Title: Development of an English to Twi Machine Translation System

Author: Emmanuel Nimo
Supervisor: Dennis Owusu, Ph.D

INTRODUCTION

In this contemporary era, where technological advancement is on the increase, it has become essential that machines such as computers behave more like humans, and one of the requirements to do so is communication. Natural language and processing is an interesting topic that fortifies the backbone of machine's natural communication; enabling computers to understand and process human languages. For a country like Ghana, which was once ruled by the British, it is of much essence to develop a digital assistant language translator for converting the English language to Twi. This paper seeks to demonstrate how translation between both languages can be done using machine language translation.

Machine translation is an automated translation process which allows computer software to translate a text from one natural language to another. It is one of the early applications of Natural language processing that is very much researched [1]. A good example of an application that use Machine translation is Google Translate. Google Translator translates languages a little over 100. To be able to perform this, it requires a source or original language and a target, thus, the translated language. It incorporates interpretation and analysis of elements in the text and compare their influences and relations. It also delves deeper into use of grammar, semantics, syntax, morphology, etc. and the cultural context of the regions of the two languages to match the source language to target language.

1. BACKGROUND.

In Ghana and beyond, Twi is popularly known as the dialect of the Akan languages, one that is largely spoken by the people in the southern and central part of Ghana. It traces its origin to Ashanti and Akuapem, and due to historical events, such as wars, trading and building of kingdoms, it made Twi a dominant language in Ghana. It is possible that one is likely to find that in every 2 out of 3 Ghanaians speak or understand Twi. Twi is easy to learn and speak due to its phonology (consonants and vowels) and its interdependencies with other languages.

English, a language originated from England, now an official language of many countries such as the USA, Ghana, Nigeria, etc. [2]. English is the main medium of communication in Ghanaian schools and has even become a requirement in occupations and professions that interfaces with other international bodies. After British colonialism, the English language became the national language and continues to have a deep impact on the Ghanaian society and will be an important issue in the shaping of its future [3].

2. METHODOLOGY

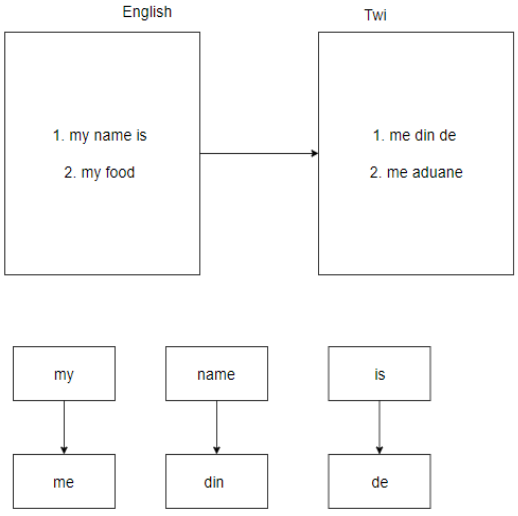
In this paper, the Corpus-based approach with machine neural network, and Statistical machine translation will be considered to show how conversion between the English text and Twi text is done.

i. Corpus Based Approach

This approach requires the alignment of two available parallel corpora, thus, English corpus (source language) and

Twi corpus (target language). According to Koehn (2010), parallel corpus is a collection of test, paired with translations into another language (p. 54) [5]. In this area, sentence alignment is performed to make the corpus useful or the type of statistical machine translation model.

Since there is no current corpus for Twi, we can create one with most commonly spoken phrases and sentences from books. The corpora become the dataset. The dataset is then prepared using tokenization, normalization, sentence fragment to check for punctuation, lowercase, uppercase, special characters. The result is then trained by matching the fragments (words or phrases) of the sentences in the source language against that of the target language. The translated fragments are then put together appropriately using grammatical and syntactic rules. During training, neural translation model is applied to the dataset to automatically aligns words or phrases with the sentence pair in a parallel corpus (word alignment and phrase extraction).



Given Twi sentences $(t_1, .. t_{nt})$ *to be converted to english* $(e_1, ... e_{ne})$. The sentence alignment will consist of list of sentence pairs $(S_1, ... S_n)$.

$$s_i = (\{t_{start-t(i),.....,}t_{end-t(i)}\}\ ,\{e_{start-e(i),.....,}e_{end-e(i)}\})$$

This implies that:

$$start - t(i) = end - t(i-1) + 1$$

$$start - e(i) = end - e(i-1) + 1$$

Although this approach is time consuming task in terms of using the corpora for both languages, it has the advantage to speedily find out the sense of words and phrases in the two languages [4].

ii. Statistical Machine Translation

With the already prepared parallel aligned bilingual text corpora (Twi and English), the approach gives the best translation of sentence using the posterior probability of a search term. Thus, the words which have the highest probability will give the best translation. The method inculcates the use of Bayes rule and it is maximized to find the probability. This approach will require the language model for both English and Twi, a translation model and a search algorithm. To find the most probable English sentence (e) given a Twi sentence (t), compute:

$$P(e|t)$$

$$P(e|t) = \frac{p(t|e)(e)}{p(t)}$$

$$\dot{c} = argmax_e p(e|t)$$

$$\dot{c} = argmax_e \frac{p(t|e)(e)}{p(t)}$$

$$\dot{c} = argmax_e\, p(t|e)(e)$$

3. Evaluation

The translator when developed will be evaluated by humans since it is very effective, but expensive. They will rate the translator based on factors such as fluency (clarity and naturalness), and fidelity (adequacy).

4. Limitation.

The major challenges in statistical translations are the decoding complexity and target language reordering [6]. Because human languages are full of special cases due to region variations between Ghana and Britain, there is likely characters that are hard to translate and therefore it will require lemmatization.

REFERENCES

[1] Dechelotte, D. et al. "A state-of-the-art Statistical Machine Translation System based on Moses". Semantic scholar. 2006

[2] S. Potters and D. Crystal (2018). "English Language". Encyclopedi Briticanna.

[3] L. Morris (1998). "The function of English in contemporary Ghanaian society." African Dispora Collection.

[4] S. Nahar, M. Nurul and Md. Nur-E-Arefin. "Evaluation of Machine Translation Approaches to Translate English to Bengali." IEEE. December 2017.

[5] Koehn, P. "Words, Sentences, Corpora". Statistical Machine Translation. 2010. Cambridge University Press: UK.

[6] S. Sreelekha. Statistical Vs Rule Based Machine Translation; A Case Study on Indian Language Perspective. 2017. Indian Institute of Technology Bombay: India