

Natural Language Processing
Prof. Pushpak Bhattacharyya
Department of Computer Science and Engineering
Indian Institute of Technology, Bombay


Lecture - 17
HMM

In this lecture, we need to understand a very important mechanism, which is the hidden Markov model, we introduce this in the last lecture and we say that this is a very powerful tool for statistical natural language processing. We mention that part of speech tagging, which comes very early in the process of textual data has to use statistical techniques, and there the part of speech tags are learnt by training from corpus data. Now, the machine which is typically used for this purpose called the hidden Markov model, and we would like to devote our attention to this particular model.

(Refer Slide Time: 01:11)

Observations leading to why probability is needed

- Many intelligence tasks are sequence labeling tasks
- Tasks carried out in layers
- Within a layer, there are limited windows of information
- This naturally calls for strategies for dealing with uncertainty

 Probability and Markov process give a way


So, as we mentioned last time that natural language processing happens at many layers, many intelligence tasks are actually sequence labeling tasks, and these tasks are carried out in layers. So, tasks themselves a sequence labeling tasks and they happen in layers; within a layer there are limited windows of information, and this naturally calls for strategies for dealing with uncertainty, the moment the information becomes limited. We know that we do not have all information to make a decision, and therefore we must be

resilient to uncertainty, we must have strategies for dealing with them. Probability and Markov processes give a way of revealing with uncertainty.

(Refer Slide Time: 02:08)

"I went with my friend to the bank to withdraw
some money, but was disappointed to find it
closed"

POS	Bank (N/V)	closed (V/ adj)
Sense	Bank (financial institution)	withdraw (take away)
Pronoun drop	But I/friend/money/bank	was disappointed
SCOPE	With my friend	
Co-referencing	It -> bank	



Now, the example that was introduced and which we could look at in closer detail to understand these business of information processing layers, the information processing in a limited window within a layer. The sentence we have in front of us is, I went with my friend to the bank to withdraw some money, but was disappointed to find it close. So, the first label of process that happens is the part of speech tagging, the word syntactic categories grammatical categories are first of all deciphered.

So, from our discussion about a part of speech tagging we know that I is a pronoun, went is a verb, with is a preposition, my is a possessive pronoun, friend is a noun, to is a preposition, the is an article, the first difficult to emit with is in the word bank. Bank can be a verb meaning thereby I bank on my friends for their, support in case of distress or bank could be a noun, a river bank or a bank, where we deposit money or we withdraw money from.

To is a preposition, withdraw once again is problematic, because it can have noun and verb categories, almost all noun in English can be used as verbs and many verbs can be used as nouns. Some is an adjective it is a quantifier money is a noun comma is comma, but is a it is a conjunction, was is a auxiliary verb disappointed can be both an adjective and a verb.

In this case it is a verb, but one could say that disappointed crowd, the disappointed spectators, where this is acting like an adjective, to is preposition, find can be both noun and verb, I would like to find my key here it is a verb this new player is a terrific find, which means it is a noun. It is a pronoun closed can be an adjective or an verb here, it is closed, the bank is closed that you could have closed bank for example, and there are closed is an adjective.

We see that when we process the sentence on our wrought to the meaning we do find that we have to disambiguate categories of the words. So, here is the two layer the part of speech layer one or two examples of this are given bank can be noun or verb, closed can be verb or adjective similarly find, find can be noun or verb. Then we see the next layer which is the sense disambiguation layer, so if you go left to right once again I is a pronoun, number sense ambiguity went can be very, very ambiguous.

Go is a verb with many, many, many go has more than 50 meanings in English language, they and went is the past tense form at east part of speech disambiguation does not apply here, because nouns do not undergo perfolological transformation in tense, but went can have many different senses, depending on the senses of go. So, for example, go with the emotions means to conform with tradition, go home this is the act of going the act of going act of moving home, moving to home and there are many other meanings.

So, go is ambiguous, bank now is a classic example of ambiguitive, because it can be river bank or the financial bank, withdraw can mean to take away or go away. Similarly, there is a pronoun drop at this place, but I was disappointed, so I has to be put in here, but from an from the discourse segment, which came before. And there are many nouns and pronouns I, friend, bank, money, these are candidate nouns, but some of them will have type mismatch for example, money and bank cannot be disappointed, so we should have a alimentative here then I and friend are the candidates I is result.

So, this is the pronoun drop result, scope ambiguities is that with is a preposition and what is the scope of with, with can form preposition phrase, but how much of text does it demand to form a preposition phrase. So, here with my friend, but if it was with my friend from America, then the whole thing becomes preposition phrase with my friend from America, so the preposition phrase has to stop with friend.

So, this is a disambiguation task, because you have to decide the amount of text to eat up to form the phrase, then the other difficulty the final difficulty is co-referencing, where it. This pronoun has to rebound particular noun, so what was closed was bank, so it has to rebound to bank, but how does the system know, how does the machine know this, it will have to search amongst the possible nouns in the previous close. So, all these are different layers of information processing or disambiguation task, and they necessarily work with limited amount of information at any level the information is limited and therefore, there is uncertainty and we need a machine to deal with uncertainty.

(Refer Slide Time: 08:26)




So, the answer to that is often a very elegant machine the hidden Markov model, so we proceed to describe these model with its mathematical machinery and description and definition.


(Refer Slide Time: 08:35)

A Motivating Example


Colored Ball choosing



Urn 1
of Red = 30
of Green = 50
of Blue = 20




Urn 2
of Red = 10
of Green = 40
of Blue = 50



Urn 3
of Red = 60
of Green = 10
of Blue = 30

Probability of transition to another Urn after picking a ball:

	U ₁	U ₂	U ₃
U ₁	0.1	0.4	0.5
U ₂	0.6	0.2	0.2
U ₃	0.3	0.4	0.3



But, before that we have a motivating example which is found in a most texts and papers on the settings is as follows the three containers are in an urns, urn 1, urn 2, and urn 3. And there are totally 300 balls distributed amongst these urns with 100 balls each in one urn, so 1 urn contains 100 balls 2 contains 100 balls 3 contains 100 balls. And these 100 balls, again are of different colors for example, in urn 1 we have 30 red balls 50 green balls and 20 blue balls, similarly in urn 2 10 red balls 40 green balls 50 blue balls in urn 3 60 red 10 green and 30 blue balls.

You can see that if you sum these numbers horizontally, so red 30 red 10 and red 60 that is a total numbers of red balls is 100, which are distributed amongst the three urns. Similarly, the proportions of green balls or the total number of green balls are 100 distributed as 50 40 and 10 in the three urns respectively, blue balls are also 100 in number distributed 20 in first urn 50, in urn 2 and 30 in urn 3.

So, this is a situation and a person picks up balls from the urns, and places them one after the other, and the color of the ball is recorded there is a probabilities associated with the urn pick up itself, which urn the ball should come from is also probabilistically decided. And what is known is that we know the transition probability of going from one urn to another, so for example, if we look at this row and the corresponding column values we see that from urn 1 you can go to urn 1 again with probability of 0.1.

That means, if a person has drawn a ball from urn 1 the probability of him drawing the ball again from urn 1 is again 0.1, similarly the probability of drawing from urn 2 after the ball was drawn from urn 1 is 0.4, the probability from drawing from 0.3 after having drawn the ball from urn 1 is 0.5. So, a row gives the next urn after a particular urn, and a column gives the previous urns for a given urn, so this is the transition probability table.

(Refer Slide Time: 11:40)

Example (contd.)

Given :

	U ₁	U ₂	U ₃
U ₁	0.1	0.4	0.5
U ₂	0.6	0.2	0.2
U ₃	0.3	0.4	0.3


Observation : RRGGBRGR

State Sequence : ??

and

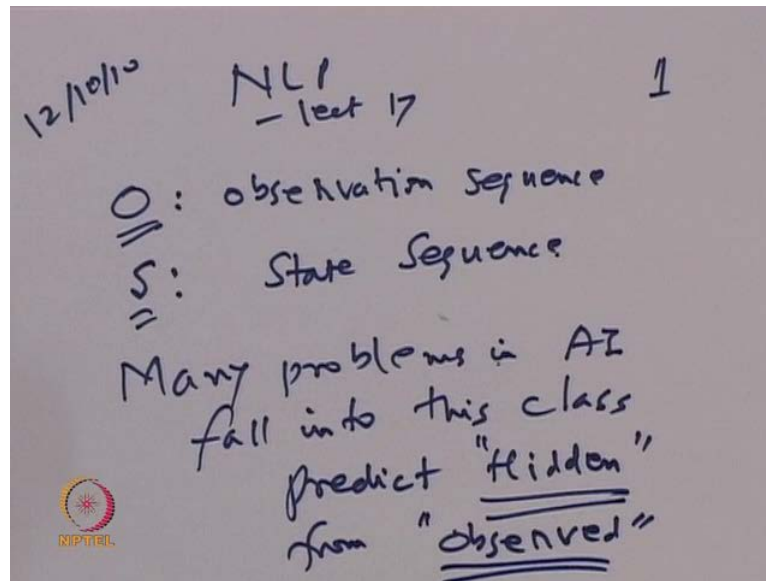
	R	G	B
U ₁	0.3	0.5	0.2
U ₂	0.1	0.4	0.5
U ₃	0.6	0.1	0.3

Not so Easily Computable.



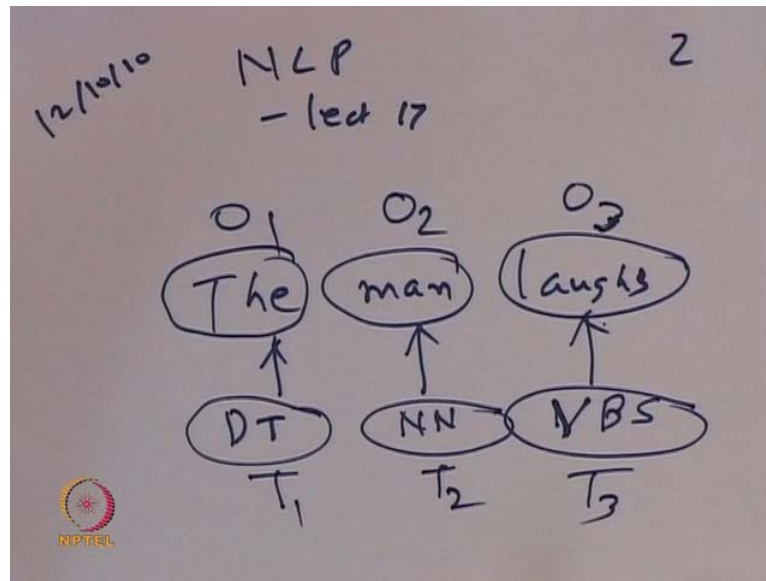
The other probability is the observation probability table, so for urn 1 the probability of drawing a red ball is 0.3, that a green ball is 0.5, that a blue ball is 0.2, and the reason for this is that the 30 percent of balls in urn 1 are red, 50 percent are green, 20 percent are blue, so the probability is this 0.3 0.5 and 0.2. Now, that the question that is asked is here is the observation of the colors of the balls drawn, R R G G B R G R that is red red green green blue red green green red. Then from this observations sequence can we identify the state sequence; that means, the urns which was chosen corresponding to each ball, so this is not easily computable we have to predict this.

(Refer Slide Time: 12:50)



Now, we make remarks, so when we discussing this problems, what we are saying is that the state has to be predicted corresponding to the observation. So, we have the observation sequence O and we have to predict the states, which is the state sequence the notation is O the observations sequence as the state sequence. Now, the point being made is that many problems in AI are of this kind fall into this class predict hidden from observed. So, this is the state of affairs we have observation sequence we have to predict a hidden sequence many problems of AI fall into this category for example, in machine learning in natural language processing in planning we will in particular concentrate on natural language processing.

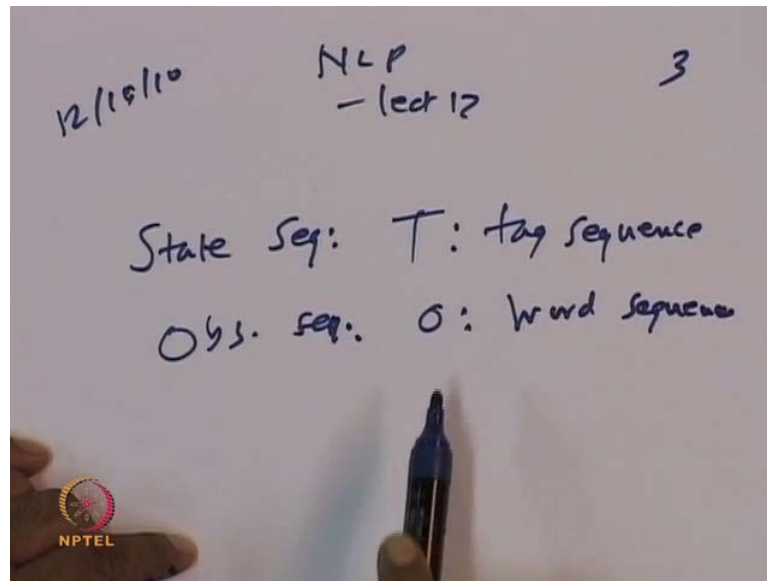
(Refer Slide Time: 13:54)



So, let me just illustrate this point, the point is that in part of speech tagging which we have understood to an extent and we have also looked at the derivations. In part of speech tagging, we have the words the man laughs in a sentence, now the part of speech for the is D T the determiner, the part of speech for man is noun, and the part of speech for laughs is verb, so determine and noun and verb.

Now, we have the sequence of words the man and laughs, the sequence labeling tasks is to produce part of speeches the D T N N V B S. So, as if you know the observations are our words, this is O 1 O 2 and O 3 and we have to predict what is hidden below this is T 1 the tag one T 2 and T 3, so that the problems, that comes in front of us now is that of predicting a state sequence from the observation sequence.

(Refer Slide Time: 15:04)



The state sequence in this case is the tag sequence, and the observation sequence is the word sequence, now this is a slightly unintuitive to a person, this may be looked up slightly unintuitive. The reason is that when you have these urns, we look at the transparency.

(Refer Slide Time: 15:52)

Example (contd.)

Given :

	U ₁	U ₂	U ₃
U ₁	0.1	0.4	0.5
U ₂	0.6	0.2	0.2
U ₃	0.3	0.4	0.3


and

	R	G	B
U ₁	0.3	0.5	0.2
U ₂	0.1	0.4	0.5
U ₃	0.6	0.1	0.3

Observation : RRGGBRGR

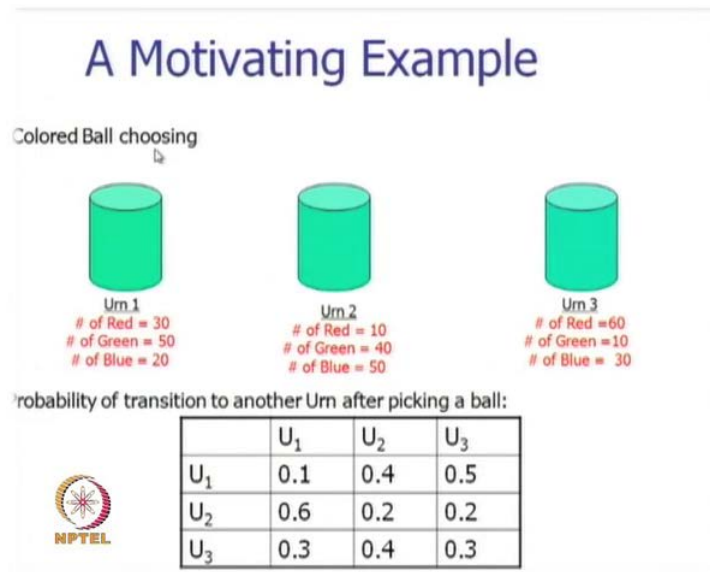
State Sequence : ??

Not so Easily Computable.



When we look at, when we see the urns, the urns can be looked upon as state, because urns are producing walls different colors walls.

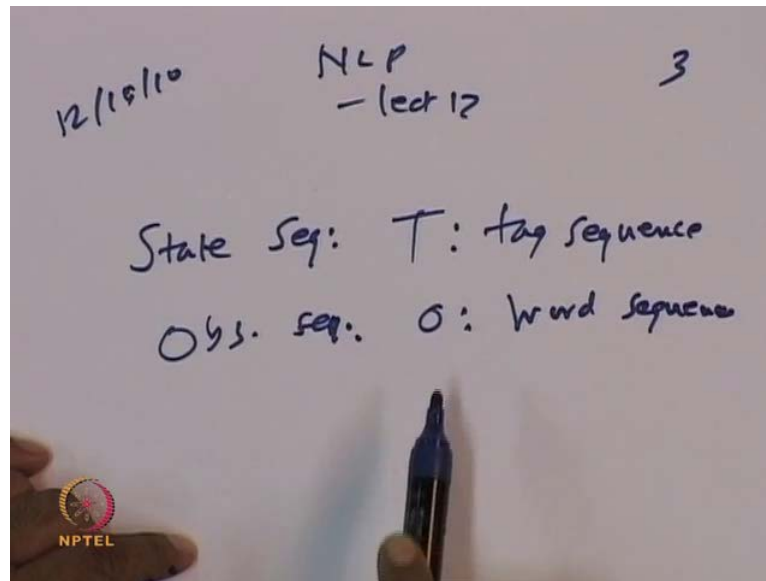
(Refer Slide Time: 15:56)



In what way can we say, that the what sequences produced by the attack sequence, how can we stretch this example to the example of language processing, where the problem is to produce tag sequence for the word sequence, the labeling task is to produce the texts. So, is not if somewhat unintuitive to think that there is a hidden sequence updates in the form of text, and the observations are words. Well, the best that can be said at this stage is that this is a convenient formulation for being able to apply a powerful tool named the hidden Markov model for a problem of part of speech tagging.

It may seem unintuitive that the state sequence or the tag sequence is producing the observation sequence, which is the words. Take a more common word is to think that the word sequence is producing the text, but that is a different approach of solving the same problem we will see that later. Now, in this case it is convenient for us to think of the tag sequencing the tag sequence generating the word sequence, so this is a generative way of solving a part of speech tagging problem we will look at a debate of discriminative verses generative model.

(Refer Slide Time: 17:39)



So, this is what we have written the state sequence is the tag sequence T observation sequence is O, the word sequence and we have seen the deduction for this as to how the tag sequence can be found out lets discuss the theory of HMM in more detail.

(Refer Slide Time: 18:02)

Example (contd.)

Given :

	U_1	U_2	U_3
U_1	0.1	0.4	0.5
U_2	0.6	0.2	0.2
U_3	0.3	0.4	0.3


and

	R	G	B
U_1	0.3	0.5	0.2
U_2	0.1	0.4	0.5
U_3	0.6	0.1	0.3

Observation : RRGGRRGR

State Sequence : ??

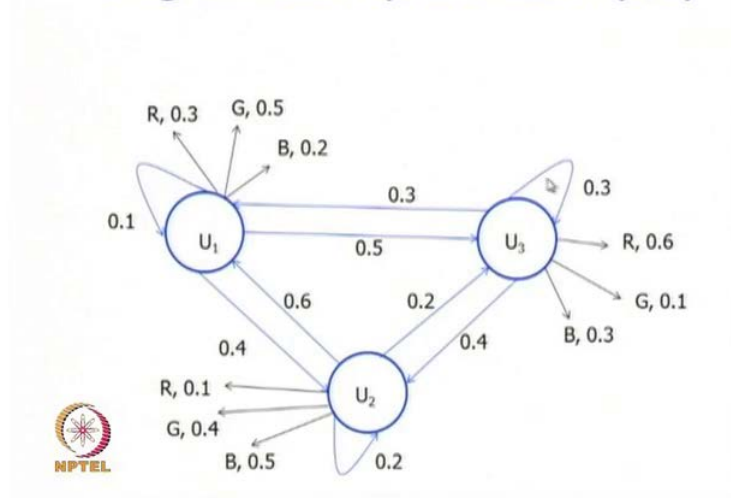
Not so Easily Computable.



So, we have got these observation probabilities the transition probabilities, the observation sequence in the form of the colors of balls are given the balls are of course, drawn with replacement. Now, the question is what is the state sequence at the urn sequence?

(Refer Slide Time: 18:19)

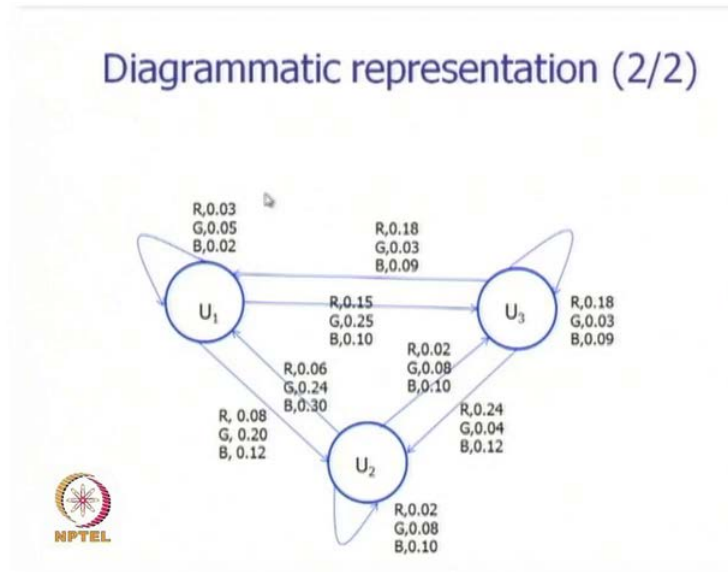
Diagrammatic representation (1/2)



From the transition table and the observation table, we can draw a machine here with transition probabilities and output probabilities, so we see here that urn 1 the probability of coming back to urn 1 immediately after urn 1 is 0.1 and the probability of drawing red green and blue ball is given as 0.3 0.5 and 0.2 respectively. Similarly, for urn 2 for the probability of coming back to urn 2 is 0.2 red green and blue ball probabilities are given here of drawing them, similarly for urn 3 these arrows here which has which has straight lines they indicate the transition probabilities.

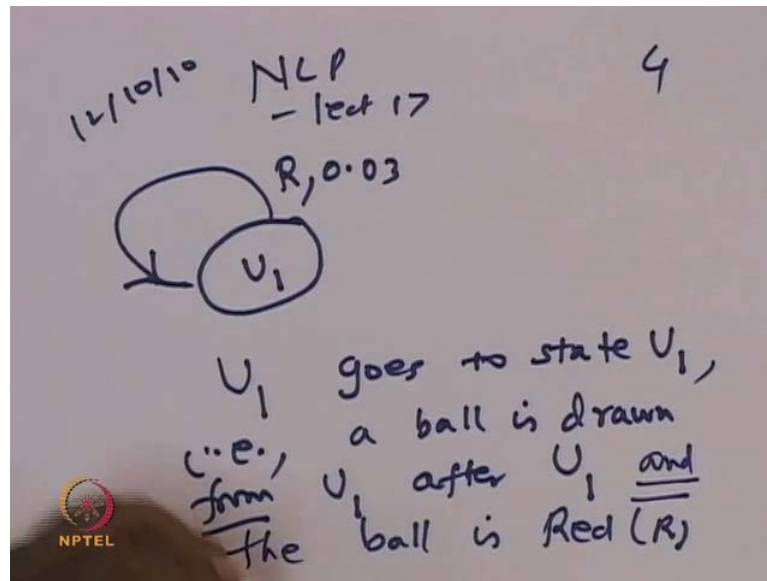
So, from U_1 we can go to U_2 with probability 0.4, so this was given in the transition probability table from U_2 I can go to U_3 with transition probability of 0.2 and so on. So, this is a convenient and very clearly representation, which embeds or combines both the output probability, probability of drawing this box and the transition probability.

(Refer Slide Time: 19:44)



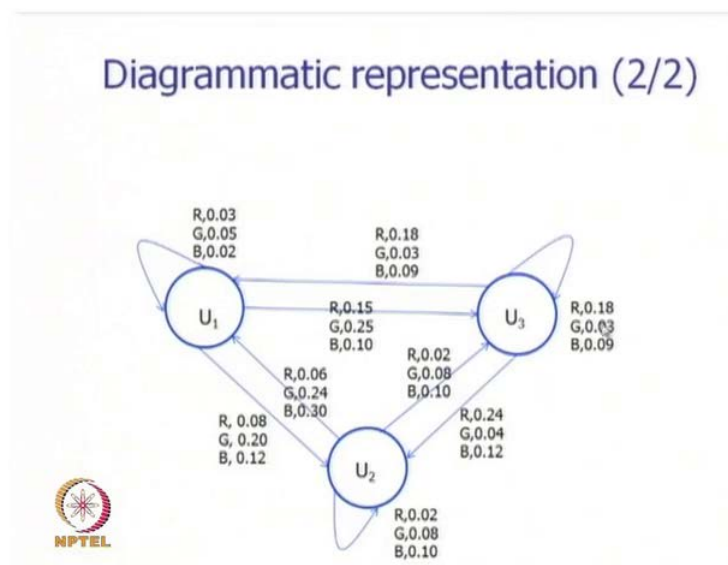
It turns out that we will not so many arrows after all, because we can combine the transition and output probabilities and we can have a more compact machine as shown here. So, for every state we have transitions from the state into another state, and the output probabilities are absorbed in this state transitions, in what way in this way. That the red ball R is shown here with comma, and then this number 0.03, what is this? This is the probability that U_1 will go back to U_1 and will produce a red ball, so this is a joint event it is a joint probability this R here with 0.03 has the following meaning. So, this is a machine which combines both transition and output probabilities, and we would like to see how it comes, but first of all let me write down the meaning of this probability R comma 0.03 with an arg here and state here.

(Refer Slide Time: 21:03)



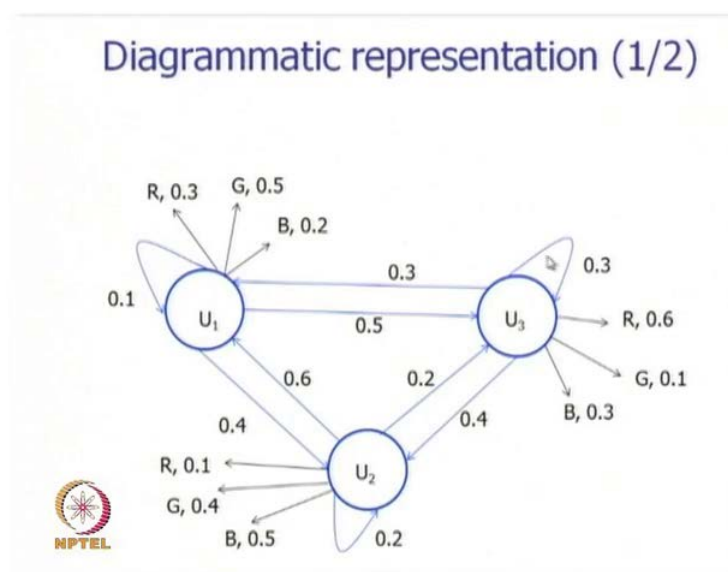
Their meaning is this that we have U_1 here and U_1 is coming back to U_1 with $R, 0.03$ the meaning of this diagram is that U_1 goes to state U_1 that is a ball is drawn from U_1 after U_1 and this is a joint probability and the ball is red. So, this is the meaning, the meaning is that we draw a ball again from U_1 , but before that we have produced a red ball arg from U_1 , so there is a probability of a composite event, the composite event is U_1 goes to U_1 and the ball drawn is red.

(Refer Slide Time: 22:04)



So, the slides show the other numbers also, so we pick up another quantity let us say suppose I take this particular quantity G comma 0.08 adjacent to these arrow. The arrow direction is from U_2 to U_3 the level is G that is green ball and the probability is 0.08 , so what it all means is that the probability of producing a green ball from U_2 and then going to urn U_3 is 0.08 . So, we produce a green ball from urn 2 and the next urn 2 choose is urn 3, so this is a composite event the probability of going from U_2 to U_3 and producing a G from U_2 .

(Refer Slide Time: 23:07)



So, this is a more compact representation compared to the previous representation, where the outgoing args where representing ball drawing probability, and the state to state args were drawing where depicting the transition probabilities.

(Refer Slide Time: 23:18)

Example (contd.)


- Here :
 - $S = \{U_1, U_2, U_3\}$
 - $V = \{R, G, B\}$
 - For observation:
 - $O = \{o_1 \dots o_n\}$
 - And State sequence
 - $Q = \{q_1 \dots q_n\}$
 - π is $\pi_i = P(q_1 = U_i)$

A =

	U ₁	U ₂	U ₃
U ₁	0.1	0.4	0.5
U ₂	0.6	0.2	0.2
U ₃	0.3	0.4	0.3

B =

	R	G	B
U ₁	0.3	0.5	0.2
U ₂	0.1	0.4	0.5
U ₃	0.6	0.1	0.3




So, this is the state sequence here this is the set U 1 U 2 and U 3, the observation sequence is from the set R G and B, observation are written as one port o 1 to o n and state sequences q 1 to q n. There is initial probabilities associated with the states what is the probability that we will draw a ball first, that is why draw a ball first from urn 1. What is the probability that we first draw a ball from urn 2 probability first draw a ball from urn 3 and so on, and then this a is transition probability table b is the observation probability table.

(Refer Slide Time: 24:03)

Observations and states

	O ₁	O ₂	O ₃	O ₄	O ₅	O ₆	O ₇	O ₈
OBS:	R	R	G	G	B	R	G	R
State:	S ₁	S ₂	S ₃	S ₄	S ₅	S ₆	S ₇	S ₈

S_i = U₁/U₂/U₃; A particular state
 S: State sequence
 O: Observation sequence
 S* = "best" possible state (urn) sequence
 Goal: Maximize P(S*|O) by choosing "best" S




So, the observation and states can be depicted in this way we have the observations R R G G B R G R, these observations are O 1 O 2 O 3 up to O 8 and states corresponding to these are S 1 S 2 S 3 up to S 8, so each states is producing a particular color of the wall a particular observation. Now, what are these S's each S i is a particular state, and it can take values U 1 U 2 or U 3; that means, one of the three args. S i is a particular state, but S whole thing is a state sequence, O is an observation sequence in this case it is these R R G G B R G R. We are interested in S star the best possible state or urn sequence, our goal is to maximize the S given O by choosing the best S and S star is the best possible state sequence.

(Refer Slide Time: 25:09)

Goal

- Maximize $P(S|O)$ where S is the State Sequence and O is the Observation Sequence

$$S^* = \arg \max_s (P(S | O))$$


So, maximizing probability value for a particular value of the independent variable this is captured by the argmax function, so this is expressed here as star the best possible states sequence, which is equal to argmax. Over all possible S, where the argmax happens over these expression P S given O; that means, we vary S and O is the observation sequence which is given, so we vary S and obtain different values of these probabilities values P S given O. Wherever, we reach a maximum of P S given O we record, that particular S and the output is given as that particular S which is S star, so this is the best S for possible for us, because for that S the probability P S given O S maximized.

(Refer Slide Time: 26:11)


False Start

	O ₁	O ₂	O ₃	O ₄	O ₅	O ₆	O ₇	O ₈
OBS:	R	R	G	G	B	R	G	R
State:	S ₁	S ₂	S ₃	S ₄	S ₅	S ₆	S ₇	S ₈

$$P(S|O) = P(S_{1-8} | O_{1-8})$$

$$P(S|O) = P(S_1|O).P(S_2|S_1,O).P(S_3|S_{1-2},O)...P(S_8|S_{1-7},O)$$

By Markov Assumption (a state depends only on the previous state)

$$P(S|O) = P(S_1|O).P(S_2|S_1,O).P(S_3|S_2,O)..P(S_8|S_7,O)$$


Now, here is a false start possibility in the mathematical treatment of this situation, where we find that given the observation sequence our interest is in the best possible state sequence. So, suppose we mathematically treat these expression P S given O, now P S given O a is nth, but P S 1 to 8 given O 1 to O 8, so this is a convenient notation shows state sequence S 1 S 2 up to S 8, and is written in a short hand 1 to 8.

Similarly, the observation sequence O 1 O 2 O 3 up to O 8 written short hand fashion O 1 dash 8, so by applying chain rule we obtain P S given O S P S 1. And the first state S 1 O into P S 2 given S 1 comma O S 1 becomes the conditioning variable then P S 3 given S 1 and S 2 which is written as S 1 dash 2 and O and lastly P S 8 given S 1 to 7 and O, so S 1 to 7 should be understood to mean S 1 S 2 S 3 up to S 7.

So, this is the whole probability is broken down into a set of probability values was product is taken, now we invoke the Markov assumption. What do we do? the Markov assumptions states that any set any state actually depends only on the previous state, so from this expression for example, P S 3 given S 1 S 2 and O, we can drop S 1 because S 1 is not the immediately preceding state.

So, P S 3 given S 1 2 and O will become P S 3 given S 2 and O, and here we have P S 8 given S 7 and O all the previous states from S 1 to S 6 are ignored. So, this is a Markov assumption a state depends only on the previous state and what is happening here is that we are making an assumption of Markov state for order k equal to 1. So, this is order one

Markov assumption; that means, only one previous state matters, similarly we could have order two Markov assumption, where a state depend on previous 2 states.


Now, the problem is when you have treated the expression mathematically this way, we find that we have this probability values $P(S_1 | O)$, $P(S_2 | S_1 \text{ and } O)$, $P(S_3 | S_2 \text{ and } O)$ and so on, but these probability values cannot be computed easily. These parameters are not obtainable easily, they are cumbersome items which need to be processed further, instead of that we can make use of another very nice theorem called the Baye's theorem, which along with Markov assumption is a very powerful theoretical construction.

(Refer Slide Time: 29:23)

Baye's Theorem

$$P(A | B) = P(A) \cdot P(B | A) / P(B)$$

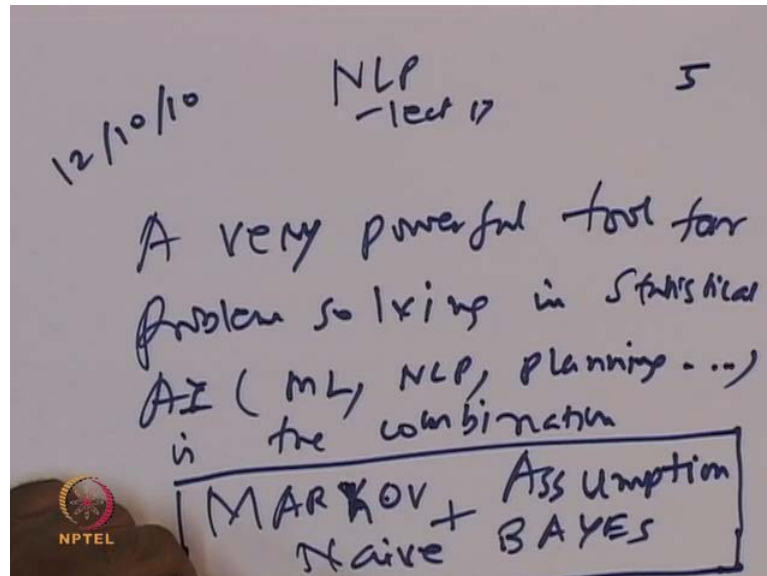
$P(A)$ -: Prior
 $P(B|A)$ -: Likelihood

$$\operatorname{argmax}_S P(S | O) = \operatorname{argmax}_S P(S) P(O | S)$$


Baye's theorem states that $P(A | B)$, where A and B are Random variables $P(A | B)$ is equal to $P(A) \cdot P(B | A) / P(B)$, so this is the way a probability value is turned around $P(A | B) = P(A) \cdot P(B | A) / P(B)$, but it is multiplied by $P(A)$ and divided by $P(B)$. So, $P(A)$ is called the prior probability $P(B | A)$ is called the likelihood, now when we take argmax over S of $P(S | O)$, we can equivalently take the expression as argmax over S of $P(S) \cdot P(O | S)$. It is not difficult to see that the denominator $P(O)$ can be ignored, because it is independent of S and it comes as a constant factor for all these expressions $P(S | O) = P(S) \cdot P(O | S) / P(O)$. So, $P(O)$ is independent of S therefore, it can be eliminated from consideration, so this is base theorem and it is very powerful

construct, so let me make some remark here and write it down in the form of the list of things.

(Refer Slide Time: 30:50)



So, we are saying that based theorem is invoked along with Markov assumption, so a very powerful tool for problem solving in statistical AI, which means machine learning nlp and planning and so on. A very powerful tool for problem solving statistical AI the combination Markov assumption plus Naive Baye's, so this becomes a very powerful tool for solving many different problems in here Markov assumption plus Naive Baye's. Naive Baye's is the utilized through the Baye's theorem Markov assumption is invoked through the dependence of states on previous dates, so first remember this Markov assumption plus Naive Baye's.

(Refer Slide Time: 32:08)

State Transitions Probability

$$P(S) = P(S_{i-8})$$

$$P(S) = P(S_1)P(S_2|S_1)P(S_3|S_1, S_2)P(S_4|S_1, S_2, S_3) \dots P(S_8|S_1, S_2, S_3, S_4, S_5, S_6, S_7)$$

By Markov Assumption ($k=1$)

$$P(S) = P(S_1)P(S_2|S_1)P(S_3|S_2)P(S_4|S_3) \dots P(S_8|S_7)$$



So, after invoking the base theorem, now we find that state transition probability the prior which is $P(S)$ can be treated in the following way, $P(S)$ is nothing but p of S from 1 to 8. So, this actually a sequence p of $S_1 S_2 S_3$ up to S_8 , and this can be written as $P(S_1)$ into $P(S_2 \text{ given } S_1)$ $P(S_3 \text{ given } S_1 \text{ and } S_2)$ $P(S_4 \text{ given } S_1 S_2 S_3)$ and so on, the final probability is $P(S_8 \text{ given } S_1 S_2 \text{ up to } S_7)$.

Now, Markov assumption of order one k is equal to 1, we can drop all those states, which are not immediately preceding, so $P(S)$ becomes equal to $P(S_1)$ into $P(S_2 \text{ given } S_1)$ $P(S_3 \text{ given } S_2)$ $P(S_4 \text{ given } S_3)$ and so on until $P(S_8 \text{ given } S_7)$. So, this is an easy expression which can be computed by some means which you have so see, so the prior probability has become a product of these kind of by graham probabilities, we call this by graham probabilities. We have here two states S_1 and S_2 interacting with each other, S_3 and S_2 interacting with each other in general S state as k interacting with S_{k+1} , so we have treated the prior probability component.

(Refer Slide Time: 33:36)

Observation Sequence probability

$$P(O|S) = P(O_1|S_{1-8})P(O_2|O_1, S_{1-8})P(O_3|O_{1-2}, S_{1-8}) \dots P(O_8|O_{1-7}, S_{1-8})$$

Assumption that ball drawn depends only on the Urn chosen

$$P(O|S) = P(O_1|S_1)P(O_2|S_2)P(O_3|S_3) \dots P(O_8|S_8)$$

$$P(S|O) = P(S)P(O|S)$$

$$P(S|O) = P(S_1)P(S_2|S_1)P(S_3|S_2)P(S_4|S_3) \dots P(S_8|S_7)$$

$$P(O_1|S_1)P(O_2|S_2)P(O_3|S_3) \dots P(O_8|S_8)$$

NPTEL

We come to the next probability component $P(O|S)$, so $P(O|S)$ can be written as $P(O_1|S_{1-8})$, $P(O_2|O_1, S_{1-8})$, $P(O_3|O_{1-2}, S_{1-8})$ up to $P(O_8|O_{1-7}, S_{1-8})$. So, now, we invoke another assumption and this complicated expression gets simplified we assume that a ball drawn depends only on the urn chosen and there is a very reasonable assumption nothing else in the current situation, because when a ball is drawn with replacement from an urn.

Presumably, there are number other factor in the world influences the color of that ball drawn, it purely depends on the urn from here it is coming because which we have a probability associated with the color coming from an arg. So, when we make that assumption we know that O_1 only depends on S_1 , O_2 depends only on it is current urn, O_2 depends on current which is current state S_2 , O_3 depends on current state S_3 , and nothing else.

And therefore, all these probability factors can be written as $P(O_1|S_1)$ into $P(O_2|S_2)$ $P(O_3|S_3)$ and finally, $P(O_8|S_8)$. So, once you have got all these expressions ready, we can see that the probability values come in the following way, $P(S|O)$ is $P(S)$ into $P(O|S)$ and this is equal to $P(S_1)$ into $P(S_2|S_1)$ $P(S_3|S_2)$ $P(S_4|S_3)$ $P(S_5|S_4)$ $P(S_6|S_5)$ $P(S_7|S_6)$ $P(S_8|S_7)$. And the probability values coming from the obsess, $P(O_1|S_1)$ given

S 1 P O 2 given S 2 finally, O 2 given S eight. So, these probability values can be grouped together.

(Refer Slide Time: 35:52)

Grouping terms


O ₀	O ₁	O ₂	O ₃	O ₄	O ₅	O ₆	O ₇	O ₈	
Obs: ε	R	R	G	G	B	R	G	R	
State: S ₀	S ₁	S ₂	S ₃	S ₄	S ₅	S ₆	S ₇	S ₈	S ₉

$$\begin{aligned}
 &P(S) \cdot P(O|S) \\
 &= [P(O_0|S_0) \cdot P(S_1|S_0)] \cdot \\
 &\quad [P(O_1|S_1) \cdot P(S_2|S_1)] \cdot \\
 &\quad [P(O_2|S_2) \cdot P(S_3|S_2)] \cdot \\
 &\quad [P(O_3|S_3) \cdot P(S_4|S_3)] \cdot \\
 &\quad [P(O_4|S_4) \cdot P(S_5|S_4)] \cdot \\
 &\quad [P(O_5|S_5) \cdot P(S_6|S_5)] \cdot \\
 &\quad [P(O_6|S_6) \cdot P(S_7|S_6)] \cdot \\
 &\quad [P(O_7|S_7) \cdot P(S_8|S_7)] \cdot \\
 &\quad [P(O_8|S_8) \cdot P(S_9|S_8)].
 \end{aligned}$$

We introduce the states S₀ and S₉ as initial and final states respectively.

After S₈ the next state is S₉ with probability 1, i.e., P(S₉|S₈)=1

O₀ is ε-transition



We will not worry about the first line and the last line in between we see that a grouping has been done, so we have P O 1 given S 1 into P S 2 given S 1 P O 2 given S 2 and P S 3 given S 2 P O 3 given S 3 and P S 4 given S 3. So, in any such line what is happening is that we have a state S k and the corresponding observation O k multiplied by the probability of the state and it is next state, probability of next state given in the previous state.

So, this is systematically done for all the lines until O 7 because S 7 is the corresponding state and the next state is also well defined S 8. So, this is shown in the picture here we have O 1 O 2 up to O 7, and we have different colored walls which are coming as observation sequence and below that is the state sequence S 1 to S 8. This S 9 here is the final state, S 0 is introduced as a fictitious state this is the state where the number urn has been chosen, and the drawing of balls has to begin.

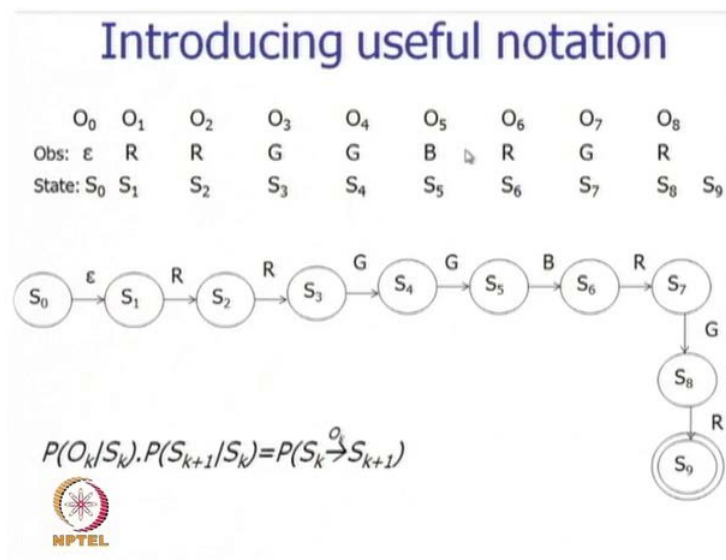
So, this is the state S 0 this is the start of the ball drawing situation, and here the observations is the epsilon, epsilon is a an empty transition which essentially means that nothing is done except that the state is changed from S 0 to S 1, epsilon is the empty streak. So, the meaning of these is that initially with there is not red green or blue ball

produced, we go straight from initial state S_0 to 1 of the urns as is expressed by the fact that each state can take a value U_1 , U_2 or U_3 .

So, now, O_0 given S_0 probability of O_0 given S_0 , so probability of epsilon transition given S_0 that is always 1, so initially number ball will be drawn, and we will go from a starting state to a state which corresponds to the urn. And similarly, we will go from state S_0 to S_1 positively S_1 can be 1 of U_1 , U_2 or U_3 this is probability, which is saying that we will begin choosing urn, after all the last line is also important we had $P(O_8 | S_8)$, now this is multiplied by $P(S_9 | S_8)$ given S_8 state.

Now, S_8 is the eight is the length of observation sequence S_8 is the last state corresponding to a ball drawn and S_9 is the final state, so $P(S_9 | S_8)$ given S_8 is definitely going to be 1. Therefore, it does not disturb the equality, when we introduced this stop, so we have taken the observation sequence we have taken the state sequence and we have done something useful here or something necessary here. We have introduced a final state and we have introduced an initial state with it is observation and the observation is epsilon, the purpose of doing this is that we want to capture the start of the process and we also we want to captured the end of the process, that is one purpose.

(Refer Slide Time: 39:47)



The next purpose is that we can combine the terms and we can introduce a useful notation, so what we have seen last time, in the last slide is that we have O_k for S state S_k and a transition S_k to S_{k+1} . We write these notation $P(O_k | S_k)$ given S_k into $P(S_k \xrightarrow{O_k} S_{k+1})$

S_k plus S_{k+1} given S_k as probability of $S_k \rightarrow S_{k+1}$, the meaning of that arrow is that the next state after that S_{k+1} on the output or input to be O_k .

So, O_k is the output and the state transition is from S_k to S_{k+1} , so this is diagrammatically depicted here we know that S_0 is the starting state there is a epsilon transition to S_1 S_0 goes to S_1 , S_1 goes to S_2 , S_2 goes to S_3 , S_3 goes to S_4 , S_4 to S_5 and so on. And everywhere, the arg is marked with the observation; that means, S_1 after producing r goes to S_2 in other words S if S_1 S was one of the urns, let us say urn number 2 and suppose S_2 is urn number 3. So, we go from urn number 2 to urn number 3 with the transition arg; that means, having drawn the red ball, so one you can see the that men's utility and in tube tubeness of this particular expression, this terms or probability expression into a actually a probability finitate state automitan.

(Refer Slide Time: 41:45)

Grouping terms


O_0	O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8
Obs: ϵ	R	R	G	G	B	R	G	R
State: S_0	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8 S_9

$$\begin{aligned}
 &P(S) \cdot P(O|S) \\
 &= [P(O_0|S_0) \cdot P(S_1|S_0)] \cdot \\
 &\quad [P(O_1|S_1) \cdot P(S_2|S_1)] \cdot \\
 &\quad [P(O_2|S_2) \cdot P(S_3|S_2)] \cdot \\
 &\quad [P(O_3|S_3) \cdot P(S_4|S_3)] \cdot \\
 &\quad [P(O_4|S_4) \cdot P(S_5|S_4)] \cdot \\
 &\quad [P(O_5|S_5) \cdot P(S_6|S_5)] \cdot \\
 &\quad [P(O_6|S_6) \cdot P(S_7|S_6)] \cdot \\
 &\quad [P(O_7|S_7) \cdot P(S_8|S_7)] \cdot \\
 &\quad [P(O_8|S_8) \cdot P(S_9|S_8)].
 \end{aligned}$$

We introduce the states S_0 and S_9 as initial and final states respectively.

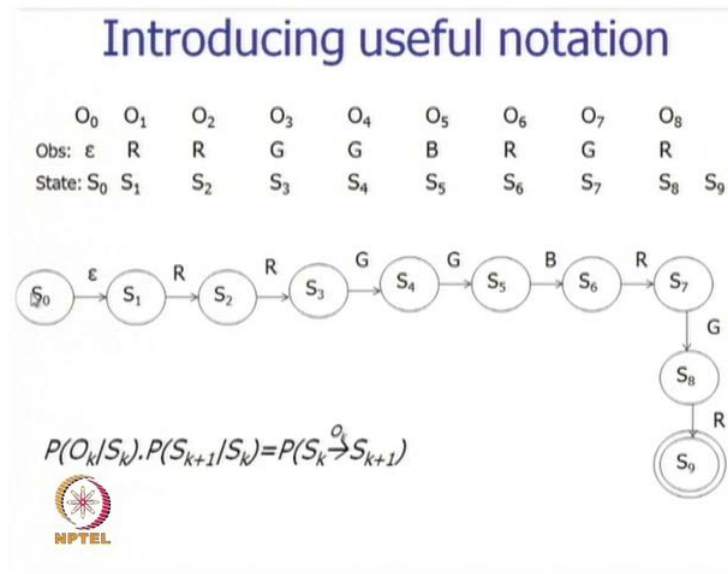
After S_8 the next state is S_9 with probability 1, i.e., $P(S_9|S_8)=1$

O_0 is ϵ -transition



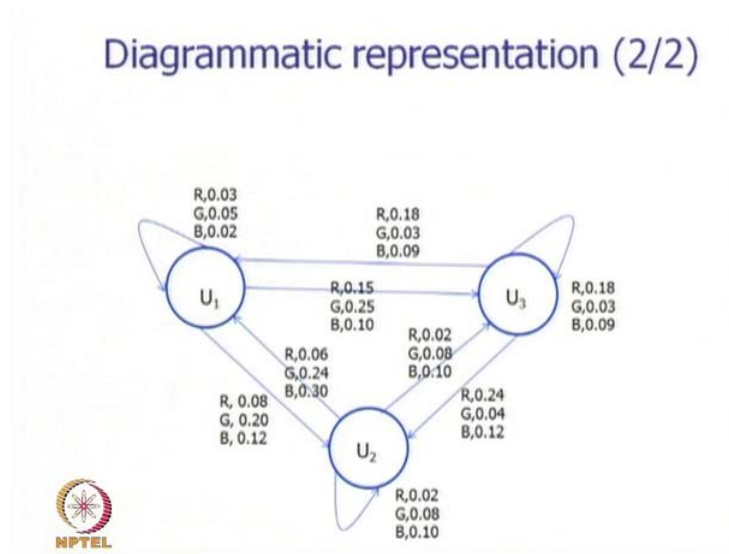
So, these expression that we had produced of observation probability and state probability and we group them together, so $P(O_0|S_0)$ into $P(S_1|S_0)$ $P(O_1|S_1)$ given S_1 into $P(S_2|S_1)$. So, these are the groups and for each of them we have now a notation for example, what will be the notation for this, the notation for this is that this is a state S_0 going to S_1 S_0 to S_1 on the symbo O_0 . So, similarly these expression if we try it goes from state S_5 to S_6 as it indicate with a conditional probability here and the output produced is O_5 , which is one of the colored balls.

(Refer Slide Time: 42:37)



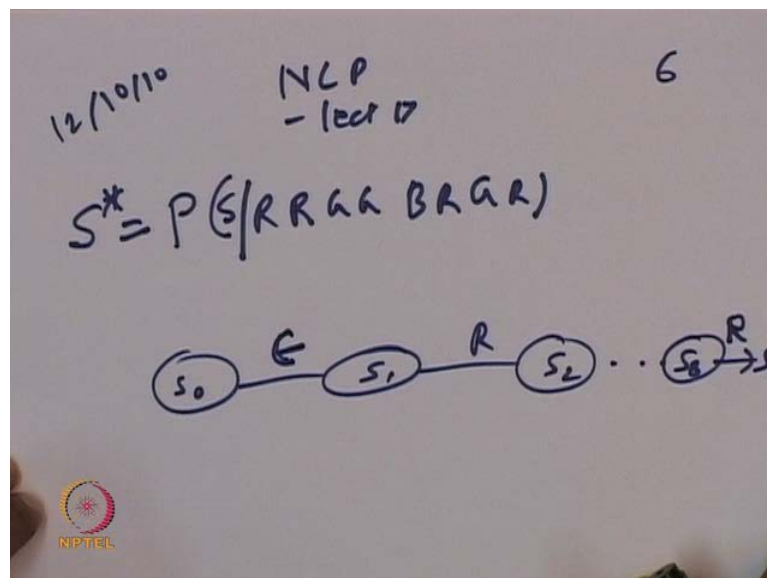
Therefore, these probabilistic expression P S into P O given S has now, become a finite state machine with probabilities associated with the args and the observations also marked on the args, so these expression has become finite state machine. So, this is a very, very useful notation, because the HMM after all is a finite machine with probabilities on the args. And the probabilities on the args come from these theory, which has been worked out in front of you P O k given S k into P S 5 P S 5 given S 4, which is equal to P S 4 to S 5 on the symbol O 5. So, these is the finite state machine depicting that situation here, so now, we would like to see if you can make use of this machine and compute the best probability value.

(Refer Slide Time: 43:44)



So, this we will draw on the paper and show you the competition for this i need the probability values i need this machine.

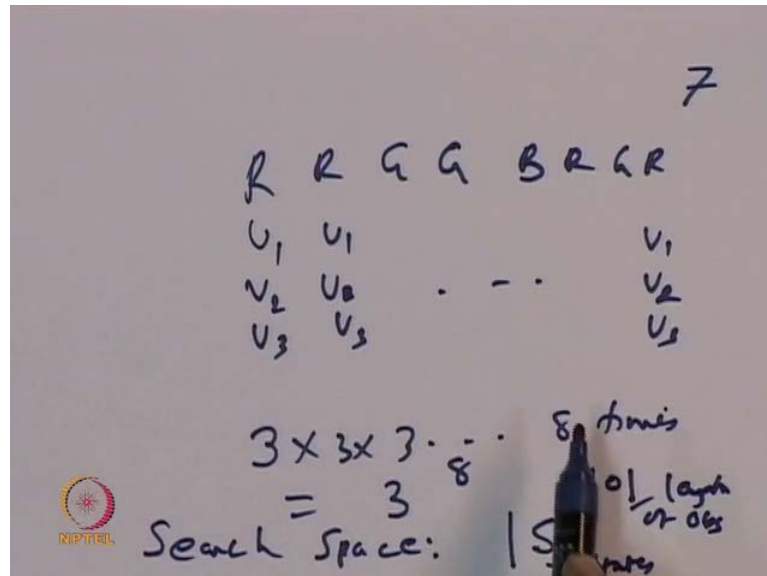
(Refer Slide Time: 43:59)



So, our goal now is to get the probability value of this sequence P R R G G B R G R, we want to get the probability of this sequence or the probability of the state sequence S given this observation sequence, and the st best state sequence probabilities is a star this we would like to get. And this we can get from the automaton that we had got which is S 0 S 1 then S 1 S 2 1 R and so on and finally, we have S 8 going to S 9 with observation

R, and there are probability values associated with them, this corresponds to the following search configuration.

(Refer Slide Time: 45:01)



If we don't have any restrictive influence and we produce the state sequence for this which is U₁ U₂ or U₃, then the number of possibilities are 3 into 3 into 3, 8 times. So, the search space becomes 3 to the power 8 or the search space size would be the number of states to the power the length of the observation states length of Obs. So, this is a tremendous amount of computation, we will see in the next class how to efficiently compute it by viterbi algorithm.