Original Article

DEVELOPMENT OF AN ANDROID-BASED NLP SYSTEM FOR ENGLISH-TO-IGBO LANGUAGE TRANSLATION

Okoro Chinedu Samson and Nwankwo David Ebuka

Department of Computer Science, Rivers State University, Port Harcourt, Nigeria
DOI:https://doi.org/10.5281/zenodo.15430438

Abstract

Machine Translation (MT) is a process, sometimes referred to as Natural Language Processing which uses a bilingual data set and other language assets to build language and phrase models used to translate text. Statistical Machine Translation utilizes statistical translation models generated from the analysis of monolingual and bilingual training data. Essentially, this approach uses computing power to build sophisticated data models to translate one source language into another. Statistical Machine Translation consists of Language Model (LM), Translation Model (TM) and Decoder. In this research, English to Igbo language translation system has been developed. The Language Model, Translation Model and Decoder is done in Microsoft Hub; training of parallel document, and the language translation system was implemented in Android studio environment, can be accessed through Android application in smartphones. English and Igbo language tokens were determined using Finite State Automata; transition in each state identified the valid token and invalid. Valid tokens were found where transition produces letters, invalid tokens occur when a transition produce combination of digit and letter. English and Igbo language sematic were determined using attribute grammar which was further expressed in parse tree showing the syntax structure. An integrated custom keyboard was developed to input the Igbo words and phrases. Result shows one to one and one to many mapping of English to Igbo words/phrases.

Keywords: Natural Language Processing, Custom Keyboard, Android OS

1. Introduction

Translation can without a doubt be thought about from the point of view of language, culture or society and translation researchers have added to the comprehension of the field all in all by considering the different parts thereof. Translation can be conceptualized as acting, sneaking or cross-acting. There should an applied space in which to relate proposals points of view to each other and inside which one could comprehend why these roads are taken in the endeavored to conceptualize translation. The applied space ought to not just concentrate on either contrast or closeness. Or maybe, it ought to be a conceptualization that persistently recognizes similitude and

Original Article

recognizes (contrasts), that is, it keeps up a theoretical oddity between parts of translation. Understanding the interrelationships between the different parts of translation could help with making a comprehension of the marvels with which we are working, (Kobus, 2014). Driven by the development of a worldwide economy and advancements in high innovation, the way toward making and deciphering specialized documentation has been developing quickly. Specifically, Machine Translation (MT) has demonstrated expanding abilities of effectually achieving the beginning times of the eight phases of translation. As an outcome, translators have figured out how to utilize machine translation as an instrument to quicken their work, yet they have likewise become careful about machine translation's potential for supplanting them. To guarantee solid job, a few translators have started broadly educating as specialized essayists; correspondingly, a couple of specialized writers have started broadly educating as translators, as the two callings give off an impression of being experiencing a progressive pattern of combination. Scholarly projects are asked to react to the advancing patterns (Maylath, 2013). (Odejobi, et al.., 2015). Propose system that can aid the instructing and learning of Hausa, Igbo, and Yoruba. The investigation considers human body parts ID, plants distinguishing proof, and creatures" names. The English to Yoruba machine translation and Yoruba number tallying systems are a piece of the fundamental system. The exploration of Yoruba language translation has gone so wide. In any case, to date there has been almost no examination concentrated on Igbo language translation precisely how, why, and to what degree documentation benefits these endeavors. The little research done on Igbo language translation prompted this exploration. The purpose of this project is to build a translator Android application which can be used by people who want to translate texts from English into Igbo. With the help of translator applications people can at least get a general understanding of a foreign text. This is particularly useful for developing countries such as Nigeria with scare written resources in science and technology.

2. Related Literature

Natural Language Processing utilizes a bilingual informational collection and other language resources to construct language and expression models used to interpret content. This definition includes representing the linguistic structure of every language and utilizing standards, precedents and sentence structures to exchange the syntactic structure of the Source Language (SL) into the Target Language (TL) (Mouiad and Tengku, 2011). (Odejobi, et al., 2015) proposed a system that can aid the instruction and learning of Hausa, Igbo, and Yoruba languages. The investigation considers human body parts ID, plants distinguishing proof, and creatures" names. The English to Yoruba machine translation and Yoruba number tallying systems are a piece of the fundamental system. The model was designed to build a system for the learner of the three languages. (Hana, 2016) proposed “Amharic to English Language Translator for iOS (iPhone Operating System)”: iOS application to make a Translation of English to Amharic and the other way around. The interpreter application utilizes a Translation system which was based on Microsoft Interpreter Center point and utilized Microsoft Interpreter Programming interface. The application can be utilized to make a Translation of writings from Amharic to English and vise versa. (Goyal and Lehal, 2010) proposed "Hindi to Punjabi Machine Translation System": This system is based on direct word-word translation method. It consists of morphological analysis, word sense disambiguation, transliteration and post processing. (Agbeyangi, et al., 2015) proposed "A rule-based approach for English to Yoruba Machine Translation System": There are three ways to deal with Machine Translation process. The creators investigated these methodologies and considered principle based methodologies for the Translation procedure. As indicated by authors, there is constrained corpus that is accessible for Yoruba language, which illuminates the standard based methodology. (Mouiad, and Tengku, 2011) proposed Rule-Based and Example-Based Machine Translation from English to Arabic: English to Arabic methodology for deciphering very much

Original Article

organized English sentences into all around organized Arabic sentences, utilizing a Grammar based and example-translation procedures to deal with the issues of requesting and understanding. The proposed strategy is adaptable and versatile, the principle points of interest are: initial, a cross breed based methodology joined focal points of guideline based (RBMT) with focal points precedent based (EBMT), and second, it tends to be connected on some different languages with minor adjustments. The OAK Parser is utilized to break down the info English content to get the grammatical feature (POS) for each word in the content as a pretranslation process utilizing the C# language, approval rules have been connected in both the database structure and the programming code so as to guarantee the honesty of information.

3. System Design

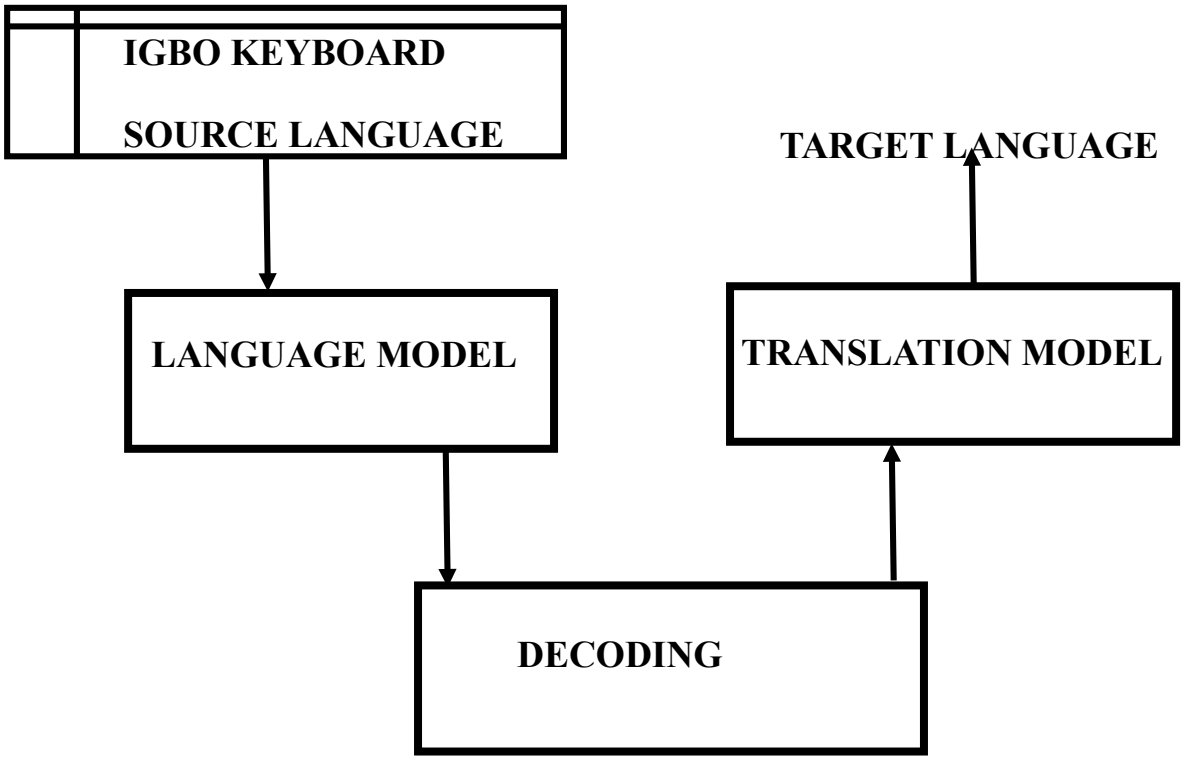

Figure 1: Architecture of Statistical Machine Translation

A parallel document was analyzed and aligned word or phrase, then a translation model will be produced. Language model is prepared from target language. The decoder produces a translation of the source language into the target language using the language model and translation model; it gives the probability of target language given the source language.

3.1 Token (Source Language)

A word in source language is recognized using Finite State Automata (FSA) for English Language Token

Original Article

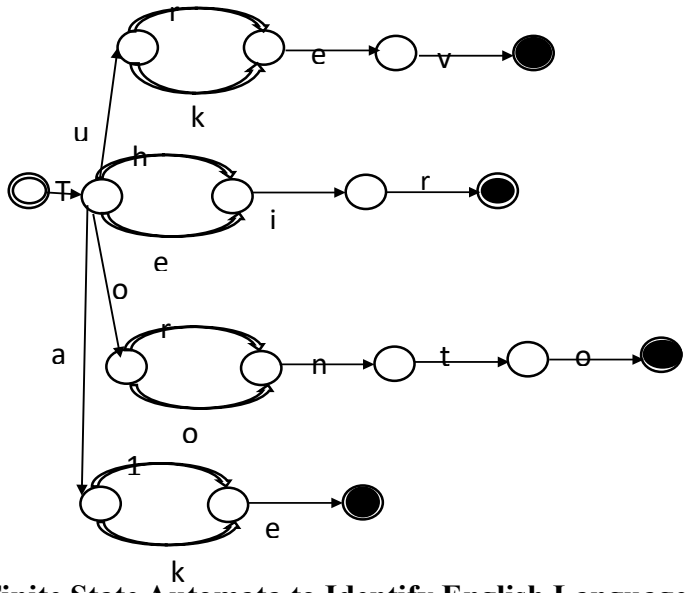

Figure 2: Finite State Automata to Identify English Language Token

Finite State Automata is used to determine English language tokens. Transition could generate the token "Their", another transition could generate the token ""Turkey", another transition could generate the token "Toronto". Another transition could generate an invalid token "t1ke", digit (1) is included in the token which makes it an invalid token.

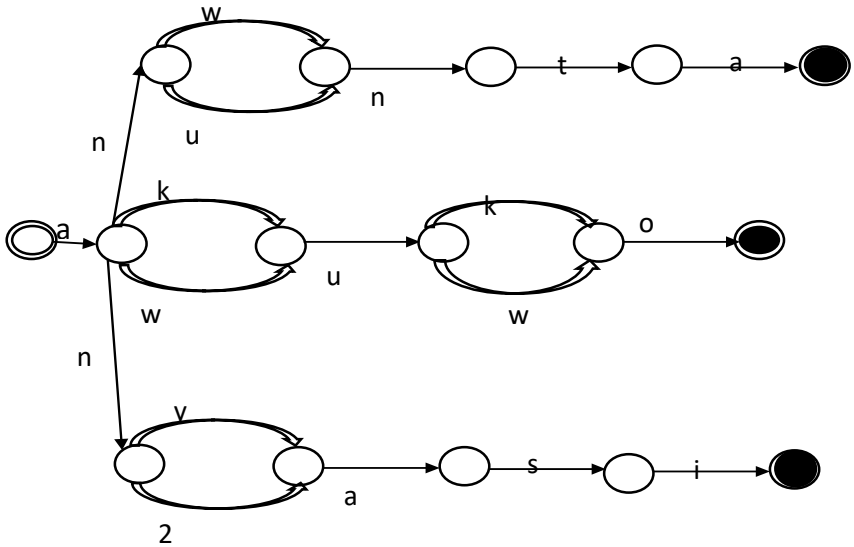

Figure 3 Finite State Automata to Identify Igbo Language Token

To determine Igbo language token; Transition could generate the token "akwụkwọ", another transition could generate the token ""anwụnta", another transition could generate the token "Toronto". Another transition could generate an invalid token "any2asị", digit (2) is included in the token which makes it an invalid token.

Original Article

3.2 Language Model

Language Model (LM) decides the likelihood of a sentence P(s); it utilizes the input of source language. The likelihood of a sentence is directly proportional to the likelihood of each word P(w).

Therefore,

$P(s) = P(w)$ (1) If 's' is a source sentence;

LM computes P(s) and input to the decoder.

3.3 Translation Model

Translation Model (TM) computes the probability of source sentence P(ss) for target sentence P(ts). Translation is done word-based, phrase-based or syntax-based.

$P(ss) = P(ts)$ (2)

3.4 Word-based Mapping

Numerous translations in a single language can allude to a solitary word in another language. For example; Igbo word Akwukwo can be translated in English as Book, Paper, Publication, and Text.

Utilizing a likelihood conveyance work, P(Akwukwo)Book = 0.4. This can be determined by thinking about the occurrence of English translations in Table 1. Words are assigned specific rank (values) which are used to compute the probabilities of each word.

Word	Word Rank
Book	8
Publication	5
Paper	3
Text	2
∑WC = 18	

Table 1 Translation of Akwukwo

$P(\sum WC) \leq 1$

P(Akwukwo) Book = 0.4

P(Akwukwo) Publication = 0.27

P(Akwukwo) Paper = 0.16

P(Akwukwo) Text = 0.1

Therefore,

$P(\sum W) = 0.9 \approx 1$

Original Article

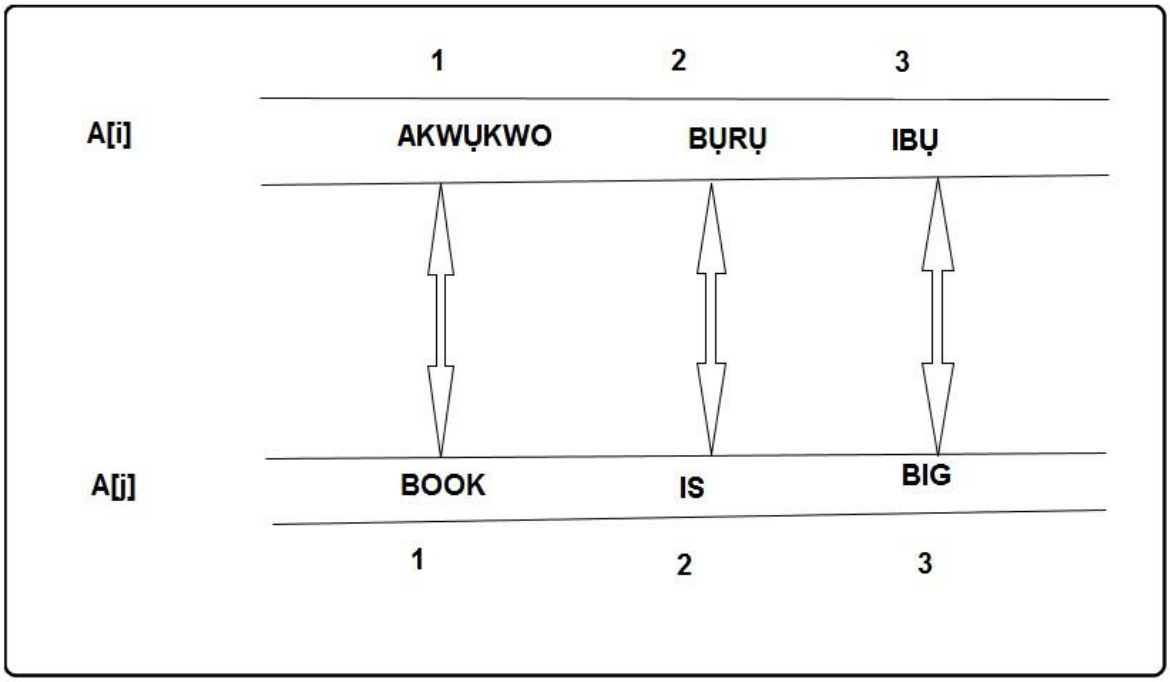

Figure 4 Word-based Mapping

In word model, the conceivable translations of one language is mapped in another language. To translate 'Akwukwo', translator picks the English word with the most noteworthy likelihood, 'Book'. Translator computes likelihood on noun and adverb, checking their elective articulation (word), however does not check probability does not check likelihood on action word as action word has no other word.

The source language at position 'I' and the target language at position j is defined as:

$A: i \rightarrow j$ (3) given

$A: \{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3 \ldots\ldots i_n \rightarrow j_n\}$ (4)

3.5 Phrase-based Mapping

The sentence is broken down into segments called phrases and aligned according to the grammatical rule of the language. Figure 5 shows how one word in a language can be mapped into two or more words in another language.

Original Article

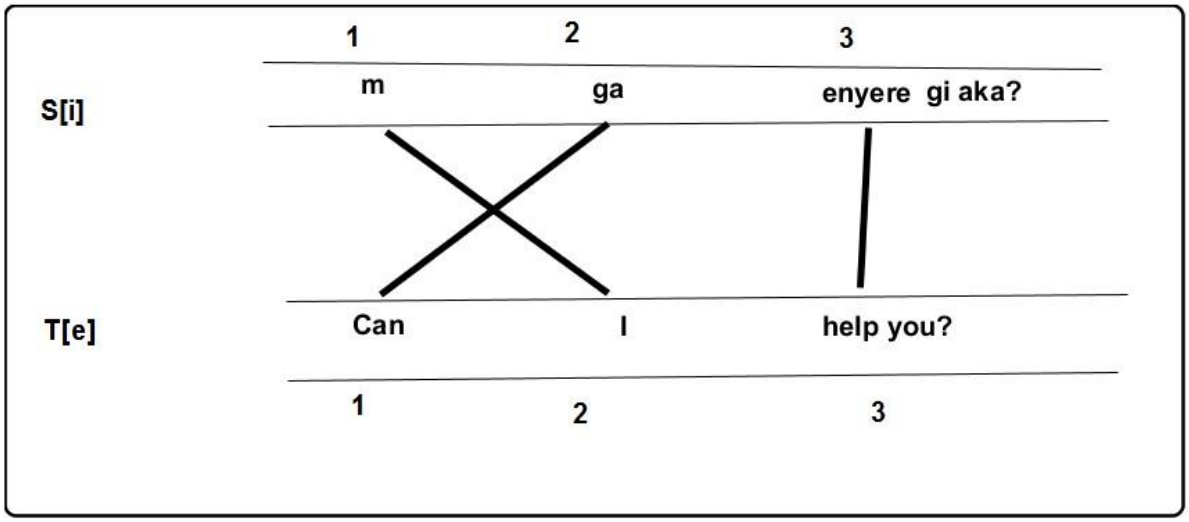

Figure 5 Phrase-based model

This model translates whole sentence which cannot be carried out in word-based mapping. Mathematical representation of a phrase-based mapping is given as: Suppose 's' is a source language and 't' is target language, we can break down a source language into a sequence of

'i' phrases and target language into a sequence of 'e' phrases.

$S = S_1, S_2, S_3,..S_i$ (5)

$T = T_1, T_2, T_3,..T_e$ (6)

Where S = T

Using a probability distribution function $\varphi(t_e/s_i)$ to translate each phrase in source language (s_i) into target phrase (t_e).

$$P(S/T) = \frac{P(s1/t1)}{P(t1)} P(t_1/s_1)P(s_1) \quad (7)$$

3.6 Syntax-based mapping

This is used to incorporate the structural difference of sentences in different languages into the Word-based model. Figure 6 shows the translation between Subject-Verb-Object (SVO) languages. Consider the English sentence: "He likes going to school" and Igbo sentence: "Ọ na-enwe mmasị ịga ụlọ akwụkwọ".

Using the production rule:

SENTENCE ----------→ <NP>< VP>
NP ----------→ <DET>< NOUN>
NP ----------→ <DET><ADJ><NOUN>
NP ----------→ <PREP>< NP>
VP ----------→ <VERB><NP>
VP ----------→ <VERB><ADVERB><NP>
VP ----------→ <VERB>< ADVERB>
VP ----------→ <VERB>

Original Article

Original Article

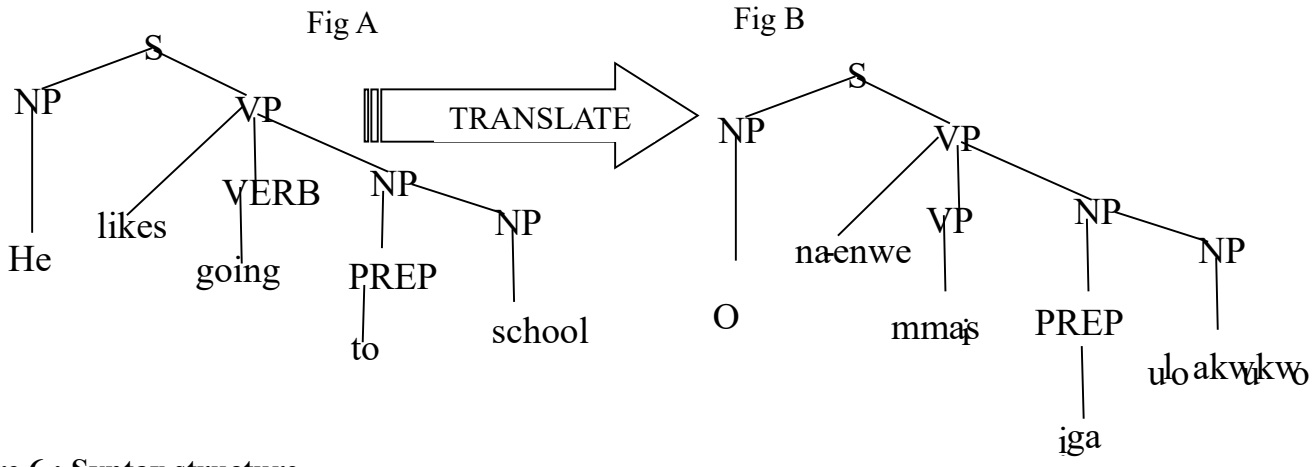

Figure 6 : Syntax structure

Grammar based acknowledges a parse tree as an input and in the preparing channel tasks, embedding additional words at every node, and translation of leaf words are performed on every node of the parse tree. The Verb Phrase (VP) has children nodes which creates a child node (Token).

4. Results and Discussion

Training of data set was done in Microsoft Translation Hub. Training setup involves choosing a project, training new system and uploading documents, tuning and testing. Language translation has to do with training of parallel documents containing source language (English) and target language (Igbo). Thus, if training extracted sentence <10,000 OR tuning set<2500 then training will fail. Dictionary is optional. Multiple training was done until a satisfactory result was found. It took several training runs in order to create a suitable translation system for the project1 and Project2.

The codes and the layout were done so that it would be straightforward to add other words and phrases of English to Igbo languages. The application has a Home screen where a user can enter a text in source language to be translated in the target language. When a user taps on the TextField "Type Your Word ", inbuilt English keyboard will appear and user types in the word or phrase and click "Translate" button to see the corresponding word or phrase in TextFiled "Igbo words". Also, English textTospeech was added between (Type Your Word) TextField and (Igbo words) TextField. The plus sign button (+) is used to add words and phrases to the system. "Three stroke lines" button is used to migrate from Igbo to English page as shown in Figure 7.

To migrate to target language (Igbo), the user presses the triple line tab to display Igbo to English page. When a user taps the TextField "Type Your Word" to enter the text, the Igbo keyboard appears. The keyboard contains a button to dismiss the keyboard named “X”, a button to clear the text of the TextField called "three lines" and a button to translate the entered text into the target language called “Translate “. The Igbo keyboard has six rows. First row consists of the digits (0-9), second row consists of alphabets (A, B Ch, D, E, F, G, Gb) and the sixth row consist of alphabets (V, W, ohere, Y, Z) the key (ohere) in sixth row is used as spacebar in English keyboard. If the user presses the translate button the translated text appears in the TextField "English words" as shown in Figure 8.

Consider an English sentence "adaeze is a school girl" the corresponding translation in Igbo language is shown in Figure 7 and Igbo sentence is shown in Figure 9. Igbo word “agwa”, the translation in English language is “beans” as shown in Figure 10

Original Article

Table 2 English to Igbo (Project 1)

No of Runs	TRAINING (Byte)		TUNING (Byte)	TESTING	DICTIONARY (Byte)	TRAINING DURATION (SECOND)	BLEU SCORE
1	Extracted Sentence count	10,625	3,369	Auto generated	10,376	2 hour 17 minutes	81.13%
2	Extracted Sentence count	10,525	3,369	Auto generated	10,076	2 hour 11minutes	79.45%
3	Extracted Sentence count	10,490	3,180	Auto generated	8,230	2 hour 10 minutes	77.68%
4	Extracted Sentence count	10,238	2,903	Auto generated	8,125	2 hour 7 minutes	76.10%

Table 4 Igbo to English (Project 2)

No of Runs	TRAINING (Byte)		TUNING (Byte)	TESTING	DICTIONARY (Byte)	TRAINING DURATION (SECOND)	BLEU SCORE
1	Extracted Sentence count	10,626	3,369	Auto generated	10,376	3 hours 2 minutes	73.73%
2	Extracted Sentence count	10,526	3,369	Auto generated	10,076	2 hours 58 minutes	69.23%
3	Extracted Sentence count	10,491	3,180	Auto generated	8,230	2 hours 50 minutes	51.65%
4	Extracted Sentence count	10,239	2,903	Auto generated	8,125	2 hours 22 minutes	46.21%

Original Article

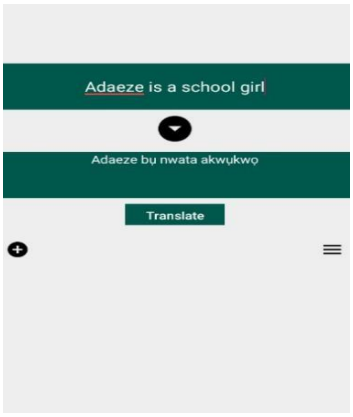

Figure 7: English Sentence Translation to Igbo

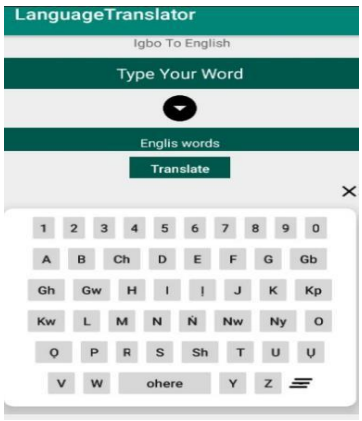

Figure 8: Igbo to English Translation Keyboard

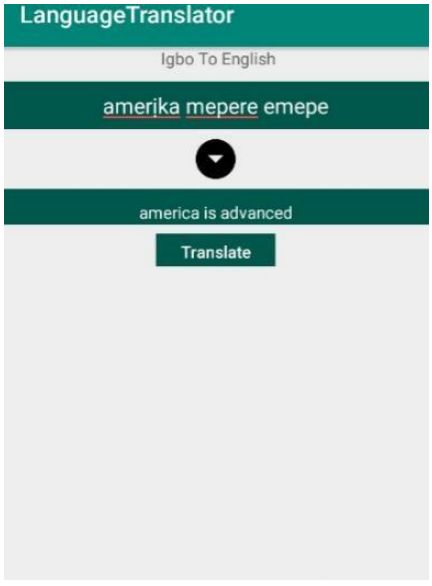

Figure 9: Igbo Sentence Translation to English

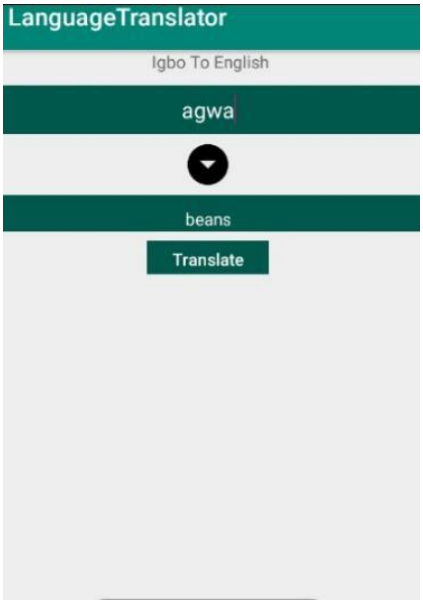

Figure 10: Igbo Word Translation to English

5. Conclusion

The application was tested in Android 5.1, Android 6 simulators and the user interface looks the same in all cases. The performance of the application in terms of speed is satisfactory. Translation quality is satisfactory for both English-Igbo and Igbo-English. The system succeeds in translating both words and phrases (one to one mapping and one to many mapping) of English to Igbo language. More words and phrases can be added as the system implemented addTranslation() method. One way to improve the translation system is to make parallel corpora by extracting sentences from websites and other documents and use the free sentence processing tolls on the internet, or use the free TMX editors and make parallel documents sentence by sentence. Despite that we have successfully built language translation system for bilingual corpus and developed Igbo custom keyboard, yet there is need for further improvement of English to Igbo statistical machine translation system on the following: The work can be

extended to speech to text recognition and to include multilingual corpus of different languages in the source-target pair.

References

Agbeyangi, A. O., Eludiora, S.I., and Adenekan, D.I. (2015). "English to Yorùbá Machine Translation System using Rule-Based Approach". Journal of Multidisciplinary Engineering Science and Technology (JMEST), Vol. 2 Issue 8, August, Nigeria. Goyal, V., and Lehal, G. S. (2010). "Hindi to Punjabi Machine Translation System". Department of Computer Science, Punjabi University, Patiala, India Hana, B.D. (2016). "Amharic to English Language Translator for iOS". Department of Information Technology, Helsinki Metropolia University, Finland

Kobus, M. (2014). "Translation Theory and Development Studies. A Complexity Theory Approach" Published by Routledge, New York, USA.

Maylath, B. (2013). "Current Trends in Translation" Department of English, North Dakota State University, USA

Microsoft Corporation (2018). "Microsoft Translator Hub User Guide". Publishes by Microsoft Corporation, USA.

Mouiad F. A. and Tengku M. S. (2011). "Rule-Based and Example-Based Machine Translation from English to Arabic".

Odejobi, O. A., Ajayi, A. O, Akanbi, L. A., Eludiora, S.I., Iyanda, A. R., and Akinade, O. O. (2015). "A Web-Based System for Supporting Teaching and Learning of Nigerian Indigenous Languages". Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria