

**Natural language processing**  
**Prof. Pushpak Bhattacharyya**  
**Department of Computer Science and Engineering**  
**Indian Institution of Technology, Bombay**

**Lecture - 1**  
**Introduction**

This is Pushpak Bhattacharyya, professor of Computer Science and Engineering at IIT Bombay, delivering a course on natural language processing. Natural language processing is a very important topic in today's world of internet. In the, in this age there is lot of information on the web in the form of text. It is a very important concern in today's world to obtain information from this text and use it for various purposes. This is the motivation for understanding natural language processing, its tools, techniques, philosophy and principle, let us go ahead.

(Refer Slide Time: 01:04)



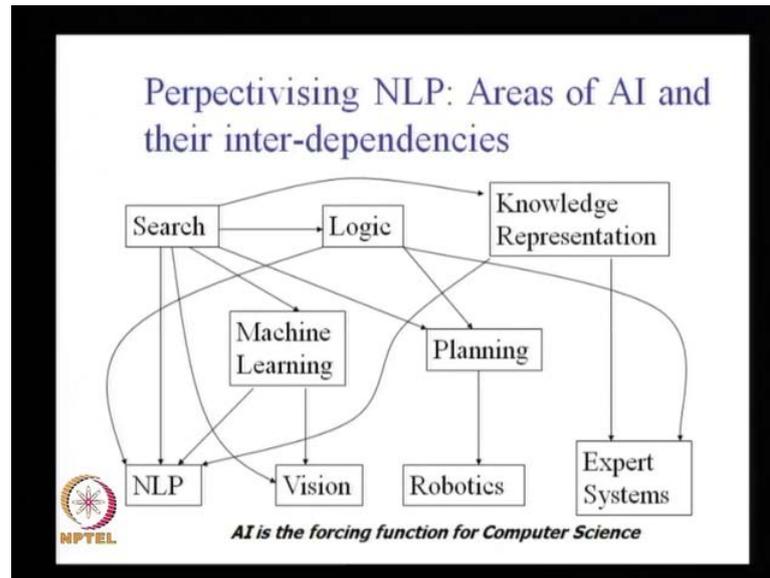
**Persons involved**

- Faculty instructor: *Dr. Pushpak Bhattacharyya* ([www.cse.iitb.ac.in/~pb](http://www.cse.iitb.ac.in/~pb))
  - Areas of Expertise: Natural Language Processing, Machine Learning
- Course home page (to be created)
  - [www.cdeep.iitb.ac.in/~nlp-2010](http://www.cdeep.iitb.ac.in/~nlp-2010)



The course instructor is myself, Doctor Pushpak Bhattacharyya. My home page is [www.cse.iitb.ac.in](http://www.cse.iitb.ac.in/~pb) tilde p b. Areas of expertise are natural language processing and Machine learning. Course home page, which will be created, is slightly to be under [www.cdeep.iitb.ac.in](http://www.cdeep.iitb.ac.in/~nlp-2010) slash tilde n l p 2010.

(Refer Slide Time: 01:40)



Moving forward, we give a perspective on natural language processing, different areas of artificial intelligence and their inter dependencies. If you look at this diagram there are this three layers: Search, logic and knowledge representation, forms the 1st layer. The next layer is: Machine learning and planning. The final layer is: Natural language processing, vision, robotics and expert systems. I would like to spend some time understanding these areas and their inter relationships. Before that, let me talk about the importance of artificial intelligence in computer science and engineering. Artificial intelligence is called the forcing function for computer science.

Computer science has grown, by leaps and bounds in recent years and one of the reasons for that has been artificial intelligence. Artificial intelligence has always pushed in the boundary of computer science and engineering by demanding more and more from the machine. It is understood that, the machine should come closer and closer to human beings in terms of its usage and its application. If we look at the lower most layer in the transparency, if you see this here, natural language processing forms the left most corner block, followed by vision, computer vision then robotics then expert systems. I would like to make some remarks on this.

Natural language processing is concerned with, the computer being able to process human light languages like: English, French, Marathi, Hindi and so on. In computer vision the machine processes seen and understands, how to operate in the seen. In

robotics there is an embedded software inside the robot, asking it to perform various actions like navigating on a terrain. Expert system is concerned with, the expert level performance of a software on a specific task. For example, the task could be diagnosis of diseases and curing this. A doctor is known to operate with a number of rules, a very large number of rules obtained by years of education and practice on patience.

So, the expert system is concerned with emulating this behavior of the expert. Let us move on to the, feeding disciplines which are at the 2nd layer. We find that machine learning and planning feed into a number of layers in the outer most category. For example, natural language processing is fed by machine learning and natural language processing is also fed by knowledge representation. The reason for this is that in current world, natural language processing is using lots of statistical techniques. Statistical techniques are machine learning techniques; they make use of the knowledge contained in the data.

In today's internet world we have a large amount of text, in the form of a number of documents: estimable pages, p d f pages, word pages, power point presentations and so on and so forth. The internet is full of textual documents. So, these textual documents have to be made use of, a program has to make sense of this textual document, that requires natural language processing technique. Now, the point here is that suppose human beings are ask to make sense of all these data. Then how many pages can human being really shift through, in let say 24 hours of time or even if you take you know working hour of 8 hours per day. How much of data can human being possibly see?

That is the reason why it is important to develop machine learning techniques, statistical techniques which look at the data and obtain the knowledge content of the data. This is the importance of machine learning. Let us go to the transparency again and see that machine learning is feeding into natural language processing. Here, we find that the 1st layer is: search, logic and knowledge representation. In Search the machine is faced with a number of choices, a number of choices as it computes. Search algorithms try to find out the best possible strategy, the optimal strategy for computer. Now, one might ask in natural language processing is it necessary to conduct search? We will see many examples, where we are faced with a number of choices, when we are processing the textual data and search is very important for this.

I will give you a very small example. Suppose, I utter the word, "I went to the bank to withdraw some money". Now, it is known that bank is a very ambiguous word. Bank contains two meanings: one meaning of bank is the financial bank where one deposits money and withdraws money from, the other meaning of bank is the bank of river, the side of the river, the land mass of the river, the land mass which is on side of the river. So, when I say I went to the bank to withdraw some money, which meaning of bank did I have in mind? This requires search. A program will have to read the sentence left to right, I went to the bank to withdraw some money. Until it comes to money it is difficult for the machine or even a human being to understand that we are talking about the financial sense of bank.

So, this a very simple example to show, that we conduct search and we solve problems of search when we understand natural language. We will have many examples coming up when we discuss ambiguity of natural language processing. So, let us look at the transparency once again and understand the importance of logic. What does logic do? Logic is a vehicle for reasoning and inference. Logic is a vehicle for inferencing in the following sense, a number of rules and pieces of knowledge are given in logic as formalism. So, in logic we are concerned with a number of constructs like if x is true then y is true. So, if x is true then y is true that says that, whenever we can satisfy the values of x, the value of y is also satisfied.

In natural language processing, logic forms a very crucial component because the textual knowledge has to be converted into logical forms which a machine can process. So, this is the importance of logic and mainly proposition calculus, predicate calculus and some forms of non monotonic logic are used for natural language processing. Finally, we see this box, knowledge representation. Knowledge representation is again critical for natural language processing because, the sentence contains knowledge and this knowledge has to be extracted and embedded in the machine. So, this is a very important problem again.

So, let me summarize this transparency. In this transparency we see the importance of natural language processing and its place in the whole business of artificial Intelligence. So, natural language processing draws from: search, logic, knowledge representation. It also draws from machine learning and different areas of artificial intelligence like: vision, robotics, expert systems also draw from many different areas of AI as shown in the

diagram. If you look at the last sentence given in the transparency, A I is the forcing function for computer science. That tells a story that artificial intelligence pushes the frontier of computer science. We can take a step forward and say that natural language processing is the forcing function for artificial intelligence itself.

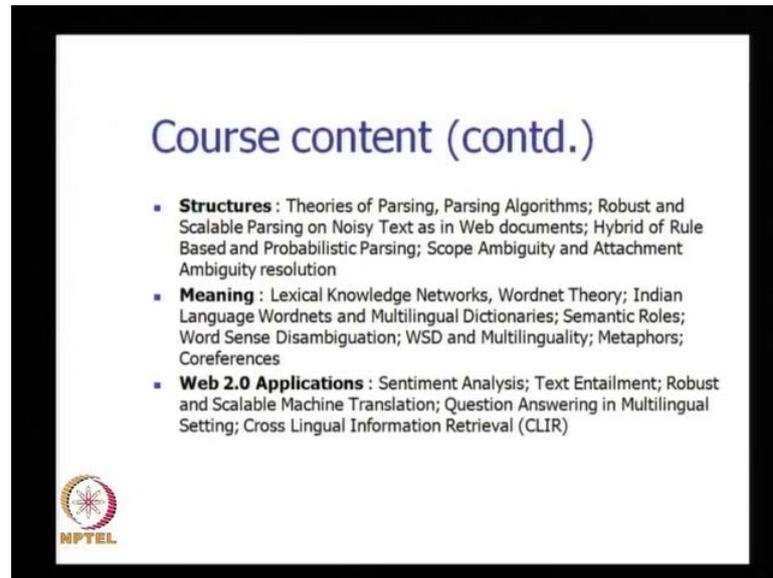
In natural language processing, we are concerned with day to day communication. Now, day to day communication requires understanding: the words, the phrases, the structures, the syntactic processing, the meaning of the sentences all these are extremely important even for artificial intelligence. And therefore, when we make it advancement industry language processing it impacts artificial intelligence as a field. An artificial intelligence in its own turn advances the frontiers of natural language processing. So, I think I am clear about the whole chain of influencing the different fields do. N L P influences A I, A I influences artificial intelligence, N L P advances A I, A I advances computer science and engineering. Let us move ahead.

(Refer Slide Time: 11:09)



I would like to spent some time on the course content. The top level topics which are mentioned here are: sound, words and word forms.

(Refer Slide Time: 11:23)



**Course content (contd.)**

- **Structures** : Theories of Parsing, Parsing Algorithms; Robust and Scalable Parsing on Noisy Text as in Web documents; Hybrid of Rule Based and Probabilistic Parsing; Scope Ambiguity and Attachment Ambiguity resolution
- **Meaning** : Lexical Knowledge Networks, Wordnet Theory; Indian Language Wordnets and Multilingual Dictionaries; Semantic Roles; Word Sense Disambiguation; WSD and Multilinguality; Metaphors; Coreferences
- **Web 2.0 Applications** : Sentiment Analysis; Text Entailment; Robust and Scalable Machine Translation; Question Answering in Multilingual Setting; Cross Lingual Information Retrieval (CLIR)

 NPTEL

In the next slide: structures, meaning, and web 2.0 applications. If we move to the 1st topic (Refer Slide Time: 11:09) sound, sound is concerned with the biology of speech processing. Natural language comes in two different forms: one is the spoken utterances, the other one is written communication. If we look at the biological processing of speech, how do human beings utter understand speech? Human beings hear words, the sound patterns which come to the ear and those speech patterns are understood by the brain. There are specific areas of brain, designated for auditory signal processing ok. And, these speech sound is converted in two patterns which are stored in the brain and they in turned activate our self's to perform some action.

We utter some sentences in response to speech or we take some action. We perform in real life world moving our hands, legs and so on. So, speech processing refers to concept formation in the brain, it also refers to taking action in the real life world. So, the topics that are mention there pertain to different areas of speech processing, the biology of speech, phonetics, phonology, place articulation and many different statistical techniques which are needed for processing of speech. Now, why is speech important for natural language process? Speech is important because speech gives many different statistical techniques consumed by natural language processing into today's world. Earlier when natural language processing used to be completely reliant on a linguist expertise, a language experts proficiency, a lexicographers proficiency and so on. It was completely human regain.

Now, it is seen that there is a lot of data, in textual form on the web and we can use machine learning techniques to make sense of these data. So, where do these techniques come from? These techniques come from two fields namely: speech and computer vision. They have been for a long time processing signals and patterns using machine learning techniques, statistical techniques. And, today's natural language processing cannot ignore those techniques. In fact, they benefit a lot from the application of those techniques. So, the main point I am making here is that speech actually provides natural language processing with its own statistical approach, statistical techniques. Moving forward word and word forms, look at the 2nd point here words and word forms.

Words and word forms I mentioned here: morphology fundamentals, morphological diversity of Indian languages, morphology paradigms, finite state machine based morphology, automatic morphology, learning, shallow parsing, named entities, maximum entropy models and random fields, I have listed out a number of topics. Let me explain to you in brief what I mean here. Words come in many different forms and our concern is to be able to process words very skillfully. Words form the 1st step, when you process language it is the words which we have to deal with in written communication for example. I took this example, very famous example in natural language processing, where the sentence was, "I went to the bank to withdraw some money". It is also possible to say: I will go to the bank to withdraw some money, I will go to banks to withdraw some money, I will go to banks to withdraw all my money.

So, look at this, look at what is happening. What is happening is that the same word is coming in many different forms. For example, banks is coming from the word bank, we will go, went, they come from the root word go. Now, English is a very simple language in terms of morphology. English produces many forms many of which are quite simple. For example, to form a future tense from go you just have to plug will before go, will is called an inflectional form. In Hindi you will have to say ((Refer Time: 15:54)) the word ((Refer Time: 15:56)) comes from two morphemes ((Refer Time: 16:00)) to go ((Refer Time: 16:03)) will go will. So, ((Refer Time: 16:06)) plus produces ((Refer Time: 16:08)). Therefore, it makes sense to take the word ((Refer Time: 16:11)) and break it into two pieces ((Refer Time: 16:15)) and, this is known as morphology processing, morphological analysis.

The opposite process is called morphology generation or morphology syntheses. We have a root word and from the root word we should be able to produce the word form. Again to take an example in English, suppose the root word is transport, we transport some material. Now, if I say that this word transport is for singular number and present tense, he transports some material. So, given the root word transport and the fact that the tense is present tense and the person is 3rd person singular number, transport becomes transports and s is added to transport. So, this is known as morphology generation or syntheses. Imagine a machine which is required to do natural language processing and natural language generation. So, what is given to the machine is let us say c, machine is suppose to describe as and it sees a human being transporting some material from point a to point b. So, the machine will have to produce the sentence, “he is transporting material from point a to point b”.

The word transport now have become transporting. This is the morphology process and a number of computer algorithms have been devised to deal with morphological processing. How can we efficiently process words and obtain their root forms? So, in this course we would like to see finial state machine base morphology. How morphology is processed by means of finial state machines. We will cover this topic in some detail just to show how language and computer science come together in the form of a very simple machine namely the finite state machines. We go to the next transparency. We come to now more advanced topics: structure, meaning and web 2.0 applications. In structures, I will mention the topic of: theories of parsing, parsing algorithms, robust and scalable parsing on noisy text as in web documents, hybrid of rule based and probabilistic parsing, scope ambiguity and attachment ambiguity resolution.

All these are technical terms which will be explained soon. But, it will make the main point here. Parsing or syntactic analysis happens to be an extremely well research topic in natural language processing. All other areas of natural language processing have been investigated but not so much as parsing or syntactic analysis. Syntactic analysis or parsing is also seen in other branches of computer science like, compilers and programming languages. But, there we are dealing with a much simpler problem. The problem of parsing a programming language, a piece of program is parsed. This has hardly any complexity compare to natural language, a C program or FORTRAN program

or paschal java program, they do not have the complexity of natural language sentences, paragraphs and chapters.

When we have a running piece of text, large amount of text we processed by a machine. And, we have to isolate the words, the morphine within the words, the morphological processing, the phrases which are present in the sentence: noun phrase, word phrase, which we will deal with after sometime. All these structures which are detected from the sentences correspond to syntactic analysis or structure processing. This is an extremely well understood area and a number of algorithms exist in this I will cover those algorithms in detail. Parsing will form an important component of our discussions. From structures we come to the next topic, if you look at the transparency again it is meaning. Meaning is the ultimate aim of natural language processing.

How do we extract the meaning of sentences? This is our main concern. I mention the topics here as: lexical knowledge networks, word net theory, Indian language word nets and multilingual dictionaries, semantic roles, word sense disambiguation, W S D and multilinguality, metaphors, co references, a very large number of topics. All of them are complex difficult topics. I would like to again spend some time on this topic. Meaning I said is the main concerned of natural language processing. How do we understand meaning? Now, in this business something that comes to big help is called word net and dictionaries and anthologies ok, they form very important components of meaning detection and meaning representation.

Now, word nets and lexical knowledge networks are nothing but meaning representations and their interconnections. To take an example, dog is an animal. We look at dogs and we also have this class of animals. What do dogs have? Dogs have: tails, eyes, legs, hair and so on. Animals also have many different properties. Most of the animals are moving, most of the animals pro create produce children's. And, most of the animals: drink water, eat food and so on and forth. Now, dog being a member of animal family in held's all these properties. So, there is an intimate meaning linkage between dog and animal. So, what are we talking about here?

We are talking about not the word, not the word dog and animal. We are talking about meanings of the words dog and animal. How they relate to each other. So, this is what is represented in lexical knowledge networks and it took a long time for natural language

processing to understand that meaning networks are crucial for natural language processing. We will spent a quite an amount of time on word nets at I I T Bombay we have done lot of research and development in word net building. And, I would like to describe our work on Hindi word net, Marathi word net, our effort at create creating Indian language word nets all over the country which is advancing the state of art in natural language processing in this country.

So, meaning representation through word net anthology dictionary will form an important component of our discussions. Coming to the next topic which is web 2.0 applications I mention the items as: sentiment analysis, text entailment, robust and scalable machine translation, question answering in multilingual setting, cross lingual information retrieval. What do I mean by all this? Those of you who are keeping track of what is going on in the internet, internet is again going through a revolution. Internet itself has caused a revolution in human life. Civilization has been profoundly influence by web, by internet, things which we were not imaginable before, the advent of internet is happening regularly today, absolutely regularly.

So, web is now coming up with its next version, which is web 2.0. I mention some topics here like: sentiment analysis, text entailment, machine translation in the large scale and so on and so forth. Let us just take the 1st topic, sentiment analysis. If you think carefully we have a completely new world order now. Never before was so much of public opinion was available in electronic form ok. Common man can access information about any organization any person, just through the click of a mouse, so much of information about persons and organization organizations are available on the web in a completely electronic form. This is a completely new scenario. It did not exist before. So, sentiment analysis is concerned with: how to look at a document, how to process the content of the document and then find out what is the document writer ok or the speaker say about a particular entity, an organization or a person.

Is the person positively oriented towards the person or the organization or is the persons opinion negative. For example, take a bank and you look at the blogs, that the users of the bank express their opinions seen blog. You have lot of textual data in electronic form and you would like to understand, is the blog praising the bank or is it expressing opinion against the bank. This is the concerned of the field call sentiment analysis. In sentiment analysis, we would like to develop programs which automatically understand the opinion

of the users from the electronic text. This is known as sentiment analysis or polarity detection. So similarly, text entailment is concerned with inferencing of text. Given two pieces of text are they consistent with each other, does one text follow from the other? In large scale machine transitions which is a very old field and extremely relevant for a country like India where multiple languages are spoken, written.

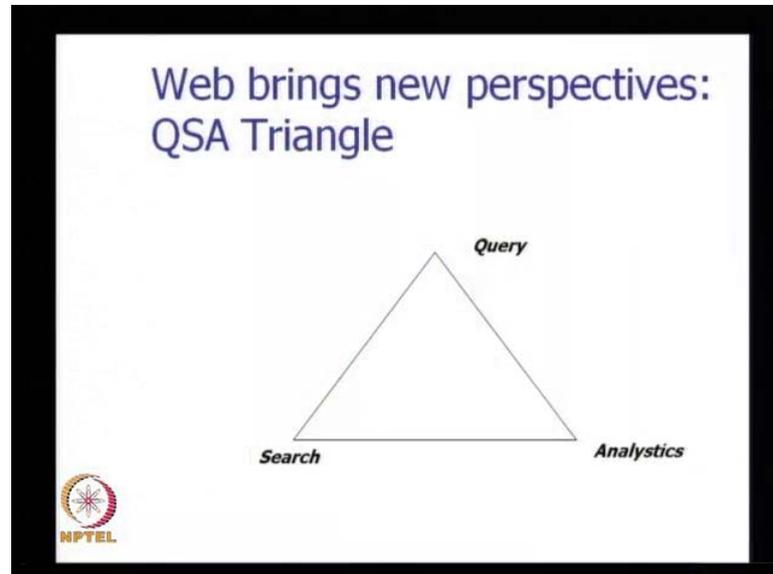
The concerned is to able to translate from one language to another like: English to Hindi, Hindi to Marathi and Marathi to Bengali and so on. At I I T Bombay, we again have large scale activity on machine transition. The final topic I mention there is cross lingual information retrieval. If you think about it in a country like India, cross lingual information retrieval forms a very important problem. Users have information it. Where do they go for their information need? Earlier, when the web did not exist users is to go to libraries, obtain their information from the library. Now, a day's people click, click a mouse or they use keyboard they go to the web and obtain the information from to the web. Now, imagine what kind of problem a user will face, if the language of the user is not English? The web, the a large a large part of a is in English. The user has to suppose the query in English, obtain information in English and then understand the document.

If the user is not comfortable with English, then the user is handicapped. This is known as the problem of language barrier. In India the comfort level of English is not very high. It is known that about 5 to 6 percent of Indian population is comfortable in usage of English. So, for such people when there is information need has to be met, they should be able to posed query in their own language, obtain information in their own language. But, at the background what is happening is the web is processing the query looking at large amount of English documents and then producing an answer in the language of the user. This is known as cross lingual information retrieval and cross lingual retrieval is a very relevant and current problem for many countries and India is no exceptions ok.

So, these in an our shell is at over view of the topics I would like to cover in summary. We will look at speech processing techniques. We will look at how words are processed and stored, morphological processing dictionaries. We will understand the techniques of syntactic analysis parsing, meaning representation is done in the dictionaries, word nets and anthologies. We would like to discuss those topics. And finally, some of the web 2.0 applications like: sentiment analysis, text entailment, machine translation and cross lingual, information retrieval will be covered. These gives an overview and I suppose

this is a an existing set of topics and, once these topics are covered one gets a good overview of what is going on in naturally language processing.

(Refer Slide Time: 28:48)

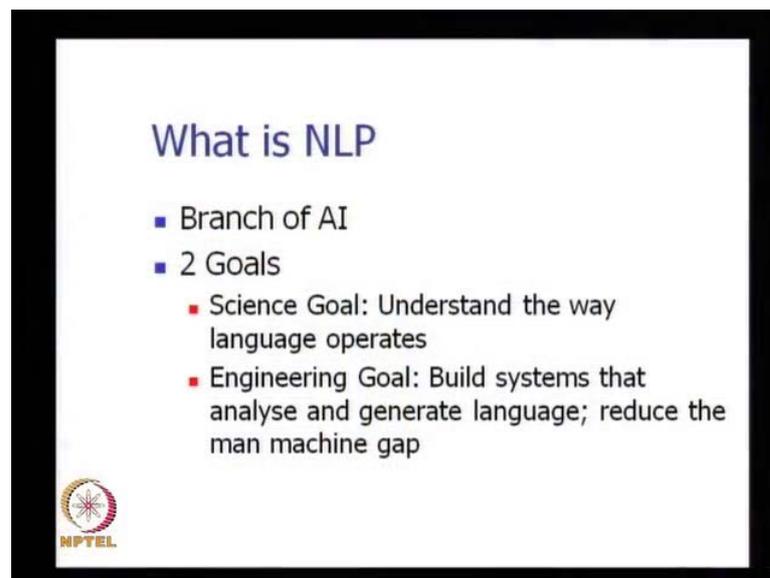


We will move forward now, with some of the central topics. We have said many times that web brings new perspective. If you look at the diagram here, we have what is called the Q S A triangle. Q is the query, S is a search and analytics is the intelligent processing of information on the web. This is known as the Q S A triangle, a very famous concept in the scenario of web. Q query, S search and A analytics ok. And, this actually brings in many fine points of discussion. We do not have enough time to going to those details. Let me just make one of the remark and this. When the user is posing a query, this query is process by the search engine and document searched. So, this is the search component. The 1st component was the query component; query has to be processed after.

And, when the documents come up, the large documents come up from the search engine, we have to understand what this document mean and do the satisfy our information need. The topic of analytics is concerned with that. It is concerned with how to process these documents intelligently. So, we can imaging a future scenario where the user gives in the query, the documents are searched and the essential to information with respect to that query is presented to the user all by the machine, all by the computer ok. So, after the query is presented search and analytics they work together, they work in synchrony. They work in tandem to produce information for the user to consume for the

user to use that information. That is the Q S A triangle and you can see that this is a documents are mainly in textual form, there lots of documents in image form. The document which are in textual form they required natural language processing techniques to be processed.

(Refer Slide Time: 31:03)



Moving forward we now proceed to define natural language processing. As the slide shows it is, natural language processing is a branch of artificial intelligence. We have already dwelt, on this particular issue in one of the transferring before where we showed many different dependencies of A I. Now, natural language processing has two goals: the science goal and the engineering goal. The science goal and engineering goal they go hand in hand. They work in tandem. In science goal, the aim is to understand how language is produced and how language is understood by an intelligent entity ok. For example, how do I process the sentence, “I went to the bank to withdraw some money”. If somebody asked me why did you go to the bank, I will answer to withdraw some money ok. How do I answer this question?

You I must of understood the question. I must of understood that the meaning of bank here is a financial bank not river bank. How did I do that? So, this is a cognitive process which is happening in the brain and human beings are extremely good at it. The science goal of natural language processing is to understand these phenomenon. This tremendous phenomenon of: how we process language, how we generate language, how we interact

with our fellow human beings through language, is the science goal. The engineering goal on the other hand, is concerned with the use of the techniques some natural language processing. When we create a natural language processing program, it has been particular use ok. For example, I could use the natural language processing program to read a text I am produce sounds corresponding to the text. This for example, could be very useful to blind person ok.

A blind person is not able to read a piece of text himself. So, we give the text to natural language processing program, it automatically reads a document, produces the speech sound and the blind person understands what the meaning of this document is. So, this is a an extremely important utility of language processing.

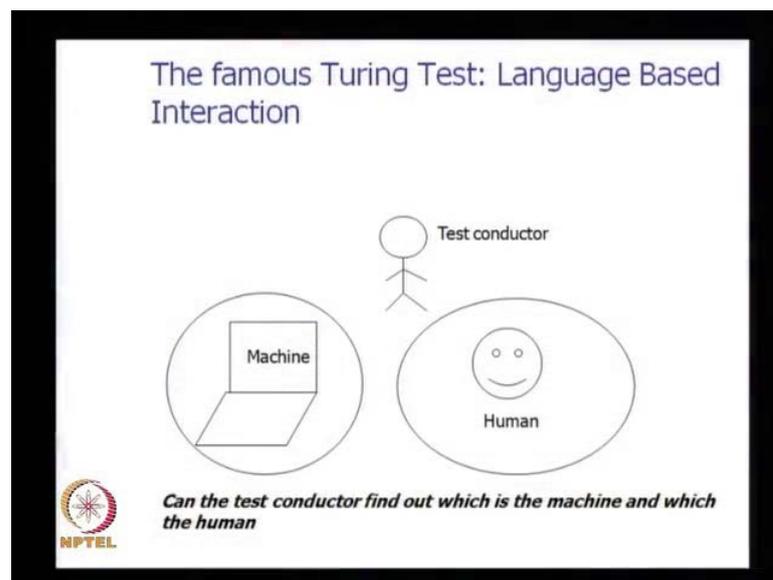
Consider another scenario where the person is not blind but he is speaking. And, as he speaks the speech sounds are interpreted by a speech processing language understanding system and those speech sounds are converted into a textual file. So, imagine a, imagine how the life of a teacher will be comes simpler if such a systematic. The teacher simply comes to the class they leavers the lecture. A software program capture those sounds and produce a textual file which can eventually uploaded, on the teachers home page and he meant on the on the students can make use of the lecture notes.

The lecture note was not written by the user, by the teacher not was anything written by the student. A program captured all the sound and produces the lecture note ok. So, this is another very important use namely: the speech understanding and speech encoding into text. There are many such applications of language processing. I mentioned sentiment analysis some time back and this problem, the problem sentiment analysis requires language processing. It is a very important problem. Organizations, now a day's are very concerned about what the public are saying about on the internet. And, it is impossible for a single human being or even a team human being, to look at the whole web and tell the organization. You know people are saying very nice things about you or saying bad things about you.

So, can the employees, software agent, can the software programs which will navigate the web, over the whole web and give to the organization peoples feedback about this. So, these are many different utilizes natural language processing and these pertains to the engineering goal. So, we mentioned two goals: science goal and engineering goal both

have to be kept in mind, for anybody working on natural language processing ok. The excitement of the field comes from this, a this wisteria phenomenon. This very you know deep phenomenon of language being processed in the brain and language being produced. The excitement comes from there, these a science goal. And, the other kind of excitement that comes is that language processing produces useful tools and resources which makes human beings life easier, so science goal and engineering goal.

(Refer Slide Time: 35:51)



Preceding further, there was this very famous test call turing test, which was proposed by one of the great, I would say one of the great man of computer science and mathematics Ailon Turing, who proposed that a particular test can be conducted to find out if a machine is truly intelligent. If you look at this diagram there is this test conductor, who is interacting with a machine and a human. The person, the test conducted does not know, where is the machine and where is the human beings. Imagine these two circles to be two different rooms and all the communication takes place through the keyboard. Now, the task is, can the test conductor find out which room contains the machine and which room contains the human being? Can the test conductor find out which is the machine and which the human? So, if the test conductor cannot distinguish between the machine and the human, then the machine is suppose to be intelligent. It should, it is suppose to be passing the test for intelligence.

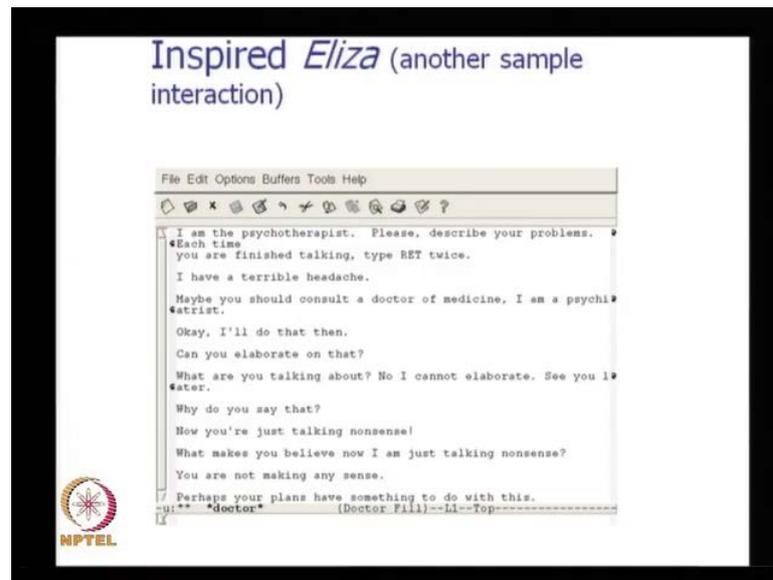
So, let us spend some time on this test. This is an extremely famous test called turing test. In turing test, our goal is to see if a machine is truly intelligent, if the program psi truly intelligent. And this is proposed by Ailon Turing to test a program for its intelligence. There are many criticisms, many philosophical criticisms against turing test. However, the test has state, over the years the test has been studied by A I students, A I researcher, so much so, now a days you have a competition. In this competition you essentially conduct a turing test to find out if a program is truly intelligent or not. The participant or subjected to something like a turing test and they are given enormous amount as money of price. If the test conductor who is human being cannot detect, if the software program who was a machine indeed or was it actually human being, which was interacting with the test conductor.

(Refer Slide Time: 38:25)



Now, the turing test actually inspired something called Eliza, which is a very famous program, one of the 1st programs to demonstrate natural language processing. Though in a very nigh way but it brought out a number of aspects of natural language processing. So, here is the u r l which can be try, it which you can try to interact with Eliza.

(Refer Slide Time: 38:51)



I show a snap shot of Eliza. Look at this screen and look at the 1st sentence here. The whole screen captures interaction with a mechanical psychotherapist. It is a software program which behaves like a psychotherapist. The 1st sentence is, I am the psychotherapist please describe your problem. Each time you are finished talking type R E T twice that is the return button twice. This is the sentence coming from the machine or the program and the human being answers saying I have a terrible headache. Here it is, I have a terrible headache. The machines response is maybe you should consult a doctor of medicine, I am a psychiatrist. So, this is a very nice answer, if you can appreciate this point is a very nice answer, very natural answer, any psychotherapist or psychiatrist will answer this way.

If you have a headache where should you go? You should go to a doctor not to a psychiatrist. Then the human being says, ok I will do that then. Now, the machine comes back and say can you elaborate on that and this is a bit of a strange response. Why would the entity respond this way can you elaborate on that? The machine is already advice the human being to go to a doctor and the human being also, the patient also as agreed, yes I will go to the doctor. Now, why should the machine come back and say can you elaborate on that. Then the human being says, what are you talking about, no I cannot elaborate see you. So, he is about to leave. I cannot elaborate I will go now, see you later.

Then the machine says, why do you say that? Then the human being says, now you are just talking nonsense, the human being is irritated. Now, the machine says, what makes you believe now I am just talking nonsense? Human being says, you are not making any sense. Then the machine says perhaps your plans have something to do with this. You know, so after can you elaborate on these and after the human being is about irritated and says that he cannot do much about all these conversations. The machine goes into a completely nonsensical mode of communication. Now, you are just talking nonsense. What makes you believe now, I am just talking nonsense and so on. So, what does this show? What does this communication show? The communications shows that initially or even after some level, the machine can fail to be intelligent, can pretend to be intelligent.

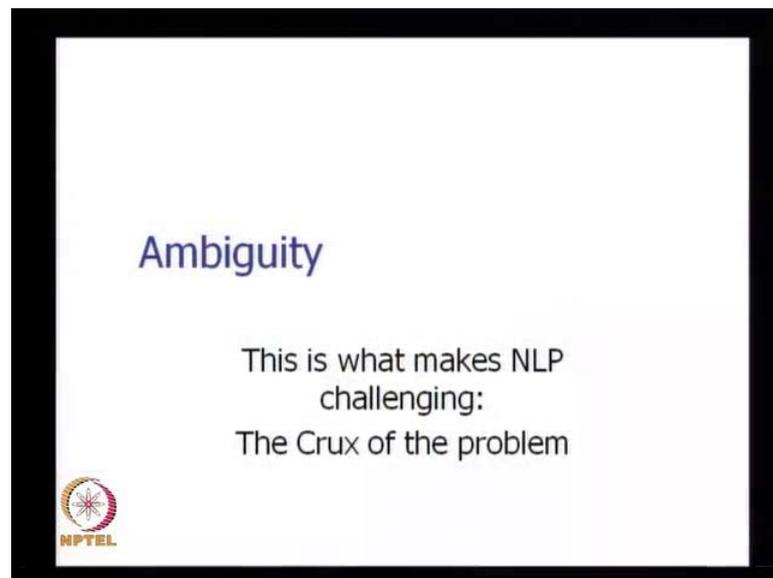
The moment in the conversation takes a deeper turn it requires more complex listening, more complex background knowledge, lot of experience, the machine shows sign of failure. It begins to ask question which a completely nonsensical and shallow ok. And why does this, why does this happen? Eliza was a program which was created by wise and bomb and the goal was to show that natural language processing actually does not have much substance in this. Nobody will agree with this point of view anymore in today's world. Natural language processing is understood to be very deep field with remain a some lot of utility. But, in those days wise an bomb set out to showing to the world, there see I can create a program, a software program which can intelligently converts with a human being, without a were informing the human being that it is actually a software agent ok.

So, that is the point, the software program is showing that it is powerful enough to deal with language whereas, it actually is not. The internal algorithm was the following: there were a number of templates and queue words which is the software program was looking for all the time and matching it. So, it has for example, a very stock answer. What makes you say dot dot dot ok. And what makes you say that you are not well. So suppose, I say I am not well and the machine response is what makes you say that you are not well. And, there is some cleverness in converting I to you, you too I and so on. But, the whole thing is template based. There is a template will says that whenever you see a sentence and you do not know what to do with this sentence, how to process the sentence?

Simply output, what makes you think, what or makes you say something. So, this is this not intelligence, this is hardly intelligent. This a completely superficial processing of the

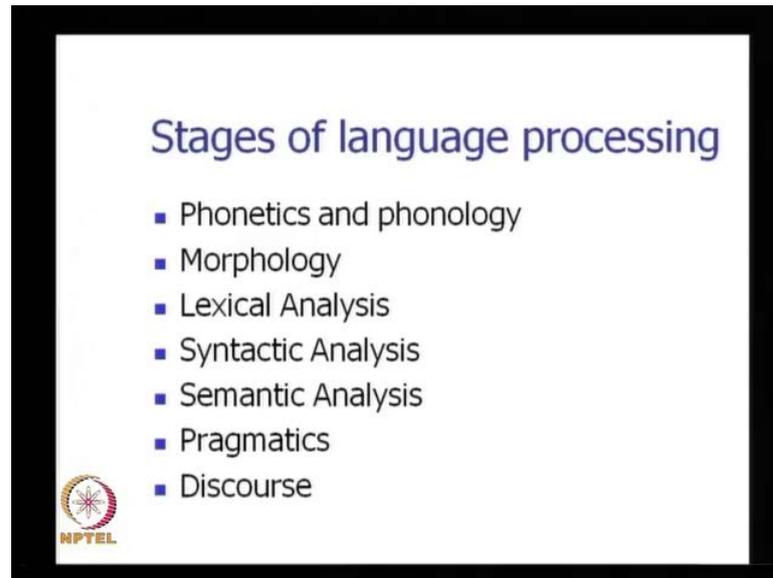
sentence. Similarly, there were many other templates which are in build in the program and vision bomb try to show that just by activating these templates a machine can meaningfully converse. But, this is not true as the example is showing here. The moment the conversation goes into deeper things the machine begins to failure. Still, the program was inspire by turing test, vision bomb wanted to constructed program which will behave like a human being and compares with a another human being.

(Refer Slide Time: 44:55)



Moving forward we come to probably the most important starting discussion of natural language processing namely ambiguity. So, we write here ambiguity. This is what makes natural language processing challenging and this is the crux of the problem. Ambiguity is there everywhere at all stages of natural language processing and we proceed to elaborate on that.

(Refer Slide Time: 45:10)



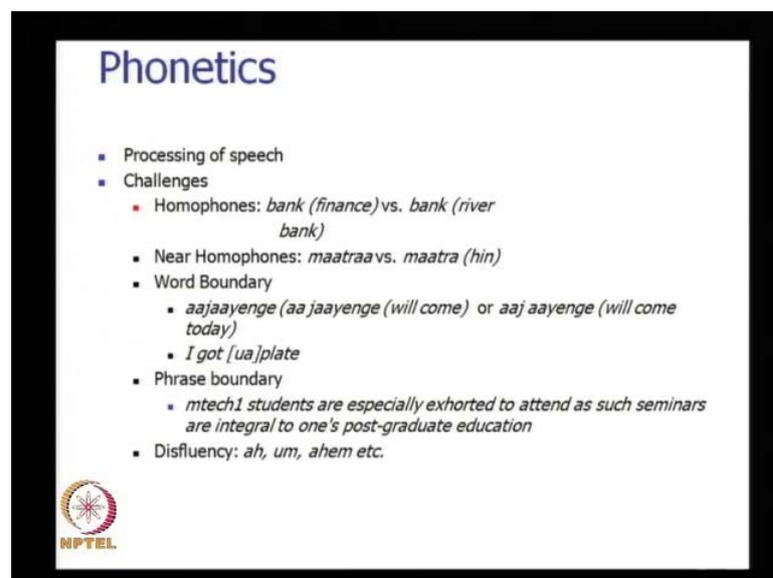
## Stages of language processing

- Phonetics and phonology
- Morphology
- Lexical Analysis
- Syntactic Analysis
- Semantic Analysis
- Pragmatics
- Discourse



It is known that there are different stages of language processing. The stages are listed here: phonetic and phonology, morphology, lexical analysis, syntactic analysis, semantic analysis, pragmatics and finally discourse. So, all this topics are very well known in natural language processing and speech. My point here a listing all this topics is that, will see everywhere we have to deal with the problem of ambiguity. Starting for phonetics and phonology up to discourse everywhere, we have the problem of ambiguity coming up and we have to find solutions to those problems.

(Refer Slide Time: 45:56)



## Phonetics

- Processing of speech
- Challenges
  - Homophones: *bank (finance)* vs. *bank (river bank)*
  - Near Homophones: *maatras* vs. *maatras (hin)*
  - Word Boundary
    - *aajaayenge (aa jaayenge (will come) or aaj aayenge (will come today)*
    - *I got [ua]plate*
  - Phrase boundary
    - *mtech1 students are especially exhorted to attend as such seminars are integral to one's post-graduate education*
  - Disfluency: *ah, um, ahem etc.*



Please take the 1st of this list, phonetics. Phonetics is concerned with the processing of speech. The challenges in this are homophones namely strings of alphabets or words which sound similar. So, bank in the financial sense and bank in the river sense. They sound similar. Therefore, they are called homophones. Near homophones are those which have very close sound, for example, the word ((Refer Time: 46:33)) and ((Refer Time: 46:34)) in Hindi. These two words also and used in Marathi, there also used in Bengali. Most Indian languages have these two words if the language is from the Indo European origin. So, ((Refer Time: 46:46)) and ((Refer Time: 46:47)) sound similar but they are not identical unlike homophones so, near homophones and homophones. We come to this very integrate problem world boundary detection.

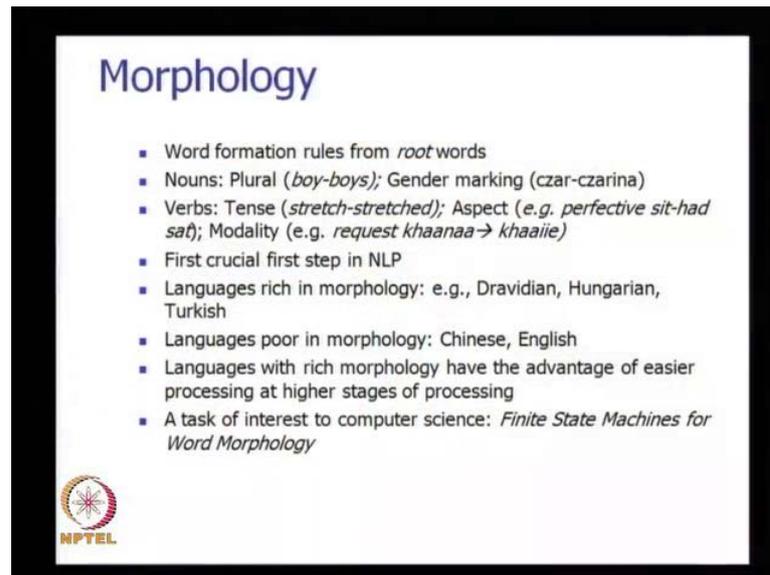
What is it I have written here, ((Refer Time: 47:02)). This Hindi word ok, this is a Hindi string. This can be broken up in two ways ((Refer Time: 47:13)) will come or ((Refer Time: 47:16)) will come today. So, if you break it at j before j then it is ((Refer Time: 47:24)) will come and if you break it after j then it is ((Refer Time: 47:29)) will come today. So, the speech understanding system has to break the word, at the appropriate place. For example, if you break it here ((Refer Time: 47:43)) and ((Refer Time: 47:44)) then this to morphemes are not making much sense. No listener will break this string in to this two parts ((Refer Time: 47:54)) and ((Refer Time: 47:55)). I take this example in English, one of my favorite examples. I got a plate. So, if I says it very quickly, you will not know what I said. I got a plate ((Refer Time: 48:09)) or I got a plate ((Refer Time: 48:13)).

So, these are the two meanings and when this whole string is given, I got a plate from the context you have to make out, is it I got a plate ((Refer Time: 48:26)) or I got a plate ((Refer Time: 48:30)). Similarly, there are these problems are phrase boundary detection. These are example here, I will it as an exercise to you to see, how the phrase boundary when broken at, when broken before as such or after as such and produce two different meanings. A very important problem in phonetics is disfluency. I show some of the strings here: ah, um, ahem etcetera. Now, these strings have no meaning, these earliest says the speaker to organize is thought, the speaker by sometime through these regions of disfluency.

If I say, I will go to school but, I do not know what to carry their, I forgot my umbrella so all this etcetera are regions of disfluency. They give the use if the speaker sometime to

organize is thought. So, what they say in this slide, what they said in this slide was that when we have phonetics and phonology we are dealing with speech sounds. We have to deal with a three problems: one homophones, two near homophones, three word boundary. So, these are the three different problems we deal with when we are dealing with speech sound.

(Refer Slide Time: 50:06)



**Morphology**

- Word formation rules from *root* words
- Nouns: Plural (*boy-boys*); Gender marking (*czar-czarina*)
- Verbs: Tense (*stretch-stretched*); Aspect (*e.g. perfective sit-had sat*); Modality (*e.g. request khaanaa → khaaiie*)
- First crucial first step in NLP
- Languages rich in morphology: *e.g.,* Dravidian, Hungarian, Turkish
- Languages poor in morphology: Chinese, English
- Languages with rich morphology have the advantage of easier processing at higher stages of processing
- A task of interest to computer science: *Finite State Machines for Word Morphology*

 NPTEL

We move forward and take the challenges involving morphology. Morphology, as described before, deals with word formation rules from the root words. For examples, the nouns: boy boys, gender marking, zar zarina. They come from the root word boy or zar, ((Refer Time: 50:28)) for example. They are also, they corresponds to gender marking. Verbs give rise to different forms through tense like stretch stretched, aspect for example, perfective sit. From sit we can obtain had, modality for example, request the root word is ((Refer Time: 50:52)). From this, we obtain ((Refer Time: 50:55)). So, 1st crucial step in natural language processing is morphology. We have to detect all the morphemes contain in a large word string.

So, languages which are reach in morphology are Dravidian languages: Tamil, Telugu, Malayalam, Hungarian and Turkish in European. Languages which are poor in morphology are Chinese and English. Chinese hardly uses any morphological suffixes. English is also not very rich in morphological variations. Languages with rich morphology have the advantage of easier processing at higher stages of processing.

Since, we have dealt with the word, in all its suffixes prefixes and so we have made a lot of progress towards word meaning. A task of great interest to computer science is finite state machines for word morphology. So, computer science comes in handy here. There are word formation rules and the suffixes which get added to the word they coming particular word.

Let me take an example from Marathi. The word ((Refer Time: 52:08)), so ((Refer Time: 52:12)) has three different components ((Refer Time: 52:17)). You can also say ((Refer Time: 52:21)). Now, ((Refer Time: 52:23)) they coming particular word ((Refer Time: 52:27)) cannot come before ((Refer Time: 52:28)) cannot come before ((Refer Time: 52:30)). So, there is a particular word, in which the suffixes I produced and inserted into the word. So, these automatically become a small parsing problem and this problem is very effectively dealt with by making use of finite state machines. computer science has invested lot of its energy in understanding finite state machine, the algorithm corresponding to that and the theory of finite estate machines. They come in handy, when we deal with morphologic processing. Here we will finish the 1st lecture.