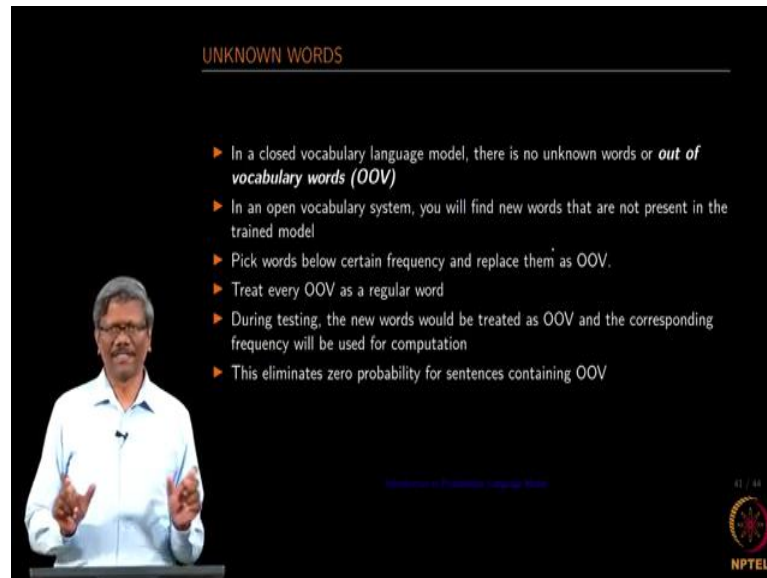


Applied Natural Language Processing
Prof. Ramaseshan Ramachandran
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Lecture - 27
Out of vocabulary words and curse of dimensionality

(Refer Slide Time: 00:15)



UNKNOWN WORDS

- ▶ In a closed vocabulary language model, there is no unknown words or *out of vocabulary words (OOV)*
- ▶ In an open vocabulary system, you will find new words that are not present in the trained model
- ▶ Pick words below certain frequency and replace them as OOV.
- ▶ Treat every OOV as a regular word
- ▶ During testing, the new words would be treated as OOV and the corresponding frequency will be used for computation
- ▶ This eliminates zero probability for sentences containing OOV

41 / 44
NPTEL

So, earlier we mentioned the unknown words, right. So, it is always possible to find a new word in a given sentence; for example, when somebody is giving a lecture, you will find a new word when he is speaking, right. So, immediately you would not know the meaning of that, but based on the context you will be able to make it up and then move forward. So, in the same fashion right, if the corpus did not have that word. So, we should not let the model fail because of the new word that is been given as an input to find either the probability of this sentence or of predicting the next word.

So, we need to create a model. So, that the unknown words are also taken into consideration, how do we do this? So, that is what we going to be talking about in this particular slide. In a closed vocabulary system there is going to be no new word, right. So, we are going to be only talking about the words that are present in the given corpus; whereas in an open system like what we are doing in Google search and so on. So, many new kinds of keywords we keep adding and we want Google to really help us in terms of

finding out where those documents are present and list them in certain orders. So, that we would go through that and then find some info or data out of that document.

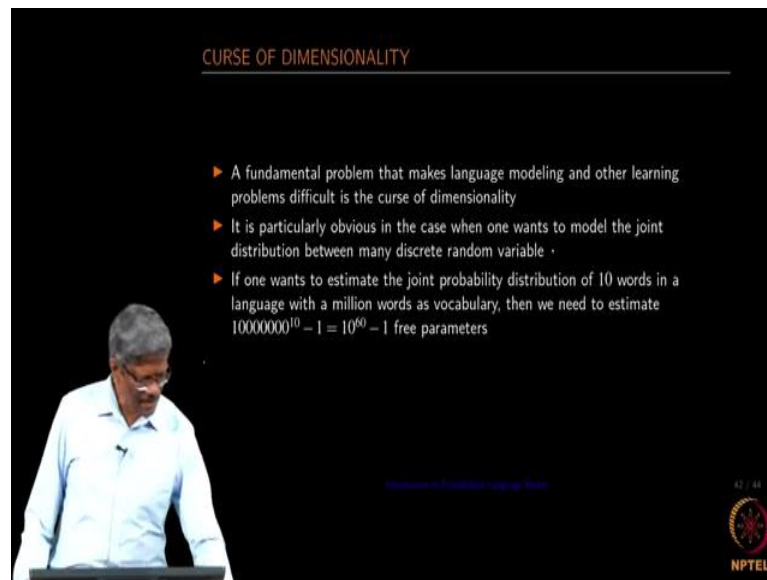
So, there are two purposes, right. So, one is about building the model second one is about querying it; when you query also you are going to be really querying with certain words, and you probably may not know what are all the vocabulary is used in the given corpus. So, you might use your own words which probably are not present in the corpus. So, even there in the query we have to really handle the out of vocabulary words, which we term as OOV. So in this case what we do is, we want to pick up certain words with certain frequencies; for example, there could be so many words in a given corpus with just one occurrence.

So, we can probably term them as, out of vocabulary words and then label them as OOV rather than the actual word itself. So, when a new word comes in as input to the model, we can label that word as OOV. When we do that and when the model has been trained to really look at the OOV based on the frequencies of the world that we have picked and called them us or label them as OOV.

Now, we are not going to have a zero probability for a given sentence, correct. So, in that way we can really address the out of vocabulary words in a given corpus; so this really eliminates our zero probability. So, we treat each and every word in the out of vocabulary as a regular word ok. So, during the model building, since we have labeled certain words with a frequency as OOV those frequencies also would be available as part of the model, right.

So, we would have built a model with those out of vocabulary words, and the corresponding probability also is available for every out of vocabulary word. So, when a new word comes, we label it as OOV; now this OOV is again available as part of the corpus or the model. And then when you use the model which had treated the OOV in the fashion that I have mentioned earlier; we will not have any zero probability for any sentence containing OOV as input ok.

(Refer Slide Time: 04:19)



CURSE OF DIMENSIONALITY

- ▶ A fundamental problem that makes language modeling and other learning problems difficult is the curse of dimensionality
- ▶ It is particularly obvious in the case when one wants to model the joint distribution between many discrete random variables
- ▶ If one wants to estimate the joint probability distribution of 10 words in a language with a million words as vocabulary, then we need to estimate $10000000^{10} - 1 = 10^{60} - 1$ free parameters

NPTEL

So, this is something that I have spoke about earlier, you know while picking up the corpus, right. So, it is really a big concern when you go for a big corpus. So, one good example that I have given as the third bullet point is; if you have a want to have a joint probability distribution of 10 words in a language that contains a million words as vocabulary. Then we need to estimate $10^6 - 1$ free parameters that many the model that I have shown right; so we have to estimate, so many of them. So, when I tried this model on a corpus that is available in NLTK, especially the brown corpus, it took more than one and a half hours to really build that model.

And you know while really testing your code and the application you do not really go for a very high model in the beginning; you just start with a smaller corpus. And then build the model and make sure that your testing is all right with a smaller model; then once you are sure that your code, testing, the model building and the predictions are coming out all right, go for a bigger corpus ok. So, we need to be really concerned about the dimensionality, when you start building the model. So, you cannot eliminate this ok.