Unit 3

Part of Speech Tagging

HMM, Rule Based PoSTagging, Stochastic PoS Tagging

Natural Language Processing (NLP)
MDS 555

Objective

- Hidden Markup Model
- Type of PoS Tagger
  - Rule Based PoSTagging
  - Stochastic PoS Tagging
  - Transformation based Tagging

Methods of PoS Tagging

- Rule-Based POS tagging
  - e.g., ENGTWOL [Voutilainen, 1995]
  - large collection (> 1000) of constraints on what
  - sequences of tags are allowable

Methods of PoS Tagging

- Transformation-based tagging
  - e.g.,Brill’s tagger [Brill, 1995]

Methods of PoS Tagging

- Stochastic (Probabilistic) tagging
  - e.g., TNT [Brants, 2000]

Rule-Based Tagging

- Knowledge-driven taggers
- Usually rules built manually
- Limited amount of rules

Rule-Based Tagging

- Standard approach (two steps):
  - Dictionaries to assign a list of potential tags
    - Plays (NNS/VBZ)
    - well (UH/JJ/NN/RB)
    - with (IN)
    - others (NNS)
  - Hand-written rules to restrict to a POS tag
    - Plays (VBZ)
    - well (RB)
    - with (IN)
    - others (NNS)

Hidden Markup Model (HMM)

- Sequence labeling algorithm
- An HMM is a probabilistic sequence model:
  - given a sequence of units (words, letters, morphemes, sentences, whatever), it computes a probability distribution over possible sequences of labels and chooses the best label sequence
- It uses markov chain

Stochastic Variable

- A stochastic variable is a random variable that is a variable whose value cannot be predicted definitely but the values of a very large number of observations follow a clear pattern.

Markov Chain

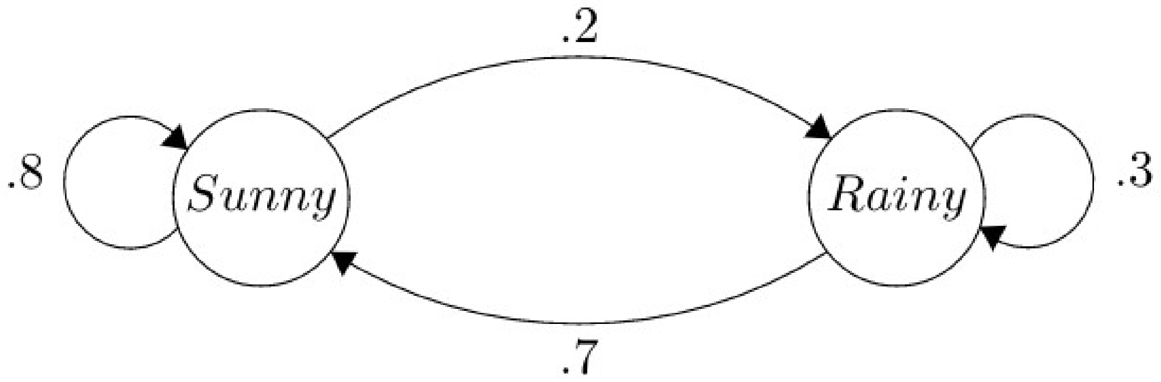
- A Markov chain is a stochastic model that uses mathematics to predict the probability of a sequence of events occurring *based on the most recent event*.
  - It is represented as a state diagram
  - Future step is only depends on the current state
  - The change of the state is based on the highest occurrence of the probability

Markov Chain

- Example of use
  - Next word suggestion:
  - A common example of a Markov chain in action is the way Google predicts the next word in your sentence based on your previous entry within Gmail.
- The main goal of the Markov process is to identify the probability of transitioning from one state to another

Markov Chain

- Markov chains may be modeled by finite state machines



- The main goal of the Markov process is to identify the probability of transitioning from one state to another.
- One of the primary appeals to Markov is that the future state of a stochastic variable is only dependent on its present state.

Markov Process - Memorylessness

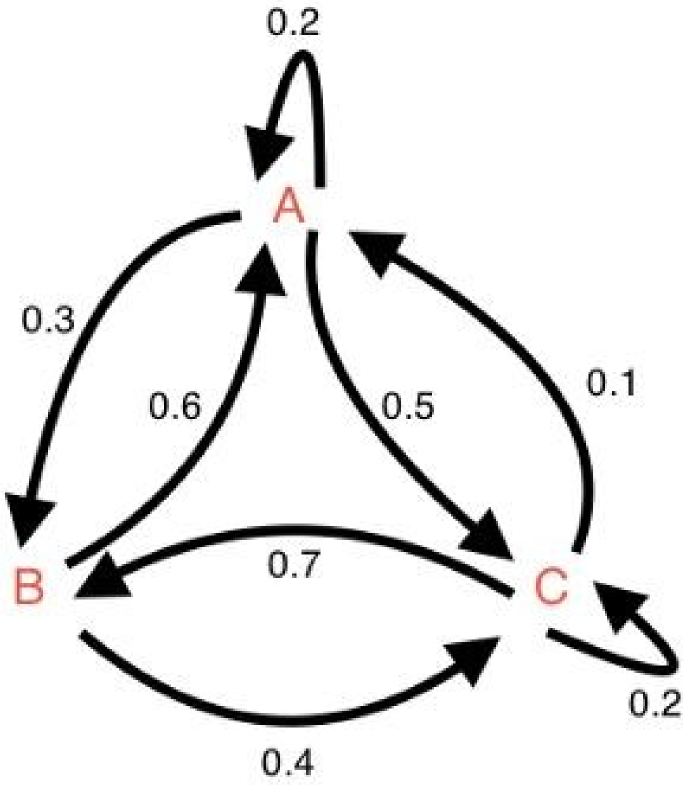
- Markov process is a stochastic process which has memoryless characteristics
- In other words, when a model has a memoryless property,
  - it implies that the model has “forgotten” which state the system is in.
  - Hence, previous states of the process would not influence the probabilities.

How to Create a Markov Chain Model

- A Markov chain model is dependent on two key pieces of information
  - the transition matrix
  - initial state vector
- Transition Matrix
  - Denoted as “P”
  - This NxN matrix represents the probability distribution of the state’s transitions.
  - The sum of probabilities in each row of the matrix will be one, implying that this is a stochastic matrix

Markov Chain Model

	A	B	C
A	.2	.3	.5
B	.6	0	.4
C	.1	.7	.2



Markov Chain Model

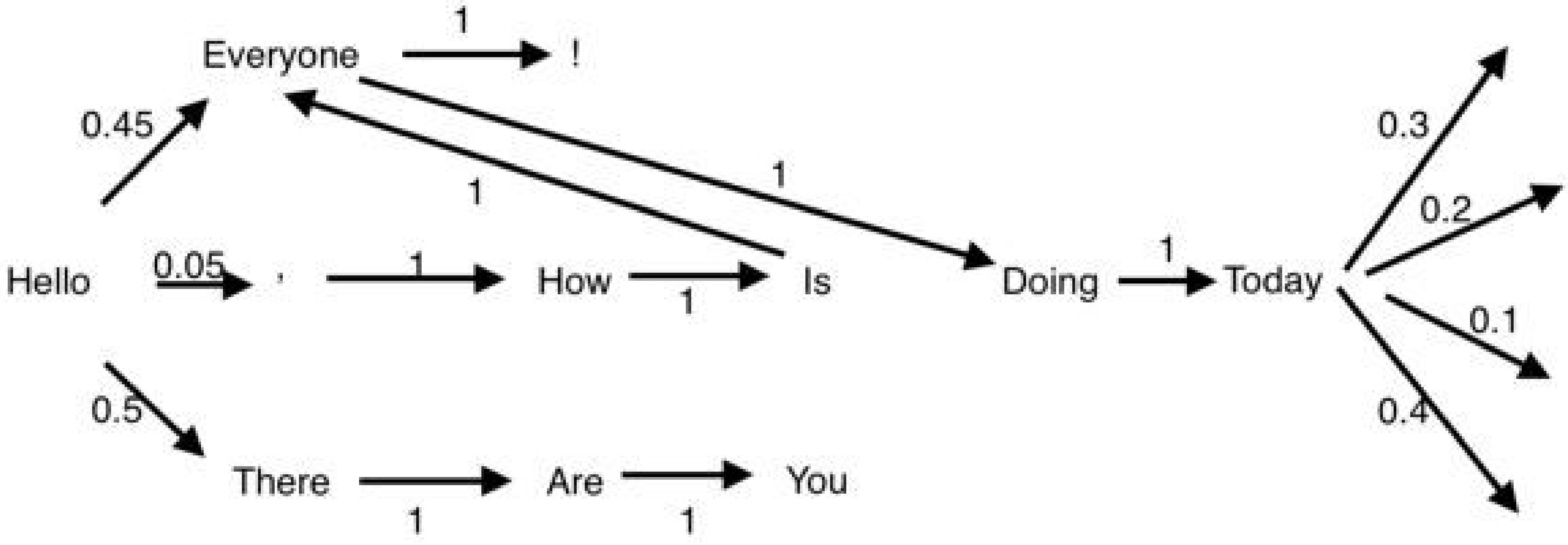
- Initial State Vector
  - Denoted as “S” this Nx1 vector represents the probability distribution of starting at each of the N possible states.
  - Every element in the vector represents the probability of beginning at that state.

Example of Markov Chain

- Text prediction
  - Suppose you had a large amount of text associated with a topic.
  - You can imagine each sentence as a sequence of words in that corpus of text.
  - Each word would then be its own state, and you would associate the probability of moving from one state to another based on the available words to which you are connected.
  - This would allow you to transition from one state to another based on the probabilities associated with the transition matrix.

Example of Markov Chain

- Text prediction



HMM Model

- HMM model consist of these basic parts:
  - hidden states
  - observation symbols (or states)
  - transition from initial state to initial hidden state probability distribution
  - transition to terminal state probability distribution (in most cases excluded from model because all probabilities equal to 1 in general use)
  - state transition probability distribution
  - state emission probability distribution

Hidden states and observation symbols

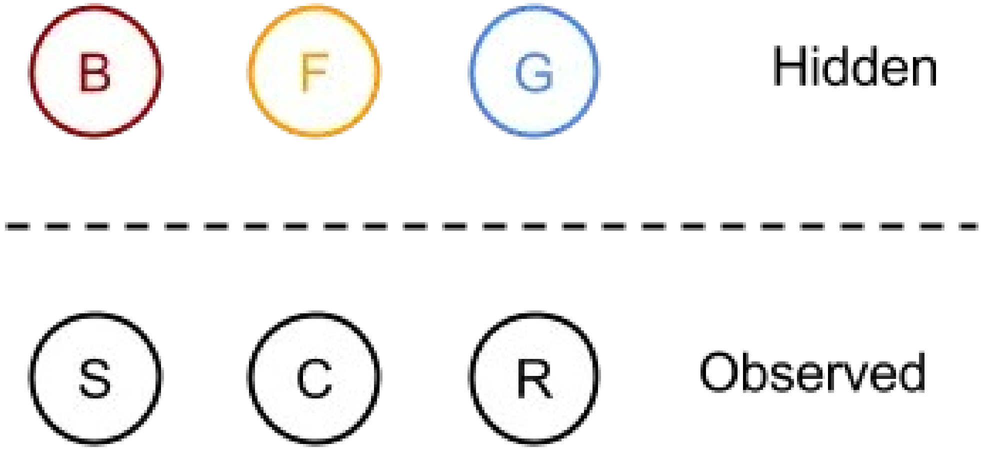
- HMM has two parts:
  - hidden and observed.
- The hidden part consist of hidden states which are not directly observed, their presence is observed by observation symbols that hidden states emits.

HMM: States

- You want to know your friends activity, but you can only observe what weather is outside.
- Your friend activities which are hidden states “emits” observable symbols, which are weather condition.
- You might think that should be other way, that weather conditions is hidden states and your friends activities are observable symbols, but the key is that weather you can observe, but your friends activity you can’t, that makes states a way it is.

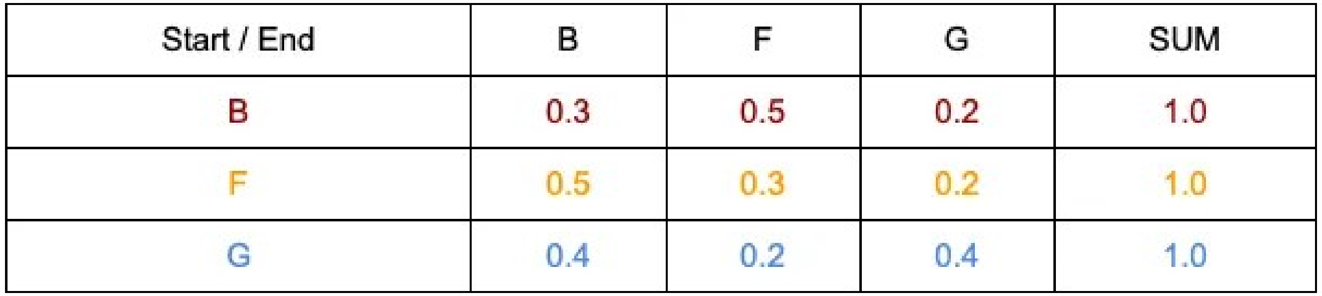
HMM: States

- Your friends activities:
  - Basketball (B)
  - Football (F)
  - Video games (G)
- Observable symbols:
  - Sunny (S)
  - Cloudy ©
  - Rainy (R)

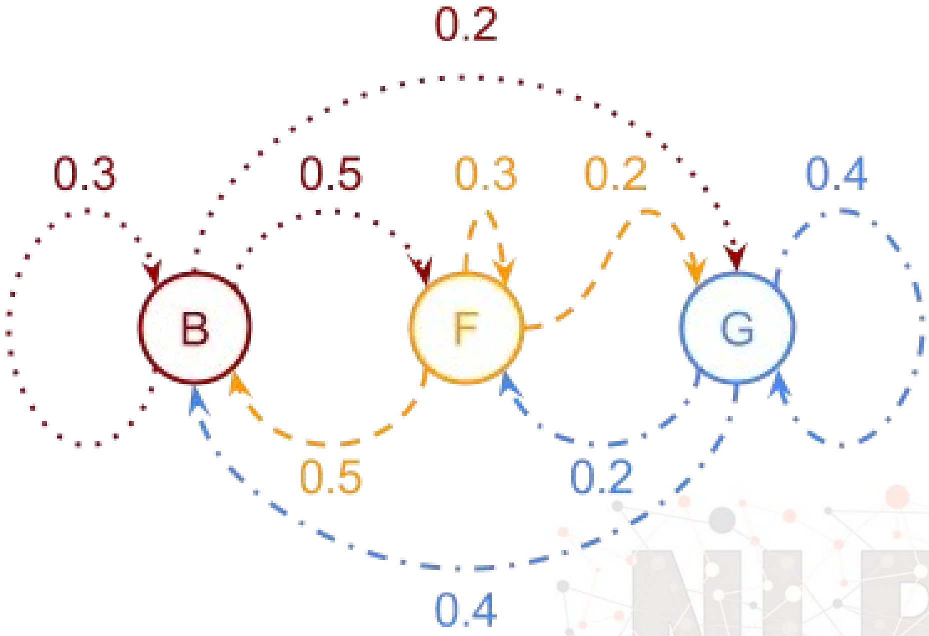


HMM: State transition probability distribution

- state transition probability distribution explains transitions between hidden states

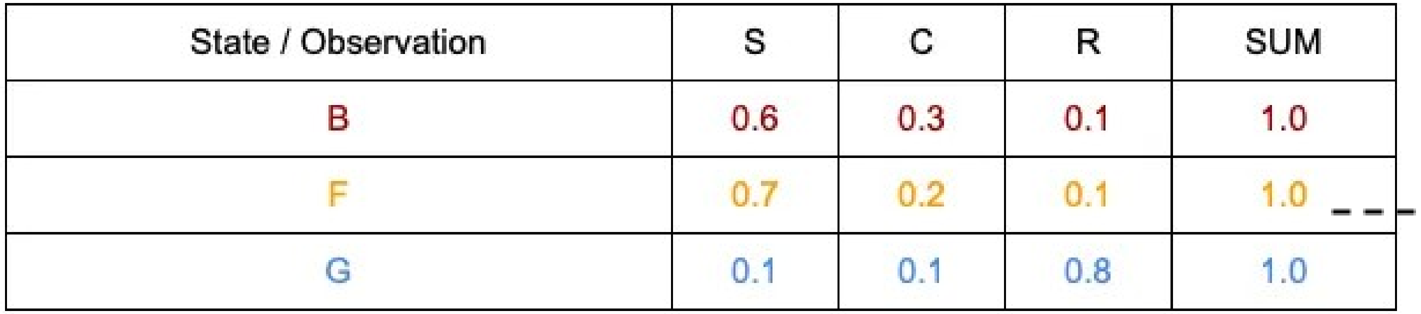


Start / End	B	F	G	SUM
B	0.3	0.5	0.2	1.0
F	0.5	0.3	0.2	1.0
G	0.4	0.2	0.4	1.0

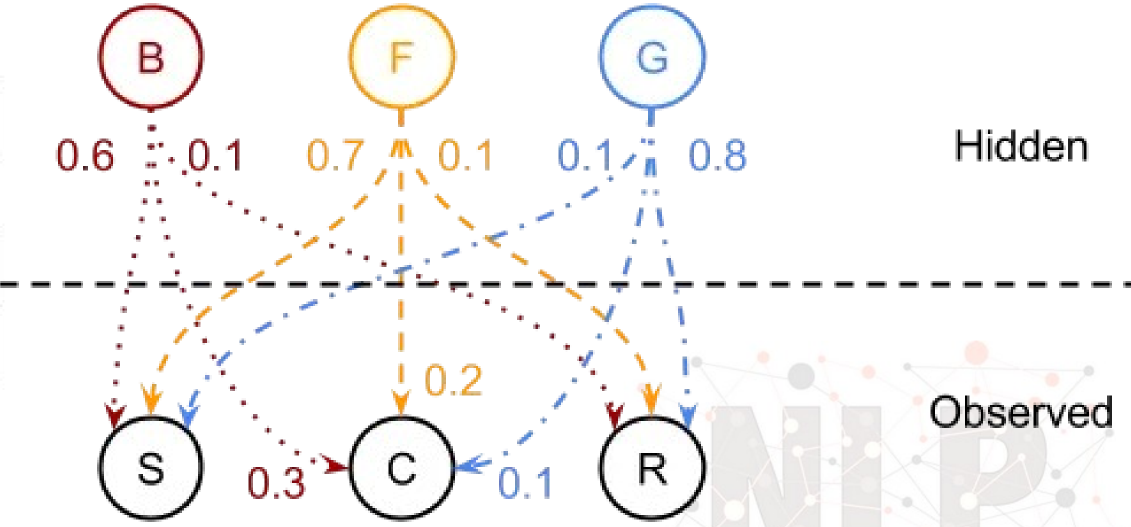


HMM: State emission probability distribution

- The hidden states and observation symbols are bind by state emission probability distribution
- Every transition to hidden state emits observation symbol.



State / Observation	S	C	R	SUM
B	0.6	0.3	0.1	1.0
F	0.7	0.2	0.1	1.0
G	0.1	0.1	0.8	1.0

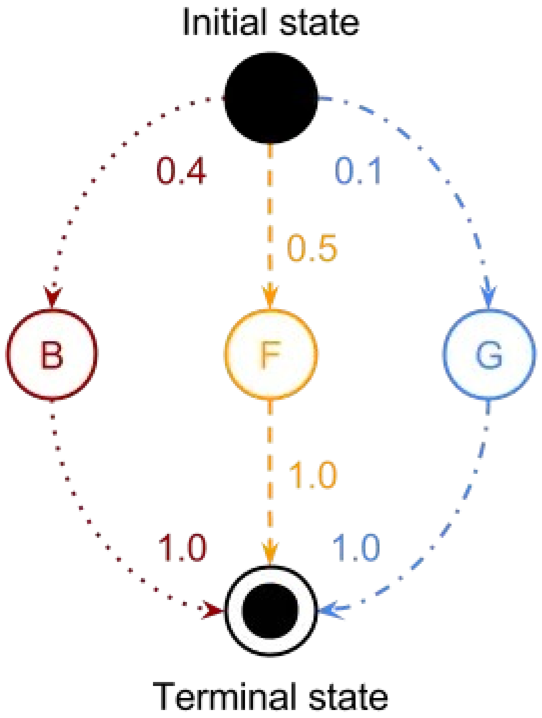


HMM: Initial/terminal state transition probabilities

- When you have hidden states there are two more states that are not directly related to model, but used for calculations
- Initial State:
  - when observation sequence starts initial hidden state which emits symbol is decided from initial state transition probability
- Terminal State:
  - When you reach end of observation sequence you basically transition to terminal state

HMM: initial and terminal state

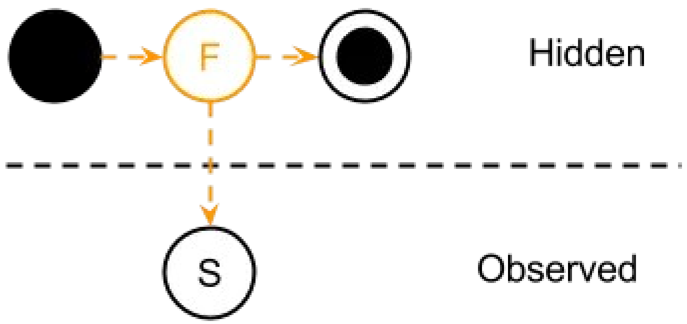
- you can see that when observation sequence starts most probable hidden state which emits first observation sequence symbol is hidden state



HMM: Observation sequence

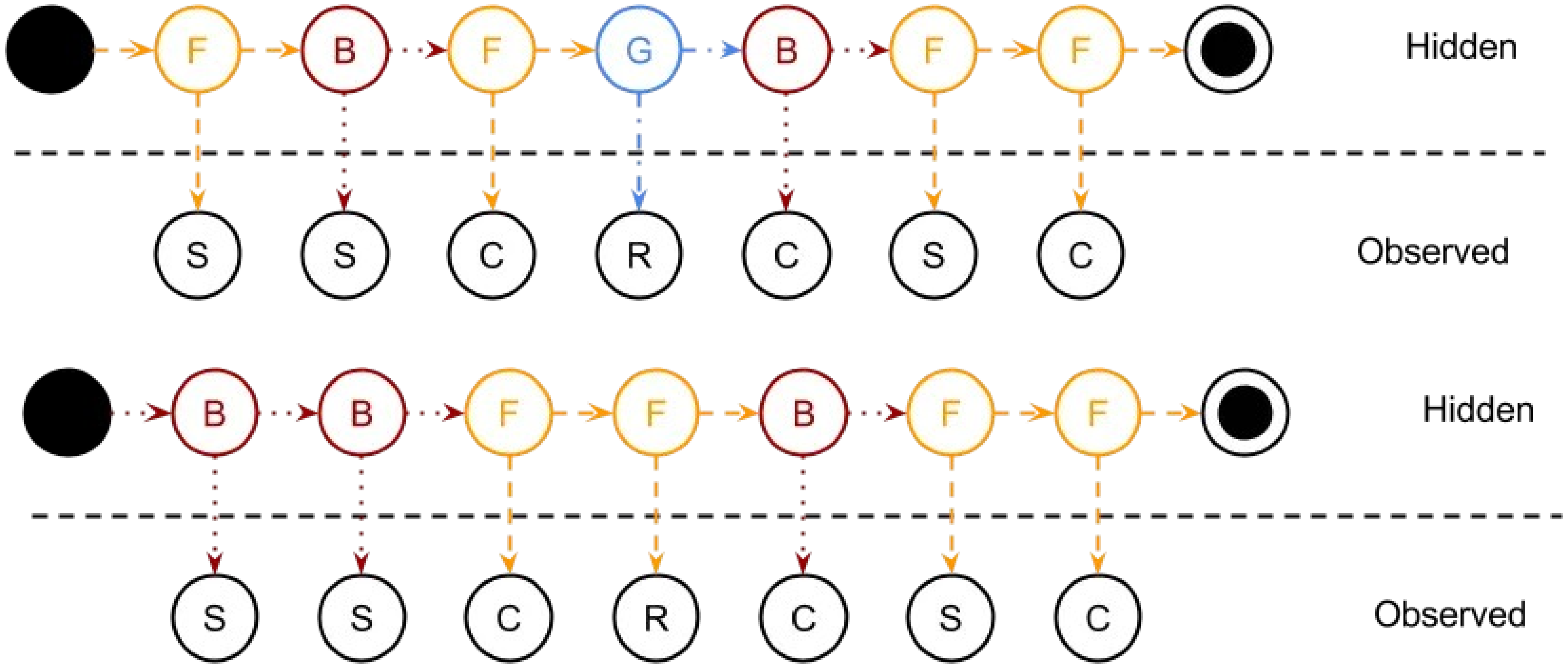
- Observation sequence is sequence of observation symbols from 1 symbol to N symbols.
- Every observation sequence is treated as separate unit without any knowledge about past or future.

Fig:Observation Sequence: S



HMM: Observation sequence

- Observation sequence can be emitted from difference hidden state sequence
- Observation sequence SSCRCSC



HMM: Formal Definition

- Finding the best sequence of tags (t1 ...tn) that corresponds to the sequence of observations (w1 ...wn)
- Probabilistic View
  - Considering all possible sequences of tags
  - Choosing the tag sequence from this universe of sequences, which is most probable given the observation sequence

$$\hat{t}_1^n = argmax_{t_1^n} P\left(t_1^n | w_1^n\right)$$

HMM: Formal Definition

- Using bayes rules

$$\hat{t}_1^n = argmax_{t_1^n} P(t_1^n|w_1^n)$$

$$P(A|B) = \frac{P(B|A)\cdot P(A)}{P(B)}$$

$$P(t_1^n|w_1^n) = \frac{P(w_1^n|t_1^n)\cdot P(t_1^n)}{P(w_1^n)}$$

$$\hat{t}_1^n = argmax_{t_1^n} \underbrace{P(w_1^n|t_1^n)}_{\text{likelihood}}\cdot \underbrace{P(t_1^n)}_{\text{prior probability}}$$

HMM: Formal Definition

- Markov Assumption

$$\hat{t}_1^n = argmax_{t_1^n} P(w_1^n|t_1^n) \cdot P(t_1^n)$$

$P(w_1^n|t_1^n) \simeq \prod_{i=1}^{n} P(w_i|t_i)$ (it depends only on its POS tag and independent of other words)

$P(t_1^n) \simeq \prod_{i=1}^{n} P(t_i|t_{i-1})$ (it depends only on the previous POS tag, thus, bigram)

$$\hat{t}_1^n = argmax_{t_1^n} \prod_{i=1}^{n} P(w_i|t_i) \cdot P(t_i|t_{i-1})$$

Two Probabilities

- The tag transition probabilities: P(ti|ti–1)
  - Finding the likelihood of a tag to proceed by another tag
  - Similar to the normal bigram model

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

- The word likelihood probabilities: P(wi|ti)
  - Finding the likelihood of a word to appear given a tag

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

Two Probabilities

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

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

I$_{[PRP]}$ saw$_{[VBP]}$ the$_{[DT]}$ man$_{[NN?]}$ on$_{[]}$ the$_{[]}$ roof$_{[]}$.

$$P([NN]|[DT])=\frac{C([DT],[NN])}{C([DT])}$$

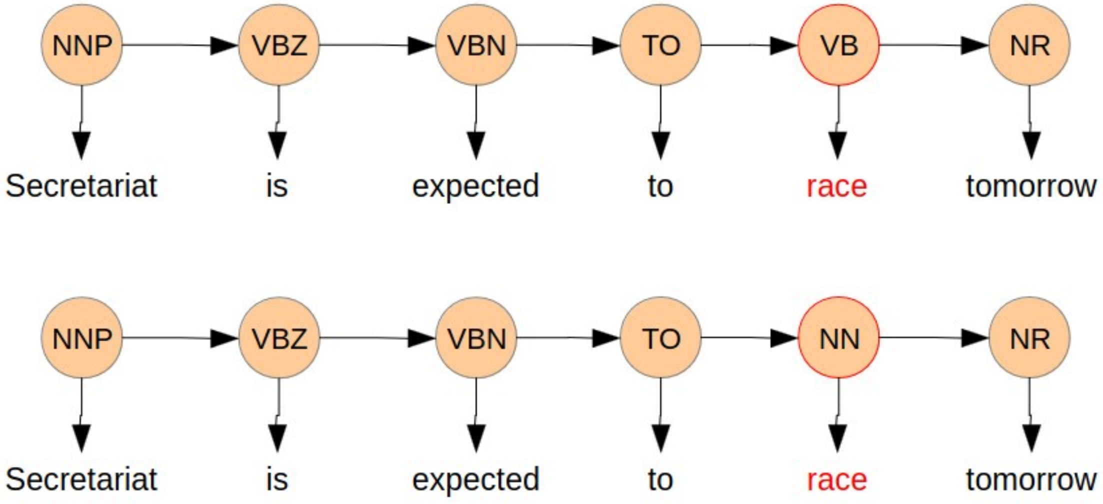
$$P(man|[NN])=\frac{C([NN],man)}{C([NN])}$$

Ambiguity in POS tagging

Secretariat$_{[NNP]}$ is$_{[VBZ]}$ expected$_{[VBN]}$ to$_{[TO]}$ race$_{[VB]}$ tomorrow$_{[NR]}$.

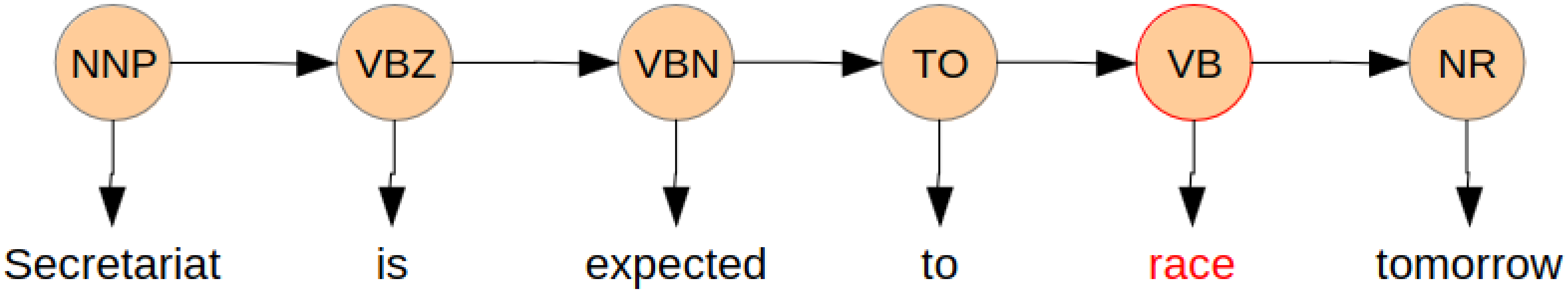
People$_{[NNS]}$ inquire$_{[VB]}$ the$_{[DT]}$ reason$_{[NN]}$ for$_{[IN]}$ the$_{[DT]}$ race$_{[NN]}$.

Secretariat$_{[NNP]}$ is$_{[VBZ]}$ expected$_{[VBN]}$ to$_{[TO]}$ race$_{[?]}$ tomorrow$_{[NR]}$.



Ambiguity in POS tagging

Secretariat$_{[NNP]}$ is$_{[VBZ]}$ expected$_{[VBN]}$ to$_{[TO]}$ race$_{[VB]}$ tomorrow$_{[NR]}$.



$$P(VB|TO) = 0.83$$

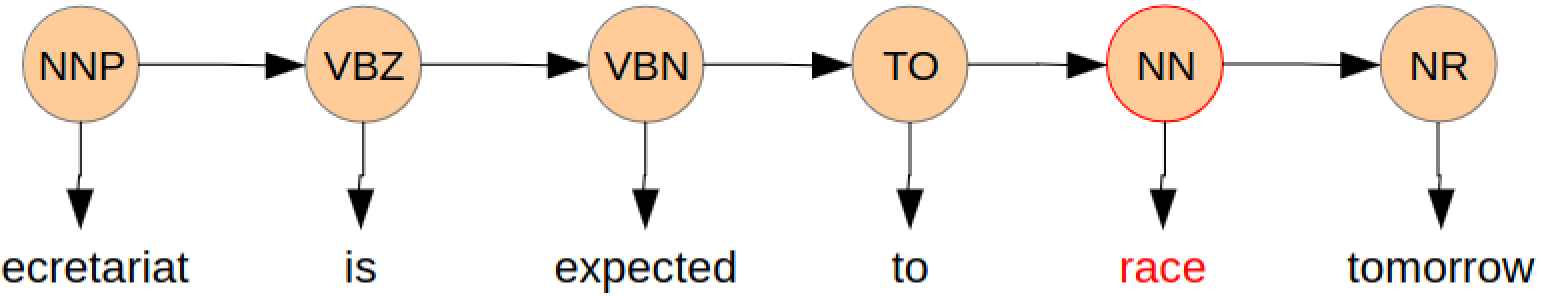
$$P(race|VB) = 0.00012$$

$$P(NR|VB) = 0.0027$$

$$P(VB|TO).P(NR|VB).P(race|VB) = 0.00000027$$

Ambiguity in POS tagging

Secretariat$_{[NNP]}$ is$_{[VBZ]}$ expected$_{[VBN]}$ to$_{[TO]}$ race$_{[VB]}$ tomorrow$_{[NR]}$.



$$P(NN|TO) = 0.00047$$

$$P(race|NN) = 0.00057$$

$$P(NR|NN) = 0.0012$$

$$P(NN|TO).P(NR|NN).P(race|NN) = 0.00000000032$$

Reference / Further reading

- https://towardsdatascience.com/hidden-markov-model-hmm-simple-explanation-in-high-level-b8722fa1a0d5
- https://www.mygreatlearning.com/blog/pos-tagging/

Thank you