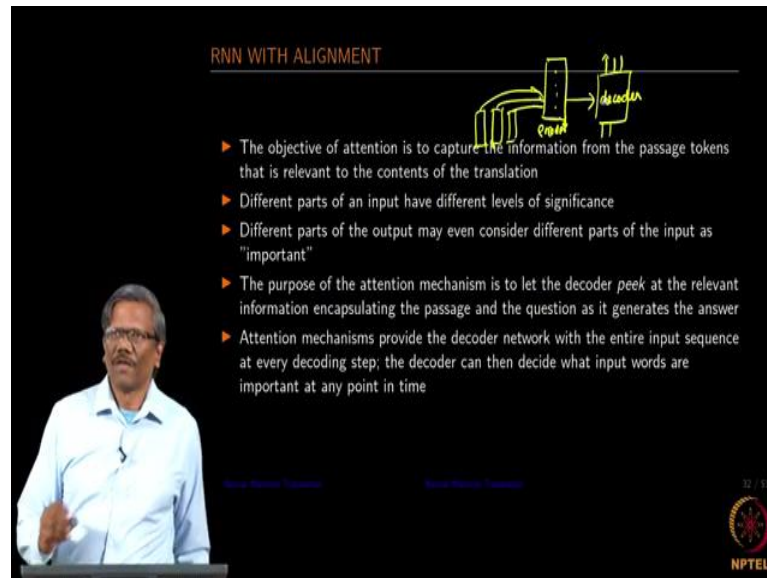


**Applied Natural Language Processing**  
**Prof. Ramaseshan Ramachandran**  
**Department of Computer Science and Engineering**  
**Chennai Mathematical Institute, Madras**

**Lecture – 78**  
**Introduction to Attention-based Translation**

(Refer Slide Time: 00:15)



**RNN WITH ALIGNMENT**

The diagram shows an encoder (RNN) processing an input sequence and producing a hidden state vector, which is then fed into a decoder (RNN) to generate an output sequence. The decoder is shown with a 'peek' arrow pointing back to the encoder's hidden state.

- ▶ The objective of attention is to capture the information from the passage tokens that is relevant to the contents of the translation
- ▶ Different parts of an input have different levels of significance
- ▶ Different parts of the output may even consider different parts of the input as "important"
- ▶ The purpose of the attention mechanism is to let the decoder peek at the relevant information encapsulating the passage and the question as it generates the answer
- ▶ Attention mechanisms provide the decoder network with the entire input sequence at every decoding step; the decoder can then decide what input words are important at any point in time

32 / 51  
NPTEL

We have seen earlier that RNN is very good in terms of understanding the translation patterns if it is a very short sentence I must say that.

And we know that missions there is design to you know do the translation part is really learning in the patterns between two languages and then use that pattern in terms of providing the translation. So, that is a fundamental aspect of any machine learning right. So, it learns the patterns which are not very clearly visible to us when the sizes very huge.

So, and the progression of the machine translation especially in the neural machine translation, where we just provided in the earlier neural net model and output from the encoder right, it is an encoded sequence of input sentence, right. So, this encoder gives a small vector which is said as the input to the decoder, right. So, this is encrypting the entire sentence that is in input into a fixit sized vector right; this is a fixed-sized vector that is output from the encoder, right.

Then this goes as input to our decoder right and then the decoder output now that we are predicting depends on the contexts that we have an encoder. So, we can call this as the context right; this context is fed into the decoder and then it is used to predict the next word during the training process, ok. So, let me erase this, ok.

So, now what is the next step? So, as we have seen in the statistical machine translation where we moved from word to word translation to alignment base translation, right. So, we created an alignment model and then using the alignment model later we try to translate the input sentence into the target sentence. And then the same alignment model was extended to phrase-based alignment; instead of just doing word-based alignment, we started aligning phrases right between the input and the output sequence.

In the same fashion, now we won't extend the neural phrase to translation also into the alignment-based model, ok. So, how do we do that? So, this is what we going to be saying in this particular session. So, the objecting of this attention is to capture the information from the passage tokens that is relevant to the contents of the translation. So, we have a sequence like what we have here, right and then we want to translate into a target language and then we want to make sure that the information is captured partially I like this and then use this partial translation or the partial tokens for adding the translation, ok.

Earlier what we did? We took the entire sentence, is not it? So, we took the entire sentence like this and then it is we created a context vector, correct. So, instead; so can we not just take part in this and then start aligning that with the target sentence and then see whether the translation would do a better job whether the NMT would do a better job or not right that is what you want to try.

(Refer Slide Time: 04:52)

**RNN WITH ALIGNMENT**

- ▶ The objective of attention is to capture the information from the passage tokens that is relevant to the contents of the translation
- ▶ Different parts of an input have different levels of significance
- ▶ Different parts of the output may even consider different parts of the input as "important"
- ▶ The purpose of the attention mechanism is to let the decoder peek at the relevant information encapsulating the passage and the question as it generates the answer
- ▶ Attention mechanisms provide the decoder network with the entire input sequence at every decoding step; the decoder can then decide what input words are important at any point in time

NPTEL

Can we know that the different parts of the input have different levels of significance, right? So, this we saw even in the statistical machine translation. So, some parts you know for example, in the third one different parts of the output may even come in different parts of the input as important. It is not going to be one phrase to the other phrase you know; one to one kind of correspondence from in terms of the length. For example, this is not going to be corresponding to the second word in the translation rather in the target language, right.

So, for example, this particular sequence could be available as part of the target language at the eighth position or ninth position. So, for example, this is in the third position let us say if we consider this 1, 2, 3 and 4 right as in the fourth position.

So, this phrase maybe belonging to the third position in the target language. So, we need to figure out, how do we really align those; how do we really align when different parts align to of the different parts of the input sentence align to different parts of the output sentence.

So, the idea also used to lift the decoder look at relevant portions of the input sentence, which is important to it during that time slice, ok. For example, if we are in the decoding process during the training. So, what we are doing is at letting us say timestamp  $t$ ; there is a hidden value that we are computing, ok.

So, at that point time decoder would like to look at what is there in the input sequence corresponding to that particular time slot or time  $t$  plus 1 slot and so on understand what I am saying. See for example, we have the decoder here and we have the encoder and then; so I am computing this hidden value, ok.

So, it wants to find out which part of this; this is let us say sentence number 1 of the sequence and this is word number 1 of that sentence 2, 3 and so on, ok. So, this is our  $y$ . So, I want to know which part of this sentence of input is useful to me. So, I want this particular note would like to look at or peek into the relevant information encapsulating the passage in the input sequence. So, I will come to this part again once I finish the last one. The attention mechanism provides the decoder network with the entire input sequence at every decoding step.

So, what it does is at every time slice in the decoder part; it knows what is the entire sequence of input you know how is it decoder and so on so. And then it can look at it and then figure out whether that is useful to it or not; let us say, for example, it can take a look at only this portion of the input or it can only look at this portion of the input and so on and decide which is important to it and then move along.

So, coming back to the fourth point right and especially the last part of the sentence the same as I mentioned earlier in the previous session RNNs can also be used for question answering.

(Refer Slide Time: 09:40)

**RNN WITH ALIGNMENT**

- ▶ The objective of attention is to capture the information from the passage tokens that is relevant to the contents of the translation
- ▶ Different parts of an input have different levels of significance
- ▶ Different parts of the output may even consider different parts of the input as "important"
- ▶ The purpose of the attention mechanism is to let the decoder peek at the relevant information encapsulating the passage and the question as it generates the answer
- ▶ Attention mechanisms provide the decoder network with the entire input sequence at every decoding step; the decoder can then decide what input words are important at any point in time

Diagram illustrating the attention mechanism: An encoder processes input  $x$  to produce a hidden state  $A$ . This state  $A$  is used by a decoder to produce output  $y$  from input  $z$ . Handwritten notes highlight  $A$ ,  $Qx$ , and  $Answer$ .

32 / 51  
NPTEL

For example you have; let us say we have questions here fed as input and then we have corresponding answers, ok. So, this is very useful especially in the IT services space where they really serve the customer and in a lot of questions coming in. It could be used for banks, it could be used for the insurance company, it could be for the health care what not right. So, they are they have customers and customers are asking questions and then you want to find out the right answer for the question that is incoming, correct.

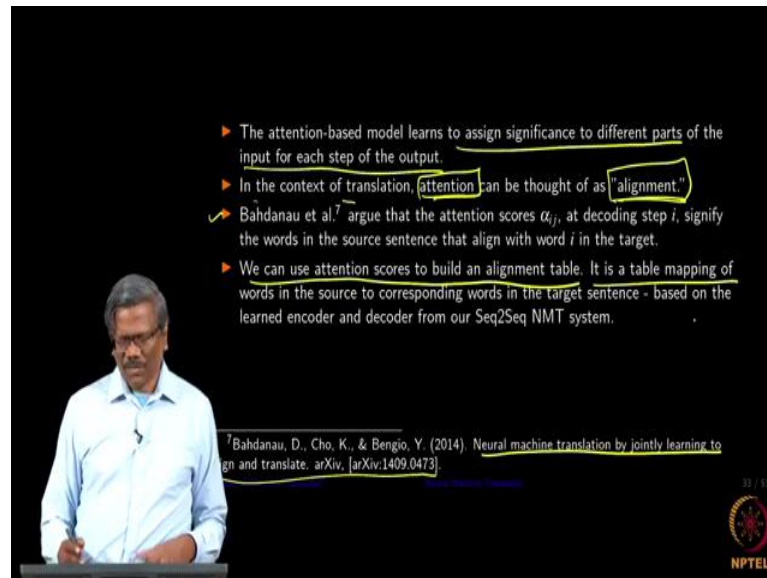
So, what you can do is; you can actually train in the network based on the corpus of questions and the relevant answers to each of the questions and train the network. So, the questions can come in various forms, for example, for the same answer there could be a variety of questions that are coming in the right you understand that right.

So, there could you can phrase the question in several ways, but the answer would be only a single answer for that; all of them mean the same, but they are constructed differently. So, while generating the answer for this ok; so you can actually look at all the relevant words related to that. It is very similar to what we had seen in the word embedding right where we have this skip-gram as well as the CIBO model. In the CIBO model, we actually provided the context and we identified the central word, correct.

So, in the same fashion we think of the CIBO model where the questions are all coming as context questions, the question is looked at as context words here. The reason why I am looking at them as the context word is in the case of this CIBO; we are able to really relate the central word with the set of context words, in the same fashion we can relate the answer with the set of contextual questions.

The neural nets are really good in terms of understanding the context understanding the pattern so that when in the next production cycle when a question of that type; you know need not be same as what it had seen comes in based on the contextual words, it would be able to predict the answer, ok. So, we can use the set of RNNs one for encoding the question and another one for answering the question there are coming in into the encoder here, ok. So, that is where this is going to be useful alright.

(Refer Slide Time: 13:05)



- ▶ The attention-based model learns to assign significance to different parts of the input for each step of the output.
- ▶ In the context of translation, attention can be thought of as "alignment."
- ▶ Bahdanau et al.<sup>7</sup> argue that the attention scores  $\alpha_i$ , at decoding step  $i$ , signify the words in the source sentence that align with word  $i$  in the target.
- ▶ We can use attention scores to build an alignment table. It is a table mapping of words in the source to corresponding words in the target sentence - based on the learned encoder and decoder from our Seq2Seq NMT system.

<sup>7</sup>Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv, [arXiv:1409.0473].

NPTEL

So, we have seen this fifth one as well right, that is moving on to the next one. So, what happens when we do the training process at; when we try to peek in from the decoder into the encoder; the attention-based model learn to assign different parts of the input for each step of the output.

So, it is able to really assign say for example, this portion of the target is aligned to this portion. Since it keeps looking at the patterns when you provide close to about 500 million pairs of sentences; it captures those patterns and disabes to really relate those very clearly. In the context of the translation here, the attention can be thought of us alignment, it is very similar to what we saw in the statistical machine translation where they where we used alignment and here we going to be using the term attention.

Bahdanau and others they actually are the creators of this attention-based model. So, this is the seminal paper you may want to read this paper. They claim that the attention is automatically learned when you are able to provide the attention mechanism in the neural net. So, let us look at the details of that little later. This we can also use the attention score to build the alignment table. So, here the alignments are not pre-computed; it is learned along the training process as well.

So, we know that it is a table of mapping of words in the source to corresponding words in the target sentence ok. So, automatically this particular model learns as well as aligns the sentences on its own. So, there is no separate module for creating attention.