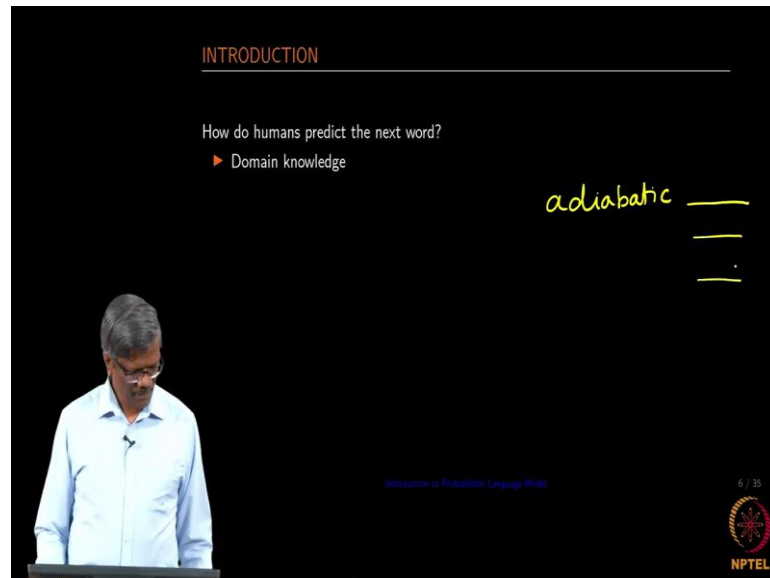


Applied Natural Language Processing
Prof. Ramaseshan Ramachandran
Department of Computer Science and Engineering
Visiting Professor at Chennai Mathematical Institute

Lecture - 21
Introduction to Probability in the context of NLP

(Refer Slide Time: 00:15)



So, how do we predict? we predict based on what we know right. If we did not know about that particular sentence, we would not have predicted anything. we need to know about that ok. we need to know about the domain very well. For example, if I give you the word adiabatic and then ask you to find what is the next word ok; if you did not understand thermodynamics or if you did not study thermodynamics it is not possible for you to fill in those words like adiabatic, expansion, compression and so on. you require domain knowledge to be able to fill in the missing word.

(Refer Slide Time: 01:12)

The slide is titled "INTRODUCTION" and asks "How do humans predict the next word?". It lists five factors: Domain knowledge, Syntactic knowledge, Lexical knowledge, Knowledge about the sentence structure, and Some words are hard to find. Why? To the right, the phrase "to come" is written in yellow, with "to" underlined and "come" written above it. The slide also features a small image of a man in a light blue shirt at the bottom left, the text "Introduction to Probabilistic Language Models" at the bottom center, and the NPTEL logo at the bottom right.

I you also require this syntactic knowledge about the language that you are studying. you need to know how the words are strung together and what is the legal way to form the sentence, it is something that you have to know. you need the domain knowledge, you require the syntax, you require the lexical knowledge for example if I write this word ok. you will be able to very quickly figure out that there is a word boundary here.

So, we understand based on the language that we speak, read, or write that there are two words here. there has to be a space right. we understand the lexical structure of the language that is we know each and every token that is written even if it is combined in as one word. And, then I spoke about this knowledge about the sentence structure; how do you really construct the sentence in the English language using verb phrase and noun phrase. And, then there are certain words which are very hard to find for the example that I gave in terms of the adiabatic expansion right.

So, it is very hard to find even though the word is known to you unless you know the domain is going to be difficult for you to find out.

(Refer Slide Time: 02:43)

The slide is titled "INTRODUCTION" in orange text at the top. Below the title, the text "How do humans predict the next word?" is followed by a list of bullet points, each preceded by an orange arrowhead. The bullet points are: "Domain knowledge", "Syntactic knowledge", "Lexical knowledge", "Knowledge about the sentence structure", "Some words are hard to find. Why?", "Natural language is not deterministic in general", "Some sentences are familiar or had been heard/seen/used several times", and "They are more likely to happen than others, hence we could guess". In the bottom right corner, there is a small red circular logo with a white star and the text "6 / 35" above it, and the "NPTEL" logo below it. A man in a light blue shirt is visible in the foreground on the left side of the slide.

It is not very deterministic in nature natural language is human created and it allows you to innovatively create sentences and form new ways of forming a sentence and on. it is not like any other programming language that you are used to; correct. some of the sentences you know are very familiar to us. we can have heard this, seen this and used them several times. it is easy for us to very quickly construct a sentence even if something is missing.

For example if you probably would have used this several times in your mobile phone, if you are sending of birthday wish to your friend right and assuming that you are typing all the words wish you many more happy returns of the day. And, I am sure you would have noticed in all your mobile phones when you start typing wish you many and then at the bottom you will see some options for you to pick words, you know we do not have to really write them; right.

So, our mobile operating systems now have the capability to really fill in what is going to be the next word. in the same way, we have in our brain that we had stored a lot of patterns of that and we know that when some words are started, we know that immediately what could be the next set of words that would come in. we want to assimilate our copy of this model. we are able to build sentences using the machine. some sentences you know if you look at it are more likely to happen than others.

So, when since we could very quickly guess what would be it for example, if I say something and then stop in the middle you will be able to fill in what I am really saying. For example, natural language processing is the study of; you can fill in some of the missing words after that right. Since, we know what the domain is and you have come to some of the classes. you would be able to know what could be the following words in that particular sentence.

I was very particular in terms of saying that we had seen this several times, you know I have not used any of the terms that we normally use in the probability you know; I just want to give you the normal word combination that we use you know without knowing that particular subject. Now, having said that probability would be very useful in the case where we would like to estimate what could be the next word probability really uses some of the observation that we had seen and then some additional knowledge is coming from the observation and then there are certain things which are missing.

So, using the observed values we would be able to estimate certain values or in this case certain words right. that is why if you look at the first three slides you had seen several words appearing you know instead of just one. it tells you that some of the words occur more frequently than others or it says that some of the words occur more frequently or the word is more likely to happen after these two words. these are the terminologies that we would continue to use from now onwards.

(Refer Slide Time: 06:48)

WHY PROBABILITY?

- ▶ Provides methods to predict or make decisions to pick the next word in the sequence based on sampled data
- ▶ Make the informed decision when there a certain degree of uncertainty and some observed data

How* you?

How* you?

7 / 35

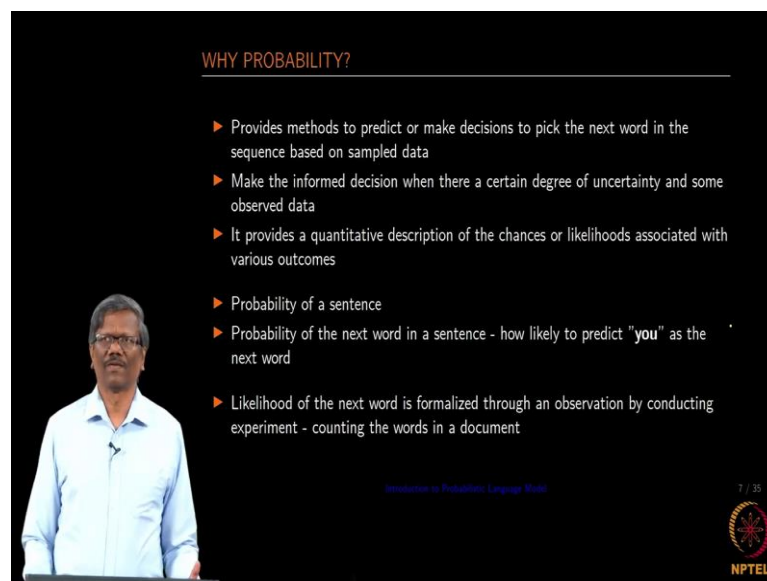
NPTEL

So, I am not going to be getting into the details of the probability theory and. I am going to be only covering what is essential for the language model ok. as I mentioned earlier it gives you some mechanism to make a decision. in the earlier case of the first three slides if you look at that the how are and then the last word you had happened so many times.

So, I can say with the probability very high probability saying that how are you are more likely to be formed than any other sentence that had happened at the bottom of the graph ok. that decision was taken based on the sampled data ok. we can make some informed a decision when there is a certain degree of uncertainty and some observed data are available to correct. if I use those again. as we had seen there of so many words that form part of this correct.

So, what we have done is we take a corpus very large corpus and then try to find out the trigrams where how occurs as the first word and then we can form a set wherein these words occur. this is one of our set which will contain how as the first one you as the last one and then the rest are available for us with respect to some kind of counting. we start counting how many times those middle words occurred and then based on the observed data we make some informed decision saying that how do you could be the probable sentence that we are looking at ok.

(Refer Slide Time: 09:17)



WHY PROBABILITY?

- ▶ Provides methods to predict or make decisions to pick the next word in the sequence based on sampled data
- ▶ Make the informed decision when there a certain degree of uncertainty and some observed data
- ▶ It provides a quantitative description of the chances or likelihoods associated with various outcomes
- ▶ Probability of a sentence
- ▶ Probability of the next word in a sentence - how likely to predict "you" as the next word
- ▶ Likelihood of the next word is formalized through an observation by conducting experiment - counting the words in a document

7 / 35

NPTEL

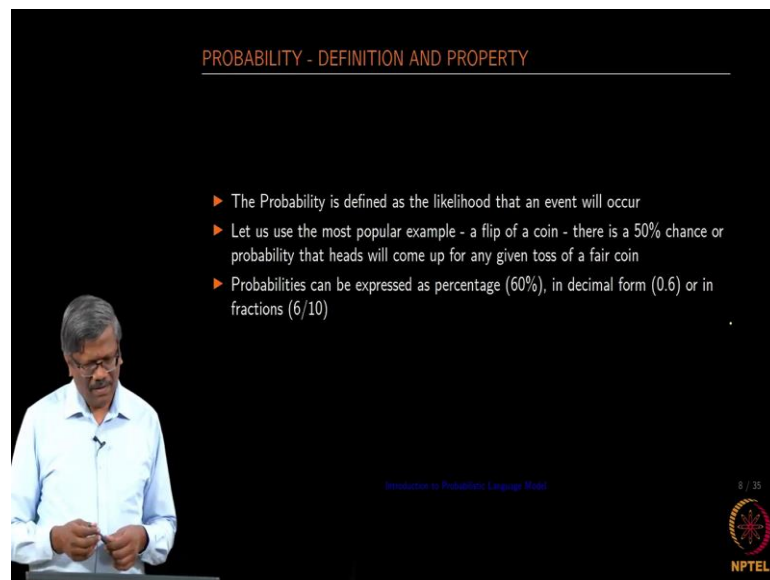
So, it gives you the quantitative description of the changes or likelihood associated with various outcomes; you got it? we can also compute the probability of a sentence you

remember earlier we had been talking only about the words and the synonyms and the context surrounding those middle word right.

So, now having acquired some more tools we should be able to form a sentence using the machine, and then we should be able to find out what could be the probability of forming this particular sentence is it high low and so on ok. again it requires a large corpus we require to do a lot of word counting and so on, probability of finding in the next word in a sentence how likely to predict you as the next word. these are all the important factors used in which we would probably utilize the likelihood of the next word is formalized through observation by conducting an experiment counting the words in a document.

So, we start with the tokenization model ok. we are given a document or a corpus we start actually tokenizing and then start counting how many times a particular word occurred what are the different types that we find or the vocabulary that we have within the document how many words are there in the corpus or a document, how many times a given type has occurred in the given document and so on. we will make those counts in the following slides ok.

(Refer Slide Time: 11:15)



The slide is titled "PROBABILITY - DEFINITION AND PROPERTY" in orange text at the top. Below the title, there are three bullet points in white text:

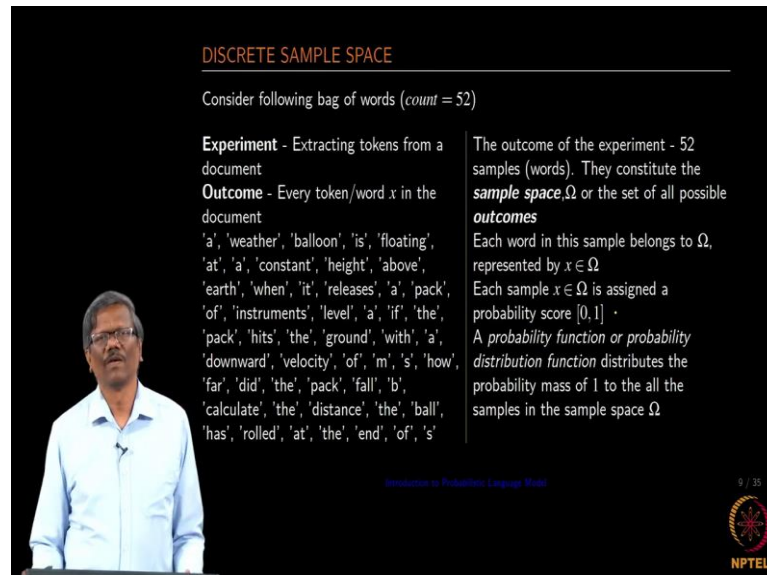
- ▶ The Probability is defined as the likelihood that an event will occur
- ▶ Let us use the most popular example - a flip of a coin - there is a 50% chance or probability that heads will come up for any given toss of a fair coin
- ▶ Probabilities can be expressed as percentage (60%), in decimal form (0.6) or in fractions (6/10)

In the bottom left corner, there is a small video inset showing a man in a light blue shirt. In the bottom right corner, there is a small logo for NPTEL and the text "8 / 35".

So, let us first define the probability. the probability is defined as the likelihood that an event will occur let us take the most popular example the flip of the coin right. there is a 50 percent chance or when you toss a coin; there is a 50 percent chance that it would be ahead or a tail correct and then the probabilities are expressed in terms of the percentage

or in the decimal form or in the fraction form. and we will be using the decimal form in most cases ok.

(Refer Slide Time: 12:02)



DISCRETE SAMPLE SPACE

Consider following bag of words (*count* = 52)

Experiment - Extracting tokens from a document

Outcome - Every token/word x in the document

'a', 'weather', 'balloon', 'is', 'floating',
'at', 'a', 'constant', 'height', 'above',
'earth', 'when', 'it', 'releases', 'a', 'pack',
'of', 'instruments', 'level', 'a', 'if', 'the',
'pack', 'hits', 'the', 'ground', 'with', 'a',
'downward', 'velocity', 'of', 'm', 's', 'how',
'far', 'did', 'the', 'pack', 'fall', 'b',
'calculate', 'the', 'distance', 'the', 'ball',
'has', 'rolled', 'at', 'the', 'end', 'of', 's'

The outcome of the experiment - 52 samples (words). They constitute the **sample space**, Ω or the set of all possible **outcomes**

Each word in this sample belongs to Ω , represented by $x \in \Omega$

Each sample $x \in \Omega$ is assigned a probability score $[0, 1]$

A **probability function** or **probability distribution function** distributes the probability mass of 1 to the all the samples in the sample space Ω

9 / 35

NPTEL

So, let us define the important terms of the probability with respect to what we are dealing with our document, sentence, or corpus; ok. Let us first define what is an experiment; the experiment is, in this case, is extracting the tokens from a given document; let us assume that you have been given a document and then you are asked to tokenize that. that is an experiment. The tokenization experiment gives you a token at the end of it as its output. you tokenize the entire document and it forms a collection right.

So, what is the outcome of the experiment is a token in this case. now we have taken a document and then tokenized this. if you count the number of terms or the outcome of the tokenization gives you a word right and there are 52 words in this experiment or in this document; the outcome of this experiment is 52 words and we can say that this particular set of 52 words form the sample space Ω and each outcome belongs to the sample space and then each sample is assigned a probability score and the score is in between 0 and 1 ok.

So, if you define it the probability function or the probability distribution function distributes the probability mass of one to all the samples in the sample space. In this case the value 1 is distributed across the token there are 52 terms. of the probability of the word whether it is going to be 1 by 52 ok.

(Refer Slide Time: 14:10)

SAMPLE SPACE - CONSTRAINTS

All the words in the Ω , must satisfy the following constraints:

1. $P(x) \in [0, 1], \forall x \in \Omega$ and
2. $\sum_{x \in \Omega} P(x) = 1$

10 / 35

NPTEL

So, we need to understand the constraints; all the words in these sample space must satisfy the following constraints ok. For all x belonging to the sample space the probability of that word would be between 0 and 1 and then the sum of all the probabilities would be equal to 1. This is the very fundamental definition and I am sure most of you would know I am just reiterating for the sake of the NLP ok.

(Refer Slide Time: 14:49)

EXAMPLE - 1

Bag of words *Count* = 52

'a', 'weather', 'balloon', 'is', 'floating',
'at', 'a', 'constant', 'height', 'above',
'earth', 'when', 'it', 'releases', 'a', 'pack',
'of', 'instruments', 'level', 'a', 'if', 'the',
'pack', 'hits', 'the', 'ground', 'with', 'a',
'downward', 'velocity', 'of', 'm', 's', 'how',
'far', 'did', 'the', 'pack', 'fall', 'b',
'calculate', 'the', 'distance', 'the', 'ball',
'has', 'rolled', 'at', 'the', 'end', 'of', 's'

If we are equally likely to pick any word from the BOW, then the probability for any word is
 $P(x) = 1/52, \forall x \in \Omega$ so that
 $P(\Omega) = 1$
 $P(\text{'weather'}) = 1/52 = 0.01923076923$

11 / 35

NPTEL

So, now let take an example. We have the bag of words count as 52. If you want to pick a word from this you know any word it is assuming that they are all put in a bag and then

you just want to pick one word then the probability of picking any word in this is 1 by 52 or 1 or it is equal to 0.01 ok. this is the fraction that we have for the word whether or any other word that you will find here ok.

(Refer Slide Time: 15:33)

EVENTS

'a', 'weather', 'balloon', 'is', 'floating',
 'at', 'a', 'constant', 'height', 'above',
 'earth', 'when', 'it', 'releases', 'a', 'pack',
 'of', 'instruments', 'level', 'a', 'if', 'the',
 'pack', 'hits', 'the', 'ground', 'with', 'a',
 'downward', 'velocity', 'of', 'm', 's', 'how',
 'far', 'did', 'the', 'pack', 'fall', 'b',
 'calculate', 'the', 'distance', 'the', 'ball',
 'has', 'rolled', 'at', 'the', 'end', 'of', 's'

Total number of words = 52. The number of unique words = 37 or there are 37 **types** of words in this BOW. 15 words have frequencies > 1. An **event** is a collection of samples of the same type, $E \subseteq \Omega$

$$P(E) = \sum_{x \in E} P(x) \quad (1)$$

Events can be described as a variable taking a certain value

12 / 35
 NPTEL

Then I am let us now define an event for the same sample space there are a total number of words that we have is 52 as I mentioned and then if you look at the unique verse there are only 37 or there are 37 types of words in this particular bag of words; that means, 57 I am sorry 15 words have frequencies greater than 1 right; that means if you look at the word the right. you would definitely find that particular one happening more number of times than any other word and event let us now define an event.

So, an event is a collection of samples of the same type; right. Since, there are 37 types. we can have 37 events; correct. if you look at the definition of that a probability of an event is defined as this sum of the probability of the individual word that is present in that particular type so; that means, it is going to be the sum of all the occurrences of the particular word let us illustrate that by an example.

(Refer Slide Time: 17:04)

EVENTS - EXAMPLE

'a', 'weather', 'balloon', 'is', 'floating',
'at', 'a', 'constant', 'height', 'above',
'earth', 'when', 'it', 'releases', 'a', 'pack',
'of', 'instruments', 'level', 'a', 'if', 'the',
'pack', 'hits', 'the', 'ground', 'with', 'a',
'downward', 'velocity', 'of', 'm', 's', 'how',
'far', 'did', 'the', 'pack', 'fall', 'b',
'calculate', 'the', 'distance', 'the', 'ball',
'has', 'rolled', 'at', 'the', 'end', 'of', 's'

In the BOW, the word type **the** occurs 6 times. Then

$$E_{the} = 6$$
$$P(E_{the}) = 6 \times \frac{1}{52} = 0.115$$

In the BOW, the word type **pack** occurs 3 times. Then

$$E_{pack} = 3$$
$$P(E_{pack}) = 3 \times \frac{1}{52} = 0.058$$

13 / 35
NPTEL

So, here I am taking 2 words ok. let us assume that we are going to have 2 you events; one is going to contain the type pack, the other one is going to contain the type the ok. you can see that there are 6 occurrences of the word; that means, the event has 6 occurrences. if you want to find the probability of that since there are 6; it is going to be $6 \times 1 \div 52 = 0.115$; ok. In the same case if you look at the word pack there are 3 counts; that means, the probability of the E pack is 0.058 ok.

(Refer Slide Time: 17:59)

RANDOM VARIABLE

- ▶ A **random variable**¹, is a variable whose possible values are numerical outcomes of a random phenomenon
- ▶ Two types - continuous and discrete - for NLP, they are discrete

To capture the type-token distinction, we use random variable W . $W(x)$ maps to the sample $x \in \Omega$.

V is the set of types and the value is represented by a variable v .

Given a random variable V and a value v , $P(V=v)$ is the probability of the event that V takes the value v , i.e.: $P(V=v) = P(x \in \Omega : V(x) = v)$

$$P(V = 'the') = P('the') = 0.115$$

Random variables are useful in describing/constructing various events

¹Random Variable -
http://www.stats.gla.ac.uk/steps/glossary/probability_distributions.html#randvar

14 / 35
NPTEL

Let us move on to the next definition. again these are all very simple for those who have understood or studied probability I am just repeating it for those who have done it a long time ago or who have not done this. this I am just trying to explain this in a very simple fashion; if you like to go and then study in detail you may find a lot of books related to this and a lot of examples; worked out samples are available for you to understand this more in detail; ok.

Let us look at the next term the random variable. A random variable is a variable whose possible values are numerical outcomes of a random phenomenon; ok. There are two types one is a continuous one another one is a discrete one. In this case I am going to be only considering the discrete one because it going to be how a discrete set of random variables that you will find; ok. To capture the type-token distinction let us assume that we have defined a random variable W ; $W \times \Omega$ maps to these samples $x \in \Omega$ again using a different variable V is the set of types and the value is represented by the variable v ok.

So, how do you put them in the formal notation we can say the probability of the random variable equal to capital in that look at the capital part of the equation to the value V is the probability of the event that the random variable V takes the value v ok.

$$P(V=v) = P(x \in R) : V(x)=v$$

$$P(v) = P(\text{'the'})=0.115$$

this is useful in terms of describing and constructing various events.