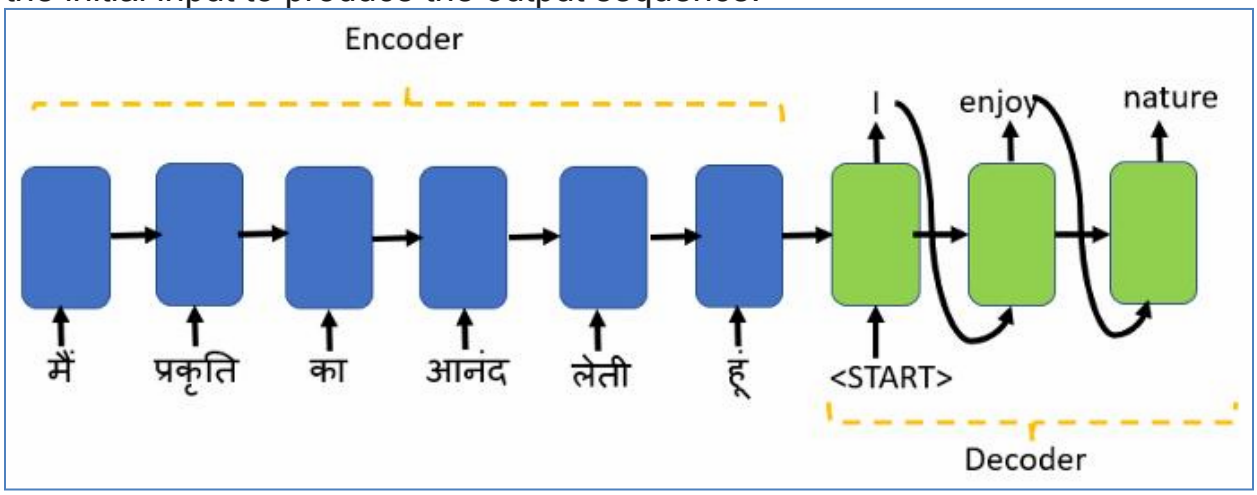
Intuitive explanation of beam search

In order to understand beam search, we will take sequence-to-sequence neural machine translation as an example.

The sequence-to-sequence model uses an encoder and decoder framework with long short-term memory (LSTM) or gated recurrent unit (GRU) as basic blocks.

The Encoder maps the source sequence, encodes the source information, and passes it to the decoder. The decoder takes the encoded data from the encoder as input, and at the same time the string start character<START> As the initial input to produce the output sequence.



This source sequence (source sentence) is a sentence in Hindi, and the target sequence is generated in English. I hope to choose the most suitable word for the translation and most likely to match the meaning of the Hindi sentence.

How to choose the best and most likely words for the target sequence?

A simple method is to build a vocabulary in the target language, such as 10,000 words, and then get the probability of 10,000 target words based on the source sequence. There may be many possible translations of the source sentence in the target language.

Should any translation be chosen randomly?

Our goal is to select the best and most likely translation word, so we choose the target word with the highest probability based on the source sentence.

Should I just pick the best translation?

The Greedy Search algorithm selects a best candidate as the input sequence for each time step. Choosing only one best candidate may be suitable for the current time step, but when we construct a complete sentence, it may be a suboptimal choice.

The Beam Search algorithm selects multiple alternatives for the input sequence at each time step based on conditional probability. The number of multiple alternatives depends on a parameter called Beam Width B B B. At each time step, beam search options B B B The best alternative with the highest probability is the most probable choice at this time step.

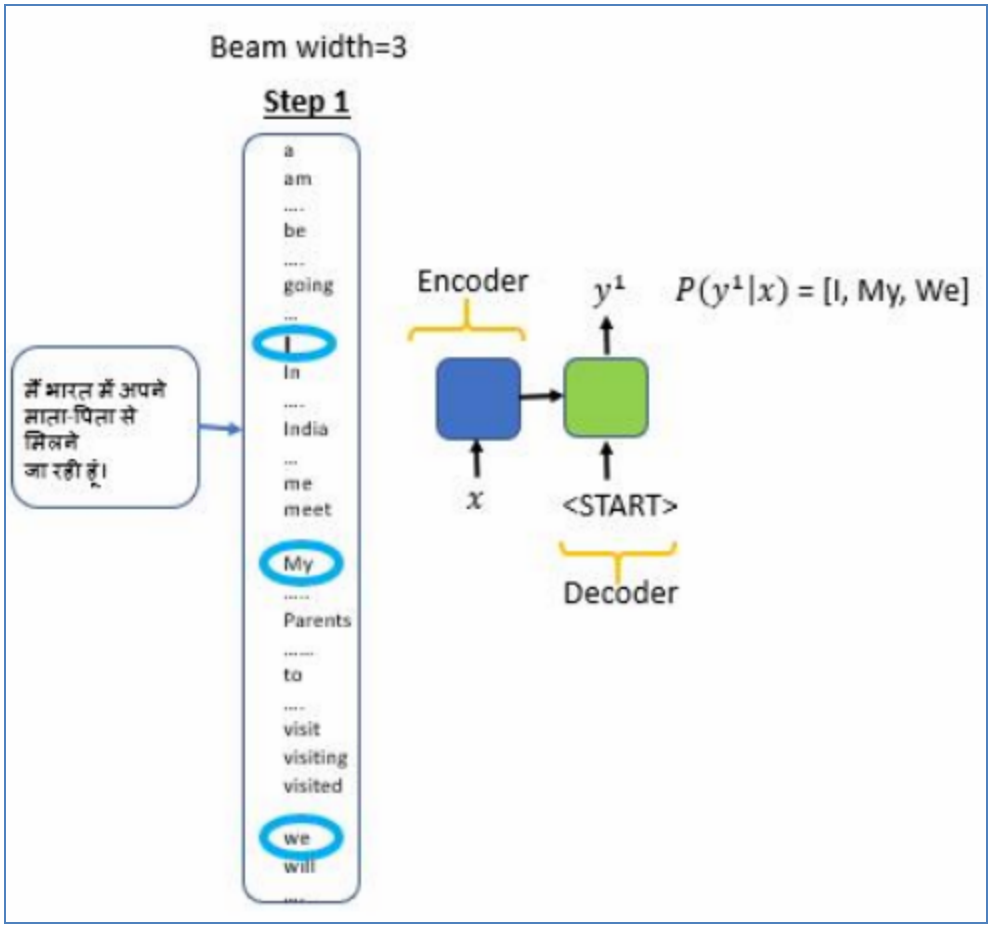
Let us take an example to understand this.

Hindi input sequence → मैं भारत में अपने माता-पिता से मिलने जा रही हूं।
English probable output sequence → ?

We will choose beam width=3; there are 10,000 English words.

Step 1: Given the input sentence, find the top 3 words with the highest probability. The most likely number of words is based on the beam width.

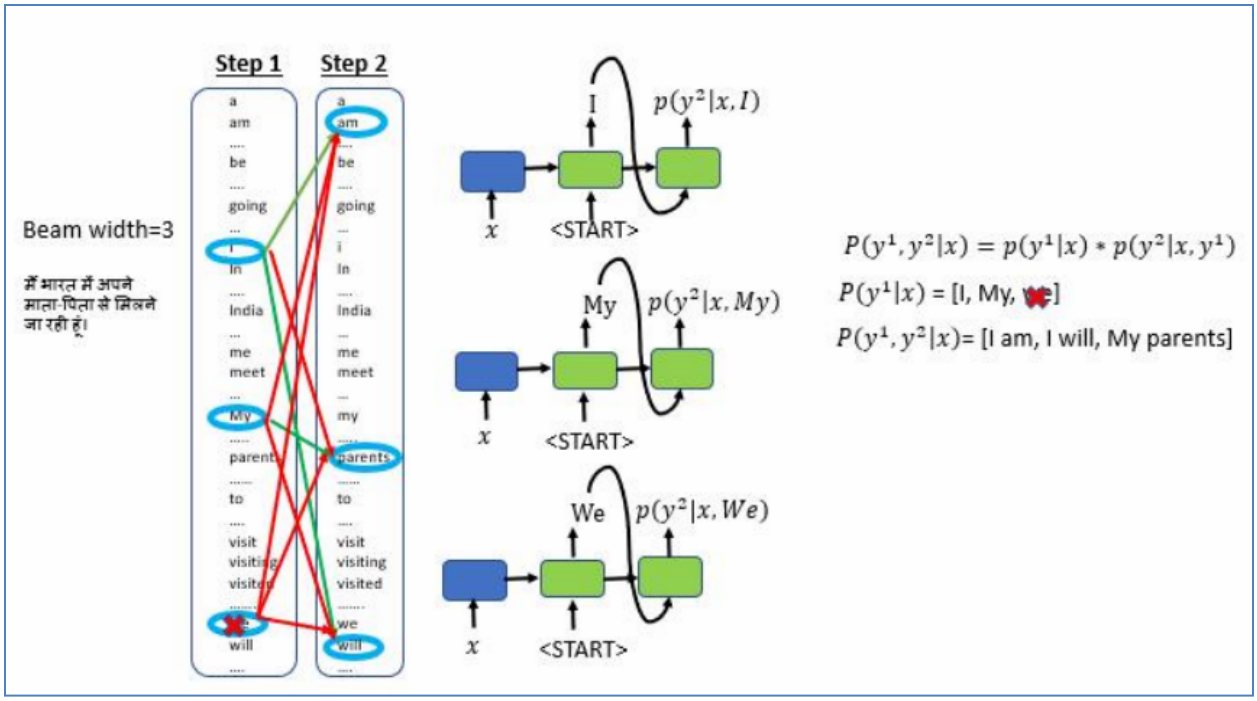
- The encoded input sentence is input to the decoder; the decoder will apply the softmax function to all 10,000 words in the vocabulary.
- From 10,000 probability scores, select the top 3 words with the highest probability.
- Consider the 3 most likely options for translating words, because the beam width is set to 3. If the beam width is set to 10, the top 10 words with the highest probability are selected.
- Store the three most important words: I, My, We in the memory.



Step 1:Find three words with the highest probability based on the input sentence

Greedy search always considers only the best option.

Step 2: Find the three best pairs of the first and second words according to the conditional probability.



Step 2: Find the top 3 pair of words for the first and second word of the translated sentence

Use the first three words (I, My, We) selected in the first step as the input in the second step.

Apply the softmax function to all 10,000 words in the vocabulary and find the three best options for the second word (am, parent, will in the picture above). While doing this, we will use conditional probability to find the combination of the first and second words most likely to form a pair.

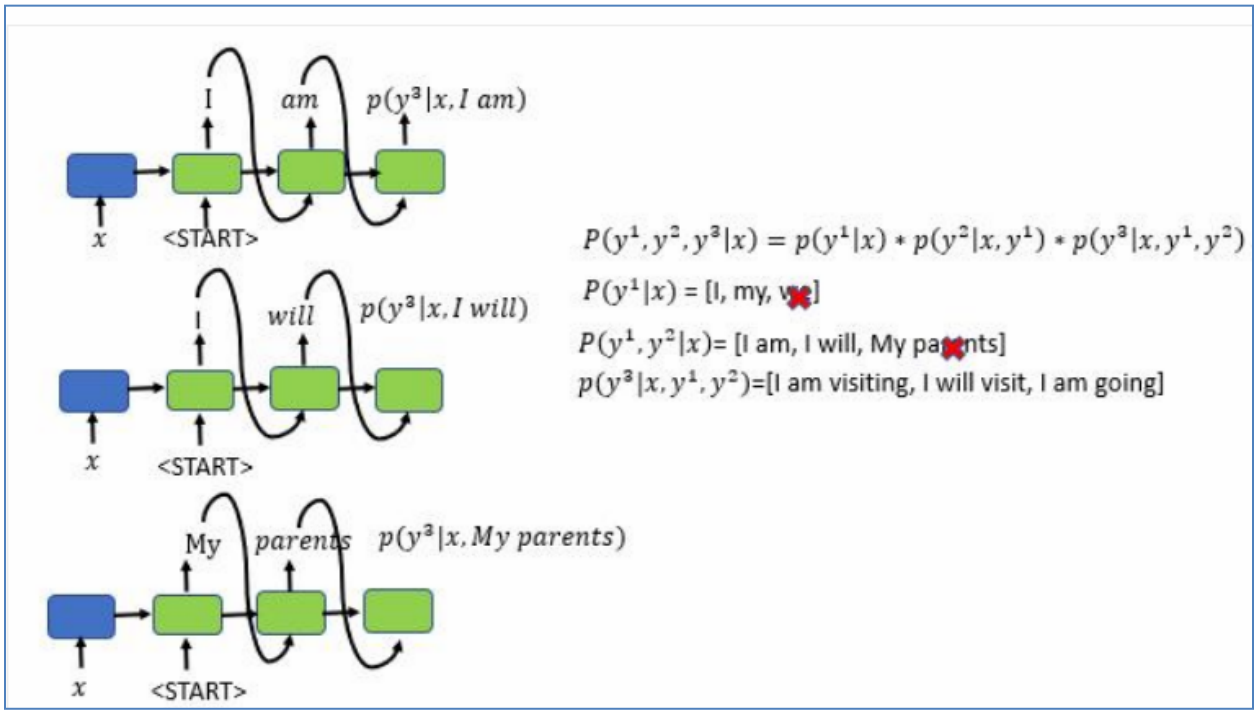
In order to find the 3 best pairs of the first and second words, we first take the first word “I”, Apply the softmax function to all 10,000 words in the vocabulary. Candidates in the first word“My” with “We” Do the same operation.

The above operation needs to run 30,000 different combinations to select the first 3 word pairs of the first and second words. The result is:“I am”, “My parents”, “I will”。

Delete the first word “We”, Because no “We” The conditional probability of the first word and the second word is very high.

At each step, three copies of the encoder-decoder network are instantiated to evaluate these partial sentence fragments and output. The number of copies of the network is the same as the beam width.

The third step: According to the input sentence and the first and second words selected, find the three best pairs of the first, second and third words.



Step 3: find the most likely choice of the first three words in the translated sentence

Together with the input sentence and the word pair composed of the first 3 first and second words obtained in the previous step: *“I am”*、*“My parents” with “I will*”, Find the word pair consisting of the third word with the highest conditional probability.

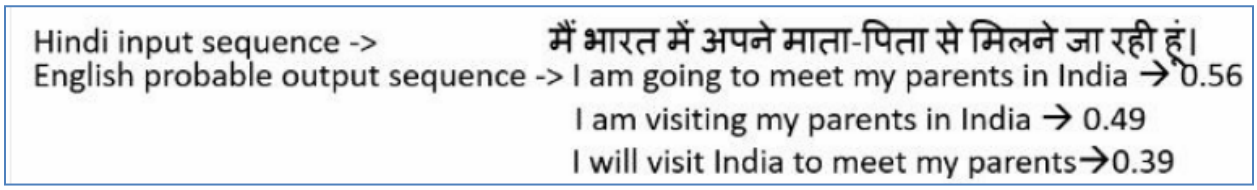
The above operation needs to run 30,000 combinations again to select the best and most likely combination of the first, second, and third words, and instantiate three copies of the seq2seq and encoder-decoder models.

The first two first, second, and third word pairs are:

“I am visiting”、“I will visit” with “I am going”。

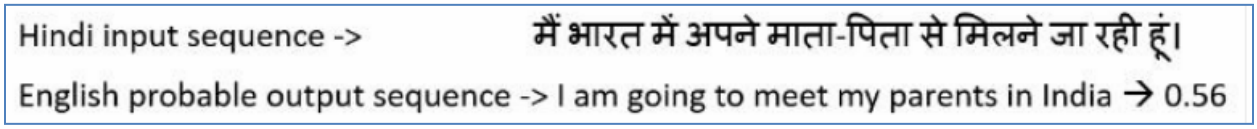
Delete “My parents” Combination, because we did not find a word pair related to it in the word pair composed of the first three words.

Continue this process and select the 3 sentences with the highest probability. The length of these 3 sentences can be different or the same.



Three output sentences with the highest conditional probabilities and different lengths

We finally choose the output of the decoder as the sentence with the highest probability.



The higher the beam width, the better the translation effect will be?

A higher beam width will provide better translation results, but will use a lot of memory and computing power.

When the beamwidth is 3 and the vocabulary is 10000, it is necessary to evaluate 30,000 combinations at each time step, create 3 encoder-decoder instances, and the maximum sentence length is 9. Creating multiple copies of the encoder-decoder and calculating the conditional probability of 30,000 words at each time step requires a lot of memory and computing power.

Lower beam width will lead to poor translation results, but it reduces memory usage and computing resources.

Conclusion

Beam search is the most popular search strategy in sequence to sequence deep natural language processing algorithms, such as neural machine translation, image captioning, chatbots, etc.

Reference :

An intuitive explanation of Beam Search
A simple to understand explanation of Beam search
Renu Khandelwal
https://towardsdatascience.com/an-intuitive-explanation-of-beam-search-9b1d744e7a0f#_=_
https://www.programmersought.com/article/92285907653/