Lexical Semantics Wrap-up and Midterm Review

How do we know when a word has more than one sense?

- ATIS examples
  - Which flights serve breakfast?
  - Does America West serve Philadelphia?
- The "zeugma" test:
  - ?Does United serve breakfast and San Jose?

Synonyms

- Word that have the same meaning in some or all contexts.
  - filbert / hazelnut
  - couch / sofa
  - big / large
  - automobile / car
  - vomit / throw up
  - Water / H_20
- Two lexemes are synonyms if they can be successfully substituted for each other in all situations
  - If so they have the same **propositional meaning**

Synonyms

- But there are few (or no) examples of perfect synonymy.
  - Why should that be?
  - Even if many aspects of meaning are identical
  - Still may not preserve the acceptability based on notions of politeness, slang, register, genre, etc.
- Example:
  - Water and H_20

Some more terminology

- Lemmas and wordforms
  - A **lexeme** is an abstract pairing of meaning and form
  - A **lemma** or **citation form** is the grammatical form that is used to represent a **lexeme**.
    - ***Carpet*** is the lemma for ***carpets***
    - ***Dormir*** is the lemma for ***duermes*.**
  - Specific surface forms *carpets, sung, duermes* are called **wordforms**
- The lemma *bank* has two **senses:**
  - Instead, a **bank** can hold the investments in a custodial account in the client's name
  - But as agriculture burgeons on the east **bank**, the river will shrink even more.
- A **sense** is a discrete representation of one aspect of the meaning of a word

Synonymy is a relation between senses rather than words

- Consider the words *big* and *large*
- Are they synonyms?
  - How **big** is that plane?
  - Would I be flying on a **large** or small plane?
- How about here:
  - Miss Nelson, for instance, became a kind of **big** sister to Benjamin.
  - ?Miss Nelson, for instance, became a kind of **large** sister to Benjamin.
- Why?
  - *big* has a sense that means being older, or grown up
  - *large* lacks this sense

Antonyms

- Senses that are opposites with respect to one feature of their meaning
- Otherwise, they are very similar!
  - dark / light
  - short / long
  - hot / cold
  - up / down
  - in / out
- More formally: antonyms can
  - define a binary opposition or at opposite ends of a scale (*long/short, fast/slow*)
  - Be **reversives**: *rise/fall, up/down*

Hyponymy

- One sense is a **hyponym** of another if the first sense is more specific, denoting a subclass of the other
  - *car* is a hyponym of *vehicle*
  - *dog* is a hyponym of *animal*
  - *mango* is a hyponym of *fruit*
- Conversely
  - *vehicle* is a hypernym/superordinate of *car*
  - *animal* is a hypernym of *dog*
  - *fruit* is a hypernym of *mango*

superordinate	vehicle	fruit	furniture	mammal
hyponym	car	mango	chair	dog

Hypernymy more formally

- Extensional:
  - The class denoted by the superordinate
  - extensionally includes the class denoted by the hyponym
- Entailment:
  - A sense A is a hyponym of sense B if being an A entails being a B
- Hyponymy is usually transitive
  - (A hypo B and B hypo C entails A hypo C)

- Why would hypernyms/hyponyms be important to constructing a meaning representation?

Why would hypernyms/hyponyms be important for meaning representation?

II. WordNet

- A hierarchically organized lexical database
- On-line thesaurus + aspects of a dictionary
    - Versions for other languages are under development

Category	Unique Forms
Noun	117,097
Verb	11,488
Adjective	22,141
Adverb	4,601

WordNet

- Where it is:
  - https://wordnet.princeton.edu/

Format of Wordnet Entries

The noun “bass” has 8 senses in WordNet.

1. bass^{1} - (the lowest part of the musical range)
2. bass^{2}, bass part^{1} - (the lowest part in polyphonic music)
3. bass^{3}, basso^{1} - (an adult male singer with the lowest voice)
4. sea bass^{1}, bass^{4} - (the lean flesh of a saltwater fish of the family Serranidae)
5. freshwater bass^{1}, bass^{5} - (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
6. bass^{6}, bass voice^{1}, basso^{2} - (the lowest adult male singing voice)
7. bass^{7} - (the member with the lowest range of a family of musical instruments)
8. bass^{8} - (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

The adjective “bass” has 1 sense in WordNet.

1. bass^{1}, deep^{6} - (having or denoting a low vocal or instrumental range)
   "a deep voice"; "a bass voice is lower than a baritone voice"; "a bass clarinet"

WordNet Noun Relations

Relation	Also called	Definition	Example
Hypernym	Superordinate	From concepts to superordinates	$breakfast^1 \rightarrow meal^1$
Hyponym	Subordinate	From concepts to subtypes	$meal^1 \rightarrow lunch^1$
Member Meronym	Has-Member	From groups to their members	$faculty^2 \rightarrow professor^1$
Has-Instance		From concepts to instances of the concept	$composer^1 \rightarrow Bach^1$
Instance		From instances to their concepts	$Austen^1 \rightarrow author^1$
Member Holonym	Member-Of	From members to their groups	$copilot^1 \rightarrow crew^1$
Part Meronym	Has-Part	From wholes to parts	$table^2 \rightarrow leg^3$
Part Holonym	Part-Of	From parts to wholes	$course^7 \rightarrow meal^1$
Antonym		Opposites	$leader^1 \rightarrow follower^1$

WordNet Verb Relations

Relation	Definition	Example
Hypernym	From events to superordinate events	$fly^9 \rightarrow travel^5$
Troponym	From a verb (event) to a specific manner elaboration of that verb	$walk^1 \rightarrow stroll^1$
Entails	From verbs (events) to the verbs (events) they entail	$snore^1 \rightarrow sleep^1$
Antonym	Opposites	$increase^1 \Longleftrightarrow decrease^1$

WordNet Hierarchies

```
Sense 3
bass, basso --
(an adult male singer with the lowest voice)
=> singer, vocalist, vocalizer, vocaliser
   => musician, instrumentalist, player
      => performer, performing artist
         => entertainer
            => person, individual, someone...
               => organism, being
                  => living thing, animate thing,
                     => whole, unit
                        => object, physical object
                           => physical entity
                              => entity
               => causal agent, cause, causal agency
                  => physical entity
                     => entity

Sense 7
bass --
(the member with the lowest range of a family of
musical instruments)
=> musical instrument, instrument
   => device
      => instrumentality, instrumentation
         => artifact, artefact
            => whole, unit
               => object, physical object
                  => physical entity
```

How is "sense" defined in WordNet?

- The set of near-synonyms for a WordNet sense is called a **synset (synonym set)**; it's their version of a sense or a concept
- Example: chump as a noun to mean
  - 'a person who is gullible and easy to take advantage of'

{chump^1, fool^2, gull^1, mark^9, patsy^1, fall guy^1, sucker^1, soft touch^1, mug^2}

- Each of these senses share this same gloss
- Thus for WordNet, the meaning of this sense of chump *is* this list.

Wordnet example

Questions?

Midterm

- Format
  - Multiple Choice questions
  - Short answer questions
  - Problem solving

- What will it cover?
  - Anything covered in class
  - From reading that supports material in class
  - Math as needed for neural nets, machine learning, smoothing

Midterm

- Closed book, no notes, no electronics
- Will avoid asking you to recall formulas
  - That said, you should know how to compute the probability of ngrams, of POS tags, basics for smoothing, language modeling, how to do computation for neural nets.
- Will cover anything from beginning through today
- Sample midterm questions posted

Top topics

- Viterbi algorithm
- Dependency parsing
- RNNs

Questions?

Viterbi and POS

Two kinds of probabilities (1)

- Tag transition probabilities $p(t_i|t_{i-1})$
  - Determiners likely to precede adjs and nouns
    - That/DT flight/NN
    - The/DT yellow/JJ hat/NN
    - So we expect P(NN|DT) and P(JJ|DT) to be high
    - But P(DT|JJ) to be low
  - Compute P(NN|DT) by counting in a labeled corpus:

$$P(t_i|t_{i-1}) = \frac{C(t_{i-1},t_i)}{C(t_{i-1})}$$

$$P(NN|DT) = \frac{C(DT,NN)}{C(DT)} = \frac{56,509}{116,454} = .49$$

Two kinds of probabilities (2)

- Word likelihood probabilities $p(w_i|t_i)$
  - VBZ (3sg Pres verb) likely to be "is"
  - Compute P(is|VBZ) by counting in a labeled corpus:

$$P(w_i|t_i) = \frac{C(t_i, w_i)}{C(t_i)}$$

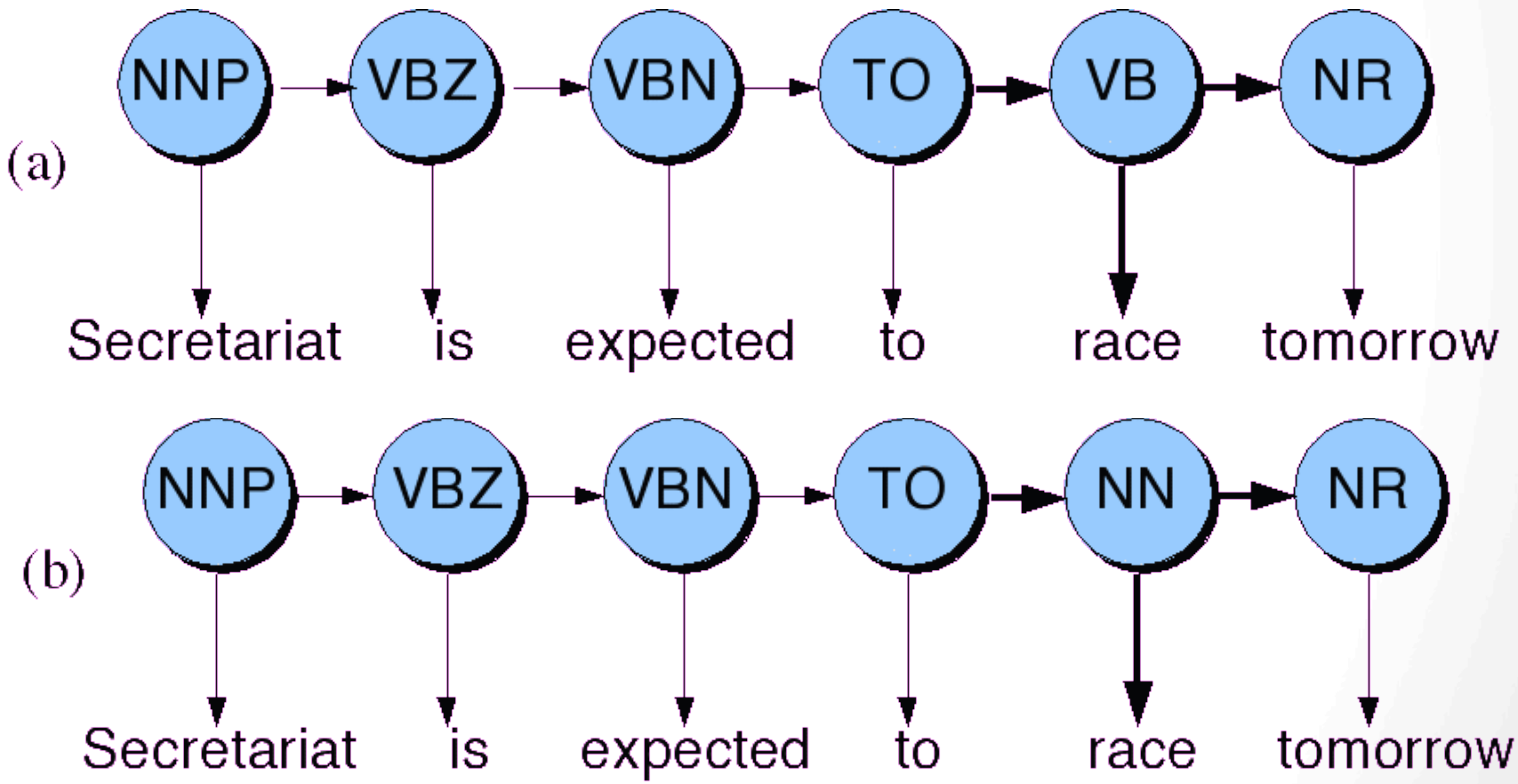
$$P(is|VBZ) = \frac{C(VBZ, is)}{C(VBZ)} = \frac{10,073}{21,627} = .47$$

An Example: the verb "race"

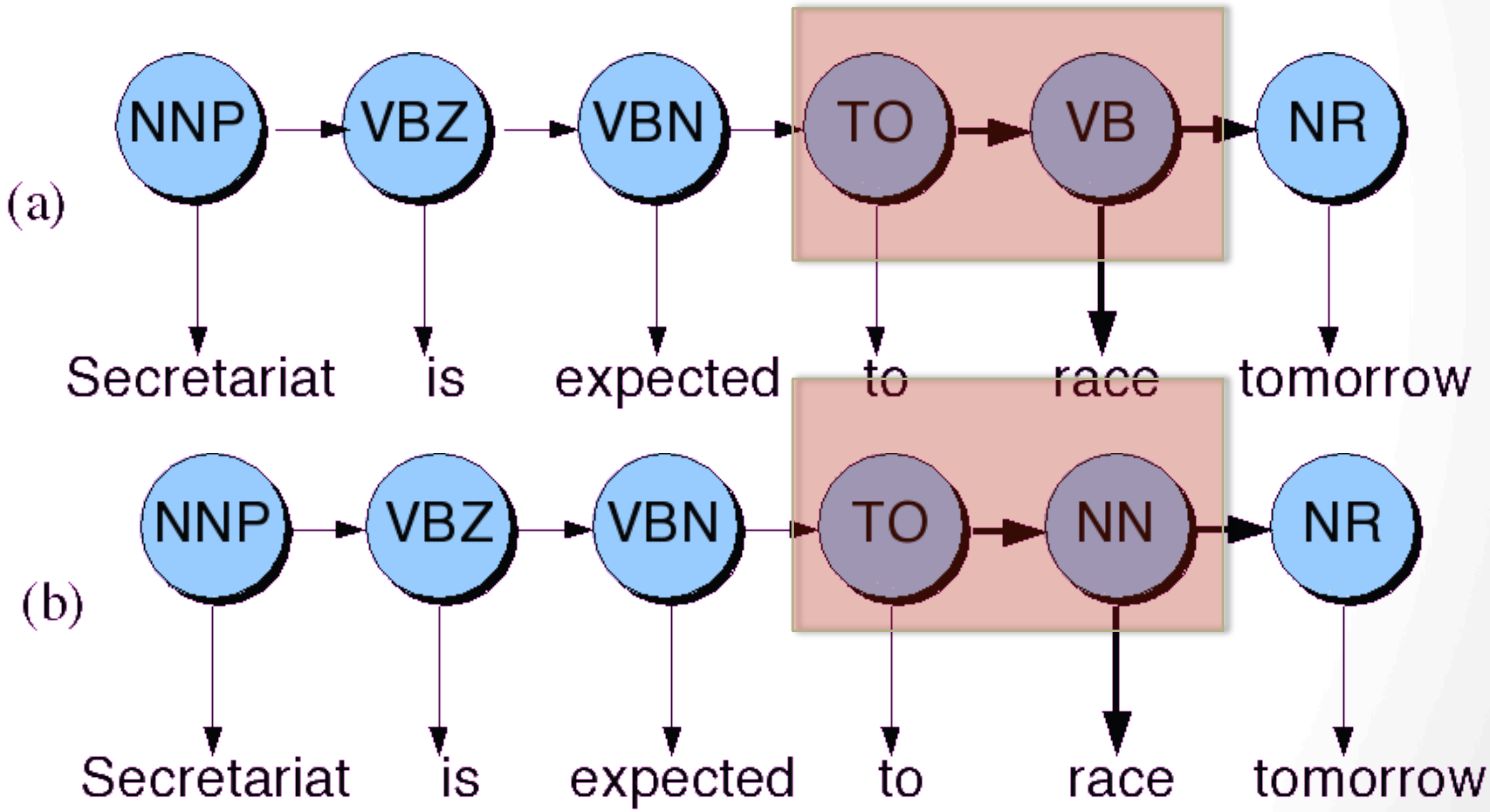
- Secretariat/NNP is/VBZ expected/VBN to/TO **race**/VB tomorrow/NR
- People/NNS continue/VB to/TO inquire/VB the/DT reason/NN for/IN the/DT **race**/NN for/IN outer/JJ space/NN

- How do we pick the right tag?

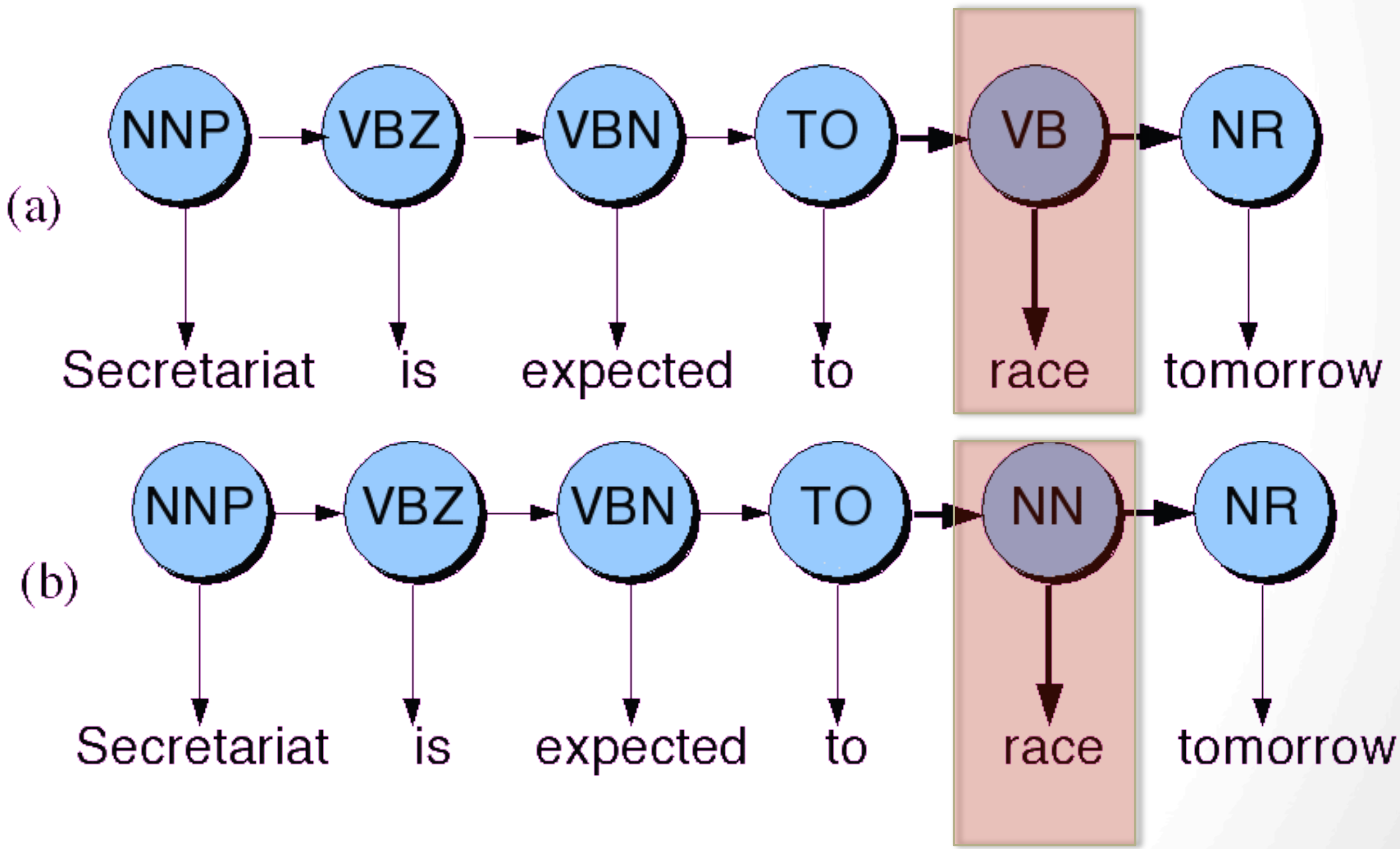
Disambiguating "race"



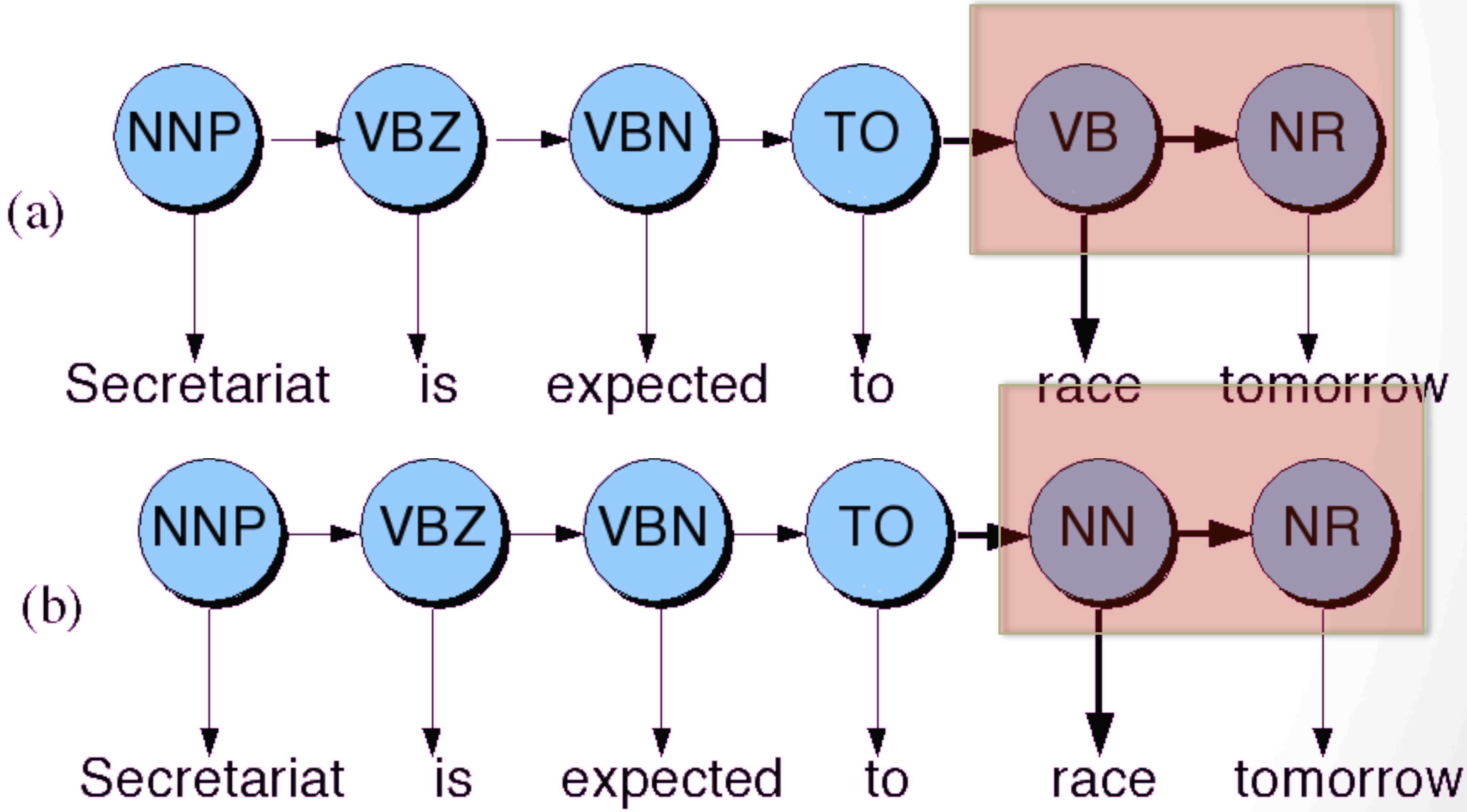
Disambiguating "race"



Disambiguating "race"



Disambiguating “race”



- P(NN|TO) = .00047
- P(VB|TO) = .83
- P(race|NN) = .00057
- P(race|VB) = .00012
- P(NR|VB) = .0027
- P(NR|NN) = .0012
- P(VB|TO)P(NR|VB)P(race|VB) = .00000027
- P(NN|TO)P(NR|NN)P(race|NN)=.00000000032
- So we (correctly) choose the verb reading,

HMMS

Hidden Markov Models

- We don't observe POS tags
  - We infer them from the words we see
- Observed events
- Hidden events

Hidden Markov Model

- For Markov chains, the output symbols are the same as the states.
  - See **hot** weather: we're in state **hot**
- But in part-of-speech tagging (and other things)
  - The output symbols are **words**
  - The hidden states are **part-of-speech tags**
- So we need an extension!
- A Hidden Markov Model is an extension of a Markov chain in which the input symbols are not the same as the states.
- This means we don't know which state we are in.

Hidden Markov Models

- States $Q = q_1, q_2 \ldots q_{N;}$
- Observations $O = o_1, o_2 \ldots o_{N;}$
  - Each observation is a symbol from a vocabulary $V = \{v_1, v_2, \ldots v_V\}$
- Transition probabilities
  - Transition probability matrix $A = \{a_{ij}\}$

$$a_{ij} = P(q_t = j \mid q_{t-1} = i) \quad 1 \le i, j \le N$$

- Observation likelihoods
  - Output probability matrix $B = \{b_i(k)\}$

$$b_i(k) = P(X_t = o_k \mid q_t = i)$$

- Special initial probability vector π

$$\pi_i = P(q_1 = i) \quad 1 \le i \le N$$

Hidden Markov Models

- Some constraints

$$\sum_{j=1}^{N} a_{ij} = 1; \quad 1 \le i \le N$$

$$\pi_i = P(q_1 = i) \quad 1 \le i \le N$$

$$\sum_{k=1}^{M} b_i(k) = 1$$

$$\sum_{j=1}^{N} \pi_j = 1$$

Assumptions

- **Markov assumption:**

$$P(q_i \mid q_1 \ldots q_{i-1}) = P(q_i \mid q_{i-1})$$

- **Output-independence assumption**

$$P(o_t \mid O_1^{t-1}, q_1^t) = P(o_t \mid q_t)$$

Three fundamental Problems for HMMs

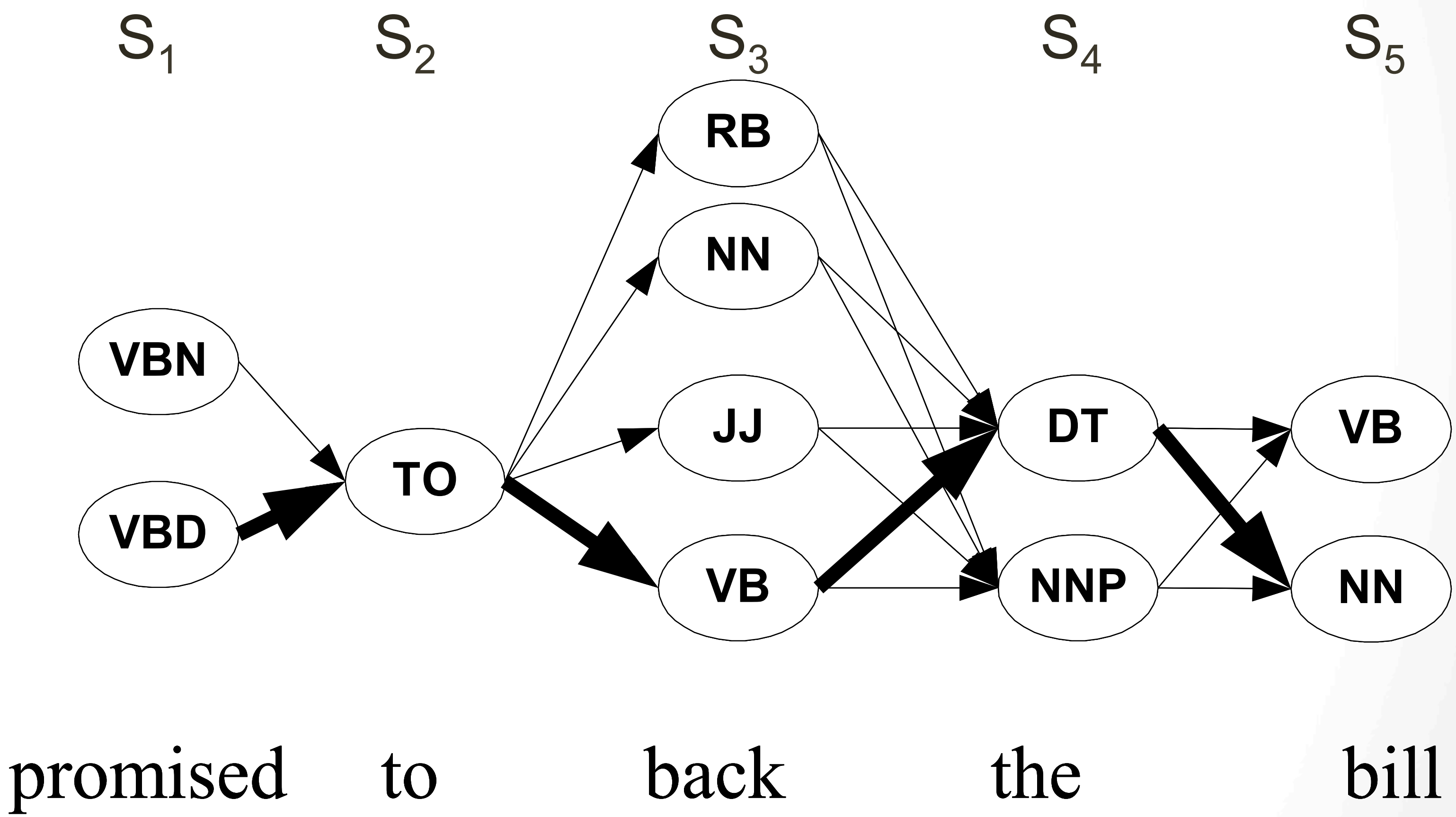
- ***Likelihood***: Given an HMM $\lambda = (A,B)$ and an observation sequence O, determine the likelihood $P(O, \lambda)$.
- ***Decoding***: Given an observation sequence O and an HMM $\lambda = (A,B)$, discover the best hidden state sequence Q.
- ***Learning***: Given an observation sequence O and the set of states in the HMM, learn the HMM parameters A and B.

  What kind of data would we need to learn the HMM parameters?

Decoding

- The best hidden sequence
    - Weather sequence in the ice cream task
    - POS sequence given an input sentence
- We could use argmax over the probability of each possible hidden state sequence
    - *Why not?*
- Viterbi algorithm
    - Dynamic programming algorithm
    - Uses a dynamic programming trellis
        - Each trellis cell represents, $v_t(j)$, represents the probability that the HMM is in state j after seeing the first t observations and passing through the most likely state sequence

Viterbi intuition: we are looking for the best ‘path’



Slide from Dekang Lin

Intuition

- The value in each cell is computed by taking the MAX over all paths that lead to this cell.

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

- An extension of a path from state i at time t-1 is computed by multiplying:

$v_{t-1}(i)$	the **previous Viterbi path probability** from the previous time step
a_{ij}	the **transition probability** from previous state q_i to current state q_j
$b_j(o_t)$	the **state observation likelihood** of the observation symbol o_t given the current state j

The Viterbi Algorithm

function VITERBI(*observations* of len T, *state-graph*) **returns** *best-path*

num-states ← NUM-OF-STATES(*state-graph*)
Create a path probability matrix *viterbi[num-states+2,T+2]*
viterbi[0,0] ← 1.0
for each time step t **from** 1 **to** T **do**
  for each state s **from** 1 **to** *num-states* **do**

$$viterbi[s,t] \leftarrow \max_{1 \leq s' \leq num\text{-}states} viterbi[s',t-1] * a_{s',s} * b_s(o_t)$$

$$backpointer[s,t] \leftarrow \operatorname{argmax}_{1 \leq s' \leq num\text{-}states} viterbi[s',t-1] * a_{s',s}$$

Backtrace from highest probability state in final column of *viterbi[]* and return path

The A matrix for the POS HMM

	VB	TO	NN	PPSS
<s>	.019	.0043	.041	.067
VB	.0038	.035	.047	.0070
TO	.83	0	.00047	0
NN	.0040	.016	.087	.0045
PPSS	.23	.00079	.0012	.00014

Figure 4.15 Tag transition probabilities (the a array, $p(t_i|t_{i-1})$ computed from the 87-tag Brown corpus without smoothing. The rows are labeled with the conditioning event; thus $P(PPSS|VB)$ is .0070. The symbol <s> is the start-of-sentence symbol.

What is P(VB|TO)? What is P(NN|TO)? Why does this make sense?

What is P(TO|VB)? What is P(TO|NN)? Why does this make sense?

The B matrix for the POS HMM

	I	want	to	race
VB	0	.0093	0	.00012
TO	0	0	.99	0
NN	0	.000054	0	.00057
PPSS	.37	0	0	0

Figure 4.16 Observation likelihoods (the *b* array) computed from the 87-tag Brown corpus without smoothing.

Look at P(want|VB) and P(want|NN). Give an explanation for the difference in the probabilities.

	VB	TO	NN	PPSS
<s>	.019	.0043	.041	.067
VB	.0038	.035	.047	.0070
TO	.83	0	.00047	0
NN	.0040	.016	.087	.0045
PPSS	.23	.00079	.0012	.00014

Figure 4.15 Tag transition probabilities (the a array, $p(t_i|t_{i-1})$ computed from the 87-tag Brown corpus without smoothing. The rows are labeled with the conditioning event; thus $P(PPSS|VB)$ is .0070. The symbol <s> is the start-of-sentence symbol.

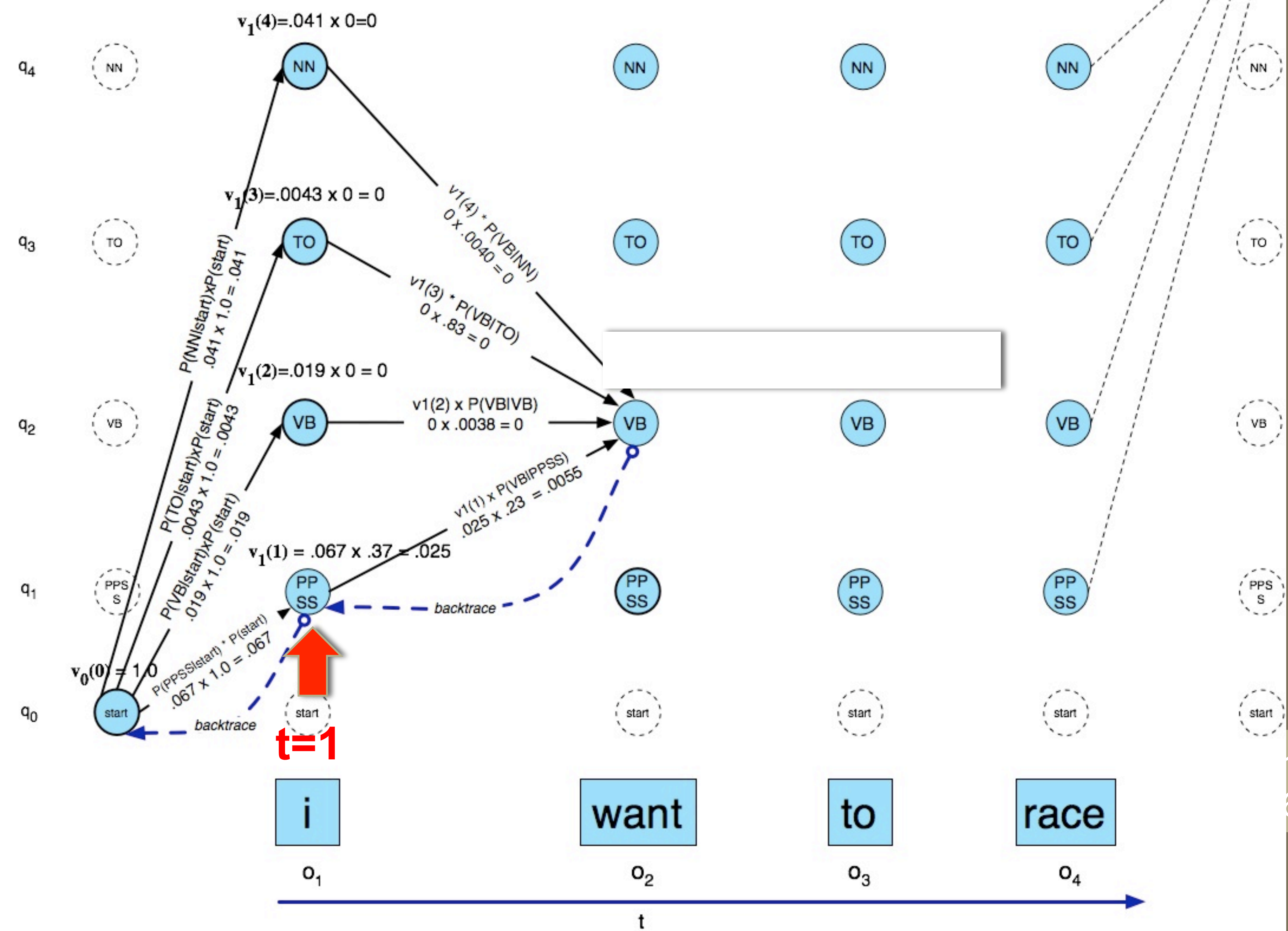
	I	want	to	race
VB	0	.0093	0	.00012
TO	0	0	.99	0
NN	0	.000054	0	.00057
PPSS	.37	0	0	0

Figure 4.16 Observation likelihoods (the b array) computed from the 87-tag Brown corpus without smoothing.

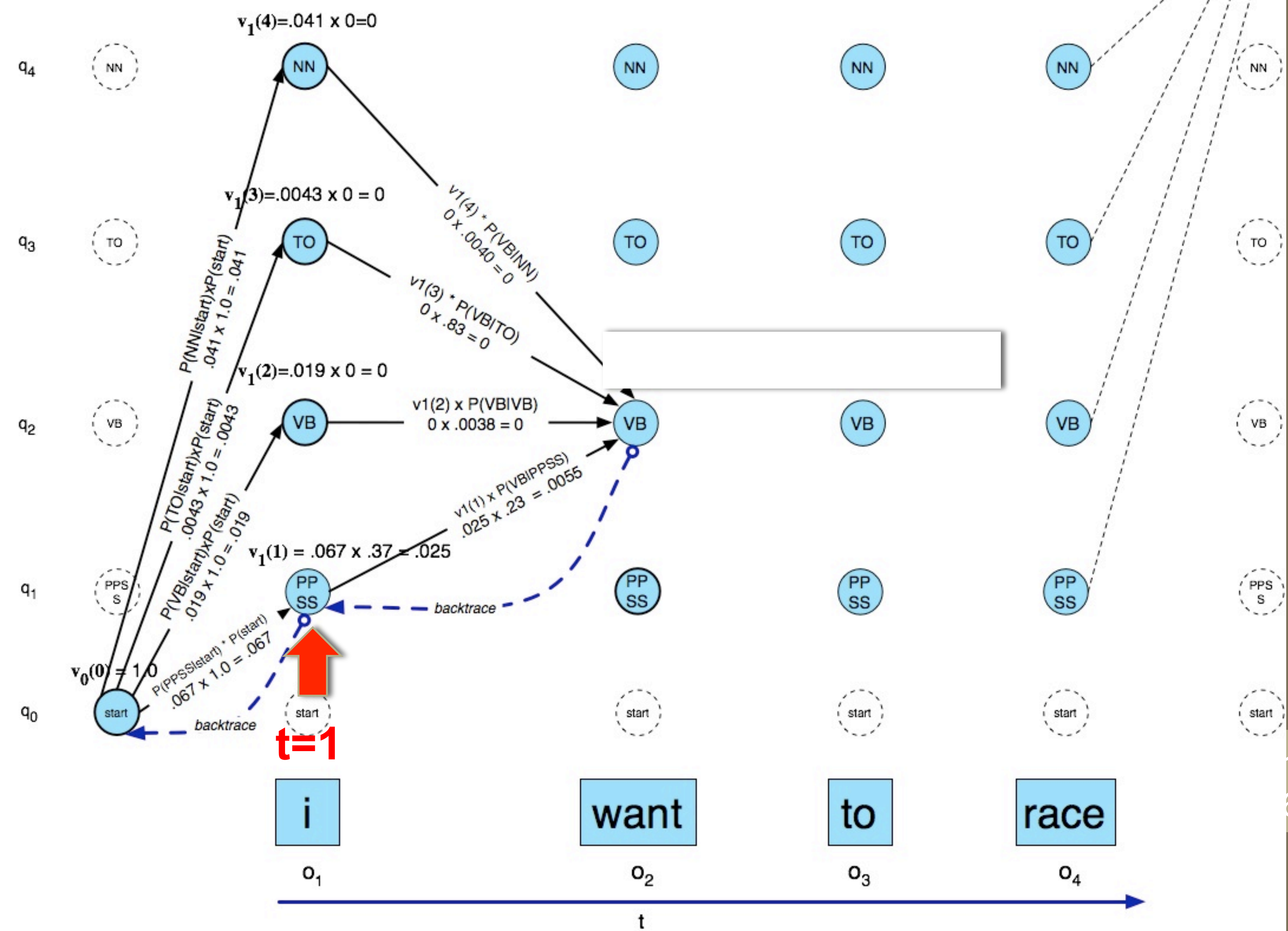
Problem

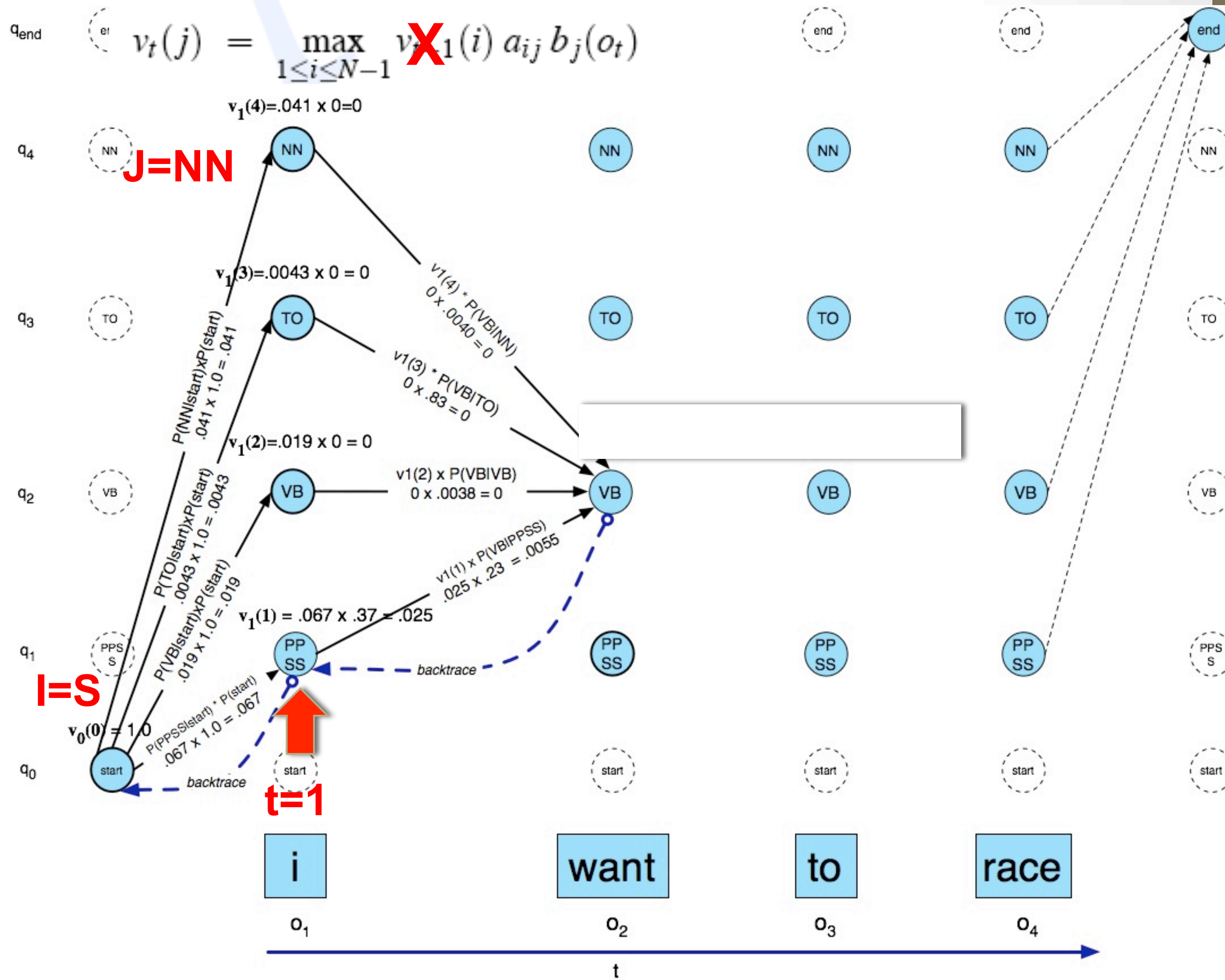
- I want to race (possible states: PPS VB TO NN)

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



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





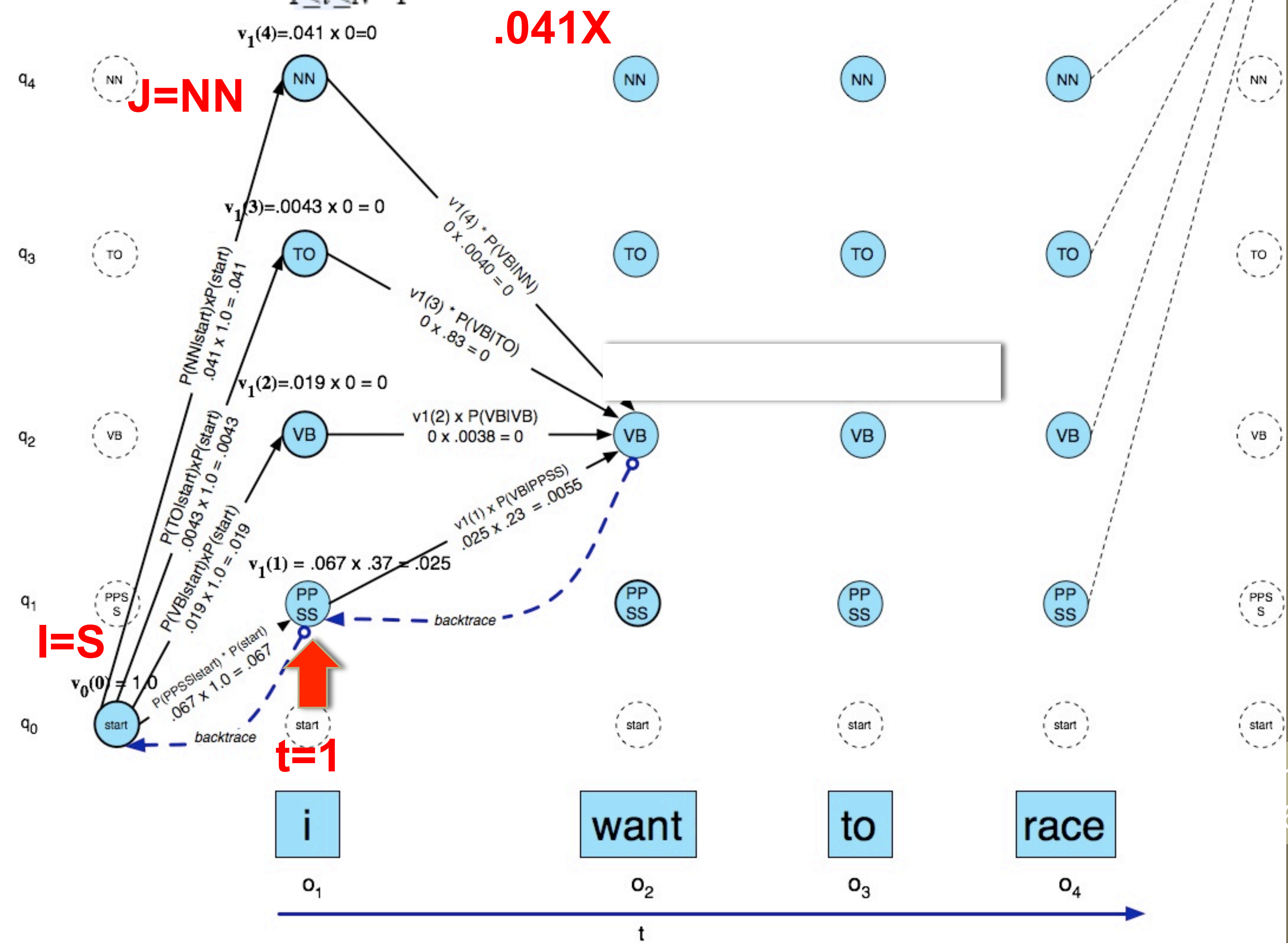
$v_t(j) = \max_{1 \le i \le N-1} v_{t-1}(i)\, a_{ij}\, b_j(o_t)$
X
J=NN
I=S
t=1
q_end
q_4
q_3
q_2
q_1
q_0
end
NN
TO
VB
PPS S
PP SS
start
v_1(4)=.041 x 0=0
v_1(3)=.0043 x 0 = 0
v_1(2)=.019 x 0 = 0
v_1(1) = .067 x .37 = .025
v_0(0) = 1.0
P(NN|start)xP(start) .041 x 1.0 = .041
P(TO|start)xP(start) .0043 x 1.0 = .0043
P(VB|start)xP(start) .019 x 1.0 = .019
P(PPSS|start) * P(start) .067 x 1.0 = .067
v1(4) * P(VB|NN) 0 x .0040 = 0
v1(3) * P(VB|TO) 0 x .83 = 0
v1(2) x P(VB|VB) 0 x .0038 = 0
v1(1) x P(VB|PPSS) .025 x .23 = .0055
backtrace
backtrace
i
want
to
race
o_1
o_2
o_3
o_4
t

The A matrix for the POS HMM

	VB	TO	NN	PPSS
<s>	.019	.0043	.041	.067
VB	.0038	.035	.047	.0070
TO	.83	0	.00047	0
NN	.0040	.016	.087	.0045
PPSS	.23	.00079	.0012	.00014

Figure 4.15 Tag transition probabilities (the a array, $p(t_i|t_{i-1})$ computed from the 87-tag Brown corpus without smoothing. The rows are labeled with the conditioning event; thus $P(PPSS|VB)$ is .0070. The symbol <s> is the start-of-sentence symbol.

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

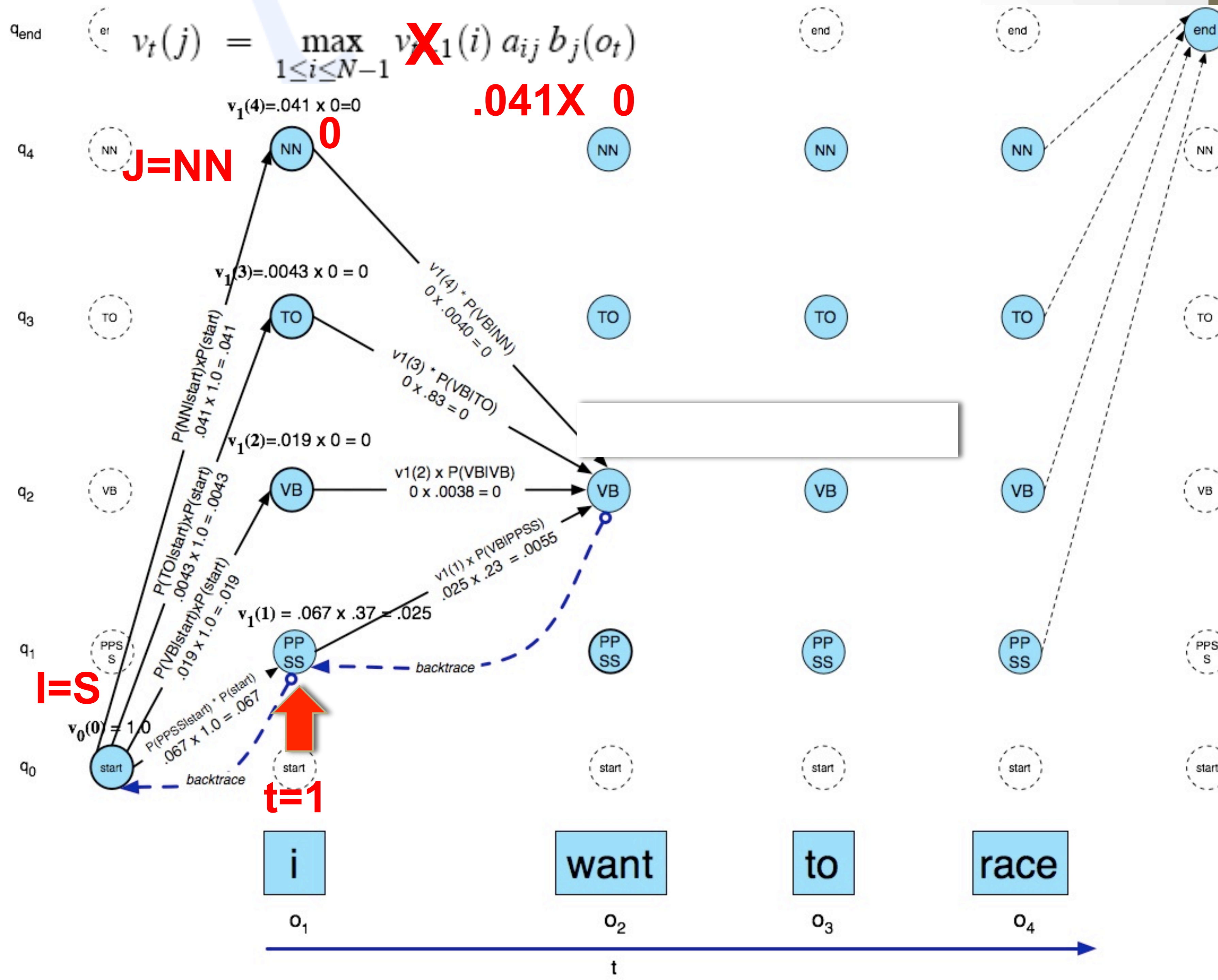


The B matrix for the POS HMM

	I	want	to	race
VB	0	.0093	0	.00012
TO	0	0	.99	0
NN	0	.000054	0	.00057
PPSS	.37	0	0	0

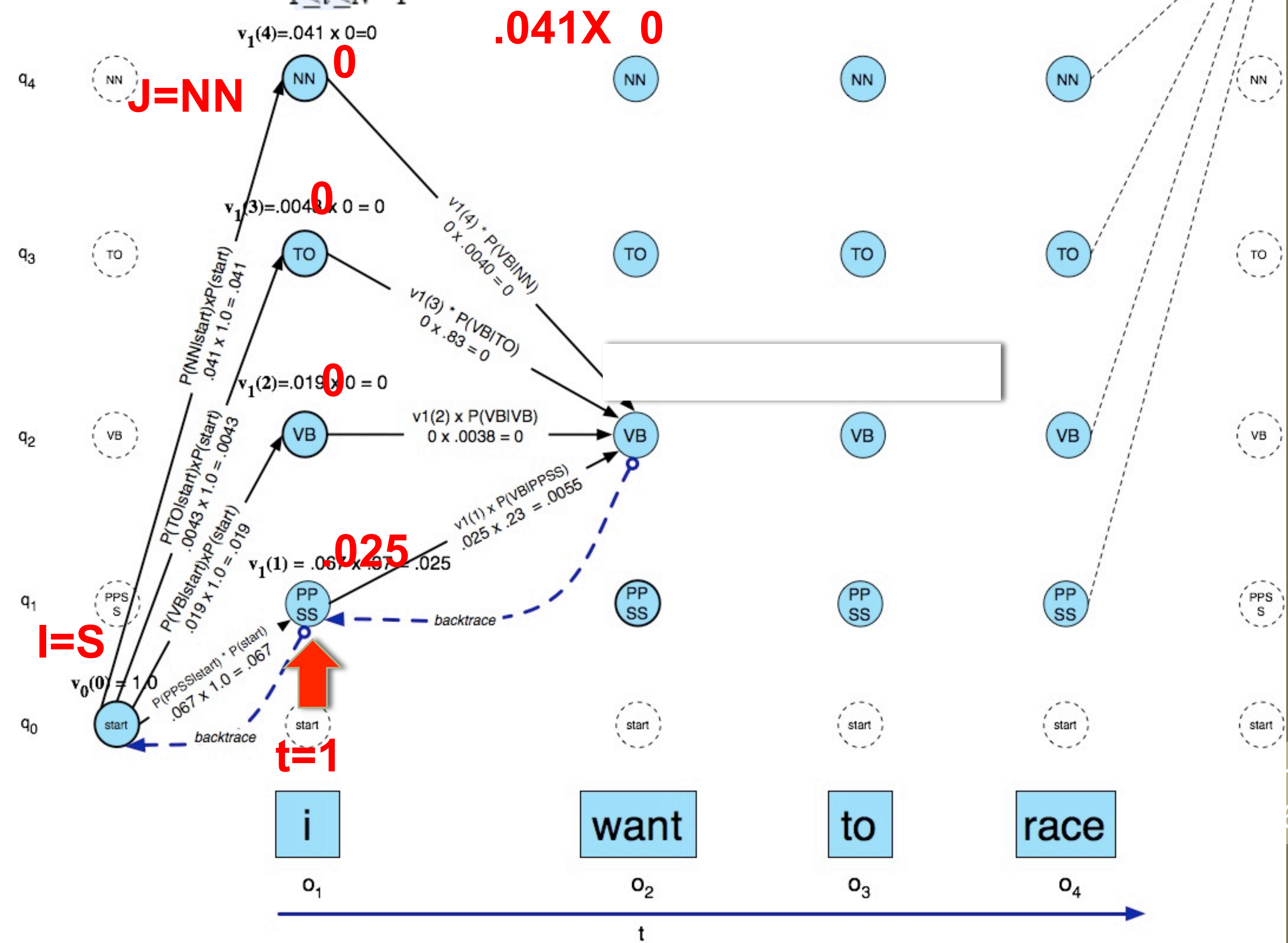
Figure 4.16 Observation likelihoods (the *b* array) computed from the 87-tag Brown corpus without smoothing.

Look at P(want|VB) and P(want|NN). Give an explanation for the difference in the probabilities.



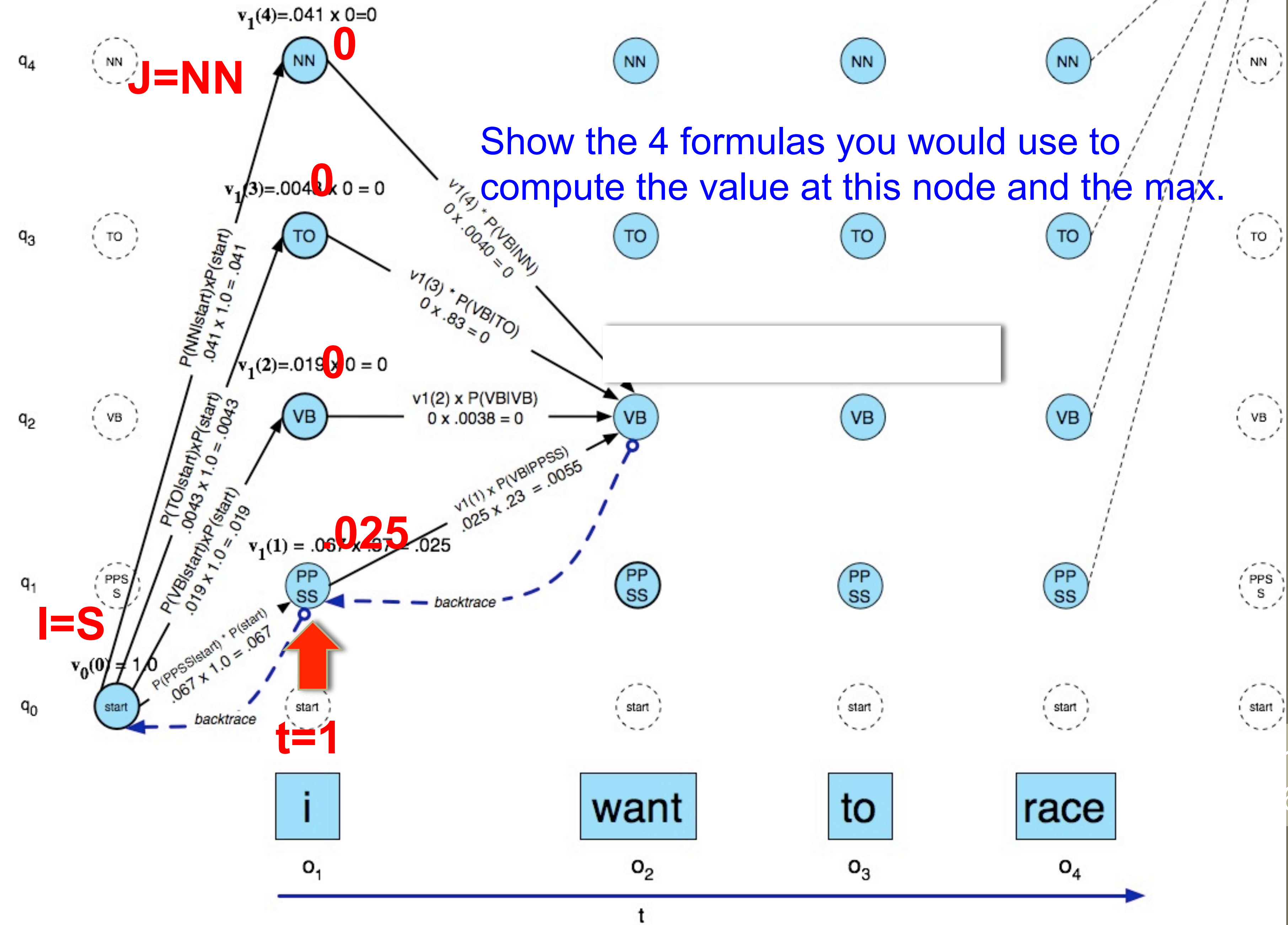
$v_t(j) = \max_{1 \le i \le N-1} v_{t-1}(i)\, a_{ij}\, b_j(o_t)$
X
.041X 0
$v_1(4)$=.041 x 0=0
0
J=NN
$v_1(3)$=.0043 x 0 = 0
$v_1(2)$=.019 x 0 = 0
$v_1(1)$ = .067 x .37 = .025
P(NN|start)xP(start)
.041 x 1.0 = .041
P(TO|start)xP(start)
.0043 x 1.0 = .0043
P(VB|start)xP(start)
.019 x 1.0 = .019
P(PPSS|start) * P(start)
.067 x 1.0 = .067
v1(4) * P(VB|NN)
0 x .0040 = 0
v1(3) * P(VB|TO)
0 x .83 = 0
v1(2) x P(VB|VB)
0 x .0038 = 0
v1(1) x P(VB|PPSS)
.025 x .23 = .0055
backtrace
backtrace
$v_0(0)$ = 1.0
I=S
t=1
q_{end}
q_4
q_3
q_2
q_1
q_0
end
NN
TO
VB
PPSS
start
i
want
to
race
o_1
o_2
o_3
o_4
t

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



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

Show the 4 formulas you would use to compute the value at this node and the max.



Dependency Parsing

Dependency parsing

- An example from the NY Times today:

Last week, on the third floor of a small building in San Francisco’s Mission District, a woman scrambled the tiles of a Rubik’s Cube

Dependency parsing

- An example from the NY Times today:

Last week, on the third floor of a small building in San Francisco's Mission District, a woman scrambled the tiles of a Rubik's Cube

Dependency parsing

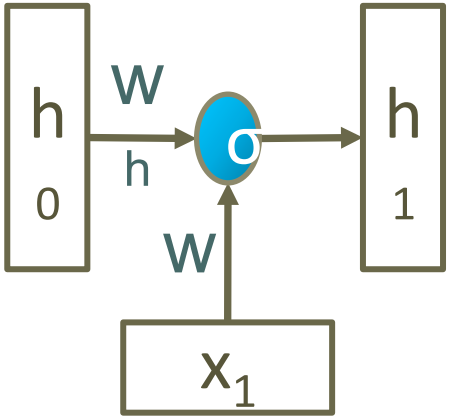
- An example from the NY Times today:

Last week, on the third floor, a woman scrambled the tiles of a Rubik's Cube

RNNs and LSTMs

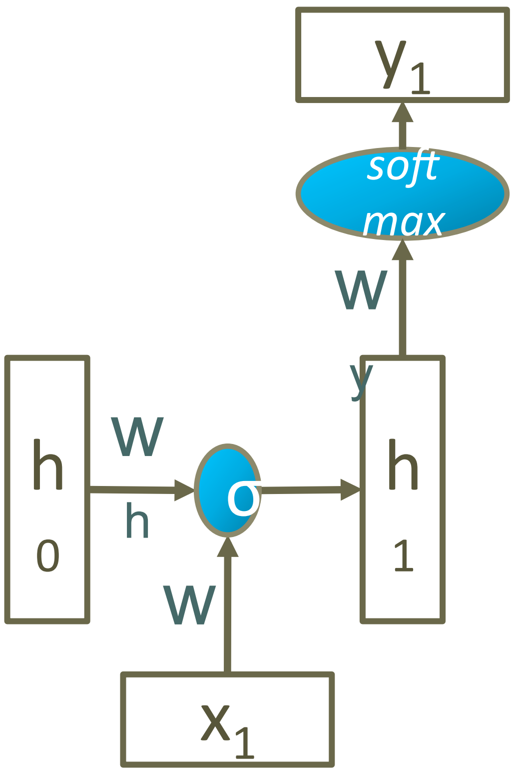
Recurrent Neural Networks

$$h_t = \sigma(W_h h_{t-1} + W_x x_t)$$



Slide from Radev

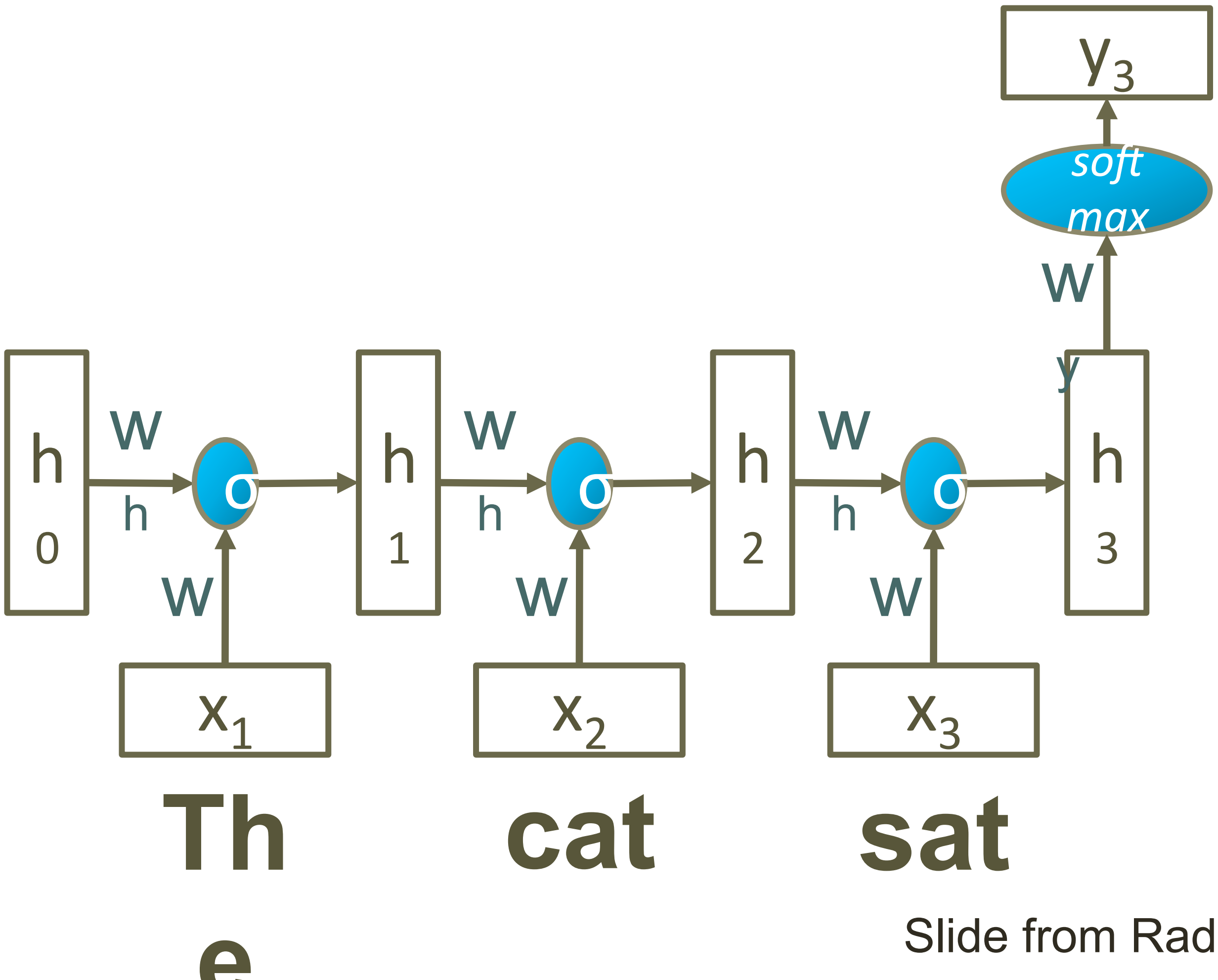
RNN



$$h_t = \sigma(W_h h_{t-1} + W_x x_t)$$
$$y_t = softmax(W_y h_t)$$

Slide from Radev

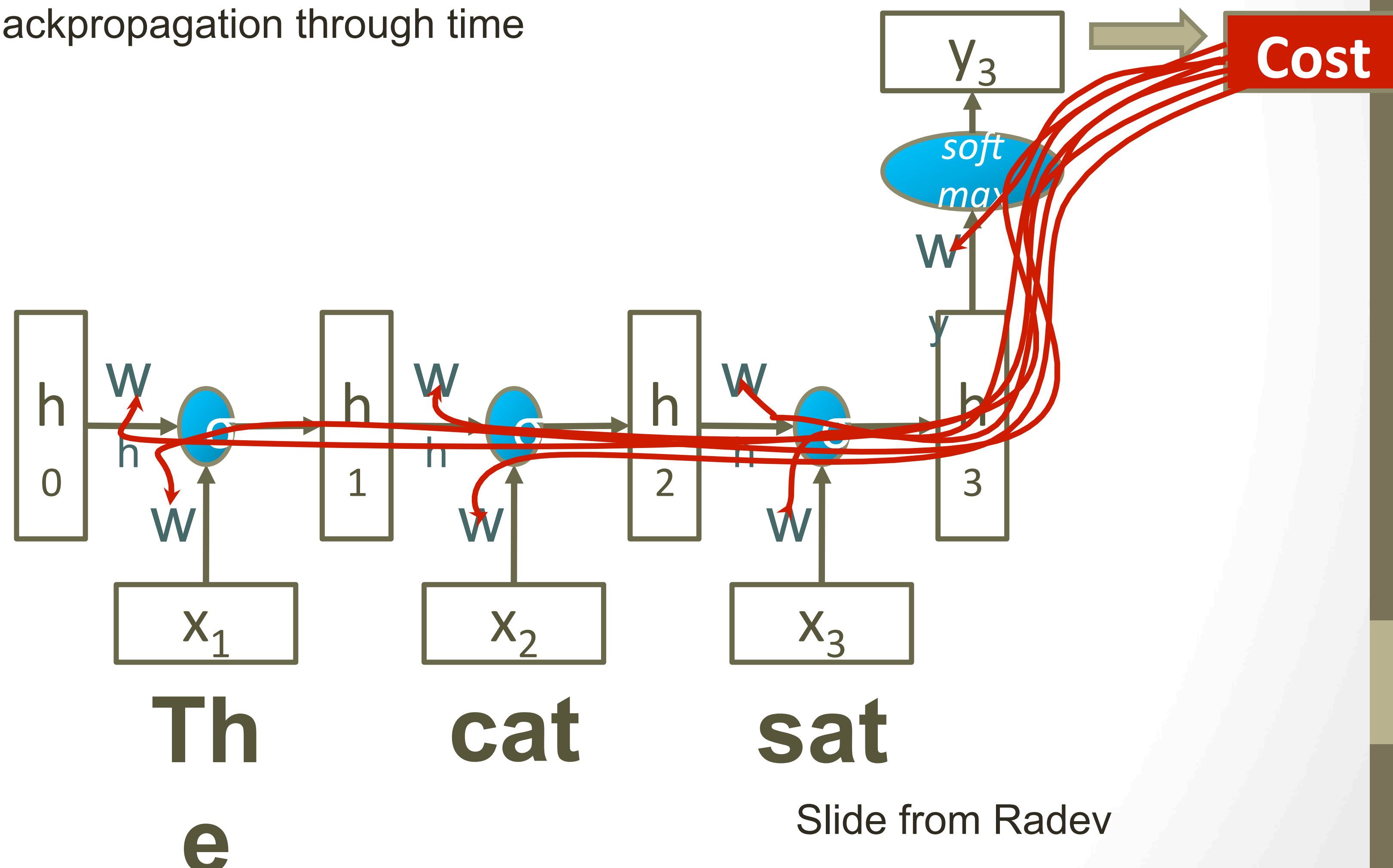
RNN



Slide from Radev

Updating Parameters of an RNN

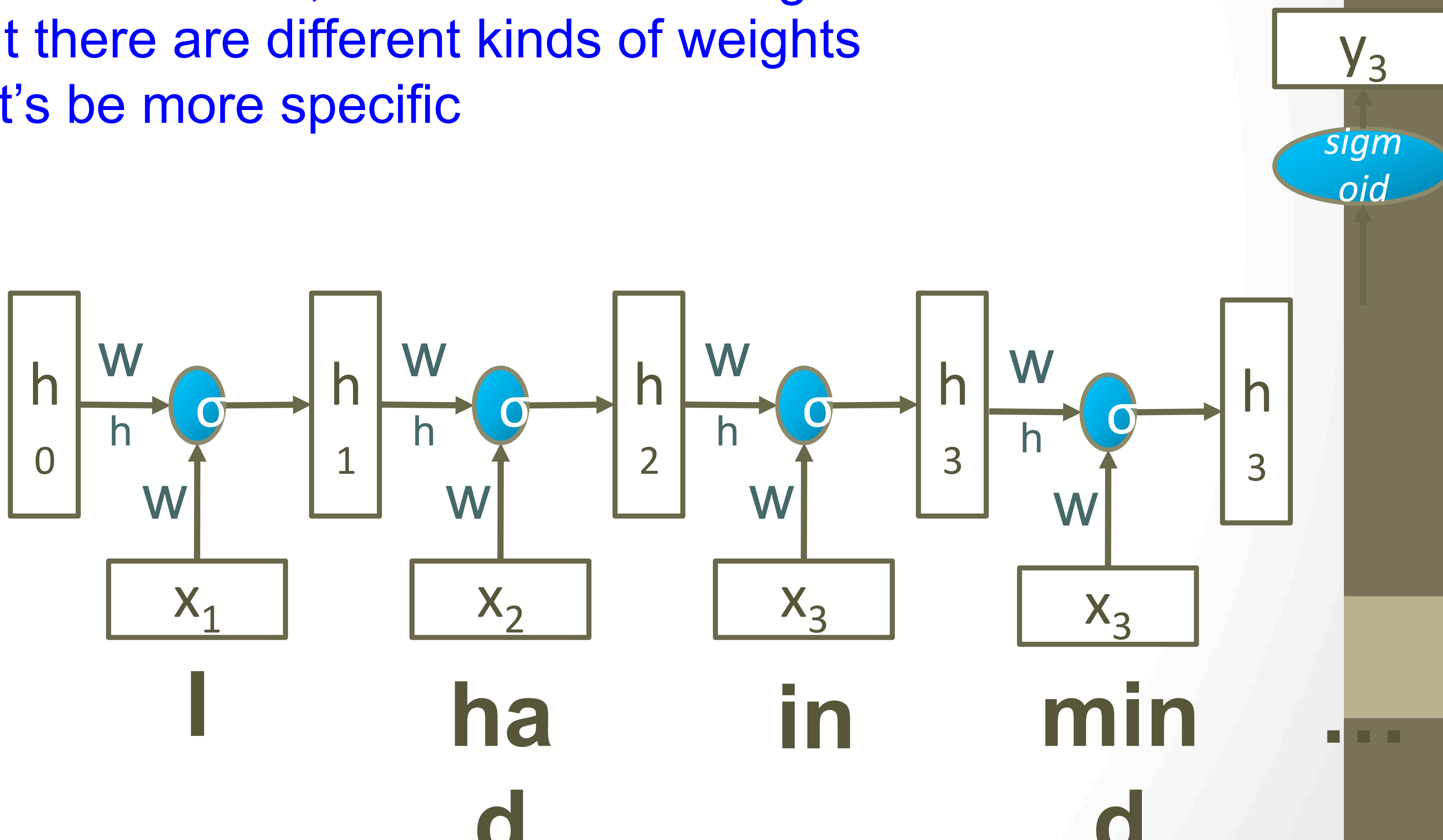
Backpropagation through time



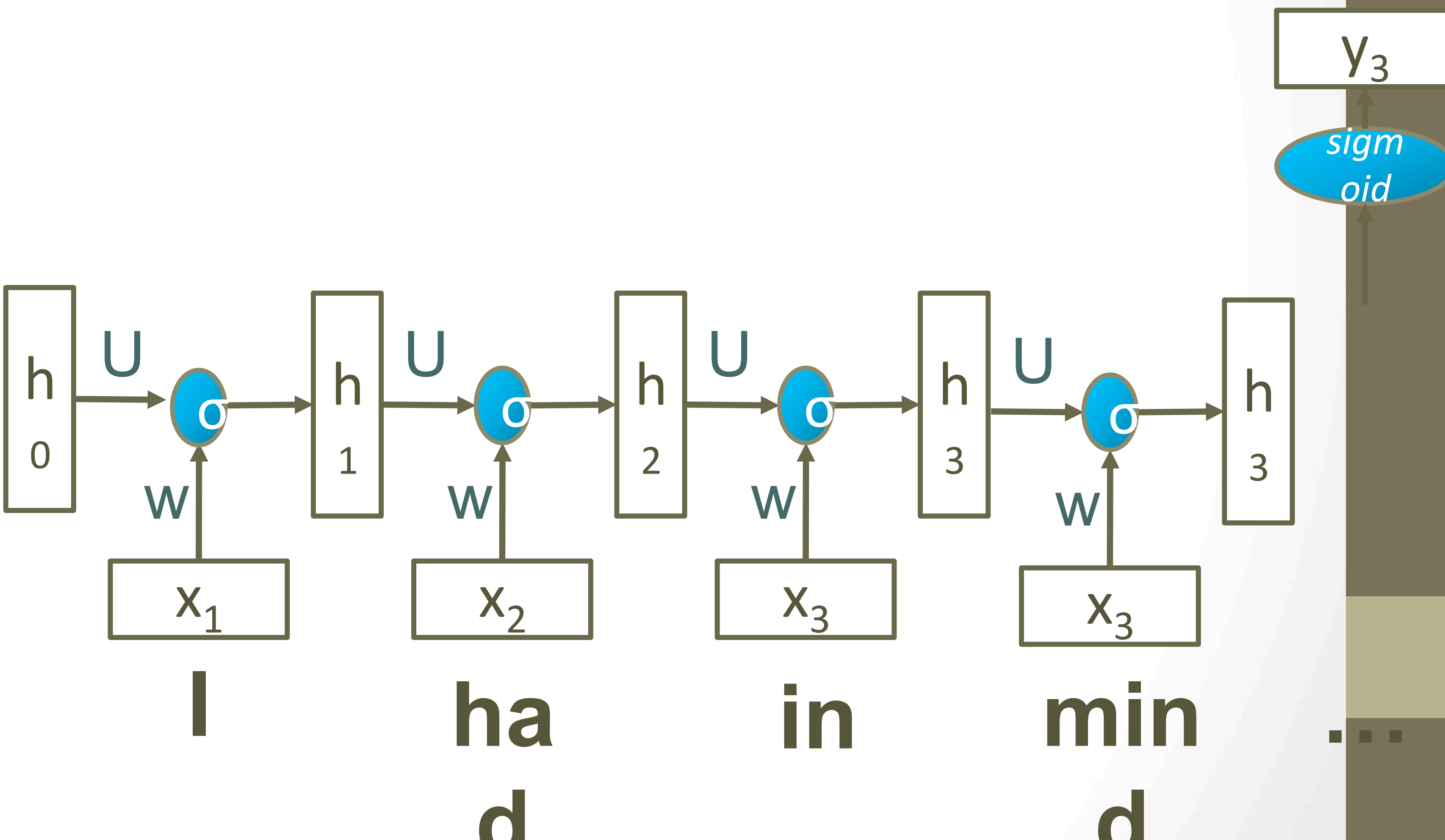
Slide from Radev

RNN – I had in mind your facts, buddy, not hers.

In this overview, w refers to the weights
But there are different kinds of weights
Let's be more specific

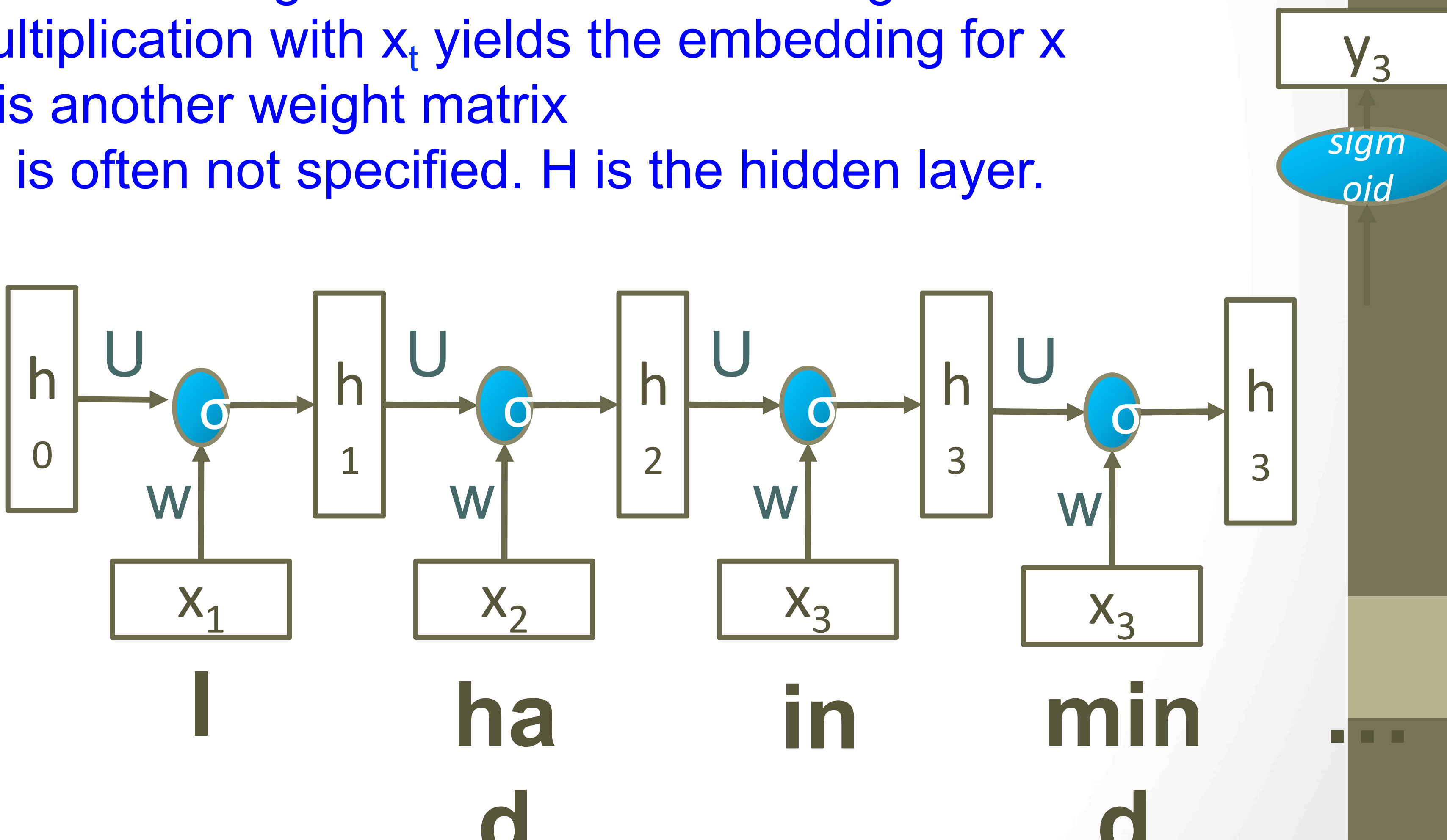


RNN – I had in mind your facts, buddy, not hers.



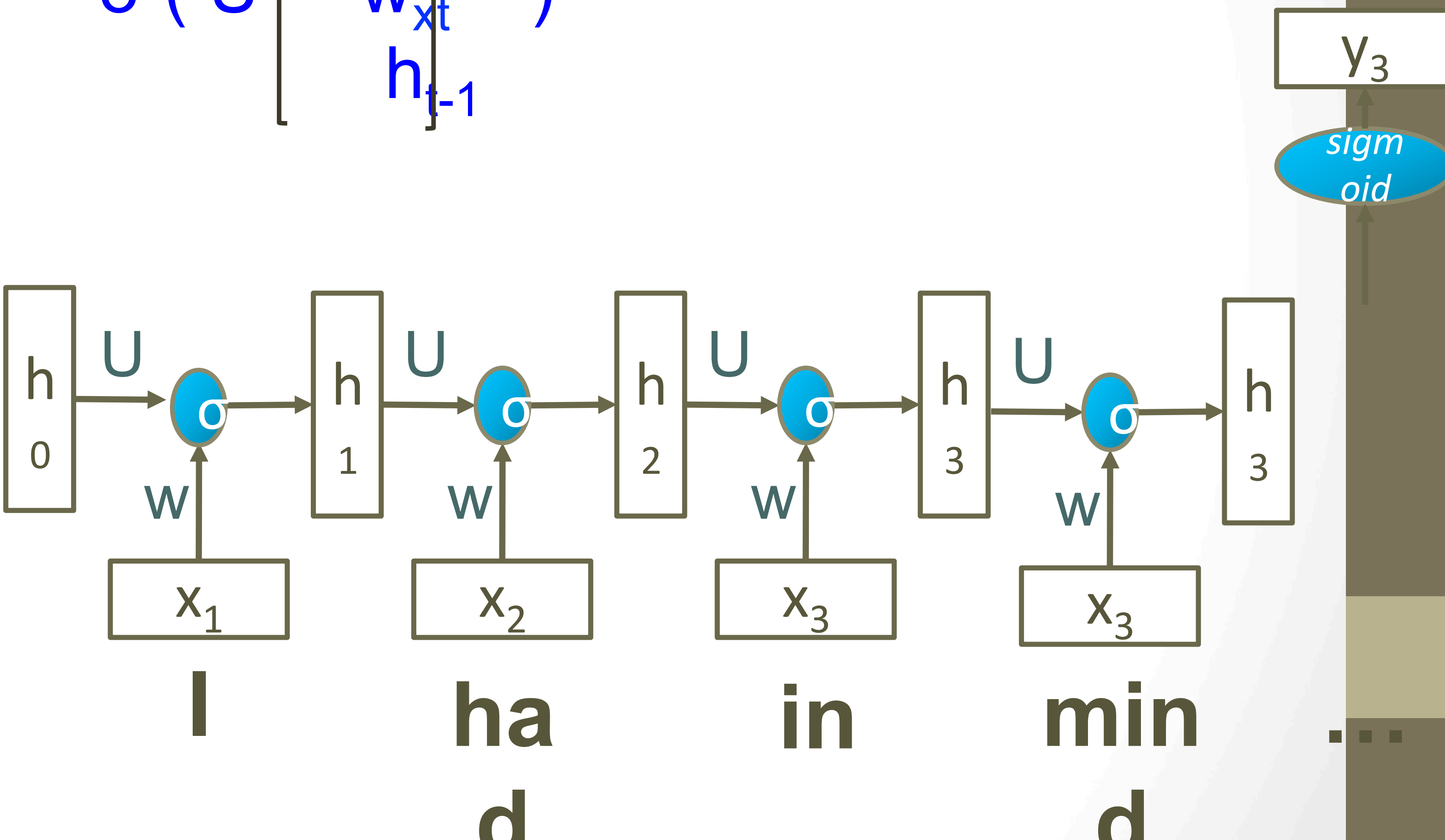
RNN – I had in mind your facts, buddy, not hers.

W are the weights: the word embedding matrix multiplication with x_t yields the embedding for x
U is another weight matrix
H_0 is often not specified. H is the hidden layer.



RNN – I had in mind your facts, buddy, not hers.

$$h_t = \sigma \left(U \begin{bmatrix} w_{xt} \\ h_{t-1} \end{bmatrix} \right)$$



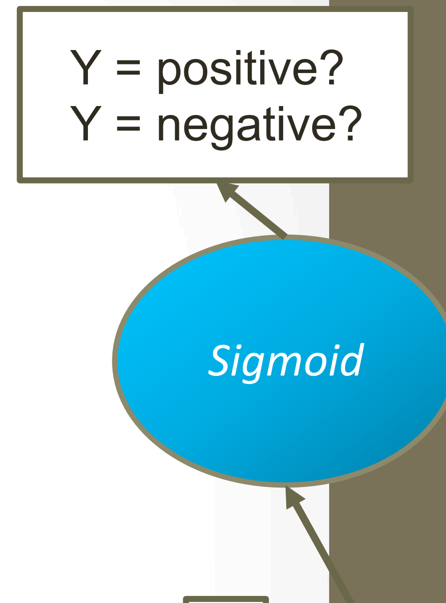
RNN – I had in mind your facts, buddy, not hers.

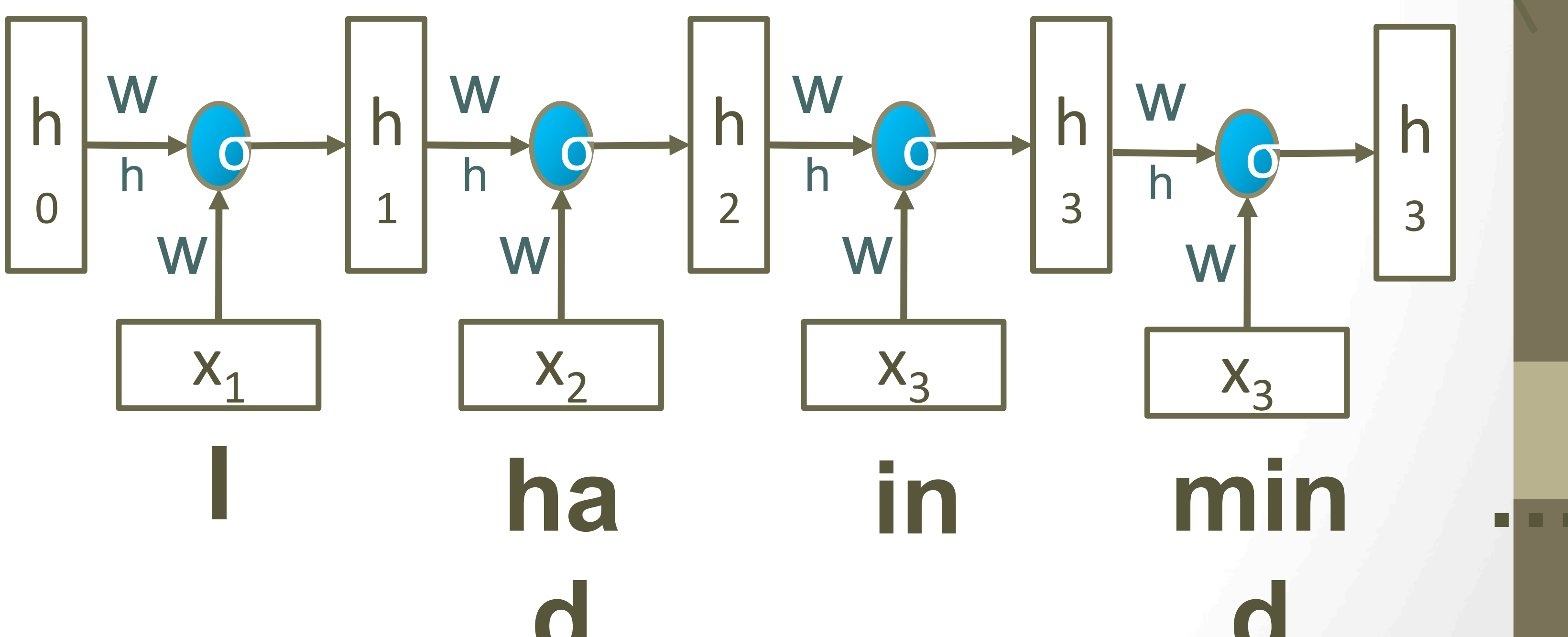
Final embedding run through the sigmoid function -> [0,1]

1 = positive

0= negative

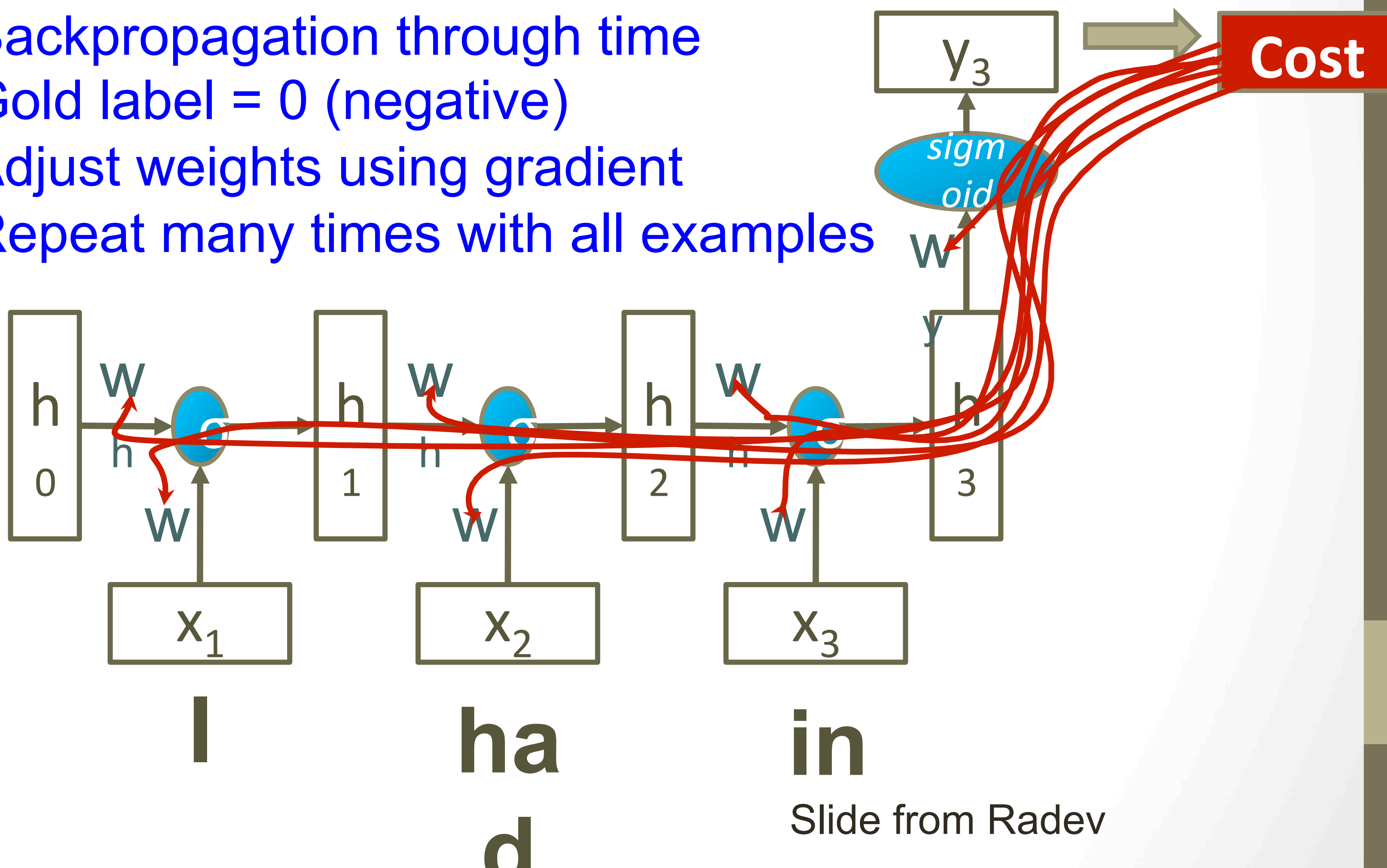
Often final h is used as word embedding for the sentence





Updating Parameters of an RNN

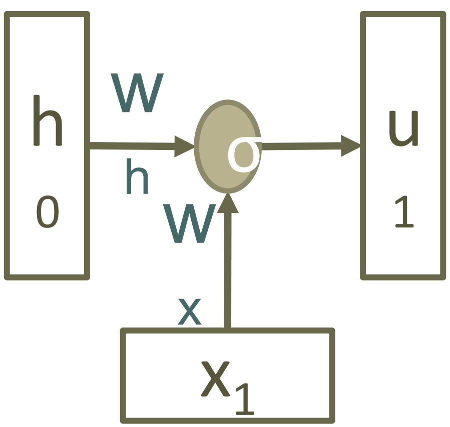
Backpropagation through time
Gold label = 0 (negative)
Adjust weights using gradient
Repeat many times with all examples



Slide from Radev

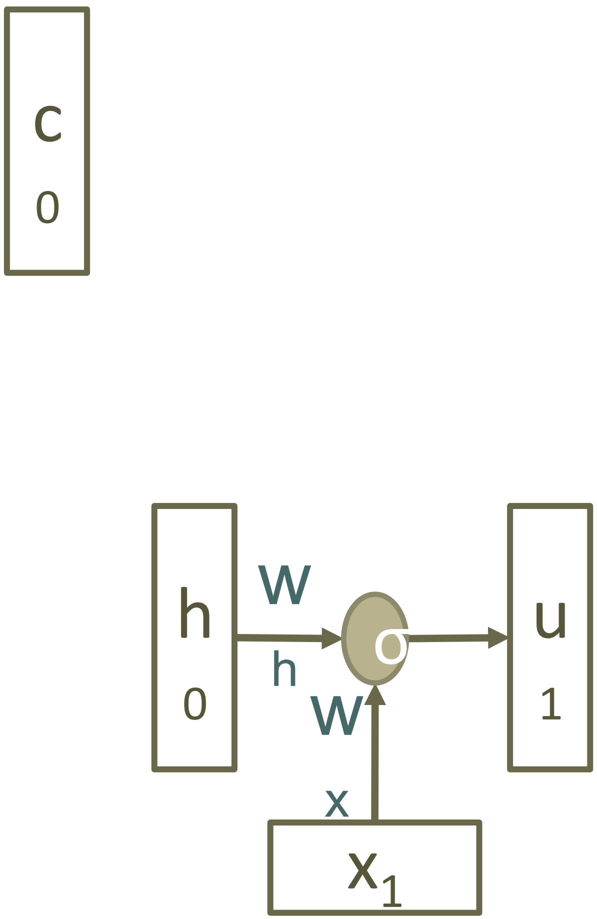
Transforming RNN to LSTM

$$u_t = \sigma(W_h h_{t-1} + W_x x_t)$$



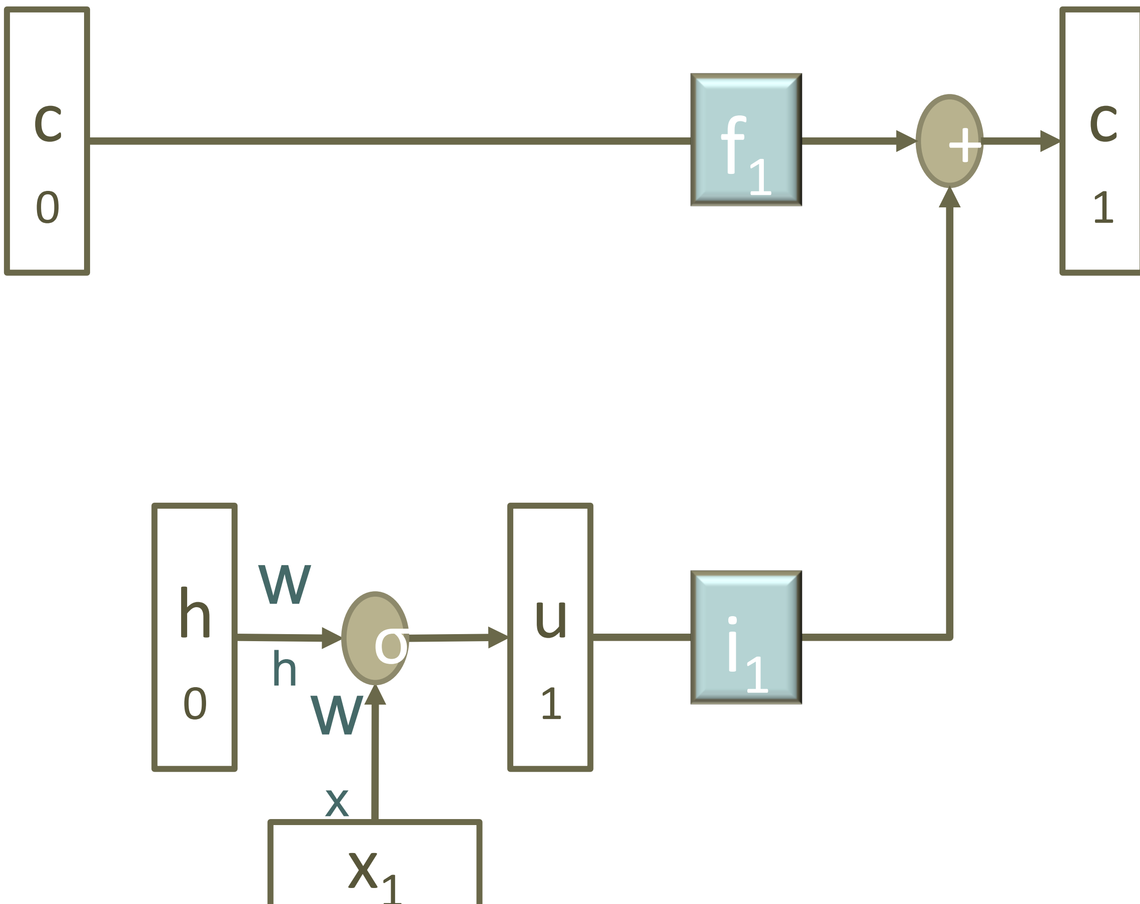
[slides from Catherine Finegan-Dollak]

Transforming RNN to LSTM



[slides from Catherine Finegan-Dollak]

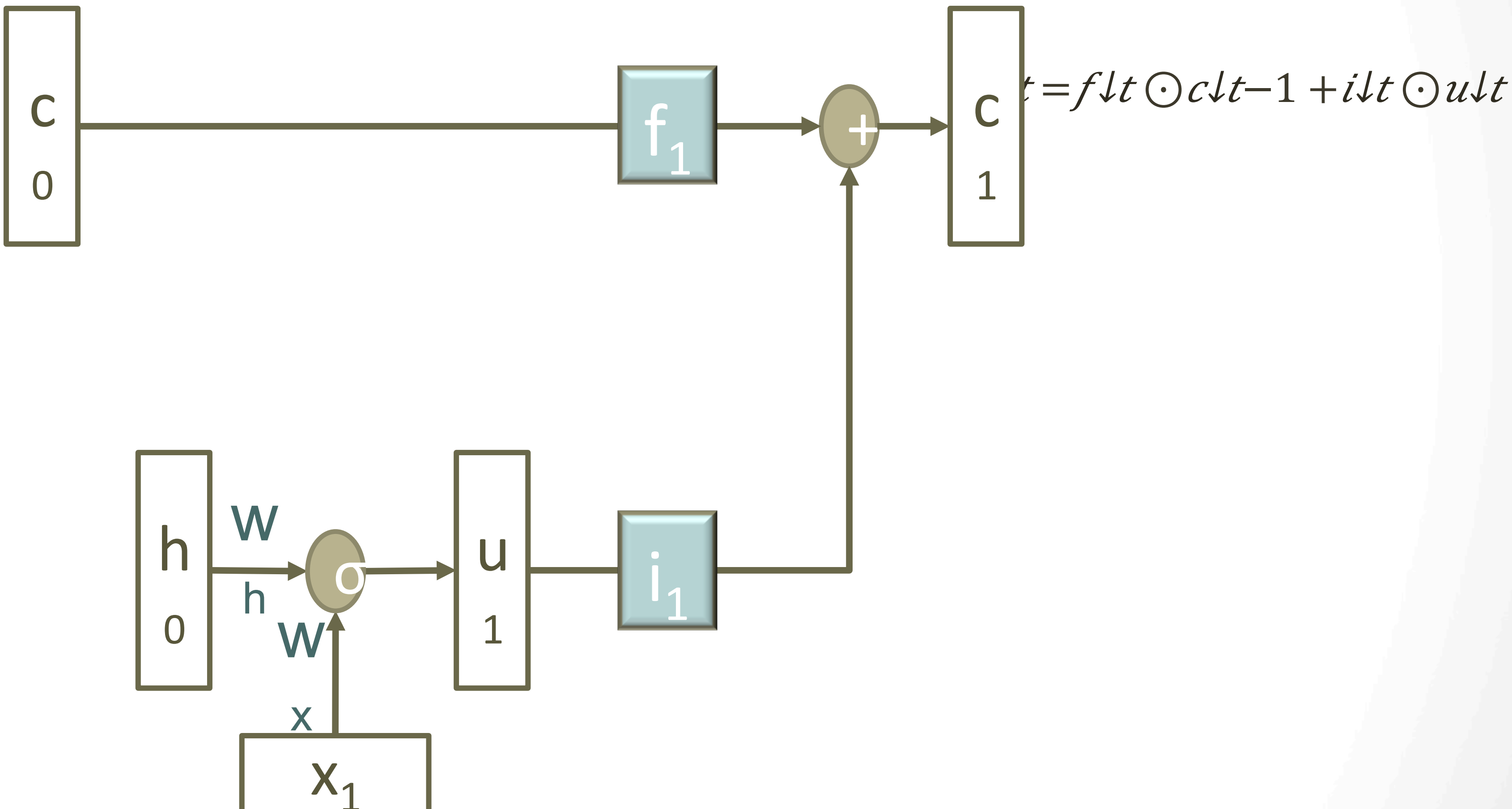
Transforming RNN to LSTM



$t = f_t \odot c_{t-1} + i_t \odot u_t$

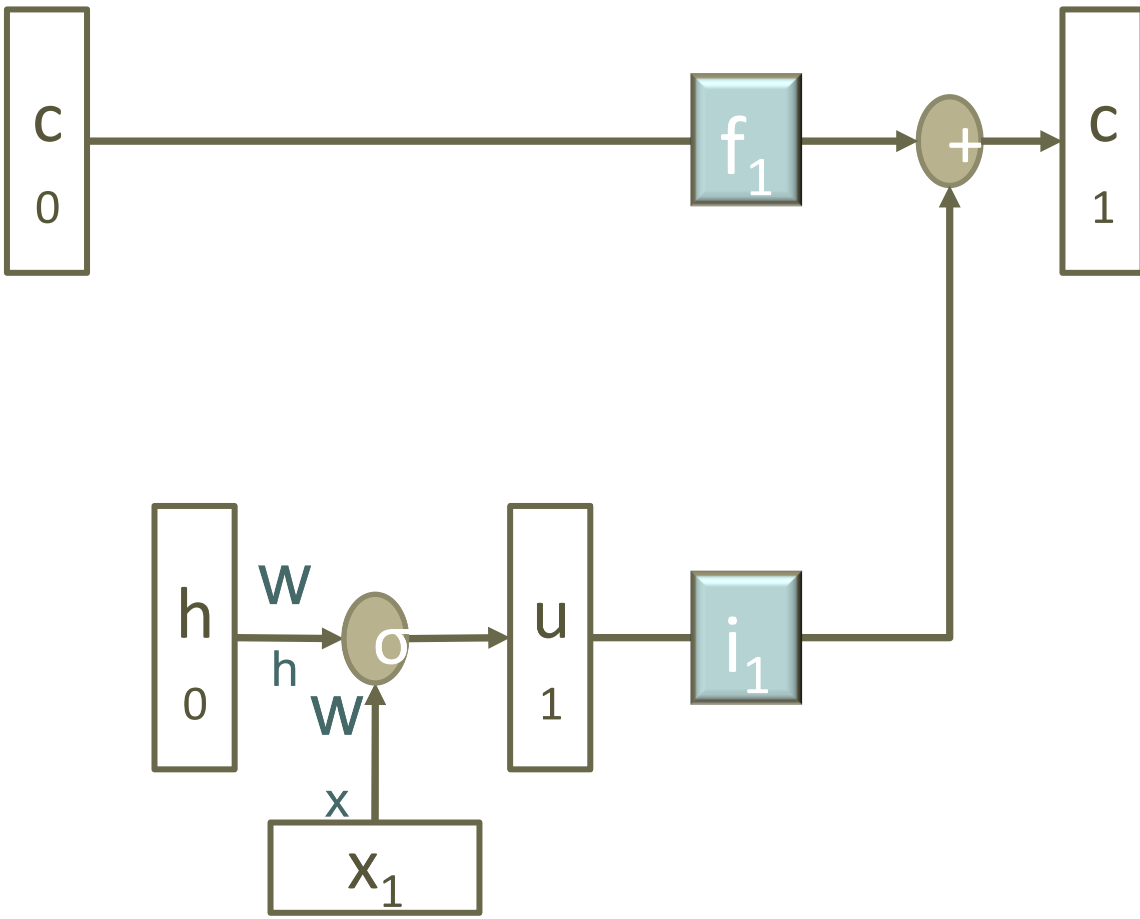
[slides from Catherine Finegan-Dollak]

Transforming RNN to LSTM



[slides from Catherine Finegan-Dollak]

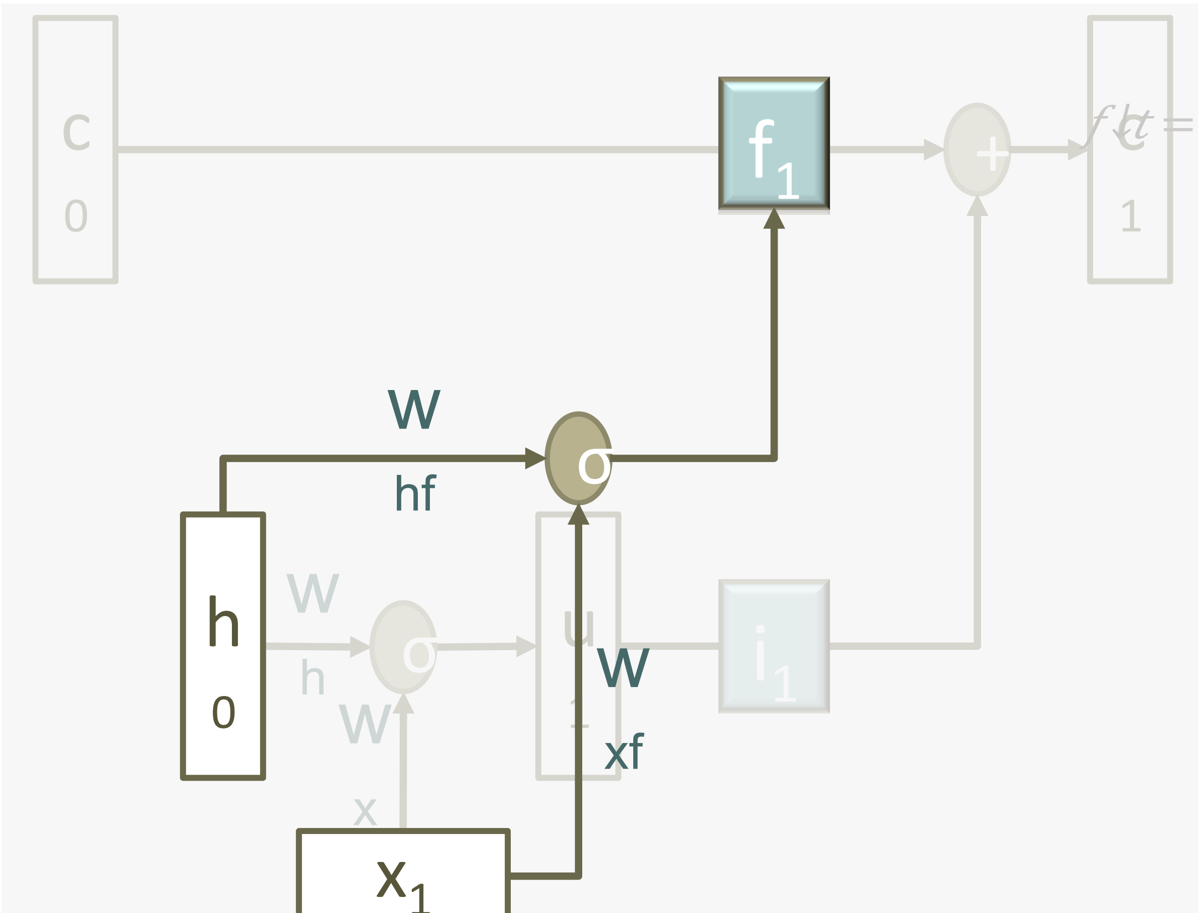
Transforming RNN to LSTM



$t = f_t \odot c_{t-1} + i_t \odot u_t$

[slides from Catherine Finegan-Dollak]

Transforming RNN to LSTM

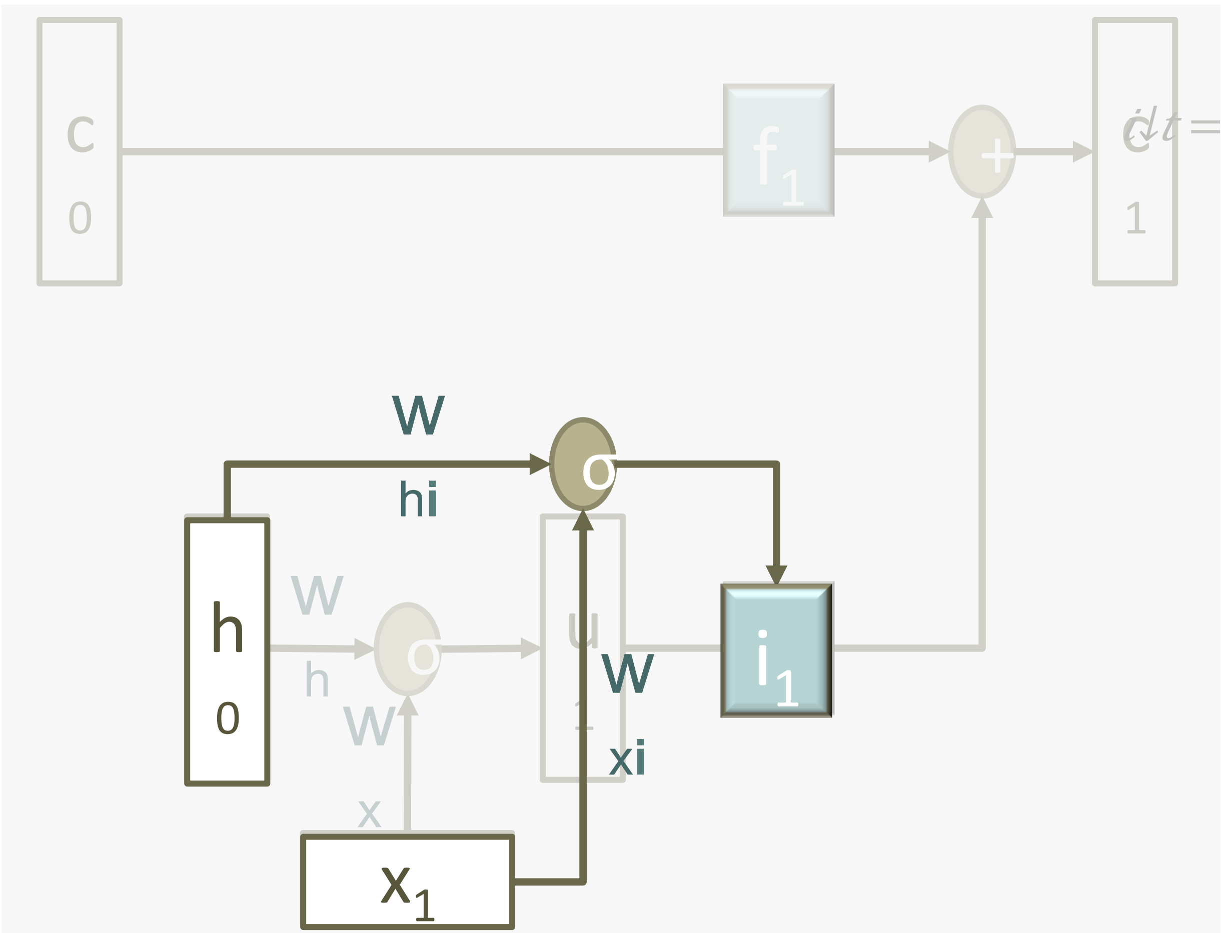


$$f_t = \sigma(W_{hf} h_{t-1} + W_{xf} x_t)$$

[slides from Catherine Finegan-Dollak]

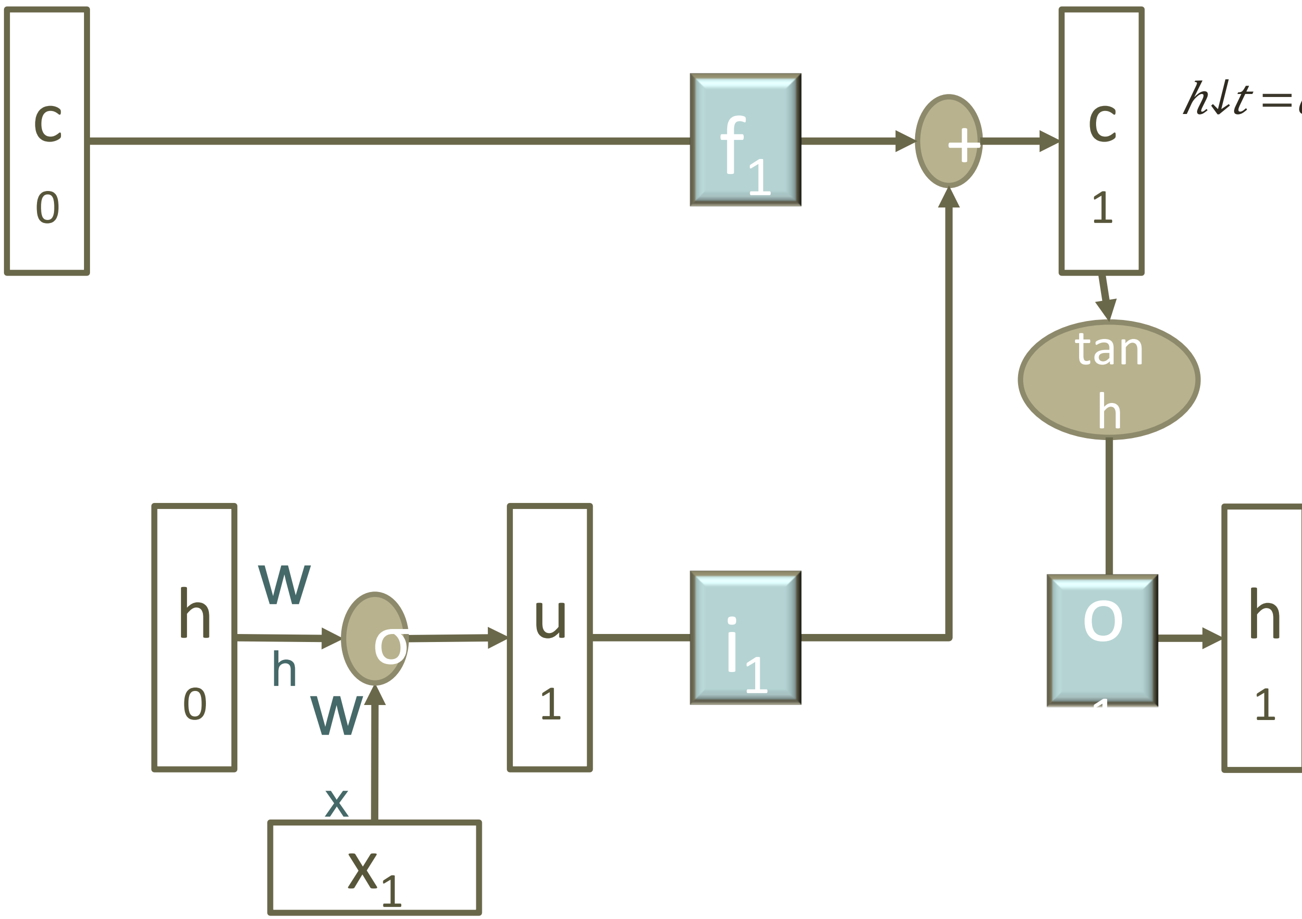
Transforming RNN to LSTM

$$i_t = \sigma(W_{hi} h_{t-1} + W_{xi} x_t)$$



[slides from Catherine Finegan-Dollak]

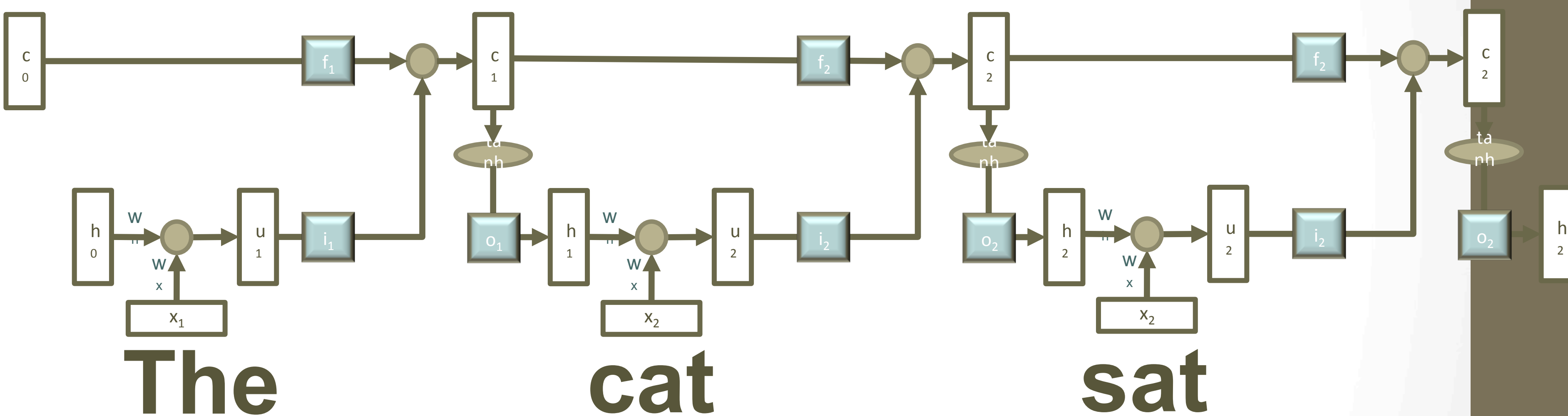
Transforming RNN to LSTM



$$h_t = o_t \odot \tanh c_t$$

[slides from Catherine Finegan-Dollak]

LSTM for Sequences



[slides from Catherine Finegan-Dollak]

Problem 10 from sample midterm questions