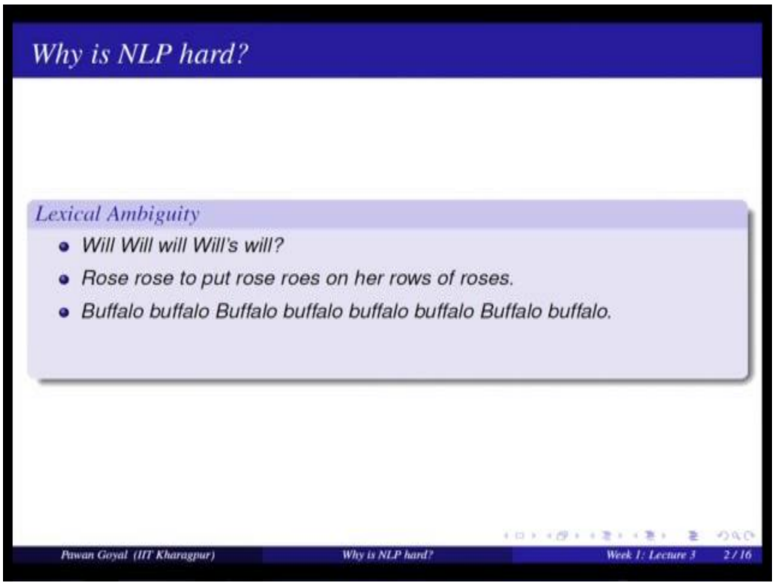
Natural Language Processing
Prof. Pawan Goyal
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture – 03
Why is NLP Hard

Hello everyone. So, welcome to the 3rd lecture of the first week. So, in the last lecture we were discussing, what do we do in NLP, what are the various applications where NLP has been used and in which been use currently, in what are some of the ferric some of the future potentials of NLP. So, today in this lecture, we will discuss why is NLP hard, what are the some of the difficult is that we face while designing algorithms for NLP and why do we need to worry about various machine learning techniques and all in NLP.

(Refer Slide Time: 00:50)



Why is NLP hard?

Lexical Ambiguity

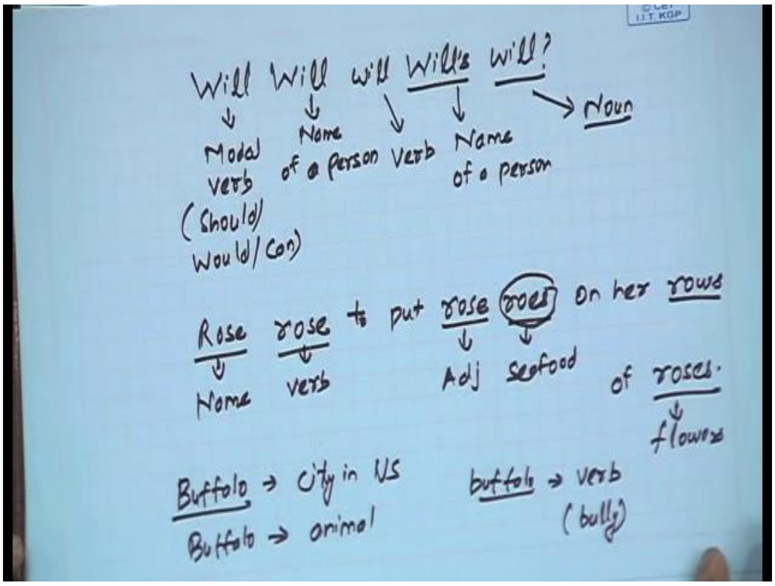
- Will Will will Will's will?
- Rose rose to put rose roes on her rows of roses.
- Buffalo buffalo Buffalo buffalo buffalo buffalo Buffalo buffalo.

Pawan Goyal (IIT Kharagpur) Why is NLP hard? Week 1: Lecture 3 2 / 16

So, which is start with a very simple case of the kind of problems that you face in NLP; we start with the case of lexical ambiguities; ambiguities as you understand is implied for the same word, meaning various interpretations.

So, take this example sentence here, so I have this sentence will will will will’s will. So, what you see here is the same word will, has been used 5 times.

(Refer Slide Time: 01:20)



So, let us try to find out what are the various interpretations of meaning the same what will has been used for; will will will will’s will. So, what do you say about the first will? So, the first will is a modal verb like should, would, can etcetera; the second will here is an name of a person; this sometimes easy to get from the English sentence because the full always been in capitals; then you have the third will, this is a verb in the sentence to express his will; the fourth one is again a name of a person you can see it by the apostrophe that is after this word, but it can be either the same will or a different will. So, against a name of person, but we denote is the same will or a different will.

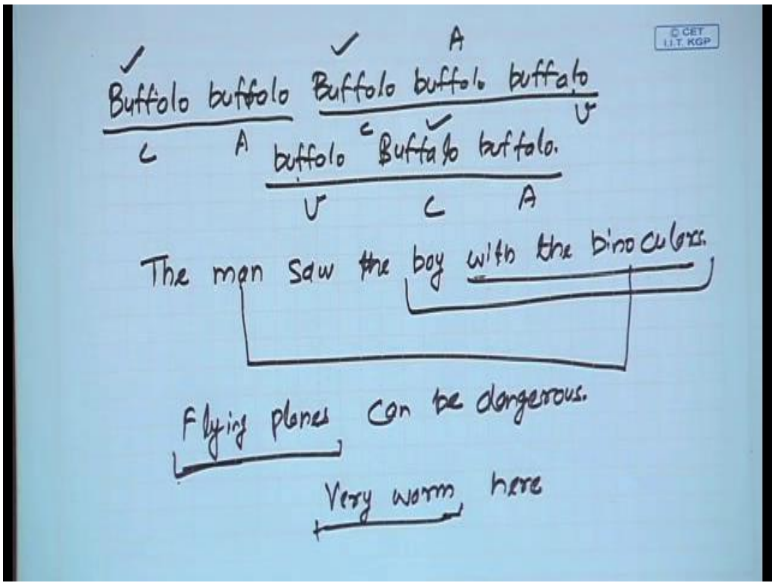
But by the way, the way we write language and the way we communicating language we can say this probably is a second person because if it by the same person I would have used probably hinge and this will is a noun, the will itself. So, what you are saying in the same sentence, will the person will, will express or use will’s will, we are using the same word will at least in four different meanings. So, this is one of the extreme cases that we see in language, but let us try to go through some other examples.

So, this is another example: rose rose to put rose roes on her rows of roses. So, here can you try to decipher what are the various meanings that I am (Refer Time: 03:17) for the same word rose. So, what you say about the first rose? The first rose is the name of the person, rose they should be a verb in the past tense to put rose roes, so what is that? This will be an adjective and rose is some sort of the seafood and then we have the next sentence on her rose of roses and these versus you can find out these are the flowers and this is a rose with you.

So, again these see here the same word rose, if you see the way it is written, in terms of autography the same word rose has been use one now as a name; as a verb, as an adjective and as a flower – 4 different sans age. But for the time just try to think of some a speech recognition system, we are your trying to pronounce this sentence and it has to transcript what are the different words that you have said in your trends. So, it will have ambiguity, even to find out whether this rose means r o e s or r o s e; this is another problem that this also handled by various models that we will see in natural language processing.

And let us take another extreme example. So, the Buffalo buffalo Buffalo buffalo buffalo buffalo Buffalo buffalo. So, now, this sentence is composed only by a single word used 8 times. So, can you try to decide for this? So, let me give your hint here. So, here word buffalo has been using 2 sense 3 senses Buffalo, one is the city in US, another Buffalo is an animal and the third buffalo is like a verb that is used in the same sense as bully. So now given these three interpretations, can you try to decipher the meaning of the sentence?

(Refer Slide Time: 06:00)



Buffolo buffolo Buffolo buffol. buffalo
buffolo Buffalo buffolo.

The man saw the boy with the binoculars.

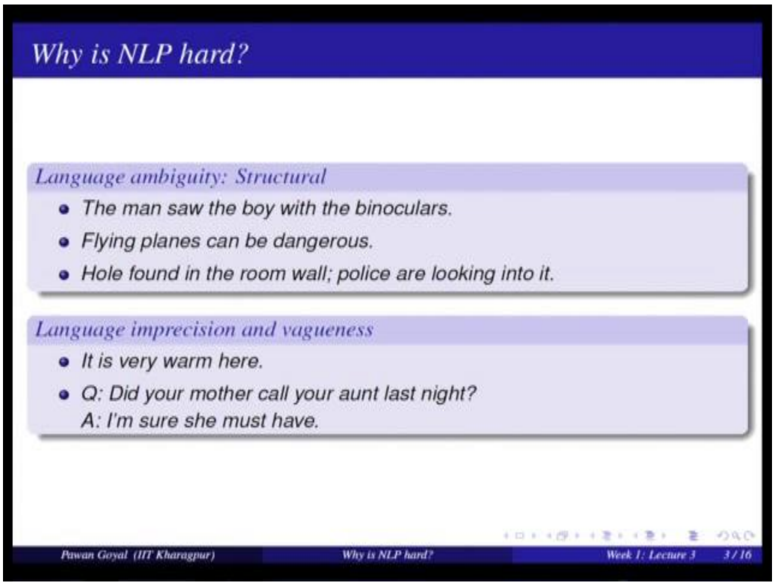
Flying planes can be dangerous.

Very warm here

So, let us try to identify various blocks here; Buffalo buffalo Buffalo buffalo buffalo buffalo Buffalo buffalo. So, let me try to give you what are the units together ok.

So, there are 3 units in this sentence. So, now, what would be the interpretation? So, these are the cities by C and the animal A, A and A, and these 2 are verbs. So, then this would be the interpretation of the sentence, buffaloes from buffalo Newyork whom buffaloes from buffalo Newyork bully, bully buffaloes from buffalo Newyork. Probably is not the sentence that you will encountered very often in the copes, but this is just to convey a point that in language you can actually use the same word and multiple different meanings and that creates a problem and this problem is called as Lexical ambiguity, the same word in lexicon is used in multiple different senses.

(Refer Slide Time: 07:46)



Why is NLP hard?

Language ambiguity: Structural

- *The man saw the boy with the binoculars.*
- *Flying planes can be dangerous.*
- *Hole found in the room wall; police are looking into it.*

Language imprecision and vagueness

- *It is very warm here.*
- *Q: Did your mother call your aunt last night?*
  A: I'm sure she must have.

Pawan Goyal (IIT Kharagpur) Why is NLP hard? Week 1: Lecture 3 3 / 16

So, now let us go to the next problem or again get to an ambiguity that is a structural ambiguity; what do I mean by structural ambiguity? So, I can have different interpretations of the same sentence. So, let us see the first sentence here. So, this is very very common example that we given NLP, the men saw the boy with the binoculars. So, can you try to see what is the ambiguity here? So, the ambiguity here is whether the binoculars are with the boy or the binoculars are with man; whether the men saw the boy with this binoculars or whether the men saw the boy who was standing with these binoculars. So, these are two different interpretations of the same sentence.

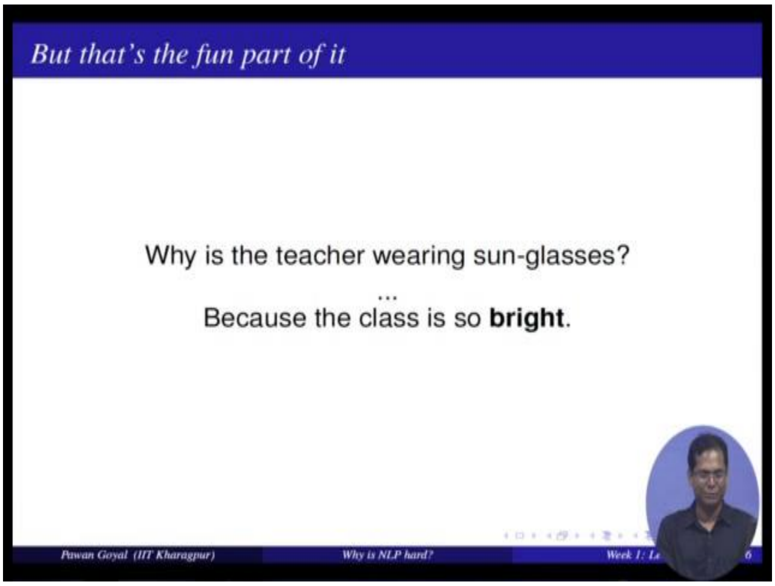
Let us see the other sentence here flying planes can be dangerous; can you see the two interpretations of the sentence, what is dangerous? Are the planes dangerous or flying dangerous? You see flying planes can be dangerous. So, whether the flying is dangerous or flying planes together is dangerous, these are the two interpretations of the same sentence. Similarly you can see a third sentence here - hole found in the room wall police are looking into it. So now, can you see the ambiguity in the sentence? So, ambiguity here is what are police looking in to or the looking into the whole or the looking in to the matter that there was a whole and we are looking in to the matter.

So, this is another problem that we face very very often with language, that this is ambiguous both in terms of lexical ambiguity; the same word can imply multiple meanings or a structural ambiguity where the same sentence can be interpreted in multiple ways.

And then we have some other problems like languages very imprecise and vague. So, what are the examples here? So, here is simple sentence, it is very warm here. So, can you see what is the vagueness or imprecision here? So, whenever I see a word like it is very warm here, I cannot tell for sure what would be the temperature there? If I am in India, for me warm many a temperature of only 40 degrees Celsius, but if I am in UK Europe for me warm might been 25 degree Celsius. And it might also depend on what was the weather in the last month and so. So this depends on a lot of context you find out what is the actual temperature that is been conveyed by this simple sentence.

Similarly if you see the other example; so have a question, did your mother called your aunt last night and the answer is I am sure she must have. So, what is the imprecision and vagueness here in the sentence? So, you see whenever I say I am sure she would have done that; that means, I do not know; if I know that I will say yes she has called, but whenever I am saying I am sure she must have. So, probably I do not know whether she has.

(Refer Slide Time: 10:53)



But that's the fun part of it

Why is the teacher wearing sun-glasses?

...

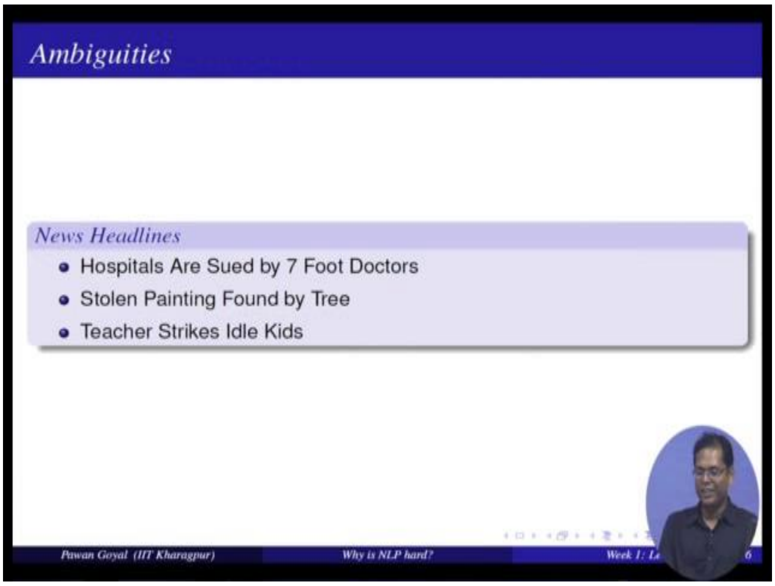
Because the class is so **bright**.

Pawan Goyal (IIT Kharagpur) Why is NLP hard? Week 1: Le 6

That is what I say this is the fun part of NLP with that helps in constructing a lot of jokes.

So, I have the symbol, this is a nice joke for the classes, so why is the teacher wearing sun glasses and you can given answer because the classes so bright and you can see the bright might mean either the class is bright, in the sense of lot of sun light and on or because the class is very bright in the sense of the students being really intelligent, fine.

(Refer Slide Time: 11:26)



Ambiguities

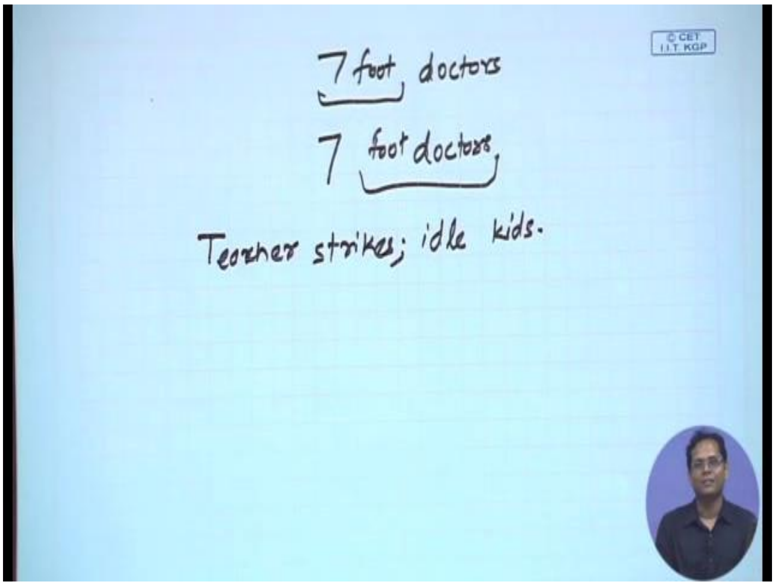
News Headlines

- Hospitals Are Sued by 7 Foot Doctors
- Stolen Painting Found by Tree
- Teacher Strikes Idle Kids

Pawan Goyal (IIT Kharagpur) Why is NLP hard? Week 1: Le 6

So, continuing on the same topic of ambiguities. So, let us see some other examples. So, that is something that we see will news headlines, which is the first sentence: hospitals are sued by 7 foot doctors, so can we see the ambiguity here? In general when you look at this sentence what comes to your mind? Probably the doctors as 7 foot, 7 foot doctors of course, we will 7 feet doctors, but yes this might you want to a interpretation, but what is implied by the sentence? There are 7 different doctors and they are all 4 doctors, so that is what is implied 7 different foot doctors.

(Refer Slide Time: 11:47)



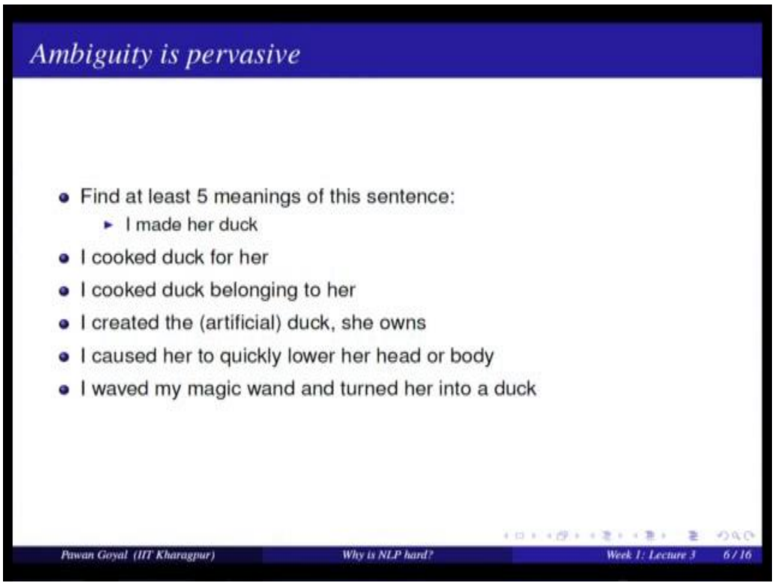
$\underbrace{\text{7 foot}}$ doctors

7 $\underbrace{\text{foot doctors}}$

Teacher strikes; idle kids.

Take the next sentence here, stolen painting found by tree. So, can you see the ambiguity? So, it looks as if the tree found the paintings, but what it means is that the paintings were found near the tree. Take the third example, teacher strikes idle kids. So, when you see the sentence as it is what comes to your mind? Teacher strikes some kids, but what the headline where meaning teacher is striking there is some semicolon and the kids are idle.

(Refer Slide Time: 12:55)

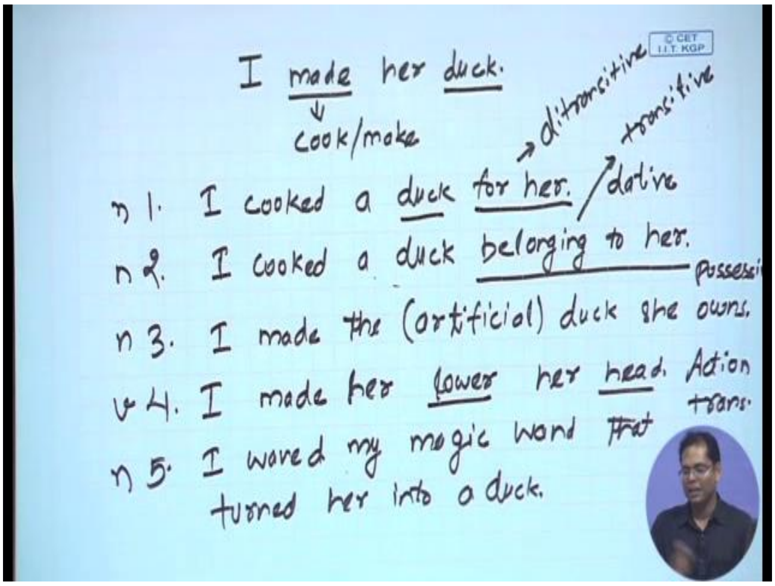


Ambiguity is pervasive

- Find at least 5 meanings of this sentence:
  - I made her duck
- I cooked duck for her
- I cooked duck belonging to her
- I created the (artificial) duck, she owns
- I caused her to quickly lower her head or body
- I waved my magic wand and turned her into a duck

So, let us take let us do a simple exercise on the same topic of ambiguity. So, I can give a simple sentence are made her duck. Now try to find, peoples there are 10 or more than 10 meanings for the sentence, but try to find at least 5 meanings of this sentence, I made her duck. So, let us do this exercise.

(Refer Slide Time: 13:20)



I made her duck.
cook/make

n 1. I cooked a duck for her. / dative

n 2. I cooked a duck belonging to her.

n 3. I made the (artificial) duck she owns.

v 4. I made her lower her head.

n 5. I waved my magic wand that turned her into a duck.

ditransitive
transitive
possessi
Action
trans.
© CET
I.I.T. KGP

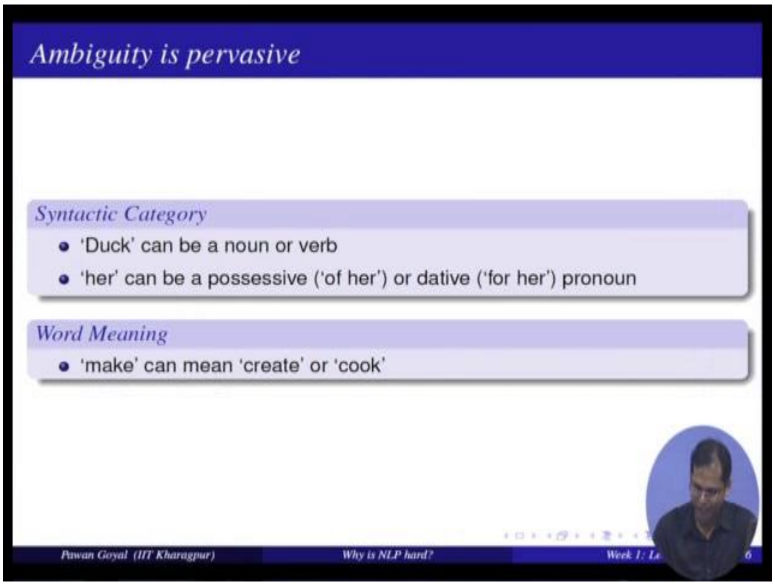
I made her duck. So, what are different meanings this sentence can take? So, we need to see what are the different interpretations each of the word can have in this sentence; so for example, the word made can mean for example, cook or make. So, one interpretation can be I cooked a duck for her. So, simple interpretation I cooked a duck for her that is one interpretation.

Now, in the same sense of cooking, I can also try to write a different interpretation what is that? So, one meaning is I made her duck another could be I made a duck that belong to her. I cooked a duck belonging to her. So, that is her duck, it may not, I might not have cooked for her, I might have cooked for myself, but I cooked the duck that belong to her I made her duck. Now made can also mean is simple making, so you can think of an artificial duck like a toy and I can say I made the artificial duck she owns of course, you can also have the second interpretation here, I made the artificial duck for her or belonging to her.

But let us take this interpretation, now what are the other two interpretations you can think of? So, now, try to think of the other meaning of duck, can duck be used as a verb? So, if you are listening to some cricket commentaries sometimes, the batsman duck whenever they were bouncer in sometimes so; that means, (Refer Time: 12:25) once head. So, one interpretation can be I made her lower her head that can be another interpretation. Now can you think of any different interpretation from all the four that we have seen till now? So, the hint is that try to go in the Harry potter mode. So, this is something like I waved my magic wand and converted her into duck, yes that is a possible would interpretation, I waved my magic wand that turned her into a duck ok.

So, you see the simple sentence I made here duck can have at least these 5 different interpretations; these are the 5 interpretations that we saw. So, now, what is in the language that gives rise to all these different interpretations? So, let us try to look closely.

(Refer Slide Time: 16:37)



Ambiguity is pervasive

Syntactic Category

- 'Duck' can be a noun or verb
- 'her' can be a possessive ('of her') or dative ('for her') pronoun

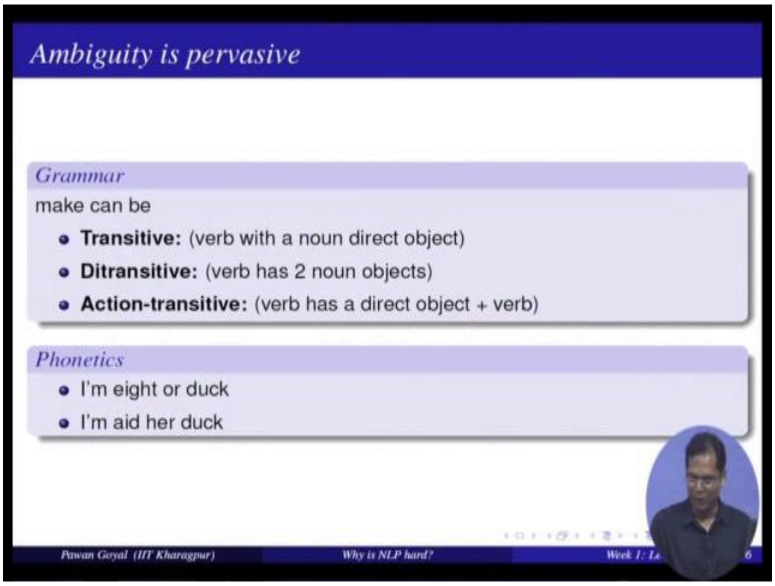
Word Meaning

- 'make' can mean 'create' or 'cook'

So, this ambiguity is pervasive everywhere, so how? So, one thing is about the syntactic category; what is the role of a word in a sentence. So, you see the word duck here it can mean either a noun or a verb. So, can you label the sentences here, where duck has been uses a noun? So, this is noun, this is noun, this is noun, this is verb, this is noun. So, fine, these are two interpretations that can be noun or a verb.

Then there is a case with her, the word her can either be a possessive of her or dative for her. So, can you see the two examples here: the two interpretations which were made because of these; this was dative I did it for her or this is possessive belonging to her. So, these are again two interpretations for this ambiguity in language then we so make can mean either to create or cook this. So, this for cooking and this was for making, what is?

(Refer Slide Time: 17:58)



Ambiguity is pervasive

Grammar

make can be

- **Transitive:** (verb with a noun direct object)
- **Ditransitive:** (verb has 2 noun objects)
- **Action-transitive:** (verb has a direct object + verb)

Phonetics

- I'm eight or duck
- I'm aid her duck

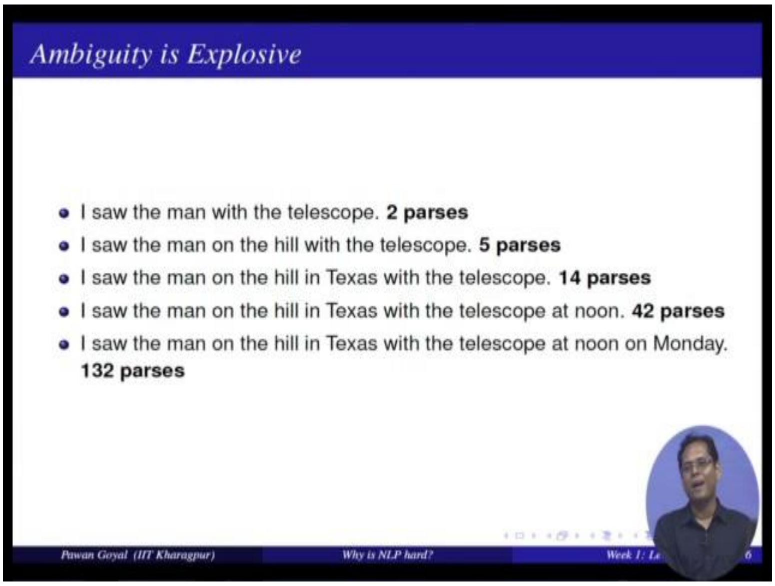
Pawan Goyal (IIT Kharagpur) Why is NLP hard? Week 1: Le 6

Then if you go to grammar, the same work make can be either transitive; that means, it will be a verb with a single noun as direct objective, it can be ditransitive; that means, a verb that is having two different non objects or action transitive, it has a direct object plus a verb.

So, in these 5 interpretations can you try to mark them, where the verb make was uses transitive, ditransitive and action transitive? So here I have the same word make which is having an object and a verb. I made her lower her head or I made her do something. So, this is action transitive; what is ditransitive? So where there are two objects of the same verb; I cooked a duck for her. So, this as two different objects and what is the single transitive? I could they duck belonging to her this is a single object. So, here there are two different objects: this is ditransitive and this is transitive. So, in language the same verb may can be use any of these 3 different vague and that gives me 3 different interpretations.

So, now, suppose you go to phonetics. So, I am speaking this sentence, I made her duck what are different interpretations you can think of? So, what happens in speech recognition? I am a spin something and you have to transcript. So, whenever I see I made her duck you might thing of all these possible transcriptions. So, like I am eight or duck, I am aid her duck all these are possible, but the problem that NLP decisions will face is going there are many different possible interpretations, which going to choose in a given context.

(Refer Slide Time: 20:04)



Ambiguity is Explosive

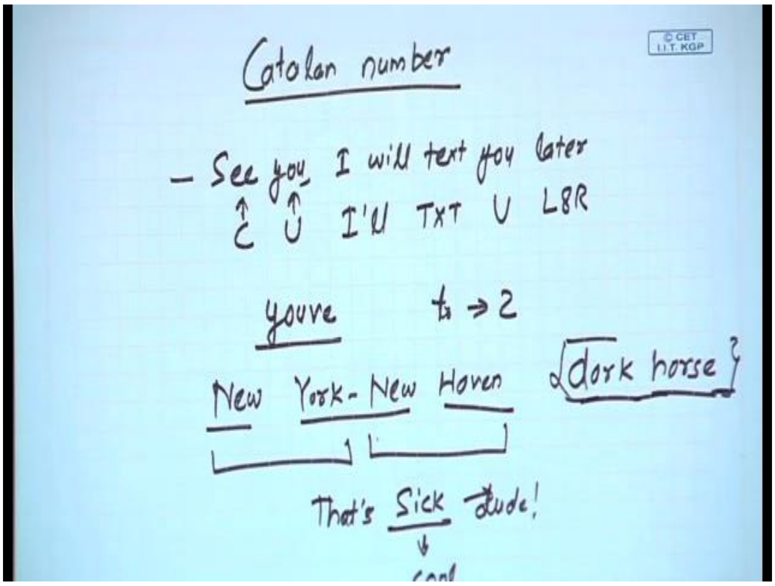
- I saw the man with the telescope. **2 parses**
- I saw the man on the hill with the telescope. **5 parses**
- I saw the man on the hill in Texas with the telescope. **14 parses**
- I saw the man on the hill in Texas with the telescope at noon. **42 parses**
- I saw the man on the hill in Texas with the telescope at noon on Monday. **132 parses**

Pawan Goyal (IIT Kharagpur) Why is NLP hard? Week 1: Le 6

So, let see something from the (Refer Time: 20:07) ambiguity now. So, I have the simple sentence I saw the men with the telescope and we saw earlier it can have two different parses. So, parses I mean different ways in which the words can be connected in this sentence.

So, we will have a complete topic on parses is give we will discuss there in detail, but I guess you can see at least the idea from the last example. So, now, suppose I try to increase the length of the sentence, I saw the man on the hill with the telescope and immediately you can find 5 different parses of the sentence; this does not stop here. So, let me have bit this sentence I saw the man on the hill in Texas with the telescope. So, now, it has 14 parses, if I say I saw the man on the hill in Texas with the telescope at noon it has 42 parses and if I say I saw the man on the hill in Texas with the telescope at noon on Monday it look 132 parses and you can actually laid these numbers to something call a Catalan number.

(Refer Slide Time: 21:06)



Catalan number

– See you I will text you later

C U I'll TXT U L8R

youre to → 2

New York-New Haven

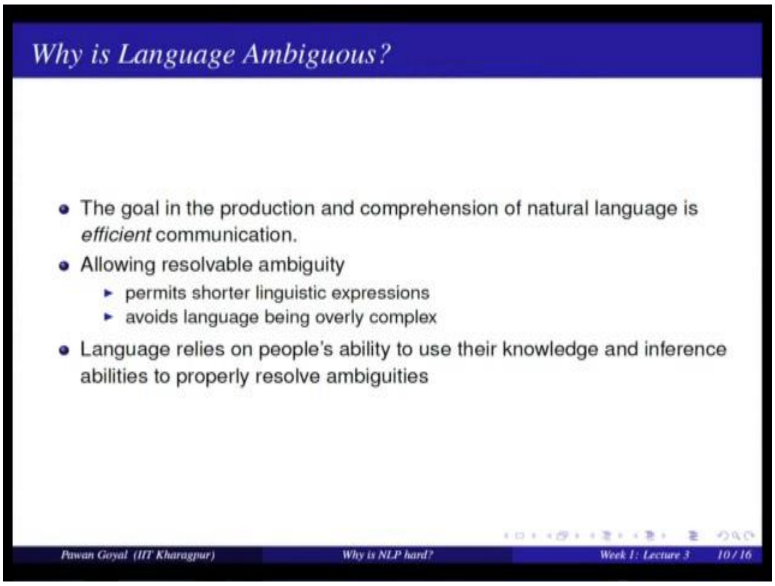
{dark horse}

That's Sick dude!

cool

So, as you keep on increasing the number of phases in the sentence, the number of interpretations in which they can be connected in the sentence increase.

(Refer Slide Time: 21:21)



Why is Language Ambiguous?

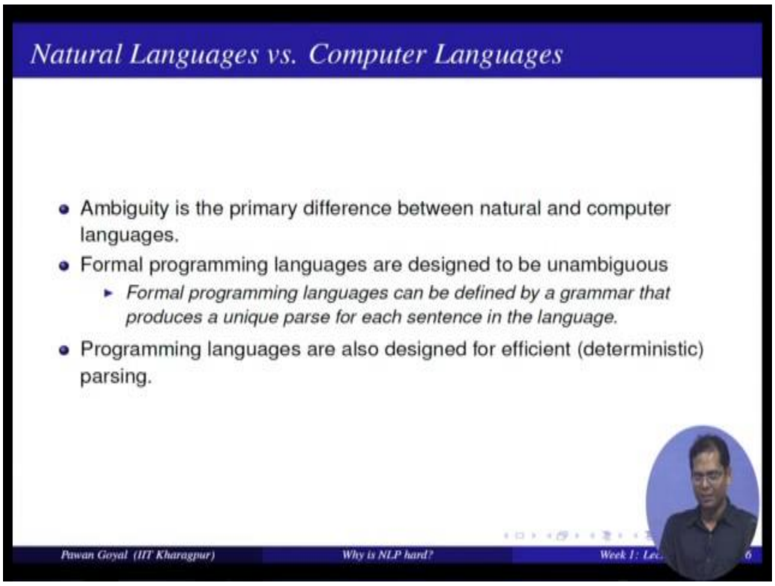
- The goal in the production and comprehension of natural language is *efficient* communication.
- Allowing resolvable ambiguity
  - permits shorter linguistic expressions
  - avoids language being overly complex
- Language relies on people's ability to use their knowledge and inference abilities to properly resolve ambiguities

Pawan Goyal (IIT Kharagpur) Why is NLP hard? Week 1: Lecture 3 10 / 16

So, now why is language ambiguous? This can be a nice question that why is language ambiguous at all. So, we need to understand what the goal of language as such. So, language is used for communication. So, the goal introduction or communication languages to be able to communicate ideas clearly, but at the same time I should clear certain restrictions or that then that a post like I should have shorter linguistic expressions. So, think of the previous example that we saw in the last slide, the same sentence was having 132 different interpretations. Now, what if I need to have a different sentence for each of these 132 different interpretations that will make by language expression really large and also the language will become very complex.

So, what happens in language? Some sort of ambiguities allowed, but is ambiguity is some that is easily resolvable; so you cannot have an ambiguity there is not resolvable at all; by some sort of knowledge that you have human being tries to resolve it, but in the case of NLP we try to we have, we have this starts for developing algorithm that kind resolve this ambiguity. So, yes language relies mostly on the people's ability to use their knowledge and some inference capabilities to properly to resolve these ambiguities.

(Refer Slide Time: 22:58)



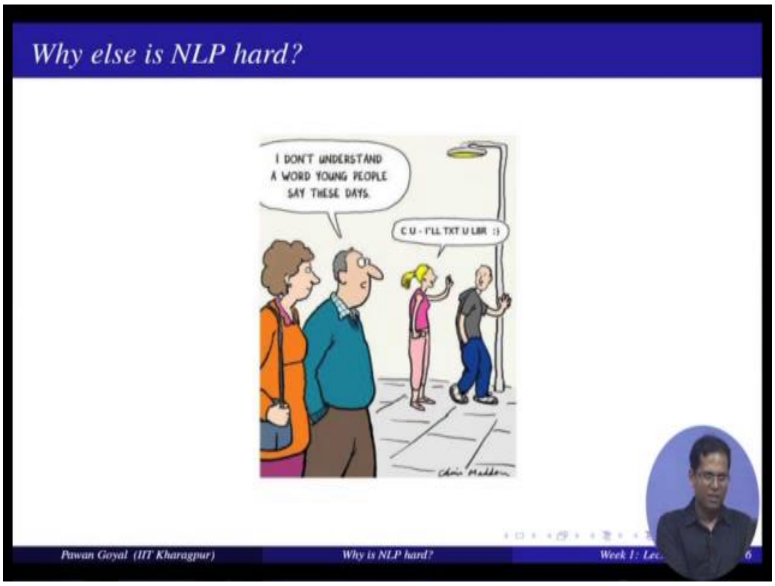
Natural Languages vs. Computer Languages

- Ambiguity is the primary difference between natural and computer languages.
- Formal programming languages are designed to be unambiguous
  - *Formal programming languages can be defined by a grammar that produces a unique parse for each sentence in the language.*
- Programming languages are also designed for efficient (deterministic) parsing.

So, now, just a brief discussion on what is the difference between a natural language and a computer language as such; one primary difference is the ambiguity. Programming language is do not have any ambiguity in, so whenever you write a program it will mean only one particular thing these no ambiguity there, but that is not the same that is not true for the case of language in natural language.

So, all the programming languages are formal and they are designed to be unambiguous so that you can have very very efficient passing for that. So, they can always defined by a grammar, that process is unique parse for each sentence or each programming construct in the language that is not true for the natural languages, so you can have parsing in linear time.

(Refer Slide Time: 23:50)

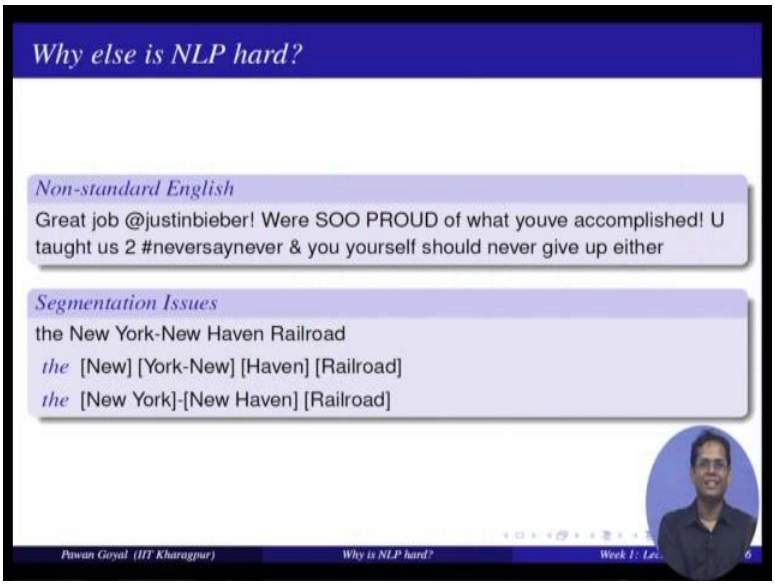


Why else is NLP hard?

Pawan Goyal (IIT Kharagpur) Why is NLP hard? Week 1: Lec. 6

Now why else is NLP hard? So, this carton gives you some idea that I do not understand a word young people say these days. So, I am talking about social media here. So, see you, I will text you later. The sentence see you, I will text you later now your writing is C U I’ll TXT U L8R. So, now the problem at hand is to understand that that you have actually meaning the sentence and find out a starting from C U mean this see, from U the mean this you and so on.

(Refer Slide Time: 24:46)



Why else is NLP hard?

Non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either

Segmentation Issues

the New York-New Haven Railroad

the [New] [York-New] [Haven] [Railroad]

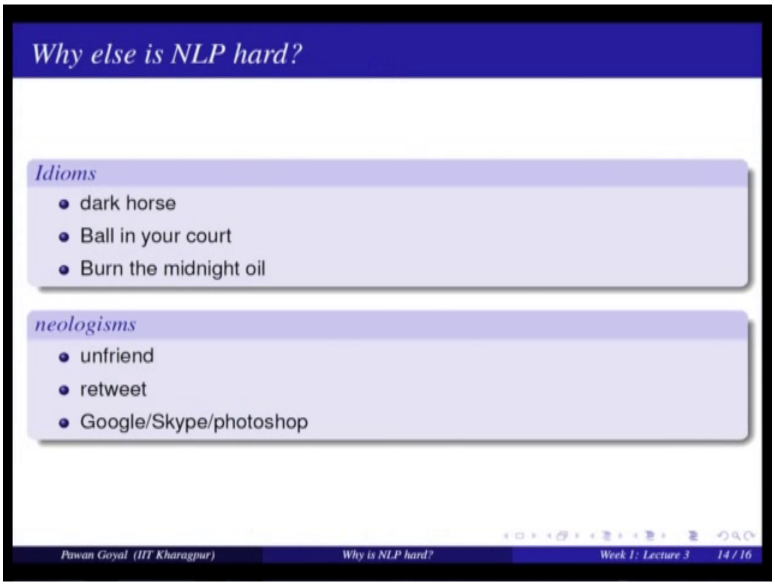
the [New York]-[New Haven] [Railroad]

So, this is also called the non standard use of English and that is very very problem with the use of social media, all the SMS and other plate forms.

So, let us see this particular tweet. So, great job Justin Bieber we are so proud of what you have accomplished, you taught us to neversaynever and you yourself should never give up either. So, in this sentence see it is a tweet. So, what are the things that you do not see in formal language? So, one example here is mentioning the use of mention at justinbieber. So, this mention is again wherever is specific to tweet and you are using hash tag neversaynever – this is in the hash tag; what is? So, we are seeing constructs like this, you have is written as you have you v e and to is written as 2. So, this is again non standard use of English that makes NLP difficult. Then there are many other problems like segmentation issue. So, I have the simple sentence the New York-New Heaven Railroad.

So, based by segment this particular sentence, especially New York-New Haven, should I segmented like that? New York- New Haven or New York, New Haven the second is the correct segmentation. So, there are two possible segmentation of the same sentence.

(Refer Slide Time: 26:31)



Why else is NLP hard?

Idioms

- dark horse
- Ball in your court
- Burn the midnight oil

neologisms

- unfriend
- retweet
- Google/Skype/photoshop

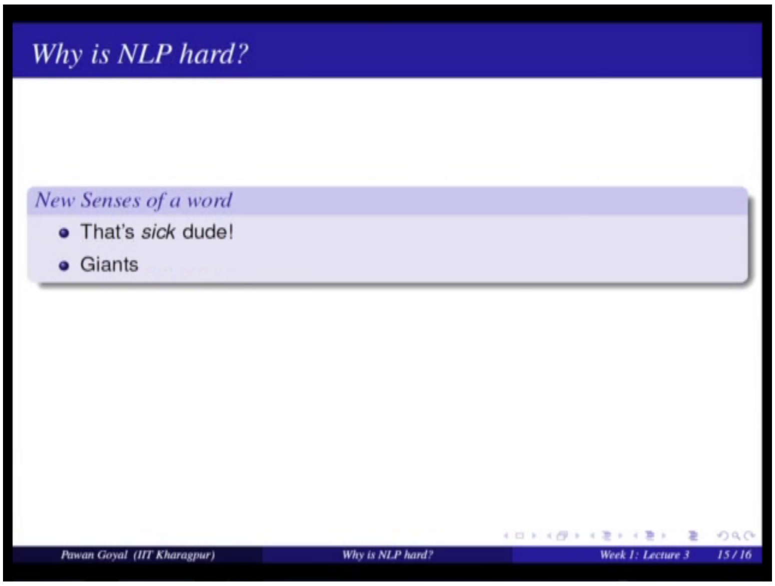
Pawan Goyal (IIT Kharagpur) Why is NLP hard? Week 1: Lecture 3 14 / 16

So, problem would another would be to find out what is the correct segmentation given this sentence, then there are other cases like the case of idioms in language. So, what happens in the case of idioms you cannot construct the meaning of the page by looking at the meaning of the individual words and trying compose them together. So, if I say idiom like dark horse, I cannot take the meaning of a horse which is dark and make the meaning of this idiom; dark house is something else, so this is some sort of idiom that is conveyed for the person who is not well known, but he suddenly becomes dark horse. So, the he is suddenly excelling in certain field.

Similarly you see this idiom ball in your court, does not mean the ball is in your court; that means the matter is at your hand and now it is your turn. Similarly this idiom is again very much used burn the midnight oil, there is not mean that I am burning the oil at midnight, but what it means is that you are doing hard work. Then there are various new words that have been floated new usages that are coming up, so they also called new (Refer Time: 27:49). So, for a one particular example, these are examples are taken from social media is unfriending. I have some facebook friend and I am unfriending. So, this has become is a verb itself; retweet - retweet has been used a lot as a verb, similar Google, Skype, Photoshop all are very much used in as verbs and Googling and all.

So, that creates another problem for NLP that is new words are coming up in different and new usages, so do not have a closed vocabulary and your vocabulary keeps on increasing.

(Refer Slide Time: 28:29)



Why is NLP hard?

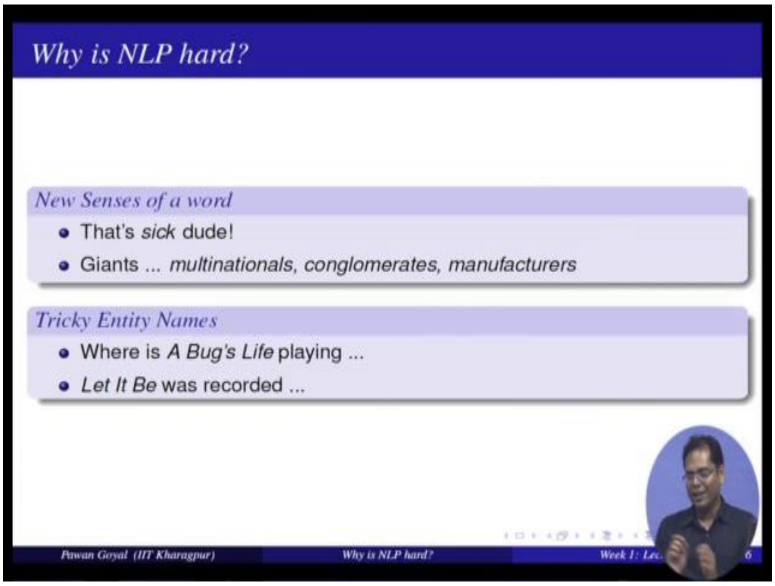
New Senses of a word

- That's *sick* dude!
- Giants

Pawan Goyal (IIT Kharagpur) Why is NLP hard? Week 1: Lecture 3 15 / 16

Then you keep on getting new senses of the words. So, see this particular example - That’s sick dude! So, I am pointing here to the word sick. So, what is the usual meaning of sick that you see? Sick I will something that is not healthy, that is someone who is ill, but in this particular sentence that is sick dude, sick mean sick means something that is cool and that meaning is coming up with social media. Similarly you see the word giants, what is the particular meaning of giants seasonally we see giants as some sort of demons that we have in storage. But recently some new meaning of giants is coming up, so that you can think of, like some sort of giant multi nationals and all that.

(Refer Slide Time: 29:20)



Why is NLP hard?

New Senses of a word

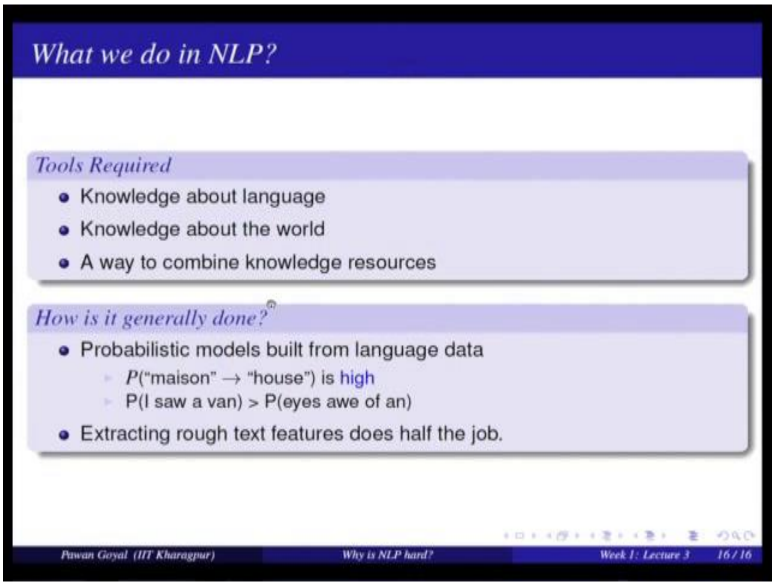
- That's *sick* dude!
- Giants ... *multinationals, conglomerates, manufacturers*

Tricky Entity Names

- Where is *A Bug's Life* playing ...
- *Let It Be* was recorded ...

Giant manufactures and then there is a problem of entity names. So, take this sentence where is a bug's life playing? You cannot understand the meaning of the sentence un less you know that this particular utterance a bugs life is single entity and then you try to understand the meaning of the sentence. Similarly the second sentence let it be was recorded you need to understand that let it be is single entity here.

(Refer Slide Time: 29:52)



What we do in NLP?

Tools Required

- Knowledge about language
- Knowledge about the world
- A way to combine knowledge resources

How is it generally done?

- Probabilistic models built from language data
  - *P*("maison" → "house") is high
  - P(I saw a van) > P(eyes awe of an)
- Extracting rough text features does half the job.

So, what we do in NLP to handle all that? So, we need to have some knowledge the language how the sentence is a constructed, what kind of words are there and so on.

We need to knowledge about the world and we need to have the way in which we will combine various knowledge resources in an efficient manner. So, how is it generally done? So, most of the times we do it by using various probabilistic models; this is the single simple example here. So, I have word like [FL] in French and may translate in English and there are multiple interpretations, so I want to use a model that (Refer Time: 30:30), the probability that [FL] goes to house is high and I should choose this particular interpretation. Similarly for the next sentences suppose this is for this speech recognition system, whenever I am saying I saw a van and there are two interpretations I saw a van or eyes of a n I should be able to say that the first interpretation is more probable to occur in the language then the second interpretation.

And we will deal with this a lot. So, many a times we have to extract lot of text a feature that does most of this job, fine. In this lecture we covered some aspects of wild languages hard. So, in the next lecture we will start talking about some of the very very basic empirical large that we see in language and we will start with doing some basic pre processing.