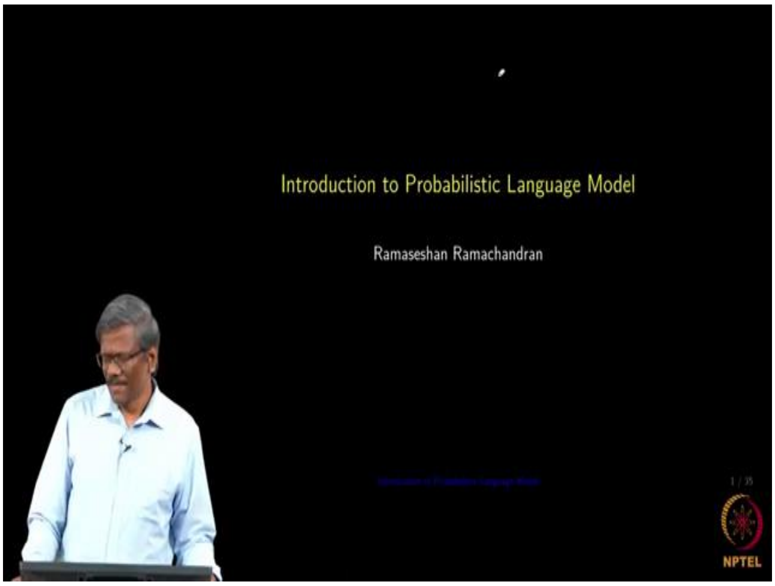
Applied Natural Language Processing

Prof. Ramaseshan Ramachandran
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Lecture – 20
Examples for word prediction

(Refer Slide Time: 00:15)



Introduction to Probabilistic Language Model

Ramaseshan Ramachandran

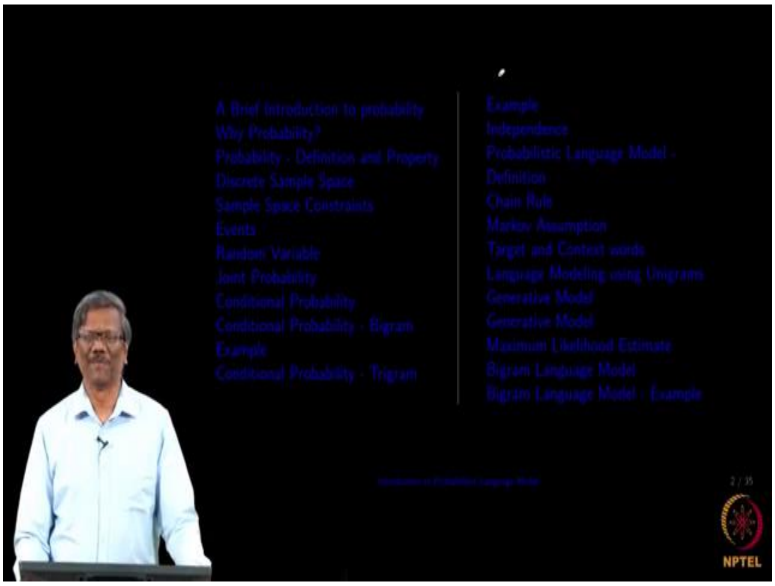
1 / 35

Hello everyone. Welcome to this class once again. Today, we are going to be talking about the Probabilistic Language Model. Before that let me give a recap of what we have done and where we are going with respect to learning and understanding words, sentences, and so on ok. Earlier we had seen that context is very important in terms of understanding the meanings of the word in a given document or the next step in terms of an understanding document is understanding a sentence correct.

So, we need to understand how these words are strung together to form a sentence. if you understand how you learned the language when you are very young nobody told you the syntax of the language, how the verbs and nouns play a part in constructing a sentence and so on. you have been listening to your peers, your teachers, your parents and then try to pick up new words and then try to repeat what you have heard earlier, correct?

So, in some way we are able to capture the patterns in our brain and then make use of that whenever there is a need to use that particular word or sentence. How do we repeat or replicate this process while making the machine understand the word and then trying to construct a sentence out of the words it has understood ok.

(Refer Slide Time: 02:12)



A Brief Introduction to probability
Why Probability?
Probability - Definition and Property
Discrete Sample Space
Sample Space Constraints
Events
Random Variable
Joint Probability
Conditional Probability
Conditional Probability - Bigram
Example
Conditional Probability - Trigram

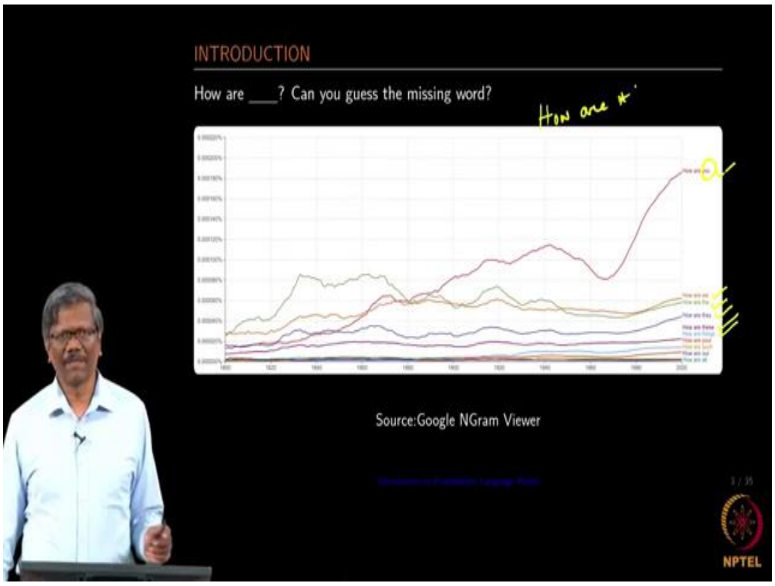
Example
Independence
Probabilistic Language Model - Definition
Chain Rule
Markov Assumption
Target and Context words
Language Modeling using Unigrams
Generative Model
Generative Model
Maximum Likelihood Estimate
Bigram Language Model
Bigram Language Model - Example

2 / 35

So, before moving on to that let me first ask you a few questions. Even before asking you those questions let me tell you what is going to be in this lecture. that you get a very high view higher-level view of what you are going to be learning in the next few minutes ok. I will be giving a very brief introduction to probability. It is very important for us to understand this. Many of you would have studied probability I am going to be just connecting the NLP with the standard terminologies of probability and then see how we can make use of that in our studies here.

And, then I will talk about examples of the sample space evens and what is a random variable with respect to the documents and words. We will talk about the conditional probability, we will talk about some examples using bigrams and trigrams and then we will also talk about the independents and how two words are independent or dependent depending on where they occur. And then later we bring in the language model introduction and I will define what is a chain rule on how to compute the probability of a sentence, then I will bring in the mark of assumption in terms of reducing the complexity of the computation so on ok.

(Refer Slide Time: 03:50)



INTRODUCTION

How are ___? Can you guess the missing word?

Source:Google NGram Viewer

3 / 35

So, let me get to the next slide where I am going to be asking you a question ok. this is a very general question most of you would know the answer to this. Find out what could be the last word in this how are ok. what could be the last word? I will give you about 10 seconds to think about the word that could be formed at the end of this sentence. If you have not got enough time you can pass this video and then think about it and then go to the next section of the slide ok.

So, how many of you have thought about you? Have you got any other words at the end of this sentence? let me give you an example of what Google Ngram Viewer does. we will talk about this little later, but I am going to be giving you only the output of what the Ngram Viewer does. if you look at the Ngram Viewer which is based on reading all the books that are available in the Google books and then the end of the word end of the sentence is picked up based on how often they occur in those books ok.

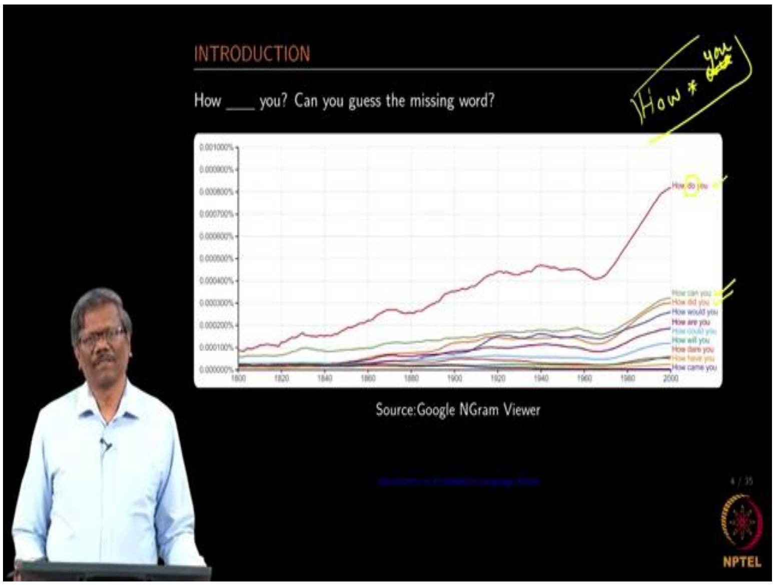
you are right I think most of you would have figured out the first one right, how are you right. How many of you have thought about how are we? And then there are also places where there is the letter the word the appears after how are and they how are these, how are things, how are you were, how are such how are our how are you and so on. it gives only the top 10.

So, what actually this Ngram viewer does is it reads all the books and then tries to figure out what could be the possible word after how are ok. See the first two words are given.

it is going and then finding out you remember the trigram; in a trigram sequence it tries to find out the first two words are how are, the third one what would be the third one. you can also do this in the Ngram Viewer by just using this option how are and then start when you do this you get a very similar graph like this.

I am sure you will get the same as this may not be a similar one because the corpus is fixed. you should be getting almost the same graph that you are seeing right now ok. let me take you to the next question ok.

(Refer Slide Time: 06:53)



INTRODUCTION

How ____ you? Can you guess the missing word?

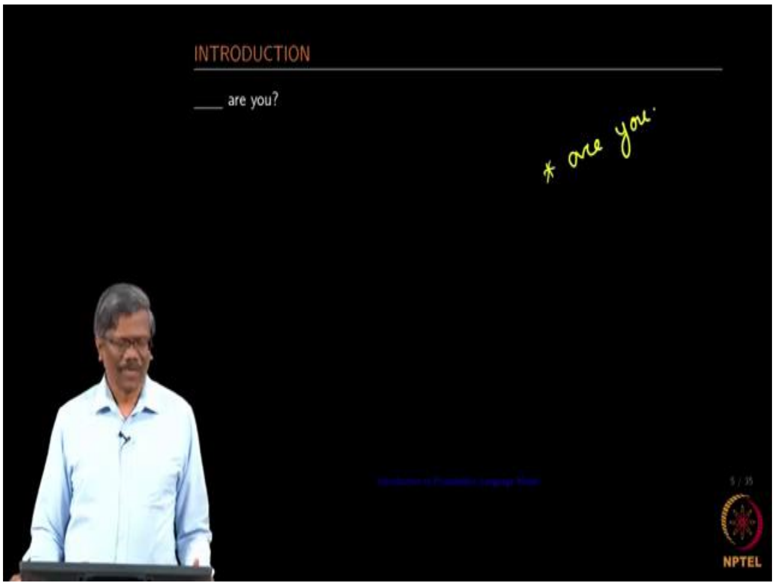
Source:Google NGram Viewer

So, now what I have done is I have taken the middle word off and now you have to find the middle word how to dash you. I will give you again about a few seconds to think about the word that could be the part of the middle ok. How many of you have come up with more than 5? If you need more time you can pass the video and then take your time and figure out how many could be formed in the middle of the sentence ok.

So, let me show you what Google Ngram Viewer had found ok. In the same fashion now we have a trigram we have the first word and the last word the middle word is anything right. you can just have you can use this phrase and then search in the Google Ngram Viewer. you will get something like how do you. you see there is a do. it is not are as the first one correct.

So, here in the Google books probably how do you phrase had occurred more times than how are you. This should be you, I am sorry and then the second one is how can you; the third one is how did you see the variations right. since you are given the first slide you know I told you that it is how are you it is possible that you would not be able to think more than 2 or 3, maybe some smart one would have thought about more than 10 right. if you take the contexts differently now you are filling the middle with different words ok. Let us go on to the next one.

(Refer Slide Time: 09:09)



INTRODUCTION

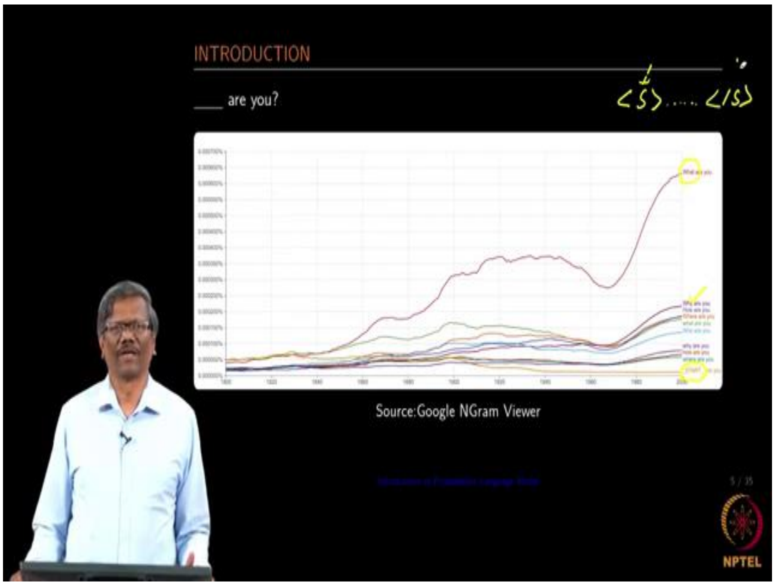
____ are you?

* are you.

5 / 35

So, I have taken the first word out and then asking you to find out what is the first word right. you can search in this fashion in the Google Ngram and let us see what we are getting.

(Refer Slide Time: 09:30)



INTRODUCTION

____ are you?

Source:Google NGram Viewer

See the variations right. we are blinded by the fact that earlier we had seen how in the first place. we would probably think all the time now and then maybe you know we will be thinking about the other ones it will be maybe a little difficult for us to find the rest of the words that are going to be appearing as the first one.

So, in this case if you look at it we have what why how where what and then the last one let me take the last one at the start. are you are also the sentence that you can form. what he Google had done is, it is using a start symbol as the first symbol ok. for all the sentences remember there will always be a start symbol and there always be an end symbol. if you want to really process it you have to look for the start symbol first and then the end you are processing when you see the end of this sentence.

So, normally we use this as our starting symbol and our ending symbol ok. what did you gain from this? what we have seen here is using a different context you will be able to find different words right. the words are coming into the play when the context or different words are coming to the play when the context changes. The context is so small here even if it is so small the world that is fitting in into the middle of the first or in the last is changing drastically. how do they really come into play?

as I mentioned earlier are based on the trigram in this case since there are three words the Google Ngram Viewer would have scanned through the entire Google books and then figured out for the first example are could be the first one because the are would have happened more number of times when you are giving how and you as the context. And then if you look at the second one if you look at the first one you as the last word correct. that happened so many times it has counted that particular trigrams had occurred so many times that comes as the first in the first graph.

The second one the middle one is missing right. it goes and then finds out what could be the middle word and then it counted all those words all those trigrams and then figured out that the number of counts for do is more when it is combined with how and you is more. it brings how do you ask the first one. there is some kind of counting involved in this and it also tries to find out what could be the first one to be placed. And then you can see that in all the cases the first one is very highly placed right than the rest, correct ok.

So, can we how do we simulate this? how is it possible? we are going to be getting into the details of this in this lecture from now onwards.