Machine Translation:

Structure and Style of the Epitome of your Research

Fynnba Biney

Department of Computer Science and Engineering

Ashesi University

Berekusu, Arizona, USA

fynnba.biney@ashesi.edu.gh

Introduction

A major problem in Ghana is the lack of important documents, articles, written material and speeches in Ghanaian dialects such as Twi, Ga and Ewe. Having text in these language would help Ghanaians to understand things better as local dialects are usually most Ghanaians first language. Machine translation makes this possible. It is making a "machine" which is really a computing device, such as a computer, tablet, or smartphone that uses translation software in order to translate written or verbal texts from one language to another.

Previously and even to a large extent today, language is seen as an art that must be mastered through hundreds and thousands of hours of practice, but for some tasks basic tasks, there are some great benefits that can be derived from the use of machine translation. It is faster, costs less and can be used to translate multiple languages at a time. The best way to produce good results involves using a Phrase-Based Machine Translation (which is the most widely used) to learn input data then using an analyzer and generator for each language to translate to and from any language.

System Model and Solution

Machine translation is using a "machine" which is really a computing device, such as a computer, tablet, or smartphone that uses translation software in order to translate written or verbal texts from one language to another.

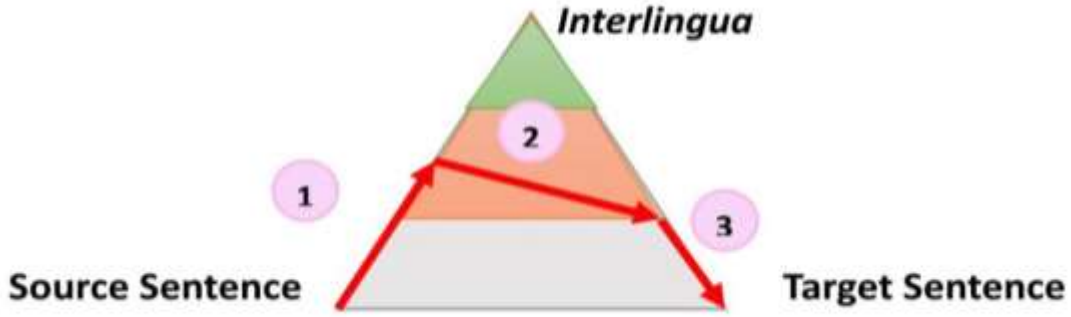
In a machine translation task, the input already consists of a sequence of symbols in some language, and the computer program must convert this into a sequence of symbols in another language. (Deep Learning, 2016)

In one of the most widely taught methods which is statistical machine translation, the following process is used. The initial step in text classification and analysis is pre-processing. Various techniques are applied to the text corpus in order to reduce the noise of text, reduce dimensionality, and assist in the improvement of classification effectiveness:

- Removing newline characters
- Removing numbers
- Stemming
- Part of speech tagging
- Remove punctuation
- Lowercase
- Remove stopwords

After cleaning the text, three technologies are used on the data in different steps of each translation process and the resources that each technology uses to translate. Then we will take a look at a few examples and compare what each technology must do to translate them correctly.

A useful representation of an automatic translation process is the following triangle which was introduced by French Researcher B. Vauquois in 1968.



The triangle represents the process of transforming a source sentence into the target sentence using three different steps. The left side of the triangle symbolizes the source language; the right side the target language. The different levels (the green, red and grey areas) inside the triangle represent the depth of the analysis of the source sentence. These layers could represent syntactic analysis, semantic analysis, and many others.

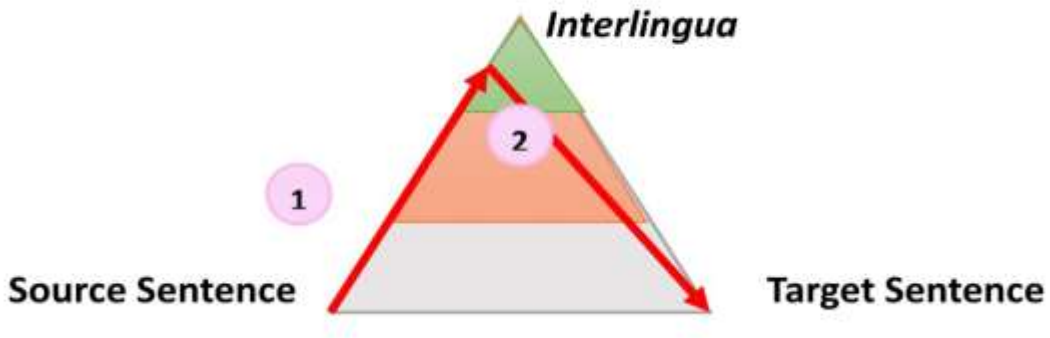
The first red arrow represents the analysis of the sentence in the source language which in this case is either English or one of the Ghanaian languages. From the actual sentence (which is a sequence or a long string of words) the program will build an internal representation corresponding to how deep we can analyze the sentence.

For instance, on one level we can determine the parts of speech of each word (noun, verb, etc.) which could involve Part of Speech Tagging, and on another we can connect words: for instance, which noun phrase is the subject of which verb. When the first stage, which is the analysis stage, the sentence is moves into the second process, a representation of equal or slightly less depth in the target language. Then, a third process called “generation” generates the most accurate target sentence from the computer generated internal representation (which is the sequence of words in the target language).

The reason for using a triangle is, the higher or deeper the source is analyzed the, the smaller or simpler the transfer phase.

Neural Machine Translation

The neural machine translation approach is can also be represented by using the Vauquois Trian



With the following platforms:

- The “analysis” is actually **encoding** and the result of the analysis is a complex sequence of vectors
- The “transfer” is **decoding** and directly generates the target form without any generation phase.

The neural network performs the encoding in its hidden layers, which are a part of the newtwork that cannot really be seen or fully understood.

This is actually a building block of the technology, and as is the case in a rule-based system where each word is first looked up in a monolingual dictionary, the first step of the encoder is to look up each source word in a word embedding table.

Part of how meanings are represented in neural machine translation are in the word embedding. The underlying idea is to map out the words in numerical representation (which are vectors) using a three dimensional graph. Words with the relation to one another are close to each other. ie, the vectorized for of dog and cat may be close to each other on the graph as they are all animals and are popular pets. Words with some common property will be near on one dimension of this space.

It is not exactly clear how word embedding are constituted, but the neural network is able to work with the vectors in the embedding.

In the second step, target words are generated. They are generated using:

- The "Target Context" generated together with the previous word, and which represents some information about the status of the translation;
- A weighted "Source Context" which is a mix of the different source contexts by a specific model called Attention Model – we will discuss Attention Models further in another article. Essentially, Attention Models select the source word to translate at any step of the process;
- The previously translated word using a word embedding to convert the actual word into a vector that the decoder can actually handle.

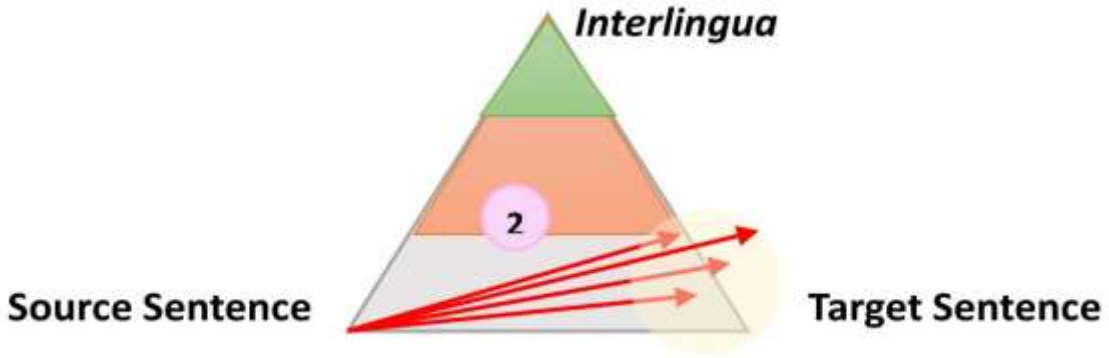
This was from the 2014 "On the properties of neural machine translation: Encoder-decoder approaches" by Cho, K., Van Merriënboer, B., Bahdanau, D., and Bengio.

The translation will end when the decoder generates an end-of-sentence special word. The complete process is enigmatic, but it is very possible to implement a model which can so this using Ghanaian local Languages.

The translation process of a Neural Machine Translation engine follows the same sequence of operations as a rule-based engine, however, the nature of the operations and the objects manipulated are different as the network requires a the words to be converted into vectors to be used in word embedding

Phrase-based machine translation

Phrase-based machine translation does not stictly follow the process defined by Vauquois. There no analysis or generation. Instead, the engine will generate multiple translations for one source sentence, and then selects the best one.



For this the model is based on 3 main resources:

- A table which contains translation option and their probabilities of the occurrence of phrases in the source language
- A table showing how words can be reordered when transferred from source language to target language
- A language model which gives probability for each possible word sequence in the target language. Discriminative models work well here.

To make this model efficient, usually, smart probability calculations and smarter search algorithms are used so that only the most likely translation will be scrutinized and the best one ot of them is chosen.

Conclusions

In conclusion, the best model to be used for Machine Translation in the Ghanaian context could be a Phrase-based machine translation model as it may be to get the data for implement the model which are two tables, one containing translation option and their probabilities of the occurrence and the other, showing how words can be reordered when transferred from source language to target language.

References

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1). Cambridge: MIT press.

Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Vauquois, B. (1968). A survey of formal grammars and algorithms for recognition and transformation in machine translation', ifip congress-68, edinburgh, 254-260; reprinted in ch. *Bernard Vauquois et la TAO: Vingt-cinq Ans de Traduction Automatique-Analectes*, 201-213.