Special Topics in Natural Language Processing

CS698O

Ashutosh Modi
CSE Department, IIT Kanpur

Lecture 10: Sequence Prediction 1
Jan 27, 2020

Sequence Models

- Language has sequential nature.
- Sentences are composed of words in a sequence
- Sequence Models try to capture the sequential nature of language
- Example:
  - POS Tagging
  - NER
  - Speech Recognition
- Sequence Models come in many flavors:
  - HMM
  - MEMM
  - CRF
  - RNN

Hidden Markov Models (HMM)

Hidden Markov Models (HMM)

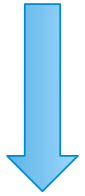
- Consider the task of Part Of Speech (POS) tagging

The dog saw a cat

Hidden Markov Models (HMM)

- Consider the task of Part Of Speech (POS) tagging

The dog saw a cat



Determiner Noun Verb Determiner Noun

The dog saw a cat

Hidden Markov Models (HMM)

This can be modeled via a generative process

Hidden Markov Models (HMM)

This can be modeled via a generative process

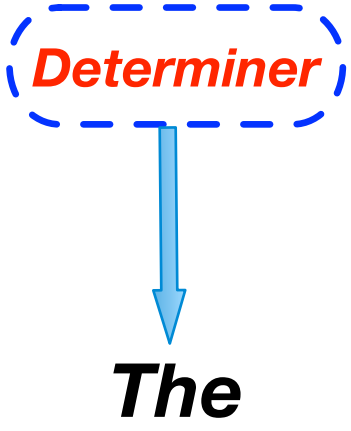
STEP 1: Select a part of speech category ($p(y)$)

Hidden Markov Models (HMM)

This can be modeled via a generative process

STEP 1: Select a part of speech category ($p(y)$)

STEP 2: Based on the part of speech select a word ($p(x \mid y)$)



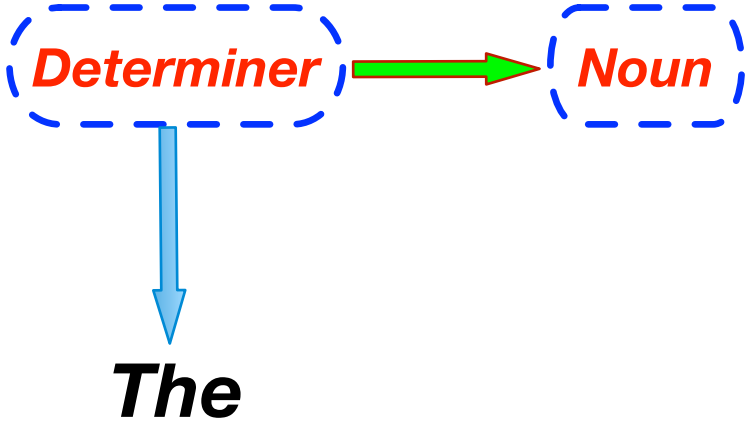
Hidden Markov Models (HMM)

This can be modeled via a generative process

STEP 1: Select a part of speech category ($p(y)$)

STEP 2: Based on the part of speech select a word ($p(x \mid y)$)

STEP 3: Select the next category ($p(y_{new} \mid all\ old\ y's)$)



Hidden Markov Models (HMM)

This can be modeled via a generative process

STEP 1: Select a part of speech category ($p(y)$)

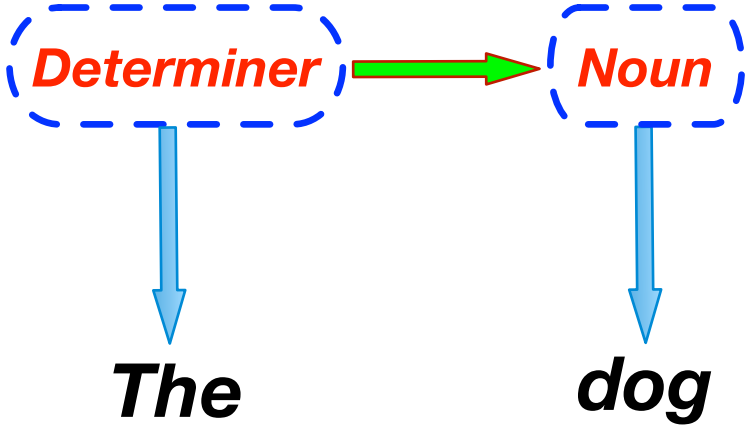
STEP 2: Based on the part of speech select a word ($p(x \mid y)$)

STEP 3: Select the next category ($p(y_{new} \mid all\ old\ y's)$)

If category is not STOP

go to STEP 2

else terminate



Hidden Markov Models (HMM)

This can be modeled via a generative process

STEP 1: Select a part of speech category ($p(y)$)

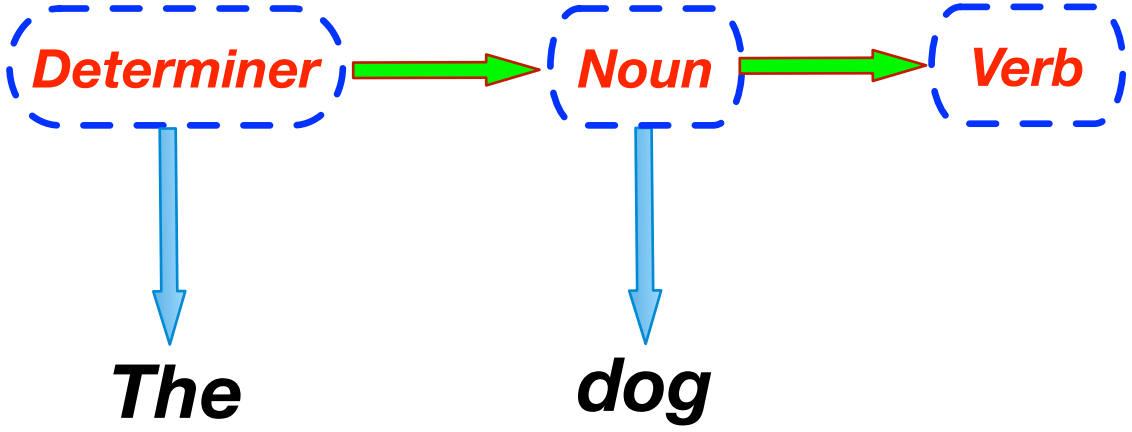
STEP 2: Based on the part of speech select a word ($p(x \mid y)$)

STEP 3: Select the next category ($p(y_{new} \mid all\ old\ y's)$)

If category is not STOP

go to STEP 2

else terminate



Hidden Markov Models (HMM)

This can be modeled via a generative process

STEP 1: Select a part of speech category ($p(y)$)

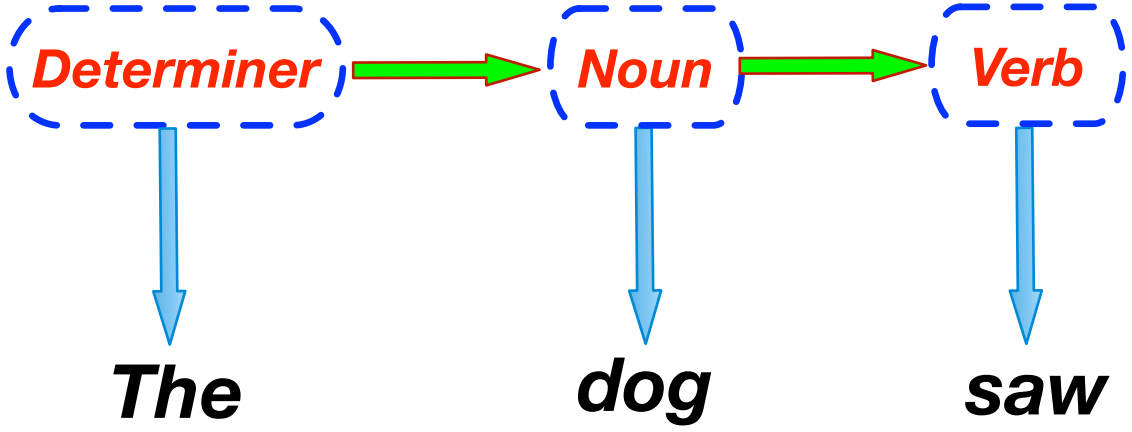
STEP 2: Based on the part of speech select a word ($p(x \mid y)$)

STEP 3: Select the next category ($p(y_{new} \mid all\ old\ y's)$)

If category is not STOP

go to STEP 2

else terminate



Hidden Markov Models (HMM)

This can be modeled via a generative process

STEP 1: Select a part of speech category ($p(y)$)

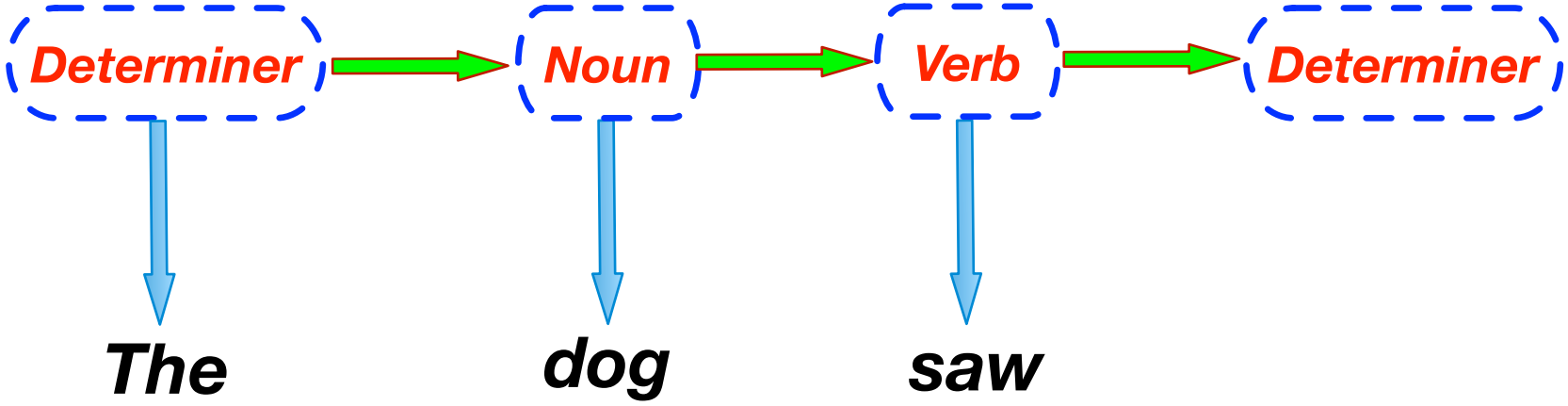
STEP 2: Based on the part of speech select a word ($p(x \mid y)$)

STEP 3: Select the next category ($p(y_{new} \mid all\ old\ y's)$)

If category is not STOP

go to STEP 2

else terminate



Hidden Markov Models (HMM)

This can be modeled via a generative process

STEP 1: Select a part of speech category ($p(y)$)

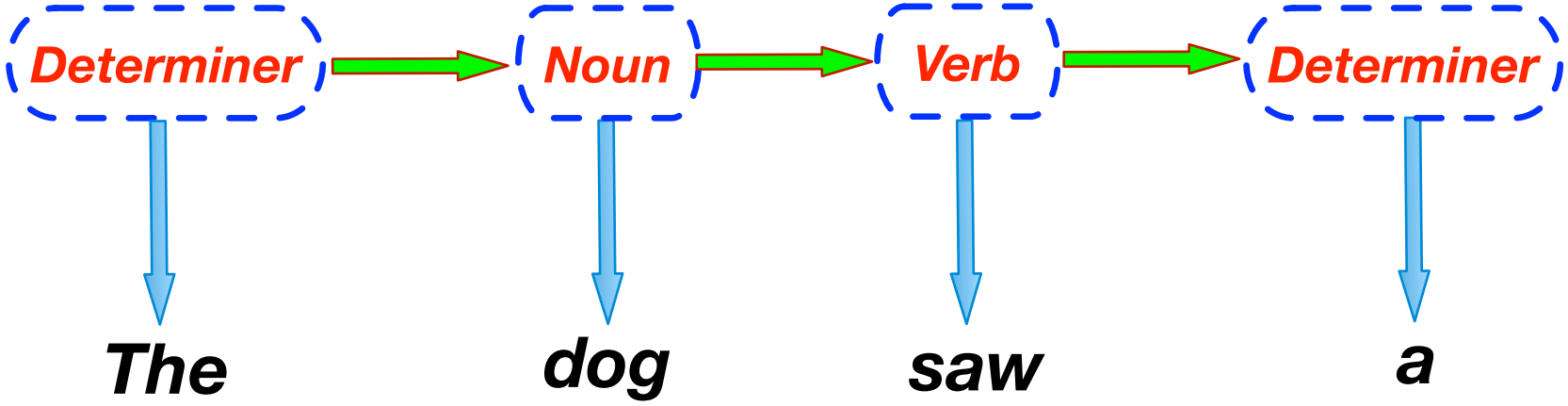
STEP 2: Based on the part of speech select a word ($p(x \mid y)$)

STEP 3: Select the next category ($p(y_{new} \mid all\ old\ y's)$)

If category is not STOP

go to STEP 2

else terminate



Hidden Markov Models (HMM)

This can be modeled via a generative process

STEP 1: Select a part of speech category ($p(y)$)

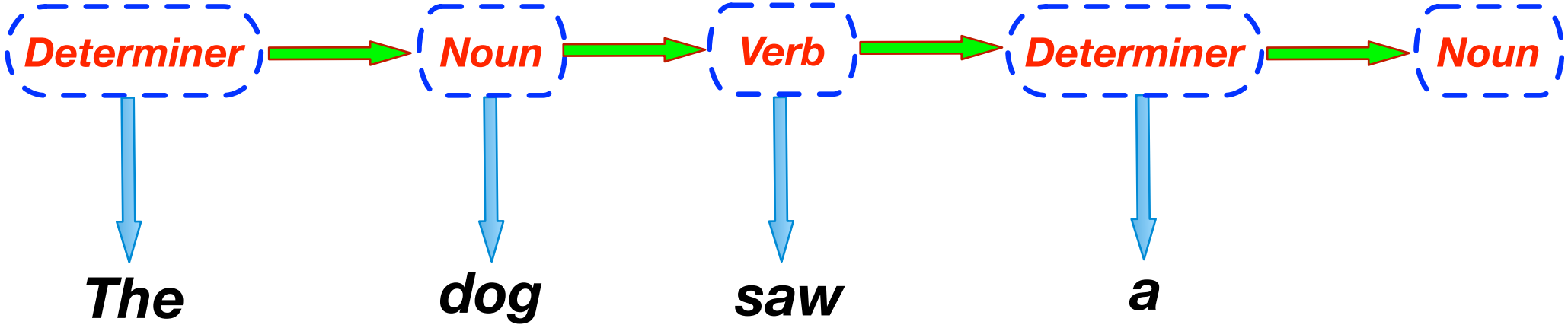
STEP 2: Based on the part of speech select a word ($p(x \mid y)$)

STEP 3: Select the next category ($p(y_{new} \mid all\ old\ y's)$)

If category is not STOP

go to STEP 2

else terminate



Hidden Markov Models (HMM)

This can be modeled via a generative process

STEP 1: Select a part of speech category ($p(y)$)

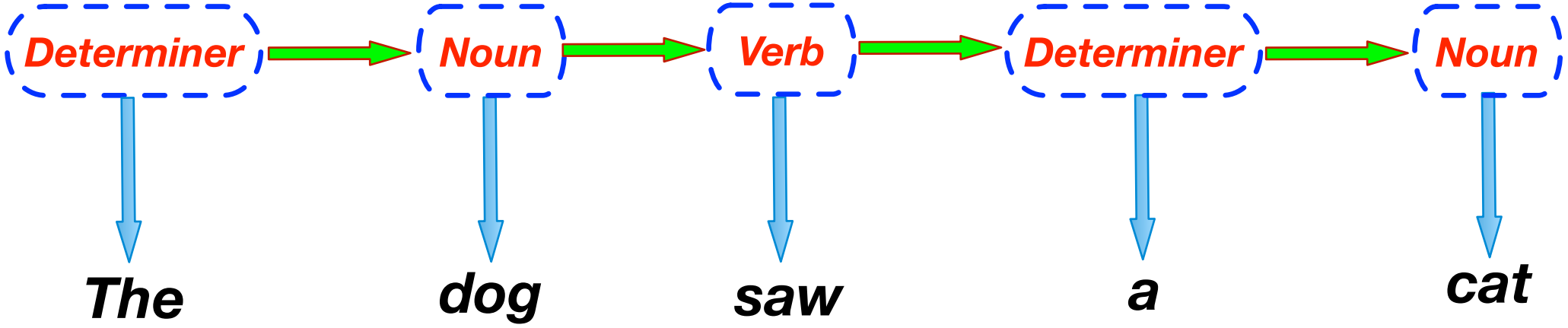
STEP 2: Based on the part of speech select a word ($p(x \mid y)$)

STEP 3: Select the next category ($p(y_{new} \mid all\ old\ y's)$)

If category is not STOP

go to STEP 2

else terminate



Hidden Markov Models (HMM)

This can be modeled via a generative process

STEP 1: Select a part of speech category ($p(y)$)

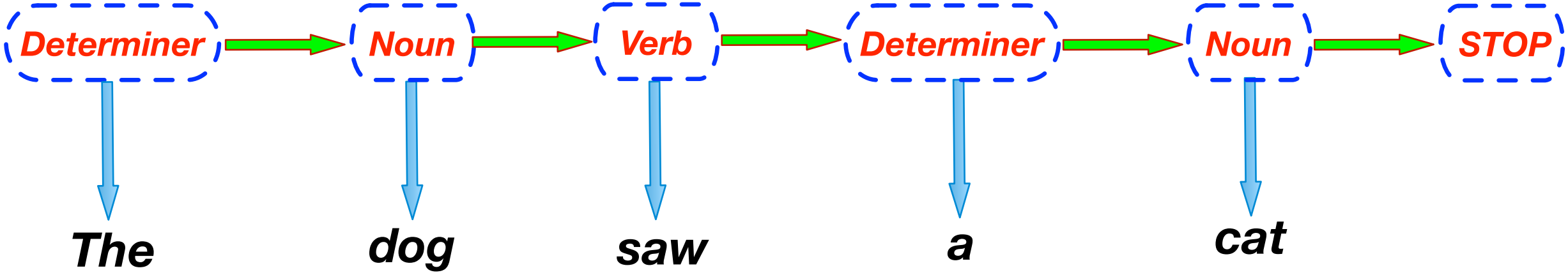
STEP 2: Based on the part of speech select a word ($p(x \mid y)$)

STEP 3: Select the next category ($p(y_{new} \mid all\ old\ y's)$)

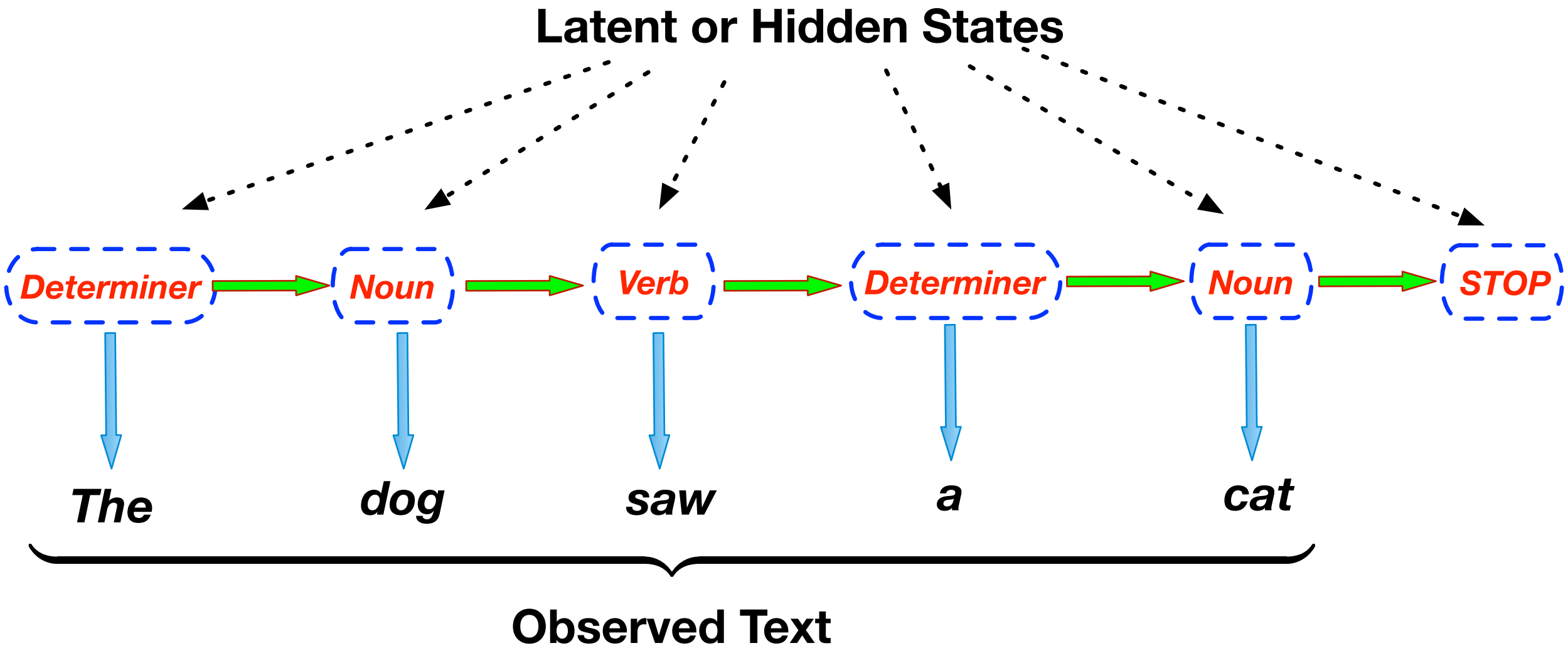
If category is not STOP

go to STEP 2

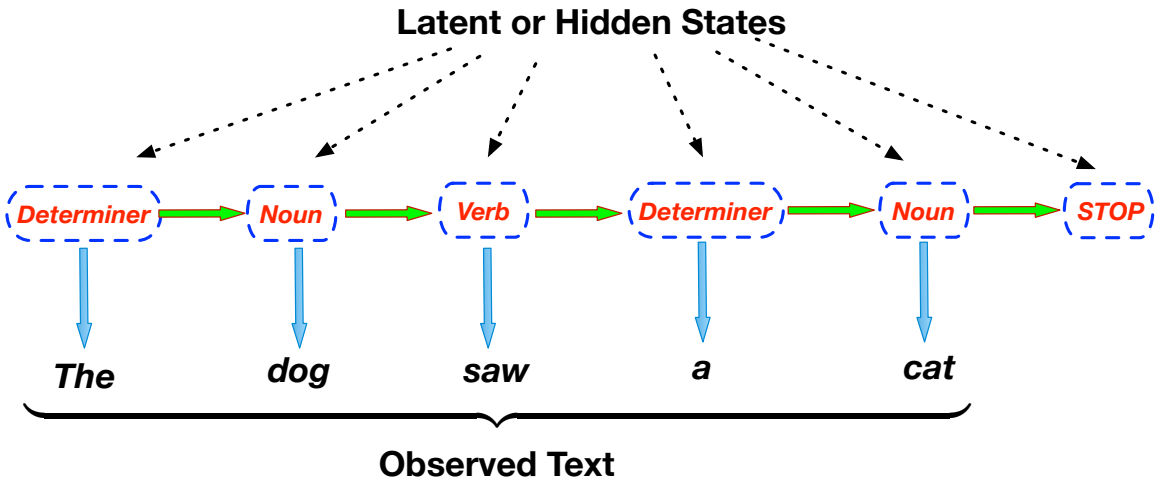
else terminate



Hidden Markov Models (HMM)

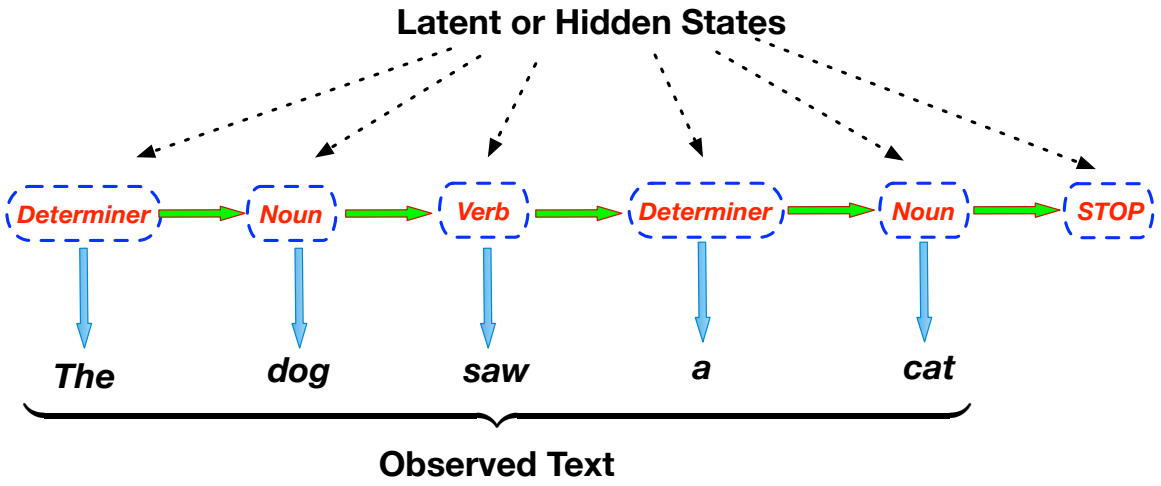


Hidden Markov Models (HMM)



Three Questions:

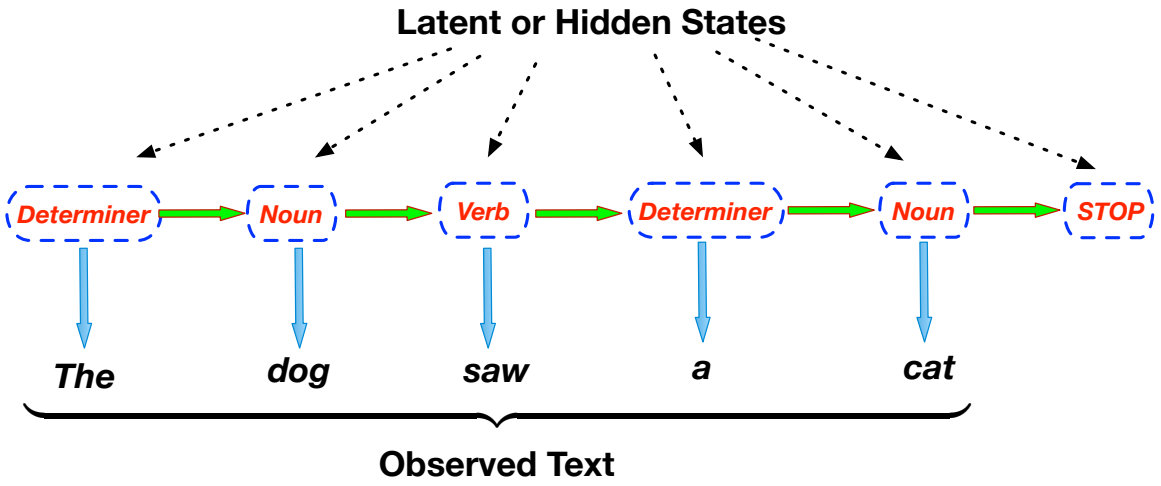
Hidden Markov Models (HMM)



Three Questions:

1. ***Learning Problem***: How do we estimate the parameters of distributions?

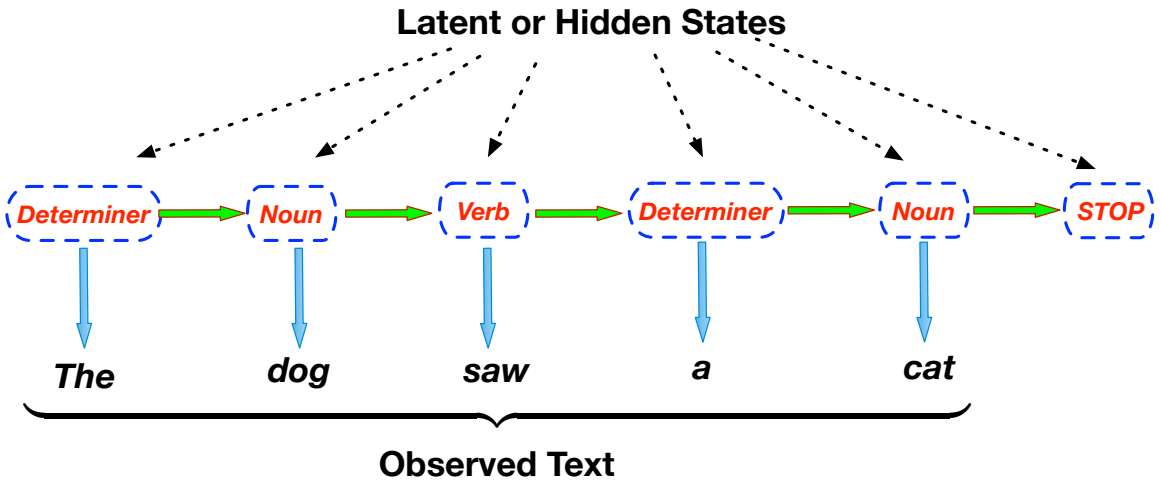
Hidden Markov Models (HMM)



Three Questions:

1. ***Learning Problem***: How do we estimate the parameters of distributions?
2. ***Decoding Problem***: Given the observed text what is the hidden POS sequence that best explains the observation?

Hidden Markov Models (HMM)



Three Questions:

1. ***Learning Problem***: How do we estimate the parameters of distributions?
2. ***Decoding Problem***: Given the observed text what is the hidden POS sequence that best explains the observation?
3. ***Likelihood Problem***: Given the parameters of the distributions what is the probability of the observed sequence?

HMM Setting

- ***Supervised Learning***: Given set of paired observation and state sequences

$$\mathcal{D}_L = \{(x^1, y^1), ..., (x^M, y^M)\}$$
$$x^i = X_1^i \; X_2^i .... \; X_N^i$$
$$y^i = Y_1^i \; Y_2^i .... \; Y_N^i$$

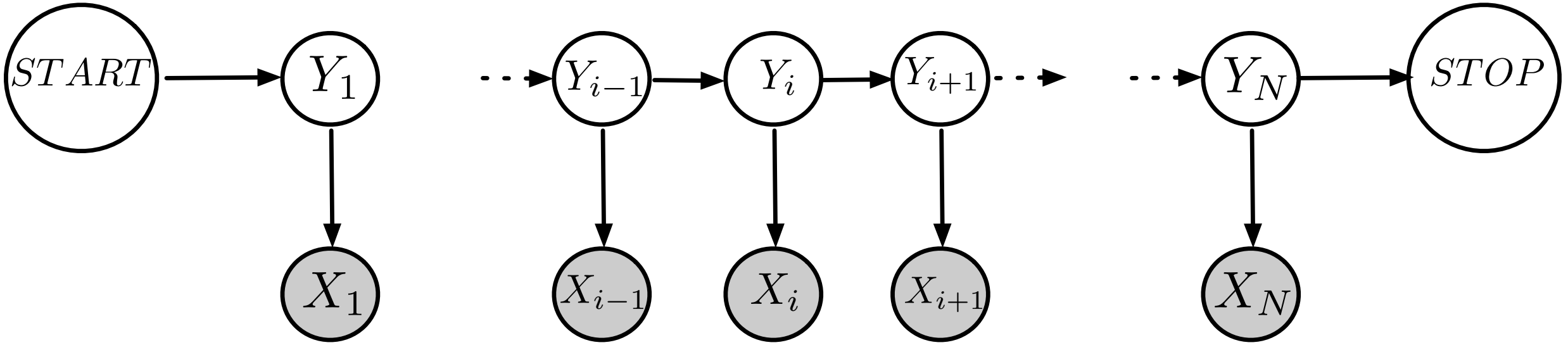
HMM Setting

- ***Supervised Learning***: Given set of paired observation and state sequences

$$\mathcal{D}_L = \{(x^1, y^1), ..., (x^M, y^M)\}$$
$$x^i = X_1^i \; X_2^i .... \; X_N^i$$
$$y^i = Y_1^i \; Y_2^i .... \; Y_N^i$$

$$X_k^i \; is \; a \; random \; variable$$
$$Y_k^i \; is \; a \; random \; variable$$

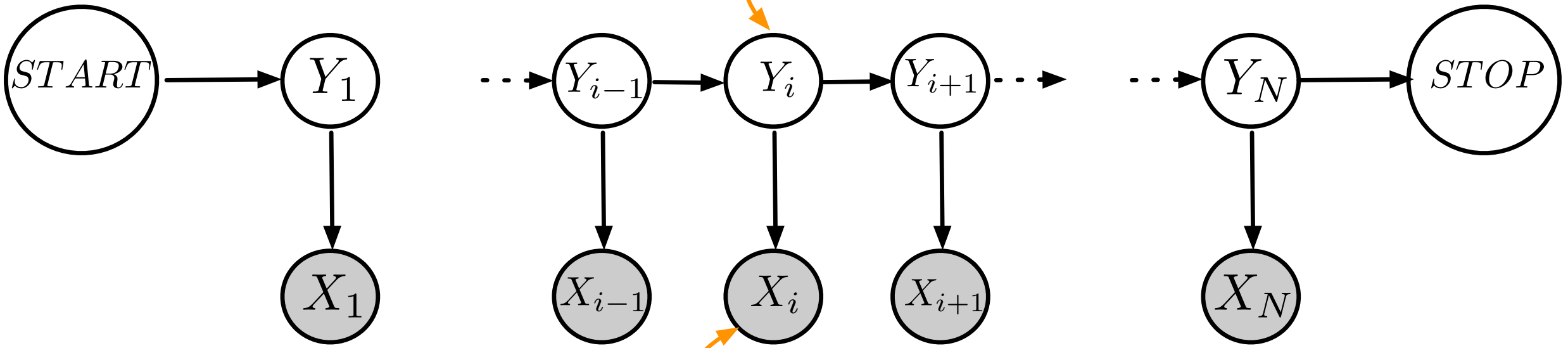
HMM Setting



HMM Setting

Set of States (Λ)

$\{c_0 = START, c_1, c_2, ..., c_K, c_{K+1} = STOP\}$



$\{w_1, w_2, ...w_J\}$

Set of Observations

HMM Setting

Set of States (Λ)

$$\{c_0 = START, c_1, c_2, ..., c_K, c_{K+1} = STOP\}$$

Set of Observations

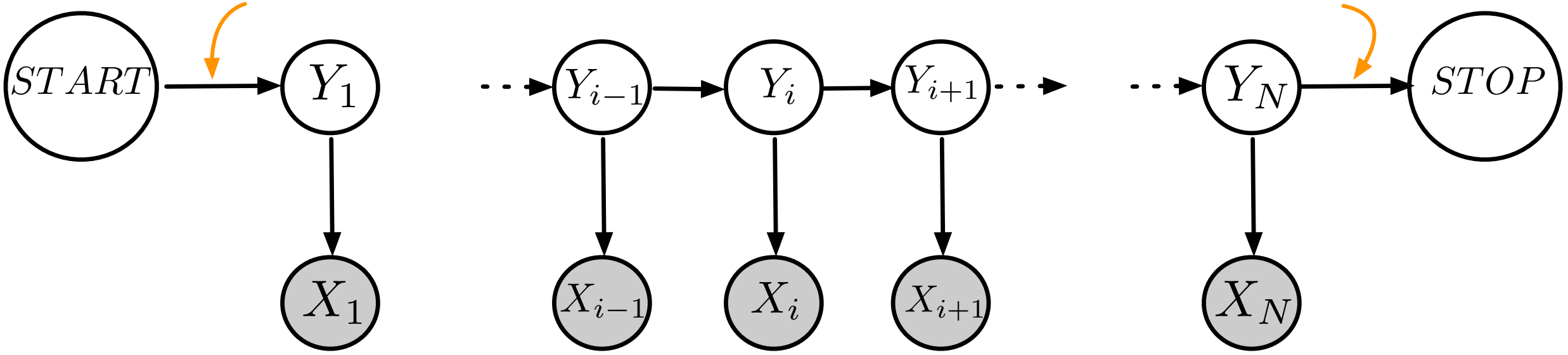
$$\{w_1, w_2, ...w_J\}$$

Start Probability

$P(Y_1 \mid Y_0 = START)$

End Probability

$P(Y_{N+1} = STOP \mid Y_N)$



HMM Setting

Set of States (Λ)

$$\{c_0 = START, c_1, c_2, ..., c_K, c_{K+1} = STOP\}$$

Set of Observations

$$\{w_1, w_2, ...w_J\}$$

Start Probability

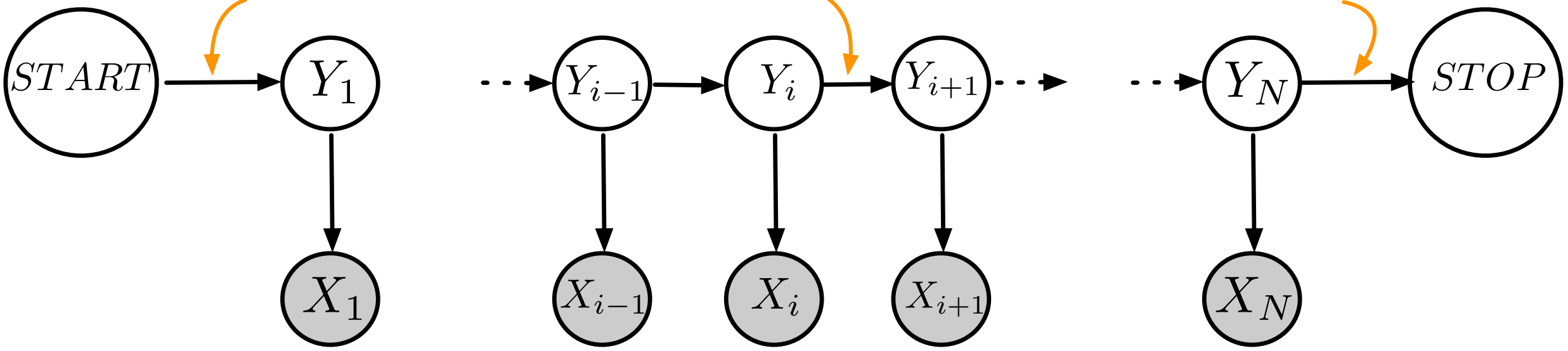
$$P(Y_1 \mid Y_0 = START)$$

Transition Probability

$$P(Y_{i+1} \mid Y_i)$$

End Probability

$$P(Y_{N+1} = STOP \mid Y_N)$$



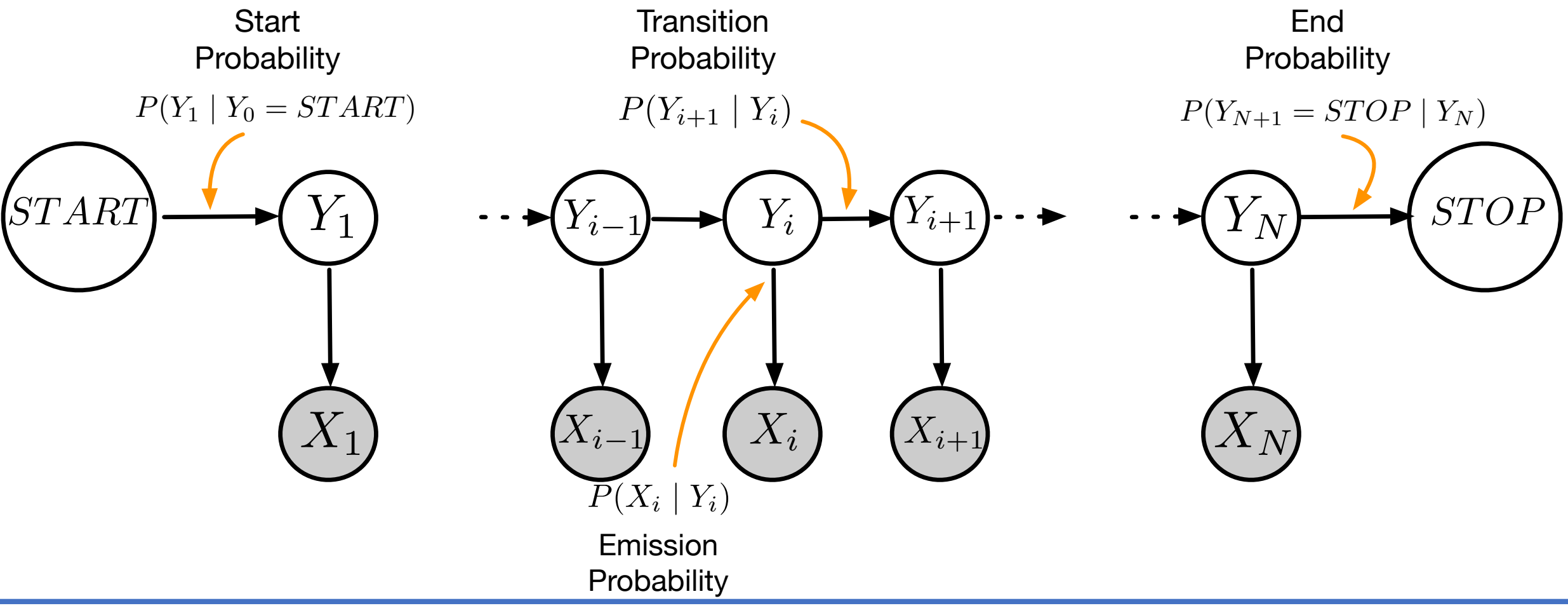
HMM Setting

Set of States (Λ)

$$\{c_0 = START, c_1, c_2, ..., c_K, c_{K+1} = STOP\}$$

Set of Observations

$$\{w_1, w_2, ...w_J\}$$



HMM Assumptions

HMM Assumptions

- ***Markov Property***

$$P(Y_i = y_i \mid Y_{i-1} = y_{i-1}, Y_{i-1} = y_{i-2}, ...Y_1 = y_1) = P(Y_i = y_i \mid Y_{i-1} = y_{i-1})$$

HMM Assumptions

- ***Markov Property***

$y_i, y_{i-1}, \ldots, y_1 \in \{c_1, \ldots, c_K\}$

$$P(Y_i = y_i \mid Y_{i-1} = y_{i-1}, Y_{i-1} = y_{i-2}, \ldots . Y_1 = y_1) = P(Y_i = y_i \mid Y_{i-1} = y_{i-1})$$

HMM Assumptions

- ***Markov Property***

$y_i, y_{i-1}, \ldots, y_1 \in \{c_1, \ldots, c_K\}$

$$P(Y_i = y_i \mid Y_{i-1} = y_{i-1}, Y_{i-1} = y_{i-2}, ...Y_1 = y_1) = P(Y_i = y_i \mid Y_{i-1} = y_{i-1})$$

- ***Observation Independence***

$$P(X_i = x_i \mid Y_1 = y_1 ... Y_i = y_i ... Y_N = y_N, X_1 = x_1 ... X_{i-1} = x_{i-1}) = P(X_i = x_i \mid Y_i = y_i)$$

HMM Assumptions

- ***Markov Property***

$y_i, y_{i-1}, \ldots, y_1 \in \{c_1, \ldots, c_K\}$

$$P(Y_i = y_i \mid Y_{i-1} = y_{i-1}, Y_{i-1} = y_{i-2}, \ldots . Y_1 = y_1) = P(Y_i = y_i \mid Y_{i-1} = y_{i-1})$$

- ***Observation Independence***

$$P(X_i = x_i \mid Y_1 = y_1 \ldots Y_i = y_i \ldots Y_N = y_N, X_1 = x_1 \ldots X_{i-1} = x_{i-1}) = P(X_i = x_i \mid Y_i = y_i)$$

- ***Homogeneity***

$$P(Y_i = c_k | Y_{i-1} = c_l) = P(Y_t = c_k | Y_{t-1} = c_l)$$

$$P(X_i = w_j | Y_i = c_k) = P(X_t = w_j | Y_t = c_k)$$

HMM Assumptions

- ***Markov Property***

$y_i, y_{i-1}, \ldots, y_1 \in \{c_1, \ldots, c_K\}$

$$P(Y_i = y_i \mid Y_{i-1} = y_{i-1}, Y_{i-1} = y_{i-2}, \ldots . Y_1 = y_1) = P(Y_i = y_i \mid Y_{i-1} = y_{i-1})$$

- ***Observation Independence***

$$P(X_i = x_i \mid Y_1 = y_1 \ldots Y_i = y_i \ldots Y_N = y_N, X_1 = x_1 \ldots X_{i-1} = x_{i-1}) = P(X_i = x_i \mid Y_i = y_i)$$

- ***Homogeneity***

random variable Y_i has a fixed value c_k

$$P(Y_i = c_k | Y_{i-1} = c_l) = P(Y_t = c_k | Y_{t-1} = c_l)$$

$$P(X_i = w_j | Y_i = c_k) = P(X_t = w_j | Y_t = c_k)$$

HMM Joint Distribution

- We have a generative model and are interested in the joint distribution

$$P(X, Y) = P\left(X_1 = x_1, \ldots, X_N = x_N, Y_1 = y_1, \ldots, Y_N = y_N\right)$$

HMM Joint Distribution

- We have a generative model and are interested in the joint distribution

$$P(X,Y) = P\left(X_1 = x_1, \ldots, X_N = x_N, Y_1 = y_1, \ldots, Y_N = y_N\right)$$

$$P\left(X_1 = x_1, \ldots, X_N = x_N, Y_1 = y_1, \ldots, Y_N = y_N\right) =$$
$$P\left(y_1 | START\right) \times P\left(x_1 | y_1\right) \times P\left(y_2 | y_1\right) \times \ldots\ldots$$

HMM Joint Distribution

- We have a generative model and are interested in the joint distribution

$$P(X, Y) = P\left(X_1 = x_1, \ldots, X_N = x_N, Y_1 = y_1, \ldots, Y_N = y_N\right)$$

$$P\left(X_1 = x_1, \ldots, X_N = x_N, Y_1 = y_1, \ldots, Y_N = y_N\right) =$$

$$P\left(y_1 | START\right) \times \left(\prod_{i=1}^{N-1} P\left(y_{i+1} | y_i\right)\right) \times \left(\prod_{i=1}^{N} P\left(x_i | y_i\right)\right) \times P\left(STOP | y_N\right)$$

HMM Joint Distribution

- We have a generative model and are interested in the joint distribution

$$P(X, Y) = P\left(X_1 = x_1, \ldots, X_N = x_N, Y_1 = y_1, \ldots, Y_N = y_N\right)$$

$$P\left(X_1 = x_1, \ldots, X_N = x_N, Y_1 = y_1, \ldots, Y_N = y_N\right) =$$

$$P\left(y_1 | START\right) \times \left(\prod_{i=1}^{N-1} P\left(y_{i+1} | y_i\right)\right) \times \left(\prod_{i=1}^{N} P\left(x_i | y_i\right)\right) \times P\left(STOP | y_N\right)$$

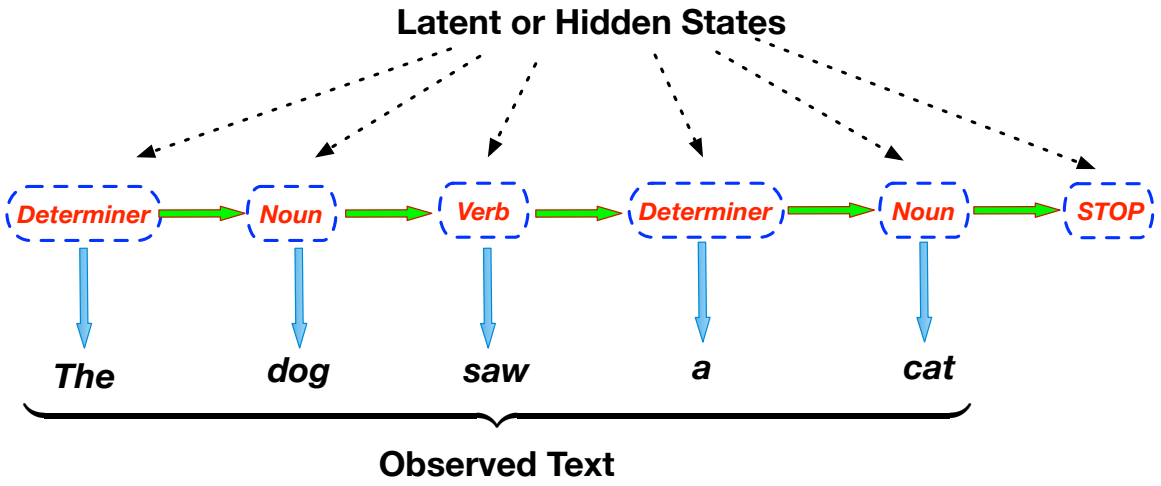
Initial / Start Probability

Transition Probability

Emission Probability

End / Final Probability

Hidden Markov Models (HMM)



Three Questions:

1. ***Learning Problem***: How do we estimate the parameters of distributions?
2. ***Decoding Problem***: Given the observed text what is the hidden POS sequence that best explains the observation?
3. ***Likelihood Problem***: Given the parameters of the distributions what is the probability of the observed sequence?

Learning Problem

- We would like to estimate probability distributions

Learning Problem

- We would like to estimate probability distributions
- Given training data, this can be estimated using Maximum Likelihood estimate i.e. frequency counts

Learning Problem

$$P_{\text{init}}\ (y_1 = c_k | START) = \frac{Count(y_1 = c_k)}{\left(\sum_{l=1}^{K} Count(y_1 = c_l)\right) = M}$$

Learning Problem

$$P_{\text{init}}\ (y_1 = c_k | START) = \frac{Count(y_1 = c_k)}{\left(\sum_{l=1}^{K} Count(y_1 = c_l)\right) = M}$$

$$P_{\text{trans}}\ (y_{i+1} = c_k | y_i = c_l) = \frac{Count(c_k, c_l)}{Count(c_l)}$$

Learning Problem

$$P_{\text{init}}\ (y_1 = c_k | START) = \frac{Count(y_1 = c_k)}{\left(\sum_{l=1}^{K} Count(y_1 = c_l)\right) = M}$$

$$P_{\text{trans}}\ (y_{i+1} = c_k | y_i = c_l) = \frac{Count(c_k, c_l)}{Count(c_l)}$$

$$P_{\text{emiss}}\ (x_i = w_j | y_i = c_k) = \frac{Count(w_j, c_k)}{Count(c_k)}$$

Learning Problem

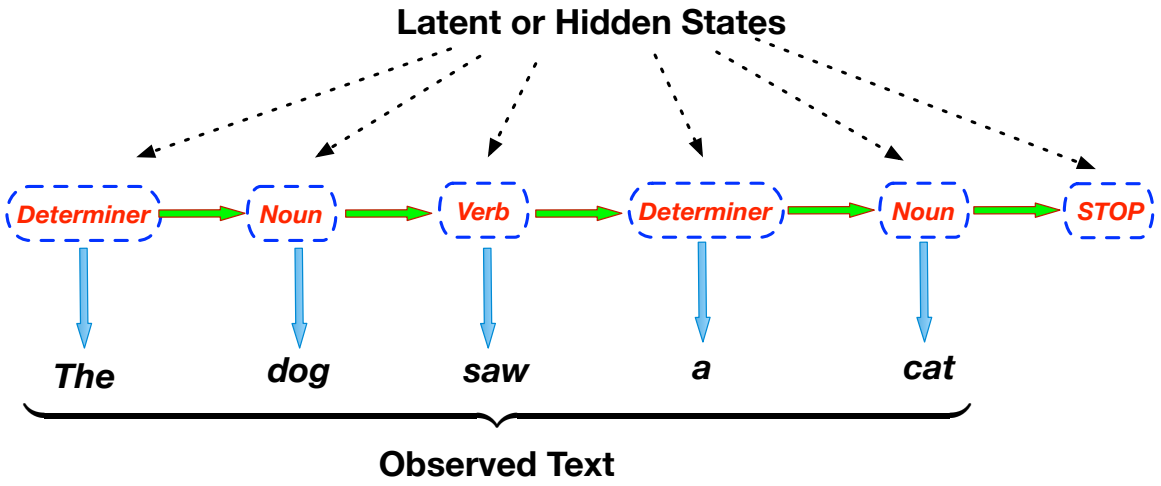
$$P_{\text{init}}\ (y_1 = c_k | START) = \frac{Count(y_1 = c_k)}{\left(\sum_{l=1}^{K} Count(y_1 = c_l)\right) = M}$$

$$P_{\text{trans}}\ (y_{i+1} = c_k | y_i = c_l) = \frac{Count(c_k, c_l)}{Count(c_l)}$$

$$P_{\text{emiss}}\ (x_i = w_j | y_i = c_k) = \frac{Count(w_j, c_k)}{Count(c_k)}$$

$$P_{\text{final}}\ (STOP | y_N = c_k) = \frac{Count(y_N = c_k)}{\left(\sum_{l=1}^{K} Count(y_N = c_l)\right) = M}$$

Hidden Markov Models (HMM)



Three Questions:

1. ***Learning Problem***: How do we estimate the parameters of distributions?
2. ***Decoding Problem***: Given the observed text what is the hidden POS sequence that best explains the observation?
3. ***Likelihood Problem***: Given the parameters of the distributions what is the probability of the observed sequence?

Decoding Problem

- Two ways possible

Decoding Problem

- Two ways possible
  - **Posterior Decoding**: Maximize probability of each state

$$y_i^* = \arg\max_{y_i \in \Lambda} P\left(Y_i = y_i | X_1 = x_1, \ldots, X_N = x_N\right)$$

Decoding Problem

- Two ways possible
  - **Posterior Decoding**: Maximize probability of each state

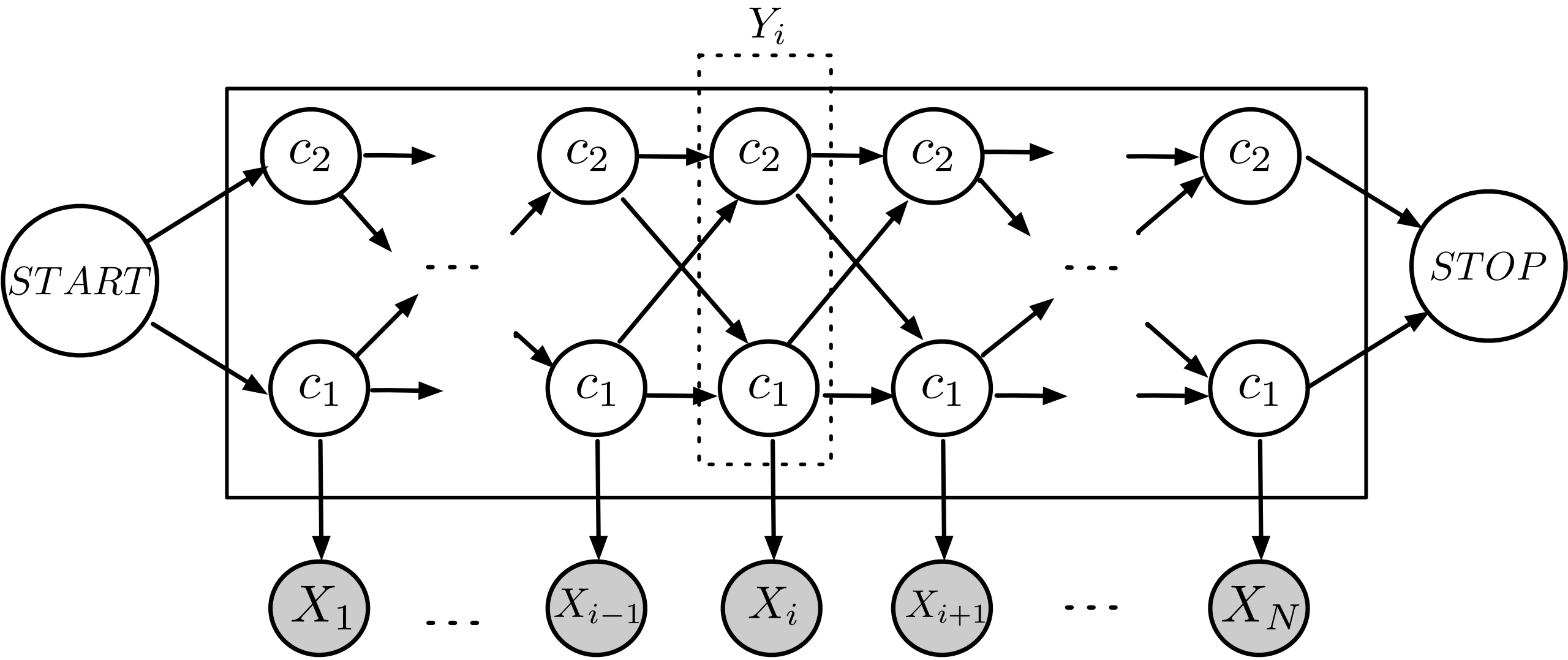
$$y_i^* = \arg\max_{y_i \in \Lambda} P\left(Y_i = y_i | X_1 = x_1, \ldots, X_N = x_N\right)$$

  - **Viterbi Decoding**: Maximize the probability of the entire sequence

$$y^* = \arg\max_{y = y_1 \ldots y_N} P\left(Y_1 = y_1, \ldots, Y_N = y_N | X_1 = x_1, \ldots, X_N = x_N\right)$$

$$= \arg\max_{y = y_1 \ldots y_N} P\left(Y_1 = y_1, \ldots, Y_N = y_N, X_1 = x_1, \ldots, X_N = x_N\right)$$

Trellis Diagram



Decoding Problem

- Decoding requires the following calculation

$$P(Y \mid X) = \frac{P(X, Y)}{P(X)}$$

$$P(Y \mid X) = \frac{P(X, Y)}{\sum_{Y} P(X, Y)}$$

Decoding Problem

- Decoding requires the following calculation

$$P(Y \mid X) = \frac{P(X, Y)}{P(X)}$$

$$P(Y \mid X) = \frac{P(X, Y)}{\sum_{Y} P(X, Y)}$$

But what is the sum in the denominator over?

Decoding Problem

- Decoding requires the following calculation

$$P(Y \mid X) = \frac{P(X, Y)}{P(X)}$$

$$P(Y \mid X) = \frac{P(X, Y)}{\sum_{Y} P(X, Y)}$$

But what is the sum in the denominator over?

Over all possible sequences $Y = y_1, \ldots, y_N$

Decoding Problem

We want sum over all possible sequences $Y = y_1, \ldots, y_N$

Decoding Problem

We want sum over all possible sequences $Y = y_1, \ldots, y_N$

But how many sequences are there?

Decoding Problem

We want sum over all possible sequences $Y = y_1, \ldots, y_N$

But how many sequences are there?

Suppose, number of states $= 2$ i.e. $K = 2$ in $\{c_1, \ldots, c_K\}$

i.e. $\forall i, \quad y_i$ can take 2 possible values: $\{c_1, c_2\}$

Suppose, length of sequence is $N = 4$

How many sequences $\{y_1, y_2, y_3, y_4\}$ are there?

Decoding Problem

How many sequences $\{y_1, y_2, y_3, y_4\}$ are there?

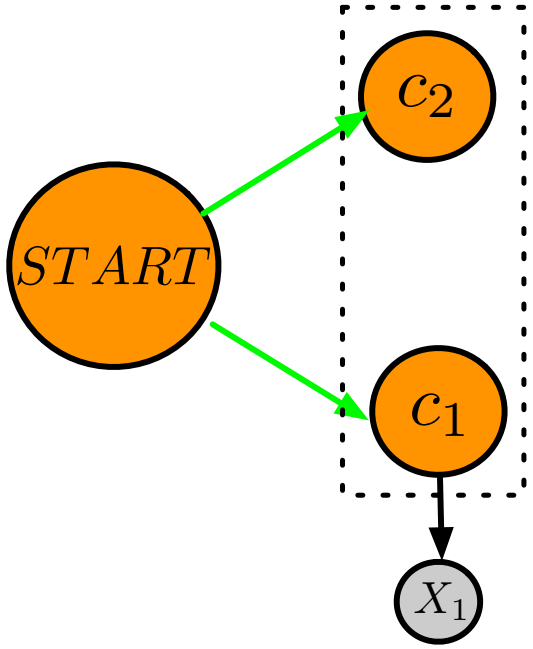
Let $count = 1$

Decoding Problem

How many sequences $\{y_1, y_2, y_3, y_4\}$ are there?

Let $count = 1$

y_1 can take 2 possible values $\implies$ count $= 2$



$\{(c_1), (c_2)\}$

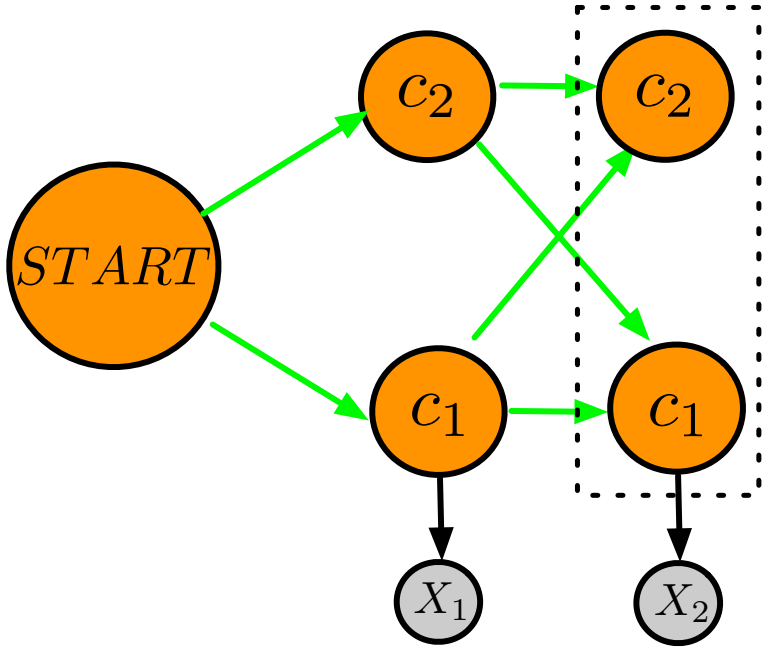
Decoding Problem

How many sequences $\{y_1, y_2, y_3, y_4\}$ are there?

Let $count = 1$

y_1 can take 2 possible values $\implies$ count $= 2$

y_2 can take 2 possible values $\implies$ count $= 2 \times 2$



$$\{(c_1, c_1), (c_1, c_2), (c_2, c_1), (c_2, c_2)\}$$

Decoding Problem

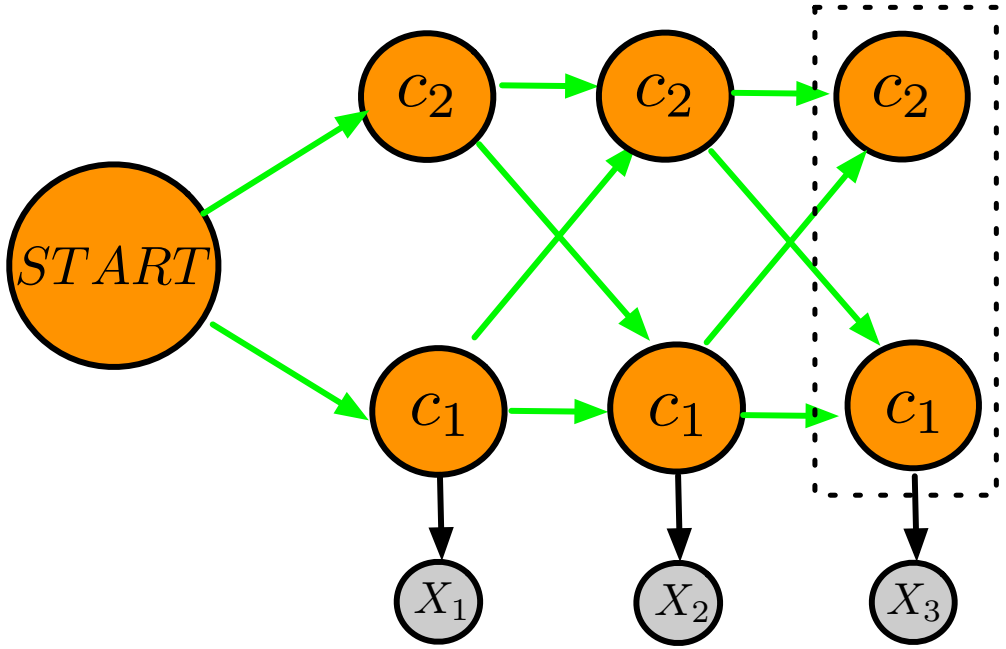
How many sequences $\{y_1, y_2, y_3, y_4\}$ are there?

Let $count = 1$

y_1 can take 2 possible values $\implies$ count $= 2$

y_2 can take 2 possible values $\implies$ count $= 2 \times 2$

y_3 can take 2 possible values $\implies$ count $= 2 \times 2 \times 2$



$\{(c_1, c_1, c_1), (c_1, c_1, c_2),$
$(c_1, c_2, c_1), (c_1, c_2, c_2),$
$(c_2, c_1, c_1), (c_2, c_1, c_2),$
$(c_2, c_2, c_1), (c_2, c_2, c_2)\}$

Decoding Problem

How many sequences $\{y_1, y_2, y_3, y_4\}$ are there?

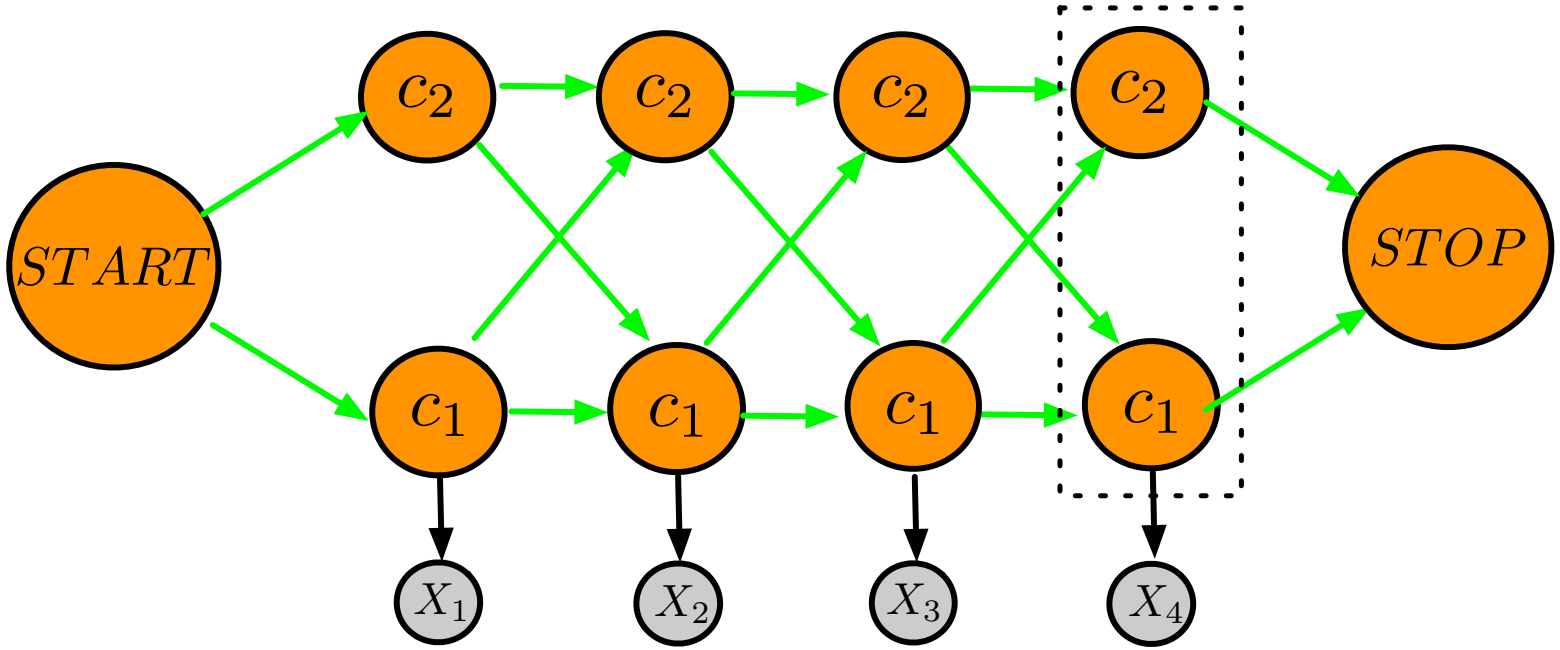
Let $count = 1$

y_1 can take 2 possible values $\implies$ count $= 2$

y_2 can take 2 possible values $\implies$ count $= 2 \times 2$

y_3 can take 2 possible values $\implies$ count $= 2 \times 2 \times 2$

y_4 can take 2 possible values $\implies$ count $= 2 \times 2 \times 2 \times 2$



Decoding Problem

How many sequences $\{y_1, y_2, y_3, y_4\}$ are there?

Let $count = 1$

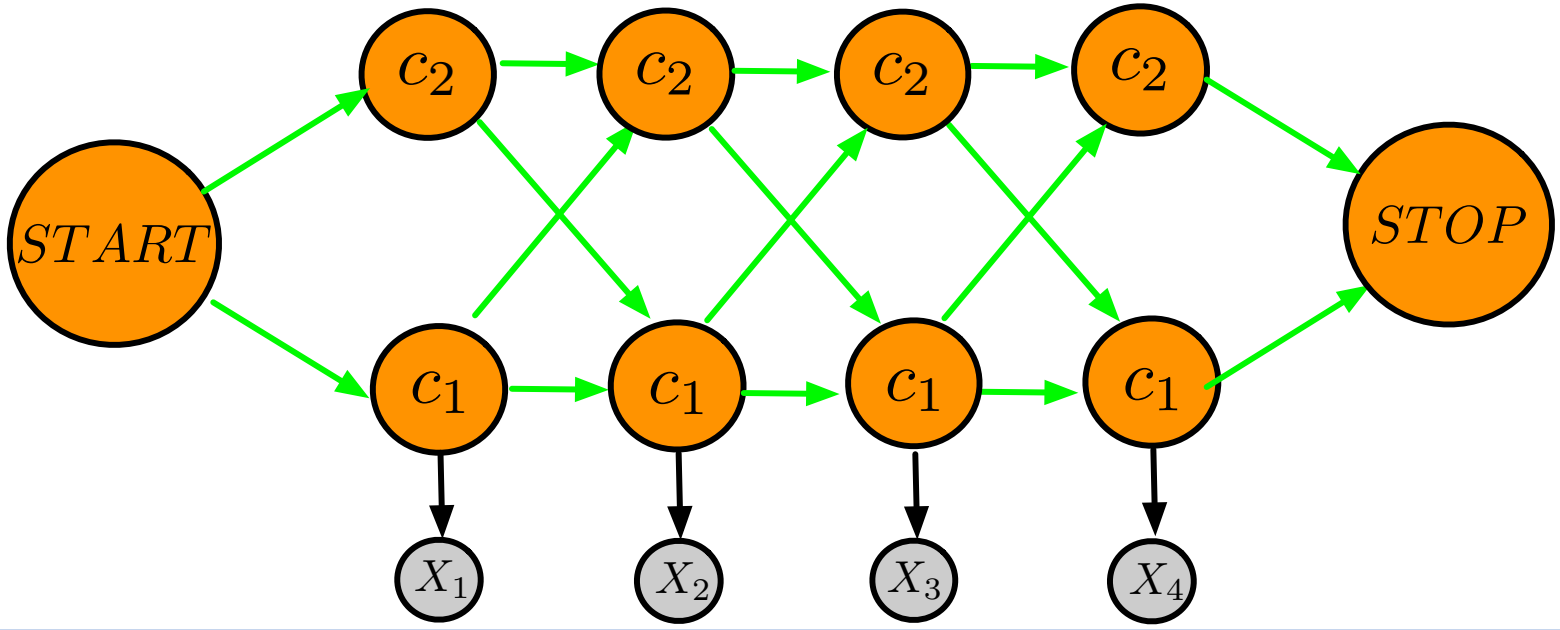
y_1 can take 2 possible values $\implies$ count $= 2$

y_2 can take 2 possible values $\implies$ count $= 2 \times 2$

y_3 can take 2 possible values $\implies$ count $= 2 \times 2 \times 2$

y_4 can take 2 possible values $\implies$ count $= 2 \times 2 \times 2 \times 2$

16 sequences



Decoding Problem

Suppose, $K = 10$ in $\{c_1, \ldots, c_K\}$ and $N = 10$

How many sequences $\{y_1, \ldots, y_N\}$ are there?

Decoding Problem

Suppose, $K = 10$ in $\{c_1, \ldots, c_K\}$ and $N = 10$

How many sequences $\{y_1, \ldots, y_N\}$ are there?

$$10^{10}$$

Decoding Problem

Suppose, $K = 10$ in $\{c_1, \ldots, c_K\}$ and $N = 10$

How many sequences $\{y_1, \ldots, y_N\}$ are there?

$10^{10} = 10,000,000,000$

TEN BILLION!!!!

Image: https://www.123rf.com/clipart-vector/no_nail.html?sti=mdopb61zzsjrv10wes|

Real Life Decoding Problem

Suppose, $K = 45$ (Number of POS Tags)

and suppose, length of sequence (length of a sentence) $N = 10$

How many sequences $\{y_1, \ldots, y_N\}$ are there?

Real Life Decoding Problem

Suppose, $K = 45$ (Number of POS Tags)

and suppose, length of sequence (length of a sentence) $N = 10$

How many sequences $\{y_1, \ldots, y_N\}$ are there?

$$45^{10}$$

Real Life Decoding Problem

$$\text{Number of possible sequences} = 45^{10} = 3.4 \times 10^{16}$$

Image: tenor.com

Decoding Problem

$$\text{In general, number of sequences} = K^N$$

Decoding Problem

$$\text{In general, number of sequences} = K^N$$

EXPONENTIAL IN LENGTH

Decoding Problem

$$\text{In general, number of sequences} = K^N$$

EXPONENTIAL IN LENGTH

What do we do?

Decoding Problem

$$\text{In general, number of sequences} = K^N$$

EXPONENTIAL IN LENGTH

What do we do?

DYNAMIC PROGRAMMING

Summary

- Sequence prediction is an important problem in NLP
- We looked an important class of sequence prediction models: HMM
- For HMM three problems need to be addressed:
  - Learning
  - Decoding
  - Observation Probability
- Learning problem can be solved using MLE
- Decoding requires dynamic programming

References

1. Michael Collin's NLP Lecture Notes: http://www.cs.columbia.edu/~mcollins/hmms-spring2013.pdf

2. Chapter 6, Speech and Language Processing, Dan Jurafsky and James Martin

3. LxMLS Lab Guide: http://lxmls.it.pt/2016/LxMLS2016.pdf

- Next class
  - Decoding Problem