AN APPROACH TO DEVELOPING A NATURAL LANGUAGE PROCESSING SYSTEM FOR THE TWI GHANAIAN LOCAL LANGUAGE – BRIDGING THE GAP OF LOW-RESOURCE LANGUAGES IN MACHINE TRANSLATION

BY: DAVID SASU

Abstract

Language is the primal form of communication among human beings. In recent times, this characteristic of language, which had solely been possessed by mankind for centuries, is being shared with machines through the advent of Artificial Intelligence. The branch of Artificial Intelligence which is responsible for this computing breakthrough is known as Natural Language Processing. Natural Language Processing mainly deals with utilizing statistical and mathematical models to train machines to understand human language. The technologies which have been developed using Natural Language Processing paradigms are among the most popular form of technologies in the world today. Even though Natural Language Processing has gained a lot of traction and influence in the current world, it has a major underlining flaw. This major underlining flaw is that many of the algorithms that are used in Natural Language Processing cannot be directly applied to "low-resource languages". This is because these languages do not provide the massive amounts of readily available data that the conventional Natural Language Processing algorithms need to feed on in order to produce efficient and effective models. Unfortunately, majority of the Ghanaian languages, including the Twi language, are low-resource languages. Since the Twi language is widely spoken within the country, this research would explore whether Twi language data obtained from unconventional sources, such as the bilingual Twi-English bible, can be used to create a Natural Language Processing System such as a Machine Translation System.

Keywords:

Natural Language Processing; bias; low-resource languages; artificial intelligence; machine translation

1 - Introduction and Background

Language is the primary tool that enables human beings to communicate. It empowers mankind to not only transmit information, but it also allows him to express his ideas and vivid imaginations. It is the common denominator that binds people together and preserves their history, culture and knowledge. It is an extremely influential and powerful tool and depending upon how it is used, it can shape and define reality.

Language constitutes of all the different sounds, gestures and symbols that can be used in the communication of information. There are currently 7,097 languages in the world and each of these languages fall under one of the six major language families, which are: Afro-Asiatic, Austronesian, Indo-European, Niger-Congo, Sino-Tibetan and Trans-New Guinea [1]. Even though language can be expressed in different forms, it is mainly either spoken or written. During the process of writing, the letters, words or symbols that embody linguistic expression within a language are put together to represent and communicate ideas.

For many centuries, the art of representing language in the form of text had been regarded as something that only human beings could do and before the advent of modern technologies, it was impossible for machines to understand the intricacies and concepts of human language. In current times, machines now possess the ability to not only understand human language but also generate intelligent expression of human language through writing.

The ability of a language to be represented in written form has facilitated the process of language translation.

Language translation can be formally defined as the process of converting the written word from one language into another in a way that is culturally and linguistically appropriate so it can be understood by its intended audience [2]. When machines, instead of human beings, are used to convert written words from one language to another, it is known as Machine Translation. Machine Translation is made possible through the use of technologies provided by Artificial Intelligence.

According to 'Artificial Intelligence – A Modern Approach', a book which was written by two Computer Scientists, Peter Norwig and Stuart Russell, 'Artificial Intelligence' is the branch of Computer Science that is concerned with the automation of intelligent behaviour [3]. Artificial Intelligence encompasses a wide range of computational fields such as: Robotics, Vision systems, Learning systems, Natural Language Processing, Neural Networks and Expert systems.

The branch of Artificial Intelligence which enables machines to understand and translate human language is Natural Language Processing. 'Natural Language Processing' can be defined as "the automatic or semi-automatic processing of human language" [4]. It involves training machines to understand how human beings learn and use natural language so that technological tools and systems can be built to utilize natural language to perform desired tasks. Technologies that have been created using the paradigms of Natural Language Processing have gained a lot of traction in current times because of their incredible intuitive feel and efficiency. Typical examples of popular Natural Language technologies are Siri, Cortana and Alexa.

To ensure high efficiency, systems that are created using Natural Language Processing usually require a lot of human language data for training and learning. Immense language data of certain languages such as English, French and German have already been recorded and made available through easily accessible databases. As a result of this, the development process of Natural Language Systems for those languages is made easier since the researchers would only have to focus on the development of the system and would not have to worry about where to obtain the language data.

However, the development process of Natural Language Systems for languages that do not have large quantities of readily available language data is far more challenging and tiresome. This is because researchers would usually have to spend a lot of time and money to gather the language data of such languages. These types of languages are referred to as 'Low-resource Languages'. A 'Low-resource Language' can be formally defined as a 'language that has very limited annotated resources' [5].

Most native Ghanaian Languages can be classified as low-resource languages because they have either very little or no annotated resources that can be used in the development of Natural Language Systems such as Machine Translation Systems. This poses as a huge problem because it implies that the benefits that come from the creation and utilization of Natural Language Systems would not be made available to the illiterate Ghanaian population who can only communicate in their native language.

Twi is the language of the Akan ethnic group and it is widely spoken within Ghana. According to the research conducted by the Rutgers University School of Arts and Sciences, approximately 44% of Ghanaians speak Twi as their first language and 80% of Ghanaians speak Twi as their first and second language [6]. However, even though Twi is widely spoken in Ghana, it is classified as a low-resource language and no Natural Language technologies have been built for it yet. Therefore, this thesis work is going to explore the question of whether we can mitigate the bias against low-resource languages in Natural Language Processing technologies by developing a Machine Translation System for translating between the Twi local language and English language with limited language data.

1.1 Research question

The development of natural language processing systems for low-resource languages is a very active field of research in the world today. Some researchers are trying to solve this problem of the lack of representation of low-resource languages in Natural Language Processing systems by exploring better supervised or unsupervised machine learning algorithms that can efficiently learn from very little language data. Other researchers are trying to solve this problem by exploring better data sources from which natural language data for these low-resource languages can be obtained. Since this research is centred around the development of a Machine Translation System for the Twi local language, the research question that would be investigated in this work is: "Can we mitigate bias against low-resource languages in Natural Language Processing technologies by developing a Machine Translation System for translating between the Twi local language and English language with limited language data obtained from unconventional sources?"

2 - Related Literature

2.1 Low-Resource Languages

The biggest problem with low-resource languages is that the resources that are needed for these languages are extremely difficult to attain. Much of the language information and description of these languages is either unpublished, exists in only paper format or simply does not exist. In most scenarios, even when the language information of these languages exists in electronic format, they cannot be used simply because they are represented in a way that makes it either extremely difficult or impossible to use. As a result of all this, even raw text in a low resource language can be difficult to obtain and use [7].

Another problem with low-resource languages is that the orthographies for these languages may not be standardized. In such languages, the word boundaries, spellings of words and even the usage of the language itself can vary from one speaker of the language to another [8].

2.2 Machine Translation Systems

Machine Translation Systems are any technological tools that are "solely responsible for the complete translation process from input of the source text to output of the target text without human assistance, using special programs, comprehensive dictionaries and collections of linguistic rules" [9]. Machine Translation Systems can either be classified as either bilingual or multilingual [10]. Bilingual machine translation systems are machine translation systems which are designed to translate from one language to another language in a single direction

[10]. On the other hand, Multilingual machine translation systems are machine translation systems which are designed to translate between more than two languages [10].

Even though different strategies have been adopted by various researchers in the creation of machine translation systems, the general architectural design used in the formation of machine translation systems can be grouped into two main broad categories, namely: (1) the linguistic architectural design and (2) the computation architectural design. [11].

Under the linguistic architectural design, there are three main basic architectural approaches that are used for developing machine translation systems. These three approaches are: (1) the direct approach, (2) the transfer-based approach and the (3) the interlingua approach [11]. The direct approach involves first taking a string of words from the source language and removing the morphological inflection from the words to obtain the base forms of the words. The next step in the direct approach then involves looking up the base forms of the source language words in a bilingual dictionary between the source and the target languages [11]. In the transfer-based approach, translation is done in three main stages. These three stages involve: (1) Converting source-language texts into intermediate representations, such as parse trees, (2) Converting the intermediate representations of the source language texts into their equivalent intermediate representations in the target language texts and (3) Generating final target language texts from the intermediate representations of the target language texts [11]. In the interlingua approach, language translation is achieved in two steps. The first step involves analysing the source language text, extracting its semantic content and using the extracted semantic content to create an interlingua form of the source language text. The second step of the interlingua approach involves using the interlingua form of the given source language text to generate a desired target language text [11].

In the computational architectural design, there are also three main architectural design approaches. These architectural approaches are: (1) the rule-based approach, (2) the corpus-based approach and (3) the hybrid approach [11]. In the rule-based approach of machine translation, different levels of linguistic, morphological, syntactic and semantic rules are used to translate a sentence from a source language to its corresponding target language. The corpus-based approach to machine translation involves the utilization of large amounts of bilingual language data to develop translation knowledge or models which can be used to transform a given source language text into its corresponding target language text [11]. There are two types of corpus-based machine translation [11]. They are Statistical Machine Translation and Example-Based Machine Translation [11]. In Statistical Machine Translation, statistical analysis and predictive algorithms are used to determine the rules that are best suited for translating a source language text to its corresponding target language text. In Example Based Machine Translation, computer systems are taught to transform source language text to target language text by sifting through large amounts of bilingual corpus, which contain examples of source language text and their corresponding target language text. The Hybrid approach to machine translation combines the transfer-based approach with one of the corpus-based approaches in the process of language translation [11]. The main idea in this approach is to learn synthetic transfer rules from limited amounts of word-aligned data and use these learnt rules to translate a given text from one language to another [11].

2.3 Machine Translation Systems for low-resource languages

Due to the vast amount of work and research done in the development of machine translation systems for low-resource languages in recent years, it is almost infeasible to conduct a detailed review of all the proposed approaches in the literature. As such, this section presents only an overview of a few proposed systems. However, a more complete review of recent developments in machine translation systems for low-resource languages can be found in [12].

Several machine translation systems have been developed for low-resource languages over the past few years [13, 14, 16].Vandeghinste *et al*. [13] developed a method for building a hybrid machine translation system without parallel corpora and an exhaustive rule-set. In the implementation of the system developed by Vandeghinste *et al*. [13], flat bilingual dictionaries were used. These bilingual dictionaries consisted of matches between source language pairs of lemmas and their associated parts-of-speech tags, and their corresponding target language pairs of lemmas and their associated parts-of-speech tags. During the process of translation, the system decomposes a given source language input text into lemmas and finds its corresponding target language lemmas from the bilingual dictionary. It does this by using the parts-of-speech tags of the input text lemmas as a guide. One drawback of this approach is that the word order of the translated text may be incorrect. Since this approach utilizes a lemma to lemma mapping from the source language text to the target language text, the word order of the translated text would reflect the word order of the source language text and this may not be the correct linguistic order of the translated text. Another drawback of this approach is that the translated text would be able to only capture the general meaning of the source language text. This is because the translated text would be expressed only in terms of the lemmas found in the source language text.

Irvine and Callison-Burch [14] used comparable corpora to develop a machine translation system to improve translation accuracy and reduce out-of-vocabulary rates when translating low-resource languages. A comparable corpus is one which selects similar texts in more than one language [15]. To improve accuracy, comparable corpora was used to estimate additional features over translation pairs and those additional features were used in decoding any given input text from the source language [14]. To reduce out-of-vocabulary rates during translation, Irvine and Callison-Burch [14] used bilingual lexicon induction techniques to learn translations for words which appeared in their test set and not in their training data [14]. Bilingual lexicon induction is the task of inducing pairs of words that are translations of one another from comparable corpora [14]. The approach adopted by Irvine and Callison-Burch [14] generated a BLEU between 0.5 and 1.7 [14]. BLEU (Bilingual Evaluation Understudy) is an algorithm which is used for evaluating the quality of translated text. It is usually measured from 0 to 1, with scores closer to 1 indicating a good translation and scores closer to 0 indicating a poor translation. One drawback of this approach is that it does not take into consideration the word order or syntax of the translated text[14].

Unlike Irvine and Callison-Burch [14], Shearing *et al.* [16] tackled the challenge of developing a machine translation system for low-resource languages by exploring the method of data augmentation. The form of data augmentation that was used by Shearing *et al.* [16] involved morphological glosses. A gloss, as defined by Shearing *et al.* [16], is a mapping between an inflected form of a word, and an in-situ translation [16]. Shearing *et al.*'s approach to machine translation involves the conversion of complex and inflected words from the source language data into multi-word English glosses, which are then appended to the training data and used to train the machine translation system [16]. Shearing *et al.*'s approach generated BLEU points of up to 1 when used in Russian-English translation tasks and also generated BLEU points of up to 2.4 when used in Spanish-English translation tasks [16].

2.4 Ethical implications of bias in Natural Language Processing Technologies

Bias with regards to Natural Language Processing technologies has to do with the fact that some natural languages are either absent or not accurately represented in the development of Natural Language technologies. One of the biggest ethical concerns with regards to biased Natural Language Processing systems has to do with the fact that the benefits that these systems provide would either not work properly or would not exist at all for people who speak certain languages.

2.5 Summary of related literature

The literature shows that the major problems with working with low-resource languages in the development of machine translation systems are that: (1) the linguistic resources for low-resource languages are generally difficult to obtain and (2) the orthographic information of most low-resource languages is not standardized.

With regards to circumventing the above problems and developing machine translation systems for low-resource languages, the literature shows approaches such as: (1) using bilingual dictionaries containing lemmas and their corresponding parts-of-speech, of both source and target languages for training the machine translation system, (2) using comparable corpora in the machine translation process and (3) using morphological glosses in place of complex source language words when training machine translation systems.

The literature did not however make mention of any machine translation approaches that have already been applied in creating machine translation systems for the Twi Ghanaian local language.

3 – Methodology

The main hypothesis for this research is that a Machine Translation System can be developed to translate between the Twi local language and the English language with limited data. To evaluate the above proposition, a parallel corpus of Twi-English texts was first gleamed from a bilingual Bible. After the textual data was obtained and pre-processed, a Machine Translation System was built to allow users to provide Twi text and have the system translate it into English language.

3.1 Collecting data

The data that was used in this research was obtained from the Genesis chapter of the Jehovah Witness online bilingual Bible. The reason why this online bilingual Bible was chosen as the primary source of data for the development of the Machine Translation system in this research was that it provides immense Twi-to-English bilingual data that could be rapidly collected and used.

After the data was collected, it was split into three distinct datasets, namely: training, testing and development datasets.

3.2 Pre-processing of the collected data

The collected data was normalized so that it would maintain a standard format that could be used efficiently in the machine translation system. The pre-processing methods that were first utilized in the pre-processing of the data were: Lemmatization and Stemming. In the application of lemmatization and stemming, repeated words, which originated from the same root word, that were found in the dataset were represented with only their root word in the dataset. This was done to prevent the occurrence of sparse probabilities that may have occurred during any probability computations with the words in the dataset. Another pre-processing method which was applied to the dataset was converting all the letters represented in the corpus to lowercase letters. This was done to prevent the scenario in which the same word would be represented as different words in the corpus just because some instances of the word contain capital letters.

3.3 Building the Machine Translation System

The Machine Translation System that was developed for this research is a Statistical Machine Translation System. The Statistical Machine Translation model was chosen for the development of this system because it not only enables a Machine Translation system to be applicable to different domains of Twi language, but it also ensures a reasonably good level of quality in the process of low-resource language translation in Machine Translation Systems.

The main idea behind a Statistical Machine Translation model is to maximize the probability P(S | T), where S represents a source language sentence and T represents a target language sentence. In this research, the source language is Twi and the target language is the English. The expression P(S | T) can be expanded further using the Bayes theorem as shown in Fig. 1 below:

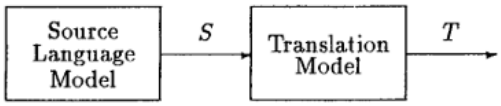
$$\mathrm{P(S \mid T)} = \frac{P(S)\ P(T \mid S)}{P(T)}$$

Fig. 1

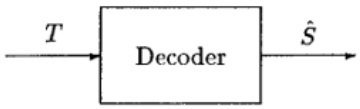
In the expansion of P(S | T), P(T) can be omitted since it does not depend on S and the remaining expression which is P(S)P(T|S) can be maximized. P(S) represents the language model probability and it suggests the order in which to place the predicted words from the source language. P(T|S) represents the translation probability and it suggests words from the source language that might have produced the words that we observe in the target language.

The overall design for the statistical machine translation system involves a method for computing language model probabilities, a method for computing

translation probabilities and a method for searching among possible source sentences for the sentence that gives the greatest value for P(S)P(T|S). The design for the statistical machine translation system is demonstrated graphically in Fig. 2 below:



$$P(S) \times P(T|S) = P(S,T)$$



$$\hat{S} = \text{argmax}_s P(S \mid T) = \text{argmax}_s P(S,T)$$

Fig. 2

3.3.1 Training the Source Language Model

In training the Source Language model, when given a word string S as '$s_1\, s_2\, s_3 \ldots s_{n-1}$' the source language model is expressed as a product of bi-grams in the following form:

$$P(s_1\, s_2\, s_3 \ldots s_n) = P(s_1)\, P(s_2 \mid s_1) \ldots P(s_n \mid s_{n-1})$$

Where for each word in the string S, we compute the probability of the word given the word that immediately precedes it in the given sentence.

3.3.2 Training the Translation Model

In training the translation model, a given Twi sentence would first be broken into a sequence of phrases or words and would then be mapped on a one-to-one basis to its corresponding target English words or phrases. For longer and more complex English sentence structures, where the word order in English would not be the same as the word order in Twi, parameters such as fertility and distortion would be computed and factored into the formulation of the Translation Model. 'Fertility' is a term that is used to showcase the number of Twi words that an English word is associated to in a given alignment. 'Distortion' is the term that is used to showcase that a Twi word will appear far away from the English word that is associated with it.

3.4 Decoding and Searching

In this step of the Statistical Machine Translation system development process, the objective is to find the Twi sentence S that maximizes P(S)P(T | S). This task is achieved by looping through the entire corpus and returning the Twi sentence that produces the highest probability.

3.5 Evaluation of the Machine Translation System

To evaluate the quality of the translated text produced by the Statistical Machine Translated System, the BLEU algorithm is used.

4 – Implementation

4.1 Implementation of Data Collection

The data that was used in the research was obtained by web-scrapping an online Twi – English bilingual Bible that can be found on a website called 'jw.org'. The web-scrapping tool that was used to obtain the bilingual language data was built using the Python programming language. In the implementation of the web-scrapping tool, methods from the following Python language libraries were used: (1) BeautifulSoup and (2) Html5lib. BeautifulSoup is a Python library that is used for pulling data out of Hypertext Markup Language (HTML) and Extensible Markup Language (XML) files and Html5lib is a Python library that is used for parsing HTML. The following series of high-level instructions represent the algorithm that was followed in the development of the web-scrapping tool:

1) Create the directory that would be used to store the web-scrapped data.

2) If for any reason the operating system is unable to create the directory, report the error that is responsible for this.

3) Obtain the URL of the Twi – English bilingual bible.

4) Make a web request to the URL obtained in Step 3 to determine all the Biblical books from which data is going to be mined from.

4) Go directly to the URL of each biblical book and use the BeautifulSoup library to mine all the information in the book.

5) While mining the biblical information, use the html5lib library to ensure that the information regarding each chapter of a Biblical book is written in a different text file and each sentence within the chapter is written on a new line in the text file.

After the bilingual data was mined, it was split into two datasets. These two datasets are the: (1) the training dataset and (2) the testing dataset. The training dataset was used to train the source language model and the translation model, and the testing dataset was used to check the translation efficiency of the developed machine translation system.

4.2 Implementation of the data pre – processing process

The Thot statistical machine translation toolkit was used to pre-process the bilingual language data that was collected. To initiate the data pre-processing process, the collected data was first tokenized. In the tokenization process, the bilingual textual data was broken down into tokens. Tokens are words, symbols or other meaningful elements which are generated from a given dataset to aid in the efficient development of machine translation systems. In the implementation of the tokenization process, the Thot machine translation toolkit was first installed, and the following commands were executed in the command line:

thot_tokenize -f ${PREFIX}/share/thot/en.train > en_tok.train

thot_tokenize -f ${PREFIX}/share/thot/en.test > en_tok.test

thot_tokenize -f ${PREFIX}/share/thot/tw.train > tw_tok.train

thot_tokenize -f ${PREFIX}/share/thot/tw.test > tw_tok.test

Note that in the above commands,

en.train represents the training dataset for English Language.

en_tok.train represents the tokenized training dataset for English language.

en.test represents the testing dataset for English Language.

en_tok.test represents the tokenized testing dataset for English Language.

tw.train represents the training dataset for Twi Language.

tw_tok.train represents the tokenized training dataset for Twi Language.

tw.test represents the testing dataset for Twi Language.

tw_tok.test represents the tokenized testing dataset for Twi Language.

After the process of tokenization, all the data was converted to lowercase. This was done using the thot_lowercase tool provided by the Thot statistical machine translation toolkit. To implement the process of converting the data to lowercase, the following commands were executed in the command line:

thot_lowercase -f ${PREFIX}/share/thot/en_tok.train > en_tok_lc.train

thot_lowercase -f ${PREFIX}/share/thot/en_tok.test > en_tok_lc.test

thot_lowercase -f ${PREFIX}/share/thot/tw_tok.train > tw_tok_lc.train

thot_lowercase -f ${PREFIX}/share/thot/tw_tok.test > tw_tok_lc.test

Note that in the above commands,

en_tok_lc.train represents the lowercased form of the tokenized English language training dataset.

en_tok_lc.test represents the lowercased form of the tokenized English language testing dataset.

tw_tok_lc.train represents the lowercased form of the tokenized Twi language training dataset.

tw_tok_lc.test represents the lowercased form of the tokenized Twi language testing dataset.

After the collected data was converted to lowercase, the dataset was cleaned, and the process of lemmatization and stemming was carried out on the data. This was accomplished through the utilization of the thot_clean_corpus_ln tool provided by the Thot toolkit. To implement the process of cleaning the dataset and applying lemmatization and stemming, the following commands were executed in the command line:

thot_clean_corpus_ln -s ${PREFIX}/share/thot/tw_tok_lc.train \-t ${PREFIX}/share/thot/en_tok_lc.train \> line_numbers

thot_extract_sents_by_ln -f ${PREFIX}/share/thot/tw_tok_lc.train \-n line_numbers > tw_tok_lc_clean.train

thot_extract_sents_by_ln -f ${PREFIX}/share/thot/en_tok_lc.train \-n line_numbers > en_tok_lc_clean.train

Note that in the above commands,

en_tok_lc_clean.train represents the cleaned and tokenized English Language training dataset.

tw_tok_lc_clean.train represents the cleaned and tokenized Twi Language training dataset.

4.3 Implementing the Source Language Model

For the implementation of the Source Language Model, the following high-level algorithm was developed and executed in the Python Programming language:

STAGE 1:

1) Open the text file containing the pre-processed and cleaned English language dataset.
2) Open the text file containing the pre-processed and cleaned Twi language dataset.
3) Create and open the text file that is going to contain the English word bigrams after the execution of the Source Language Model.
4) Create and open the text file that is going to contain the Twi word bigrams after the execution of the Source Language Model.
5) For every sentence in the English language dataset, split it based on the empty spaces between the words in the sentence to generate a list of words.
6) For every word in the generated list, append it to the preceding word in the list to create a bigram and write this bigram to the English word bigram text file.
7) Repeat Steps 5 and 6 for the Twi language dataset. Write the bigrams that are generated for the Twi language dataset into the Twi word bigram file.

STAGE 2:

1) When given a Twi source sentence, split the source sentence into bigrams based on the empty spaces between the words in the Twi sentence.
2) For each bigram obtained in Step 1 of Stage 2, calculate the probability of the bigram by finding the number of times that the bigram appears in the Twi word bigram file and dividing that result by the number of times that the preceding word in the bigram appears in the Twi word bigram file.
3) Multiply the probabilities of all the bigrams in the given Twi source sentence to obtain the probability of that Twi sentence, given the Twi language training dataset.

4.4 Implementing the Translation Model

For the implementation of the Translation Model, the following high-level algorithm was developed and implemented in the Python programming language:

1) Create a dictionary by mapping the various Twi word bigrams, generated during the execution of the Source Language Model, to their corresponding English word bigrams.
2) For each Twi source sentence given, break the sentence into a sequence of bigrams and find the various corresponding translations of each bigram. This can be done through looking through the dictionary to find the varied English translations of each bigram.

3) Link the varied English translations of all the bigrams together to form different English translations of the Twi source sentence.

4) Calculate for the probabilities of each of the English translations. Select the English sentence with highest probability as the translation of the Twi source language.

5 – Experiments and Results

This chapter discusses the experiments that were conducted to access the intrinsic performance of the Statistical Machine Translation system that was developed in this research with limited data. This Chapter also reveals and discusses the experiments that were performed.

5.1 Experiments

In this research 3 different experiments were performed on the developed Statistical Machine Translation system. In each of the experiments, 10 Twi sentences were used to test the system.

In the first experiment, all the Twi sentences that were used to test the system were sentences that had already been used to train the system. In the second experiment, 5 out of the 10 Twi sentences that were used to test the system were sentences that were not used to train the system. In the third experiment, all the Twi sentences that were used to test the system were sentences that were not used to train the system. All the Twi sentences that were used in the experiments can be found in the Appendix.

5.2 Results from the experiments

The results of the experiments were obtained on the basis of calculating the Bilingual Evaluation Understudy (BLEU) value of the Statistical Machine Translation system. The BLEU value can be calculated by using the following formula:

$$P_n = \frac{\sum_{n-gram \epsilon y} Count(clipped\ n-gram)}{\sum_{n-gram \epsilon y} Count(\ n-gram)}$$

Where

P_n is the calculated BLEU score of the translated sentence

$\sum_{n-gram \epsilon y} Count(clipped\ n-gram)$ is the number of words in the machine translated sentence which can be found in the reference sentence.

$\sum_{n-gram \epsilon y} Count(\ n-gram)$ is the number of words in the machine translated sentence.

In the first experiment, out of the 10 sentences that were used to test the system, only 3 of them had a perfect translation from the Statistical Machine Translation system, with a BLEU score of 1. For the rest of the sentences that were used to test the system, the Statistical Machine Translation system generated BLEU scores that ranged between 0.25 and 0.5.

In the second experiment, out of the 10 sentences that were used to test the system, only one of them had a perfect translation from the Statistical Machine Translation system, with a BLEU score of 1. For five of the test sentences that had not already been used to train the system, BLEU scores that ranged between 0 and 0.2 were generated. For the remaining four test sentences that had already been used to train the system, BLEU scores that ranged between 0.25 and 0.5 were generated.

In the third experiment, the BLEU scores that were generated for the test sentences ranged between 0 and 0.15.

5.3 Discussions of results

For the first experiment, the perfect translations that were obtained for 3 of the test sentences could be attributed to the fact that those sentences could be found in the data that was used to train the Machine Translation system. Also, in the first experiment, the BLEU scores that ranged between 0.25 and 5 for the rest of the test sentences could be attributed to the fact that the translations that were provided by the Statistical Machine Translation system, even though they were good, did not match up to human translation quality.

In the second experiment, the perfect translation that was obtained for one of the test sentences could be attributed to the fact that that sentence could be found in the training data of the Machine Translation system. Also, in the second experiment, the BLEU scores for the test sentences that ranged between 0 and 0.2 could be attributed to the fact that those sentences had not been encountered before in the training set and as a result the translations provided were of very poor quality. However, the BLEU scores for the test sentences that ranged between 0.25 and 0.5 could be attributed to the fact that the translations that were provided by the Statistical Machine Translation system were not of very high quality when compared to human translation quality.

In the third experiment, the low BLEU scores that ranged between 0 and 0.15 for the test sentences could be attributed to the fact that the Statistical Machine Translation system had not encountered any of the words in the test sentences during the process of training.

6 - Conclusion and Future Works

6.1 Summary

This research explored the question of whether the bias against low-resource languages in the development of Natural Language Processing systems could be mitigated through the development of a Statistical Machine Translation system for translating between the Twi and English language. The Statistical Machine Translation system that was developed for this research was made up of two main components, namely: The Source Language model and the Translation Model. The Source Language model was built to obtain the probability of any given Twi sentence and the Translation Model was built to obtain the probability of a target English sentence given the source Twi sentence. The limited data that was used in this research was obtained from the Genesis chapter of the Jehovah Witness online bilingual Bible.

The empirical results that were obtained in this research showcased general positive outcomes and demonstrated that even with limited amounts of data, a functional Natural Language Processing system may be developed for a low-resource language like the Twi language. The development of this system also proved that the process of collecting data for the development of Natural Language Processing systems does not have to be expensive or time-consuming. This research proved this point by demonstrating that the language data of low-

resource languages could be cheaply and easily obtained from unconventional data sources such as bilingual bibles.

Since the difficulty in obtaining language data had often been one of the principal excuses for not developing Natural Language systems for low-resource languages, this research has demonstrated that this should no longer be a barrier in the development of these systems.

6.2 Ethical considerations regarding the unintended consequences of the developed system

Even though the Machine Translation system that was developed in this research may serve as the first step to bridging the gap between low-resource languages and Natural Language Processing technologies, it can also be an avenue for certain problems. This section explores the various possibilities in which the Statistical Machine Translation system that has been developed in this research may be applied in unethical or sinister ways. One of the major unethical applications of this Statistical Machine Translation system is using it to obtain the Twi language translations of derogatory or insulting statements in the English language. Since the Machine Translation system developed in this research translates between the Twi language and the English language, it may be used to aid in the verbal assault of people who understand and speak the Twi language, by providing the Twi translations for certain insulting words and statements in the English language.

6.3 Suggestions for future works

This section presents a suggested extension to this research. By employing the suggestion presented in this section, this research study can be improved upon and applied in a wide variety of ways. The main suggestion offered to improve this research is that, more bilingual language data can be mined from more unconventional language data sources such as movies or music and used to improve the development of Natural Language machine learning models for low-resource languages. This could not only lead to the development of better Natural Language Processing systems, but it could also help to bridge the gap between low-resource and high-resource languages in the development of these systems.

Reference List

[1] Ethnologue. Summary by language family. Retrieved from: https://www.ethnologue.com/statistics/family

[2] Gelavizh Abbasi, Saman Saleh zadeh, Elenaz Janfaza, Arezoo Assemi and Siamak Saadat Dehghan. 2012. Language, Translation and Culture. Retrieved on March 20, 2019 from http://www.ipedr.com/vol33/017-ICLMC2012-L00062.pdf

[3] Peter Norwig and Stuart Russell. 1995. Artificial Intelligence – A Modern Approach. Prentice Hall, New Jersey, NJ.

[4] Ann Copestake. 2017. Natural Language Processing. Retrieved on March 20, 2019 from https://www.cl.cam.ac.uk/teaching/2002/NatLangProc/revised.pdf

[5] Benjamin P. King. 2015. *Practical Natural Language Processing for Low-Resource Languages*. Ph.D. Dissertation. University of Michigan, Ann Arbor, MI.

[6] Rutgers School of Arts and Science. 2019. Akan (Twi) at Rutgers. Retrieved on March 20, 2019 from https://amesall.rutgers.edu/languages/128-akan-twi

[7] Steven Bird and Gary Simons. 2003. Seven Dimensions of Portability for Language Documentation and Description. Retrieved March 19,2019 from https://www.researchgate.net/publication/220486407_Seven_Dimensions_of_Portability_for_Language_Documentation_and_Description

[8] F. Lüpke. 2011. "Orthography development," in *Handbook of endangered languages* (P. Austin and J. Sallabank, eds.). Cambridge University Press.

[9] Jaden Wu. 1985. A survey of machine translation: its history, current status, and future prospects. Retrieved on March 20, 2019 from https://www.academia.edu/1139860/A_survey_of_machine_translation_its_history_current_status_and_future_prospects

[10] W. John Hutchins and Harold L. Somers. 1992. An Introduction to Machine Translation. Academic Press, Cambridge.

[11] Mohamed Amine Chéragui. nd. Theoretical Overview of Machine translation. Retrieved on March 20,2019 from http://ceur-ws.org/Vol-867/Paper17.pdf

[12] AMTA. 2018. The 13th Conference of The Association for Machine Translation in the Americas. Retrieved on March 20, 2019 from https://amtaweb.org/wp-content/uploads/2018/03/AMTA_2018_Workshop_Proceedings_LoResMT.pdf

[13] Vincent Vandeghinste, Ineke Schuurman, Michael Carl, Stella Markantonatou and Toni Badia. nd. METIS-II: Machine Translation for Low Resource Languages. Retrieved on March 20, 2019 from https://s3.amazonaws.com/academia.edu.documents/38800097/258_pdf.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1554760739&Signature=7EBJ%2BztnQf4B4Rl0ADcpcGZKWKg%3D&response-content-disposition=inline%3B%20filename%3DMETIS-II_Machine_Translation_for_Low_Res.pdf

[14] Ann Irvine and Chris Callison-Burch. nd. Combining Bilingual and Comparable Corpora for Low Resource Machine Translation. Retrieved on March 20, 2019 from https://www.aclweb.org/anthology/W13-2233

[15] Belinda Maia. nd. What are comparable corpora? Retrieved on March 20, 2019 from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.197.1279&rep=rep1&type=pdf

[16] Steven Shearing, Christo Kirov, Huda Khayrallah and David Yarowsky. Improving Low Resource Machine Translation using Morphological Glosses. Retrieved on March 20, 2019 from **https://www.cs.jhu.edu/~huda/papers/shearing2018AMTA.pdf**

APPENDICES

The 10 Twi test sentences that were used for the first experiment

1) Na hann bae
2) Ade kyee
3) Da edi kan ni
4) Mfiase ne asase
5) Ewim mmra
6) Na ade sae
7) Nsu hann no
8) Wo nsu no ani
9) Ade kyee no ye

10) Hann no ye

The 10 Twi test sentences that were used for the second experiment

1) Me din de
2) Yefre me
3) Bra ha
4) Ko na meba
5) Kita wonsa
6) Na hann bae
7) Ade kyee
8) Da edi kan ni
9) Mfiase ne asaase
10) Na ade sae

The 10 Twi test sentences that were used for the third experiment

1) Mepese me koda.
2) Meti ye meya
3) Meko aba
4) Ewose meyi n'aye
5) Mani agye
6) Me do wo
7) Nya gyidi wo awurade mu
8) Me pe woasem
9) Esesee yeko fie
10) Me wo ahoto