

Applied Natural Language Processing
Prof. Ramaseshan Ramachandran
Department of Computer Science and Engineering
Chennai Mathematical Institute, Madras

Lecture – 01
Introduction

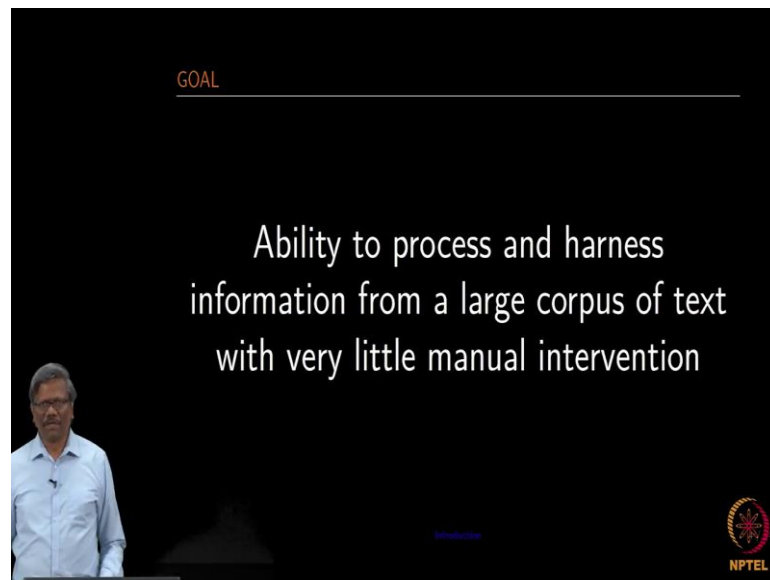
Hello everyone, this is Ramaseshan. I will be taking a through the course on Applied Natural Language Processing for the next 12 weeks. I think you would be really interested to know what are the real requirements or the basic requirements for this course.

(Refer Slide Time: 00:33)



If you know some computer science, probability and statistics at 101 level, linear algebra at the same level, and some machine learning it will really help, and we have lots of common sense it is going to be really useful. If you are really interested in these subjects, you can learn anything along the way, you do not need to really know the entire computer science or probability or statistics, you will definitely learn them along the way during this course. I will also be teaching some of the fundamentals which are required for this course as we move along all right ok.

(Refer Slide Time: 01:15)

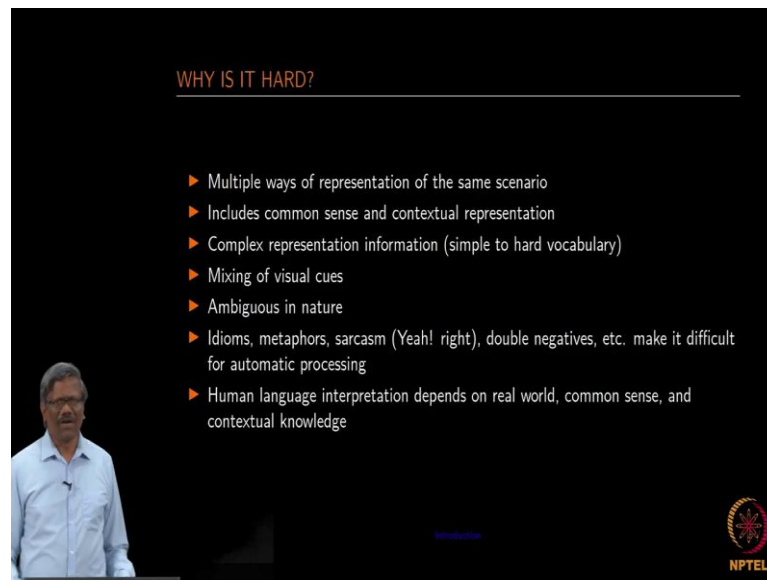


So, let me start with what is the goal of this course ok. The goal of this course as given in the slide is the ability to process and harness information from a large corpus of text with very little manual intervention ok. It is a long goal and it is a very interesting topic to deal with because natural language processing is always about human intelligence right. So, when we speak to someone, the other person on the other end really understands what you are speaking if he understands your language and then tries to respond and so it is an intellectual way of communicating to different humans right.

So, we want to be able to make the machine do the same job, is it possible for the machine to really understand what we speak, what we write and what is available as part of the large corpus of text that we are going to be providing here? Will it be able to understand the meaning of every word, will it be able to understand the sentence, will it be able to understand the context in which we are speaking all that. So, this is going to be a very long drawn journey in order to make the machine really understand everything that we do in the natural language processing right. There is a way to it, there is a what to it, and there is also a how-to it right.

So, why is it about why do we want to do this? And then what is it, what are we doing to make a machine understand this, and then for us to make the machine do something we also need to provide how you should proceed in certain fashion right. So, we will be answering all these three questions at a certain level during the course.

(Refer Slide Time: 03:05)



WHY IS IT HARD?

- ▶ Multiple ways of representation of the same scenario
- ▶ Includes common sense and contextual representation
- ▶ Complex representation information (simple to hard vocabulary)
- ▶ Mixing of visual cues
- ▶ Ambiguous in nature
- ▶ Idioms, metaphors, sarcasm (Yeah! right), double negatives, etc. make it difficult for automatic processing
- ▶ Human language interpretation depends on real world, common sense, and contextual knowledge

NPTEL

So, is it really hard natural language processing is it really hard? So, we need to really ask this question to a toddler who doesn't know any language right, but unfortunately, the toddler would not be able to communicate back to you, because he or she does not know what we are asking, but the way they learn quickly the language you know from the parents is an amazing one right. But unfortunately, we don't know how we really learned those languages when we were very small ok.

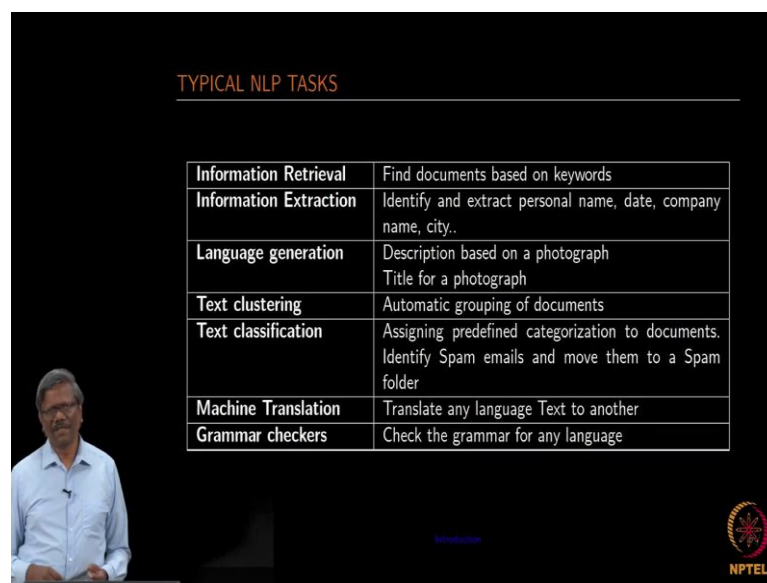
So, leave that apart, there are when you talk about any of the topics or any subject, there are multiple ways of representing the same thing in one sentence right. So, we are so innovative in a way that we create sentences every time and which are new and so on. We include a lot of common sense and contextual representation as we speak. For example, we would have seen one accident or some incident yesterday, and then when we talk to a friend today about that we will not be again taking the entire context into this and then stop talking right, right from the beginning start from somewhere. So, both understand what we are talking about, and then we move on from there right.

We represent the information in a very complex way using some hard vocabulary and so on. We make some visual clues, we use our hand symbols to communicate a few things you know without really saying in certain words. Ambiguous in nature, since these languages are created by humans you know the way you represent, the way you talk

there could be some ambiguity in it. So, we somehow managed to understand even if there is some ambiguity in what we are speaking right.

Idioms, metaphors and sarcasm make it hard especially for the machine to understand those things. So, how do you make the machine understand that there is a touch of sarcasm in what we are saying right? And then the interpretation depends on the real-world common sense and contextual knowledge. So, all those kinds of meta-knowledge that we have, we need to have somehow fed into the machine in order for the machine to really take on this job right. So, it is kind of very complex mechanism in order for us to really represent that in a certain algorithmic fashion. So, if you are able to really represent what we do using our language in an algorithmic fashion, then it is going to be easy for us to make the machine do what we do automatically.

(Refer Slide Time: 06:07)



TYPICAL NLP TASKS

Information Retrieval	Find documents based on keywords
Information Extraction	Identify and extract personal name, date, company name, city..
Language generation	Description based on a photograph Title for a photograph
Text clustering	Automatic grouping of documents
Text classification	Assigning predefined categorization to documents. Identify Spam emails and move them to a Spam folder
Machine Translation	Translate any language Text to another
Grammar checkers	Check the grammar for any language

So, what are the typical task that we perform, I am sure you all know about this because we have been using search engine right from the time we started using our mobile phones or computers right. A piece of information that we will find the document that we are looking for based on certain keywords.

And then there is another task that we do which is called the information extraction. Identify an extract personal name, date, name of the city, and then say certain other insightful information that we want to extract from documents know. We just provide the keywords for example find which country this particular player belongs to the right. So,

when is provide that information as a set of keywords, you get some Wikipedia results or some portal that provides you the details of what you are looking for.

Then the next one is language generation. This is something that we might be using on a regular basis. For example, if you are using a mobile phone and texting it to someone right, the keywords start appearing as you start typing right. When you start typing the first word or when you start typing a few characters at the keyboard, you see certain predictions or suggestions right. And then once you type a keyword, and then once you type one word or keyword or whatever in the search box or in the normal textbox, you see another prediction depending on what you have typed earlier right. Again there is some kind of a suggestion that you see, so that is something called the language generation we will also talk about that.

And then there is another task that is performed which will automatically classify documents into groups. Supposing if there are about 10 groups in the documents that we have captured, text clustering will classify them into 10 different buckets on its own without really naming them. This is an automatic way of classification it only uses what is there inside the document, and then depending on this keyword combination we will start putting document A in bucket A, and then document B in bucket B and so and so forth.

And then another one is text classification this is something that you say classify these documents based on these types of keywords. If these types of keywords occur together, then it belongs to let us say physics documents. If these keywords occur together, then it could be a spam, email and so on.

And the next one is the machine translations very very interesting this is about the translation of one language to the other. To further; for us to do that innovation should be able to understand both languages. So, how do you make the machine understand this is part of the NLP tasks? I am sure some of you also would be using grammar checkers you know as to when you type certain sentences in the textbox, grammar checkers automatically would start finding spelling mistakes and then maybe know some syntax corrections will provide you some syntax corrections and the sentences that you are typing. So, these are at a very high-level typical task that we would be looking at.