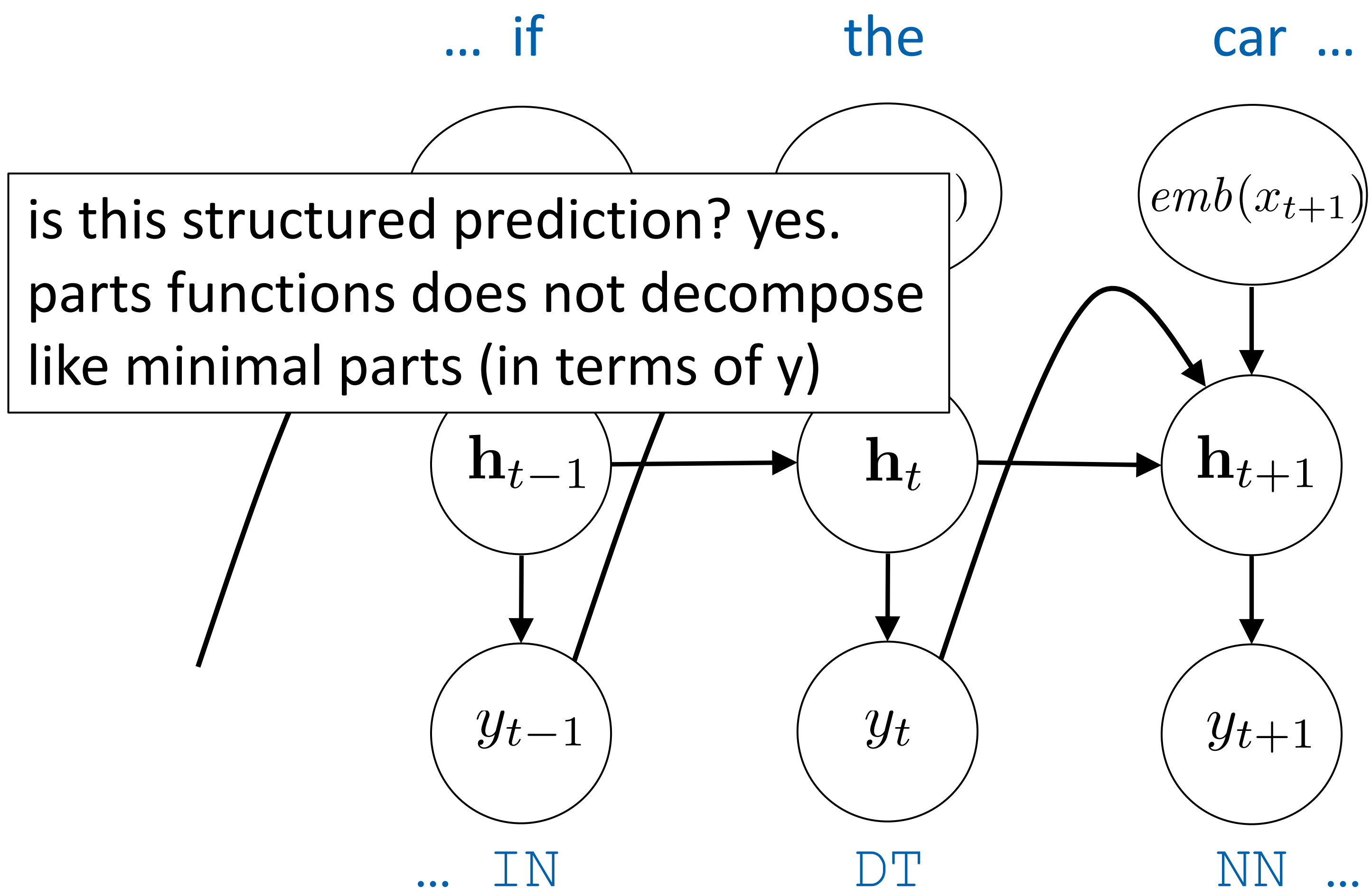
Forward RNN for Part-of-Speech Tagging with Previous Label



what is the parts function for this model & loss?

$$\text{parts}(\boldsymbol{x}, \boldsymbol{y}) = \{\langle \boldsymbol{x}_{1:t}, \boldsymbol{y}_{1:t} \rangle : 1 \leq t \leq T\}$$

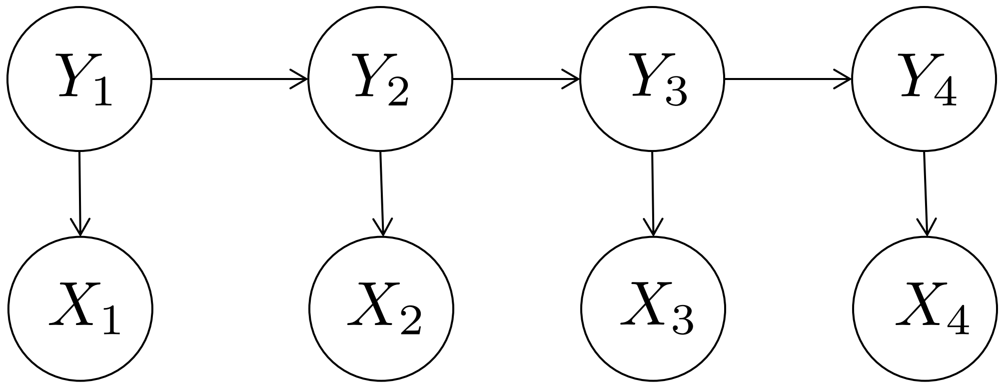
Forward RNN for Part-of-Speech Tagging with Previous Label



$$\text{parts}(\boldsymbol{x}, \boldsymbol{y}) = \{\langle \boldsymbol{x}_{1:t}, \boldsymbol{y}_{1:t} \rangle : 1 \leq t \leq T\}$$

Hidden Markov Models (HMMs)

Graphical Model for an HMM for a sequence of length 4:



$$p_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1}^{|\boldsymbol{x}|} p_{\boldsymbol{\tau}}(y_i \mid y_{i-1}) p_{\boldsymbol{\eta}}(x_i \mid y_i)$$

transition parameters: $p_{\boldsymbol{\tau}}(y_i \mid y_{i-1})$

emission parameters: $p_{\boldsymbol{\eta}}(x_i \mid y_i)$

*for now, we are omitting stopping probabilities for simplicity