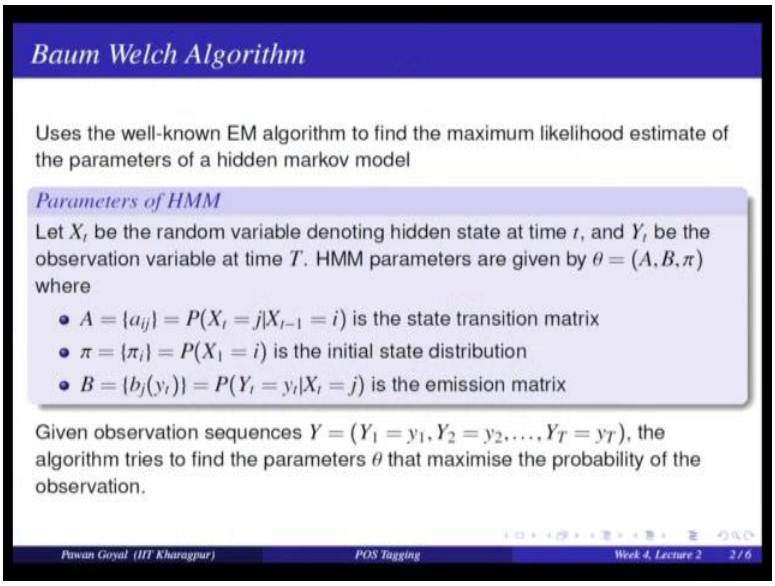
National Language Processing
Prof. Pawan Goyal
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture – 17
Baum Welch Algorithm

So, welcome back for the second lecture of this week. So, in the last lecture, we were discussing Viterbi decoding for HMMs. And in the end, we discussed the problem of learning the parameters of HMMs. And we say that when the label data set available, we can simply estimate using maximum likelihood estimate by the label data set. But if the label data set is not available then how do we actually learn the parameters of the system and we said we will be using some sort of expectation algorithm and in this particular case this is called Baum Welch algorithm.

(Refer Slide Time: 00:57)



Baum Welch Algorithm

Uses the well-known EM algorithm to find the maximum likelihood estimate of the parameters of a hidden markov model

Parameters of HMM

Let X_t be the random variable denoting hidden state at time t, and Y_t be the observation variable at time T. HMM parameters are given by $\theta = (A, B, \pi)$ where

- $A = \{a_{ij}\} = P(X_t = j | X_{t-1} = i)$ is the state transition matrix
- $\pi = \{\pi_i\} = P(X_1 = i)$ is the initial state distribution
- $B = \{b_j(y_t)\} = P(Y_t = y_t | X_t = j)$ is the emission matrix

Given observation sequences $Y = (Y_1 = y_1, Y_2 = y_2, \ldots, Y_T = y_T)$, the algorithm tries to find the parameters θ that maximise the probability of the observation.

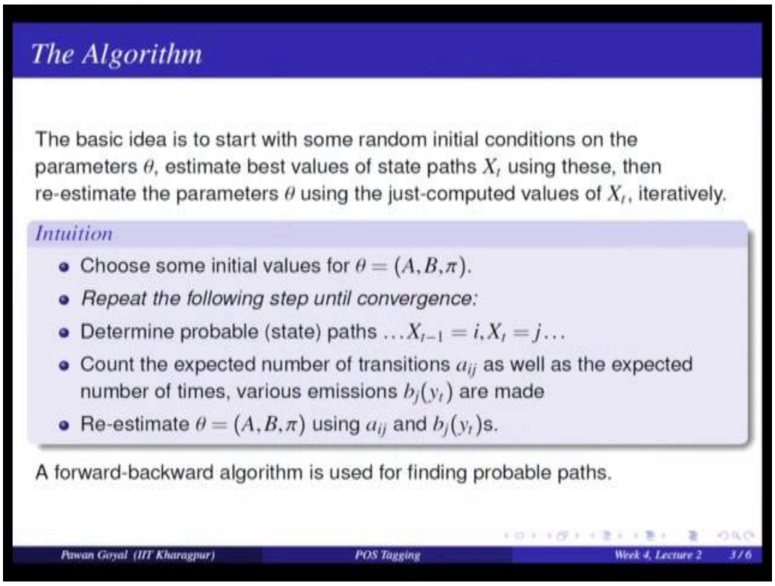
Pawan Goyal (IIT Kharagpur) POS Tagging Week 4, Lecture 2 2 / 6

So, Baum Welch algorithm as I was saying this is a well known EM algorithm to estimate the maximum likelihood for the parameters of the HMM. So, what are the parameters of HMM that you want to estimate let we just formally define them and then we will see how do we estimate them using Baum Welch. So, in the HMM model, so we have a hidden state at every time point and an observation variable. So, here I am using capital X t to denote the random variable that is the hidden state and capital Y to denote the observation variable. So, both X t and Y t can take many different values; for our part of speech tag is each X t can take one of the part of speech tags, and each Y t can take one of the possible words. So, this my hidden variable hidden state X t and the observation Y t.

The parameters are A, B and pi that we discussed last time. What is A? So A is my state transition matrix. So, given that the previous state was i what is the probability that the current state will be j? So, here the t as such does not matter what matters is that the state j occurred after state i at any given time point. Then I have the parameter pi that is what is the probability that this particular state is starts the sequence. So, probability of X 1 that the first state being the i x state that is my pi i. And then thirdly I have the emission matrix. So, where the entries are what is the probability of observing this particular word or this observation given the current state. So, this I am denoting like b j y t; at the state j what is the probability of emitting the observation variable or observation y. So, either three parameters three set of parameters that I have learnt.

Now, what is given to me? So, I was saying that we are given the corpus not labeled, but the word sequences are given. So, this can be seen as my observations are given different sentences are given, I knew what are different words occurred in the sentences. So, I can say that I am given a set of observations in data. So, here I can say that I am given a set of observation sequences that the first observation was the word Y 1 then Y 2 up to some Y t, where Y t denote the end of the sentence. So, I am given a set of observation sequences and my aim is to find out the optimal set of parameters theta that maximize the probability of this observation. So, these observation sequences, so that is where I am using Baum Welch algorithm, find out what are the optimal set of parameters theta that will maximize the probability of likelihood of observing this observation that is why I am using expectation maximization algorithm.

(Refer Slide Time: 04:31)



The Algorithm

The basic idea is to start with some random initial conditions on the parameters θ, estimate best values of state paths X_t using these, then re-estimate the parameters θ using the just-computed values of X_t, iteratively.

Intuition

- Choose some initial values for $\theta = (A, B, \pi)$.
- *Repeat the following step until convergence:*
- Determine probable (state) paths $\ldots X_{t-1} = i, X_t = j \ldots$
- Count the expected number of transitions a_{ij} as well as the expected number of times, various emissions $b_j(y_t)$ are made
- Re-estimate $\theta = (A, B, \pi)$ using a_{ij} and $b_j(y_t)$s.

A forward-backward algorithm is used for finding probable paths.

Pawan Goyal (IIT Kharagpur) POS Tagging Week 4, Lecture 2 3 / 6

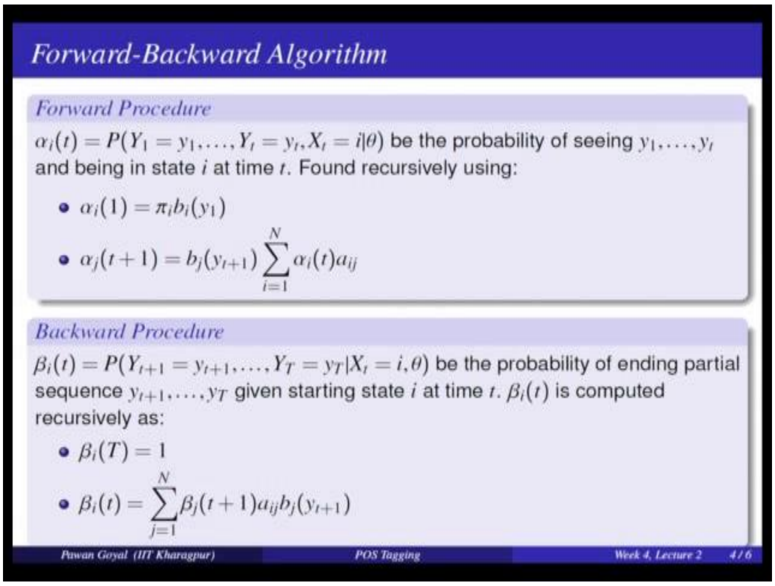
So, what is the main intuition of EM? So, idea would be I want to estimate the parameters A, B, and pi. So, what I will do? I will start with some random initial probabilities for all these parameters; I will initialize them with some random values. So, I have all these probabilities. Now, using these initial values, I will try to find out what are the probabilities of various state paths. So, on a given sequence, what is the probability that a particular state would have occurred given these parameters? So, once I have obtained the probabilities of different state paths or different states a different point, now I got this probability, I will use that again to compute to re-estimate my parameters. So, I will get started with theta zero, I get some theta prime. Now, I will again use the theta prime to complete my state paths probabilities again use that to re-compute my parameters until it somehow converges.

So, this is my iterative algorithm. First use some parameters theta to get some likely from the data for the hidden variable here that is the states. Once you have that likelihood or the probabilities use that to compute my theta. So, what we are doing here? So, I am start starting by choosing some initial values of the parameters A, B, and pi. And then I have to repeat the following steps until convergence. What is that? Firstly, determine what are the probable paths that X t minus 1 that t minus 1 object point I see the state i; at t-th, I see the state j and. So, on what is the probability of various state paths.

Now, once I have these will teach count the number of times. So, what is important expect a number of times, because we are only computing probabilities, we are not finding the actual paths. So, count the expected number of transitions a i j as well as the expected number of times various emissions b j y t m 8. So, using these state paths can you do that, yes, I know what is the probability of X t minus 1 is equal to i; X t is equal to j; I can use it to compute an expected value for a i j, and this is what we will be doing. So, we will compute a i j, similarly I already know the observation. So, I can also computes unexpected vanish for b j y t what is the probability that I see objection y t for the particular state j. So, once I computed this a i j and b j y t, I will again estimate my parameters theta using these computed values.

Now, I will go back in the loop, I have this theta, I will compute the probabilities of various state paths again compute a i j, b j y t and again compute theta and I repeat that until convergence. So, now so all these are interesting here like, firstly, once you are given some set of parameters theta it can be either the initial parameters or some intermediate parameters how do you actually compute different the probabilities of different paths, how do you actually do it. And for that we use a forward backward algorithm that is the main concept of this algorithm.

(Refer Slide Time: 08:26)



Forward-Backward Algorithm

Forward Procedure

$\alpha_i(t) = P(Y_1 = y_1, \ldots, Y_t = y_t, X_t = i|\theta)$ be the probability of seeing $y_1, \ldots, y_t$ and being in state i at time t. Found recursively using:

- $\alpha_i(1) = \pi_i b_i(y_1)$
- $\alpha_j(t+1) = b_j(y_{t+1}) \sum_{i=1}^{N} \alpha_i(t) a_{ij}$

Backward Procedure

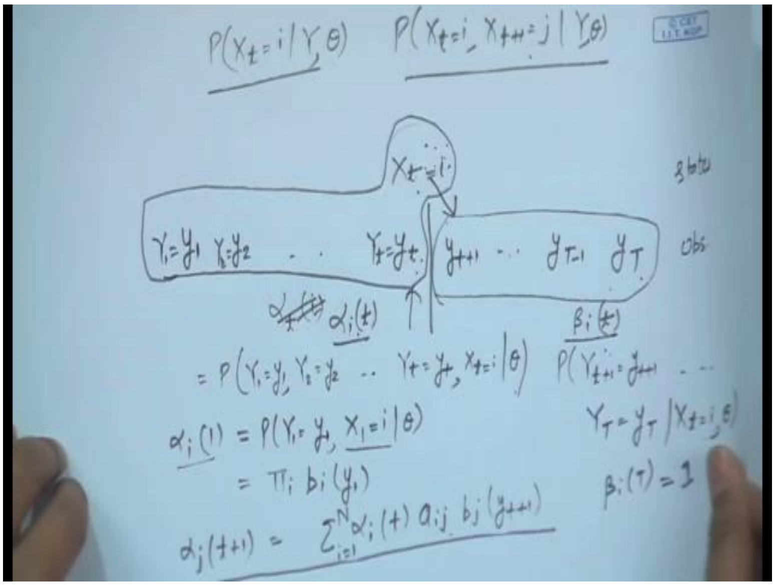
$\beta_i(t) = P(Y_{t+1} = y_{t+1}, \ldots, Y_T = y_T|X_t = i, \theta)$ be the probability of ending partial sequence $y_{t+1}, \ldots, y_T$ given starting state i at time t. $\beta_i(t)$ is computed recursively as:

- $\beta_i(T) = 1$
- $\beta_i(t) = \sum_{j=1}^{N} \beta_j(t+1) a_{ij} b_j(y_{t+1})$

Pawan Goyal (IIT Kharagpur) POS Tagging Week 4, Lecture 2 4 / 6

So, let us see what is this forward backward algorithm. So, in forward-back backward algorithm, we have a forward procedure and a backup procedure. So, so far for explaining let me just take show it on paper once.

(Refer Slide Time: 08:47)


$P(X_t=i \mid Y,\theta)$
$P(X_t=i, X_{t+1}=j \mid Y,\theta)$
$X_t=i$
states
$Y_1=y_1$ $Y_2=y_2$ . . $Y_t=y_t$
y_{t+1} .. y_{T-1} y_T
obs
$\alpha_i(t)$
$\beta_i(t)$
$= P(Y_1=y_1, Y_2=y_2 \,..\, Y_t=y_t, X_t=i \mid \theta)$
$P(Y_{t+1}=y_{t+1} \ldots Y_T=y_T \mid X_t=i,\theta)$
$\alpha_i(1) = P(Y_1=y_1, X_1=i \mid \theta)$
$= \pi_i\, b_i(y_1)$
$\beta_i(T)=1$
$\alpha_j(t+1) = \sum_{i=1}^{N} \alpha_i(t)\, a_{ij}\, b_j(y_{t+1})$

So, what is happening? I am observing this sequence Y 1, Y 2, small y T minus 1 y t, this is my observation. And correspondingly, so you can say this is Y 1 is equal to y 1 first observation Y 2 is equal to y 2 and so on. And correspondingly there are states, so this is my observation and then there are states. So, there will be some state x t this is y t observation, this is my state x t is equal to y it can be any of the possible states x t. Now, what do I do in this forward backward algorithm, I compute two different probabilities; one I called as the forward probability. This is the probability of observing y 1 to y t plus and x t is equal to i give my parameter theta. So, this is my alpha t for the ith the state. The probability that I am observing y 1 to y t and at t-th point the state is i given my parameters theta; Y 1 is equal to y 1, Y 2 is equal to y 2, Y t is equal to small y t and x t is equal to y given parameters theta. This is my forward probability.

Then I compute a backward probability. What is that? This is the probability of observing this given this state. So, this is my alpha i t, and this is beta i t that is probability that Y t plus 1 is small y t plus 1 up to capital Y t is y t given X t is equal to i and theta. So, these are two different probabilities that I stored. So, what do you see here, so I am taking just a state at a particular time t a particular state i. Now, why do you actually compute this alpha i t and beta i t, how does that help? Now I have these probabilities, I can simply multiply these alpha i t times beta i t to get the probability that I observe the sequence and the ith the state. And to actually compute the probability of saying that at tth time point the state was i, I can marginalize it by all the possible states that can happen at time t. So, I can multiply sigma i t, beta i t and divide by all the possible sigma j t alpha j t and beta j t to compute the probability of seeing state i at time point t and this is that is how this forward and backward probabilities help. So, you will see that in detail.

So, right now just the formulation. So, I am this higher the sequence, I divided in two parts for a given t. This is state i, I have the forward probability alpha i t and a backward probability beta i t; alpha i t is this particular sequence and the state, beta i t is the ending sequence given this is state; the previous state X t is equal to i. So, now so this is my alpha i t. So, now, how do I compute these values of alpha i t? I want to compute this for all possible i's and all possible values of t. So, this is a forward probability.

So, I have to start with the first point. So, how do I actually compute alpha i 1, what is that? So, alpha i 1 from this equation which probability of observing Y 1 is equal to y 1 and X 1 is i given theta. So, what is the probability of X 1 is equal to i given theta that is pi i; this is my pi i, and this probability would be emission from X 1 from the either state. So, this will be b i y 1. So, this is how I will compute the initial one alpha i 1 for all possible states i. Now, at some point of time, at some point t plus 1, how will I compute alpha j t plus one given the previous alpha i t, I have computed. So, alpha j t plus 1, so t plus 1 means the probability of seeing the sequence up to Y t plus 1 and the state X t plus 1 is j.

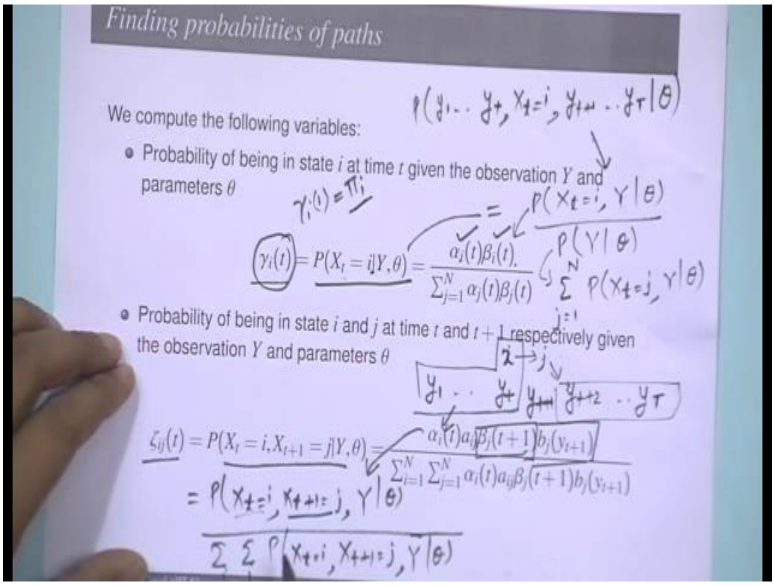
So, how can I compute that using the previous alphas? So, I have the previous alpha that gives me this. So, alpha i t at the previous time step I had some state, I have the transition from i to j, so times a i j and now I have the emission for the t plus 1th observation that is b j y t plus 1. And this I can obtain from any of the previous states. So, I have to sum over all the possible states. So, from all the previous states at the t-th time step, I can be any state i, I have the transition probabilities and the emission probability for t plus 1 and that is how I can compute sigma alpha j t plus 1 from all the alpha i t-th. So, this is how you can compute your alpha in a recursive manner starting from alpha i 1.

Now what will you do for your betas. So, this is a backward procedure. So, what is my beta i t this gives me the probability of ending the sequence with y t plus 1 to y t given the state i at time, this is my backward procedure. Now, this I again compute recursively in alpha i was starting from the first word I was computing alpha i t. In betas because the backward procedure, I will first compute beta i capital T, T is the end of the sequence and what will the beta i t, so beta i t would be probability of so given that previous the t-th point state is i having a sequence on t plus 1, but there is no word from t plus 1. So, this will always be. So, this is always be 1, this will be a null sequence. So, this is always one. Now, I have to compute it recursively for all other beta i at time point t. So, here I am going backwards.

So, you assume that you have computed all the beta j at time t plus 1 and you want to use that to compute beta i at time point t, how will you do that? So, this is very similar to what you did in the case of alpha. So, we can see this. So, this is beta j at time t plus 1 transition probabilities from i to j and the emission probabilities probability for the t plus 1th observation, and you sum over all the possible states at times j so very, very similar to what you did in the case of forward procedure. So, that is what you will do in the backward procedure for computing beta h in a recursive manner. So, now, you have a way that given the set of parameters minus theta you can compute all this alpha and beta. So, all the alpha i t is and beta i t you can compute.

Now, remember what is the next step in the algorithm, you want to find out what are the best possible paths or what is the probability of various paths that is what we intended by doing this alpha and beta, let us see how exactly we can compute these probabilities of various paths. So, I want to compute the probabilities of say probability X is equal to y given my observation and parameters theta, also I want to compute X t is equal to y and X t plus 1 is equal to j given y and theta, I want to compute both of this. So, if I can compute both of this, I can compute all the parameters, I can complete my transition probabilities using that, I can compute my initial probabilities using that by taking t is equal to 1 and you can also compute my ambition probabilities. So, now, my problem is once I computed alphas and betas can I compute these probabilities, so that is what we will see next.

(Refer Slide Time: 19:42)



Finding probabilities of paths

$P(y_1 \dots y_t, X_t = i, y_{t+1} \dots y_T | \theta)$

We compute the following variables:

- Probability of being in state i at time t given the observation Y and parameters θ

$\gamma_i(t) = \pi_i$

$$\gamma_i(t) = P(X_t = i | Y, \theta) = \frac{\alpha_i(t)\beta_i(t)}{\sum_{j=1}^{N} \alpha_j(t)\beta_j(t)} = \frac{P(X_t = i, Y | \theta)}{P(Y | \theta)}$$

$$\sum_{j=1}^{N} P(X_t = j, Y | \theta)$$

- Probability of being in state i and j at time t and $t+1$ respectively given the observation Y and parameters θ

$i \rightarrow j$

$y_1 \dots y_t \; y_{t+1} \; y_{t+2} \dots y_T$

$$\zeta_{ij}(t) = P(X_t = i, X_{t+1} = j | Y, \theta) = \frac{\alpha_i(t) a_{ij} \beta_j(t+1) b_j(y_{t+1})}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i(t) a_{ij} \beta_j(t+1) b_j(y_{t+1})}$$

$$= \frac{P(X_t = i, X_{t+1} = j, Y | \theta)}{\sum \sum P(X_t = i, X_{t+1} = j, Y | \theta)}$$

So, I am giving these some names. So, firstly, I have this gamma i t that gives me the probability that at t-th point the state is i given my observations y and theta parameters. And similarly, I have zeta i j t that is probability that at t-th point, I have state i; t plus 1 point, I have state j given y and theta. So, gamma i t is written terms of multiplication of alpha i t and beta i t divided by summation over all possible sigma z alpha j t and beta j t over all possible change. So, now how do we actually come up with this equation of gamma i t and zeta i j t, let us look at it.

So, here I want to compute probability x t is equal to y given y n theta. So, now, I will computed using probability X t is equal to y, Y given theta divided by probability Y given theta, this is simple conditions. So, if you forgot that theta, this is nothing but X t is equal to i given Y, X t is equal to y, Y divided by probability y, this simple the condition probability rule. Now, this one I am writing as summation j is equal to 1 over N, j is equal to 1 to N probability X t is equal to j, y given theta. So, now, you can see the symmetry in this equation.

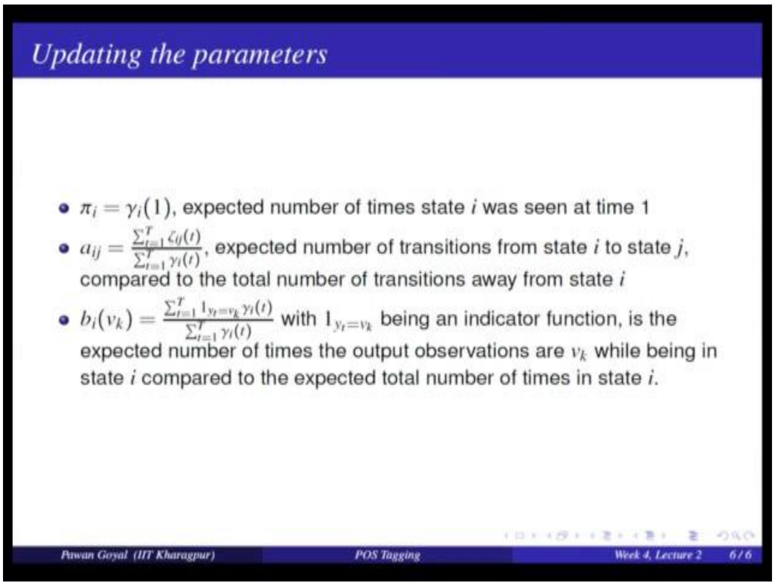
Now, what I have to show that alpha i t times beta i t is probability X t is equal to y, Y given theta and that is very easy, because what is alpha i t this is probability y 1 to y t plus x t is equal to i. What is beta i t probability of y t plus 1 to y capital T and now X t is equal to y is already given for beta, and this is given my parameters theta and this is what we are seeing here. This is my alpha i t times beta i t and this is marginalizing over all possible states are time by t. So, this will be sigma, j is equal to 1 to N, alpha j t beta j t. So, this gives me the equation for gamma i t.

Similarly, you can see for zeta i j t that is for probability of X t is equal to i X t plus 1 is equal to j Y given Y n theta. So, in the same way, you can write it as probability X t is equal to i X t plus 1 is equal to j, Y given theta divided by probability Y given theta that will marginalize over all possible i and j. And you will say probability X t is equal to i x t plus 1 is equal to j y given theta for all possible values of phi and j.

Now, what we have to show that this is actually this formula now. So, what is this? So, you are given state at time point t and time point t plus 1. So, what you are actually doing is that using the forward procedure from y 1 to y t, and you will also get the state i at time point t. And you will use the backward procedure because you are given the state at time t plus 1 state j, backward procedure for from y t plus 2 up to y capital T. So, this is what is captured in alpha i t, and this is captured condition on j in beta j t plus 1. Then you will compute the probability of transition from i to j that is a i j and y t plus 1 given j that is your b j y t plus 1. So, this equation is nothing but probability of X t is equal to i, X t plus 1 is equal to j Y given theta, this whole sequence is there and these two states are also there. And similarly you can see for here this is over all the possible values of i and j.

So, I hope this is clear that once I have computed all the possible alpha and beta and the previous parameters of there, we can compute this gamma i t and zeta i j t-th that are probability of state at a given time and a sequence i and j at time t and t plus 1, this we can compute using alpha and beta. Now, coming to the last part that is once I computed some probability is over by possible state paths how do I estimate my parameters again, I have started with some initial parameters computed this, now I want to re-estimate my parameters theta. So, how do I do that?

(Refer Slide Time: 26:14)



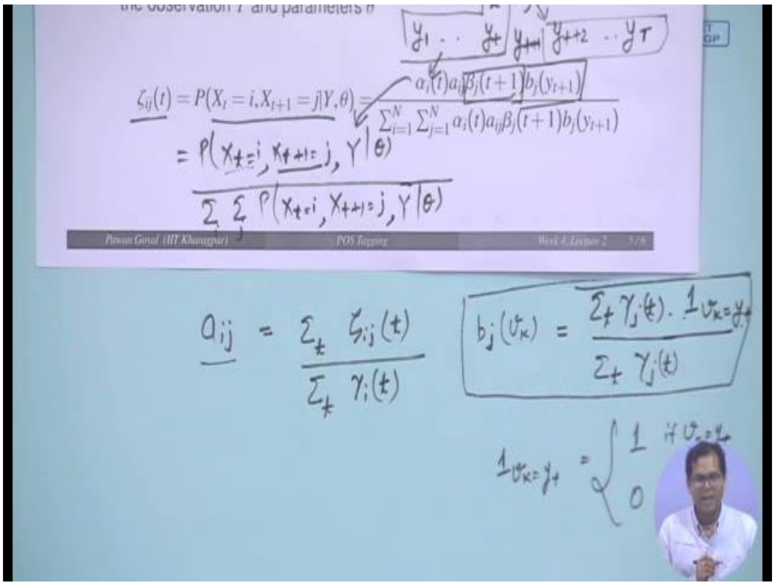
Updating the parameters

- $\pi_i = \gamma_i(1)$, expected number of times state i was seen at time 1
- $a_{ij} = \frac{\sum_{t=1}^{T} \zeta_{ij}(t)}{\sum_{t=1}^{T} \gamma_i(t)}$, expected number of transitions from state i to state j, compared to the total number of transitions away from state i
- $b_i(v_k) = \frac{\sum_{t=1}^{T} 1_{y_t=v_k} \gamma_i(t)}{\sum_{t=1}^{T} \gamma_i(t)}$ with $1_{y_t=v_k}$ being an indicator function, is the expected number of times the output observations are v_k while being in state i compared to the expected total number of times in state i.

Pawan Goyal (IIT Kharagpur) POS Tagging Week 4, Lecture 2 6 / 6

So, now I have to estimate my pi i, a i j and b i v k. So, how do I estimate this pi i is probability of seeing the state i at time one. So, now, from these parameters which one can give me this value probability that x 1 is equal to i; this I can get from gamma i 1 that is that can give me my pi i is probability that at X 1 is i. So, this is to estimate my pi i. Next thing I have to estimate is transition probabilities probability of going from state i to state j. Now I have computed this.

(Refer Slide Time: 27:18)



the observation Y and parameters θ

$y_1 \dots y_t \; y_{t+1} \; y_{t+2} \dots y_T$

$$\zeta_{ij}(t) = P(X_t = i, X_{t+1} = j | Y, \theta) = \frac{\alpha_i(t) a_{ij} \beta_j(t+1) b_j(y_{t+1})}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i(t) a_{ij} \beta_j(t+1) b_j(y_{t+1})}$$

$$= \frac{P(X_t = i, X_{t+1} = j, Y | \theta)}{\sum_i \sum_j P(X_t = i, X_{t+1} = j, Y | \theta)}$$

Pawan Goyal (IIT Kharagpur) POS Tagging Week 4, Lecture 2 5 / 6

$$a_{ij} = \frac{\sum_t \zeta_{ij}(t)}{\sum_t \gamma_i(t)}$$

$$\boxed{b_j(v_k) = \frac{\sum_t \gamma_j(t) . 1_{v_k = y_t}}{\sum_t \gamma_j(t)}}$$

$$1_{v_k = y_t} = \begin{cases} 1 & \text{if } v_k = y_t \\ 0 & \end{cases}$$

So, how do I compute my a i j, a i j would be at all possible times how many times you are having the sequence i n j. So, this will be summation over t zeta i j t, zeta i j t is i followed by j divide by over all the times when you are seeing this t. So, when you are seeing the state i and this you can compute using your gamma i that is summation over t gamma i t, and that will give me my a i j. All the possible cases where all the probabilities of i and j occurring together at any point divide by all the probabilities of i occurring anywhere, this is my a i j.

And similarly, how can you compute your b j say v k that is at state j you are emitting v k how can you estimate that. So, for that you have to find out whenever you are in state j is the observation actually v k, how many times you are in state j by the observation is we can divide by number of times you are in state j. So, this is so here the denominator is easy that is number of times you are, so this b j v k number of times you are in state j this is gamma j t numerator number of times or the probability that you are in state j times the observation is actually v k. Now, so at time t, the observation is small y t. So, I can use some sort of a simple notation or indicator function 1 v k is equal to y t, so indicator function. What is indicator function? So, 1 v k is equal to y t will be 1 if v k is actually y t and 0 otherwise this will be my indicator function.

So, now, whenever if v k is y t whenever I observe y t, I will add it, otherwise it will be 0. And this is this will give me the emission probabilities. So, I could estimate my pi i, a i j as well as b j v k using this zeta i j and gamma i which I computed using my forward backward probabilities. And this is one complete pass of this algorithm. Now, I have these parameters, I will plug these again to compute alphas and betas; again compute my gamma and zetas and again estimate the parameters until it converges, and this is the Baum Welch algorithm for learning the parameters of HMM when the label data set is not available. So, this gives you a very nice handle on HMMs.

So, you can apply HMMs, when you have the label data set very easily by learning the parameters directly from memories, maximum likelihood estimate, and data beta decoding. But even if you get a new problem or a new language furtherer you have to compute part-of-speech tag. But you do not have the labels you can use it this Baum Welch algorithm to estimate the parameters of a HMM, and then each bit have de coding to actually given a new sentence find at part of speech tag sequence. So, you can use that for any other sequence label task.

So, this completes our discussions on HMMs that were one of the very popular models for this part of speech tagging in many other sequence labeling task, but they have certain limitations and that is what we will discuss next then that what are these limitations, and how do some other models. So, we will go to now some discriminative classifiers like maximum entropy models they get rid of these limitations. So, this is for HMMs. Let us look at maximum entropy models in the next lecture.

Thank you.