MT(machine traslation) is a process when a computer software translate some text or sentence from one language to other language without human involvement. MT give huge benefit for human such as:

* Automated translation assistance: help a lot for professional human translators. Many translation management systems integrate one or more machine translation models into their work. When there are to many document to translate, they can run translations automatically, then send them to human translators for post-editing.
* Speed and volume: MT work faster, it can translate large amount of data in small time. Also, the quality of MT raises in recent year so that it very useful for everyone.
* Large language selection: While human translators can translate text in some specific language, MT support for 50-100 language or more nowadays.
* Cost-effective translation: it reduce time and the cost to translate language.

**How machine translate work?**

In MT, we have original text or language is called source language and the language you want to translate call target language. Most MT works by following a basic tow-step process:

1/ Decode the source language meaning of the original text

2/ Encode the meaning into the target language.

There are three types of machine translation methods:

1/ Rules-based machine translation (RbMT): it use grammar and language rules, developed by language experts and dictionaries which can be customized to specific topic or industry.

2/ Statistical machine translation(SMT): the first idea of this method come in 1949. It learn how to translate b analyzing large amount of existing human translations. Although it use effectively data and resources, the translation often superficial and don’t flow. It doesn’t work well with language pairs with different syntax so it need help from linguistic **(Nhà ngôn ngữ học)**.

3/ Neural machine translation(NMT): Starting from 2016, it become more and more popular as it provides better results with language pairs. By using many resources, database and translation suggestions approved by the translator, it teaches itself on how to translate by using a large neural network.

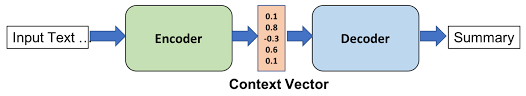
Neural Machine translation is the most accurate, versatile and fluent machine translation approach.

**Determined Machine translation quality:**

Every machine translate need high quality and accurate so that some test are been make to check the quality of MT methods. There are some test such as BLEU( Bilingual Evaluation Understudy), ROUGE, NIST and METEOR use to check the similarity between machine and human translations of the same text and BLEU is most often used.

**Encode and decode in Machine translation:**

First thing in Machine translation is that let machine understand the text. But because computer only know number, so that encode and decode appear. Basically, encode is transfer input text to list of number call vector. Then when we finish translate, we use decode to transfer vector to sentence output.



(<https://medium.com/nerd-for-tech/encoder-decoder-model-for-machine-translation-8a90be12ac32> )

**How does NMT work?**

NMT is a state of the art method that help to translate words from one language to another.

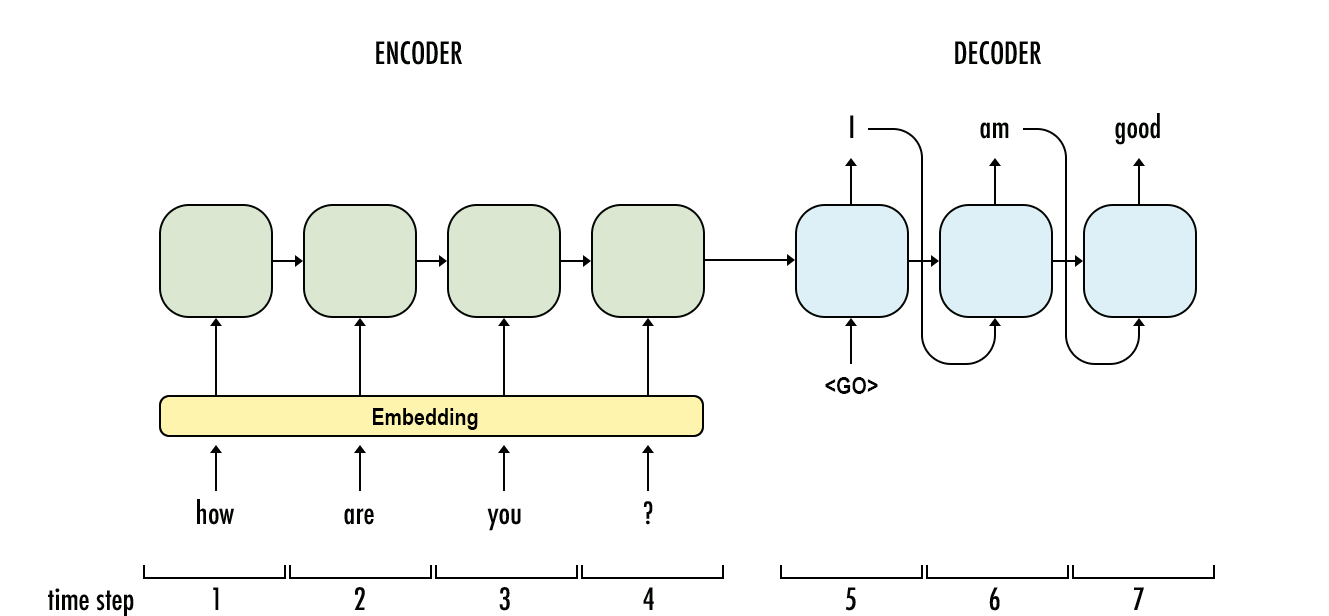
First of all, the input text is encoded into numbers. We put these numbers to neural translate model and result is output number. Then we will decode these into the language we want to translation.

Inside neural translate model is a complex mathematical formula and the output number are create and refined via training the network with sentence pairs. The result output will be find in the model with most accurately translating base on the input sentence.

**What is transformers?**

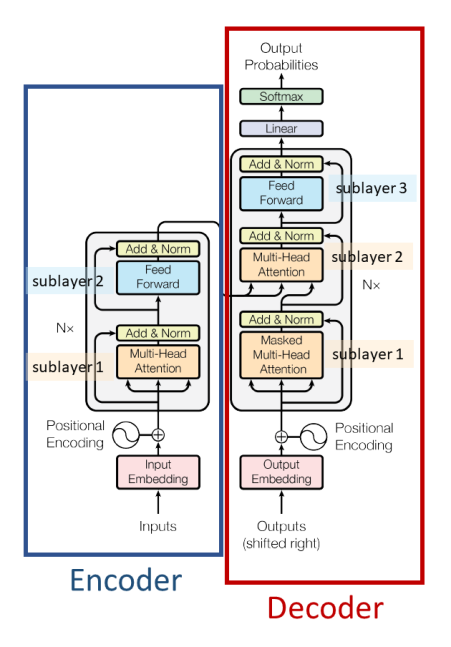
Before Transformers release, most of the neural machine translation are using Recurrent Neural Networks(RNNs). Beside the benefit of this model is that it can find the linkage and dependency constraints between the implementations in the sentence. The weakness of this method is that it difficult to catch the dependence between words in the sentence if the sentence to long. Also the training speed is slow due to sequential input processing. Transformers appear to solve these problem.

Some method using RNNs is Sequence-to-Sequence, it let the input and output as a sequence



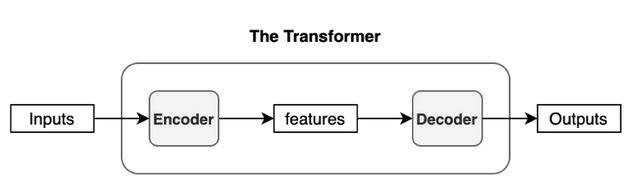
The reason why time when RNNs training slow is that it use only CPU, not use both CPU and GPU same time.

Another disadvantage of this method is that for long sentence it will lost information/attention between words or we can call Gradient Vanishing/Exploding phenomenon. Long-short Term Memory (LSTM) appear to fix this, but the training time slower than normal RNNs. Because of all that, Transformers appear as a solution. Transformers is leverage GPU parallelism to accelerate training.



**(** [**https://viblo.asia/p/transformers-nguoi-may-bien-hinh-bien-doi-the-gioi-nlp-924lJPOXKPM**](https://viblo.asia/p/transformers-nguoi-may-bien-hinh-bien-doi-the-gioi-nlp-924lJPOXKPM) **)**

Transformers first proposed by Vaswani and team in the 2017 paper title “Attention is all you need”. Transformers are among the newest and one of the most powerful classed of models invented to date. This is a deep neural networks that replace CNNs( and RNNs(Recurrent Neural Networks) with self-attention. Same as some machine translation model, transformer use encoder and decoder architecture.



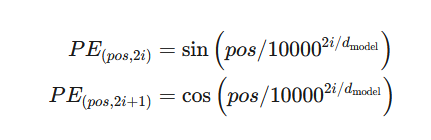
(<https://kikaben.com/transformers-encoder-decoder/> )

Encoder have multiple encoder block, the input sentence goes to encoder block, then at the end it generating a representation of it. The output of encoder go into decoder block and the output word is generated by running the stacked LSTM layers. A SoftMax activation function applies to the last layer.

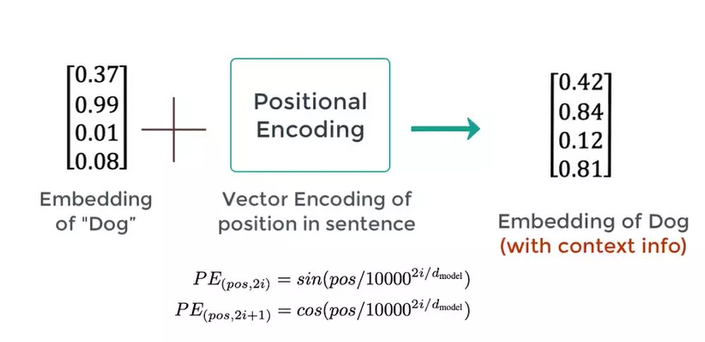
**Encoder:**

1/ Input embedding: to let computer understand words or sentences input, we must convert it to number as vector and we call it input embedding. With this step, all words with similar meanings will have vector close to each other.

2/ Positional encoding: Word that place different in sentence can have different meaning. So that transformer have positional encoding to inject information about where word in sentence.



Base on information, pos is position of word in sentence, PE is the value of I element in embedding of length dmodel . After that we add PE vector to Embedding.



(<https://viblo.asia/p/transformers-nguoi-may-bien-hinh-bien-doi-the-gioi-nlp-924lJPOXKPM> )

3/ Self-Attention: