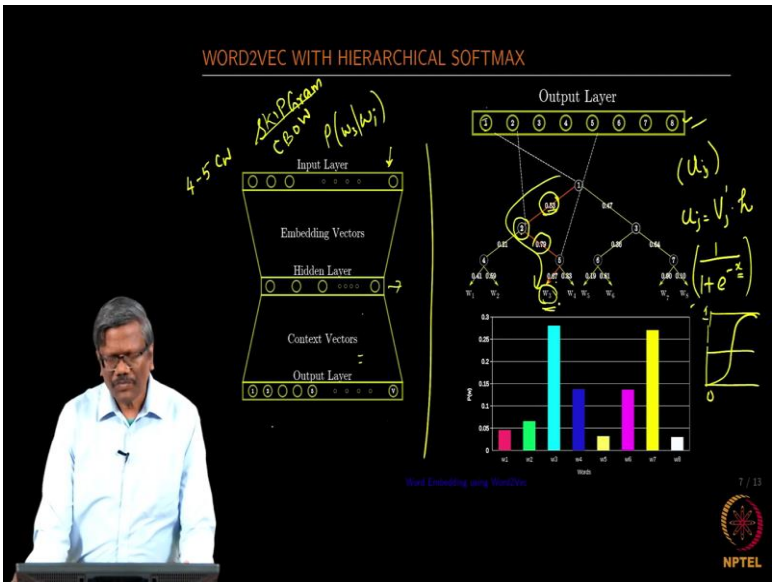
Applied Natural Language Processing
Prof. Ramaseshan Ramachandran
Department of Computer Science and Engineering
Chennai Mathematical Institute

Lecture – 48
Mapping the output layer to Softmax

(Refer Slide Time: 00:15)



So now, we have seen how the words can be arranged in the hierarchical fashion and then see how it fits into the model that we discussed earlier alright. So, we are going to be taking a neural network with 8 words as the vocabulary ok. So, I have given for refreshing your memory the architecture to the left, and then I have taken the notes on the tree structures which we used earlier and I computed the values as well and they are found in the path ok.

So, the idea is to we feed value through the 1 half vector and then we need to compute the probability of the word W_3 as seen in that. For the input word I need to find the probability of, so this is what we are looking at right.

So, how do you really compute? Now, since the structure of this softmax layer is completely changed. And now, the output layer does not really represent the words in that order it just contains some values for 8 neurons ok. So, how do we map them? So, there is it is a slight trick here that we have to apply in order to use those output layer

values and then update those changes in the context vector and the embedding vector alright.

So, I have just shown here also has a graph, so that you will know what is the probability of each of the words which are distributed in the 8-word vocabulary alright. So, now, let us say that we have an input W_1 let us call it the input vector in a skip Gram let us we are taking a skip Gram or C BOW. So, if we are going to be using a skip Gram, so there is going to be only one input and we are going to have multiple context words to be estimated correctly.

And if you are going to be using a CBOW model the input will have let us say 4 or 5 context words ok. So, if you take the dot product of the embedding vector corresponding to that one hat vector a value you get the hidden value here right, so this we explained earlier. So, we have the hidden values, we have the context vector and then compute, so we get some values in the output layer; as we mentioned earlier this does not correspond to the word ok.

So, how are we now going to update the weights if it is not going to be the word if each one of the neurons is not corresponding to the word ok? Let us now look at this when it travels from the root node to the second node what you do is your first look at where it starts from we always start from the root node that is equal to 1 right

So, we assume that this node is this node that corresponds to the node in the output layer 1, so instead of taking this softmax we are going to using a sigmoid. So, u_{ji} is nothing but right this is where it is, so here it is where you calculate each value I am using this value as u_j not that index the value that is computed as u j ok.

So, take u j value from neuron one apply sigmoid, what is a sigmoid I am sure you will remember this it is $\frac{1}{1+e^{-x}}$ right. So, applying the value of u_j in this you get value and then sigmoid is a function that translates the values of u_j between 0 and 1 correct the value would be between 0 and 1.

So, when you have the value between 0 and 1, we call that as the probability of the node here ok. So, once this is computed which is corresponding to the node in the output layer and then we now go to node 0, So, what do you take the value from here and then compute the value, so this corresponds to the probability here ok.

So, this is the relative probability of the word 3 or node 5 and then this is the relative probability of node 2 and then this is the relative probability of the leaf node. So, when we do that what we also do is we compute this value right. So, this is nothing, but 1 minus right got it, so in this fashion we now have computed the probability of the word.

So, we have shown earlier that by doing this we are creating a value W_3 which is normalized right across the vocabulary. So, 1 it is done we have the value of W_3 which is a normalized value and how do you now get back to the values here and then correct them. So, now, we need to find out what is the error, so since it does not correspond to any of the words directly how do we make the decision let us look at that.