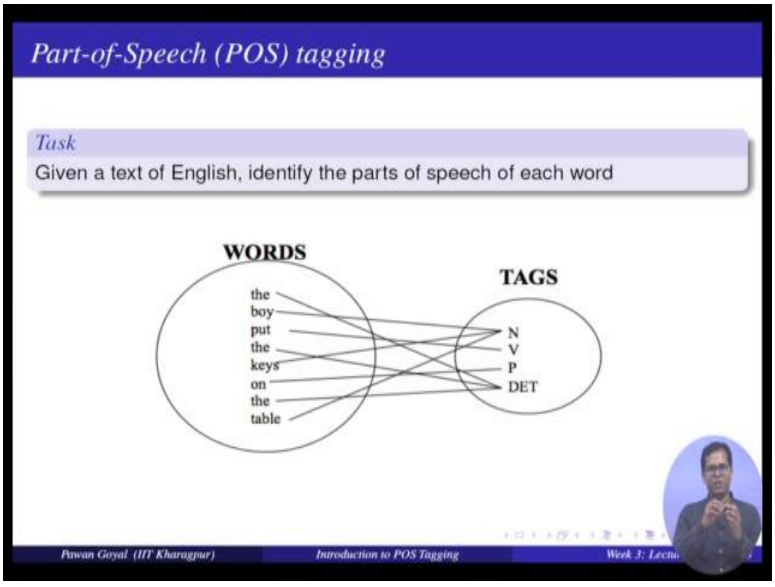
Natural Language Processing
Prof. Pawan Goyal
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture - 14
Introduction to POS Tagging

Welcome to the 4th module of this week. So as we said last time, we will be starting with a very very popular problem of part of speech tagging, if you remember this was one of the processes that we talked about when dealing with morphology. So, you can do lamentation, you can do morphological analysis, you can also tagging. So, what was the thing that was different in tagging that you are not doing morphological analysis, if you remember, you will also finding out among all the possibilities what is the actual particular morphological degree that this 4 things.

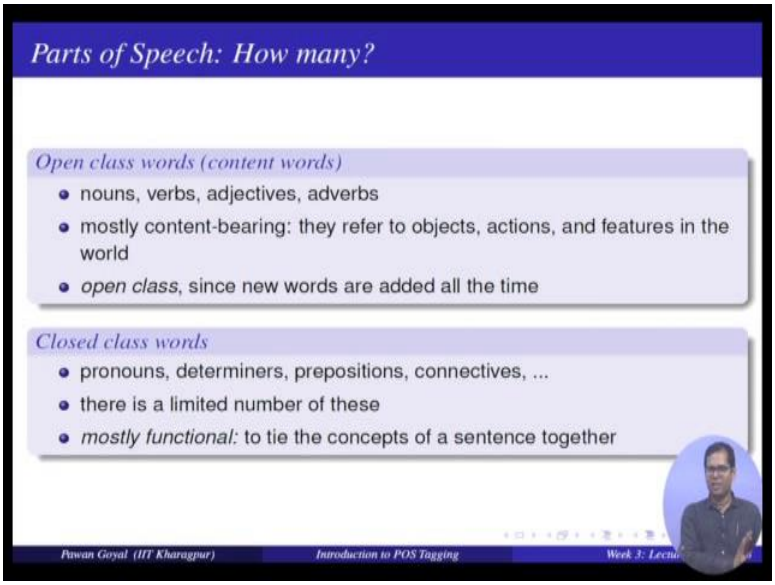
You also are doing the disambiguation here. So, this is my problem of part of speech tagging that given a set of words so; that means, a sentence or document whatever can I identify what is the actual category for each of the word, I have to give the unique answer for each one. So, if we take example for English text. So, given a text of English can be sentence or whatever can identify the part of speech of each and every word that is the problem of part of speech tagging.

(Refer Slide Time: 01:17)



If I have a sentence like the boy put the keys on the table and I have 4 possible part of speech tags. So, very very coarse label like noun and pronoun v for verb and p for pronoun the first p here is for prepositions and d DET for the determiner. So, you can map – the, is a determiner and boy, is a noun; put, is a verb; the, is a determiner; keys, a noun and on, is a preposition and the, is a determiner and table, is a noun.

(Refer Slide Time: 02:19)



Each word you can map to one of the 4 tags. So, this is a problem that given a sentence, you have all the individual words can identify and define what is in grammatical category for each of this word.

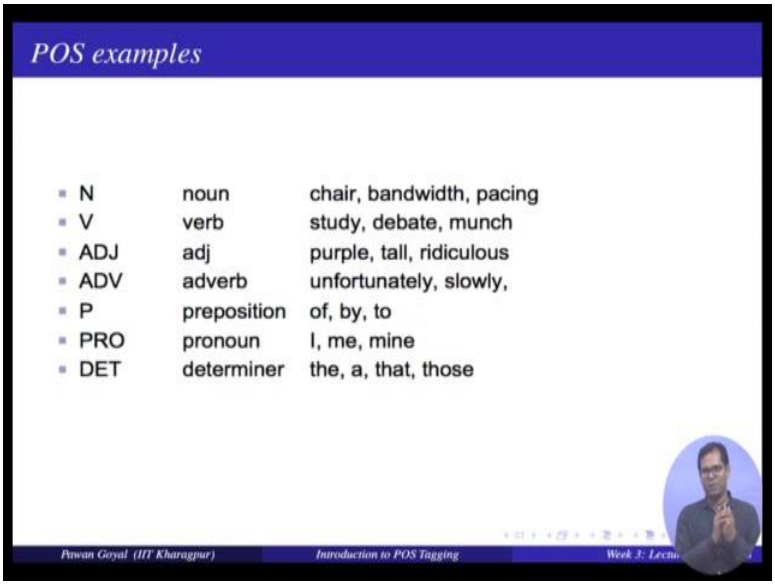
Now, so, one natural question that you might have is so, how many different part of speech tags are there? Firstly, let us see, what are the different things that will go in the part of speech part of speech tag categories? Firstly, you will have the open class words that are sort of content words. So, they are like nouns, verbs, adverbs, adjectives they are they are there, they will have their own part of speech text. So, they are the open class words are those that we are the content mostly and, so why do we call them open class because you keep on adding new and new words in the language we have coin we have seen a term for this.

You have words like Google, Photoshop, Skype; they come into the language all. So, these are coming into nouns or verbs in the language. So, this is open class you can keep on adding new and new words and then you have the closed class words like pronouns

here very fixed set of pronouns determiners very a fixed then prepositions connectives and all that they are very fixed they are all functional words and they are closed class and as you know they are they are used to tie various concepts the sentence together.

I need some part of speech categories for open class words some categories for closed class words. So, one possibility can be I can choose very very coarse grained categories.

(Refer Slide Time: 03:53)

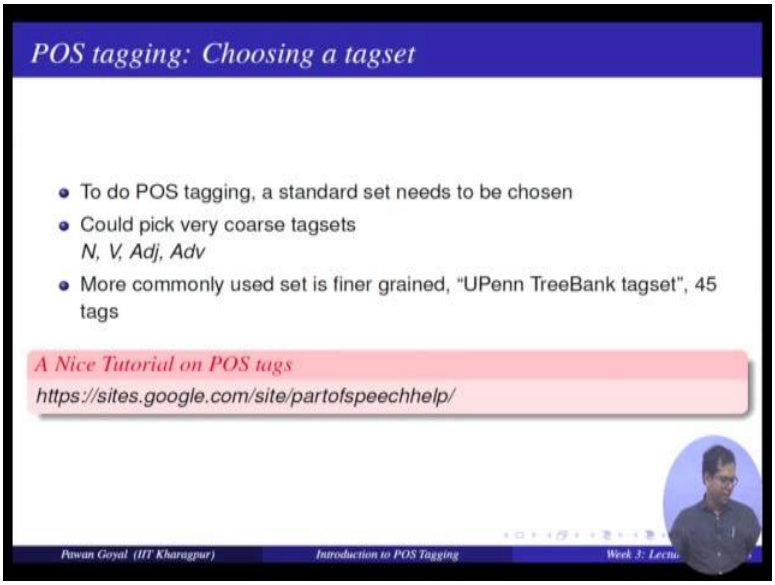


Like here, so I can take noun so I am also giving some examples here like chair bandwidth pacing all these are nouns, I can take verb, study, debate, munch, all these are verbs then adjective like purple tall ridiculous they are adjectives adverbs unfortunately slowly proportion of by, to, they become my preposition pronoun I, me, mine and then determiner, the, a, that, those.

Maybe I can just use it very very coarse label part of speech tags. But think of the perspective when of language pronouncing, so what is helpful to me? To take a very coarse label or to go to the more fine grained label. So, here what will happen? I can find out which word is a noun, but if I to tell it is a singular noun a plural noun I cannot tell if I have even if I have the tagging information. So, I might want to prefer a tagging scheme where I also go to some sort of grammatical details of the word, it is a noun and is it a verb is it a the third same person singular verb is it a verb plus a past tense and so on even noun, is it a proper noun?

I might want to go into finer detail, but again I cannot go into very very finer details. So, otherwise there will be too many part of speech tags.

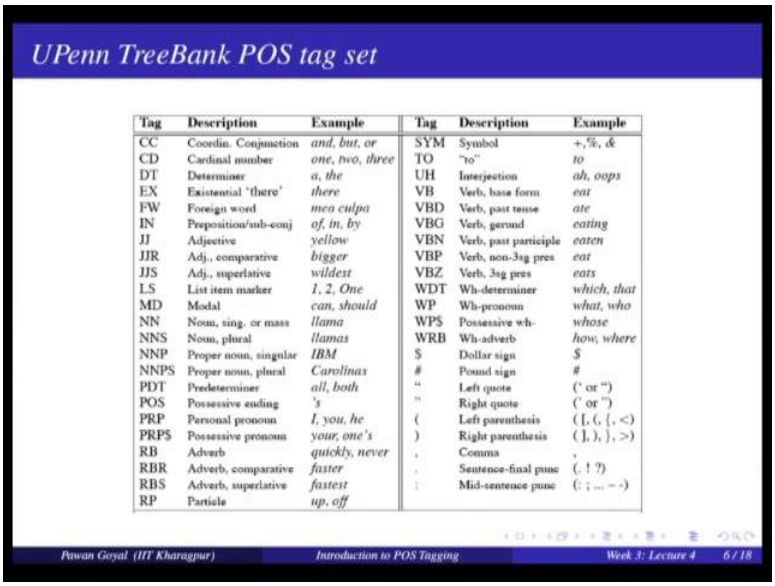
(Refer Slide Time: 05:21)



There are various schemes that are available, they have their own part of speech tags definitions, but given a sentence if you have to do part of speech tagging, you need to first define what is your tagset. What is all the possible categories among which you have to choose it and there are various standards available.

I can choose very very coarse tagsets like only taking noun verb adjective adverb or I can take a good enough set that gives me some additional grammatical information with each of the word each of the word. So, one for very popular part of speech tagset is university Pennsylvania, UPenn, Treebank, tagset that contains every 45 part of speech tags and we will see examples also and there is a very nice tutorial on these part of speech tags in terms of what is the difference between this part of speech tag versus at part of speech tag, particle versus adverb how are they different by giving examples. So, that I recommend as if you want to get more information in the sense of how they are used in various sentences you can look at this site.

(Refer Slide Time: 06:34)



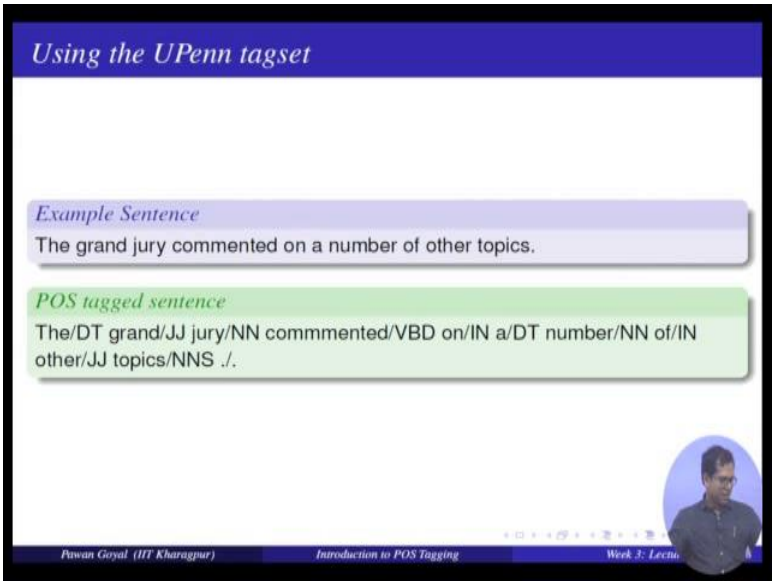
Here are some examples of the Treebank, the UPenn Treebank part of speech tag. So, what are the part of speech tags they used and what are the various sort some example words here. So, for example, if I see conjunction so, and, but, are, all the conjunction they give a tag like CC, 1, 2, 3, they are cardinal numbers, they give a tag like CD a, the, they are determiner - give a tag of DT, then additional there foreign word then, off, in, by - they are preposition, yellow - adjective, then bigger so, again with adjective you have comparative, superlative. So, you are adapting some more information there (Refer Time: 07:16).

Now, if you see at the nouns, you have a tag NN for only the noun, it is a singular noun, a masculine, sorry, it is a singular noun, for plural noun you have a tag like NNS, but, now you see there an extra text, for proper noun you had separate tags. So, once you know the tag you know it is a noun or it is a proper noun and whether it is a singular or plural you get all this information. So, that is why getting it into some finer details is helpful. So, you have 4 different texts for nouns.

Similarly, so if I am escaping those I am going to the verb directly. So, here text like VB for the base form of the verb then VBD for the past tense, VBG for gerent, VBN for past participle and then third singular present is VBP and third singular and is like VBZ non third singular versus first singular there are 2 different forms. So, you have different

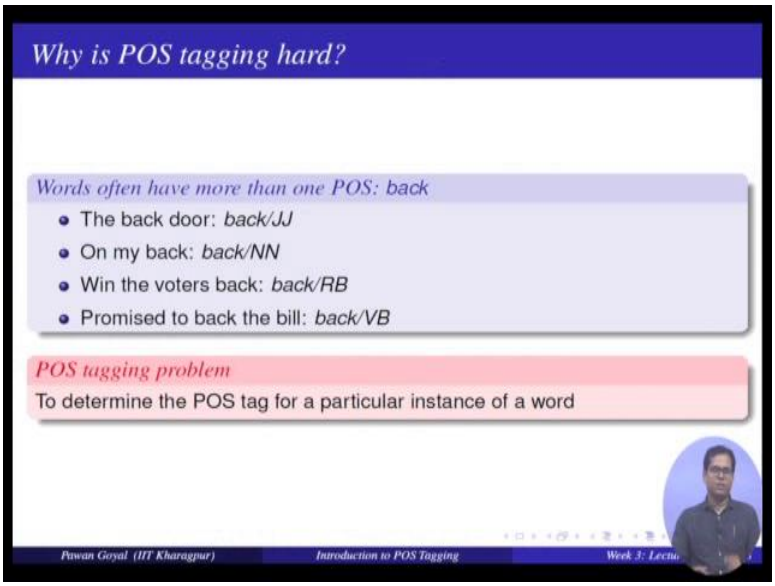
forms of verbs, again that they give you various grammatical information and then you can see other sort of part of speech tag that are available in to UPenn trees tagset.

(Refer Slide Time: 08:34)



Let us take an example then. So, we have taken a taking a sentence and if we tag it by UPenn tagset, how does it look like. So, the sentence is the grand jury commented on a number of other topics. So, now, here the word does determiner grand jury adjective. So, jury would be a noun commented becomes a verb - you have seen the part of speech tags, on is a preposition i n a is determiner number is a noun of is again a preposition other adjective and topics they come in noun in plural form. So, this is the information that you get by the part of speech tags.

(Refer Slide Time: 09:27)

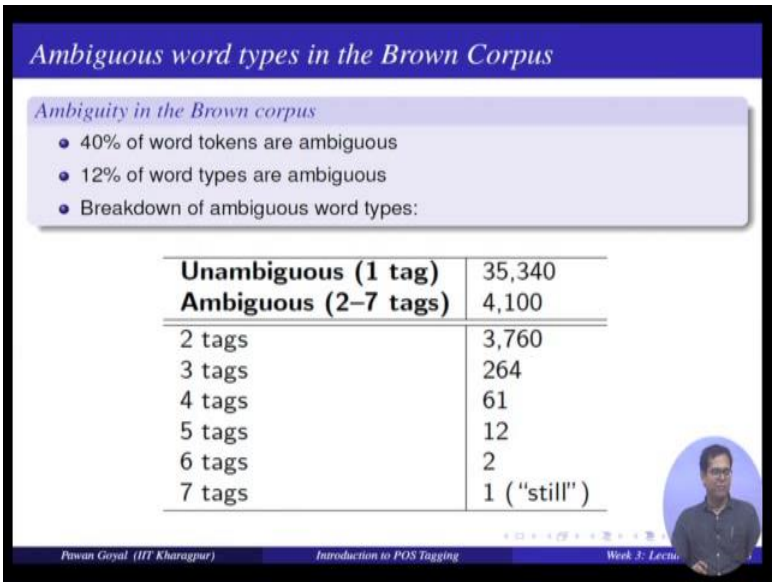


Now, the question might come that is why are we worried about solving this problem of part of speech tagging, it is a hard problem or a trivial problem. Can I define for each word in my lexicon what will be it is part of speech tag it is a very very simple problem no. So the problem occurs because each for does not have a unique part of speech tags and it might depend on the context in this the word is being used. So what I am saying is that a word might have multiple part of speech tags and only by seeing the context you might be able to identify what is the actual part of speech tag is being used in this particular context.

Let us take an example. So, I have the word like back. So, what are the parts of speech tags that this simple word like back can help. So, if I take the sentence like the back door what is back here, what is the part of speech tag of back? So, it is an adjective now if I take a word like, sentence like on my back it is not an adjective anymore it becomes a noun in this case yes. Now I have it win the voters back of the sentence it becomes in win back, so it becomes in adjective and if I have a sentence promised to back the bill now this becomes a verb. So, the same word back can be used in multiple part of speech tags.

Immediately you can see the problem that given the context find out what is the appropriate part of speech tag is be used. So, to determine the part of speech tag for a particular instance of a word is my part of speech tagging problem.

(Refer Slide Time: 11:08)



Now, what are the various information that we can use for doing this, for how co how common is this first of all how common is this problem how many words are actually ambiguous in terms of part of speech tags. So, if I see if I want to see that from the data. So this is your this, so, this point I just want to say something like once you encounter problem in language or any other field, the first thing that you might want to see is that how common is that problem if it is a very very rare problem maybe it is not worth to spend too much time you can have simple rules for solving that, but it says if it is a very very common problem then yes you might have to tackle it using certain models.

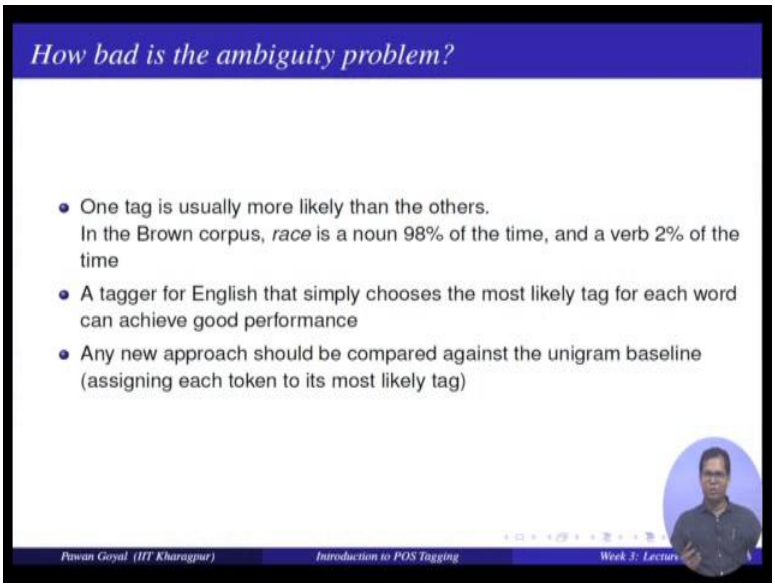
Similarly here the disambiguation part is speech tags. So, then it happen only for 10 words if it happens for only for 10 words, I need not worry about building some models and all that for solving this problem, but if it happens for say 10 percent of the words then yes there is a real problem and I need to think of some model for solving it. So, now, if let us see from the corpus how common is this ambiguity problem. So, if I take the brown corpus, what we see here 40 percent of the word tokens are ambiguous and 12 percent of the word types ambiguous.

I hope you remember the distinction between types and tokens are the all the passes all the occurrences of different words. So, same word occurs multiple times are multiple tokens. So, 40 percent of word tokens are ambiguous. So, what we are saying? Saying is that in a corpus if I am encountering and tokens 40 percent of them are ambiguous that is

a huge number 40 percent of all the words that occurred in the corpus ambiguous so; that means the real problem.

Now, if we want to just break down of the ambiguity type that how many unique words have different number of tags. So, what we see here. So, roughly 35000 types have only 1 part of speech tag, now 3760 types have 2 tags and 264 have 3 tags and so on and there is 1 word like is still that has got 7 different part of speech tags in the brown corpus. So, yes, getting 6 to 7 tag is very very rare, but getting 2 and 3 tags is quite common especially 2 tags is very very common in the brown corpus.

(Refer Slide Time: 13:36)



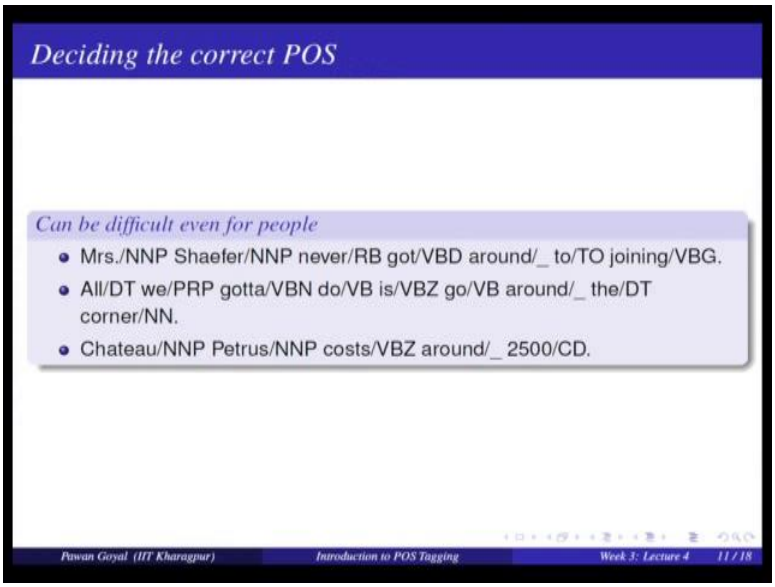
Now the next problem so we have solve that this problem is frequent yes this is not a real problem, but how bad is this problem what do I mean by this. So, can we always identify for a given word is one tag more likely than another. So, let us take an example of the word race - race can be a noun and a verb when the brown corpus race at race occurs as a noun 98 percent of the time, but as a verb 2 percent of the time.

If I have a simple model that always assigns the word race to a noun that will immediately achieve a 98 percent accuracy for this particular example at least; that means, whenever I am trying to design a model I should be able to think of what is simple baseline uninitialized compare if I am making computational model that is working even worse than this baseline it may not be very very helpful. So, despite this

can be my simple baseline for testing any of the model that I will propose for this particular task.

A tagger for English that simply chooses the most likely tag for each word can achieve good performance. So, it can be even more than 90 percent. So always it is not good to look only at the final numbers, it is also good to see how much you are improving over maybe some of the simplest baselines and some other models that are there in the literature. So, at least the simple baseline how much you are able to do better than the simple baseline.

(Refer Slide Time: 15:25)

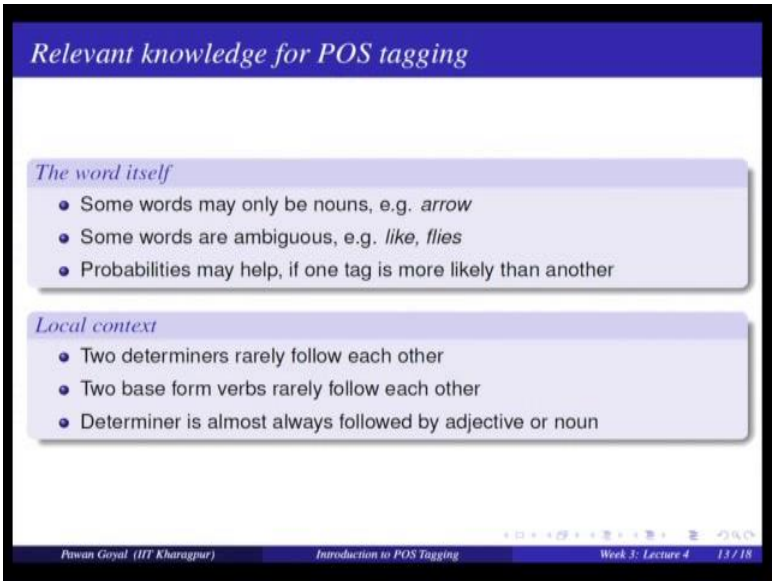


Say any new approach should be able to should be should be able to compare against the simple baseline now how do I decide the correct part of speech tags. So, yeah one thing is a it might even be difficult for people in some cases right it is not a very very easy problem always. So, here in this slide we have 3 sentences this is Shaefer never got around to joining all we go to do is to go around the corners and chateau Petrus costs around 2500 and all the 3 words, 3 sentences, you have the word around and you will find out what is the part of speech tag and you will see this is not very very easy if you will just look at these sentences is not easy to find it what are speech tags.

I will just suggest that you go to the tutorial once and so the tutorial that I talked about earlier. So, that all talks about what are the differences between various part of speech that will give you some idea on how to find out the actual part of speech tags of around

in this case. So, what we will see here? Before in the first case, it is a, the particle second case it is a preparation and third case, it is in adverb.

(Refer Slide Time: 16:30)

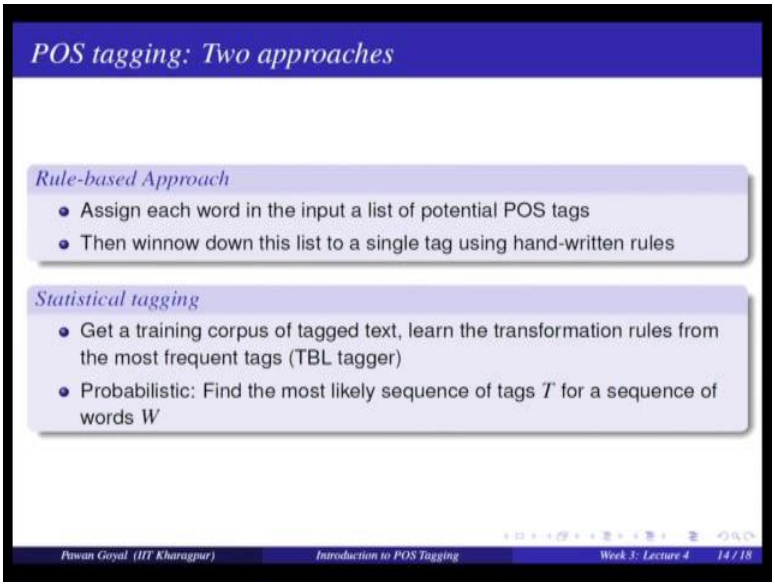


Now, what is the relevant knowledge that we might need for this part of speech tagging problem that we might need to give to our models? So, for example, some words might always be in one part of category like arrow is always a noun. So, I can have this knowledge some words are ambiguous like fly, flies and what like it can it is again ambiguous. So, what might help is what is the probability that a word occurs as a particular part of speech tag the baseline that you are talking about this might be helpful for our model also that given a word what is the most probable tag for this word this is one thing that we might use.

Now, what can be the other information that is useful? So, take the word leg flies, if you do not give me any other word, I may never be able to tell with full confidence what should be the corresponding correct part of speech tag unless you tell me the sentence where it occurs. Same was with the word like saw that was one of the earlier examples, unless you give me the sentence like Peter saw her I cannot tell the saw is actually a verb and not a noun so; that means, I need I need to do something of the word itself how come how common it occurs with some part of speech tag then another, I also need to know about the context in which the word is occurring and how can I use the context to desegregate the actual part of speech tag. I need the local context in the sentence.

So, for example, the information that to determiners really follow each other so if my model is saying 2 words, the determiner which is not allowed. Similar 2 based forms of the word they do not follow each other they should not come together similarly my model can tell me that a determiner is always followed by an adjective or now this can be useful information that if the previous word is a signed determiner it is a highly likely the next row will be adjectives or noun. So, all this we want to encode using our models as well.

(Refer Slide Time: 18:45)



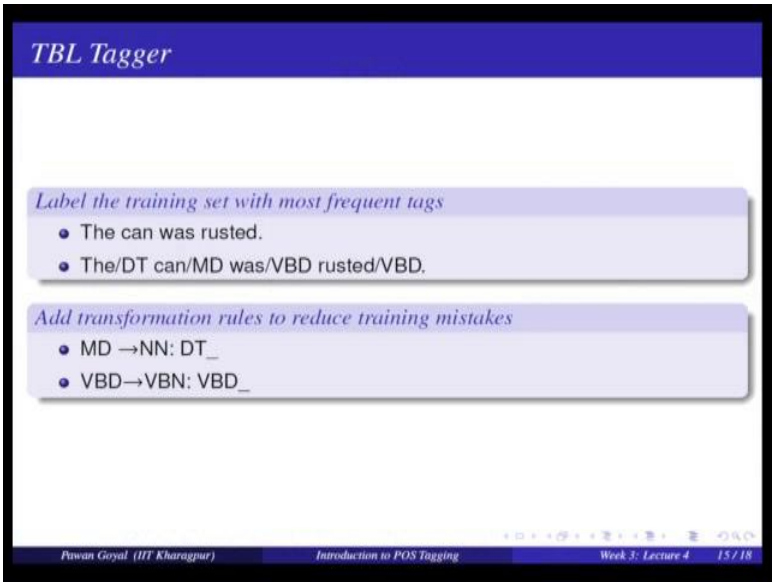
What are the various approaches? So we will try to use all this knowledge, but what are the various approaches that we can use? So, for all the problems that we will be is dealing with, in this course mostly, you can always use a rule based approach. So this visual some of the earlier models that work that were used in NLP. So, we are given a problem some language will sit together find out what are the symbol patterns or simple kind of if then else rules that one can write down to solve this particular problem.

What would be a rule based solution to this problem for doing the part of speech tagging disambiguation problem. So, what I will do? I find out for all the words that can have multiple part of speech tags what is one particular thing that that can help me decide whether to use one tag or the other and this and given a new context I can again use this is that particular contexts appearing here or not; then I can have a statistical tagging that is I get a training corpus. So, this is a standard model. So, I have a corpus training corpus

that has the tag text by tag text I mean I have the sentence and with each word I also have the actual part of speech category.

Now, using that can I learn, what is the actual part of speech tag for each individual word and this in a new sentence, can I learn some model? So, this is my statistical tagging. So, they are again different various models. So, one simple model is TBL tagger that was one of the earlier model proposed for part of speech tagging. So, that is I am given a any corpus of tag text can I learn some sort of transformation rules that this word has a particular main category most probable category of part of speech tag, but given the context if this most probable text should be changed to some other tag. So, can I use some rules from a corpus training corpus itself? And then the probabilistic models where I will have a probabilistic in the position of what is the most likely sequence of tags for a sequence of words.

(Refer Slide Time: 21:33)

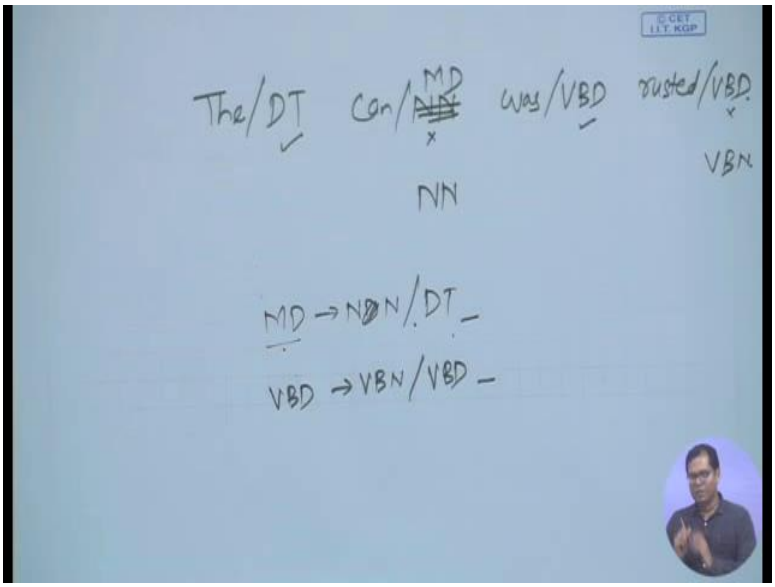


We will go through TBL tagger very very briefly and then we will devote a lot of time on the probabilistic taggers. So, what is TBL tagger transformation based learning? So, let us try to understand the idea first. So, I have a sentence like - the can was rusted. So, how the model will process? So, this is my training data set. So, I also know, what is the actual part of speech tag for each of the individual word?

Now, the first thing you will do is to find out for each individual word, what is the most likely part of speech tag. So, in other words the can was rusted. So, if you find out a

more slightly tag for the word can it will be a model work and mostly occurs as a model work similarly last word rusted it would be a verb in the past tense that is the most popular probable tag.

(Refer Slide Time: 22:24)



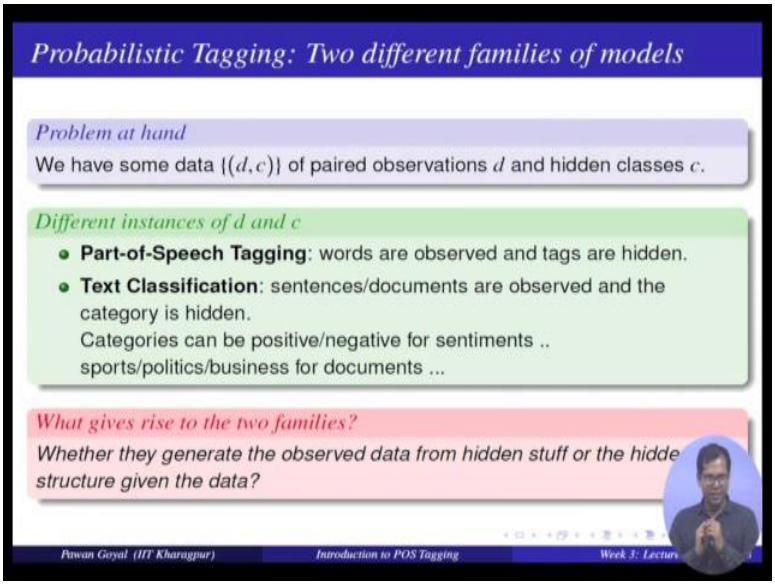
I will write down the can was rusted, now I will see in my actual training sentence what are the actual or thus rejected are assigned to different words. So, I find this is correct this is incorrect this I am sorry, this the most probable tag will be MD moreover is incorrect in this and as I will have a noun this is fine, but this will be instead VBN is a participant past participial. So, these 2 are incorrect. So, I need to write some moves so, that these can be transformed.

One symbol can be MD changes to NN sorry to NN when preceded by DT remember the rules that we saw in the previous lecture. So, it needs to be if preceded by something and there is nothing so, we are not putting any restriction what is being followed. So, this is a rule that I can use. So, now, what will happen? In a new case whenever a word is assigned MD, if I see the previous word is determiner I will change MD to NN, this is the rule I am running from this example similarly I can rule a learn, a rule here VBD goes to VBN. Whenever preceded by reading this can be another rule.

So, this is the idea - I have training corpus I find out for each for what is the most likely tag whenever they do not match I will write down some set of rules and then I keep on

doing that and I might put down a separate data set, simple small data set for testing how good my finally, tagger text. So, this is my TBL tagger. So, this is what we have seen.

(Refer Slide Time: 24:21)



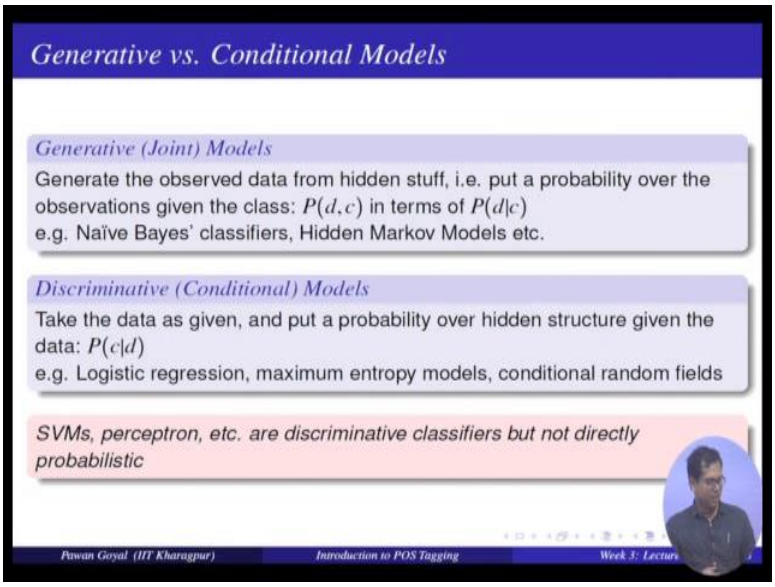
Now, what is probabilistic tagger? In probabilistic tagging I will have a probabilistic interpretation of what is the most likely sequence of tags that should be used for a given sentence. So, let me just briefly talk about what in general the probabilistic tagging models do. So, in these models we have some data that is some observations d and a certain class. So, in the case of part of speech tag what is the example of d and c? So, d are the words and then tags are my classes I want to assign various classes that that tags to different words. So, tags some classes.

Now, you take a different problem like text classification what will happen the documents all the sentences if you are doing classification or documents a sentence is they become your data and you have your classes. So, suppose using sentiment analysis. So, the sentence is your data and the class is positive negative or neutral if you are doing text classification in terms of sports, visage, politics and so on, there is categories, the document or the vertical instance of the text becomes your data and that becomes your class a sports and politics and so on. So, you have a paired observation of a data and a class.

Now, they are 2 different families of probability models that can be used for solving these problems I will just briefly talk about these 2 families and we will take examples

from both of these for this problem. So, and you will be able to use that for many other applications NLP also. So, what gives rise to 2 different families that is whether you generate the data from the class or the class from the data. So, what is the philosophy of your model?

(Refer Slide Time: 26:14)



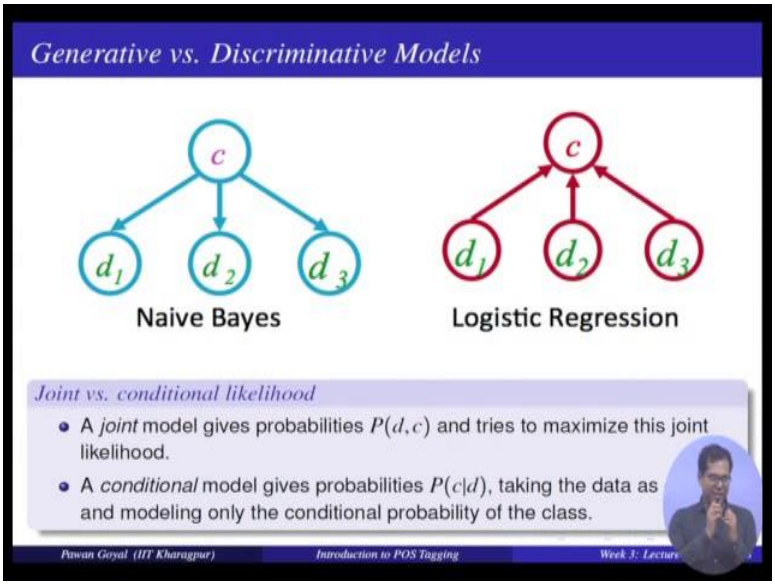
Let us try to understand that from this slide, so, we have 2 different types of models, one is a generative model and second is a conditional model. So, in generative model what will happen? So, you have appeared data and the class. So, generative model you will assume that the class is there and the class generate the data. So, this is the philosophy from the class that it I generated. So, examples are like nigh page modular hidden Markov models. So, your class is given and data generated from this.

In the discriminative condition models what will happen, that your data is there and you assume that hidden state is a generated from the data. So, that is the minor difference between that. So, in the condition, in the joint model you first at the class you first generate the class and then you generate the data from the class. So, that is so to give a simple example if you want to use a joint for generating modules for text classification what will you assume, if I am going to take a document I will first think over what is the topic I want to write say politics I think about the class and now given this class what is the probability that I write all these words. So, from that class I find out the probability of different different words. So, that is how my model is defined. In the case of

conditional model given the observation directly I want to find out the probability of the class.

This difference tells me whether so how do I actually go about solving these models. So, they are the models like SVMs perceptron that are not probabilistic. So, they are not in division one of these. So, they are also discriminative classifiers.

(Refer Slide Time: 28:05)



Now, this picture will help you understand that 2 models again. So, you see the directions are different in the pictures in the first direction, in the first in the left hand side picture. So, direction is from the class to all the data points. So, first the class is generated then all the data points so you generative model.

In the second one is a conditional model. So, given the data you want to find out directly the probability of the class. So, first one example is Naive Bayes, second one example is logistic regression. So, a joint model will give you probability of, d l the data and the class together and you will try to maximize the joint probability and condition of model you will directly want to find out the probability of the class given a rate. So, we will take examples of both of these in the next modules. So, what is a joint model that can help me solve the part of speech tagging problem and what is the condition model that can help me solve the part of speech tagging problem.

Thank you.