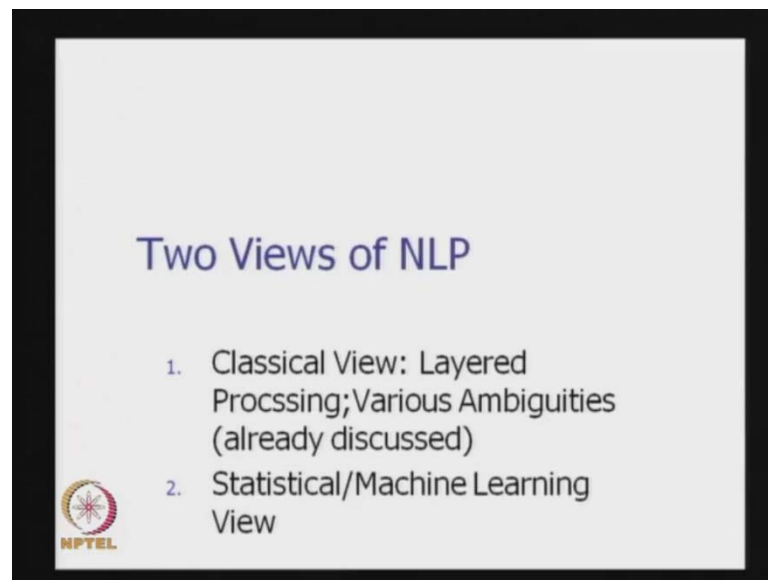


Natural Language Processing
Prof. Pushpak Bhattacharyya.
Department of Computer Science and Engineering
Indian Institute of Technology, Bombay

Lecture - 4
Two approaches to NLP

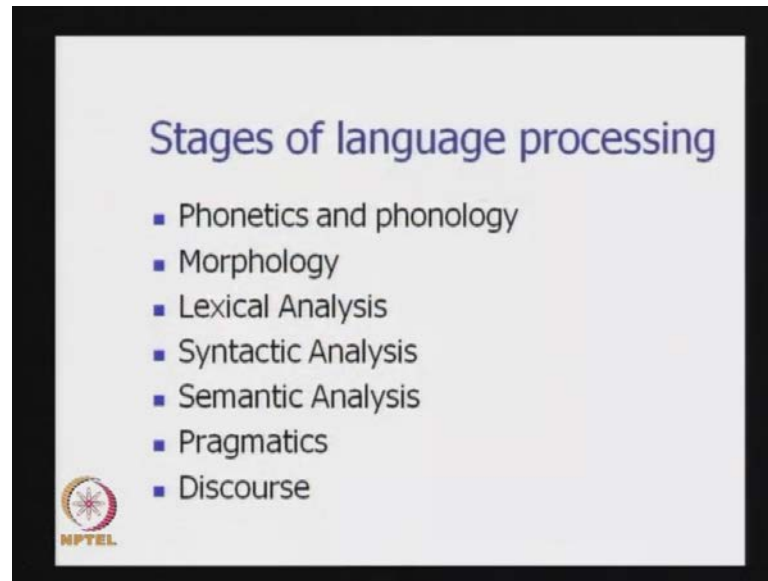
This is lecture 4 on natural language processing; we will talk about 2 views of language technology. Let us look at the presentation this slide, and what we see here is that there are 2 views of natural language processing.

(Refer Slide Time: 00:29)



The first view is classical view, layered processing, and various ambiguities which we have been discussing over last few lectures a last 3 lectures actually, and the other very predominant view is statistical or machine learning view. Let us spend some time on understating the difference between these 2 views. Why is that that there are 2 views of natural language processing? There are 2 predominant approaches the first approach is the classical approach we have seen many different stages of natural language processing namely phonetics, phonology, and so on, at every stage there are ambiguities which have been discussed extensively. In classical view of natural language processing, the owners of processing is on human beings. In this view a machine essentially executes the instructions given by a human being.

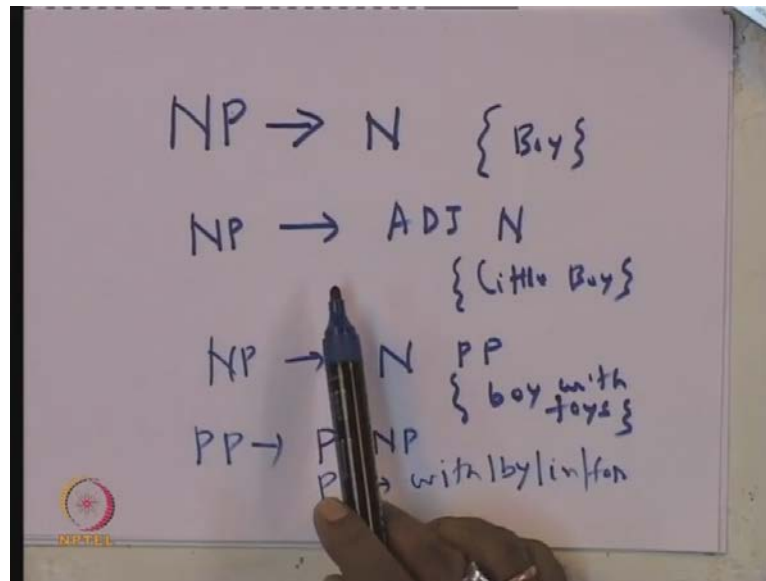
(Refer Slide Time: 01:48)



Let us look at this slide and remember what we saw on various stages of natural language processing. The first stage was Phonetics and Phonology and then came Morphology, Lexical analysis, Syntactic analysis, Semantic analysis, Pragmatics and Discourse. In each of these stages, there are human beings involved who create rules guided by linguistics, lexicography, and knowledge of language and so on which makes the machine process natural language data or information.

Just to take an example, if we take the example of syntactic analysis, where one needs to phrase a sentence what happens is that? A linear structure is given, a sentence is given and from the sentence we obtain a tree corresponding to the sentence. We identify the noun phrase and verb phrase within the verb phrase we find out the verb and so on. This whole processing happens by means of grammatical rules which a human being has encoded somebody who understands the language well has shut down and produced the grammatical rules. Now, when this grammar is written, the person producing this grammar has to anticipate all possible language phenomenon which exists in that language, and try to capture them in turns of grammatical rules. So, let me show you an example of a grammatical role by writing it on the paper.

(Refer Slide Time: 03:47)



So, suppose I say that in noun phrase N P, this is the symbol for a noun phrase N P goes to a noun. So, that means a noun phrase can be expressed by a single noun or a noun phrase can be an adjective and a noun. Let me give an example noun phrase going to noun could be boy, noun phrase going to be adjective and noun could be a little boy. Noun phrase can also go to noun and preposition phrase for example, boy with toys. So, all these are noun phrases let us look at the lines once again noun phrase can be noun.

For example, boy noun phrase can be adjective and noun little boy, noun phrase can be a noun with a preposition phrase boy with toys. Preposition phrase again can be expanded as a preposition P and a noun phrase coming after that. So, preposition again can be preposition phrase can be P at N P that means preposition at noun phrase and P in turn can be things like with, by, in, for and so on. So, you can see what I wanted to give a grammar for a noun phrase. So, what are the possibilities of a noun phrase, noun phrase can be a single noun, noun phrase can be an adjective and noun, noun phrase can be noun followed by a preposition phrase.

Preposition phrase on the other hand is preposition followed by a noun phrase, preposition can be with, by, in, for and so on. So these shows that to capture a language phenomenon and to give its grammar, we need to anticipate all possible situations we have to understand all possible situations. Now, this is what a language expert does, language experts knows various language situations and produces rules for them. Noun

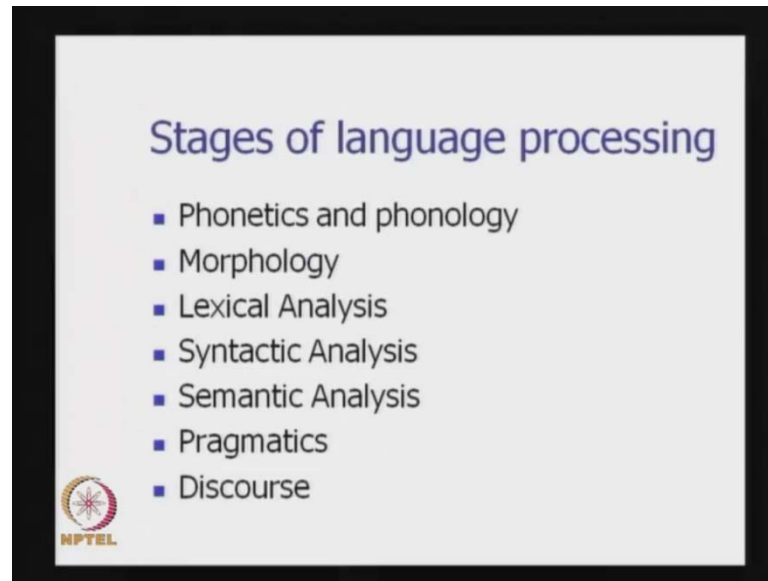
phrase going to noun, noun phrase going to adjective and noun these are actually rules expressing language phenomenon.

So, in classical view of natural language processing the complete owners or the burden is on this kind of rules which are created by human beings, this was the scenario in natural language processing. Rules and knowledge come from human beings the advent of wave change the scenario in a very dramatic way. Because of the internet a large amount of text in electronic form became available on the web and this text also can be processed by a machine. So, machine process able text in large volume became available. And this kind of text was a very reach repository gold minds are to say of language phenomena.

Now, here was a possibility were these language or text could be processed by a machine. And the regularities in the language or the constrains could be uncovered from this text. There is a technical name for the text let me write it down, the technical name for text is **CORPUS** I write in an very bold font and large font, because corpus is highly important very important in N L P in natural language processing. So, when we have **CORPUS** in electronic form large amount of text in electronic form we can apply what is called machine learning algorithms, machine learning text techniques on these data?

And understand that there are regularities which are waiting to be uncovered and which can then be used for natural language processing. So, going back to our transparencies what we said was there are these 2 views of natural language processing the classical view, classical we have explained. Now in detail which is completely rule governed and rules are given by human beings these rules contain knowledge. And the second approach is the possibility of using statistical or machine learning approach to uncover these rules and regularities underlying the **CORPUS** or the electronic text. And those rules and regularities can be discovered and used later for natural language processing.

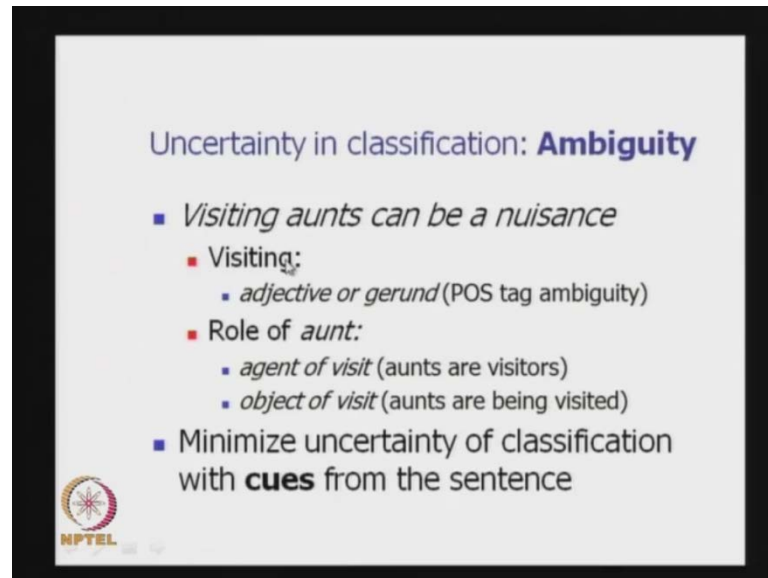
(Refer Slide Time: 09:33)



Proceeding further, we find that there are stages of natural language processing and everywhere one could make use of either the classical approach natural language processing or one could make use of statistical techniques the data driven approach to do natural language processing. For example, if you take morphology, we discussed syntactic analysis some time back if we discuss morphology. Now morphological rules are driven given by language experts, but there are lines of work where different word forms are given.


And from the word forms one identifies the suffixes and tries to uncover the rules which govern morphology. So, it is possible to create a morphology analyzer by making use of word forms and these word forms can be processed by machine language techniques for creating a morphology analyzer. So, all the stages that we have discussed morphology, lexica analysis, syntactic analysis, semantic analysis, pragmatics, and discourse everywhere one can make use of these 2 approaches.

(Refer Slide Time: 10:50)



Uncertainty in classification: **Ambiguity**

- *Visiting aunts can be a nuisance*
 - Visiting:
 - *adjective or gerund* (POS tag ambiguity)
 - Role of *aunt*:
 - *agent of visit* (aunts are visitors)
 - *object of visit* (aunts are being visited)
- Minimize uncertainty of classification with **cues** from the sentence

 NPTEL

Let us look at the problem of ambiguity, what could be the data driven machine learning based approach to ambiguity resolution? If we see this sentence here visiting aunts can be a nuisance this is an ambiguity sentence. Let us understand the ambiguity in this was discussed before I just repeat what the ambiguity is, the word visiting can be an adjective or it could be a gerund the ing from of a verb which is gerund. And this ambiguity can be caught at the part of speech tagging level and this ambiguity may or may not be resolved, but ambiguity is really the part of speech ambiguity.

So, this is the first ambiguous entity, the second ambiguity comes from the role of aunt. Visiting aunts can be nuisance; the question is who the agent of visiting is? The agent of visiting can be aunts are the visitors or the agent of visiting can be the speaker himself or herself. The speaker is complaining that visiting aunts can be nuisance the speaker does not want to visit aunts in which case if the speaker is the agent then the object of visit is aunts. So, aunts are either objects of visit or agent of visit, so we have what is called the semantic role ambiguity aunt can be agent of visit or object of visit.

So, if that is the case if aunts are visiting, if aunts are the visitors then visiting becomes an adjective what kind of aunts? Who are visiting, so these are visiting aunts and if aunts are being visited if the speaker is visiting the aunt? Then visiting is adjunct it is a verb and it is denoting the act of visiting. So, we can trace the ambiguity of this sentence as either visiting being an adjective or a gerund and also role of aunt being ambiguous.

Depending on this ambiguity these sentence as 2 meanings whether aunts are being visited or aunts themselves are visiting.

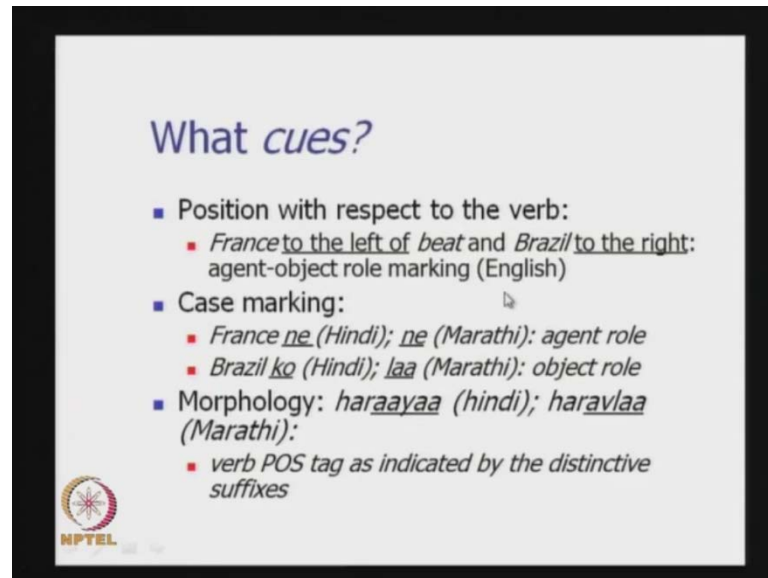
So, this ambiguity the classical view of this ambiguity is that these ambiguity exists and the ambiguity come from different Symantec roles and different part of speech of visiting. Statistical natural language processing would admit it that there is ambiguity, but would like to state that this ambiguity is coming from the uncertainty in classification. One of the important tasks of machine learning is classification there are 2 to there are 2 different kinds of machine learning, one is classification the other is class telling.

These are 2 main tasks of machine learning, in machine learning when we talk about classification we say that entities are given class labels. For example, the objects in this class room can be given different levels depending on what these entities are. For example, I am sitting on a chair the entity on which I am sitting as been given the level of chair, the object in front of me is a table.

So, this entity is given the level of table. So, in classification we have entities and we give them labels how does it apply to the situation in front of us namely the ambiguity? Looking at the slide once again visiting aunts can be nuisance we can give a label of adjective on visiting or the other level of gerund on visiting. These will make the word belong to one or the other class; it can either belong to the adjective class or to the class gerund. So, this is a classification problem and this classification is a class is un result whether or gerund is unresolved.


Similarly, the role of aunt is the entity and the classification for these is agent or object, again we are talking of 2 levels agent and object and the role of aunt will be long to one of these 2 classes' agent or object. Now this is a nice point of view, because we are making use of machine learning paradigm. We are making use of the terminology of machine learning and saying that ambiguity is nothing but uncertainty in classification and this ambiguity is resolved by making use of cues from the sentence.

(Refer Slide Time: 16:48)



What cues?

- Position with respect to the verb:
 - *France to the left of beat* and *Brazil to the right*: agent-object role marking (English)
- Case marking:
 - *France ne* (Hindi); *ne* (Marathi): agent role
 - *Brazil ko* (Hindi); *laa* (Marathi): object role
- Morphology: *haraayaa* (Hindi); *haravlaa* (Marathi):
 - verb POS tag as indicated by the distinctive suffixes



Proceeding further we ask what kind of cues is available to resolve this ambiguity. When we do classification in machine learning we work on what is called the features of the entities we classify the entity depending on the features. For example, a chair is classified as a chair based on features like it has four legs, there is back rest, and there is an area where the person sits and so on. If this entity does not have the back rest even if the person can sit on it is no longer a chair.

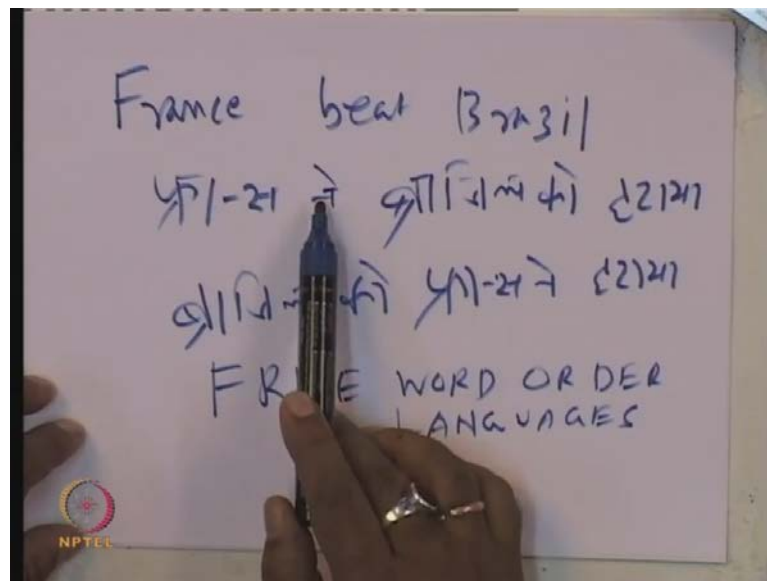
So, there are features some features are distinguishing for that particular entity other features may be common with other entities. For example, the back rest is a critical feature for the chair to be a chair the entity must have a back rest. So, by looking at this kind of features we identify the object or the entity and we give it a class label. So, features can produce our decision so we understand that the classification happens by making use of the features.

And the features actually come from the sentence itself and the sentence contains the tokens or the words. So, these words and tokens are used as cues let us look at the slide and we try to investigate what cues are used for disambiguation. So, one of the cues especially for English sentences is the position of the word with respect to a verb. So, if we take the sentence France, beat, Brazil in again then France is to the left of the beat and Brazil is the right.

So, this tells us that France is the agent and Brazil is the object. This is off course discounting the possibility that the sentence could be a passive voice sentence and the entity to the left of the verb could be an object. So, Brazil was beaten by France in this case Brazil is object even though it is to the left of the beat. However when not considering that particular fact we are considering normal active voice sentence France is to the left of beat and Brazil is to the right. So, therefore, there is no ambiguity classification France is the agent Brazil is the object.

So, agent object marking in English is done by means of these very important cues namely the position of the gerund with respect to the verb left or right. In Indian languages and many other languages of the world where the word order is relatively free we have to make use of some other cues. So, France, beat, Brazil in the football game in this case France and Brazil have fixed positions in English language, but for an Indian language these need not be the case. Let us look at the Hindi sentence corresponding to this sentence.

(Refer Slide Time: 21:06)



So, suppose we take this sentence France, beat, Brazil the Hindi sentence should be France [FL] Brazil [FL], but you can also write Brazil [FL] France [FL]. So, the same meaning is conveyed by these 2 different orders France [FL] Brazil [FL], Brazil [FL] France [FL]. Whereas for English the order is fixed France, beat, Brazil, if you change the order Brazil, beat, France then the meaning also gets changed, not so in case of in

this sentence, not so in this case many Indian language sentences, France [FL] Brazil [FL], Brazil [FL] France [FL].

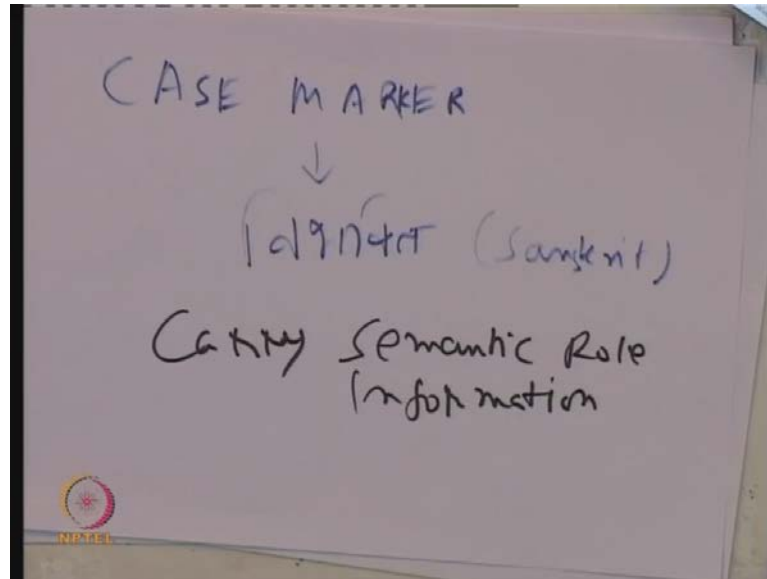
These kind of languages where the word order can be changed are called free word order languages. So, Indian languages are quite prominently free word order languages. But if you are designing if you are observing carefully then you can make out that this free word order came because of a particular factor what was that factor? Notice that France beat Brazil there were no other language particles in this sentence only those entities which are actors in these situations.

France and Brazil are the actors in this situation France is the agent, Brazil is the object, beat is the activity there is a beat activity in this. When we come to Indian languages the actors are same, but their expression in the sentence are done with the mediation of other language particles there is this [FL] which is coming after France there is this [FL] which is coming after Brazil. This [FL] and [FL] has language particles are critical for the meaning of the sentence. [FL] shows France is the agent [FL] shows Brazil is the object since [FL] and [FL] have this very crucial role to play.

They can be moved along with the nouns without changing the meaning of the sentence. Now since [FL] is the agent indicator the position is now less important, if you carry [FL] with France then you know that France is agent if you carry [FL] with Brazil you know Brazil is the object. So, you just pay some attention to this point this is a very crucial point in the English sentence France beat Brazil the, who is the agent? who is the object? This information is encoded in the position of the nouns. Noun is to the left of beat Brazil is to the right of beat and that shows who the agent is and who the object is.

So, you cannot take liberty with the position of the nouns otherwise the agent and object roles are disturbed and that disturbs the meaning of the sentence. In case of Indian language sentences in particular Hindi here the agent and object information are indicated by [FL] and [FL]. So, these make the position information redundant and therefore, we can play with the order of the words. So, I hope this point is clear to you in English position encodes semantic role information in Indian languages case markers typically encode the semantic role information.

(Refer Slide Time: 25:59)



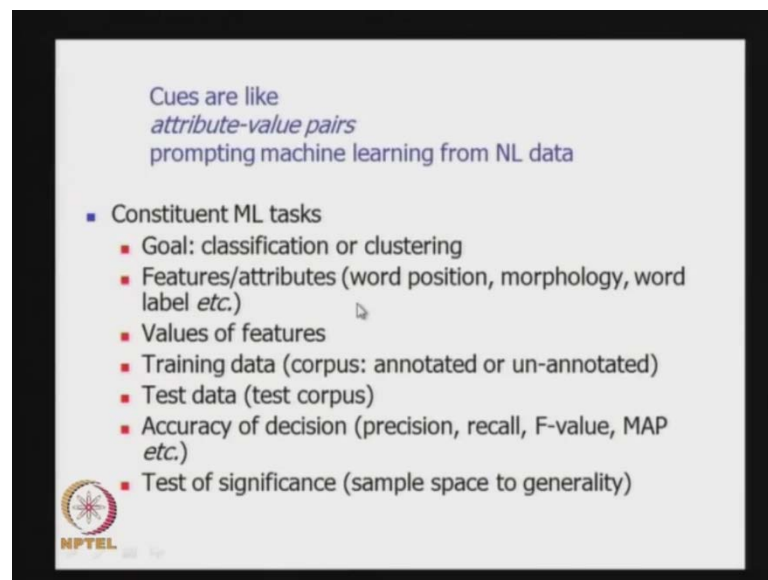
Let me write down a very important term for you, we have introduced this term just now. The term is called case marker also called [FL] in Indian languages. This came from Sanskrit tradition [FL] so in a Indian languages case marker or [FL] carry the semantic role information. So, that is it so we asked how is the classification solved classification problem solved, because classification problem needs to be solved to resolve the ambiguity. Here we find that for English sentence the cue was the position for Indian language sentence the case marker is the cue France [FL].

In Marathi also you can use the same case for the particular [FL] that indicates the agent is role Brazil [FL] in Hindi Marathi Brazil [FL]. So, this shows the object role thus the ambiguity is resolved by means of case marker it. We can also see that in Indian languages morphology can do disambiguation France beat Brazil in this case the word beat has ambiguity, because beat can be a noun. For example, heart beat for heart beat is a noun France beat Brazil beat is a verb these ambiguity does not exist for Indian language.

Because for Indian language the Hindi sentence will be France [FL] Brazil [FL] [FL] [FL] this is a verb and this is a verb is simply shown by the fact that it takes the suffix [FL]. This [FL] is a past tense marker it is an indicator of past tense that gets attached to the root verb [FL] [FL] is to defeat or to beat and Marathi example is shown here [FL] [FL]. So, these suffixes show that this word is a verb it does not have to face the


ambiguity that English beat has become noun or verb. In Hindi there is no such ambiguity so let us recount all possible cues which have been described. For English position is the cue or the clue for this ambiguities for Indian languages or nouns these are the case markers and for verbs it is a morphological suffices.

(Refer Slide Time: 25:59)



Cues are like
attribute-value pairs
prompting machine learning from NL data

- Constituent ML tasks
 - Goal: classification or clustering
 - Features/attributes (word position, morphology, word label *etc.*)
 - Values of features
 - Training data (corpus: annotated or un-annotated)
 - Test data (test corpus)
 - Accuracy of decision (precision, recall, F-value, MAP *etc.*)
 - Test of significance (sample space to generality)

 NPTEL

These are the cues proceeding ahead we consider this cues or clues as very critical for our classification task or the disambiguation task. Cues are like attribute value pairs and we can make use of this attribute values of pairs for the installing for launching machine learning algorithm on the natural language data. Lets recount here the various constituents of machine learning tasks, any machine learning task first has to specify what the goal of the task is does it belong to classification or is it a clustering kind of task?

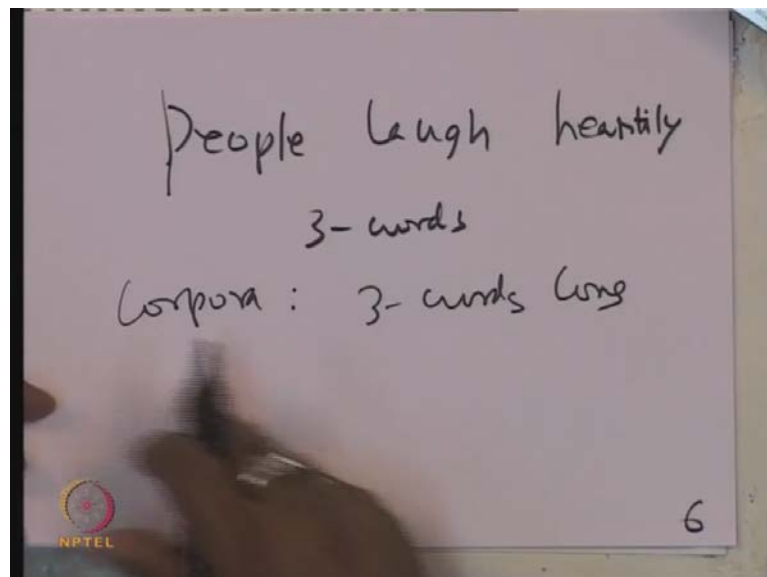
Then we have to clearly demarcate the features and the attributes for example, in natural language the feature should be word position the morphology word label that means the word category noun verb etcetera. The actual values of these features for example, for what position it could be the left or the right position with respect to the word the value of the morphological feature could be a particular suffix. For example, [FL] for past tense in the word label or word category the value for that could be noun, verb, etcetera.

Then having look at this three features we take the next most important constituent which is the training data, the corpus or the electronic text which is annotated or an

annotated will explain annotation or an annotation in a minute. These forms an important constituent then there is this test data the test corpus which is important for evaluation the machine learning algorithm.

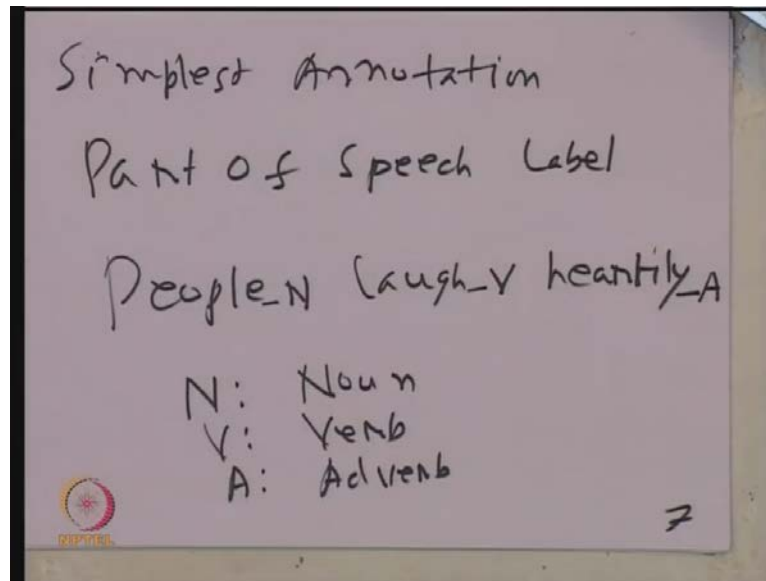
The accuracy of the of the learning situation that means the classification the accuracy is measured by means of precision recall F values map code etcetera which again will explain. And also mean times we perform what is called test of significance we go from a small sample space to a general class and that requires test of significance. Let me know describe a very important concept the concept of annotation.

(Refer Slide Time: 32:09)



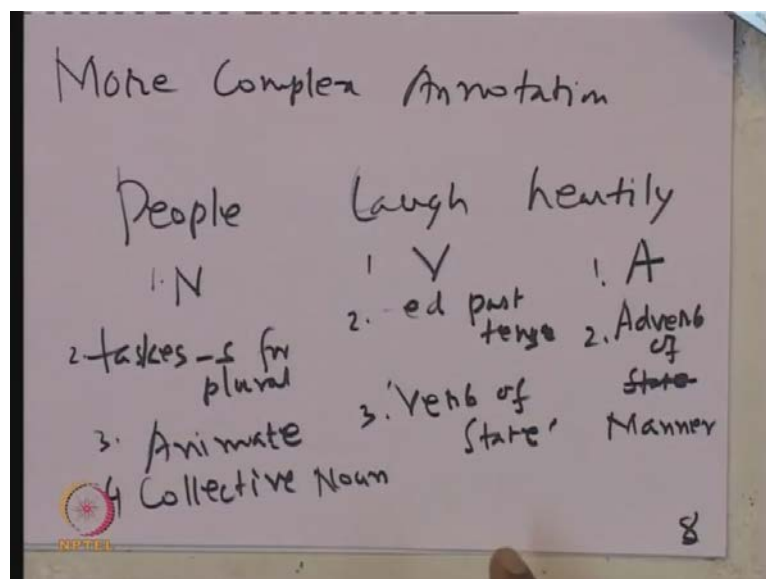
Annotation we considered this word annotation. Let me describe annotation for a few minutes. Annotation is very critical for statistical natural language processing types on annotation so we illustrate annotation means labeling producing labels. Let me give an example of this suppose we have the sentence people laugh heartily. They may be laughing at a joke or a particular situation people laugh heartily so there are 3 words. Our corpus now is 3 words long on these if you produce annotation can be of many different kinds.

(Refer Slide Time: 33:50)



So, let us first do the simplest possible annotation. Simplest annotation, part of speech label annotation we said is a labeling task we are doing part of speech label. So, people laugh heartily we produce the following annotation we say that this is a noun underscore noun, this is a verb underscore verb, and this an adverb let say we produce a very character A. So, N is noun, V is verb, A is adverb so what we have done we have annotated our corpora people laugh heartily with the labels N V A respectively. This is simplest kind of annotation the part of speech annotation. Let us do a little more complicated annotation, what could be a more complex annotation?

(Refer Slide Time: 35:14)



More complex annotation we say that people now we will produce the annotation below the words. Because we are creating more complex annotation people laugh heartily, so in this case people is a noun we can say takes S for plural these an annotation animate this is a semantic annotation. So, people are a noun it takes s for plural it is animate it is also a collective noun. So, we have produced these four pieces of information for the word people noun takes S for plural animate collective noun.

So, one could imagine many different kinds of annotation many different kinds of annotation and can enrich words with these kind of labels to make it a very informative text. Let us go to laugh is a verb all of us know that what more annotation can you produce this laugh takes e d for past tense this is N annotation laugh. It is a verb of state this is a verb of state that means when people are laughing they are at a particular state. As oppose to let us say verb of motion if you say people run speedily, instead of people laugh heartily people run speedily. In this case the word is run and the word run is a verb of motion the word run is a verb of motion, but when we say people laugh heartily people laugh heartily. So, when they are laughing they are in a particular state so it is a state verb so it is a verb indication a particular state. For heartily when you go to word when you go to the word heartily this is an adverb A produce a letter A for this is adverb of state.

So, this is adverb of state we record all this annotation, we have produce 2 annotation marks for heartily 3 for laugh, 4 for people. Now adverbs we know can be of many different kinds adverb can be space adverbs of space, adverb of time adverb of manner, and so on. So, this particular adverb people laugh heartily it will be more correct to say this is an adverb of manner instead of adverb of sate we call it adverb of manner. So, this is the task of annotation and I hope your clear about what happens in annotation, let me summarize this particular point once again it is a very important point.

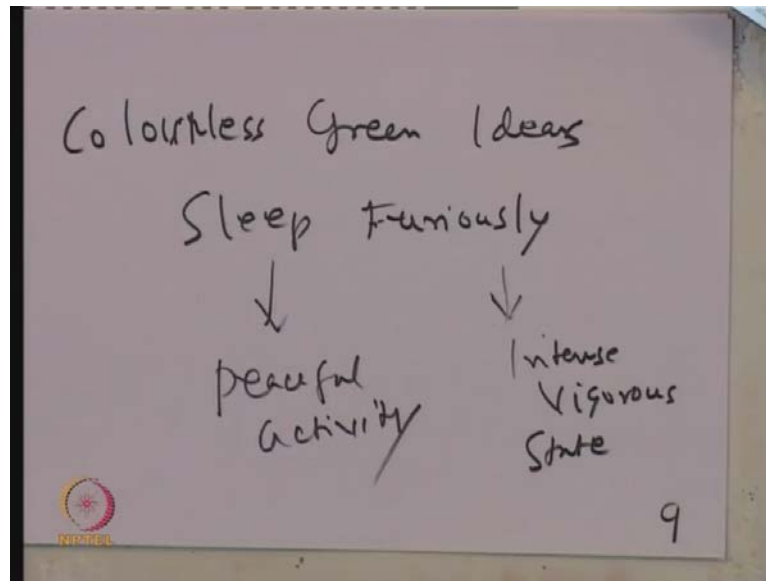
We start with words of a language the words of a text or sentence, so these words are placed one after the other to produce a meaningful sentence, sentences are produced one after the other to produce a meaningful paragraphs. Paragraphs are produced one after the other to produce meaningful chapters; chapters form a book and so on. So, gradually from words we go on building bigger and bigger textual units and ultimately produce a very large textual entity. So, when these words are taken up for processing what we have in front of us is a raw piece of text.

This piece of text is meaningful to a human begins on a expert in the field somebody who has world knowledge, somebody who has a knowledge of language can see that these symbols having meaning not so for a machine. A machine may not be equipped with so much of world knowledge so much of expressive language a machine need annotation. So, when we have a taken 3 words people laugh heartily we had to produce labels on them saying that people is a noun, people is a collective noun, people is a animate.

People takes s in the form of plural and all these and many other pieces of information can be put down on the text to enrich it, and these kind of enrichment of the text with labels is called annotation. So, I hope the notion of annotation is very clear to you, this is very critical for natural language processing you have to understand annotation very clearly. Let me make another point which is related to this and a part of great interest, who produces annotation? When annotation is done it is done by human beings people who understand the language and the people who understand the meaning of the words.

So, as a human being I produced N on people, I produce V on laugh, I produce A on heartily with the understanding that these words have noun role, verb role and adverb role respectively. How did I know that people is a animate? That is the world knowledge part. As a human being who is a part of society we know people are animate entities and not only that there are more complex issues we know that this is a meaningful sentence this laugh and heartily form a meaningful verb, adverb pair. So, there are it is possible to have some meaningless pairing this very famous sentence from which has been made immortal in natural language processing.

(Refer Slide Time: 42:22)

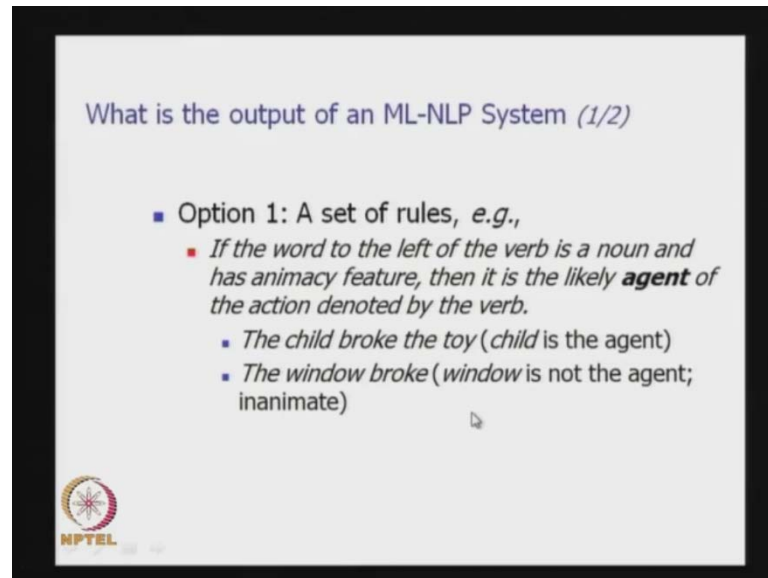


This sentence, colorless green ideas sleep furiously. This is a sentence which is very famous in NLP literature has lot of historic value. Now this sentence is change in many different ways, but one of those strange things about this sentence is this very unusual verb adverb combinations sleep and furiously. Sleep is a peaceful activity and furiously is a is an intense vigorous state and these 2 words are therefore, mutually incompatible.

So, if you put sleep and furiously together it is a strange combination and people would raise eyebrows looking at this particular sentence. So, people laugh heartily was not a strange sentence we made use of all word knowledge and also knowledge of the language knowledge of the properties of the words. And we could understand the meaning of this sentence annotated with proper labels so much for annotation.


Proceeding further, let us just remember what we say it for machine learning the constituents are goal classification of clustering, features or attributes are the clues for classification, values of the features they decide what the decision would be? Training data where the corpus is marked with different levels of training test data, the test corpus or the test text which are used for evaluating the algorithm. And there are accuracy of the classification in the form of precision, recall, F value, MAP Score and test of significance. We will have occasion to explain precision and recall which I will do after some time.

(Refer Slide Time: 44:56)



What is the output of an ML-NLP System (1/2)

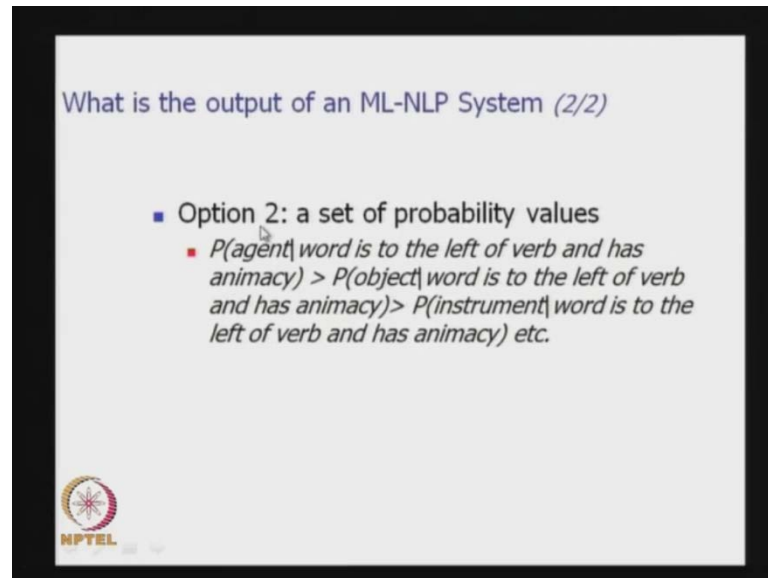
- Option 1: A set of rules, *e.g.*,
 - *If the word to the left of the verb is a noun and has animacy feature, then it is the likely **agent** of the action denoted by the verb.*
 - *The child broke the toy (child is the agent)*
 - *The window broke (window is not the agent; inanimate)*

 NPTEL

Now let us understand the output of machine learning N L P system. We have said that the statistical approach natural language processing makes use of statistical technique it makes use of classification algorithms to produce levels. And now let us understand when we apply machine learning algorithm on textual data what is the output we get from this machine learning system. The first option is we could get a set of rules, so the rule can be in this form if the word to the left of the word is a noun.

And has animacy feature that means the noun is animate then it is the likely agent of the action denoted by the verb. So, since people the word people was to the left of the word laugh and people is also animate that it is very likely the agent of the action denoted by the verb namely laugh. So, taking more examples the child broke the toy here child is the agent because child is to the left of break and it is also animate. In the sentence the window broke the window is not the agent because even though it is left of break it is in animate. So, both the conditions have to be satisfied the word has to be the left of the verb and it must have the animacy feature.

(Refer Slide Time: 46:39)



What is the output of an ML-NLP System (2/2)

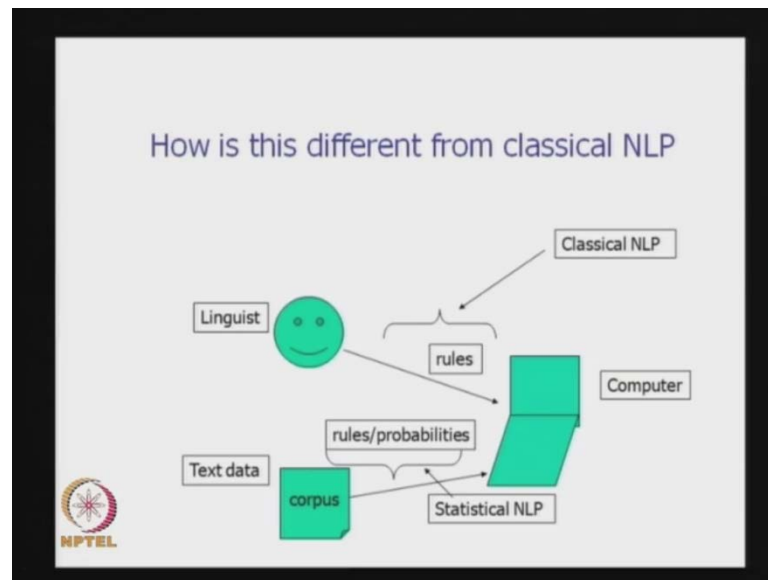
- Option 2: a set of probability values
 - $P(\text{agent} | \text{word is to the left of verb and has animacy}) > P(\text{object} | \text{word is to the left of verb and has animacy}) > P(\text{instrument} | \text{word is to the left of verb and has animacy}) \text{ etc.}$

NPTEL

There is another way the output of the machine learning N L P system can be produced. This is option number 2 and the output can be a set of probability values what we saw earlier was a rule and actual rule where as in this case the output is the set of probability values. So, it is expressed in this form the word is to the left of the word and has animacy. So, in this case the probability of the word being an agent is much more than the probability of the word being an object which again is more than probability of the word being and instrument.

So, everywhere you can see we are expressing the probability by means of a conditional expression. Probability agent given that word is the left of the word and has animacy, the word being an agent is the event we are considering. So, probability of being an agent is more than the probably of being an object is more than the probably of being an instrument. So, this is the way the probability value can be produced, this is to be contrasted with the output that we saw last time this was the rules.

(Refer Slide Time: 48:21)



Now, we will finish this lecture with a very quick remark on the difference between classical NLP and the statistical NLP. In classical NLP, we obtain the rules and these rules are embedded in the computer the rules come from the linguist, who is the human being. In statistical NLP this rules are the probabilities they come from the textual data namely the corpus and the machine works with those rules. So, in both cases there are rules and it could also probability values, but classical NLP they come from human beings and in statistical NLP they come from textual data by means of machine learning algorithm. We will make more insightful remarks on these things in the next lecture.