

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

Lecture – 03
Probability and NLP

(Refer Slide Time: 00:15)

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 ✓

Discrete Sample Space, experiment, joint and conditional probability

NPTEL

The second one is you know we need to go a little deeper into those documents. You know it is not enough that if we just find here it is not enough if we just find the frequency or term frequency or making those accounts and then trying to figure out those ratios and so on. We still want to get something out of the document right. So, what all thing that we can do? So, that is where probability comes into the picture. So, we will talk about this little detail with respect to the document not with respect to the coins.

I am going to be giving a very high-level reason for why we are looking at probability in the natural language process. So, it actually provides methods to predict or make decisions based on what we have counted right. So, you can say for example, I have counted the terms New York right, though they occur always together co-occurrence of those two words. So, next time when I give a word new to the system and if New York occurred so many times than New and any other word the system will be able to predict that the next word for the New would be New York. So, that is where probability comes into the picture ok.

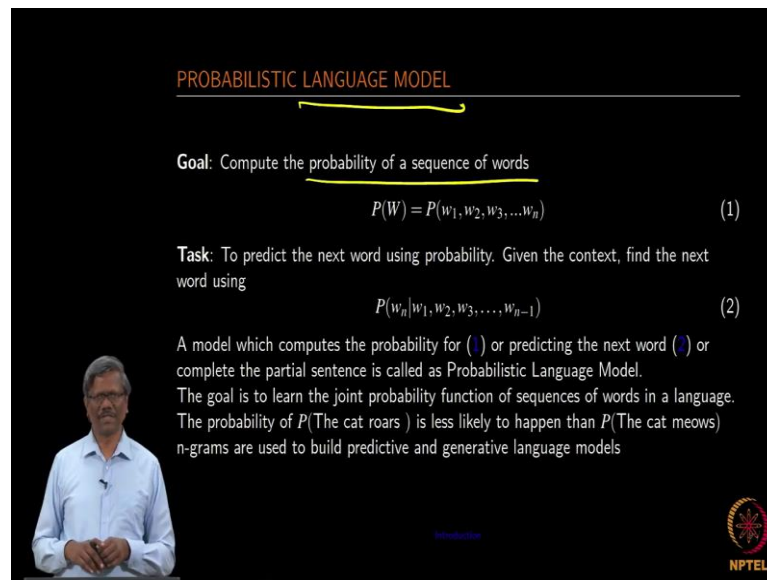
So, we can make informed decisions based on the data that is available to, based on the history that we have captured right. So, this is what we can do with probability. So, we can make some informed decisions right. So, we can also get a quantitative description or chances or likelihood associate with various outcomes. We can look at the probability of a sentence for example, assuming that I have looked at the entire corpus made all the statistical counting and so on and so forthright.

And then I am just giving a new sentence which consists of letting us say four words. So, what is the probability of this sentence being part of the corpus is something that we can deal with if we know the entire details of the corpus and we have counted all the terms and frequencies; we have counted the co-occurrences of words and so on right. So, given a new sentence so, we would be able to get the probability score for that particular sentence whether it is a very high probable sentence according to the corpus or it is not possible to have such a sentence based on the corpus that we in hand. So, that is something we will be talking about.

As I mentioned the earlier probability of the next word in the sentence – how likely to predict you as the next word. So, let us say I am going to be typing ok. So, when I type this word wish how likely we are going to be finding you as the next word? So, we will be bringing in the conditional probability. The probability of this word coming in after wish is so much, you know there is some number that will come along with that and there could be several combinations or several words also that could be coming after wish right.

So, based on which one has the higher value that particular word would be picked of the picked up as the next word. So, this is something that we already spoke about. So, we will be talking about discrete sample space, the kind of experiment we do joint and conditional probabilities and so on. So, we will talk about this during the course and then how really we use all these in natural language processing.

(Refer Slide Time: 04:41)



PROBABILISTIC LANGUAGE MODEL

Goal: Compute the probability of a sequence of words

$$P(W) = P(w_1, w_2, w_3, \dots, w_n) \quad (1)$$

Task: To predict the next word using probability. Given the context, find the next word using

$$P(w_n | w_1, w_2, w_3, \dots, w_{n-1}) \quad (2)$$

A model which computes the probability for (1) or predicting the next word (2) or complete the partial sentence is called as Probabilistic Language Model. The goal is to learn the joint probability function of sequences of words in a language. The probability of $P(\text{The cat roars})$ is less likely to happen than $P(\text{The cat meows})$ n-grams are used to build predictive and generative language models

So, we will also look at as I mentioned earlier how we can compute the probability of a sequence of words. The sentence is nothing, but a sequence of words right. So, we are given three or four words or ten words strung together to form a sentence. What is the likelihood that a particular string of words could be a legal sentence? So, that is something that we can compute using the language model and we will talk about this in detail in one of the weeks. Is explained below

$$P(w) = p(w_1, w_2, w_3, \dots, w_n)$$

$$P(w_1 / w_1, w_2, w_3, \dots, w_{n-1})$$

So, see where we are going right. So, initially we just looked at only the words without really looking at the meaning of any of those. Then, next, we started looking at the statistical counting that we have in hand and then use the probability to start predicting what could be the next word whether the sentence that I am going to be creating is really possible to have as the sentence that I am going to be creating is going to be really probable or it is what is the likelihood of that particular set of words to form as a sentence and so on right. So, those are things that we are going to be doing as a next step.

So, based on the counter that we have done earlier, so, now, we have moved up further to figure out a few things using the probability right. So, that is the second step you know keep remembering that; first counting, second we are just trying to make some sense out of the historical information that we have gathered using the numbers and third we are going to be looking at the sequence of words and then try to figure out whether that particular sentence could be a legal sentence or not based on what history that we have captured right. So, that is the third one.