Applied Linguistics
Dr Amit Sethi
Electrical and Electronics Engineering
Language, computers and Applications (Spell Checker)
Indian Institute of Technology Guwahati
Lecture 30
Language Computers and Applications

We have looked at the relationship between language and computers. We studied language, we have seen several components of language and in this course we are looking at the application aspects of those components of the study of language in various domains of society and our lives. In order to understand that, last time we discussed an interaction between language and computers that is language and machines.

We touched upon very fundamental relationship between the 2 and we only looked at some of the areas where such an interaction has contributed to some extent. We also tried to understand the encoding and decoding and production and understanding system and the requirement of interpretation at pragmatic level and the capabilities of machines in dealing with the structural ambiguities involved and the passing of chunks of the structure.

The way human brain does when we study human language, when we, we have developed theories in the study of human language that is natural language, which can contribute to a great extent. We have seen some of these aspects, today I invite you to listen to a talk by Dr Amit Sethi who teaches at the Department of Electrical and Electronics engineering at the Indian Institute of technology Guwahati.

He works in computer vision and making machine more intelligent for variety of functions. He is going to talk about how spellcheckers have developed in this area of language and computers and how it works. That will give us an insight into applicability of the interactions between language and computers.

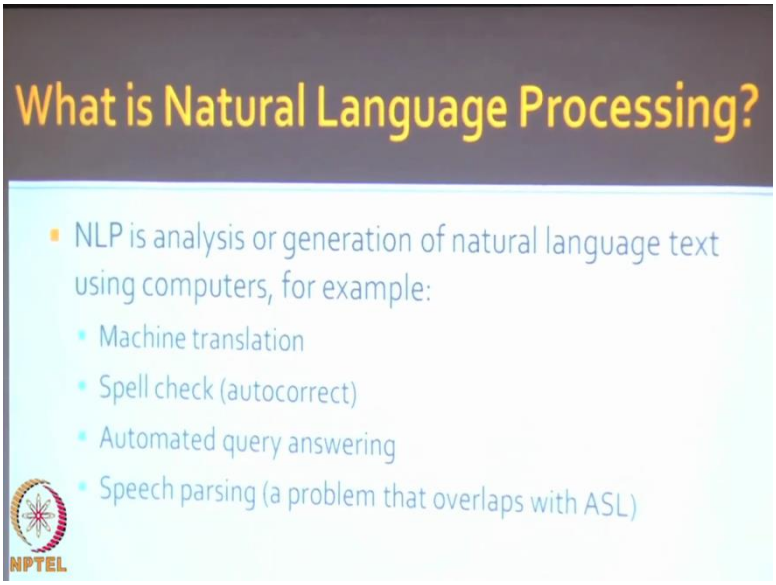
So listen to this and like I mentioned in the last module that you come up with more areas you know as exercise for this domain of Computer Science and the study of language that, what are those areas to some extent new ideas where such an interaction has potential to contribute in making machines more intelligent and changing the lives further, thank you.

(Presentation of Dr Amit Sethi starts from here 3:33)

Good morning everyone and thank you for having me here. I will give a talk on natural language processing through an example, how spellcheckers work and it will I will try to

make the talk keep it simple partly because I am also not an expert in NLP. I do teach little bit of NLP in my machine learning for, but hopefully this will be accessible to everyone.

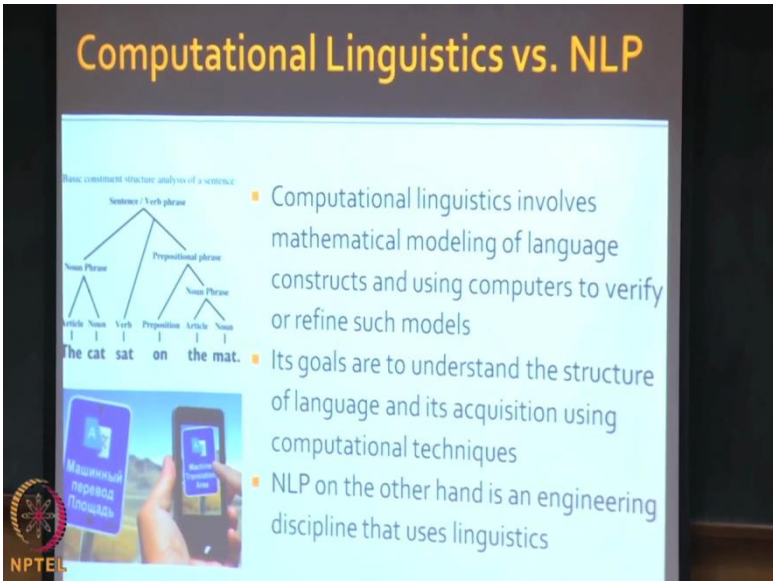
(Refer Slide Time: 04:30)



So what is NLP? What do you understand by NLP? Basically, I want to get to what is the difference between NLP and computational linguistics, okay. So, NLP is actually an engineering discipline. It is basically solving the problems of understanding natural language text, okay using engineering method. So it uses linguistics.

So for example, it could be used for machine translation spell check or auto correct automated query analysis, you ask computer a question and it will give you an answer, okay. Even things like um automated speech recognition uses natural language processing, so that overlaps with ASL. And it is based on probability of statistics, machine learning, linguistics and of course commonsense.

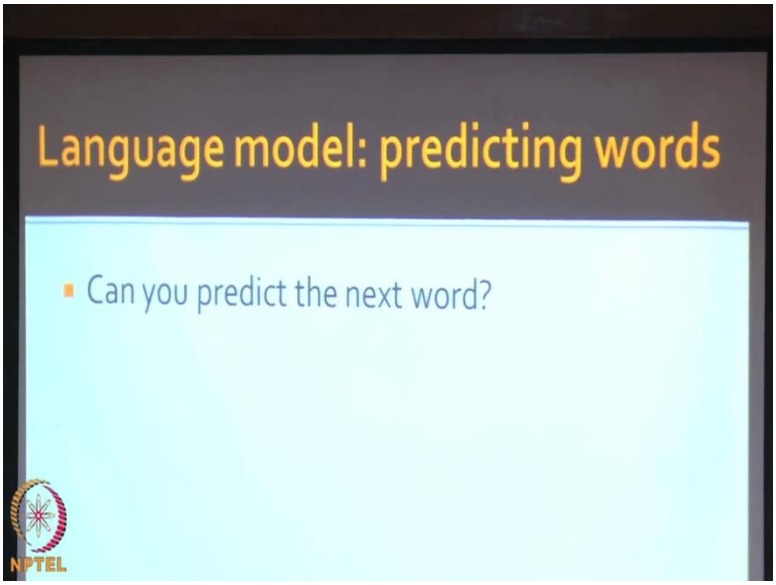
(Refer Slide Time: 04:59)



How is it different from computational linguistics? Computational linguistics is a linguistic, it is a it is a discipline in linguistics where you are forming mathematical and computational languages computational models for language. And you are using computational methods to understand the nature of language itself and to understand how language is acquired.

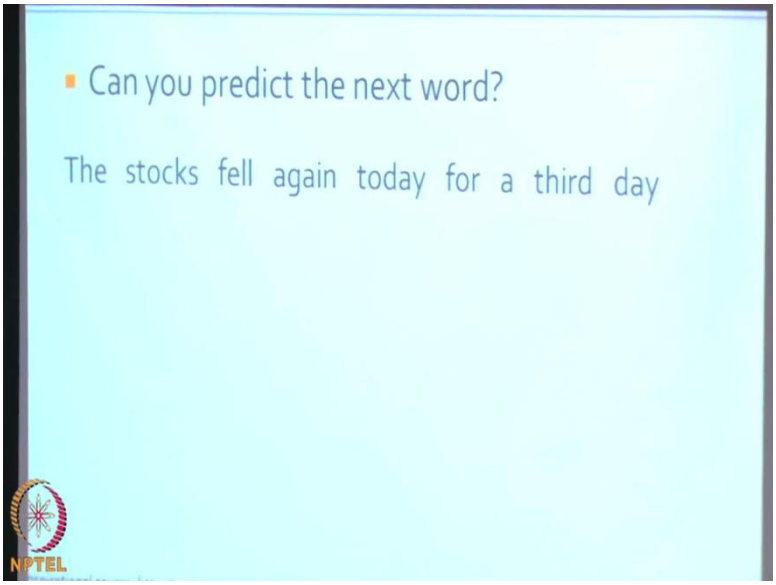
So it is like computational biology, you are actually doing biology but using computational method. So that is computational linguistics, doing linguistic using computational method, so their goals are very different. NLP on the other hand, is engineering discipline, so I just gave some examples, so like this pass tree will be very familiar to some of people here that come from linguistics.

(Refer Slide Time: 05:58)



Whereas, this is NLP where you are taking some text in a language and translating it into another language. So we will start with a discussion on what is a language model and how it is useful in our main topic which is spell checkers, but before we get to spell checkers I want to talk about some computational models of a language which may be which are mainly coming from NLP said, not from computational linguistics.

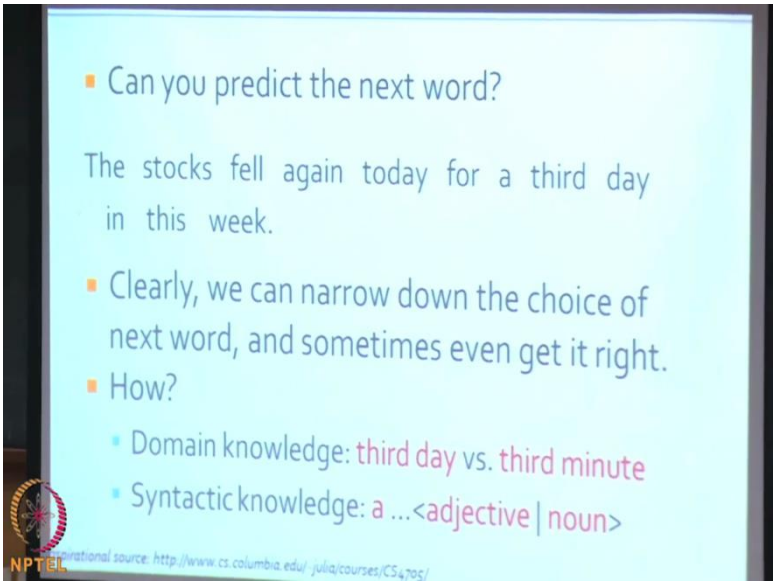
(Refer Slide Time: 06:48)



So we will next start with an exercise, can you predict the next word? I will give you a word in a sentence and I will give you word at a time and let us see if you can predict the next word, okay. What do most sentences start with? '"The" mostly right? So "the, stocks, fell, again, today, for, a, third," what? Third time? Could it be third minute? No, right? Nobody looks at stocks every minute unless you are high frequently trader.

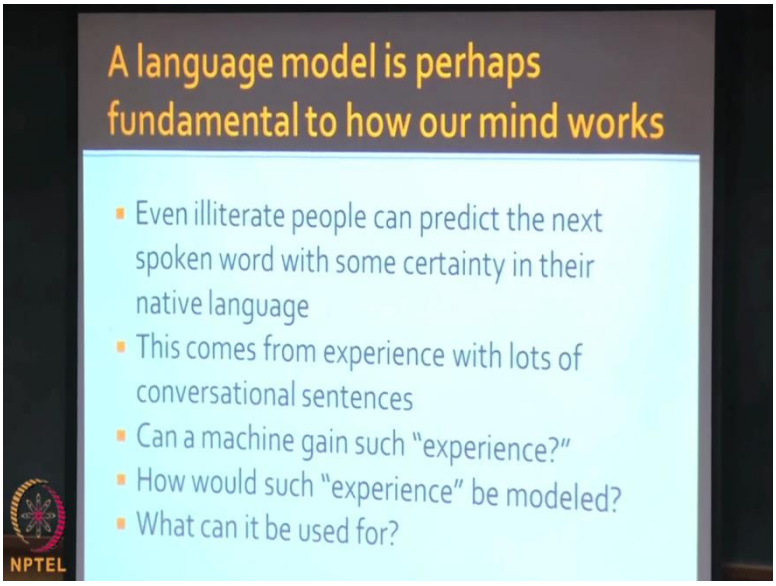
"Third Day or third time" "in" "this" "week". So you can guess what the next word is coming, right? It is not it is not um given the context, given the previous word; you have very good idea of what is coming, okay. But how do we do that? We do that based on several types of knowledge, right? We have domain knowledge for example, third Day versus third minute.

(Refer Slide Time: 07:18)



You did not say third minute because you know that when you read news about stocks, it is day to day and not minute to minute. And then syntactic knowledge, so here what I mean by this is like there is an 'a' which is here and then you talk about this "third". "Third" is an adjective and you know that, what is coming next is a noun, okay. So by because of that you know what word is coming next.

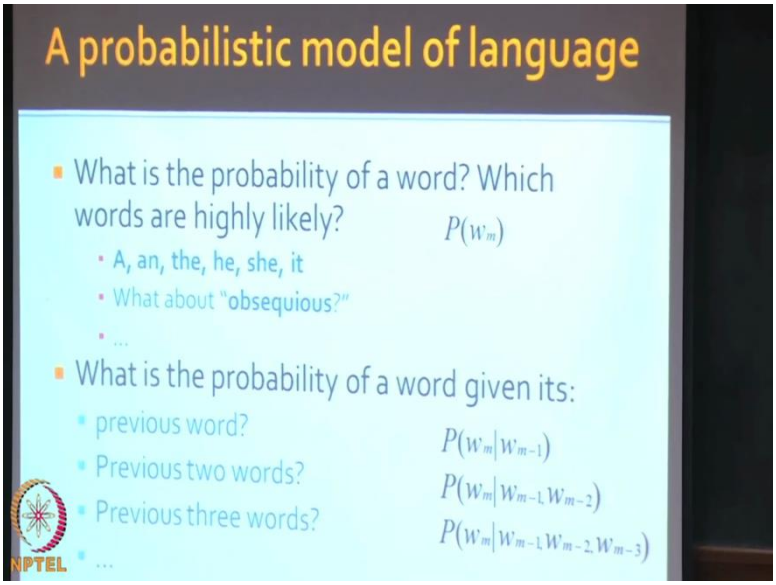
(Refer Slide Time: 07:48)



So language is fundamental to how our brains work. Even an illiterate person, they do not have to be Panini or Noam Chomsky. Even an illiterate person, when they hear parts of a sentence, they know what is coming next. Very true, it is very fundamental to us and this

comes from experience of lots of conversational sentences, right? And our goal is to find out how a machine can gain such an experience, okay.

(Refer Slide Time: 08:27)



And how will that experience the modeled and what will it be used for and of course it will be used for spell checking also. So I want to start with a very simple probabilistic model of a language. So please do not get um too bothered by the mathematical concept that you see there, they are pretty simple, I will explain very quickly okay. So, we will start with the notion of probability of a word, what does that mean?

Which words are more frequent, those words have high probability. So we are just looking at count of word and let us say all the sentences that you have heard in your life, the count of if you take all the words, the ones that will be most frequent will probably be some of these right, "a, an, the," article and prepositions and so. A word like "obsequious" will probably be very improbable; you probably have heard it may be 2 times in your life, right.

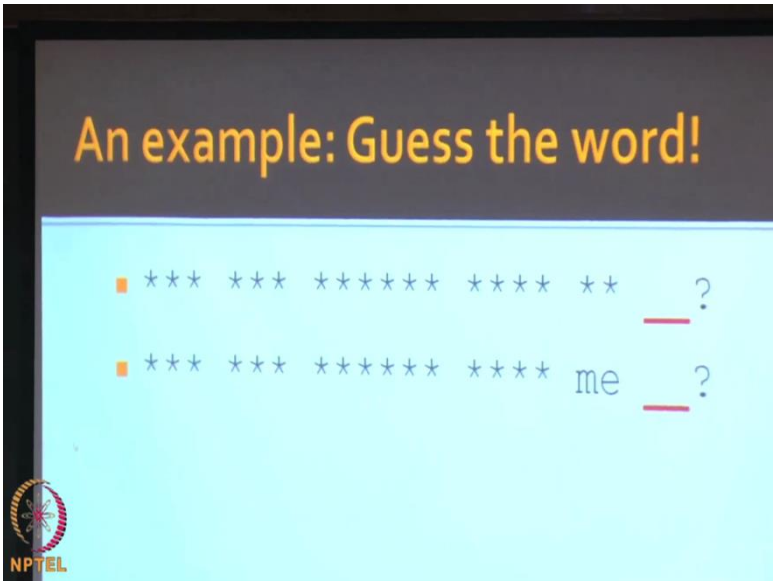
So we already know that if we have word W then it has a probability associated with it. What is 'm'? 'm' is let us say you have a sentence and or a document or a conversation, it is the mth word, starting with W 1, W 2, W 3 up to m. So the probability of W word 'Wm' is some probability, but this does not use any language of what has come before this, it is just based on the frequency of words that you have seen, so we will refine the model.

So, we definitely have a lot of information because of the things that we have heard before this coming word mth word, right? So let us look another refined model which is, P of W m given W m minus 1. What does that mean? Probability of word at mth location this line

means given, given the word at the previous location. So if you know the previous word is a is an adjective, you know that this word will most likely be a noun.

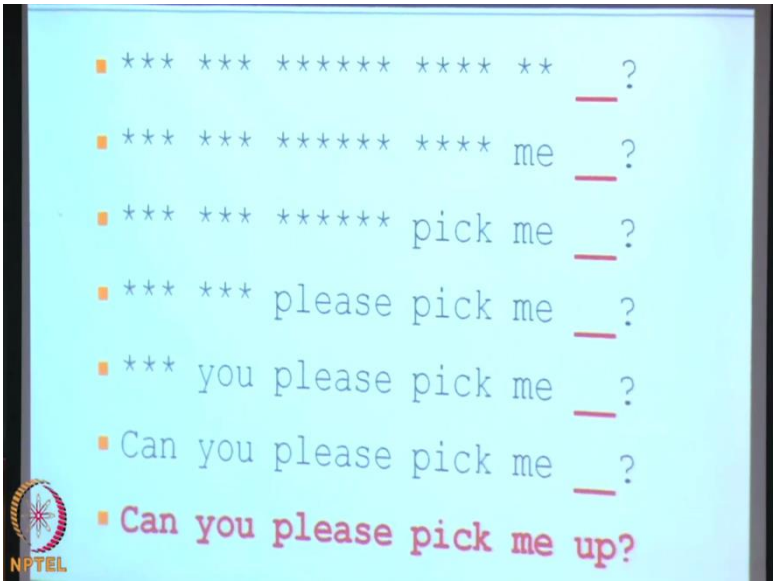
If it is an adverb, then it will be a verb, so that narrows down your choices of words, okay. So now you will discard all the other words that do not fit this model. So when you have a pair of words, a model based on a pair of words is called a bigram model 2 gram, bigram model. Similarly if you look at 2 previous words, for the current word m if you look at m minus 1 and m minus 2, then it will be a trigram model, okay, so and then 3 previous words.

(Refer Slide Time: 11:06)



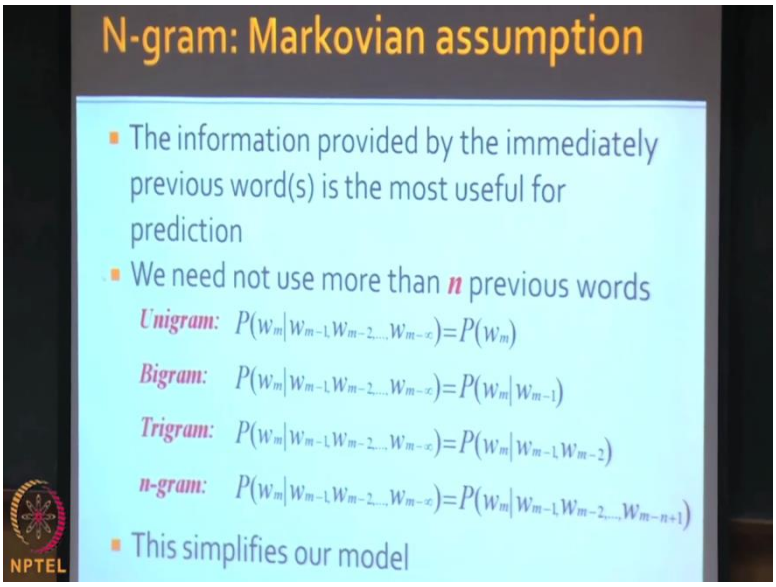
So here you are trying to guess the current word based on 3 previous, okay that is a Quadrigram model. So another game of a similar type let us say you are given a sentence and I have I have deleted all the words, you have to guess the last word, okay. It is very difficult to guess it without knowing the previous word. Now if I include one more word, you probably have some of what that would be.

(Refer Slide Time: 11:35)



What it could be? It is a, let me give you a hint, it is a 2 letter word. So I have narrowed it down very much. Yeah, something like that right, "up" "pick me up" right. So you needed to go only up to here to guess the word, you did not even need to go all the way down okay. So trigram model itself is very powerful in guessing the next word. You do not have to go to Quadrigram and higher order probability model.

(Refer Slide Time: 11:51)

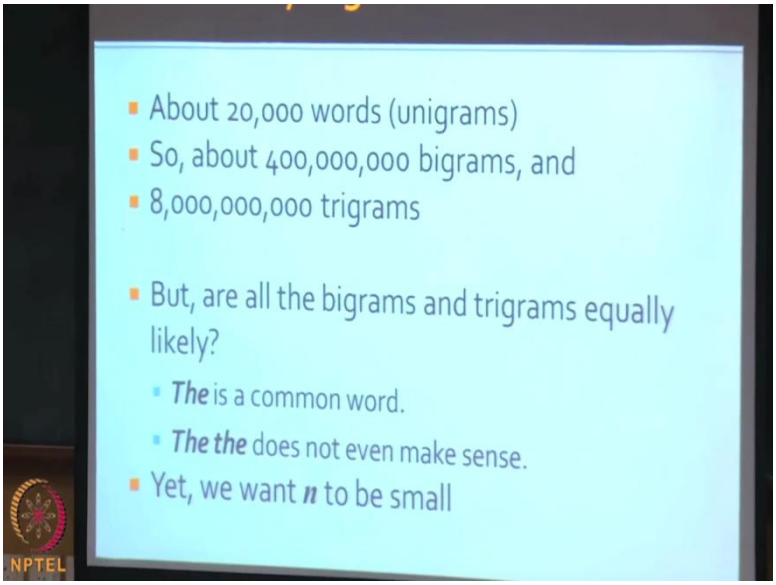


So, when we think of these models what we are saying is that we need to know only previous n words and n is a small number. In this case, we needed only the first only the 3 previous words or 2 previous words, okay. So if you look at a word and you are trying to guess it, you

are trying to guess its probability from all possible words, so let us say in conversational English we have 20,000 words.

So out of 20,000, which word it is and you base it on all the previous words, but you approximate this just that word itself, that is a unigram model. If you approximate it using the previous word, then it is a bigram model. If you approximate it using 2 previous words, then it is a trigram and n minus 1 previous one, then it is a n-gram model, okay so this simplifies the model.

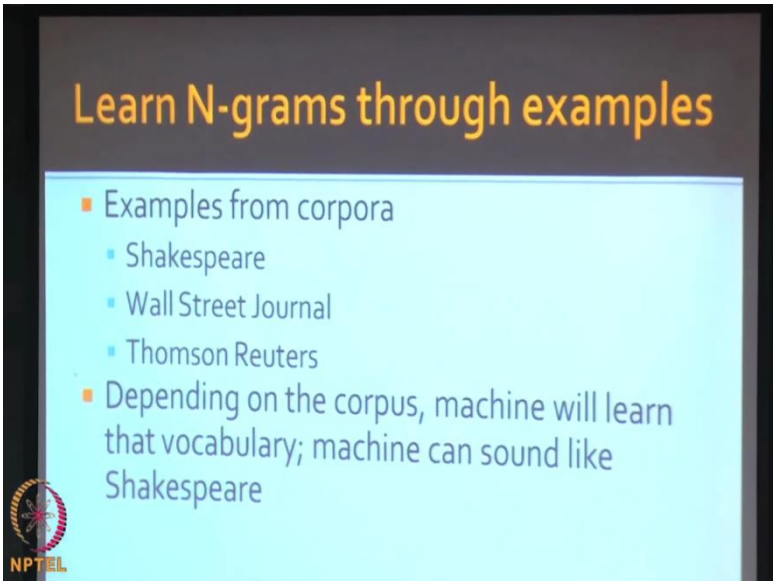
(Refer Slide Time: 12:42)



So, but how many n-grams are there okay, so if you have 20,000 conversational words, for a bigram you will have square of 20,000 bi bigrams, right? You take any word from and any word for the second position. So you have 400 million bigrams, but they are not all use in language, right. For example, “the” is a common word, but “the the” does not even make sense.

So that bigram actually does not even figure in your language, so your actual space of all bigrams and trigrams is much less than what is all possible combinations of 2 words and 3 words. So when you make language models, you take that into account and you only learn the ones that you have encountered, okay. But where do you encounter them?

(Refer Slide Time: 13:37)

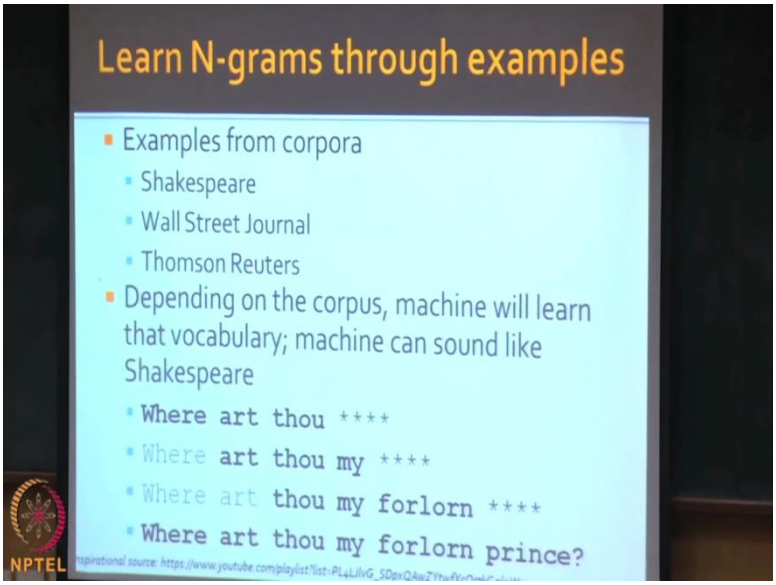


You encounter them when you train these language models, you look at some corpora. Corpora are basically large collection of facts, texts. So for example, all works of Shakespeare is a corpora corpus. Then you have Wall Street Journal, lots of articles of Wall Street Journal are available online for people to train machines on and Thomson Reuters okay.

So, now if you train your bigram model or a trigram model on one of these corpuses, then and you start to generate random sentences using that bigram model, so by random sentences I mean, you give it 3 words and you ask it to get the next one and whatever is the most probable one that put it there and then the next one and next one, it will construct something that looks like a sentence, but it may not be a complete sentence.

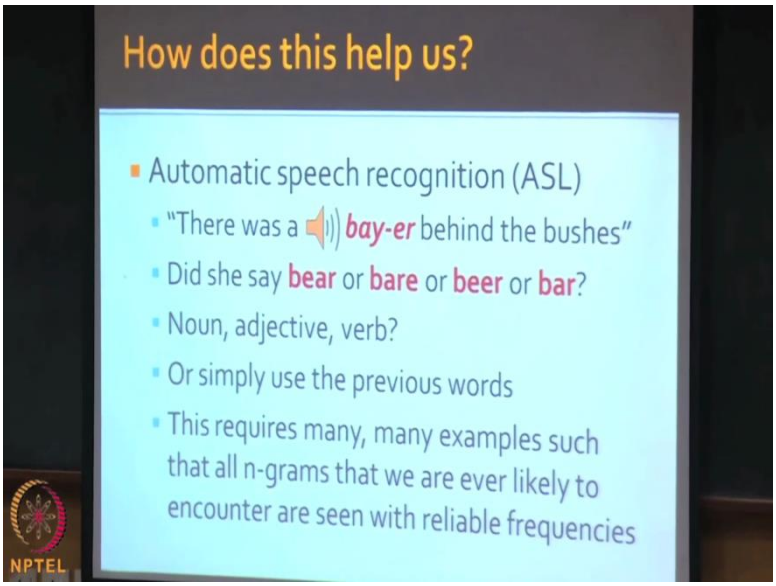
It turns out that if you use a Quadrigram gram model, where you are looking at only 4 possible 1 word depending on its previous words, you start to generate things that kind of start to make sense. Not only that, interestingly if you train it on Shakespeare, it sounds like Shakespeare, it would not sound like Wall Street Journal okay. So this is an example of machine generated.

(Refer Slide Time: 15:14)



So machine was given 3 words and it had to guess the next word okay. So it came with "my" and using these 3, it has to generate the next one, okay. So then it in the next step, it came up with the word "forlorn" and then "prince" okay. So now this may or may not be in Shakespeare, but it kind of sounds like Shakespeare now. So that is the machine generated random sentences based on just knowledge of 3 previous words, okay.

(Refer Slide Time: 15:46)



And it does not sound like Wall Street Journal at all because it was trained on Shakespeare, not on Wall Street Journal. So how does it help us? Of course it helps us in automatic speech recognition, right. So for example, let us say there was a sentence that you that you have a recording for and it says "bear was a" it sounds like "bay-er behind the bushes" right?

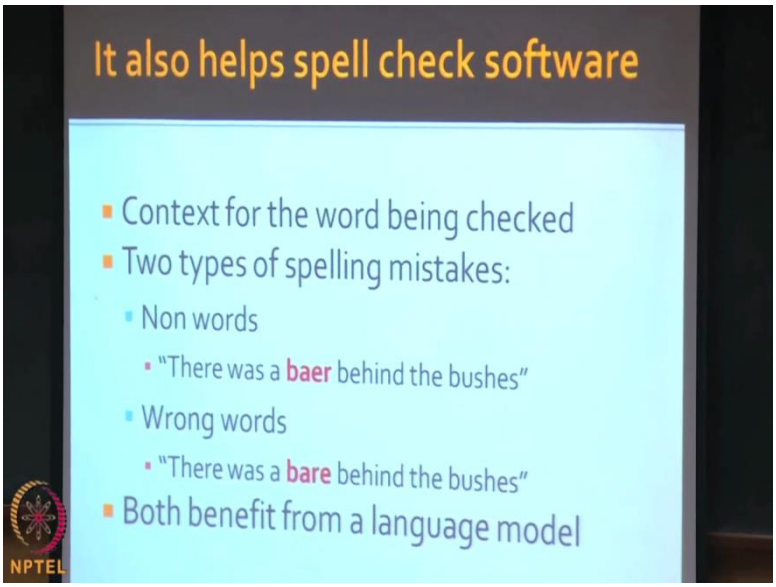
So now the job of automatic speech recognition software is to assign a word to what it has heard, sorry I did not use phonetic symbols here, but you guys know what phonetic symbols are, better than me. So, let us say it sounded like “bay-er” okay. Now is it B E A R or B A R E or was it slight mispronunciation and was it B E E R or B A R okay.

So the jobs of speech recognition software is, to first narrow down to these 4 choices and then pick the right choice okay and it does that based on these language model. So it can use for example, things like part of speech, knowledge if it has already passed previous sentence and it knows that it is an article, it knows that a noun is coming next right.

And based on that, it can be noun or an adjective, so based on that it can it can narrow some choices that may not help. Then it can it can just use previous words and n-gram model, okay. But to actually resolve this, this requires that the machine has seen many examples of all possible n-grams. It is different from how humans work; we generalize a lot better, right.

We do not need to go through the entire Shakespeare to make intelligent sentences, right. But the machine needs to go through entire Shakespeare, entire Wall Street Journal and so for to start to make some intelligent sentences. Now it also helps spell check software okay. So here a language model will give you context for the words that are coming next, that the words that is being checked. So are many types of mistakes are there in spellchecking?

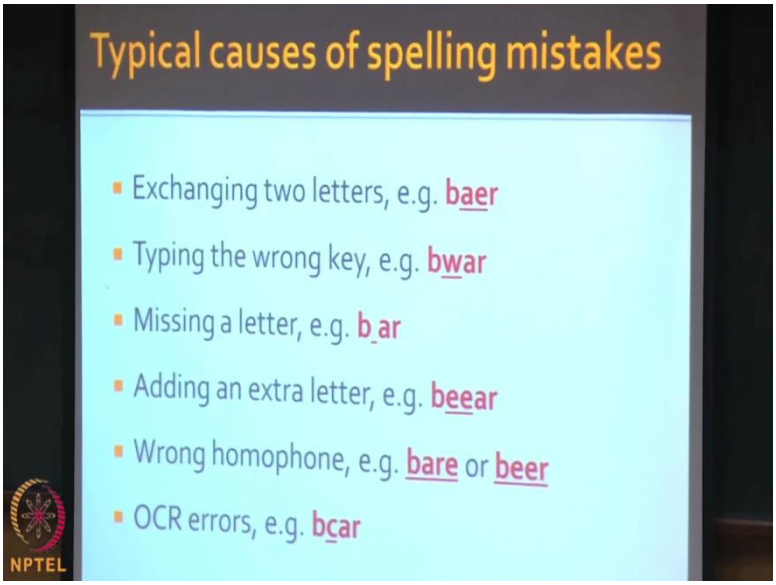
(Refer Slide Time: 18:11)



There might be different ways of categorizing mistakes, one way is that there are non-words and then there are wrong words. So what are non-words? For example, if someone types this sentence, “baer B A E R” B A E R is not a word, right? And wrong word is that this word is

actually there in dictionary, but it is not the right word okay. “There was a bare behind the bushes”.

(Refer Slide Time: 18:44)



So both of these will benefit from a language model, the n-gram model that I talked about or more sophisticated model to resolve both of these mistakes. So different types of or different causes of spelling mistakes everyone who has done typing know that they have made such mistakes with some frequency, right? So for example, exchanging 2 letters, if it is B E A R, A and E have been exchanged, okay.

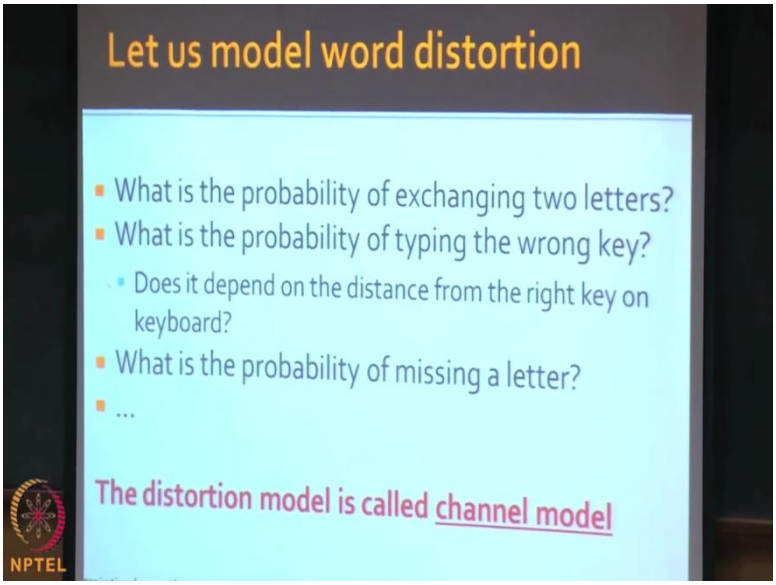
Then typing the wrong key, so why W? W is close to e on the keyboard. So you just, your finger does not land at the right at the right place, okay. Adding an extra letter, you press E too long, okay. And or you miss pressing E okay. And then wrong, homophone, words that sounds alike and in your hurry or because someone did not have the right knowledge, they type the wrong word.

So for example, for B E A R bear, B A R E is also a homophone and in some Indian pronunciation, B E A R is also called bear, so B E E R is also beer okay. And then OCR can also cause problem, optical character recognition, so if you have printed text and it is being scanned by a machine, the machine that recognizes the visual characteristics of a letter, sometimes make mistakes, E and ‘c’ look alike for example, B E A R can become b c a r.

Okay, so they are different types of mistakes, why do we need to know about these different kinds of mistakes? Is because that will help us tell if we have a wrong word, what would be the nearest right word. We have to have a sense for what mistake could have caused this.

Right. Let us look at what mistakes cause these, okay. So we need a model for word distortion what causes the mistake, okay.

(Refer Slide Time: 20:48)



What is the probability when you are typing to exchange 2 neighboring letters in a word, okay? So let us say it is once every 20 words you exchange 2 letters okay. So that is your probability is 1 over 20. What is the probability of typing the wrong key, okay? And does it depend upon the distance from the right key on the keyboard? Of course it does, right. So if you want to type E and what is next to E?

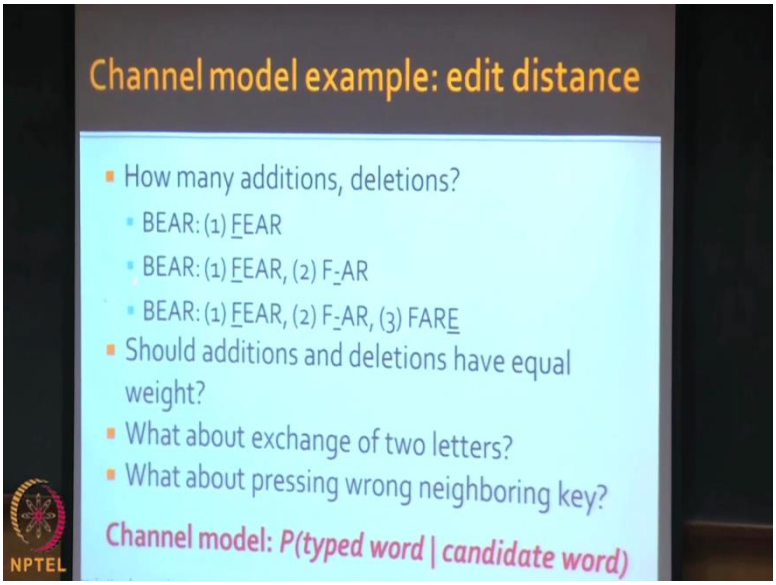
Let us say W is next to E, you are very likely to press W as opposed to X which is far away from E okay.' And then what is the probability of missing a letter? So what this means is that if you are writing B E A R and instead of E you press W which is highly likely compared to pressing X okay. So these 2 have different differences, the 2 wrong words have different distances from the actual right word.

So that notion of distance between words is important because that helps you find what is the probable mistake you made, the one that is closer to the wrong word, okay. The one that is far away is less likely to become a mistake, okay. So this distortion model that assigns some sort of a distance metric or a probability of wrong word given the right word is called the channel model.

And this channel word comes from communication, where you are sending a signal over a communication channel and the received message is slightly different from the sent message because of atmospheric noise or electromagnetic noise and so forth. So that the same term has

come here, where we are saying that it is not the electromagnetic noise, it is your finger noise right, on the keyboard that has caused a distortion.

(Refer Slide Time: 23:09)



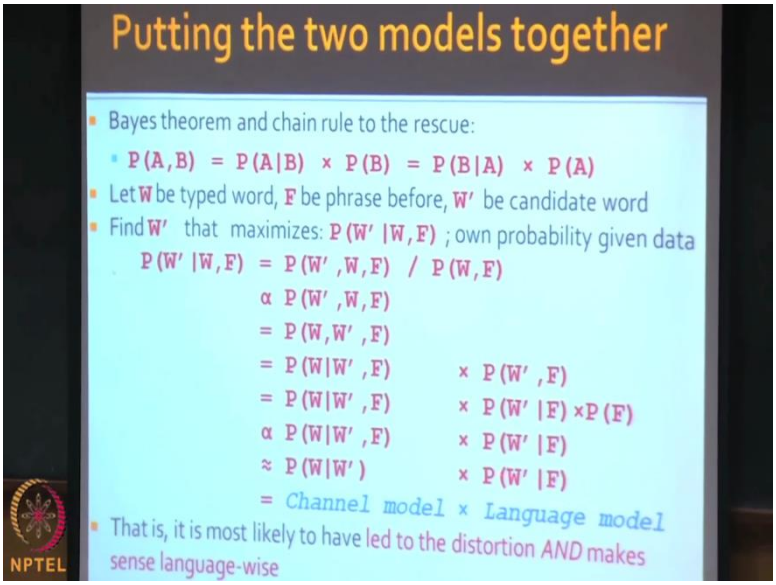
So here are some examples of starting with B E A R bear and having different edit differences to different words, okay. So let us say we define an edit distance as either one addition or 1 deletion of a letter, okay. One addition, one deletion and let us say replacing 1 letter with another for example. So let us in this case you have B E A R with just 1 replacement, you can go to F E A R, okay.

And if you want to go from B E A R to f a r, you need to first replace a letter and then delete a letter, okay so that is distance of 2 away. Then from B E A R to “fear” to “far” to “fare” is 3 distances, right. So if your actual word was bear, F A R E is less likely compared to F A R and even less likely compared to F E A R. So that is the notion of relative distance between 2 words.

Addition and now there are different questions when you model this that you can answer based on people’s behavior by observing many typing. That should additions and deletions have the same weight same distance? What about exchanging 2 letters B A R and B E A R right? What about pressing the neighboring key?

So when you answer all these questions, you will have a model of probability of a type word given what was the candidate right word, okay. So you can put all wrong words here and you can put the word that you actually saw independent and put all the possible right words and see which one is more likely, okay. So they give a distortion model, a channel model.

(Refer Slide Time: 24:36)



So we will put it together, I should not have put mathematics in here, but I will show you that it is not that difficult. What I want to show in the end that if we take both models into account, the channel model which is how words get distorted and the language model, which word combinations are more likely, then we will get a sense for how to find the right word for a spell check, okay.

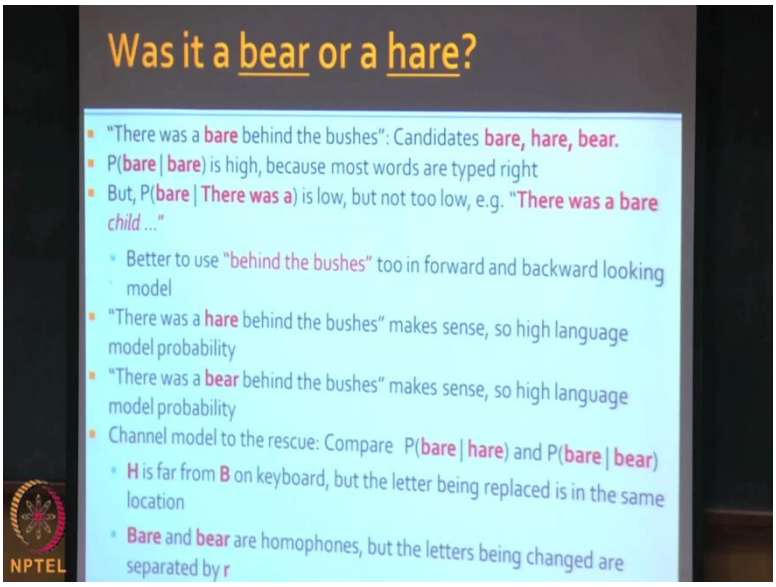
So it all depends on, it all starts with something called Bayes bayes Theorem. So Bayes Theorem, forget about that let me go through the through the definition here first. So let us say the word that was typed was W okay. F was the phrase that came before; let us say if it is if it is a trigram, then the 2 letters that came before.

And W dash is the candidate word, so W dash can be different possible corrections suggestions and you have to pick the right W dash, okay. So which one will you pick? You will pick the one whose probability is highest given the phrase and the wrong word that you already know, okay. So probability of W dash given the actual number that was typed in the word that was typed before that, okay these 2 together.

So given is shown by this line. So it turns out with some mathematics if you start with this, you can actually show that the word that maximizes this is the one that gives you the maximum product of these 2 terms. What are these 2 terms? What is the probability of the typed word possibly wrong word, given the candidate word times the probability of the candidate word given the phrase, okay?

So this is your channel model and this is your language model, so you have to find the balance between products of both of these, right. So it has to have the most likely word is the one that lead to distortion and makes sense language wise. So was it a bear or a hare, okay? Let us start looking at different possible um suggestions, okay.

(Refer Slide Time: 28:54)



So let us say the sentence was "there was a bare B A R E behind the bushes" and your candidates are B A R E which is the same word, so which means that you also take into account that there was no mistake made, specially because the wrong word, not the non-word, right. It was not B A E R, it was not the wrong word, but this is also an English word, so this is also a candidate. And then you also have hare and bear, okay. So which one is it?

Let us look at what factors you will consider, so we know that probability of bare given, what that means is that we did not make a mistake, probability that we did not make a mistake is high. Why, because 90% of the times in words we do not make a mistake. 90% of the time we type the right things that we want, so this is high, right.

But if you look at the previous words, which is "there was a" so "bare given there was a" is low, but not too low because it can also be used as an adjective "there was a bare child" right. But what else us that it is not bare, is rest of the sentence, right. So if you have more sophisticated language model, then it will also be forward looking, not just backward looking. I did not cover that the lecture, but you can always have the more forward-looking model.

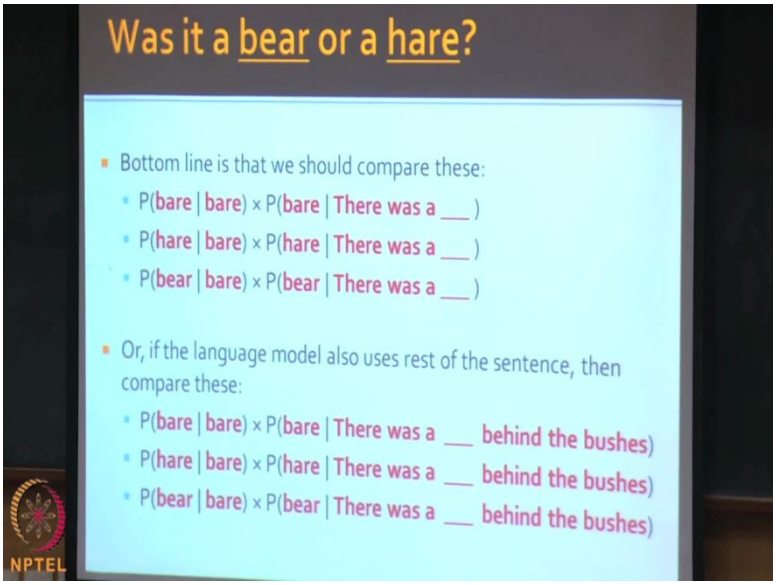
Now, "There was a hare behind the bushes" also makes sense, right. So it has high, what probability, language model probability, we have not commented on the channel model yet, it

has high language model probability. “There was a bare behind the bushes” also makes sense because it has a high language model probability, right. So now we have to look at the channel, okay.

So in channel what is the probability of bare given hare verses bare given bear? And we compare these 2, and not just these 2, but we have product with their language models also, right. So we know that H and B are far on the keyboard, so hare given bare has a small channel probability, right. But we know that bare and bear homophones, okay. But their letters that are changed are separated by R.

So they are home homophones, so they might they might have high channel probability, but the reason to make a mistake because typing error is small. It could be an error because of the Homophone itself. So we do not know that, so the way to resolve this I I do not have an answer for this, but I know how we will get this is that we should compare these 3 things.

(Refer Slide Time: 30:21)

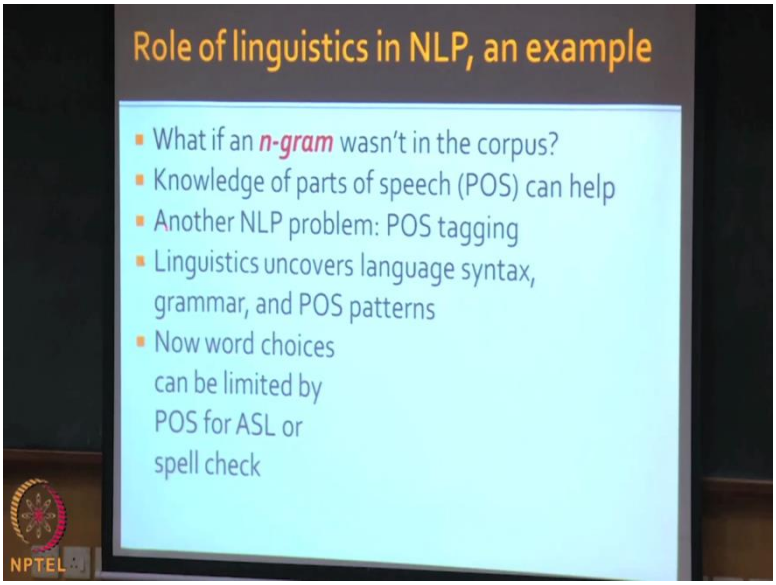


We should compare the product of bare given bare times bare given “there was a”. We should compare hare given bare times hare given “there was a” okay. bear given bare times bear given “there was a”. Or if you have more sophisticated model, then you will use the rest of the sentence also, okay. So whichever of these 3 computes to be the highest will be the highest suggestion on your spell checker, okay?

And the one that gives you the next highest will be the next suggestion on the spellchecker and so forth, okay. So this is in a nutshell how spellcheckers work. So what we have not used in this so far, we have not used any linguistic knowledge, right? We have trained these

language models; we have talked about how to train these n-gram language models just on corpora, right on large corpus.

(Refer Slide Time: 31:37)



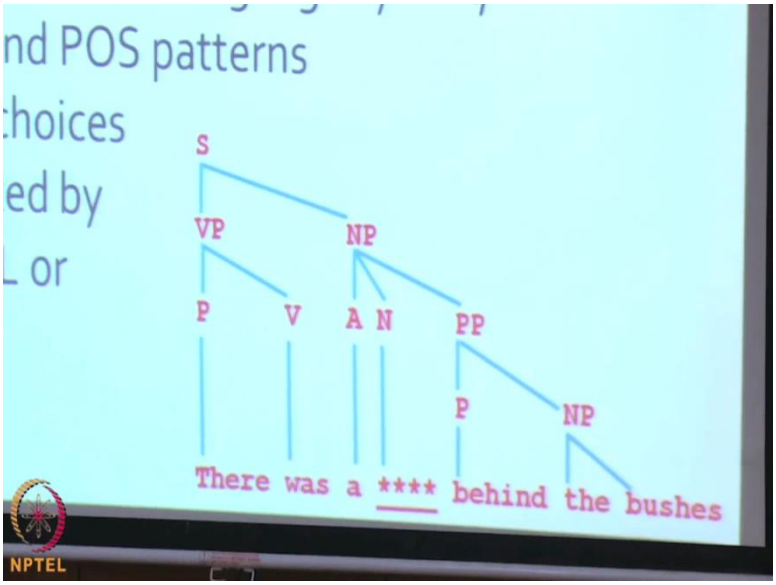
So we can also bring linguist into it, okay. So what if there was an n-gram, combination of n word that was not in the corpus, then what will you do? Because then your language model probability for that n-gram will be 0, right. So there are mathematical techniques to say that every n-gram even if it is unseen should have some nonzero probability, just give it some small number and use it in your spellchecker that is one way.

The other way is that you actually use knowledge of linguistics. So for example, you can use knowledge of parts of speech tagging, right part of speech can help. So that is another problem in natural language processing, so you are given a sentence, how do you tag a part of speech for each word that is there in the sentence?

So I will not go into that and you use linguistic knowledge to uncover Syntax, grammar and part of speech pattern, right. So you know that there will be a noun phase, there will be a verb phase and all that. So you use that to narrow down the choices of the word that you are trying to correct, okay. And that word and its combination of the n-gram may not even be in the corpus.

But you know that there has to be some noun here a verb here or an adjective here, right. So based on that you narrow it down and then from those narrowed down the choices, you find the one that has the smallest distortion between the intended word the typed word.

(Refer Slide Time: 32:50)



So for example, if you had a missing word and you make Syntax, please pardon me if my Syntax is wrong. what I know now is that this is noun and because of noun I know that B A R E will not be a part of this, okay. So that is how I can use linguistic knowledge also to correct the spellchecker here, okay so that is all, thank you very much. If you have questions, I can take questions.

(Presentation of Dr Amit Sethi ends here at 33:06 and question and answer session starts)

Participant: The first question is related to the difference between handwriting and keyboard, the second question is related to the impact of technology on the mind particularly, computers. Are we losing our intelligence or memory because we depend more on computers?

I see more these problems in young children or younger generation than in the older generation particularly for example, T H I N K and T H I N G, some of the students do not make the difference. Similarly W H E R E and W E RE, they do not make the difference. Have the children become computers like?

Dr: To answer your first question how is how will we treat handwritten language differently from computer language, the main difference will be in the channel model, right? The channel model will be very different for handwriting compared to compare to typed txt because the kinds of mistakes that you make in both are characteristic to how you are writing that.

The second question, whether the use of computers is making us less intelligent, I do not know if I have an answer to that perhaps, it is shifting our intelligence from one domain to another. I think that is probably what is happening more whereas, more types of knowledge to be kept in mind these days because there are so many things that you can actually indulge in terms of even language, interacting with language for example.

So, I think the probably the focus of the mind is shifting. I do not know if it is making us with intelligent or more intelligent, but it is definitely shifting. And this SMS type language, it is definitely a pain point personally for me also when I see that in emails from students and I just advise them that do not do this in a professional setting, people will judge you. But that is the just a friendly advice for them.

Participant: I think also in the case of handwriting, your sound system becomes very effective, right so when you are not looking at scripts right, and you are using it, but initiative is actually a sound, alright or a cluster of sound. So in that case if you have W E R E and W H E R E coming together or replacing each other incorrectly, you still are able to in put the responsibility on the sound system.

So it is totally not in era in that sense. Whereas, in the case of a machine that information unless you have really put in, the sound there along with the rest of the information right? About “where” and “were”, if you put that and I think that will be probably a big parameter spoken language versus that.

Anyway, my question is I and I am trying to understand your system because it seems to be more like working on a linear array, right. And which is why I was very concerned when you kept on saying previous word or previous words and then you said well you can also look forward, so it is not just backtrack, it is looking forward.

So I I am I am still trying to understand where good dependency come in from because what I know of computational and linguistics from one of my colleagues in my department in Delhi, I think dependency grammar is very useful right? And dependency grammar is all about the verb very centered on the work because rest of the things seem to come from there, of course you have dependencies within noun phase as well.

And I was thinking, could you explain if dependency grammar could be useful for this kind?

Dr: So there are long range dependencies in language, right. the word that you are trying to spell check would be in form of by something that was even in the previous sentence, not just this sentence for the next sentence or even like beginning of this sentence of a long sentence. So the models that I have explained are very rudimentary, this is what I would give in a class assignment.

The state of the art spellcheckers use much more long-range model. And the other thing is that all these spellcheckers, they are coming from commercial ventures, right. So they want to get a product out and then the product is it does 90% of the job with just like trigram or a Quadrigram model, they are happy to put that out.

And that helps in a lot of situations. Your experience with MS word will tell you that it is not always, it does not always give you the perfect answer that is because it does not have the same kind of model that we have probably in our brain of long range dependencies, better knowledge of language, internalizing language in some sense as opposed to like just thinking of it as a statistical problem of given 5 words and what is the next word and so forth.

Participant: See, dependency model or these Syntagmatic relations are captured in this model. When you look at languages which allow a lot of scrambling where this linear rigidity is not there, so what will happen there?

Dr: You probably need a different language model of course.

Participant: Languages which are very rich in morphological Morpheus synthesizer, morphologically very rich. They allow scrambling, they allow you know, there is no linear rigidity for example, Sanskrit or let us say French, they are morphologically very rich, syntax Syntactic structures are not rigid. Focalization is allowed, scrambling is allowed ordering of the word is not a rigid parameter.

Dr: Yeah, even in automatic speech recognition for example, Indonesian right, in Chinese.

Participant: For example, Chino-Tibetan languages, they are full of (())(38:58) markers are not there simply there is very messy.

Dr: No, I was saying that even in automatic speech recognition, it is a problem; you cannot model all languages easily. There are some languages that are toned age right, so it is very like you need tone recognition and so forth to resolve the same items in different contexts.

Participant: You were talking about corpora, when the computer relies on corpora, you have one large corpus or is it based on the topic like you said Shakespeare and Wall Street Journal are different. So would there be a separate field for Shakespeare and a separate field for…

Dr: We will probably have a separate model for Shakespeare and separate language model for Wall Street Journal, yes.

Participant: Okay, yeah thanks.

Participant: I just want to know how this Google translator and all, those kind kinds of things work.

Dr: I do not know their technology exactly because it is a proprietary secret they have, but some of the things that they use are basically these language model itself like these n-gram models probably that would they use. And they will probably have these models for 2 different languages and then they will also learn interconnect between the 2. Based on lots of examples of Spanish text translated to English by professional translators, they will learn this.

So a lot of work Google does is, large scale statistical machines learning, which means they have access to a lot of translated text, a lot of OCR and so forth, so and the way they get it many times is at some point when you are logging into Gmail, they will ask you a question to solve an OCR, what they are using is they are giving it to 3 different people and if 2 people agree, then they think that it is the right answer and that goes to train their machine.

Okay, so you are helping them train their machine by using Gmail and all these other services that they give you, so they give you good services to get data from you.

Participant: Sir, is there any corpus or anyway channel model to analyze direct speech and not the whole text? Like the corpus from that like you have a corpus or corpora and from that I can analyze…

Dr: From like speech?

Participant: From just direct yeah, from the whole text I can…

Dr: From like a speech signal, like in audio signal?

Participant: No, no, no, from written text. Is there a possibility or is there a model where I can analyze only…

Dr: I I am not aware of it, but we can search I mean there must be like a repository of like all speeches of some sound, that is politicians speeches or something, then you can computer sound like a politician.

Participant: Like from written text, suppose I have a short story or something, what I want to analyze is just the direct speech of the speakers that is it.

Dr: Okay, okay. There are techniques through that also and there might be corpora that are available. A lot of public Shakespeare plays would fall into that.

Participant: Yeah, that is what I want to like, if there is something of that time that would be.

Dr: Yeah, I am not particularly sure which one, but I I think there must be some.

Participant: For an ATK, there is corpus dedicated to what you called the text that is used in internal conversations and forums, so maybe you can use that for direct speech that will be helpful. You were talking about training and all that.

So I was thinking, when we are doing spell check very often Google keyword will take what you are typing to use to train itself for to like specialized to you, what do you think about that data being sent to Google and being kept by larger companies just, so that is the issue that and will be does not addressed most of the time.

Dr: So NLP would not address ethical and business issues like that right, it is solving an engineering problem.

Participant: Yeah, I am just asking it should NLP should address an issue like that, where the entire data privacy will be an issue.

Dr: You can have NLP solution that respect your privacy for example, so you can have let us say a software that learns it only on your phone and keeps the data only on your phone, that is possible. But, whether someone will give you that software that depends upon your demand and number of people demanding that.

Participant: Sir, I have 2 questions sir. Number 1, the authentic the authenticity of these kinds of stuffs. On what base, especially the purpose what you said is it reliable or something like that? I need authenticity. Number 2, if we go by this kind of stuffs to the students, I think they may use the thinking ability, they will not think on their own, so what is your suggestion?

Indeed they will get into that Google or whatever it is, they will not think, there is no time for them to think to check the exact meaning something like that and even the guess.

Dr: I sorry, from authenticity means has the computer being trained on authentic corpus, is that the question? So, these some of these corpuses, these corpora are pretty standard, people know that you know you can actually open Shakespeare and see that the purpose also has the same lines. So like Thomson Reuters and these Wall Street Journal and this Shakespeare corpus they are they are pretty authentic.

But in terms of losing thinking ability that is true. But we are also losing our thinking ability while doing calculations as we still use the calculator that that is true and like going back to that question about handwriting versus typing, you know when you when you are doing and writing specially cursive.

The processes in the brain when you are doing actually joining letters is very different from when you are discreetly typing keys, right so are using one mental faculty and actually going into another territory altogether. So the question whether one is more intelligent than the other, I do not have an answer to that, right. That all depends on where the humanity is going whether the evolution is taking us.

Participant: Join you in answering my good friend.

Dr: Please free, yes definitely.

Participant: It is in the nature of human beings to be anxious; anxiety is part of our biological make up. And anything that is new including shoes and bride, sort of create threat. So for example, when books came, letter was critical, when Fountain pens came in place of quilt pens or in my memory when jotter refills came, banks would not accept checks written with jotter refills, examinations could not be written in the 1960s with jotter refills.

I am sure Professor Patnaik will bare me interestingly as the only adult, a not so young person. So it is in the nature of human beings to be anxious about new technology, but if you look at the history of writing in all probability, Panini was illiterate. The history of writing does not go beyond 3 millennium and in India; it does not go beyond 2.

Asoka barely got his edits engraved on some rocks at great expense and which only Bill Gates today or Asoka could afford it then. So Panini in my opinion was not likely to have written (())(46:27). He might have composed and the students might have got them right. Dhanwel's

anxiety is totally right, when you relegate the work to some other tool, to that extend your body becomes weak.

But I think you correctly answered in, you know the gain is somewhere else, you do something else. When a driver drives the car, you sit down and enjoy music or you think of newer ways of beating your wife or anything. So I think you know, recently I attended a meeting of educational administrators who were debating whether classroom teachers ought to have a laptop and 9out of 10 of them failed like Dhanwel has done.

You know that, it will weaken the students and this is the example I gave them that when Jotter refills came, actually if I see it right, I do not want to get into technology forecasting, but you know writing and the entire technology that came with it, the hardware and software. Software will be letters, hardware would be tablets or pen or paper.

And another 50 years at the most, when we will have very commonly voice driven, speech driven word processing packages, you would not really need to write because the words could be recorded. Now there was a time when somebody in Bangalore offered a bribe to someone and court said, "No, recorded speech will not be taken as evidence".

Now it is routinely taking, a very celebrated journalist is facing trial in court because she offered to help some industrialists and her words were recorded. Technology and civilization have had a running war where eventually civilization except new technology. So we have assured Dhanwel, there is nothing much to worry about.

Dr: Thank you so much.