

Artificial Intelligence
Prof. Anupam Basu
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur
Lecture - 39
Natural Language Processing - I

In the last two lectures of this course I will be presenting to you one important aspect of intelligent behavior. A machine is considered to be really intelligent only when it can understand and interpret or speak for that matter natural language. Natural language is the language we speak day-to-day to ourselves. Different languages like English Hindi Bengali that has evolved through generations among the people through which they communicate is called natural language. And the capability to understand, interpret and communicate through natural language is a very important criterion of intelligent behavior.

(Refer Slide Time: 2:20)

The slide features a dark blue background with white and yellow text. At the top left is the IIT Kharagpur logo. The title 'Why Natural Language Processing?' is centered in white. Below the title, there are two main bullet points in yellow, each followed by a list of applications in white. The first bullet point is 'Huge amounts of data' with a sub-point 'Internet = at least 2.5 billion pages'. The second bullet point is 'Applications for processing large amounts of texts require NLP expertise'. The applications listed include: 'Classify text into categories', 'Index and search large texts', 'Automatic translation', 'Speech understanding' (with sub-point 'Understand phone conversations'), 'Information extraction' (with sub-point 'Extract useful information from resumes'), 'Automatic summarization', 'Question answering', 'Knowledge acquisition', and 'Text generations / dialogs'.

Here is an overview of natural language processing:

The first question that we need to answer is why we should at all study this? Why natural language processing is important?

If we look from the practical side then the huge amounts of data available in the internet for example at least 2.5bn pages are available and there is lot of information that is spread all over which is in natural language. So, if a computer has to learn it will have to utilize and understand such expressions which are available in the form of natural language. And there are applications for processing large amounts of text. We can see some of the typical applications here. There are so many text data available in the internet.

Is it possible to classify these texts automatically into categories?

For example, all the texts on electricity for school students. Can the computer program really understand the extent, the depth of different texts and identify that well? Like this is on electricity and these are the contents of this text or suitable for some school children. Indexing and searching large text we need to search the texts and also another very important application is automatic translation. We live in a multilingual world. So, if you consider the Indian scenario it is not the fact that all Indians specially those who are deprived of good school education that all of them understand English. So, in order that they can also benefit from some of the knowledge that is available in English it needs to be translated to their local language that they understand.

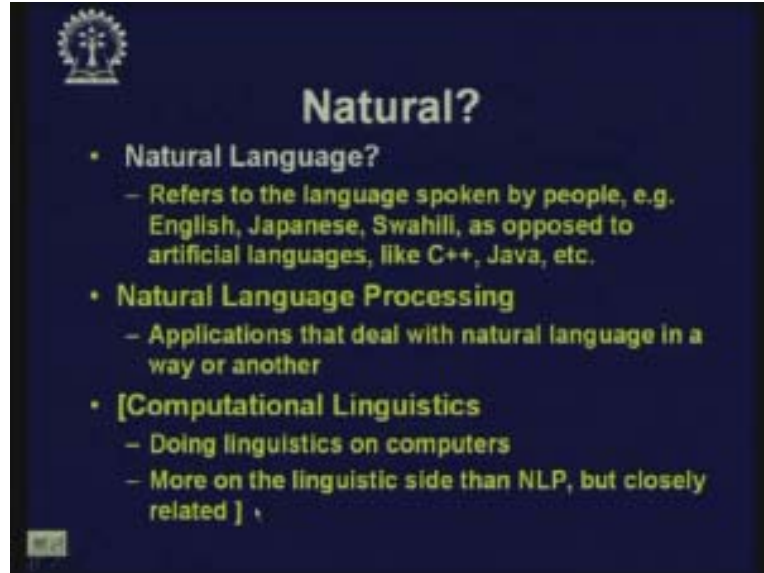
For India it is even more important because India is very much multilingual. Speech understanding is another important application where, here only one example is given; understanding telephone conversations but besides that there can be so many other applications. if we could have a computer which could understand the speech of a person and say in Hindi and that is automatically translated into text and is being recorded in the database of a computer then it will be very much for useful not only for telephone conversations but also many other applications.

Similarly, if a computer can understand the text that is written and speak it out then for those who cannot speak it will be a really great benefit and this system is known as a text to speech system. Similarly automatic speech recognition is a very important application. Now-a-days information extraction is gaining more and more importance so that one application can be that we extract useful information for the different resume that we are getting.

Automatic summarization of the text: Suppose the entire content of the book if it could be summarized in four or five pages then somebody can go to internet and have a brisk review of the basic content of the book before he decides to purchase that book or read it. Now-a-days also it is being done but it is done manually. If we had computers who could really understand natural language then this thing could be done in an automated way.

Question answering: is also another very important application. Similar other applications are knowledge acquisition. Acquiring knowledge from the experts, text generations, and dialogs and for all these applications we require natural language processing expertise.

(Refer Slide Time: 7:26)



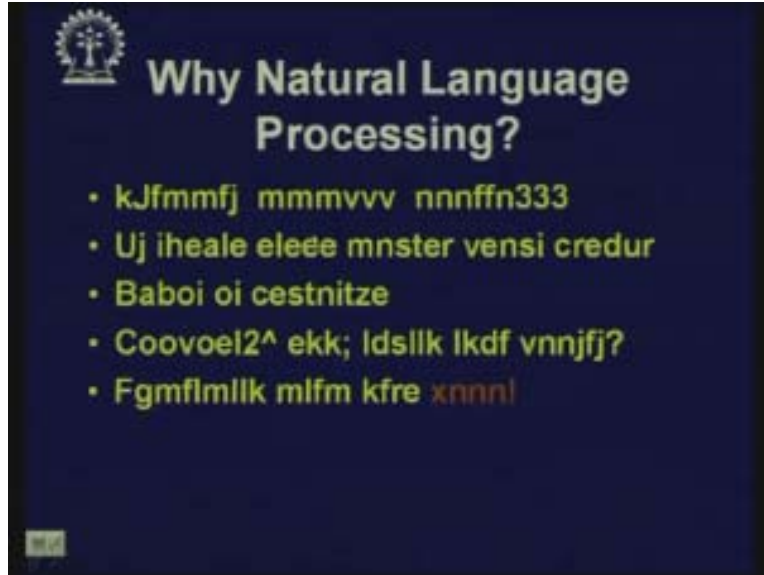
Why do we call some languages to be natural language?

These are the languages we use in our day-to-day life. It refers to the language spoken by the people naturally. For example, English, Japanese, Swahili is opposed to the artificial languages which also we also call formal languages like C++, java etc. These languages like C++, java etc have got a very strict syntax. If I say Tom go to school you will understand that I have made some grammatical mistake that instead of saying Tom goes to school I have said Tom go to school but the meaning will be immediately clear to you. Or if I say gives me the glass of water then also you will give me the glass of water although the sentence I have said is syntactically wrong. But such syntactically wrong statements will never be accepted in a formal language environment.

If we just a miss a semi colon in a c program they will be passing error. So, the formal languages are strictly guided by the rules that are there in grammar. In natural also we need the rule we have got the grammar but we are freer and minor syntactic errors does not deter the communication of the meaning or semantics. that is why we are interested in natural language processing and since in natural language communication we can take recourse to some ambiguous statements still human beings can perceive it, understand it and can act accordingly and because of this flexibility natural language processing all the more difficult than formal languages.

Natural language process is required for applications that deal with natural language in some way or other. On the other hand, we often hear the term computational linguistics. Computational linguistics means carrying out the linguistic studies. It is studying the different features of a language using computers. There is a subtle difference between these two. This is more on the linguistic side than on natural language processing but obviously they are very closely related. Now, if we continue our discussion on why natural language processing let us look at some of these sentences.

(Refer Slide Time: 10:00)



If we look at these sentences they are not carrying in any meaning to you at all. These are mere symbols which are not interpreted as yet. Now, when we feed such set of symbols to a computer the computer will not understand anything. However, whatever data we feed to the computer you know are symbols and ultimately they get converted into ASCII characters and the computer interprets that. However, in a C language whenever we type in the C program what goes in is a string of ASCII characters. But in that case we have got a program C compiler that is based on the strict grammar of the C language which can interpret these strings of symbols. So, using that compiler based on that grammar we have imparted to the computer the knowledge of the C language.

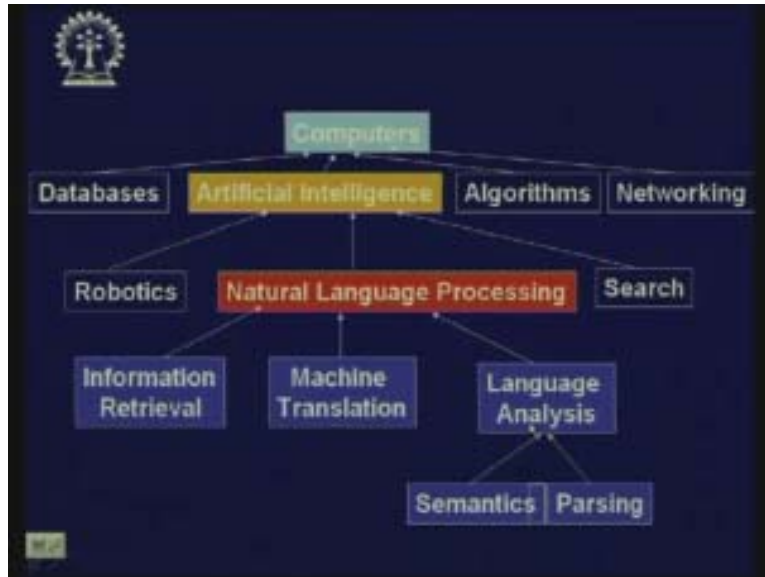
Similarly, computers by itself lack the knowledge so they see the text in English the same way you have seen in the previous text, nothing but a bunch of symbols and what is required is interpretation. However people have no trouble in understanding the language, we use common sense knowledge. For example, when this was shown you said this is all garbage. Might be if this statement is from some tribal language I do not know then they would have immediately understood that because of their common sense reasoning and because of their knowledge about that language.

Similarly, if you show me a French statement and if I cannot recognize that I will not be able to understand it. So what is required is essentially common sense knowledge, reasoning capacity and experience. And computers have by itself no common sense knowledge and no reasoning capacity unless we teach them. That is the overall scenario so that makes a natural language processing a little difficult. Now, if we look at the different areas of Computer Science research you will see that Computer Science deals with database, algorithms, networking etc.

Artificial Intelligence has got several components in it like robotics, search and expert systems that we have seen and natural language processing is one very important area of

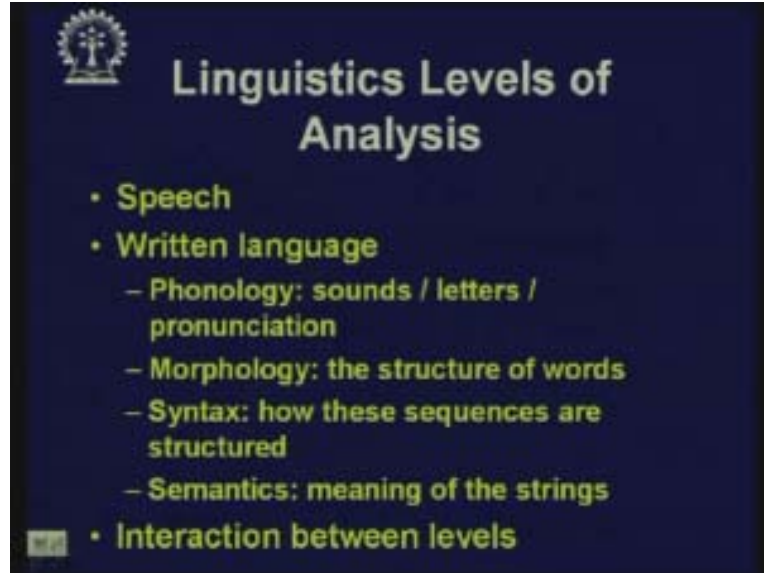
Artificial Intelligence. And why is natural language processing important? It is important because of many important applications like information retrieval, machine translation, language analysis etc. When we carry out language analysis that again includes understanding the semantics of language of any text that is written.

(Refer Slide Time: 14:00)



Parsing: Just as we write compilers to parse C statements similarly we will see the difficulties that are inherently there for parsing natural language sentences. Let us look into the problems of natural language processing and some approaches that are adopted. Information retrieval, machine translation and understanding the language together form a part of natural language processing. When we look at natural language processing there are different levels in which we can think of it.

(Refer Slide Time: 7:26)



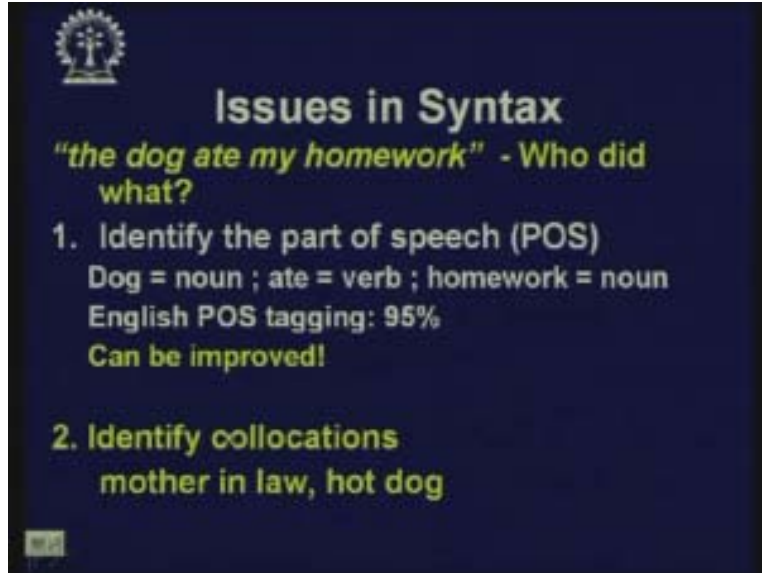
We can say that these are the different linguistic levels of analysis. Speech is one level and is at the top level. If we can generate text from **spoken speech** then all these utterances the waveforms that we generate when we talk have to be interpreted and captured and translated in the form of written text. That is one level. Even after that we have got the written level. And in the written language there is again different components phonology.

What is phonology?

Phonology deals with sounds, letters, pronunciation etc. When we say god or got the meaning totally changes because of the last phoneme. Instead of pronouncing d we are pronouncing t and the entire meaning changes. The other thing is morphology. That is the structure of the words. It is like child becomes children when we make it plural. But book becomes books. So, based on this we have to also understand that this morphological change, morphology is the structure the look it has changed and with that change the meaning has also changed. Therefore this is also another level.

Syntax: Syntax is essentially dealing with grammar. How these sequences are structured? And the semantics deals with the meaning of the strings. At the ultimate level of understanding we have to understand the meaning of the strings. The syntax is telling us what, whether those strings the symbol strings are well formed. We have to categorize different components of those symbol strings when we do syntactic analysis. That also helps in understanding the meaning. But with that we will understand some part of the meaning. Therefore we have two levels at the abstract level speech and written language. And inside the written language there are the chronological aspects, morphological aspects, syntactic aspects and semantics aspects and there is a continuous interaction among all these levels.

(Refer Slide Time: 18:19)



Issues in Syntax

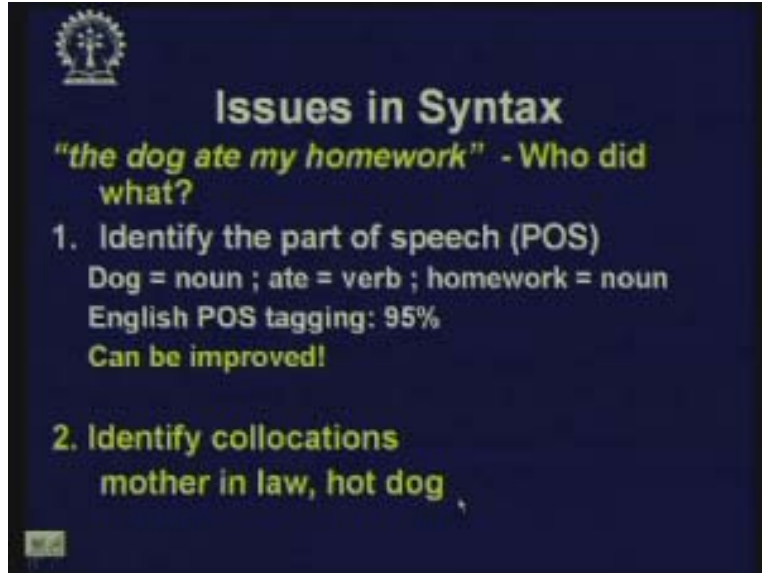
"the dog ate my homework" - Who did what?

- 1. Identify the part of speech (POS)**
Dog = noun ; ate = verb ; homework = noun
English POS tagging: 95%
Can be improved!
- 2. Identify collocations**
mother in law, hot dog

Issues in syntax: We take a very funny statement: The dog ate my homework. Now I am sure all of you will laugh at this sentence the dog ate my homework. What does it mean? But syntax per say is not dealing with the meaning of the sentence. It is trying to see whether this sentence is syntactically or grammatically well formed. The meaning may be useless here but if it is a correct syntax then it should allow me to understand the answer. Questions like who did what? Who ate? The dog ate, the dog ate what? The dog ate homework, whose homework? My homework. So, in all these you can see that the components of the sentence have got a relationship among themselves and that relationship is preserved in this sentence structure so strictly speaking this sentence is syntactically correct.

So, when we perform syntactic analysis one very important part is to identify the parts of speech of a sentence. Parts of speech is whenever we talk about a sentence then we talk about noun, verb, pronoun, adjective, adverb etc all these are known as part of speech and they identify the relationship or the role of that word in that particular sentence. For example, in this case the dog is noun, ate is verb, homework is noun. Now if I could write a program that would accept a sentence and will be able to level each component of the sentence with proper part of speech then that program will be called a part of speech tagger, POS tagger or part of speech tagging.

(Refer Slide Time: 20:26)



Issues in Syntax

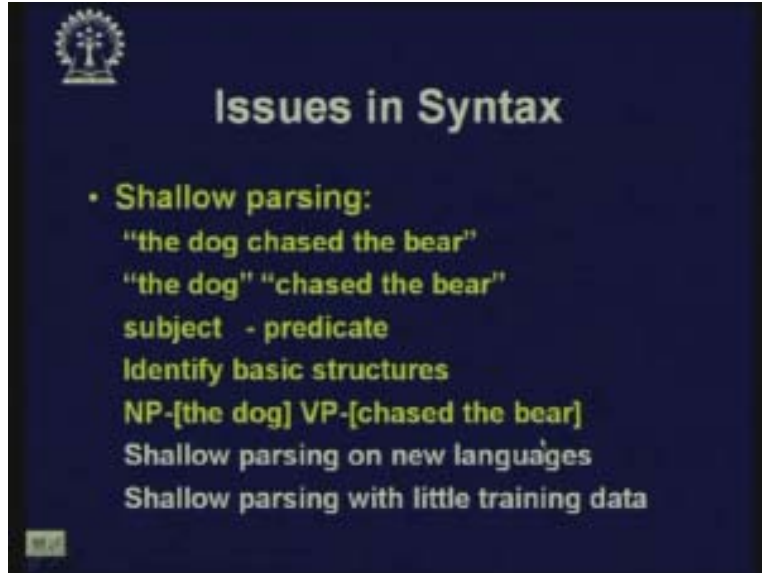
"the dog ate my homework" - Who did what?

- 1. Identify the part of speech (POS)**
Dog = noun ; ate = verb ; homework = noun
English POS tagging: 95%
Can be improved!
- 2. Identify collocations**
mother in law, hot dog

In English part of speech taggers exist with around 95% accuracy but its work is still going on where it can be improved. For Indian languages also now-a-days we are working on part of speech tagging and here also we are getting more than 95% accuracy by now. Also, the words should not always be looked in isolation, in order to understand meaning we may sometimes require to look at group of words.

For example, if there are three words mother-in-law then we cannot just look at mother and try to find out its part of speech mother is a preposition and law is another noun. But in order to understand mother-in-law is actually mapping into another concept. Dog means something that is an animal, hot has got a different concept but when in a sentence these two words are put side by side when they are collocated then these should be looked upon together because hot dog is a particular type of fast food. So, in order to understand the sentences we have to understand the parts of speech. But that is not the only thing and there are many other issues involved; one is collocation and this identifying of collocation is also known as chunking or local word grouping. There are more issues in syntax.

(Refer Slide Time: 22:57)

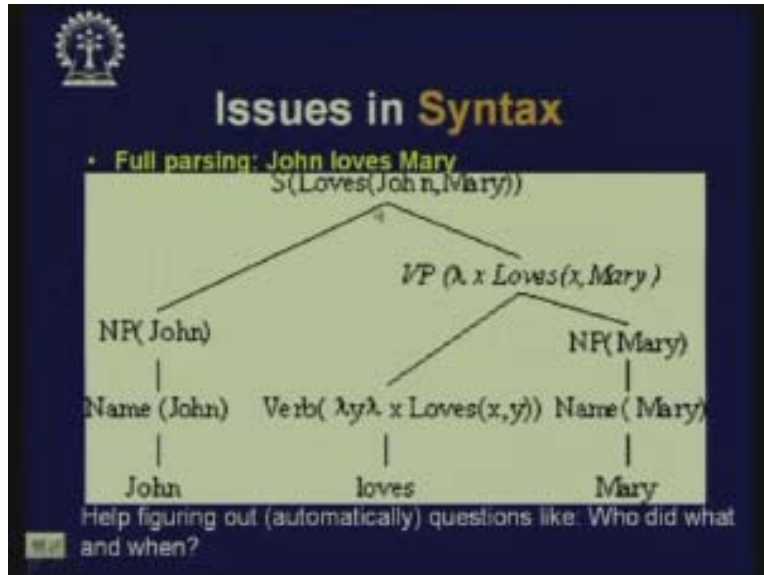


The slide features a dark blue background with a white logo in the top left corner. The title 'Issues in Syntax' is centered at the top in a white, sans-serif font. Below the title, a list of bullet points is presented in a yellow, sans-serif font. The first bullet point is 'Shallow parsing:', followed by the sentence 'the dog chased the bear' in quotes. The next line shows 'the dog' and 'chased the bear' in quotes, separated by a space. Below this, the words 'subject' and 'predicate' are listed with a hyphen between them. The following line says 'Identify basic structures'. The next line shows 'NP-[the dog]' and 'VP-[chased the bear]' in brackets. The final two lines are 'Shallow parsing on new languages' and 'Shallow parsing with little training data'. A small white box with the number '14' is in the bottom left corner.

- **Shallow parsing:**
"the dog chased the bear"
"the dog" "chased the bear"
subject - predicate
Identify basic structures
NP-[the dog] VP-[chased the bear]
Shallow parsing on new languages
Shallow parsing with little training data

We would like to do a shallow parsing where we just like to find out which one is the subject, which one is the noun phrase etc. So what we can do is, the dog chased the bear this is a sentence now when this sentence is shallow parsed we will first break up this sentence and identify that the dog is the subject and chased the bear is the predicate. And then the basic structure of this subject is that it is a noun phrase the dog and predicate is a verb phrase chase the bear. Shallow parsing is often adopted when we try to understand a new language and now the works are on, in order to have a quick and good parsing one very important approach is again a statistical approach where we need a lot of training data to understand that given a word what is its probability of being a noun, what is its probability of mapping it to a particular concept.

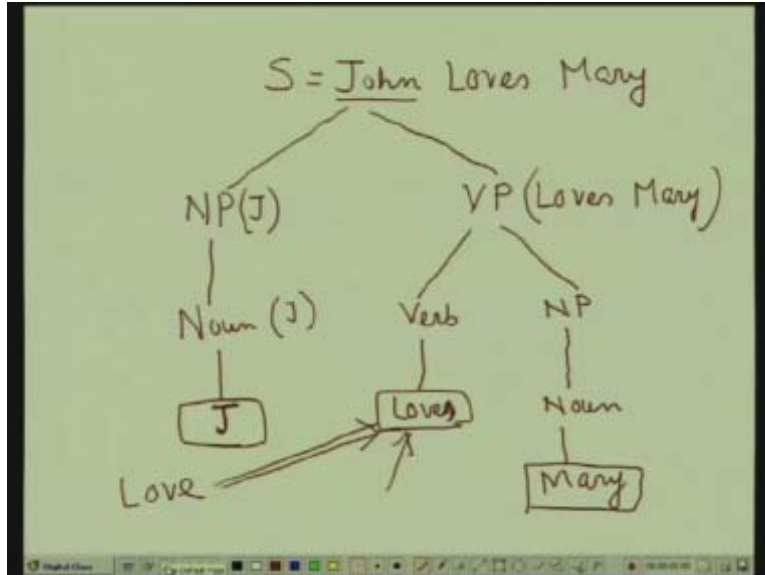
(Refer Slide Time: 24:45)



Here is an example of parsing of a sentence John loves Mary. My sentence with which we start is John loves Mary. Now, when we parse this sentence first thing is I will break it up as a noun phrase. The parse will break it up as a noun phrase and a verb phrase. And this noun phrase is this part John and the verb phrase is loves Mary. And this noun phrase is further broken down as a noun and that noun is John. So that is coming to the terminal symbol John and the verb parse can be further broken down into the verb and the noun phrase. And this verb is loves and the noun phrase is the noun which is Mary. So this is the full parse tree and this is a terminal symbol, this is a terminal symbol, and this is a terminal symbol. So, this is a parse tree.

Now one interesting thing is, look at this loves, we know the meaning of love as a verb. Now an intelligent system should be able to understand that when it is loves it is essentially a morphological change of this same verb love and because it is a third person singular number this addition of s has come and these two are basically [..... 27:02].

(Refer Slide Time: 27:07)



Here is another example of similar parts: The dog or the boy went to school. Now how do you parse it?

Again it will be broken down to noun the phrase the boy and it will be a verb phrase which will be went to school. And the boy can be further broken down as a determiner or article and a noun. And this determiner is the terminal and the noun is boy. The verb phrase can be broken down to a verb which is went and it will have another noun phrase which is this two school and this noun phrase can be further broken down as a prepositional phrase and a noun. And this noun is terminal school and this preposition phrase has got a preposition to. This is how we parse it but still there are many issues.

For example, again let us see went, what is this went? The went is the same thing as go in the past tense that has to be known that knowledge is required in order to understand this sentence. But by this parsing what have you got? We have been able to understand the role of each of these words that the boy is a noun phrase so that is the subject of the sentence. This went is the verb. So parsing is one, this helps in automatically figuring out questions like who did what and when. If we ask John loves Mary then if we ask the question who loves Mary from this parsing we can straight away come to the subject and say John loves Mary. There are more issues in syntax. One interesting thing is, this is a very interesting example anaphora resolution. We always have this sort of sentences; the dog entered my room it scared me.

(Refer Slide Time: 30:17)

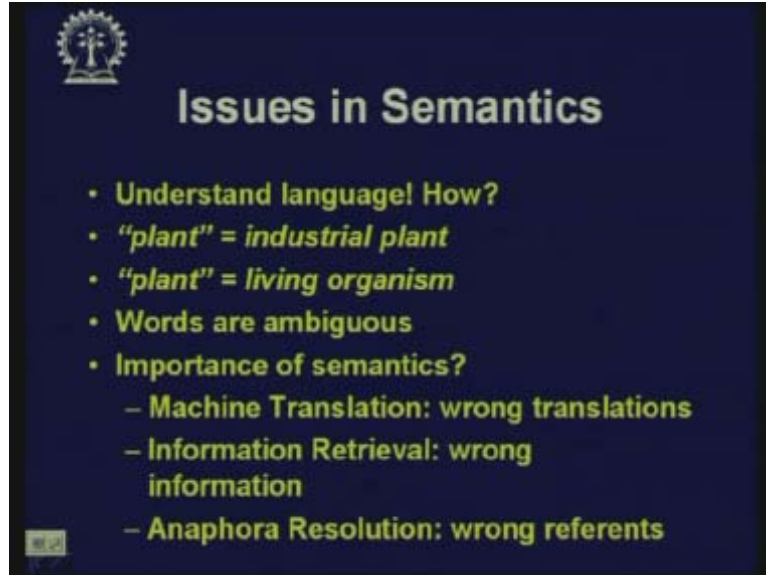


The slide features a dark blue background with a white logo in the top left corner. The title 'More Issues in Syntax' is centered at the top in a white, sans-serif font. Below the title, there are two bullet points in a yellow-green font. The first bullet point is 'Anaphora Resolution:' followed by the sentence 'The dog entered my room. It scared me' with 'dog' and 'It' underlined. The second bullet point is 'Preposition Attachment' followed by the sentence 'I saw the man in the park with a telescope' with 'with' underlined. A small white icon is visible in the bottom left corner of the slide.

Now when you look at it which is a pronoun, all of you know that a pronoun stands for a noun. In this case if I ask the question who scared you or who scared me the system should not answer it, the system should answer the dog. So the system must understand that this pronoun is actually related to this noun the dog. So these sorts of problems are called the anaphora and the second one is a more interesting problem; I saw the man in the park with a telescope. There is a lot of problem here. I saw the man in the park with a telescope. I can interpret it in different ways that I could see the man because I had a telescope I saw with a telescope and I saw the man in the park. Or this could be related to the man that I saw the man with a telescope in the part, a man was carrying the telescope.

Or the other thing is, I saw the man in the park which had a telescope. These are the three possibilities. Now which one of these is really meant? Even for us we can understand that it is very improbable to have a park with a telescope but it is possible that you can have in a special park for some astronomic fair you can put in a telescope that is also possible. But probably I saw the man with a telescope I saw with a telescope the man in the park, so I used a telescope to see the man or the other possibility was that I saw the man was carrying a telescope. Therefore the scope of lot of ambiguity is here. Now similarly there are several issues in the semantics. The earlier one was syntax. But even if we resolve some parts of the syntax there are several issues in semantics.

(Refer Slide Time: 32:57)



For example, semantics means understanding the meaning of the sentence. How do we understand the meaning of the sentence? The same word plant can mean an industrial plant, a chemical plant, electrical plant or whatever. Plant is also a living organism. So often the words we use are ambiguous.

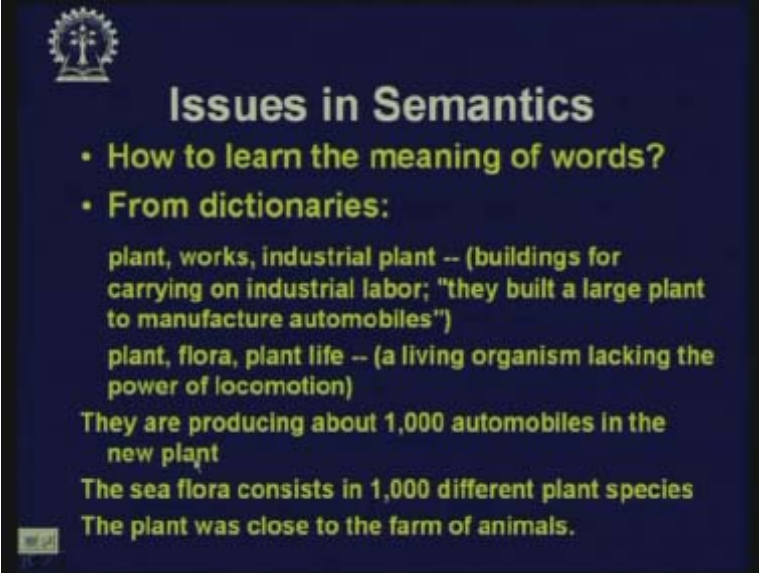
What is the importance of semantics?

Unless we really understand the meaning we will not be able to translate them there will be wrong translations. If we try to carryout information retrieval there will be wrong information that will be extracting. And if we cannot do anaphora resolution properly then we will also have wrong reference.

Now how do we learn the meaning of the words?

One thing is, from the dictionaries we can have the plants, plant works, industrial plants all these can be putting in as buildings for carrying on industrial labor. They built a large plant to manufacture automobiles. Therefore these sort of sentences or clauses we have will map to this. But again there is an entry in the dictionary plant, flora, plant life, a living organism lacking power of locomotion is a plant. Now when we state a sentence they are producing about 1,000 automobiles in the new plant then which one of these my meaning would my system choose?

(Refer Slide Time: 34:00)



Issues in Semantics

- How to learn the meaning of words?
- From dictionaries:
 - plant, works, industrial plant -- (buildings for carrying on industrial labor; "they built a large plant to manufacture automobiles")
 - plant, flora, plant life -- (a living organism lacking the power of locomotion)

They are producing about 1,000 automobiles in the new plant

The sea flora consists in 1,000 different plant species

The plant was close to the farm of animals.

I have to resolve that and I have to really understand that since I am talking of automobiles and automobiles require some industrial labor and a large plant obviously this new plant means an industrial plant where cars are being manufactured. I will obviously not interpret it as a plant in the form of a small tree. Again for this sentence the sea flora consists in 1,000 different plant species. Now we have to understand that these plant species is not the industrial plant but this sense of the plant.

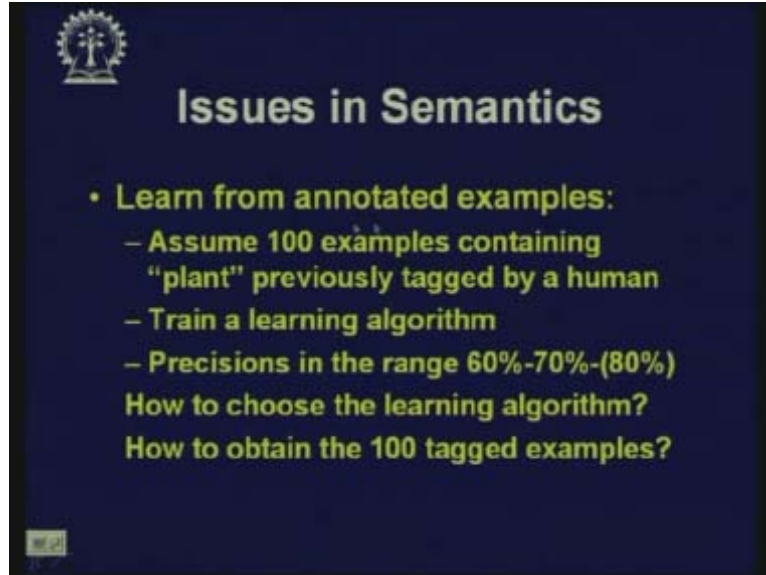
Now what happens when the plant was close to the farm of animals?

It may be an industrial plant or I am talking about a particular plant. Therefore gradually the context is also becoming important. And when I do parsing merely the role of the sentence is not enough. The components of the sentences the part of speech is not enough.

So how do we really understand meanings?

There are several issues which are involved and it is rather complex. One approach as I said which is very popular now-a-days is the statistical approach. There we start with a set of examples which are annotated. That means in a sentence like this; they are producing about 1,000 automobiles in the new plant.

(Refer Slide Time: 24:45)



In this sentence I manually annotate it and say the meaning of this plant is actually industrial plant. I annotate many such sentences. And a system may be designed using a number of tools like the Hidden Markov Model and other technology. There is technology available using which the system can learn the probability of a meaning given a particular structure of a sentence.

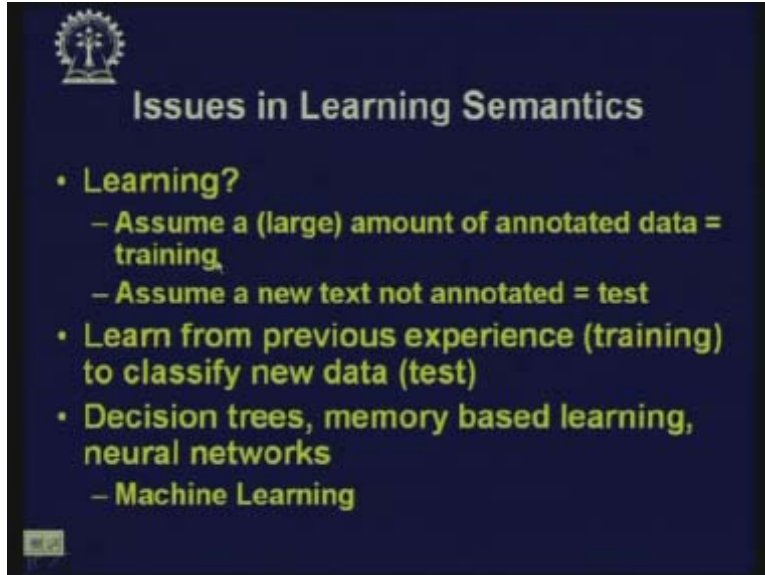
Suppose we take hundred examples containing the word plant which are previously tagged by a human with the meanings then we frame a learning algorithm. there are several learning algorithms including the neural net algorithms there are different classification algorithm, there are Markov Models, Hidden Markov Model algorithms etc where there are different approaches. So we will train a learning algorithm and using that algorithm when new sentences are fed we expect that this algorithm will try to find the closest meaning statistically. Right now the precision in this is a little more than 80% than lot of research is going on in this area and how to choose a learning algorithm is an issue.

One thing you must realize is that this sort of approach requires human tagging so you cannot have any soft path to success. There is a necessity of putting in the labor of tagging them properly at least a small set on which we can later on put **scrap** but we will have to make a small set properly annotated on which we can run our statistical learning algorithms.

Another issue is, how to obtain the hundred tagged examples?

We were talking of using learning algorithms where we learnt the semantics. Now what are the issues in learning the semantics? We can assume a large amount of annotated data for training.

(Refer Slide Time: 39:28)

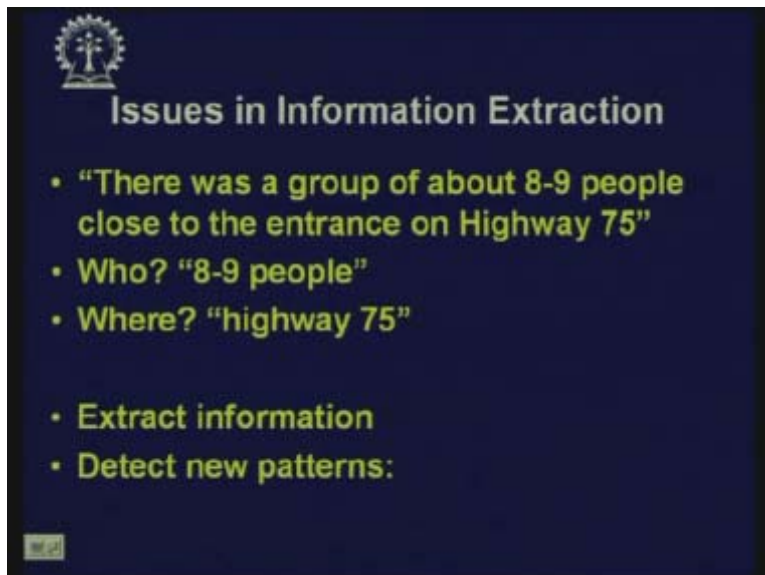


The slide features a dark blue background with a white logo in the top left corner. The title 'Issues in Learning Semantics' is centered at the top in a light blue font. Below the title, there are four main bullet points in yellow, each with a sub-bullet point. The sub-bullet points are indented and also in yellow. A small white icon is visible in the bottom left corner of the slide.

- Learning?
 - Assume a (large) amount of annotated data = training
 - Assume a new text not annotated = test
- Learn from previous experience (training) to classify new data (test)
- Decision trees, memory based learning, neural networks
 - Machine Learning

We train our algorithm and then we take a set of new texts which will be used for the test. So our approach is learning from the previous experience training to classify the new data. The test data will be annotated automatically based on whatever our system has learnt from the corpus of already annotated data. So there are different algorithms for that like decision trees, memory based learning, neural networks etc.

(Refer Slide Time: 40:23)



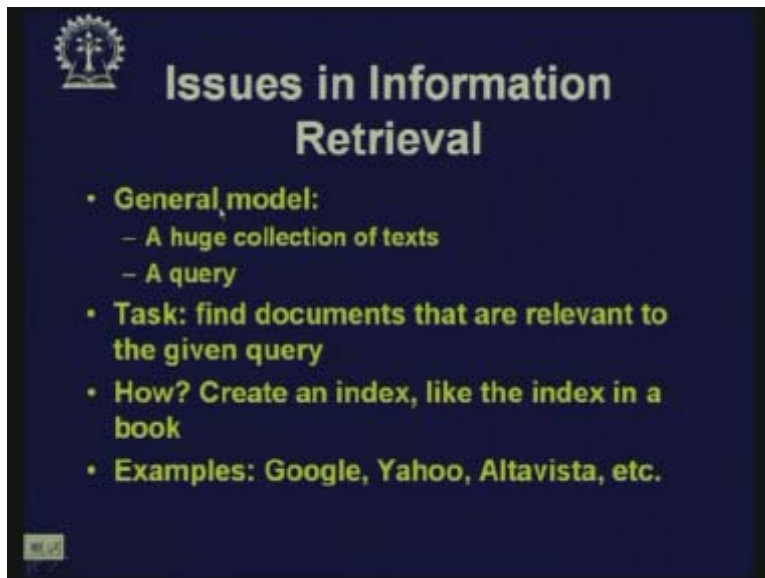
The slide features a dark blue background with a white logo in the top left corner. The title 'Issues in Information Extraction' is centered at the top in a light blue font. Below the title, there are five main bullet points in yellow. The first three points are related to a specific sentence, and the last two are general tasks. A small white icon is visible in the bottom left corner of the slide.

- “There was a group of about 8-9 people close to the entrance on Highway 75”
- Who? “8-9 people”
- Where? “highway 75”
- Extract information
- Detect new patterns:

There are several other issues in information extraction. Here you would learn the different aspects of natural language processing which are very useful now in the present day technology. We are given a sentence; there was a group of about 8 to 9 people close

to the entrance on highway 75. Now, from this a computer program should be able to extract the information like the answer to the questions who? 8 to 9 people, where? On highway 75. So this sort of answers should be extracted from this sentence where it is not just a very straight way case, where? Close to the entrance on highway 75 extract the information and also it requires detecting new patterns. There are issues in information retrieval as well.

(Refer Slide Time: 41:40)



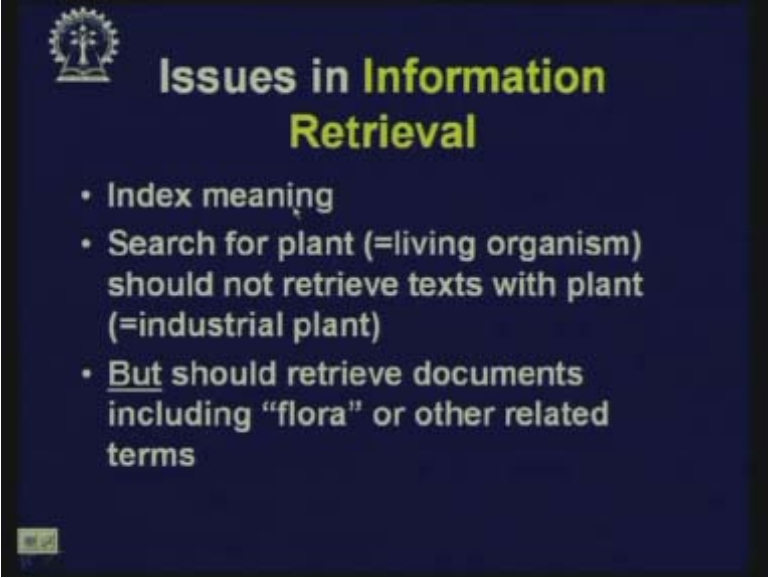
From the huge bank of information we have in the internet how we extract information that is relevant is an issue to think of. General model of information retrieval is that we start with a huge collection of texts and we have got a query and our task is to find the documents that are relevant to the given query. We have to find out the documents which are relevant to the given query. There are different measures to really compute which are relevant.

For example, there are texts on cricket and I want to find out the texts only which talk about Sachin Tendulkar. And because my query is Sachin Tendulkar in that case out of all these texts that are there I have to evaluate or rank all these texts based on the keyword of my search which is Sachin Tendulkar. A very common sense approach would be the text which has got larger number of occurrences of the keyword Sachin Tendulkar will be more relevant. May be there is a text in which a lot of things about cricket has been said and Sachin Tendulkar's name is mentioned only twice and it is talking about many other thing so that particular text will not be that relevant for me.

Therefore this is a very simplistic approach that we select the keyword and based on that we find out which one is relevant. So we usually find out the documents through creating the index like the index in a book and application of such information retrieval we will find in most of the search engines like Google, Yahoo, AltaVista etc. The main issues in

information retrieval are, we have to index the meaning because as we have seen one word can have multiple meanings.

(Refer Slide Time: 45:45)



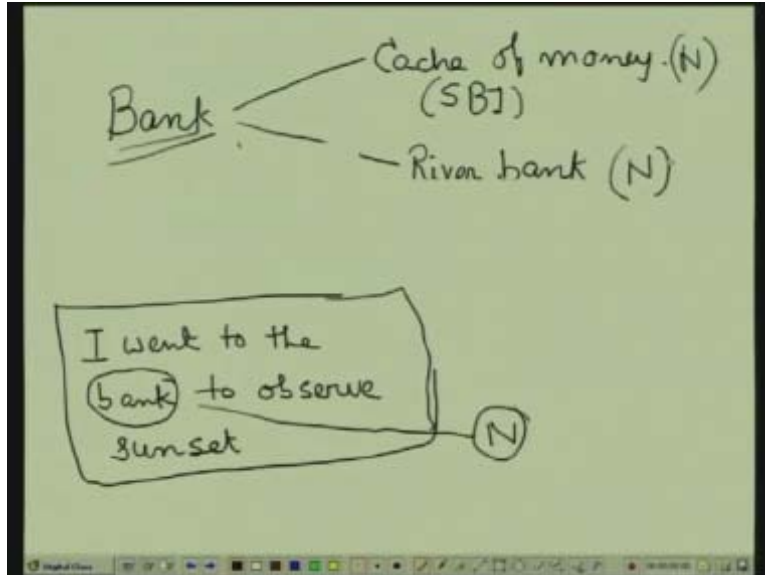
The slide features a dark blue background with a small logo in the top left corner. The title 'Issues in Information Retrieval' is written in a bold, yellow-green font. Below the title, there is a bulleted list of three items in white text. The first item is 'Index meaning'. The second item is 'Search for plant (=living organism) should not retrieve texts with plant (=industrial plant)'. The third item is 'But should retrieve documents including "flora" or other related terms'. A small green icon is visible in the bottom left corner of the slide.

- Index meaning
- Search for plant (=living organism) should not retrieve texts with plant (=industrial plant)
- But should retrieve documents including "flora" or other related terms

In that case which meaning should I assume this word to refer to? For example, if I say I went to the bank to see the sunset, bank may be the typical bank where we keep money and bank can be a river bank also. So this is one of ambiguities which is known as poly semi. Poly semi means when a particular word can have multiple meanings.

Consider the word 'bank'. You can see that we will have different meanings like bank can be cache of money like the State Bank of India, another thing can be a river bank, another thing can be bank can mean depend for example I banked on him. When I am given the sentence I went to the bank to observe sunset then this is my sentence which my machine is trying to understand. Now you can see that just by syntactic analysis through the normal parsing I will find out that this bank is basically a noun whereas where this meaning depend is a verb so this meaning is not applicable. Therefore here you can find that by syntactic analysis itself without going into any semantics analysis we have been able to eliminate some of the meanings which are not applicable. Therefore here we eliminate this, I eliminate is part, still I am left with a problem of two possible meanings that both of them are noun, this is noun, this is also noun and this part of speech is also noun.

(Refer Slide Time: 47:43)



Now which of these meanings should I take?

Here I need some more knowledge in order to solve this problem and that knowledge will be from the context that sunset is usually observed near a river or sea. So the probability of this bank being a river bank is higher in this context because this bank is associated with this phrase to observe sunset. If the sentence was I went to the bank to withdraw money then obviously if this sentence was not there. And instead if I had written I went to the bank to withdraw money then in that case this bank would have related to this meaning because of the association.

One knowledge representation structure that we have already learnt comes in very handy over here that is a semantic net. The semantic net represents the meaning. Therefore if in our knowledge base we have our semantic net then in that case we can handle these sort of situations. For example, I can have, bank is an institution which stores money, bank is close to river, bank is visited by people, watch sunset. Say I have got this sort of structure; I went to the bank to withdraw money.

When I try to understand the meaning of this sentence I can utilize this semantic structure and I will find that there is a better relevance because it is leading me to the concept money which is already over here and so I will take that meaning of bank in order to interpret the sentence. So this sort of knowledge is required for retrieving information from the retrieving information from the **web**. There are other issues like this same thing can be very useful for question and answering.

For example, what is the height of Mount Everest?

11,000 ft. **Now in the current state of the art of information retrieval accurately precision and recall are the two very important measures which is around 40 to 50%.** There is a lot of scope of research and lot of scope of work in this area.

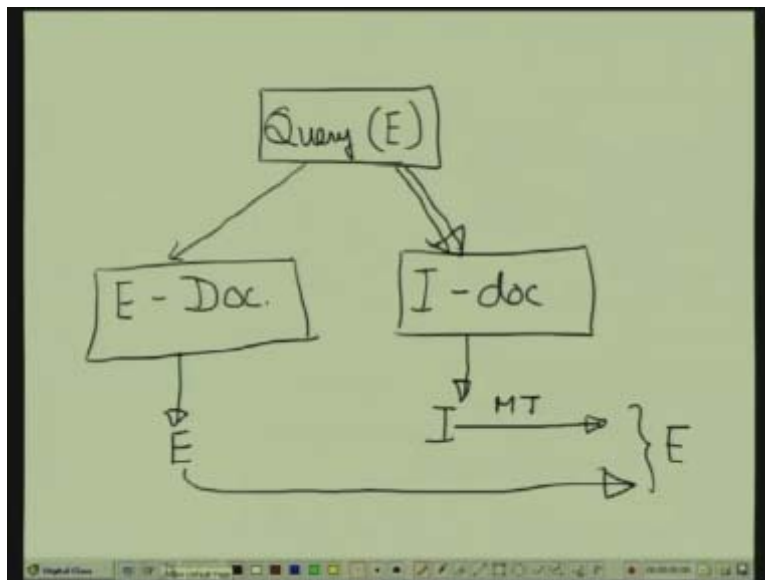
We improve the precision with the use of more common sense knowledge and often we capture this common sense knowledge in the form of knowledge representation. There are other issues like cross language information retrieval. That means we want to find the information across languages.

For example, what is the minimum requirement for car rental in Italy?

Now with this query when I search the documents that are available in the web it is not sufficient to restrict myself to the documents which are written in English only because since I want to get the information about some law in Italy there may be some documents written in the Italian language which may come in very handy for me. But I understand English and my query has been given in English. Therefore what is needed is that I will accept the query in English and I will go to the search documents which are English documents. I will also search Italian documents.

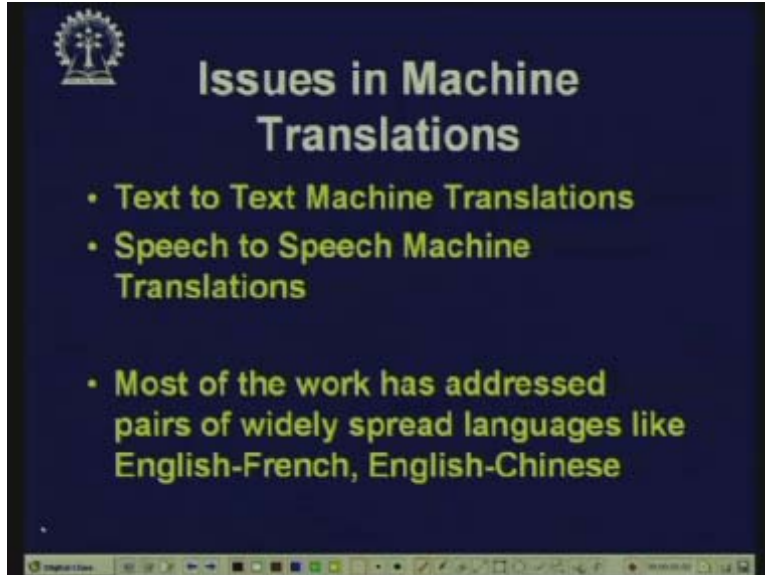
Now the query is in English so I need to have some sort of translation at this point or I must use some other technique. So, from here I will be getting some documents which are Italian documents and from here I will be getting some English documents. Now my answer should be these English documents and these Italian documents will require to be translated MT means machine translation and ultimately I want to have the answers in English because I know only English. This is cross language information retrieval. So we will have to convert this to some Italian text in order to do this sort of translation.

(Refer Slide Time: 24:45)



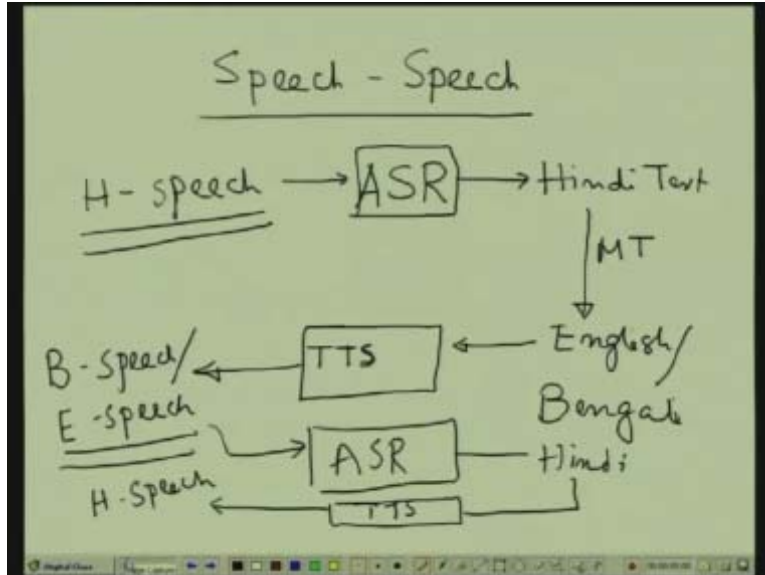
Issues in machine translations: There are text to text machine translations. That is, given English I want to get Hindi or given Hindi I want to get that document in English or speech to speech machine translation that is the ultimate dream that all of us cherish specially in a country like India. For example, speech to speech is ideal but still we are still quite far from reaching the desired goal here.

(Refer Slide Time: 54:47)



Here I accept some Hindi in the speech and that Hindi speech is recognized by automatic speech recognizer ASR and from there I get Hindi text. And that Hindi text can be translated by some machine translation to English and then using some Text To Speech system TTS English or may be other Indian languages like Bengali and there is a Text To Speech system TTS that will speak out this English of the Bengali sentence to another person so I am getting Bengali speech or English speech. So you just try to imagine if these were possible then in that case a person speaking Hindi would be able to communicate with the person speaking English and vice versa. That is, when he speaks English then it goes to an English ASR and I get a Hindi and that Hindi is again converted to a Hindi TTS and that can be spoken out as Hindi speech. You can see the advantage of this.

(Refer Slide Time: 56:45)



Therefore we needed automatic speech recognition, we needed text to speech, we needed machine translation and all these are in the domain of natural language processing and our present day challenges.