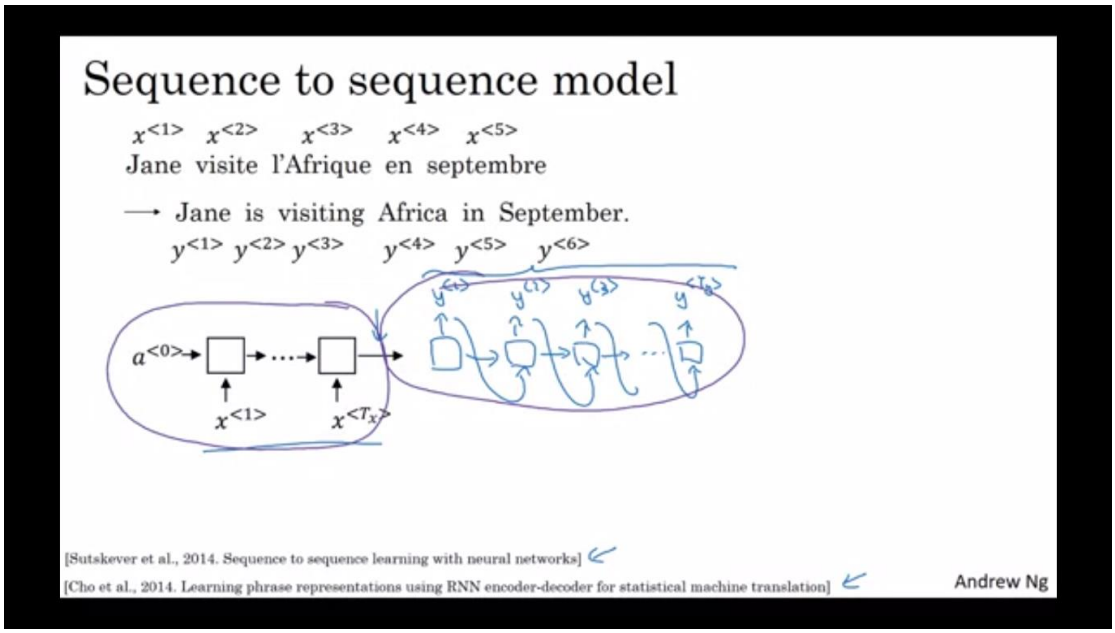
Ahmad Hussameldin Hamed Hassan

Shared Git-hub link: https://github.com/ahmadhassan1993/sharing-github

Sequence-to-Sequence

Sequence-to-Sequence is generating sequence from sequence. Like Machine translation, where we use Encoder then Decoder (Language Model). This is called Conditional Language Model (conditional probability).



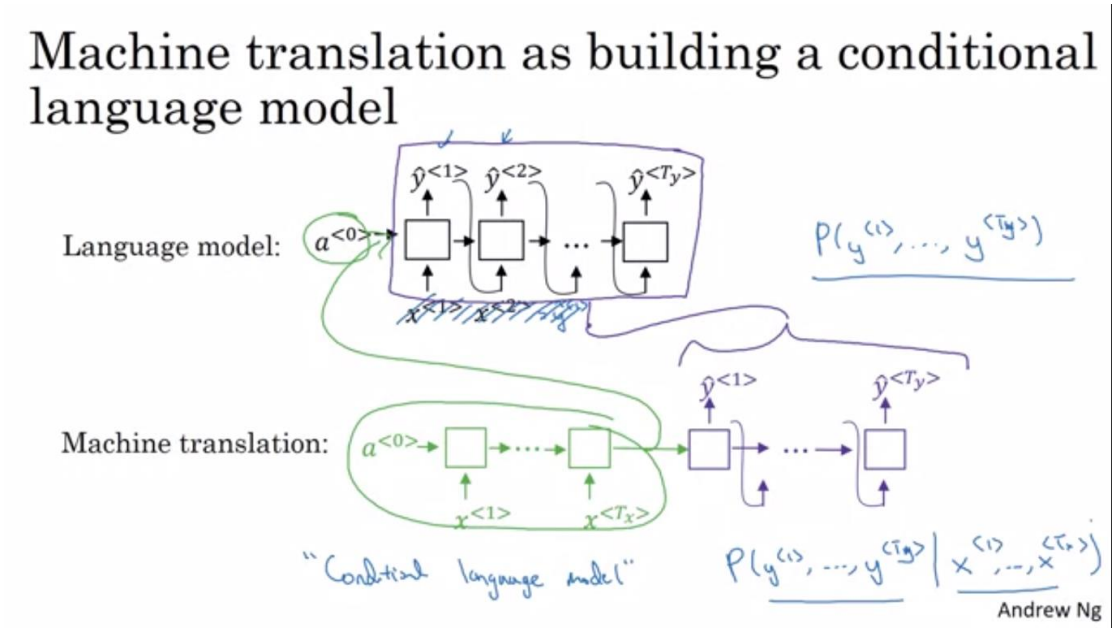
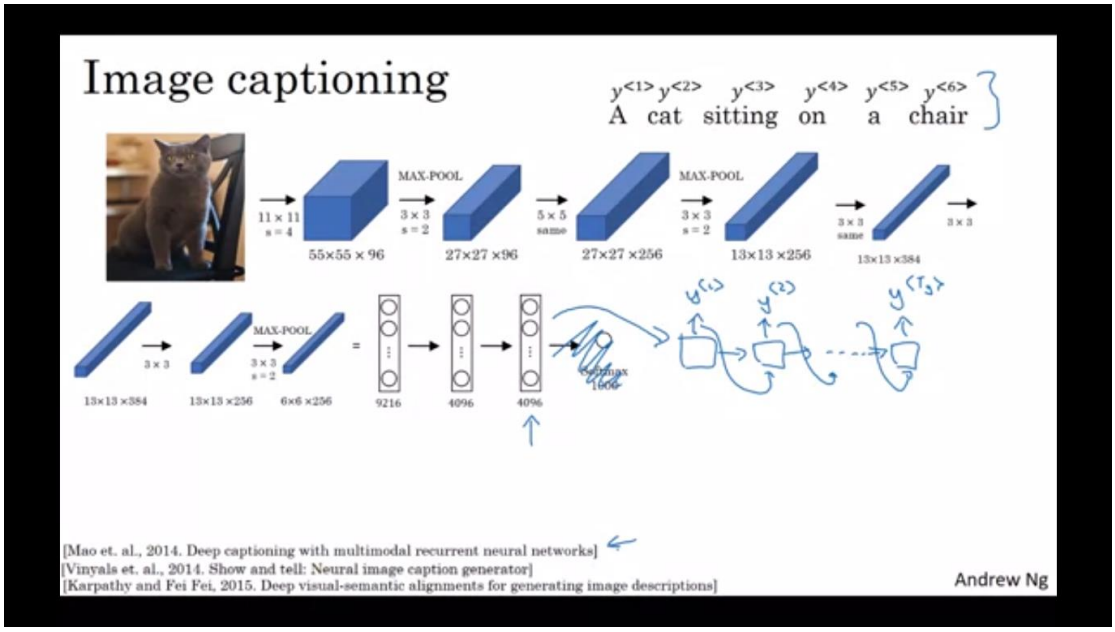
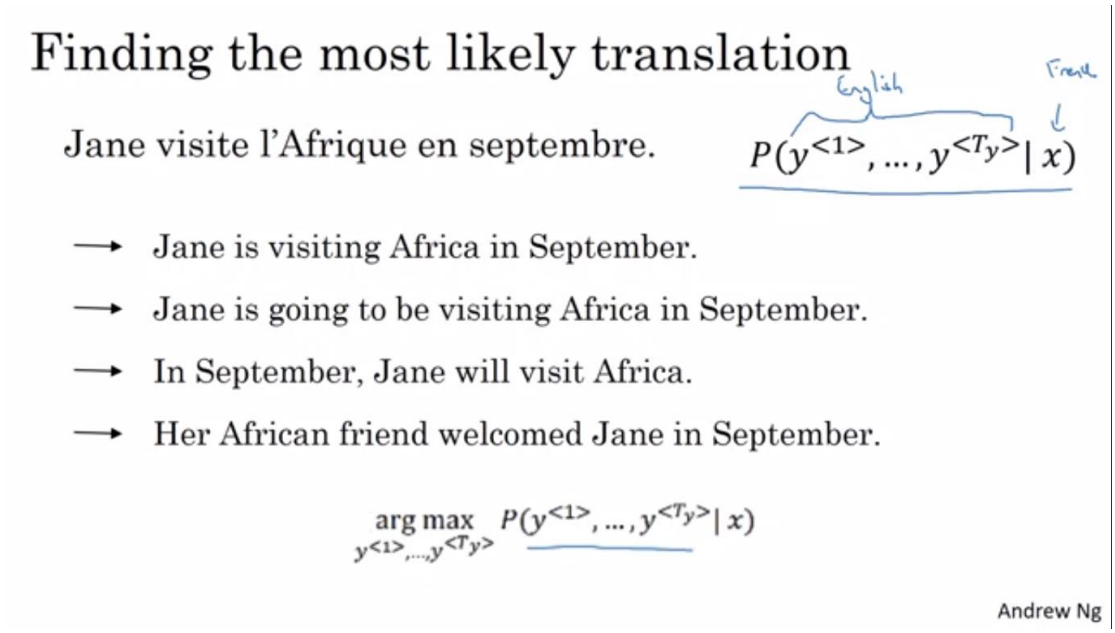


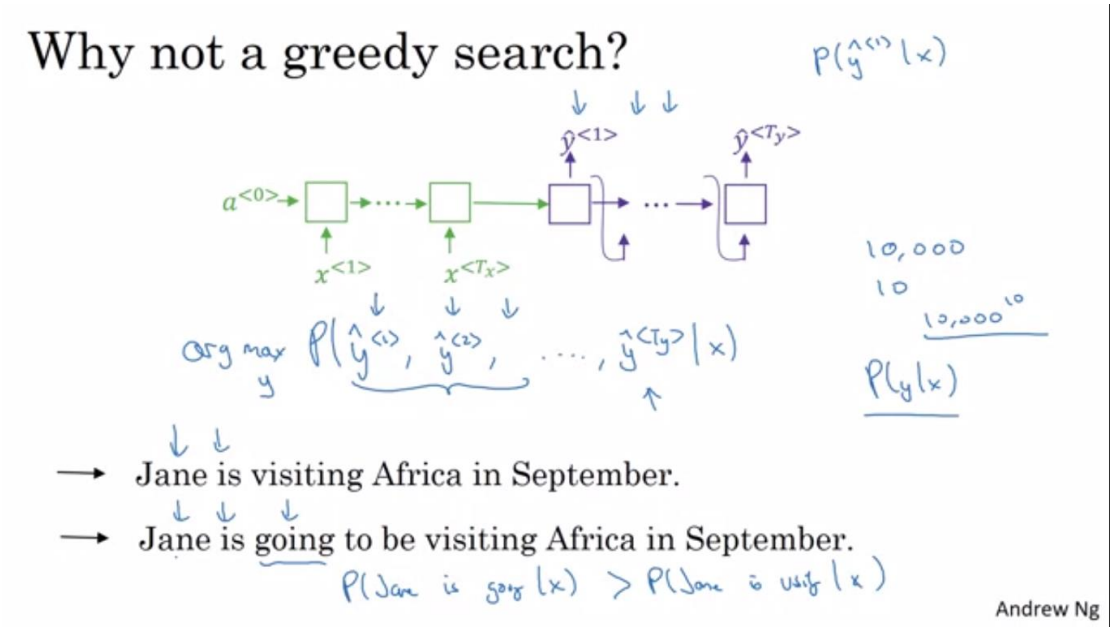
Image-to-Sequence is like making caption to an image, so we use image to generate sequence.



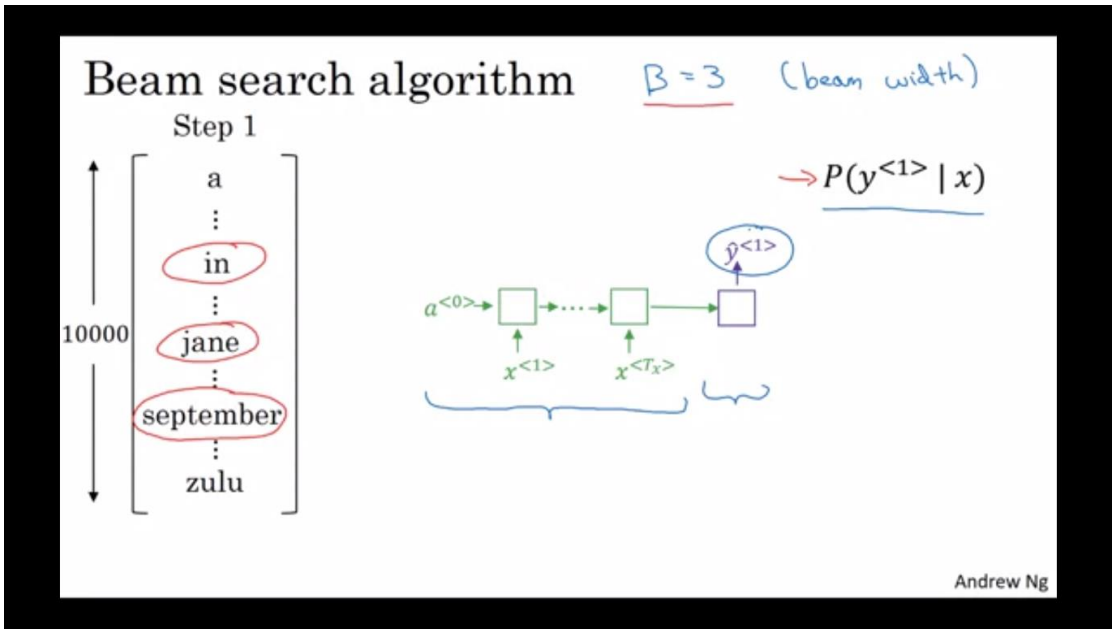
Beam Search is taking maximum of that conditional probability.

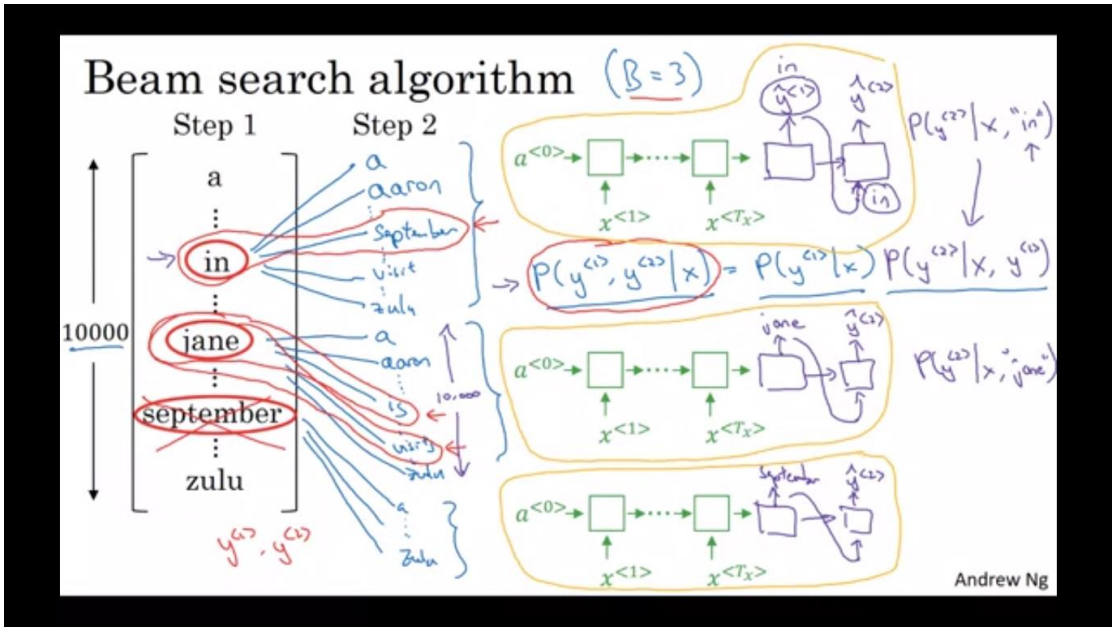


Rather than Greedy Search algorithm, where we just generating sequence without investigating whether it is the most suitable translation or not.



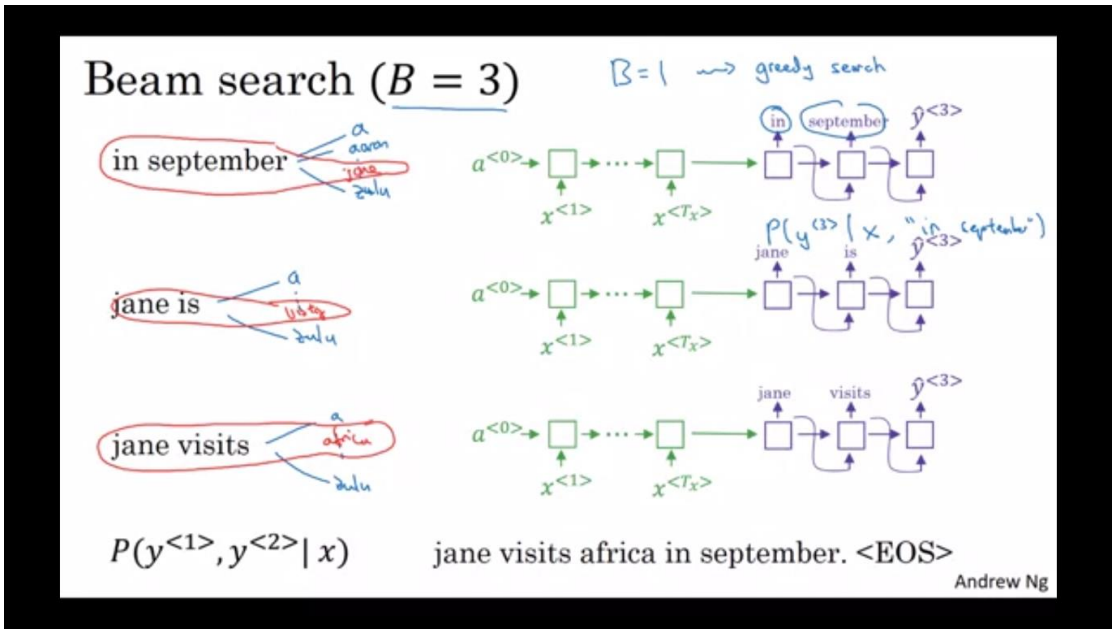
In Beam Search, we search the vocabulary vector for the number of words defined by parameter B (Beam Width) to find (from them) the most suitable words.



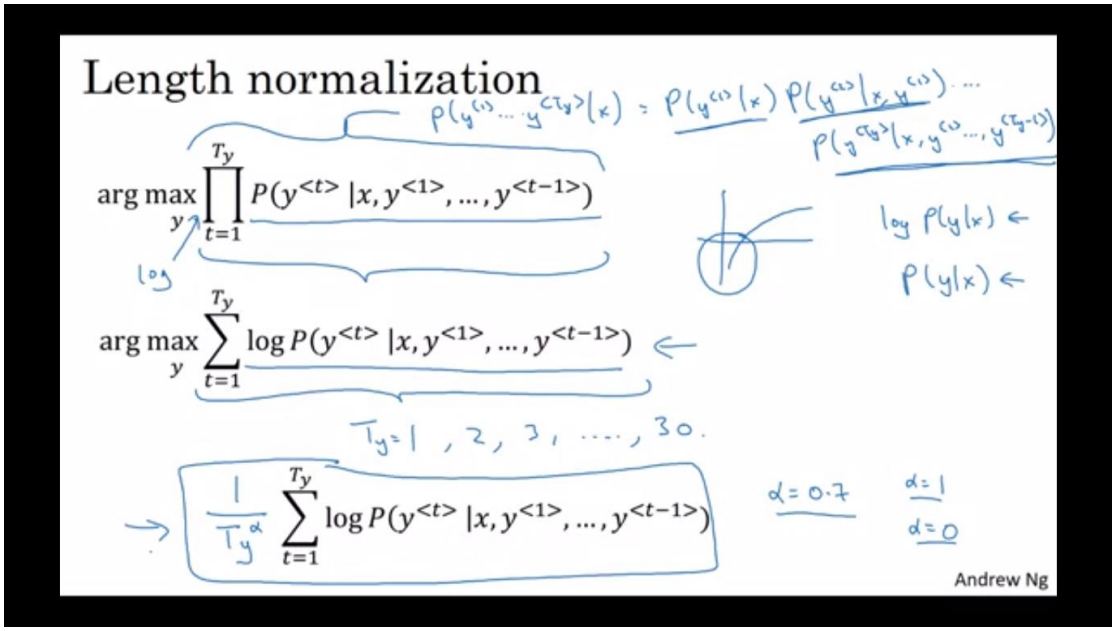


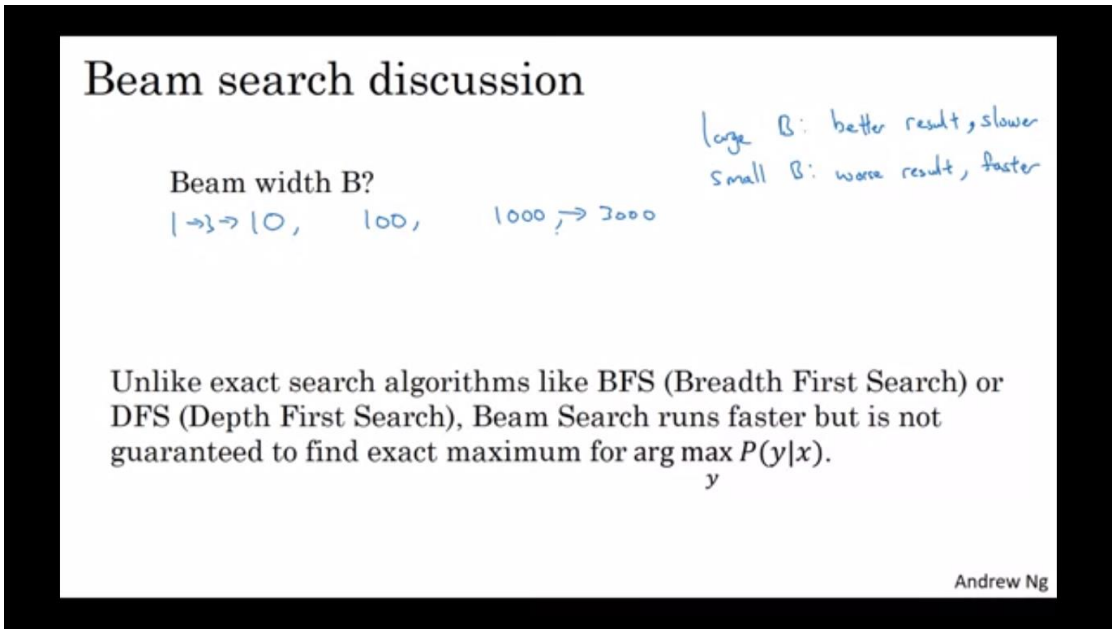
We need to use only B copies of the network to find the suitable words from (B x Vocabulary vector length) words.

Greedy search is Beam Search but with B=1.



To enhance the Beam Search, we use length normalization and log function. The more words, the more negative results in log output (Normalized log probability objective). We can make full length normalization or partially normalization depending on the application.





Detecting the error, whether it is in Beam Search algorithm or RNN Model, is based on comparing the conditional probabilities between human and machine translations. If human translation probability is higher than that of machine translation, the Beam Search is at fault.

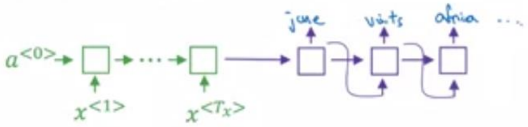
Example

Jane visite l'Afrique en septembre.

Human: Jane visits Africa in September.

Algorithm: Jane visited Africa last September.

RNN computes $P(y^*|x) > P(\hat{y}|x)$



Error analysis on beam search

Human: Jane visits Africa in September. (y^*)

Algorithm: Jane visited Africa last September. ($\hat{y}$)

Case 1: $P(y^*|x) > P(\hat{y}|x)$

Beam search chose $\hat{y}$. But y^* attains higher $P(y|x)$.

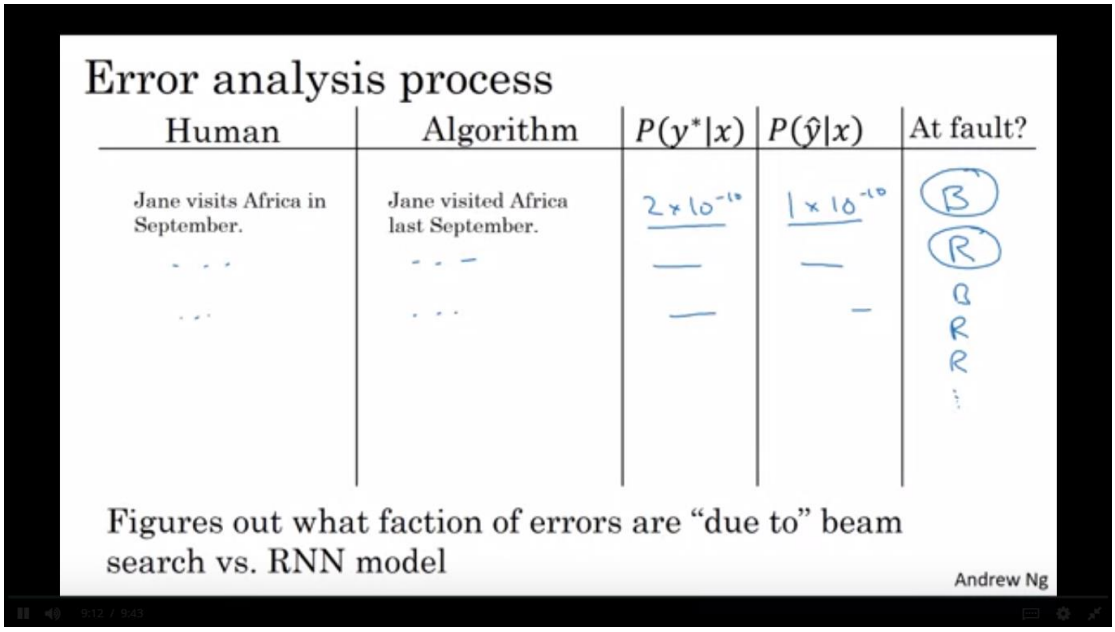
$\arg\max_y P(y|x)$

Conclusion: Beam search is at fault.

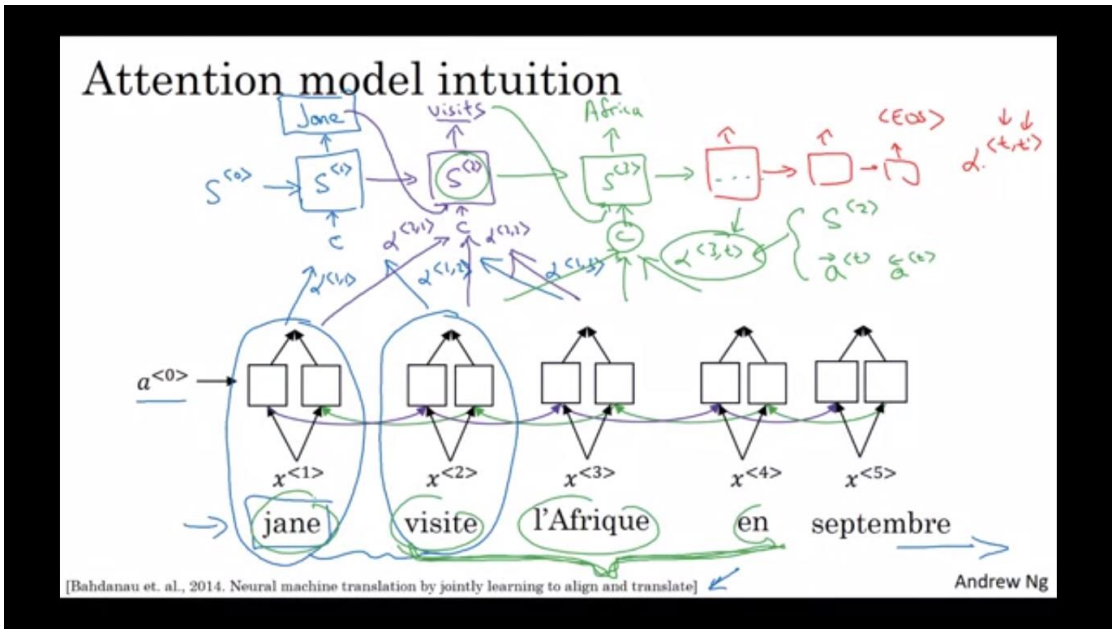
Case 2: $P(y^*|x) \leq P(\hat{y}|x)$

y^* is a better translation than $\hat{y}$. But RNN predicted $P(y^*|x) < P(\hat{y}|x)$.

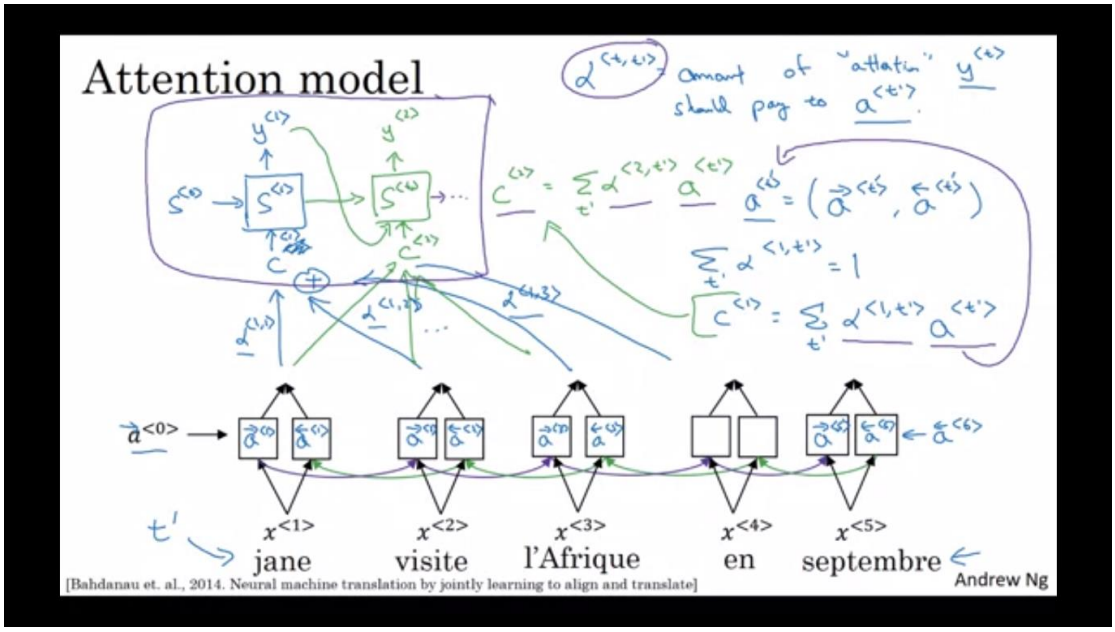
Conclusion: RNN model is at fault.

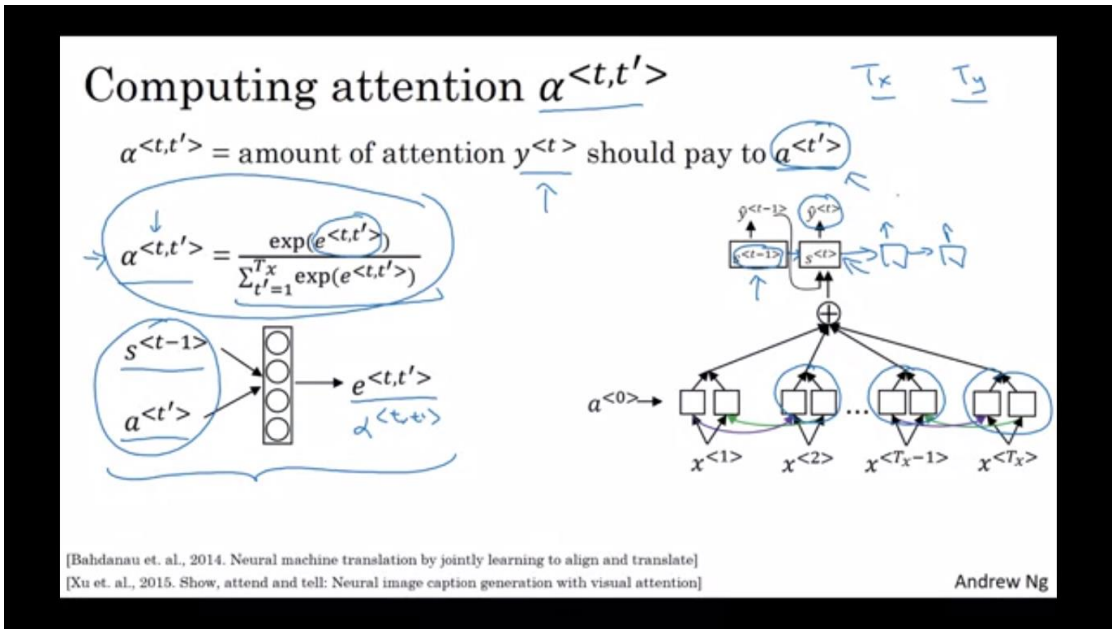


Attention Models: How much attention should we give for input sequence words near specific word to generate a relative specific word in output sequence.



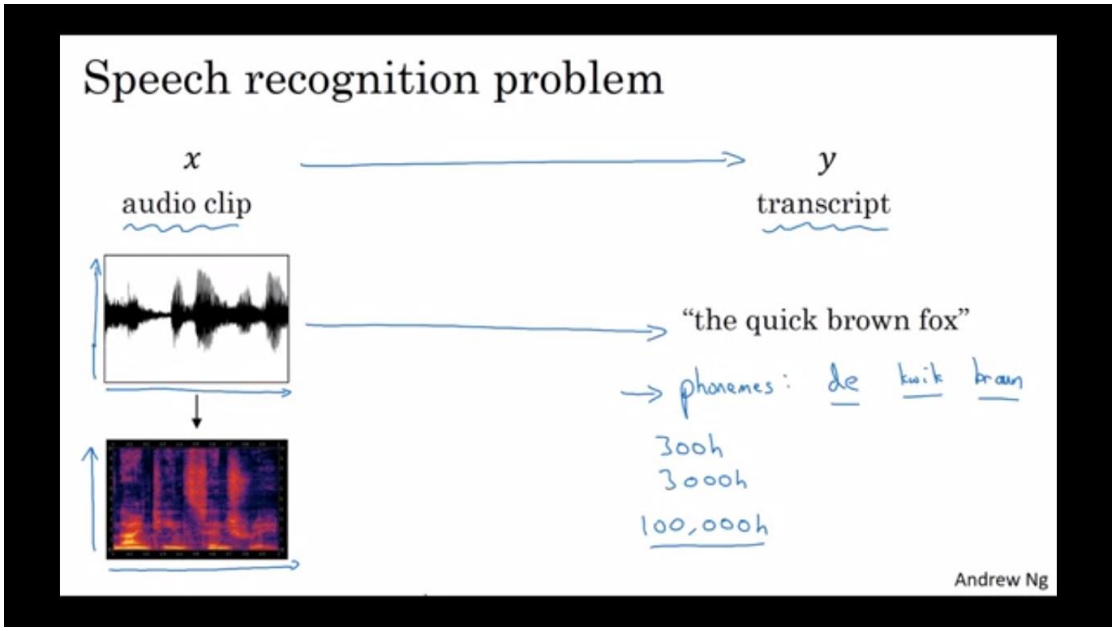
These attention weights are proportional to activation functions and state value of previous stage. Multiplying attention weights with activation functions generates context (c).

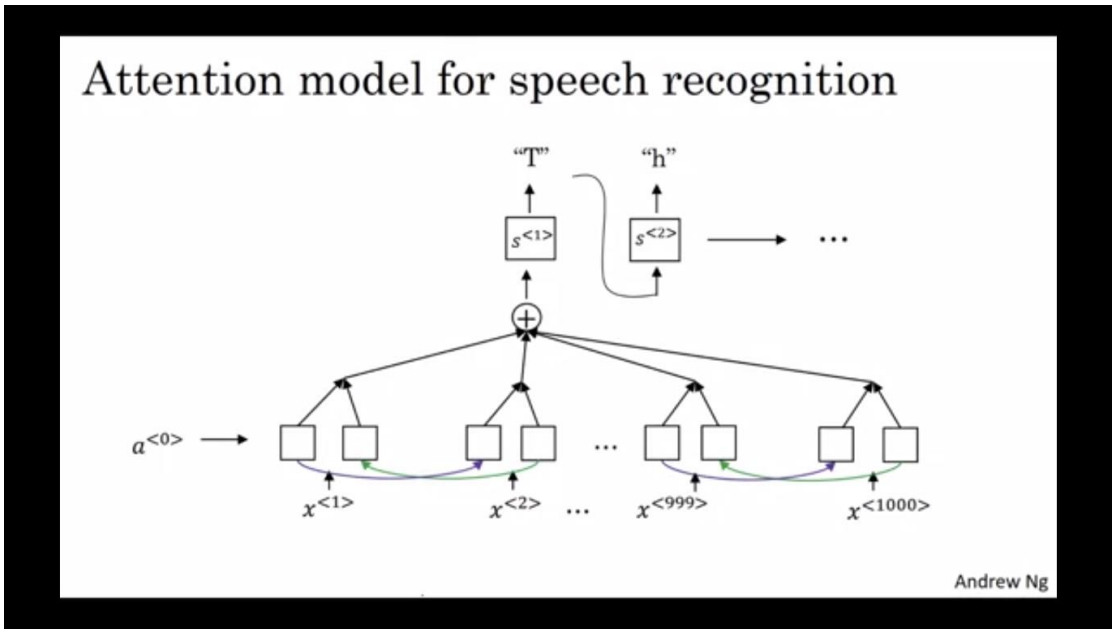




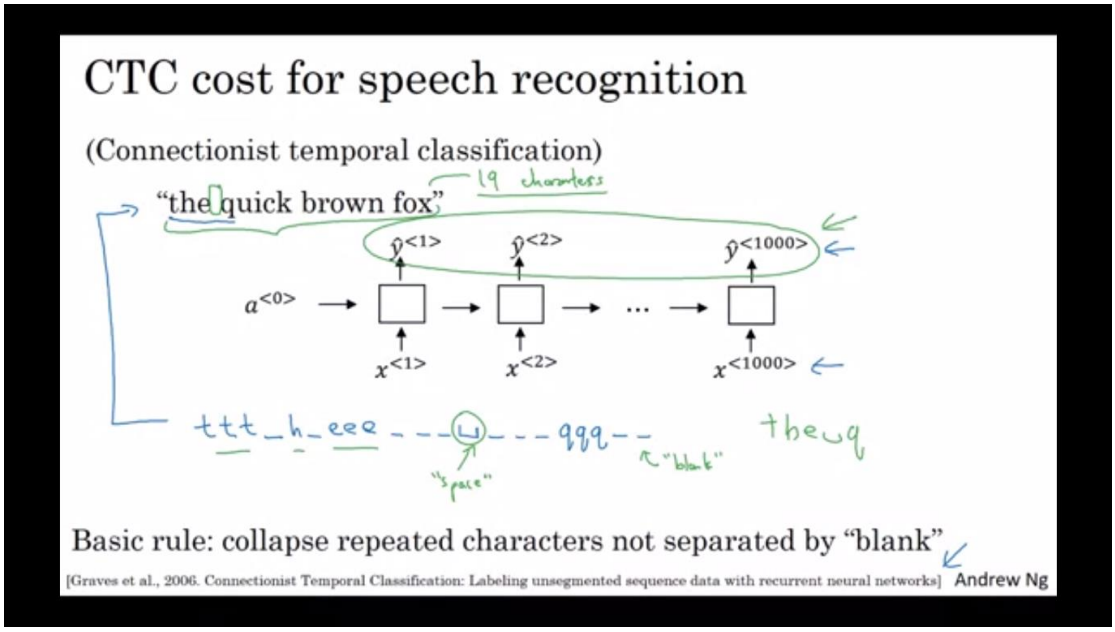
We use very small standard NN to compute the attention function by using gradient descent.

Speech Recognition:





We can use attention model or CTC (Connectionist Temporal Classification) model. In CTC, we collapse repeated characters not separated by blank after ensuring that output sequence is same length as input speech sequence.



Trigger word application: It is speech recognition application where we output 1 in output sequence if trigger word was detected.