LAUTECH Journal of Computing and Informatics (LAUJCI) – ISSN: 2714-4194
Volume 2 Issue 1, May 2021 – www.laujci.lautech.edu.ng

Development of a Syntax-Based Model for English - Igbo Statistical Machine Translation

*[1]Adebimpe Esan
Dept of Computer Engr. Federal University Oye-Ekiti, Ekiti state, Nigeria
adebimpe.esan@fuoye.edu.ng

*[2]John B. Oladosu
Dept of Computer Engr. LAUTECH, Ogbomoso, Oyo State, Nigeria
jboladosu@lautech.edu.ng

[3]Ibrahim Adeyanju
Dept of Computer Engr. Federal University Oye-Ekiti, Ekiti state, Nigeria

[4]Nnamdi Okomba
Dept of Computer Engr. FederalUniversity Oye-Ekiti, Ekiti state, Nigeria

[5]Shadrach Oforkansi
Dept of Computer Engr. Federal University Oye-Ekiti, Ekiti state, Nigeria

(Corresponding Author: jboladosu@lautech.edu.ng, adebimpe.esan@fuoye.edu.ng)

ABSTRACT

Semantic errors occurred due to syntactical difference between English and Igbo languages in existing statistical machine translators. Therefore, a syntax-based model was developed in this research for English-Igbo statistical machine translation. Parallel corpus was obtained from the religious domain and word alignments were made on the English and Igbo corpora with GIZA++. The Hidden Markov Model uses the word alignments produced by GIZA++ to estimate a maximum likelihood translation table. The Language model for the target language was built using IRSTLM toolkit and the model was tuned using Minimum error rate training (MERT). The developed SMT system was evaluated using BLEU and NIST and the results were compared to an existing related work. Results showed that the developed model outperformed the previous system by up to 0.3 BLEU score and 3.0 NIST scores respectively.

Key words: Syntax, Religious, Language model, Translation model, Domain and Corpora

1.0 INTRODUCTION

Communication is the act of conveying meanings from one entity or group to another, through the use of mutually understood signs, symbols, and semiotic rules. It is the act of transferring information from one place, person or group to another. Communication is achieved by either text or speech using a language that can be understood by the communicating parties. Human communication is unique for its extensive use of abstract language. Language is an efficient medium of communication which expresses the human mind (Oladosu, et al., 2016). There are over 6,800 languages in the world today and this reflects the scope of linguistic diversity. Traditionally, human linguists help in translating from one language to another and the limitations such as: high cost and slow speed of translation led to the emergence of machine translation.

Machine Translation (MT) is a sub-field of computational linguistics that investigates the use of software to translate text or speech from one language to another (Kenji 2010). Machine Translation approaches can be divided into two: single approach and hybrid approach (Oladosu, et al., 2016). Single approach include: rule-based (symbolic) and empirical (data-driven) (Ayogu, Adebayo , & Bolanle , 2018), Subalalitha *et al.* (2018). The blend of the single approaches with one another makes up the hybrid approach (Ayogu, *et al*., 2018) (Chéragui1, 2010). The combination of statistical approach and rule based engine produces the following approach: word based (Brown *et al* 1993), phrase based (Koehn *et al*. 2007) and syntax based models (Daniel & Helena 2013).

However, previous researches have developed English – Igbo translators but the systems are not suitable for the target usage because they yield poor quality output. An example is a phrase based statistical machine translator developed by Ayogu, *et al*. (2018). The shortcoming of the approach is the lack of syntactically motivated units which makes it difficult to model changes

in word order effectively thereby yielding poor quality output. Hence, this research developed a syntax based model for Igbo-English Statistical Machine Translation.

2.0 RELATED WORKS

Many approaches such as: rule-based, statistical, knowledge-based and direct approaches have been employed by previous research to develop machine translation (MT) systems. Statistical MT approach is based on statistical models that have been structured from the parallel corpora (Folajinmi and Omonayin, 2012). Also, the rule based approach involves analyzing input text morphologically, syntactically and semantically, and generating text via structural conversions based on internal structures (Chéragui1, 2010). The approach has been used in automatic translation of English sentences to Nigerian languages such as Yoruba, Igala, Hausa, Okun and Igbo. The approach was employed by Ayegba *et al.*, (2015) to translate English language to Igala language, Esan et al., (2018) to translate English adjectival phrases to Yoruba and Agbeyangi *et al.* (2015) to translate English words to Yoruba language as well as English-Foreign language translation (Subalalitha *et al.* 2018)
.

Furthermore, direct approach to machine translation is referred to as the most primitive approach to machine translation because it replaces words in the source language with words in the target language in the same sequence without much linguistic analysis and processing. The bilingual dictionary is the major resource used by the approach (Oladosu *et al*., 2016). Abikoye *et al*. (2016) employed direct approach for translating English language texts to its equivalent Yorùbá language. In addition, the Knowledge-based approach was employed by Oladosu and Olamoyegun (2012), to develop a multi-lingua machine translator. However, the shortcomings of the approaches include: inability to give accurate translations of full sentences, inflexibility for large-scale application and poor quality output.

Previous works have been done to improve the quality and fluency of machine translation output. Research show that two machine translation approaches (mostly rule based and statistical) have been combined to form a hybrid MT system with the aim of improving the quality of translation in this area. One of the approaches is the text to text or word-based statistical machine translation model. Word-based models were designed by Brown *et al* (1993) to model the lexical dependencies between single words. The approach was proposed by Abiola et al., (2015) to translate English language sentences to yoruba language. The shortcoming of the approach is its inability to handle word order effectively thereby yielding low quality output. Another approach that was introduced to improve the quality of machine translation of Nigerian languages is the phrase-based approach. This approach employed phrases as the basic unit of translation and be any substring. The approach allow local re-orderings, translation of short idioms, or insertions and deletions that are sensitive to local context. As a result, the context of words tends to be explicitly taken into account and the difference in local word orders between source and target languages can be learned explicitly. Ayogu *et al*. (2018) developed a phrase based MT system for translating English language to two Nigerian languages: Igbo and Yoruba as well as Igbo language to Yoruba language. Esan et al. 2020, developed a recurrent neural network model for machine translation and the approach used improved the quality and fluency of machine translation output greatly. The limitations of previous Phrase-based MT system include: lexical, grammatical and semantic errors due to the inability of the system to model long distance reordering of words thereby yielding poor quality output. Hence, a syntax based statistical machine translation system was developed in this research to improve the quality of translation of Nigerian languages.

3.0 METHODOLOGY

The methods used in developing the Syntax-based English-Igbo statistical machine translation involve: design of the model, translation process and software design.

3.1 Design of the Syntax-based Model

The syntax based-model was designed to include the following components: Source Text, tokenizer, translation model, re-ordering model, decoder and language model as shown in Figure 1. The source text refers to the text to be translated and the tokenizer divides the source text into unit of words called tokens in order for the words to be parsed to the translation model. The translation model is made up of the phrase table and rule table where the phrase tables contain words and phrases recognized in the parallel corpus and the rule table contained probability rules for joining one or more phrases by reordering the phrases. The phrase table receives the parsed words from the tokenizer and recognizes some sentence phrases amongst the words. The recognized phrases are passed to the rules table which then arranges the phrases into a correct grammatical structure through the rules derived from the rule table.

The Decoder helps to find the highest scoring sentence in the target language (according to the translation model) corresponding to a given source sentence. The language model ensures the output is fluent in the target language and also

ensures letter case rules for the target language is observed with the aid of a re-caser model. The reordering model ensures that the output target sentence is grammatically correct by reordering the translated sentence to follow the grammatical structure of the target language. Target text refers to the translated text produced by the model in the target language. System GUI connects the User and the translator and provides space for the user to easily type-in what they intend to translate and also view the result of translation.

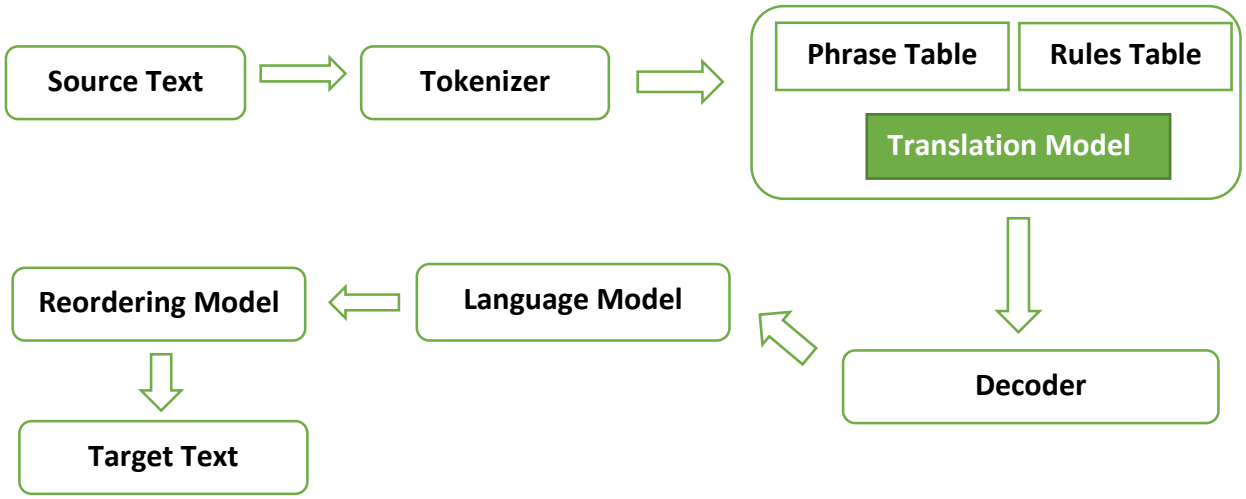

Figure 1: Architecture of the developed Model

3.2 Translation Process

Mathematically, the translation process is expressed as:

$$p(en|ig) = \underset{phrase\ table}{\emptyset(ig|en)^{weight}} \times \underset{language\ model}{LM^{weight}} \times \underset{reordering\ model}{D(en,ig)^{weight}} \times \underset{word\ penalty}{W(e)^{weight_{\emptyset}}} \quad (1)$$

From equation 1, the model composed of four (4) major components that are involved in translation, which are:

1. Phrase translation table: the table ensures that the English phrases and the Igbo phrases are good translations of each other. The table also serves as database of fetching Igbo and English phrases with their correct translations of each other/
2. Language Model: This model ensures fluency in the target side. The model ensures that the target side of the translation is fluent in the target language by using proper letter cases where necessary, use of proper punctuation marks and tonal marks.
3. Reordering Model: This model ensures proper re-arrangement of the target translation by following the grammatical structure of the target language.
4. Word penalty: The word penalty ensures that the translation does not get too long or too short by replacing long phrases with single words from the phrase table.

3.3 Software Design

Software was designed using Java.JavaFX with other material design components (JFoenix) was used to develop the front-end of the app. The backend was developed by linking the front-end to the model by creating local connections between the UI and model developed.

A trigram (3 gram) model was developed for both English and Igbo language models. The Kneser-Ney algorithm was used to smoothen and improve the accuracy of the developed language models.

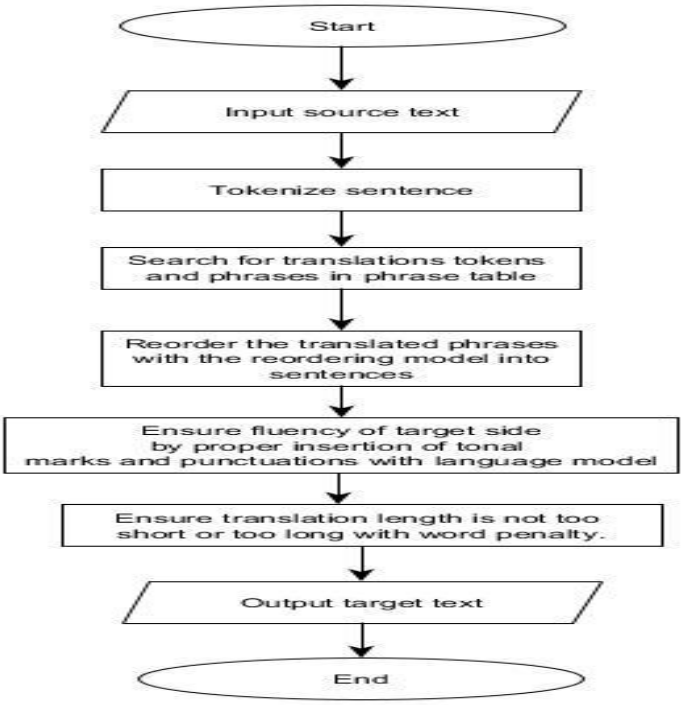

Figure 2: Flowchart of the translation process

3.4 Model Implementation

The designed model was developed through the following processes: acquisition of parallel corpus, corpus preparation, tokenization and categorization, language model training, word alignment with Giza++, translation model training, tuning, software design.

3.4.1 Data Acquisition

The parallel corpus used was acquired from the religion domain. The religious domain has been one of the most reliable sources for the acquisition of error free parallel corpus. The New World Translation (NWT) bible was used as one of the parallel corpus because it has both the English translation and Igbo translation of the Bible. The biblical books used are Genesis, Exodus, Numbers, Mathew, Galatians, and Revelations. The parallel corpus was ensured to be sentence aligned. The total number of sentences in both corpora used are 30953.

3.4.2 Corpus Cleaning, Tokenization and Categorization

The parallel corpus used was acquired from English-Igbo New World Translation (NWT) bible. The corpus was ensured to be sentence aligned with a total number of thirty thousand sentences in both corpora.The following processes were carried out during the cleaning stage: long sentences and empty sentences were removed as they can cause problems with the training pipeline, misaligned sentences were removed to ensure the equality of the parallel corpus, numerical digits were removed and normalization of the Igbo sentences was carried by appropriate insertion of tonal marks in each word present in a sentence, so as to convey appropriate meaning. The parallel corpora were also tokenized using the MOSES tokenizer script. During the process of tokenization, the sentences were broken down into units called tokens. The tokens could be a word or

character present in the sentence. The corpora were then divided into different segments namely: training corpora, tuning corpora, and test corpora, in the ratio: 94%, 3%, and 4% respectively. Therefore, the training corpora has 28982 sentences, tuning corpora has 969 sentences, and test corpora has 1000 sentences.

3.4.3 Language Model Training

The language model was used in this MT model to ensure that the output of translation is grammatically correct, letters are properly recased, and the output conforms to the grammatical structure of the output language. A language model was trained for the two different language i.e. Igbo and English. IRSTLM toolkit was used to train both language models using both an English corpora and Igbo. A trigram (3 gram) model was developed for both English and Igbo language models. The Kneser-Ney algorithm was used to smoothen and improve the accuracy of the developed language models.

3.4.4 Word Alignment

GIZA++ toolkit was used to carry out word alignment process on both language corpora. This toolkit is an implementation of the original IBM models that started statistical machine translation research. Word alignment with GIZA++ is carried out in the following process: First, the parallel corpus is aligned bidirectional. This generates two word alignments that has to be reconciled. To establish word alignments based on the two GIZA++ alignments, a number of heuristics were applied. The default heuristic; grow-diag-final starts with the intersection of the two alignments and then adds additional alignment points.

3.4.5 Translation Model Training

Word alignment with GIZA++ is the first process of training the translation model. Given the word alignments with GIZA++, two lexical translation tables are being developed using a Hidden Markov Model (HMM). The Hidden Markov Model uses the word alignments produced by GIZA++ to estimate a maximum likelihood translation table. The two lexical translation tables are English-Igbo table and Igbo-English table. From the two lexical translation tables developed, phrases are extracted from both tables with their equivalent translation and stored in a single file. This single file is called the phrase translation table. The phrase translations in the phrase table are scored to obtain the following computations: inverse phrase translation probability, inverse lexical weighting, direct phrase translation probability, direct lexical weighting, and phrase penalty (always exp(1) = 2.718).

In order to allow for long distance re-arrangement of translation sentences, a reordering table is being generated. The reordering model gives a cost linear to the reordering distance i.e the model is a compilation of the rules of the target language and allows arrangement of the source sentence based those rules to the target sentence during translation. Koehn *et al*. (2007) model was implemented and this model determined the orientation of two phrases based on word alignments at training time, and based on phrase alignments at decoding time. Finally, a configuration file for the decoder is generated with the entire correct paths for the generated model and a number of default parameter settings. This file is called model/moses.ini.

3.4.6 Tuning

Tuning is the final step in the model development which improves the efficiency of the model. The Minimum error rate training (MERT) was used as the tuning algorithm and was implemented in the MOSES toolkit. A total of 15 iterations of MERT was implemented in tuning the model.

3.4.7 Evaluation of Model

The developed MT model was evaluated using Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002) and NIST (National Institute of Standards and Technology) evaluation metrics. Using a test data of 1000 sentences which were acquired from the book of Genesis, the developed MT model was used to translate the test data from both English – Igbo and Igbo – English. From those translations, the BLEU scores and NIST scores were computed. Five sets data which consist of 200 sentences each were also used to compare results of the translation of the developed Machine Translation model and Google Translate model.

4.0 Results and Discussion

Figures 3 and 4 show screenshots for English-Igbo and Igbo-English translations respectively. Table 1show results obtained from automatic evaluation of the developed syntax based MT system. From the table, The BLEU scores for English-Igbo MT and Igbo-English MT are: 0.6439 and 0.7725 respectively while NIST scores are 10.3935 and 11.8984 respectively.

Table 1: Analysis of BLEU and NIST scores of test data

Developed MT Model	BLEU score	NIST score
English – Igbo	0.6439	10.3935
Igbo – English	0.7725	11.8984

It was observed from Table 1 that Igbo-English translations was more accurately performed than English-Igbo and the BLEU and NIST scores computed for the developed MT model generally performs above average since its scores are above the average BLEU and NIST scores. The BLEU scores obtained from the developed system shows that the system has good and fluent translation because research by Klein (2017) revealed that BLEU scores higher than 0.50 shows fluent and good translation. The results from automatic evaluation of the developed system shows that syntax based MT systems outperformed previous Phrase based MT systems (Ayogu et al., 2018).

Table 2 shows the results obtained from comparison of the developed syntax based MT system to Google translate (an existing MT system) for English-Igbo machine translation.

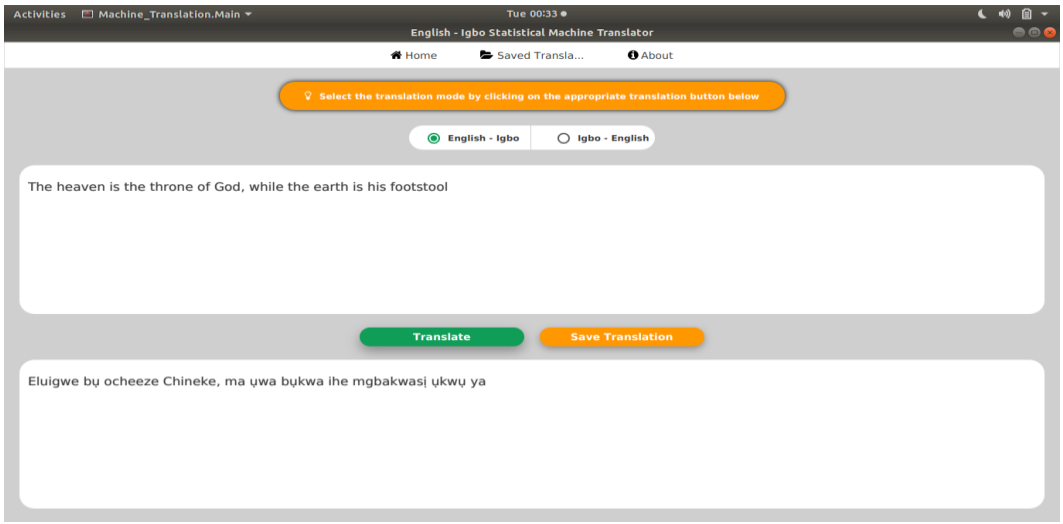

Figure 3: Translation of English to Igbo

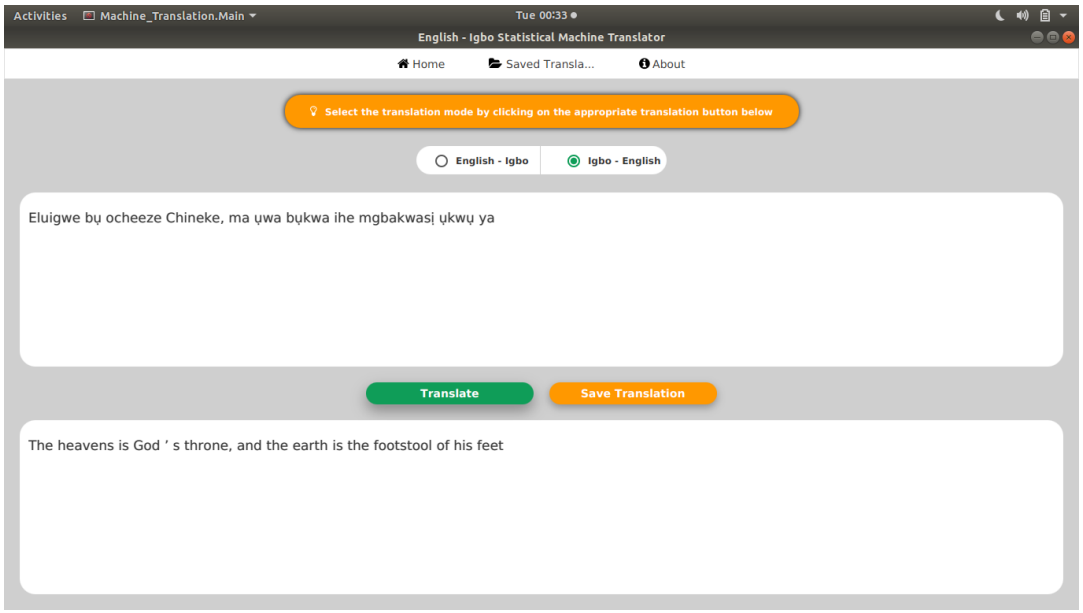

Figure 4: Translation of Igbo to English

Table 2*:* Comparison of the developed MT system and Google Translate for English-Igbo translation

Sentences	English - Igbo		Igbo - English	
	BLEU (Google Translate model)	**BLEU (Developed MT model)**	**BLEU (Google Translate model)**	**BLEU (Developed MT model)**
Test data 1	0.3792	0.6655	0.3617	0.7546
Test data 2	0.3106	0.5779	0.4133	0.7542
Test data 3	0.3140	0.6714	0.3807	0.7882
Test data 4	0.3125	0.6538	0.3447	0.7999
Test data 5	0.3512	0.6549	0.3397	0.7666

From table 2, it was observed that the developed syntax based statistical machine translation system has higher BLEU scores for all sets of test data than Google translate. Also the result shows that the developed model has higher BLEU scores for Igbo-English translations. This shows that the developed MT system produces better translations that has a higher correlation to the reference translation than the existing MT system, Google Translate. Also, the developed system is capable of producing translations that may not 100% follow verbatim to the original target translation, but make translations that depicts the meaning of the original source text.
Table 3 recorded the results obtained from evaluation of the developed MT system and Google translate.

Table 3: **Comparison of NIST scores of the developed MT model and Google Translate model for English-Igbo translation**

Sentences	English - Igbo		Igbo - English	
	NIST (Google Translate model)	**NIST (Developed MT model)**	**NIST (Google Translate model)**	**NIST (Developed MT model)**
Test data 1	6.8715	9.0745	6.4482	9.8079
Test data 2	6.1770	8.3284	7.0374	10.0620
Test data 3	6.2370	9.2405	6.9029	10.4193
Test data 4	6.1136	9.0649	6.5515	10.5304
Test data 5	6.5717	9.1147	6.1911	9.9919

From table 3, it was deduced that the developed syntax based statistical machine translation system has higher NIST scores for all sets of test data than that of Google translate model for Igbo-English translations than English – Igbo translations. Therefore, the results show that the developed MT model produces results that are more informative than the Google translate model, because NIST does not give credit for matching the word order of the reference text. Therefore, the NIST results show that the word order is not taken into consideration, rather the information being passed across is more considered.

5.0 CONCLUSION

Machine translation satisfies a great need for the translation of texts and documents in a faster, more accurate and convenient approach. This work developed a hierarchical phase-based English-Igbo Statistical Machine Translation model. A user interface was developed to serve as a means for providing inputs to the model and also displaying translated outputs of the model. The model was tested and evaluated by computing the BLEU and NIST scores from a set of test data. The developed system was also compared to Google translate and results show that the system outperformed the previous system in the religious domain. Hence, it is deduced in this research that with a moderate amount of corpra, a statistical machine translator can be developed which is capable of providing quality translations on the domain it is trained. It is recommended that future work considers a large vocabulary size in training statistical machine translation models.

REFERENCES

Abikoye, O.C. Ojo, I. M. Akintola A.G. (2017). Text to Text Translation of English Language to Yorùbá Language. *Proceedings of the International Conference on Science, Technology, Education, Arts, Management and Social Sciences (iSTEAMS Research Nexus)*.

Abiola, O.B, Adetunmbi, A.O and Oguntimilehin, A. (2015). Using hybrid approach for English-to-Yoruba text to text machine translation system (proposed)". *International Journal of Computer Science and Mobile Computing, 4*(8): 308-313.

Adebimpe Esan, John Oladosu, Christopher Oyeleye, Ibrahim Adeyanju, Olatayo Olaniyan, Nnamdi Okomba, Bolaji Omodunbi, Opeyemi Adanigbo (2020). Development of a Recurrent Neural Network Model for English to Yorùbá MachineTranslation. *International Journal of Advanced Computer Science and Applications (IJACSA),11(5):602-609 (Indexed in Scopus)* http://dx.doi.org/10.14569/IJACSA.2020.0110574

Agbeyangi, A.O., Eludiora, S.I. and Adenekan, O.A. (2015). English to Yorùbá Machine Translation System using Rule-Based Approach. *Journal of Multidisciplinary Engineering Science and Technology* (*JMEST*), *2*(8): 733-741.

Ayegba, S.F., Osuagwu, O.E. and Okechukwu, N.D. (2014). Machine Translation of Noun Phrases from English to Igala using the Rule-Based Approach, *11*(1):17-26.

Amr Ahmed, G. H. (2010). Syntax-Based Statistical Machine Translation. *Language Technologies Institute*, 1-30.

Amr Ahmed, G. H. (2010). *Syntax-Based Statistical Machine Translation:.* Language Technologies Institute,Language Technologies Institute.

Ayogu, I. I., Adebayo , O. A., & Bolanle , A. O. (2018). Developing Statistical Machine Translation System for English and Nigerian Languages. *Asian Journal of Research in Computer Science*, 1-8.

Berdica, A. (2017). The positive impact of technology in translation. (pp. 1-12). University "Aleksandër Moisiu" Durrës.

Bonet, C. E. (2010). *Statistical Machine Translation A practical tutorial.* Barcelona.

Chéragui1, M. A. (2010). Theoretical Overview of Machine translation. 1-10.

Chinenyeze, E., Bennett, & Taylor. (2019). A Natural Language Processing System for English to Igbo Language Translation in Android. *International Journal of Computer Science and Mathematical Theory ISSN 2545-5699 Vol. 5*, 1-12.

Daniel , B., & Helena , C. (2013). Tree-Based Statistical Machine Translation: Experiments with the English and Brazilian Portuguese Pair. *Learning and Nonlinear Models - Journal of the Brazilian Computational Intelligence Society*, 1-14.

Franz, J. O., Richard, Z., & Hermann , N. (2014). Phrase-Based Statistical Machine Translation. *Human Language Technology and Pattern Recognition* (pp. 1-16). Germany: RWTH Aachen – University of Technology.

GIL, J. R., & PYM, A. (2015). Technology and translation. 1-15.

Gudivada, V. (2018). Natural Language Core Tasks and Applications. 1-26.

Huang, L., Qun , L., & Haitao , M. (2016). Forest-Based Translation. 1-9.

Hutchins, J. (2009). Uses and Applications of Machine. *Principles of machine translation*, 1-64.

Kenji Yamada, K. K. (n.d.). *A Syntax-based Statistical Translation Model.* Marina del Rey: Information Sciences Institute.

Kumar, A., Dhanalakshmi, Kp, S., & Sankaravelayuthan, R. (2014). Factored statistical machine translation system for English to Tamil language. *Pertanika Journal of Social Science and Humanities*, 1-24.

Maxim, K., José, F., & Mark, D. (2016). A new subtree-transfer approach to syntax-based reordering for statistical machine translation. *Statistical Machine Translation*, 2-9.

Mukesh, Vatsa, Nikita, J., & Sumit, G. (2010). Statistical Machine Translation. *DESIDOC Journal of Library & Information Technology, Vol. 30, No. 4*, 1-8.

Ngozi , P. I., & Olivia , E. (2015). Translating technical texts: The Igbo language example. *African Educational Research Journal*, 104-110.

Nihar , S. (2018). Applications of Artificial Intelligence in Neural Machine Translation. *International Research Journal of Engineering and Technology (IRJET)*, 1-3.

Oladosu, J., Adebimpe, E., Adeyanju, I., Adegoke, B., Olaniyan, O., & Omodunbi, B. (2016). Approaches to Machine Translation: A Review. *FUOYE Journal of Engineering and Technology*, 1-8.

Onyenwe, I., Uchechukwu, C., & Hepple. (2014). Part-of-speech Tagset and Corpus Development for Igbo, an African Language. 1-7.

Pa, W. P., Thu, Y. K., Finch, A., & Sumita, E. (2016). A Study Of Statistical Machine Translation Methods For Under Resourced Languages. *Procedia Computer Science*, (pp. 250-257). Yogyakarta,Indonesia.

Rebecca, K., Marina, S.-T., & Philipp, K. (2019). A user study of neural interactive translation prediction. 1-20.

Sani, F. A., Osuagwu, & Njoku, O. D. (2015). Machine Translation of Noun Phrases from English to Igala. 1-11.

Subalalitha, Aarthi, V., & Baqui, B. S. (2018). STATISTICAL MACHINE TRANSLATION FROM ENGLISH TO HINDI. *International Journal of Pure and Applied Mathematics*, 1-8.

T´ynovsk´y. (2008). Hybrid Approaches in Machine Translation. *WDS'08 Proceedings of Contributed Papers*, (pp. 124-128). Prague.

Yulian, H. (2014). Statistical machine translation based on translation rules. *Journal of Chemical and Pharmaceutical Research*, 1-8.

Koehn,P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, Zens, R., Dyer, C., Bojar, O., Constantin, A. and Herbst E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, Prague, Czech Republic**.**