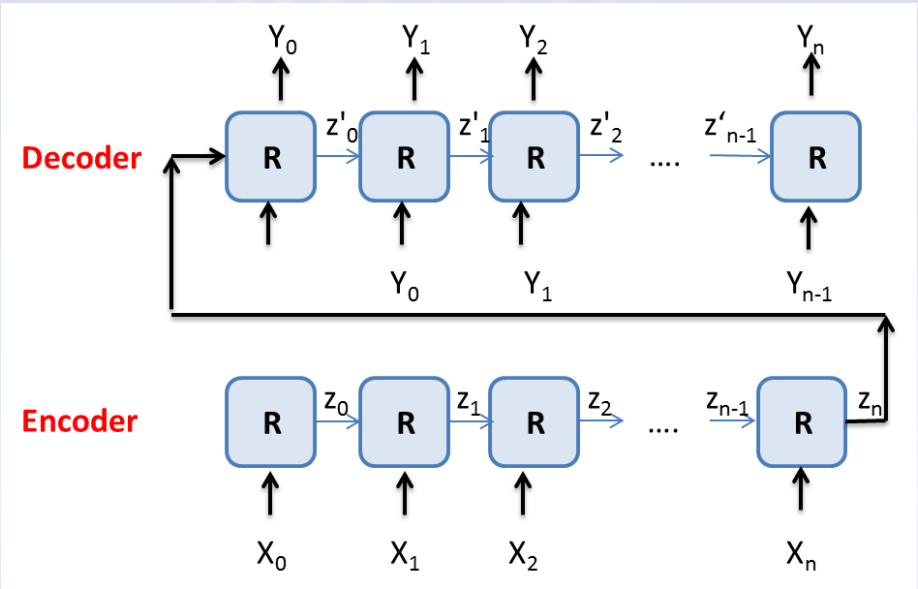

Let's look at language translation as an example (the model involves the use of two LSTMs). The first encodes the input language and outputs a fixed dimensional output z made from embeddings of the input. The final z is fed into another LSTM as its initial z' to output the predicted translated sentence. How does a decoder process inputs and generate outputs? Our decoder inputs are the target language inputs with a "start" token in front and an "end" token in back, followed by padding. The decoder unit takes as its initial input the final hidden state produced by the decoder unit.

(D) ENCODER-DECODER with attention:

Now let's take a look at encoder-decoder models with an attention unit. The difference is an additional input from the attention units received by the decoder, telling it how much attention to give to each of the encoder inputs when predicting the next output.