Machine Translation

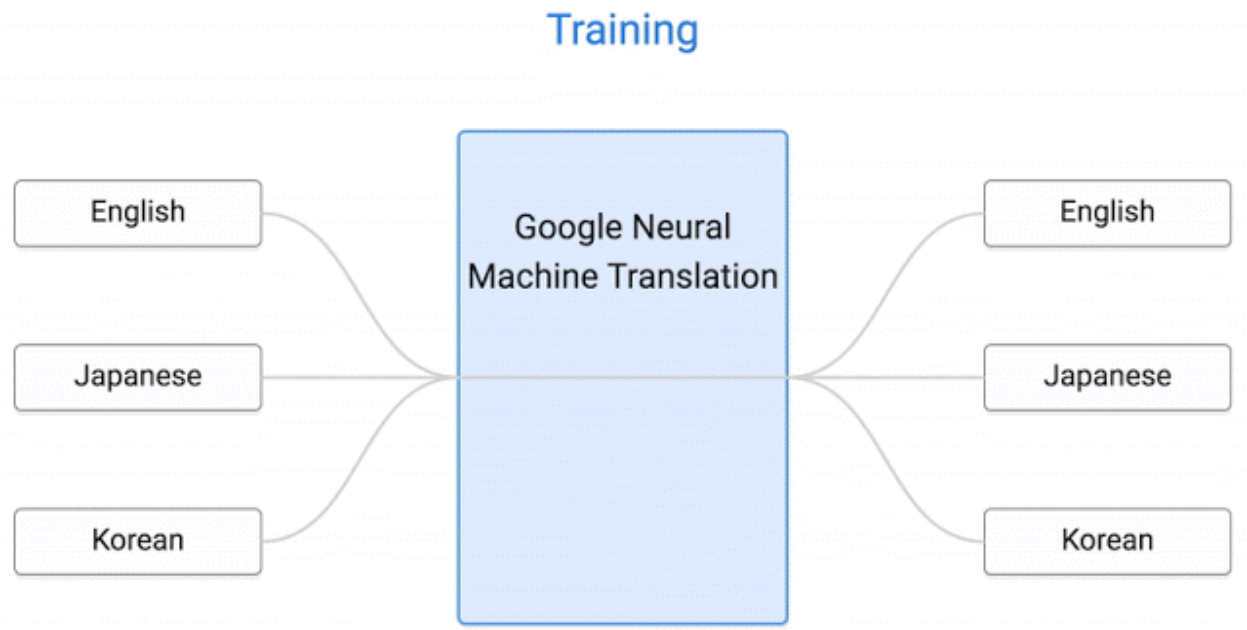
1. Introduction – What is does?

Years ago, it was very time consuming to translate the text from an unknown language. Using simple vocabularies with word-for-word translation was hard for two reasons:

- The reader had to know the grammar rules.
- Needed to keep in mind all language versions while translating the whole sentence.

Now, we don't need to struggle so much as we can translate phrases, sentences, and even large texts just by putting them in Google Translate.

The basic principles of machine translation engines

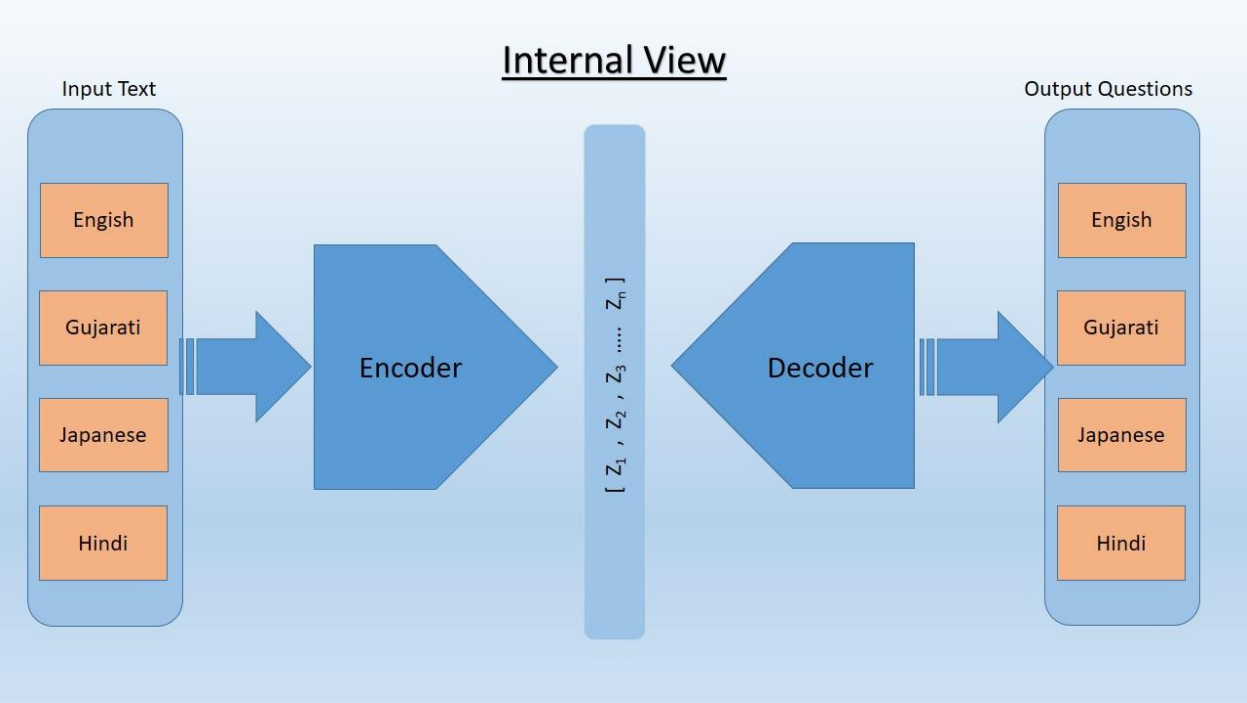


Google Machine Translation

Neural Machine Translation (NMT) is an end-to-end learning approach for automated translation, with the potential to overcome many of the weaknesses of conventional phrase-based translation systems. Deep neural networks can achieve excellent results in very complicated tasks (speech/visual object recognition), but despite their flexibility, they can be applied only for tasks where the input and target have fixed dimensionality. This issue is known as Deep Learning translation problem and it can be partially solved with the help of Recurrent Neural Network (RNN).

2. Algorithm Details – How it does?

In 2016, Google introduced the Encoder-Decoder neural network architecture for the language translation which outperforms all other language translation tools till this date. This architecture is also known as Sequence to Sequence models. Here is the overview of this architecture.



Basic overview of working of Encoder and Decoder Network

There are two parts of the sequence to sequence model. They are as follows:

2.1 Encoder

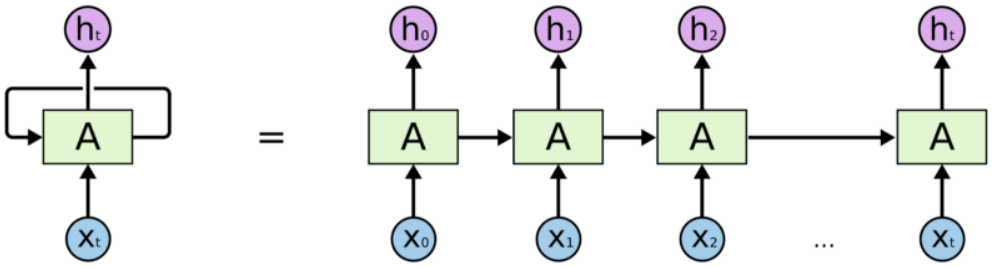
The encoder takes a preprocessed data from the input text and converts it according to the weights of the hidden layer. This hidden layer creates an intermediate representation of the input text and passes it to the decoder.

2.2 Decoder

The decoder takes this information from the hidden layer and converts it into another language. It produces the output in any desired language.

2.3 Long Short-Term Memory networks (LSTMs) and Recurrent Neural Network (RNN)

Here is where Long Short-Term Memory networks (LSTMs) come into play, helping us to work with sequences whose length we can't know a priori. LSTMs are a special kind of recurrent neural network (RNN), capable of learning long-term dependencies. All RNNs look like a chain of repeating modules.

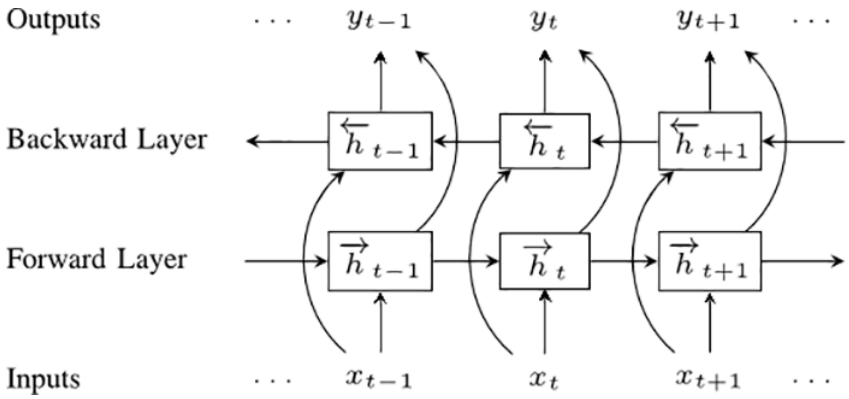


An unrolled recurrent neural network.

So the LSTM transmits data from module to module and, for example, for generating Ht we use not only Xt, but all previous input values X.

2.4 Bidirectional RNNs

Our next step is bidirectional recurrent neural networks (BRNNs). What a BRNN does, is split the neurons of a regular RNN into two directions. One direction is for positive time or forward states. The other direction is for negative time or backward states. The output of these two states is not connected to inputs of the opposite direction states.

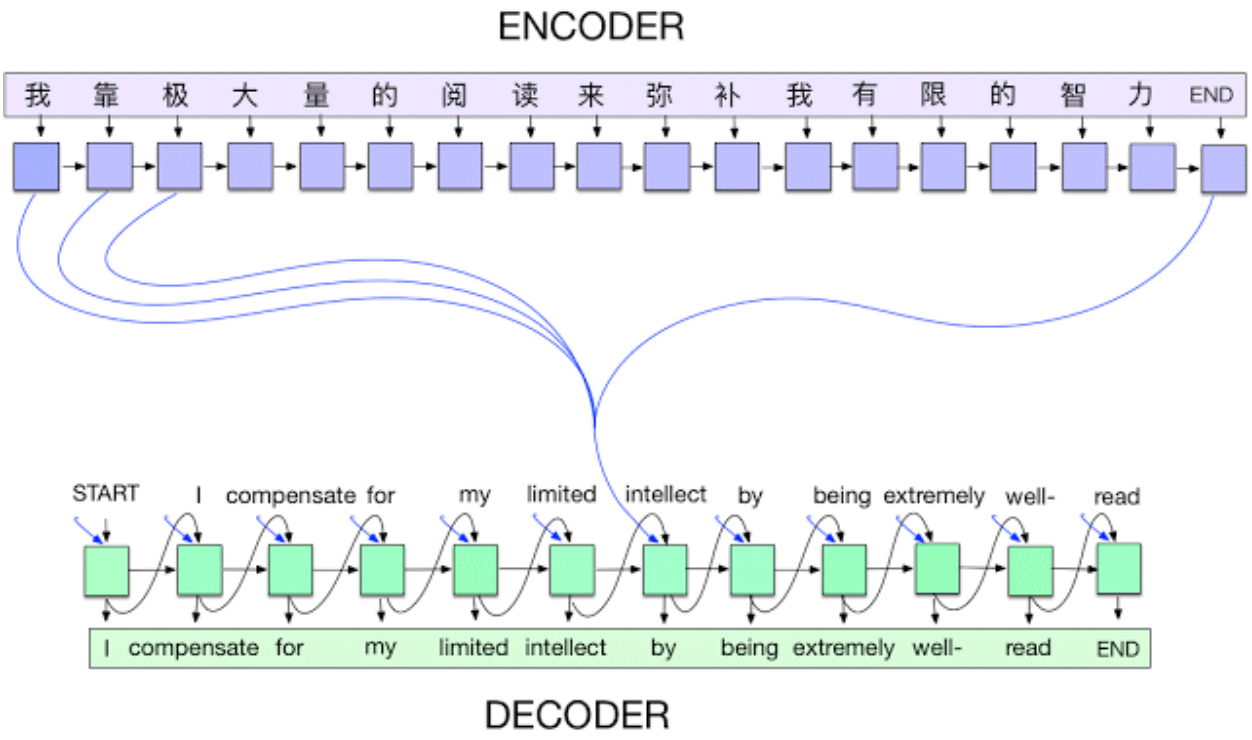


Bidirectional recurrent neural networks

To understand why BRNNs can work better than a simple RNN, imagine that we have a sentence of 9 words and we want to predict the 5th word. We can make it know either only the first 4 words or the first 4 words and last 4 words. Of course, the quality in the second case would be better.

2.5 Sequence to Sequence Model

Now we're ready to move to sequence to sequence models (also called seq2seq). The basic seq2seq model consists of two RNNs: an encoder network that processes the input and a decoder network that generates the output.



Sequence to sequence model

Hence as the output of decoder net, we get the translated information in targeted language. However, let's think about one trick. Google Translate currently supports 103 languages, so we should have 103x102 different models for each pair of languages. Of course, the quality of these models varies according to the popularity of languages and the amount of documents needed for training this network. The best that we can do is to make one NN to take any language as input and translate into any language.

2.6 Google Translation

That very idea was realized by Google at the end of 2016. The architecture of NN was built on the seq2seq model. The only exception is that between the encoder and decoder there are 8 layers of LSTM-RNN that have residual connections between layers with some tweaks for accuracy and speed. Moreover, this system requires a "token" at the beginning of the input sentence which specifies the language you're trying to translate the phrase into.

This improves translation quality and enables translations even between two languages which the system hasn't seen yet, a method termed "Zero-Shot Translation."

3. Stepwise algorithm

Machine Translation is performed in following steps:

1. Get and Load data set
2. Split training data and test data
3. Preprocessing
   - Take care of missing data
   - Convert string and character input to some numerical values
   - Feature Scaling
4. Build Encoder-Decoder (Seq2Seq) Model
5. Fit training data to this model
6. Evaluate the efficiency of the model on test data and fine tune model till acceptable accuracy is achieved.
7. Deploy model for actual use

4. Applications

Machine translation is used extensively at many places as it serves the as the bridge between different languages. Some of the application are as follows:

- Machine translation is used by Google for translating one language to another.
- Zomato (Food Delivery Company) uses machine translation to understand customer's feedback given in other languages.
- Machine translation can also be used in Question Generation problem which is entirely different NLP (Natural Language Processing) problem. Here we are interested in make questions based on the given content. In this problem, we can use input as one language and corresponding possible questions as another language and retrain the system with the help of transfer learning to get desired output.
- Machine translation can be used as a bridge between existing technologies in some language to some other language. Such as if Chatbot is made for English language interactions but it can now be extended to other languages too as simply convert other languages to English first then get a reply from the Chatbot and convert that reply again to the source language. Another example is IBM Watson, it is primarily made for English language only but now it can be extended to other languages too by the help of machine translation.
- The general public uses language translator for inter-language communication.

5. Links

Google translator: https://translate.google.com/

TensorFlow's version for Google's Neural Machine Translation:
https://github.com/tensorflow/nmt

Application that uses Machine Translation for Question Generation :
https://github.com/adityasarvaiya/Automatic_Question_Generation