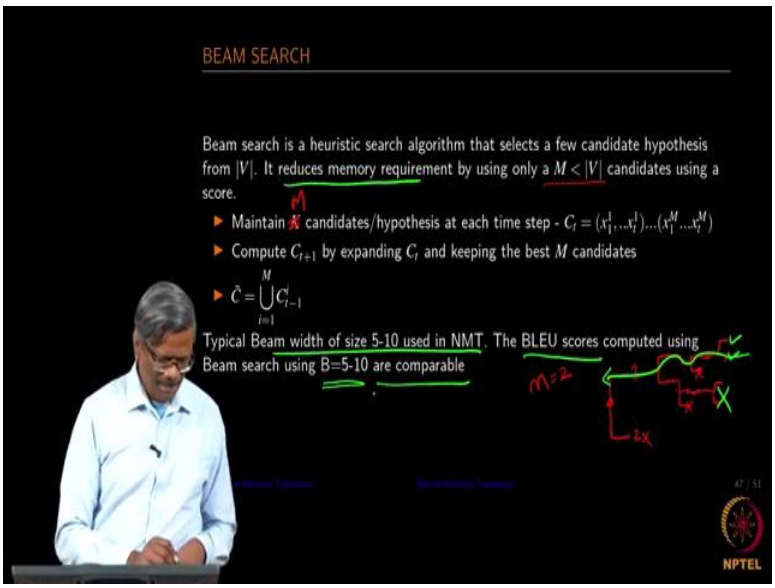
Applied Natural Language Processing
Prof. Ramaseshan Ramachandran
Department of Computer Science and Engineering
Chennai Mathematical Institute, Madras

Lecture – 81
Beam Search

(Refer Slide Time: 00:15)



Ok. In this session we going to be talking about the Beam Search. I am sure you heard about this you know several times during your neural machine translation sessions; we did not elaborate this earlier. So, now, we going to be talking a little bit about what is beam search is all about. This is a heuristic search algorithm that selects a few candidate hypotheses from a set of hypotheses. Is shown below

$$c_t = (x_1^1 \ldots . x_t^1) \ldots (x_1^m \ldots x_t^m)$$

$$\acute{c} = \bigcup_{i=1}^{m} c_{t-1}^i$$

So, why is this required? I am sure you would all know right. So, every time when we want to predict the word. So, we have the vocabulary size of about let us say 50 k right.

So; that means, the prediction gives you about 50000 values, out of which only about 10 or 15 would be very useful to us, reset would not be very useful. So, using that as a heuristic we want to reduce the number of steps involved in terms of finding out our sentences. At the end of every decoding process, we will end up with a lot of sentences and we want to pick only what is important to us with respect to the data that is going to be coming from the translation models right.

So, this reduces the memory size drastically, because we are only picking up a very small number of candidates or a small subset from the original set. And every time we maintain the number of steps, they should be M, number of candidates hypothesis at each time step; for example, when I start from here, I will have let us say if I have 2 as my m equal to 2 here; I will have 1 and then 2.

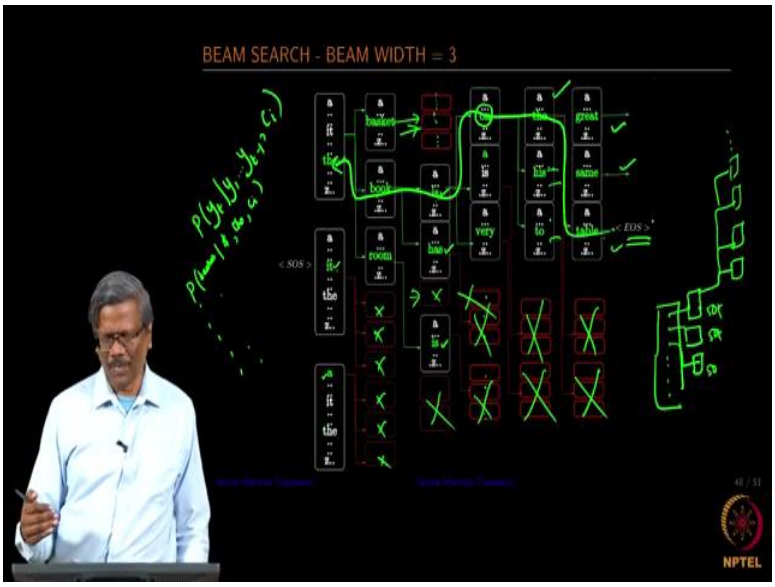
So, whichever value is good enough for me to move to the next stage, let us say that these values are obtained based on some calculations. So, I will get two outputs from this and this is removed; and then again there are possibilities I can have two here and two here all right. So, out of this again you can figure out which is best suited to you; let us say that this one and this one is suited and these are eliminated, again we will have two from here, two from here and so on.

So, in this process, we try to take say for example, if I say this value is good enough and this is good enough and this branch is broken. And then based on certain conditions we will stop and then do the backtracking of this to find out what is the original sequence that we have computed using this algorithm.

And then it is possible to have two here and then those two sentences or these sequences are picked up and then using some computation, let us say we are going to be using the conditional probability and then whichever probability is higher we will pick that sentence and then move on. So, this is the very standard heuristic algorithm set up, which helps us in terms of reducing the memory size for picking up our translated sentences ok.

Let us take some examples and then see how this works. In the case of machine translation the beam size is usually around 5 to 10 and then BLEU scores computed using these 5 to 10 right, so they are very comparable. So, most of the time you will see in the newer machine translation the beam size to be around either 5 or 10 OKs.

(Refer Slide Time: 04:39)



So, I will take some examples and then see how it is computed, in the case of a machine translation. So, when you start decoding the input sentence, right, so the decoder gets this is the start of this sentence, the decoder has given us all the context based on the attention or no attention or not it is not worried about that part; let us assume that we have received some context and then we are going to be computing the or predicting the first word right.

So, in the case of the prediction of the first word depending on the context that is coming in, depending on the first dollar or start of this sentence word, we are going to be getting a softmax that gives you all 50000 values right. So, there is going to be, let us say that there are going to be about 50000 values that are coming in through the softmax and based on this starts of the sentence; there are going to be only about a small subset of those words that will start as the starting word right. Let us say that it and a or picked up and these three words have a higher probability than any other word. So, we picked up this ok. So, now, I am just putting them as three different stuff.

So, since our beam size is three, width is three, I have now three possibilities. So, instead of taking all 50,000 and then start forming these sentences or predicting the next word for each one of them, imagine the size of that right. So, if I take all 50000. So, I will be having another 50,000, another 50 k, another 50 k each one will have 50 k right, and then this again will have 50 k and so on. So, instead of picking up and then trying to figure out all possible combinations of sentences, we are restricting our attention to only the top three right. So, assume that these three words are at the top of the list.

So, since our beam width is three, we picked up this and then put them as separate ok. And then start of this term, see our what is our goal is to predict or do the get the conditional probability of y_t depending on y_1 to y_t minus given the input context correct; this is what we are predicting every time. So, in this case we have three of this and then we will start with those three OKs. So and let us see how many sentences that we can form with these. So, now, what we do is, the next in this is the case this is. So, we have found out these three, the conditional probability of these things are higher.

So, they are listed here the next one is, so I will just have the basket, dollar comma the and the context. So, in the same fashion, we will have for it, a and so on right. So, we can keep computing this every time and then the product of the probability at the end of this sentence once we have done this; supposing if we have created this sequence, the product of each of the conditional probability will give you the total score right for that sentence that we have picked up. So, in this case now basket, book and room; let us assume that for it and all those values did not really succeed in terms of coming into the bucket of three OKs.

So, now the basket, the book, the room we have got; and then trying to predict the next one. So, again you will have several of these right 50000, 50000, 50000 and then for the book again we have 50000 values and so on right. So, out of this what is the conditional probability of picking up this, where boo and the context are given. Let us say that this is higher, this is higher, and this one is higher than many of what you have right in the basket case as well as in the book, you know there is one more that it is not picked up. So, these three values are higher. So, we picked up, again there is a beam of three here. So, now, the basket is gone, the book is, the book has, the room has, the room is.

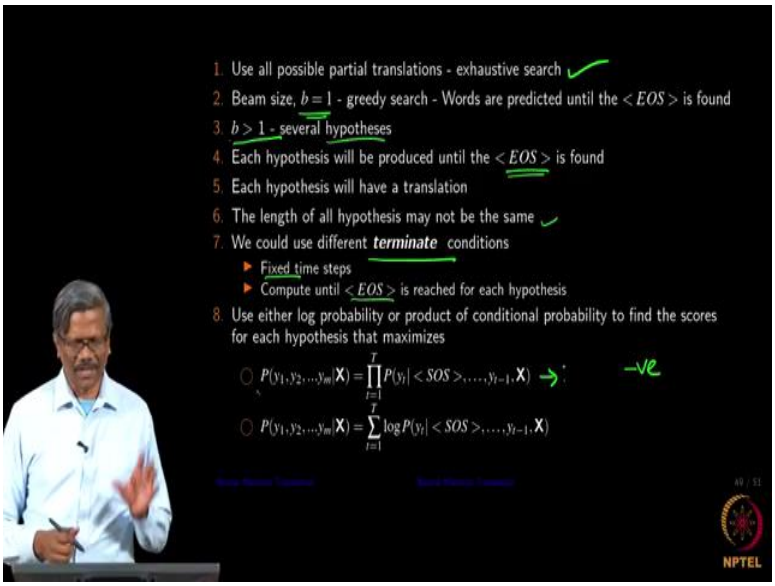
So, these are the possible sentences that we going to be having ok. And then now based on the words, we going to be predicting the next word on, a, very based on these three words rights. So, in this case we have on, a, and very; let us assume that all those the conditional probability values are a lot lower than what we have picked up here right, and they are eliminated. So, now, these sentences that you could form are the book is on, the book is a book is very; and then this has also is moved out of the picture. So, now, we have again another three and we can start picking up or predicting the next word based on, is, book, the, and the context that is coming in from the encoder.

So, we get the, his, to as three words that are predicted to be having the higher value, when compared to these right. And then now the sentence could be the book is on the book is, no these are all gone right. So, now, the book is on the book is his, the book is on to and so on. So, again using these three words, we get three more and then for the, we have all the possible words picked up; and for his and to we have eliminated right. And then now the sentences could be the book is on the table ok, the book is yeah that is it. So, the rest of them do not come into the picture or we can just say his, this could be yeah that is it right.

So, now the sentences could be based on what you see, they say the end of this sentence is found. So, we can backtrack from here, like this right. So, in this way until the end of the sentence you can keep going; for example, even for these, you can still do it. In this case I just made up those words so that it ends here. In many cases you can continue to have more number of sentences until the end of sentences reached ok.

So, this is one way of finding out the possible sentence that is translated from this source to the target. So, this need not be the best, this is based on what you have achieved depending on this course that we have computed, the training model, and so on. So, this need not correspond to the translation the human would translate into ok.

(Refer Slide Time: 14:46)



$$\mathbf{P}(y_1, y_2, \ldots . y_m|X) = \prod_{i=1}^{T} P(y_1 | < sos > \cdots y_{t-1}, x)$$

$$\mathbf{P}(y_1, y_2, \ldots . y_m|X) = \sum_{i=1}^{T} \log p(y_t | < sos > \cdots y_{t-1}, x)$$

So, if you use all possible partial translations instead of what we have done it, becomes an exhaustive search. If the beam size is equal to 1, we only take an agreed search for example; for the highest value from the top always takes. For example, if this is the highest value, you keep moving into this right, the basket is something correct. So, that will not give you the right translation and if beam size or the width is greater than 1, we get several hypotheses depending on what is your beam size. So, each hypothesis will be produced until the end of the sentence is formed or you can chop it off with a certain time slice.

Each I hypotheses will have a translation when you use the beams search. The length of all hypotheses may not be the same; for example, in the case previous case, if there are three sentences one could be shorter, one could be longer, one could be in between those two right. We could use several different terminate conditions; one is the end of a sentence, another one is fixed at the time step ok. We can either use log probability or product of conditional probability, find the scores of each hypothesis that maximizes the probability here right.

So, in this case, we have just mentioned that we are picking up that based on the highest value that is formed from the softmax and when we want to form the sentence, you can either use the log probability so that every conditional probability you can add up to and finally, get a score. And when you use the log probability, the score would be on the negative side; because it is within the range of 0 to 1 right.

And then, in this case, you will have a very small number because you are I am sorry; in this case, it will be a very small number because we are doing the multiplication and the values would be too small as well. So, in many models, you will see the log probability