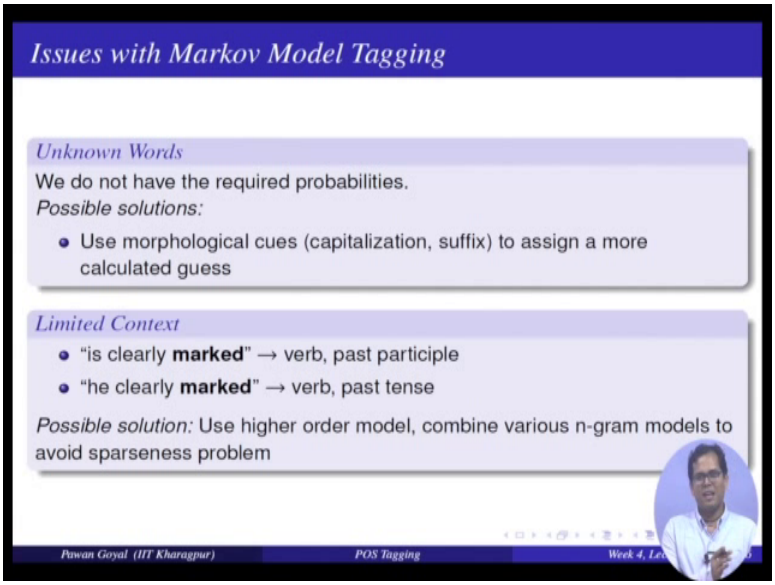
Natural Language Processing
Prof. Pawan Goyal
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture - 18
Maximum Entropy Models – I

Welcome back for the 3rd module of this week. So, in the last week we have; in the last module we had finished the discussions on hidden Markov models how do we learn the parameters and how do we use that for estimating the part of speech tags sequences for a given sentence. Now hidden Markov models for generative models and if you remember we discussed this distinction between generative model and discriminative model. So, we will see a discriminative model; maximum entropy model for this sequence labeling task, but before that let me start with some sort of motivation and why do we move from HMMs to maximum entropy model what are the limitations of HMM.

(Refer Slide Time: 01:00)



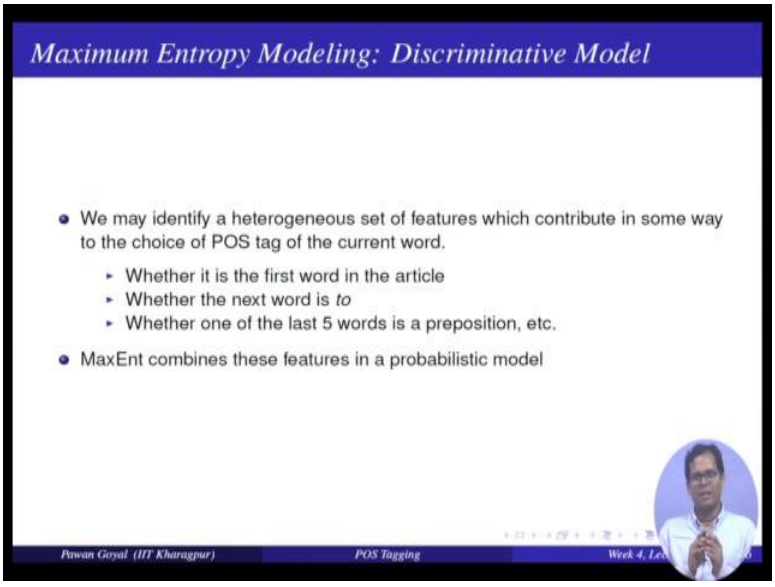
Then problem with the HMMs also, this is known as the Markov model tagging is handling unknown words suppose at run time, you get some unknown word that was not there in the corpus because the emission probabilities will not be known, there is no where you can come to you can actually use the Viterbi decoding. So, what do you do in that case? So, you might want to use some sort of morphological cues, suppose the word we word is ending with e d, you might think that over this might be a past tense of a

word that we did not encounter in the corpus or suppose you are finding capitalization in the middle of the set of the sentence you might think that this might be a proper name.

But how do we actually include that in hidden Markov models second. So, we use the first order Markov assumption where each tag dependent on the previous only on the previous tag. So, this does not take care of certain situations for example, here I have 2 sentences is clearly marked was such he clearly marked. So, what you seeing here marked in one case is a past participle in the first case second case it is a verb in past tense, but just the previous context is not sufficient to tell this. So, because in HMMs, we are only using the previous context this might not be sufficient to handle this particular case.

How will you take care of this issue? So, you might want to move to a higher order HMM say it second orders. So, you can use a high order model, but this is just one some limitations suppose you want to use what whether this is the first word in the sentence or you want to you want use any arbitrary think like whether, there is, a verb ending with e d previous to this word you want to use this information. There is no way you can use that in the case of achievements so that is where we move from achievements to my explanatory model that allows very very heterogeneous set of features. So, you can use a lot of different of your wishes from the data for your model.

(Refer Slide Time: 03:35)

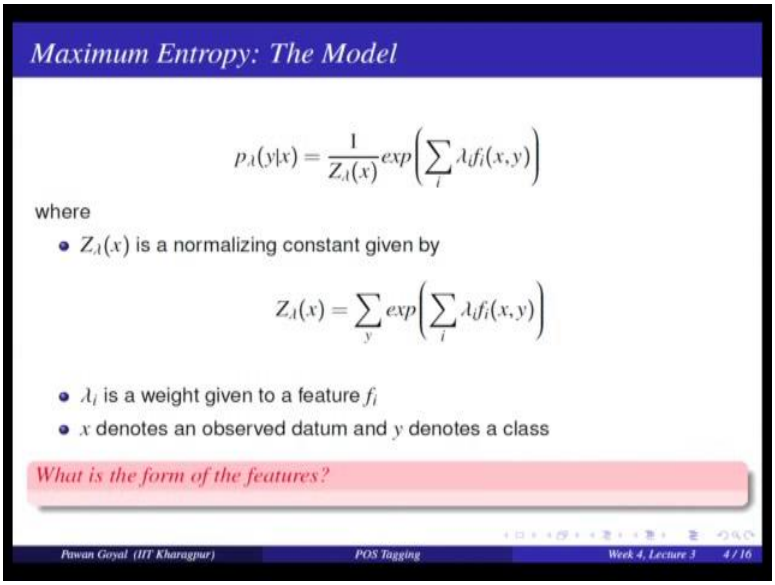


What is the important distinction? So, now, in the maximum entropy model so they are discriminative as compared to generative models which are as means you are allowed to identify very very heterogeneous set of features not only the tag given, the previous tag or the word given, the tag you can use, any sort of features which might which you think might contribute to the choice of part of speech tags.

For example, whether this is the first word of the article, this is something that you can use as a feature in maximum entropy model or whether the next word is 'to', you can use the next word as a feature whether one of the 5 previous 5 words is a preposition this is again something that you can use as a feature in maximum entropy model, but here HMM cannot make each of this.

What maximum model does it combines all this various heterogeneous features in probabilistic framework to be able to come up with the actual part of speech tag sequence for a given sentence.

(Refer Slide Time: 04:47)



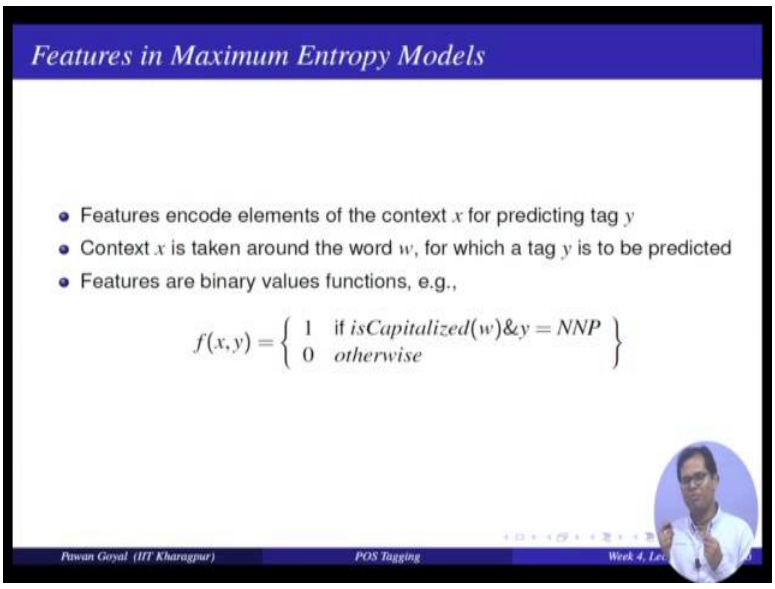
This is the maximum entropy model. So, this is how you write the probability of so y is the class that is one of the probability for any of the tags, given x is my observation it can include anything in my in my context it can load the current word, previous word, next word, anything in the sentence you can so you can define your context while you are dealing with maximum entropy model, you are saying that I can use all this information in my features, given this I want to compute the probability of my class and here

discriminative model. So, I can directly compute the probability of class given the observation. So, I can compete y given x and how do I compute that? It is simply x is a logging in model. So, summation over lambda i and f i x y, now f i x y are my various features that I have identified over my observation x and my class y and lambda is the weight given to the feature f i. So, each feature f i is given a weight lambda.

I summed over all the possible lambda is f i's and taking the exponent and then I do some sort of normalization. So, that probability y given x adds up to 1 for all the possible classes y, this is a normalization constant. So, this is the very very simple form of maximum entropy model. So, so let me tell again so z x condition on the parameters lambda is normalization constant. So, this is nothing but this exponent you sum over all the possible class x y for all the possible class x y you compute this and add this. So, once you do that you are ensuring that probability y given x will added to 1 for all possible values and lambdas are my features the weights given to my different features.

Now x denotes my observed data and y and y denotes my class now, what is interesting is that? What is the form that the features in maximum entry model can take as opposed to something like HMM, what are the features? So, what is interesting here is you are seeing f is a function of x and y both in the observation in the class. So, I am defining my feature on both of this and we are all binary features.

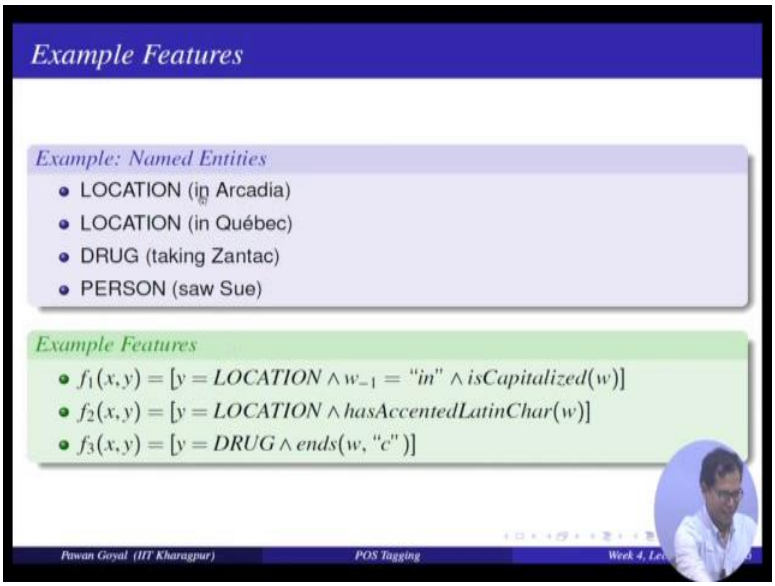
(Refer Slide Time: 07:36)



Let us see, what are my features? So, in maximum entropy model, my feature send code elements from the context x and the tag y now context x is now this is something that you must be careful about context is not my current word w, but is taken around the word w i can define my context it can be the sentence that contains w it can be it is previous 3 words next 3 words it is something that can be defined on the on the current word and y is the tag that I have, I am predicting

Now, what are my features these are binary valued functions over x and y and this is like f x y is 1 if a certain conditions condition hold on x and y and 0 otherwise and here is one example what can be some conditions like I am having the current word w if the word w is capitalized and my tag y is NNP then my feature f x y will take a value of 1 otherwise it will take a value of 0 that is one sort of feature. Similarly you can define any arbitrary set of features that we have said, it can be f x y is equal to 1 if one of the previous 5 words is as a tag of preposition and the current tag is save up this is one sort of a 0 otherwise. So, like that you can define an. So, you have a lot of freedom here in defining and choosing your features and once you complete all these features the model learns the weights for all these feature features the weights lambdas and then this is very very simple function to compute the probability of plus given any observation.

(Refer Slide Time: 09:18)



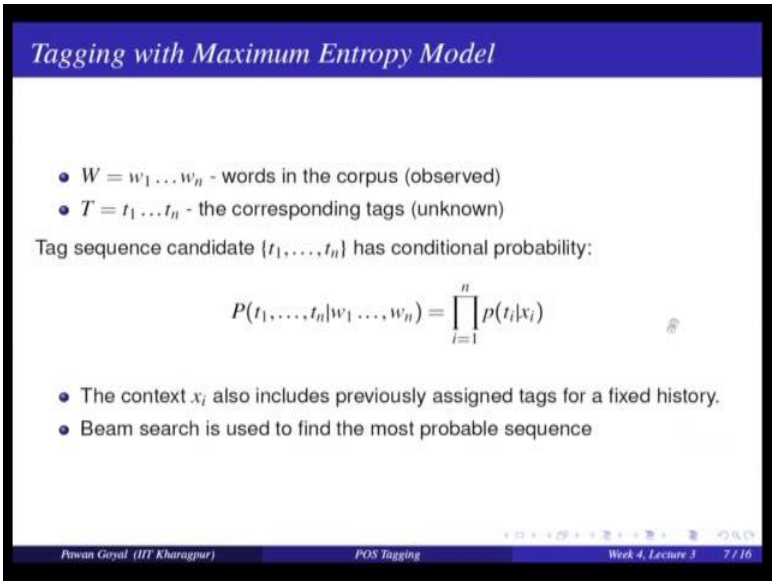
Let us see some other examples for the features. So, this is again for any generic sequence labeling task. So, here it is for finding named entities. So, it can be here like

location drug and person 3 different named entities and some examples are given. Like in arcadia, arcadia is a location in Quebec location taking Zantac, Zantac is drug and saw sauce you. So, you see you as a person and now you are defining certain features. So, let us see certain features like - the first feature is my tag is location. So, my class y is location previous word that is on x previous word is in and my current word is capitalized.

What are all the cases we have? This feature will be one, it will be one here in arcadia previous what is in and it is capitalized and the tag is location, it will also be one for second case is capitalized previous word is in and the tag is location. Next feature, y is location and w has an extended leg in Latin character. So, you can see that you are using a very very arbitrary set of features, but possibly you have some hypothesis that this features will be helpful for this task operating the predicting the tag for this word. So, this is this feature. So, for which of the case it will be? One, it will be one for the second case it has an extended Latin character and the tag is location.

Third feature, y is a drug and the word ends with c and there will be one for taking Zantac yes the word W ends to the c. So, you can see some examples of what kind of features you can take, but important thing is to know the intuition that you can choose any feature around the word w and that involves also the current class.

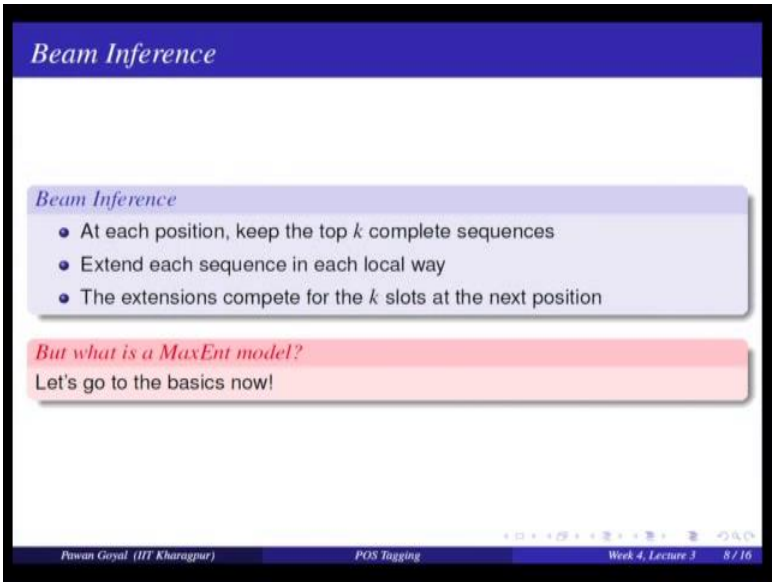
(Refer Slide Time: 11:26)



How do we? So, now, once I have all these features suppose the learning has taken place I know all the parameters lambdas, how do I go about tagging my sentence with that x? So, I have the words in my corpus w 1 to w n and I want to find out the tags document tags. So, what I will do? I want to find out tags sequence t 1 to t n. Any sequence will have this condition, probability right. So t i conditioned on x i for i is equal to 1 to I, now this is x i not w i, x i is my context around the word w i and this is independent.

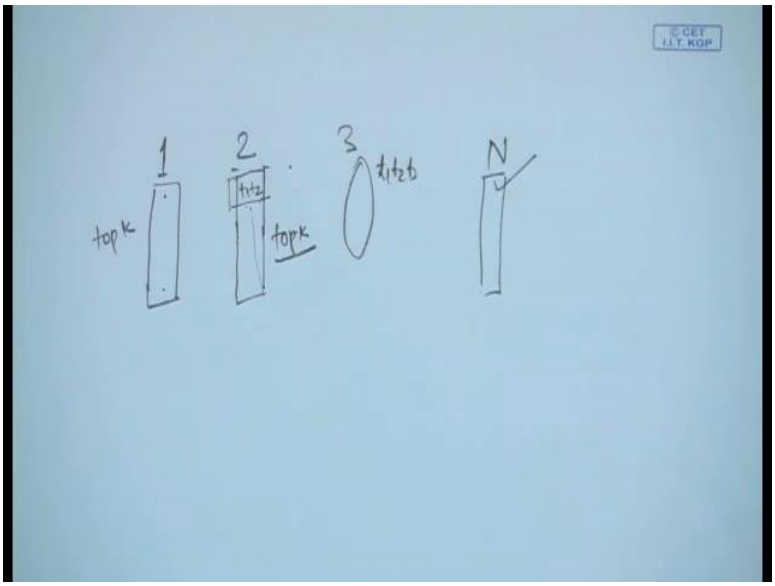
So, I can use this model to find out what is the optimal tag sequence, I can do it independently for each word and then I will find out tag sequence. So as we have said my context x i includes, may also include some previously assigned tags for a fixed history and I will be using a beam search algorithm for finding the optimal or the most probable sequence this is something we will take up in the next lecture in detail that how do we actually come go about doing beam search in maximum entropy model.

(Refer Slide Time: 12:51)



Yeah just a brief idea, here in beef search beam search what you do actually. So, you have positions in the sentence a starting from one to the end. So, at each position we will keep only the top k possible sequences and next time starting from this k; k complete sequences you find out what are the next k based sequences.

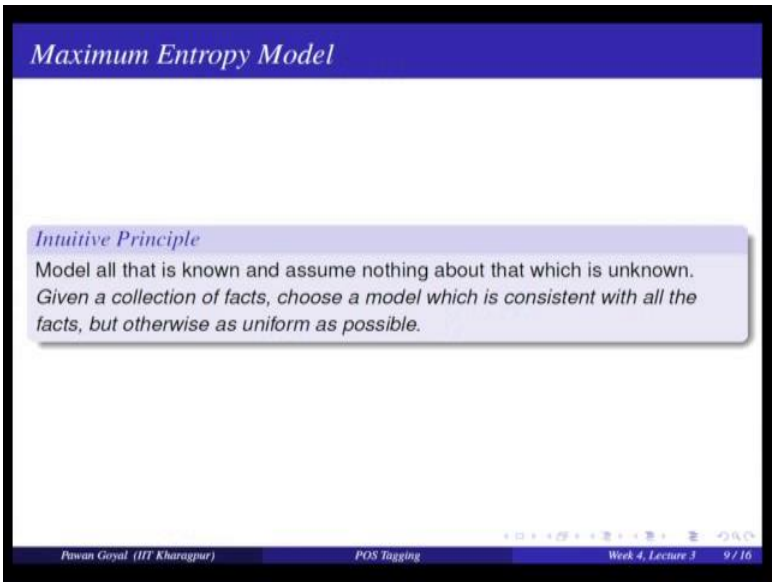
(Refer Slide Time: 13:28)



Let me just quickly say, so I have from, suppose my sentence as N words so, what is the idea? Suppose my size beam size is k, at point 1 I will keep top k sequences. So, at point 1, I will have only 1 part of speech tags so; that means, I will keep top k part of speech tags at it is position 1, now when I go to position 2, now I will have a sequence, now 2 might have some other possibilities of part of speech tags. So, what I will do? I will find out the probability of each of this k with all part of speech tags here and you store top k sequences again. So, tag 1, tag 2 some case top k sequences again at 3, I have some possibilities top k times these probabilities this possibilities again you store only top k out of this and you keep on doing that until you are here.

Here you will to choose the best one and that will give you the complete sequence. So, important here is that at each point you are is storing top k full sequences which starting from the initial point. So, here you will find some t 1, t 2, you will find some t 1, t 2, t 3 and so on top k at it is time. So, we will look at this in detail in the next lecture. So, for now let us look at the basics of maximum entropy model. So, what is a maximum entropy model?

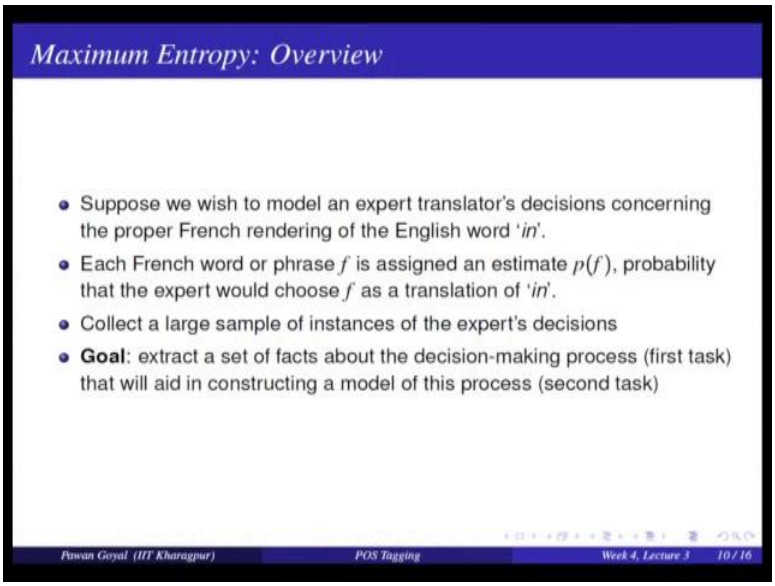
(Refer Slide Time: 15:02)



Maximum entropy model have this particular principle that is you are given the observation and you have certain hypothesis. So, you take what is given to you, but do not make any additional assumptions. So, that is take the model that satisfies all the questions that you having, but does not take any other assumptions.

What does that mean? So, you are given a collection of facts yes and you want to take a model that is satisfying all these facts or you are saying that is consistent with all these facts, but otherwise it is not making any further assumption that is after being after satisfying all these constraints it is as uniform as possible or in other words for all the possible models that satisfy a given set of constraints choose the one that is the most uniform of all.

(Refer Slide Time: 16:09)

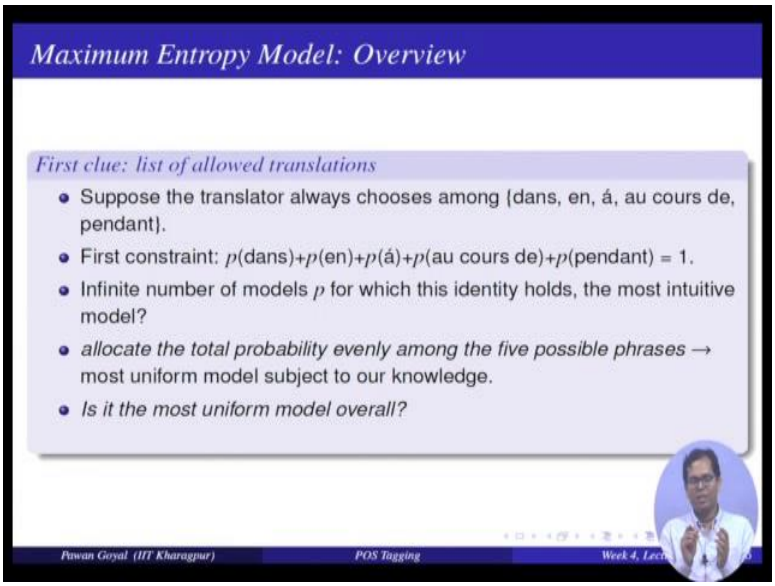


Now, what do I mean by this? Let us take a simple example. So, here I have an, this I have a simple example I have an English word in and I want to make some transition system that tries to mimic some. So, it is translating from English to French and I have some data already with me from some actual translator who translated some sentences from English to French.

Now, so I want to model this expert translator's decision that how do I translate this word in to any French word? Now how do I start? So, suppose you have a vocabulary of all possible French phrases and you want to estimate probability p f. So, that is, what is the probability that that French phrase will be a translation of the English word in; so you want to compute that for all the possible French phrases, they might be say, millions of French phrases, so what you will do? We will start by collecting a large sample of instances of expert decisions.

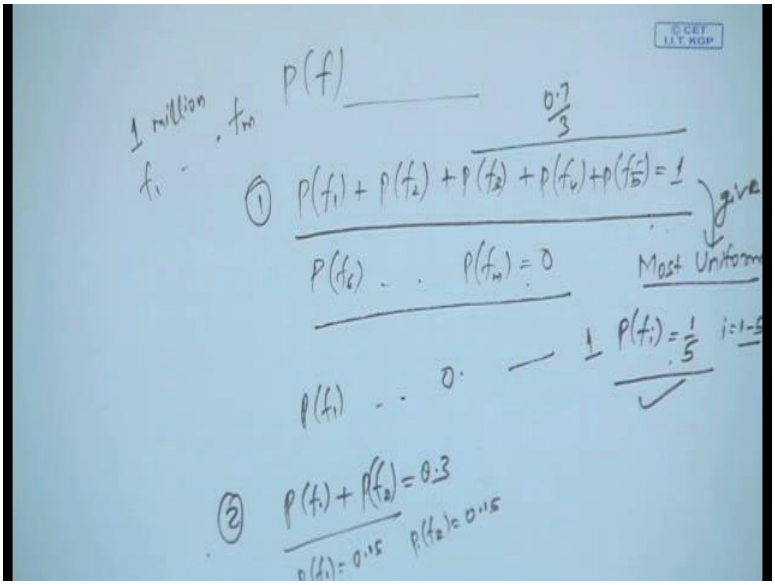
Now, from this facts you will try from this sample you will try to observe some facts for example, so the goal would be extract some facts from the sample of instances and then use these facts inside your model so. Firstly, you have to extract the facts. So, in our scenario this will be equivalent to finding out the features that is my facts and once I found these features or the facts I use them rightly in my model.

(Refer Slide Time: 18:06)



Let us see what are the different sort of facts I can find out for example, this can be my first clue. So, I find that in all the samples I have seen the translator always translates in into one of these 5 phrases, dans, en, a, au cours de, pendant. So, they are 5 different French phrases where the translator is translating. So, now, this is helpful to explain what is my maximum entropy model.

(Refer Slide Time: 18:40)



Now, I want to find out the probability of the probability distribution over French phrases all. So, suppose I have say 1 million French phrases, so, f 1 to f million. So, I

have 1 million French phrases, I want to compute this probability distribution, I want to find out the probability distribution. Given this constraint that is the translator always chooses among from one of these 5 phases. So, what is my constraint? This would be so let me call these f 1, f 2, f 3, f 4 or say dans and also let me call these probability say f 1 plus probability f 2 plus probability f 3 plus probability f 4 plus probability f 5 is equal to so, this is my constraint.

Now, given this constraint I have to choose 1 model. So, suppose I have only this much information from the data now what is the model that I should choose among all the possibilities now? Firstly, do you think there are infinite number of possibilities in which I can choose my model that satisfy this constraint. So, of course, probability for f 6 to probability f million will be 0 because if any if any of this is non 0 this constraint will not be satisfied.

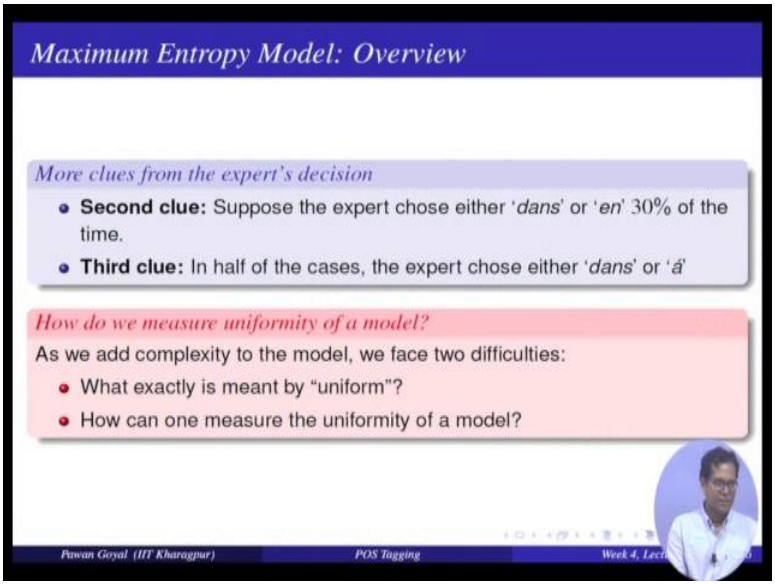
Yes, so, because all the probabilities need to add up to one so; that means, all this has is this is anyways easy, but what are all the models that will satisfy this constraint. So, you can see there are infinite solutions for this equation. So, you can choose p f 1 any number from 0 to 2 1 and you can always find some f 2, f 3, f 4, f 5 that will satisfy and same goes for f 2 and so on. So, there are infinite models that satisfy this constraint.

Now, what maximum entropy model does among all the models that satisfy this constraint it takes the one that is most uniform, now among all these models what is the most uniform model. So, that is you are not making any dimensions given this constraint take the model that is most uniform. So, it is easy to see in this case that the most uniform model will be one that gives P f i is equal to 1 by 5 for i is equal to 1 to 5, for all these 5 phrases gives the probability 1 by 5 and everything else it gives a probability 0. So, this is the most uniform model given this constraint. So, allocate the total probability evenly among all these 5 phrases that is the most uniform model subject to our knowledge.

Now is this the most uniform model overall, suppose I did not have this constraint what was the most uniform model as such, the most uniform model will be that gives equal probability to each of the million French phrases. So, it is gives the probability of 1 by million to each of these that is a most uniform model. So, the model that I have obtained which is giving probability 1 by 5 to each of these 5, it is not the most uniform model, it

is most uniform given this constraint given this constraint this is the most uniform model and this is the idea.

(Refer Slide Time: 22:31)



I will take some facts some observations given the observations I will find out what is the most uniform model.

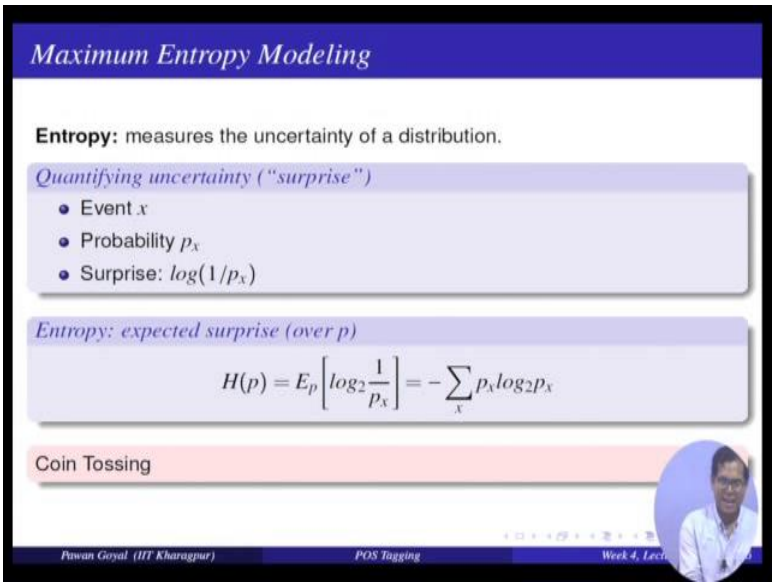
Let us take; let us go further. So, this was easy in this case. So, where we had 5 phases only knows i got a second clue that the expert chose either dans or en 30 percent of the time. So, we were supposed f 1 and f 2. So, now, I have, this is my constraint one. So, now, i have constraint 2 P f 1 plus P f 2 is equal to 0.3. Now given these 2 constraints, what would be the most uniform model? So, if you just take a quick look you can see that the most uniform model will be P f 1 is equal to 0.15 P f 2 is equal to 0.15 and f 3 f of 5 f 5 which will get 0.7 divided by 3 and that will set both the constraint this will be the most uniform model. So, this is not as uniform as this model. So, as you are getting more and more observations, you are going far from the most uniform models, but you are coming up with some other models that I as uniform as possible, but satisfying these constraints.

Now, what happens if you get a third clue like in half of the case; cases expert shows either dans or an. So, that is P f 2 plus P f 3 is equal to 0.5 that is 3, you know immediately you will see that it becomes difficult to just look at this equation and try to solve for what is the distribution that is most uniform and that is why we need to

understand the concept of maximum entropy. So, I want to find out the model that is most uniform doing the constraints that is equivalent to finding the model that is having the maximum entropy among model.

Now, yes how do we? So, as we get more and more clues it becomes difficult to see which is the most uniform model. So, for that you have to understand how do we measure the uniformity of a model and that is where we use the idea of entropy.

(Refer Slide Time: 24:49)



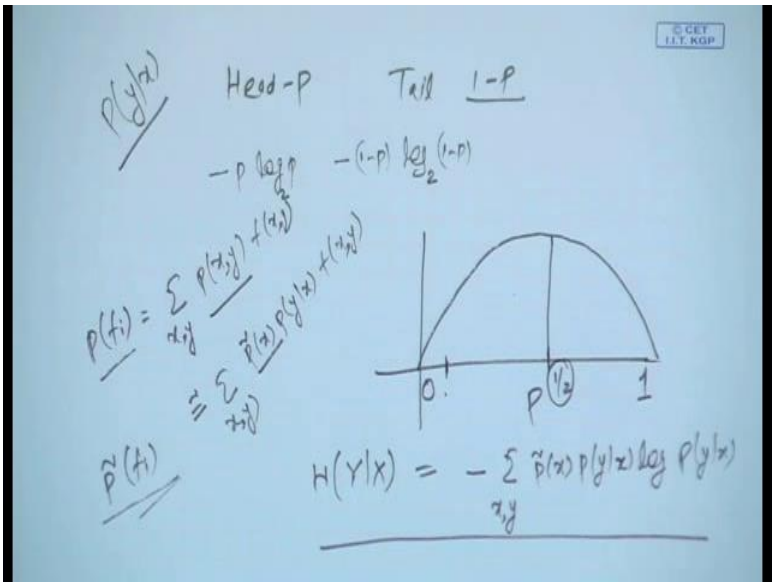
So, some of you might have already come across this term of entropy. So, what are entropy measures? What is the uncertainty of distribution? So, going to the basics, so suppose I have an event x that occurs to the probability of P x.

Entropy measures what is the amount of surprise in this event. So, if it is very very surprising then the entropy will be high, if it is like very obvious then entropy will be low so, for an event when it is more surprising when the probability is low, probability is high and the probability hi high it is like obvious. So, then the entropy will be low, but when the probability when the; so the interview will be low, but when the probability is low. So, there is a lot of surprise in that even if that happens then the probability entropy is high. So, that is why entropy is log of 1 by probability.

The probability is low entropy will be high. So, for a distribution how do we compute the entropy for a distribution? So, I just take the expected value of surprise that is summation

over p x log 1 by P x. So, I take the expected surprise over all the possible values of P. So, this is a very very a standard formula for entropy minus summation P x log P x now. So, here we are talking about distributions in the case of distributions when will be the entropy maximum. So, if you look at this formula.

(Refer Slide Time: 26:58)



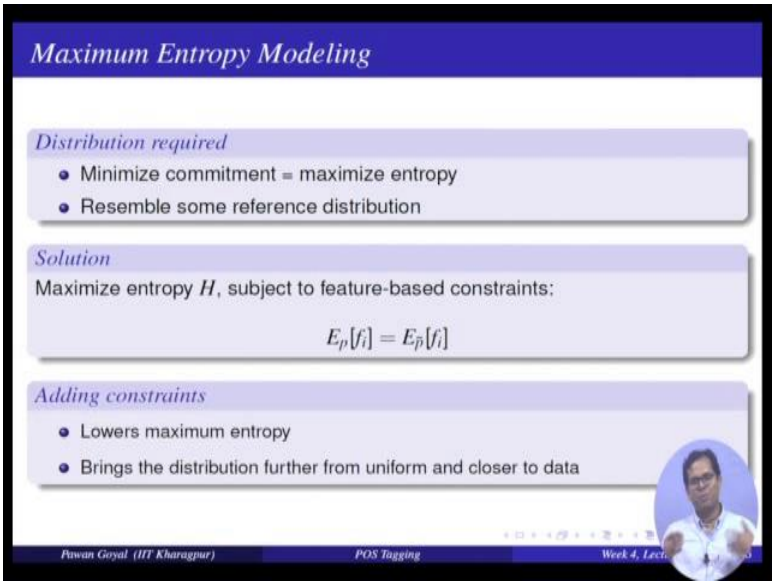
Let us take 1 simple example of coin tossing. So, I am tossing a coin. So, it can be head or tail let us say that head has a probability of P and tail has a probability of 1 minus P. So, what is the entropy of this model? It will be minus P log P minus 1 minus P log 1 minus P and if you actually. So, these are all based on too, if you actually go on plotting this as a function of P from 0 to 1 you will find that the entropy it starts increasing it becomes the highest at P is equal to half and it starts decreasing afterwards, now can we intuitively understand this?

What we are saying when both head and tail have equal probability then the entropy of them of my model is the highest, what is entropy? It measures how much is the surprise. So, let us take a point like this. So, at this point what will happen? So, head has a probability very low probability, but tail as a very high probability. So, you more or less no it will probably tail when you toss a coin it is probable most probably be a tail. So, there is not, there is not much more surprises that are coming, but when both head and tail have equal probability then you are very very confused you do not know whether the it will be head or it will be tail. So, the amount of so, expected surprise that you will get

is the maximum. So, this is the idea. So, this is a distribution over only this 2 variables P and this P N Q are you can say this have P N 1 minus P.

But even if you take any n number of variables the reservation will be having the maximum entropy wave whenever it is the most uniform. So, this gives such the intuition that why we chose the most uniform model for getting the maximum entropy.

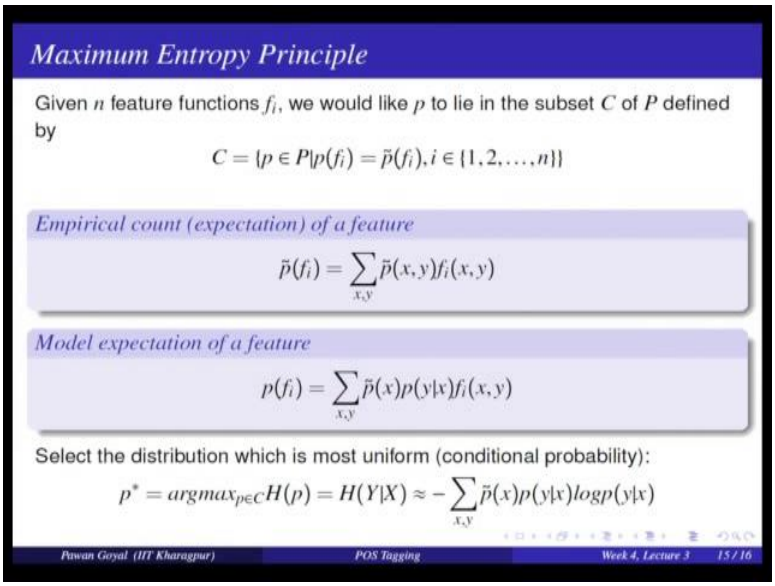
(Refer Slide Time: 29:06)



Now, coming back to a maximum entropy model, so we want to, among all the possible distributions we want to choose a model that maximizes by entropy.

Now, so I want to maximize entropy subject to certain constraints. So, we will have certain constraints from the data and these constraints are nothing, but my features. So, I will select some features and I will make sure that the ambiguous probability from the data that I am finding for the features is similar to the probability that my model is giving that and subject to this constraint I will maximize my problem my entropy. As you also seen as we keep on adding constraints one by one, the entropy of my model decreases, it comes close to my data initially might start with the most uniform model that is giving the equal probability to everything, but as you keep on adding constraint one by one you come close you go far away from the most uniform model. So, you decrease your entropy and you come closer to data. So, this is the idea. So, give us some constraints we want to come as close as possible to the data, but otherwise we want to obtain the most uniform model in terms of maximum entropy.

(Refer Slide Time: 30:36)



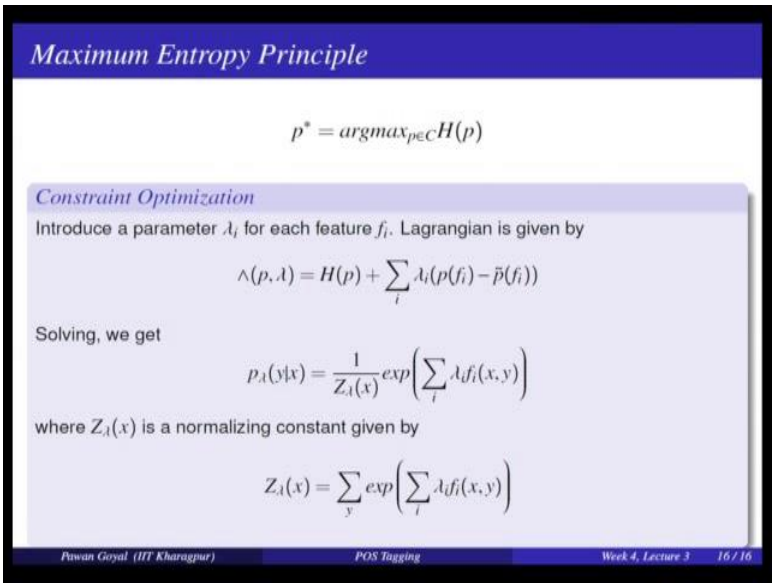
Coming close the data in this case means finding the empirical probabilities of these features and I am trying to make it trying the model probabilities to be close to this. So, this is my maximum entropy principle I am given some n different feature functions f 1 to f n, I want my probably distribution p to lie in the subset in the subset of all the possible distributions that satisfy this constraint. So, all the probability distributions that satisfy this constraint among those I want to choose the one that is having the maximum entropy. So, what do I mean by satisfying this constraint. So, what is the empirical count of or expectation of a feature? So, empirical expectation means how many times this features actually this x y actually occurred and f i x y was 1 in my data. So, P hat is my empirical probability.

Now, what will be my model probability? So, it will be so model probability P f i, it will be summation x y P x y, f x y now this P x y because we are computing finally, P y given x I can write as summation x y p x p y given x f x y and because we are not computing P s n l model I can probably approximate it with the empirical count of x. So, this gives me the model probability of the feature. So, I want to make it closer to the P tilde f i that is the empirical probability of the feature.

I have the empirical count model count and I want to select the distribution that is most uniform subject to this constraint. So, I am computing this probability distribution. So, what will be the maximum entropy model that is the one that gives me the high entropy

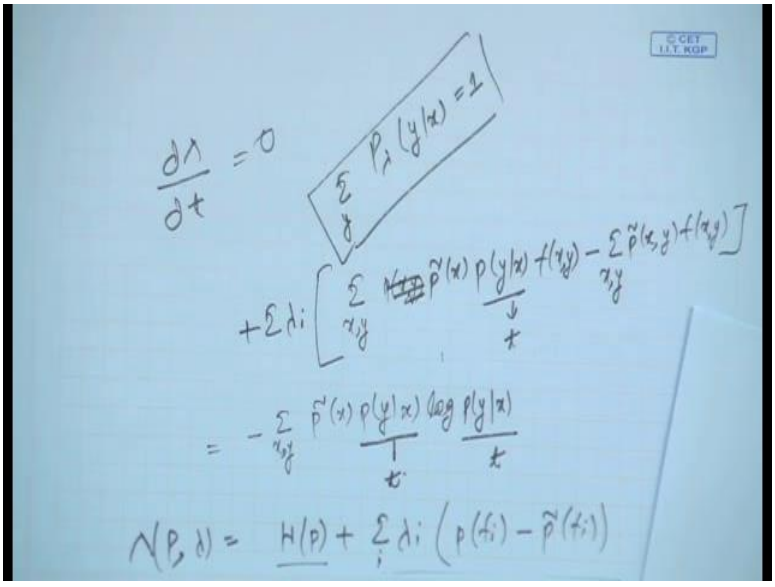
H y given x and that if you do some quickness that will be minus sigma x y P x P y given x log P y given x and again we will approximate this by P tilde. So, this is my maximum entropy model.

(Refer Slide Time: 33:25)



And I want to maximize it subject to constraint that this and these are equal. So, this gives me the models maximize, this H subject to these constraints.

(Refer Slide Time: 33:48)



For that we use the trick of being the lagrangian. So, I want to now a k lagrangian P or lambda that is my H P plus summation i lambda i P f i minus P tilde f i. So, that is I want to maximize this sorry. So, I want to maximize this yes subject to this constraint.

Now if I put these values inside this so I will get - plus sigma lambda i summation x y P x y stepwise P tilde x P y given x f x y minus summation x y P tilde x y f x y, this is my lagrangian and I want to maximize that. So, I want to find out this distribution that maximizes this. So, let me just substitute this (Refer Time: 35:58) t everywhere, so in turn what you are doing? You are maximizing it with respect to this t and that you can do by differentiating it with respect to t and then putting it into 0 and that will actually give you the equation that we have seen; that is p lambda y given x that is t over in this place it will come out to be exponential over summation lambda f i x y divided by normalization constant.

This is simple exercise that you can do you can just differentiate it with respect to this parameter t and you will actually come up with something close to what we have seen in the formulation of maximum entropy model. And this normalization constant is nothing but something that makes this probability p lambda y given x summation over y to 1 and that gives me the normalization constant. This is my maximum entropy model.

So, you have to, the important part here is to find out the set of features once you have a set of features you can use this particular equation to compute the probability of a class y given the observation x.

Next, in the next lecture, I will try to take up is a problem that will make the idea of maximum entropy model for that clear then we will see how given a new sentence we use this maximum entropy model to come up with the tag sequence and there we will study this beam algorithm, beam search algorithm. So, see you then.