Development of an automated English-to-local-language translator using Natural Language Processing

Iwara ARIKPO[1], Iniobong DICKSON[2]

Abstract— Residents in Nigeria and other multi-lingual English-speaking countries are limited to communicating with English despite existing local languages. Although English as a lingua franca facilitates communication between residents irrespective of tribe, communication challenges still exist between residents with limited English language proficiency and a different local language. This paper reports the development of an automated English-to-local-language translator as a model to bridge the communication gap in Nigeria and other multilingual settings. Object-oriented methodology was used in the design and implementation of the machine translator. Java technology was used for the development the software. Computational linguistics dispositions and potential methods to automate natural language were explored and applied to the development of a machine translator. The translator was modelled using the transfer-based approach to machine translation with the objective of restoring the meaning of an English text in a translated verse. The scope of implementation was English-to-Efik language translation. Five English–Efik generative rules were declared in the translator system's shell. A bilingual dictionary containing 500 words and 25 corpuses was designed to provide optimum direct translations regarding the Efik language. The quality of translation is influenced by the size of bilingual dictionary provided and accuracy of generative rules declared in the system's shell. The software successfully translates English text to Efik using the words and corpuses available in a bilingual dictionary.

Index Terms— translator, language, corpus, multilingual, machine translation, grammar, dictionary

1 Introduction

People's languages are vital to them. Through language, people communicate, share meaning and experience their sense of individual and community identity [1]. In 2012, the United Nations held a forum on "The Study on the role of languages and culture in the promotion and protection of the rights and identity of indigenous peoples". The importance of language is summed up in the following quote: "Language is an essential part of, and intrinsically linked to, indigenous peoples' ways of life, culture and identities. Languages embody many indigenous values and concepts and contain indigenous peoples' histories and development. They are fundamental markers of indigenous peoples' distinctiveness and cohesiveness as peoples" [2]. Nigeria as a multi-lingual country is faced with challenges such as local language revitalisation, language preservation and sharing knowledge and information in pursuit of development goals in rural areas. The cost of human interpretation and translation is high. However, the improvement in computer-aided software engineering tools coupled with availability of cheap technologies such as mobile phones and personal computers has reduced the cost for machine translation. Computer technology can be a powerful tool for providing materials in local languages to foster participation and inclusion of minorities in national development. Technologies that offer speech-to-speech or text-to-text communication from one language to another are one of the many ways that residents in multi-lingual societies can bridge communication gaps. For Nigeria to achieve her national goals, there needs to be effective communication among the diverse people [3], whilst her different local languages and cultures are preserved. This paper reports the development of an automated English-to-local-language translator, with the aim of providing solutions to language barriers and improving the understanding of how technology can be used to bridge the communication gap among residents in Nigeria.

2 Literature Review

Locally in Nigeria, not so much has been done in machine translation of local languages. However, it is worth reflecting on some previous efforts. [4] developed a machine translator for the English and Yoruba languages using classic syntactic and semantic analysing algorithms. [5] described a Statistical Machine Translation (SMT) system that translates English sentences to Yoruba sentences. The resulting software provides tools to tackle the problem of language translation between Yoruba and English language. The software employed a machine translation paradigm where translations are generated on the basis of statistical models whose parameters were derived from the analysis of bilingual text corpora. [6] gave an account of Yoruba text-to-speech (TTS) system development using the concatenation method. The

analysis of the results their work showed that, 70% respondents accepted its usability. In 2015, a few Natural Language Processing research projects aimed at producing automated language translators and providing unique and important insights into the Natural Language Processing (NLP) of African Languages commenced in Nigeria: These projects were focused on the development of automated machine translation software for Igbo - English as well as Yoruba-English, and the development of a functional corpus of computer-readable Yoruba texts in standard orthography and a Statistical Language Model (SLM) of Yoruba. These projects were sponsored by the African Languages Technology Initiative (ALT-I) – *a research and development agency with a mission for taking African cultures into the Knowledge era* (http://www.alt-i.org/?page_id=31). Generally, there are many programs now available that are capable of providing useful output within strict constraints. Most of these programs are available online, such as Google Translate and the SYSTRAN system which powers AltaVista's BabelFish [7].

3 Materials and Methods

There are several approaches to automating the language translation process. [8] classify machine translation approaches as: dictionary-based, rule-based, knowledge-based and corpus-based. Other approaches (e.g. interlingua, transfer, statistical, context-based, etc.) exist as subsets of these approaches classified by [8]. The objective of translation is to restore the meaning of an original text in the translated verse. In this study, the transfer and corpus-based approaches to machine translation were fused to develop an automated English-to-local-language translator. The transfer model aims at creating linguistic homogeneity across the globe. [8] describe the transfer model as "a model which transforms source language into an abstract, less language-specific representation. An equivalent representation (with same level of abstraction) is then generated for the target language using bilingual dictionaries and grammar rules". All natural languages have different grammatical structures, although the magnitude of difference in grammatical structure between two languages is relative, that is, some language pairs might have a low difference magnitude of grammatical structure while others have higher. Translating languages with low difference magnitude is little challenge, but the reverse poses a great challenge. Fortunately, a hybridisation of the transfer-based approach with the corpus-based approach could reduce the complexity of translating languages that have a high difference magnitude of grammatical structure. The developed translator is a Java-based desktop application that uses the Stanford CoreNLP (Core Natural Language Processing) Application Programming Interface (API) to analyse English texts morphologically. The Stanford CoreNLP API is a Java (or at least JVM-based) annotation pipeline framework, which provides most of the common core natural language processing (NLP) steps, from tokenization through to coreference resolution [10].

3.1 System Features

The automated English to local language translator contains the following key features:

1. Morphological, syntactic and semantic analysis of English texts.
2. Transform English texts grammatical structure to fit the Efik grammatical structure.
3. Replace source text (word or corpus) with Efik synonyms in the bilingual dictionary.
4. Accept and store English texts and its respective Efik translation in a dictionary. This feature enables the system to learn new translations and correct erroneous translations.

3.2 System Architecture and Design

The proposed automated English-to-local-language system architecture is designed according to the transfer-based approach to automating translation. At the first layer of the architecture is a natural language processing tool that performs morphological analysis: sentence tokenization, part-of-speech tagging, phrase formation and figure of speech tagging. The second layer of the architecture comprises of a grammar generator that is responsible for converting English language structure to target local language structures. At the third layer of the application, results produced by the grammar generator are mapped with matching terms in the bilingual dictionary. The architectures for a transfer-based model and automated English-Efik translator are respectively presented in Figures 1 and 2, while the Use case and Activity diagrams are presented in Figures 3 and 4, respectively.

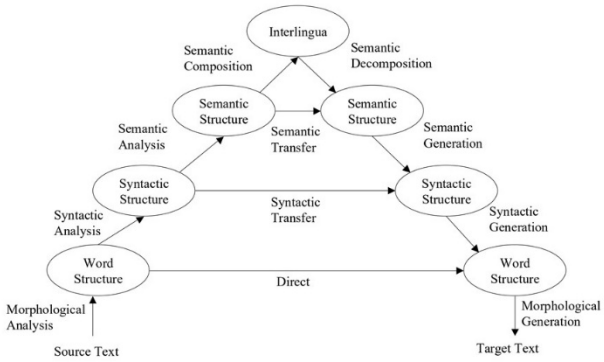

Figure 1: Architecture of transfer-based model (adapted from [9])

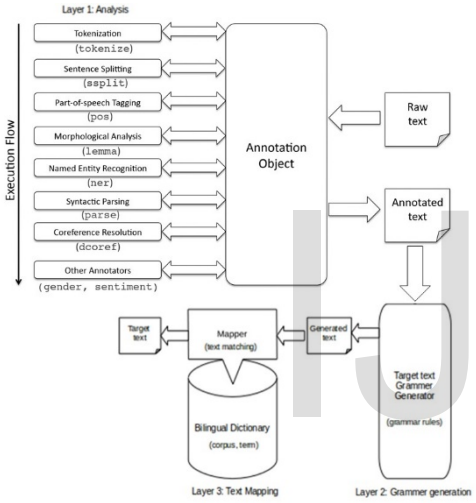

Figure 2: Architecture of the automated English–Efik translator (*Adapted from [10]*)

The system's design facilitates user interaction by providing a menu for text document translation, sentence translation, dictionary customization and corpus correction.

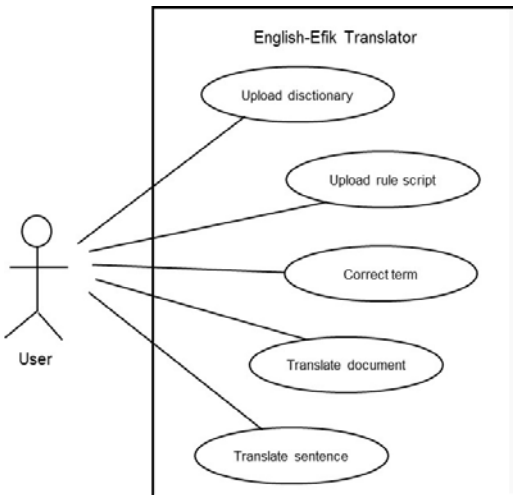

Figure 3: Use case diagram of automated English–Efik translator

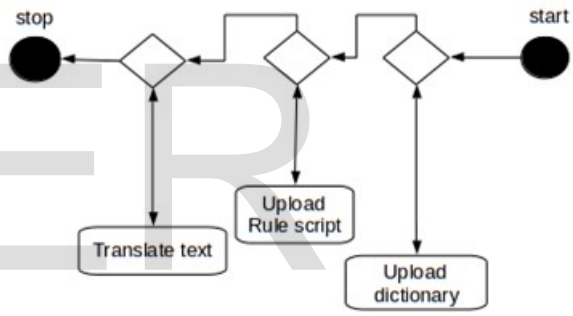

Figure 4: Activity diagram of automated English–Efik system

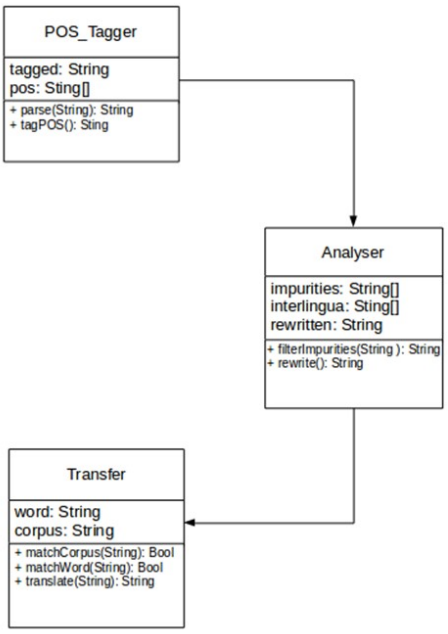

Figure 5: Class diagram of automated English-Efik translator

4. RESULTS

The system provides unidirectional translations for sentences and text documents. Sentence translations are provided on text entry. Tests were performed on the system in two modes using 15 well-formed English simple and complex sentences. In the first mode a text document containing these sentences was created and uploaded to the system while each sentence was translated interactively in the second mode. Results showed that, simple sentences generated optimal translations whilst complex sentences generated very poor translations. The system produced accurate translations for source sentences in which all corpuses and words were defined "as is" in the bilingual dictionary. Corpuses that were not defined in the bilingual dictionary, but had an equivalent analysis grammar defined in the systems rule engine translated accurately. Words and corpuses that were not defined in the bilingual dictionary could not be translated and were returned "as were" in the source text. These results can be credited to the hybridisation of the transfer and corpus fused approach to machine translation. Figures 6 – 9 are screenshots of the results from the English-Efik language translator.

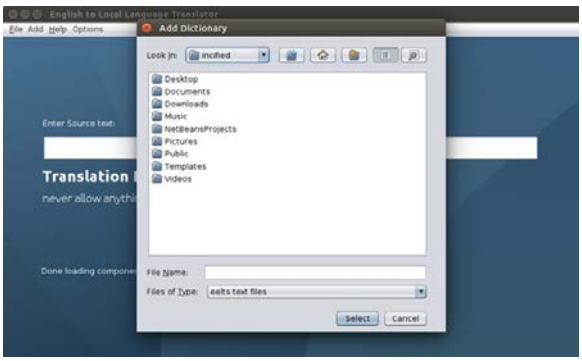

Figure 6: Panel for dictionary addition

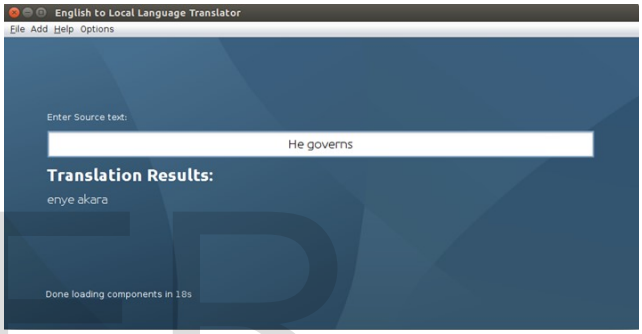

Figure 7: Screen 1 after translation

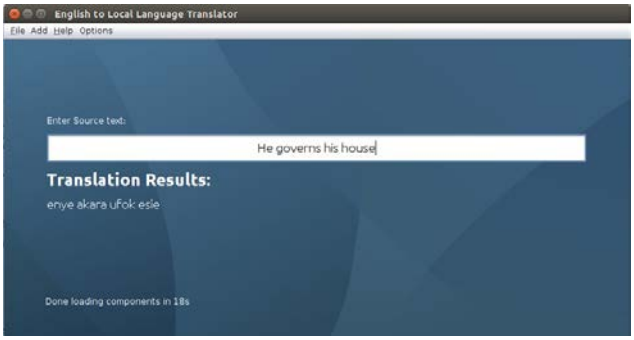

Figure 8: Screen 2 after translation

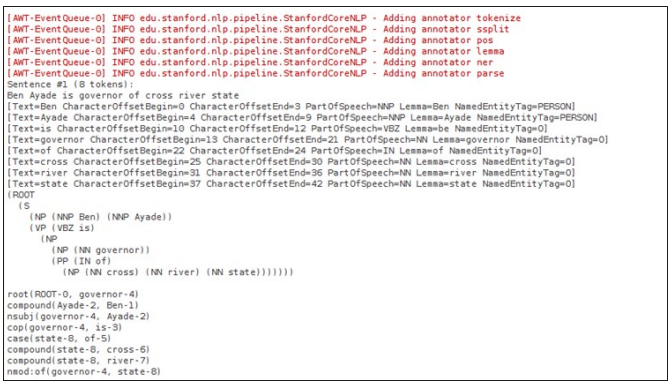

Figure 9: Screen showing part of the Java code segment

5. Discussion

Quality translations are dependent on experience; which is very costly because, experience is influenced by time. However, machines are capable of reducing the time spent on performing certain activities viz. the time spent learning different languages, and the time spent translating one learned language to another. The results from this study in line with similar studies (e.g. [5]) imply that the developed translator can provide fair translations for non-complex sentences. The non-availability of English-to-Efik translators presented difficulties in comparing the developed system's performances. Due to ambiguity, it is impossible to provide all the possible corpuses regarding a particular natural language for a translator. However, an intelligent translator could learn translation rules, whence translating. Developing translators as intelligent systems capable of learning, may be the key to bridging the communication gap between residents of bi-lingual and multi-lingual countries.

6. Conclusion

From this study, it can be concluded that, simple sentences produced fair translations because the local language grammatical analysis was simple sentence-based. However, complex and composite sentences produced poor results. The system can be improved upon by providing a means of breaking complex and composite sentences into simpler forms before translation, or simply improving its local-language grammatical analysis. Since the software produces fair translations, the automated English-to-local-language translator, when expanded may be used as a tool to facilitate local language learning outside schools.

Acknowledgment

The authors wish to thank Professor Offiong A. Offiong of the Department of Linguistics, University of Calabar, who provided the expert information during the development of the translator.

References

[1] UNESCO, (2012). *Why Language Matters for the Millennium Development Goals,* United Nations Educational, Scientific and Cultural Organisation Bangkok Asia and Pacific Regional Bureau for Education. ISBN 978-92-9223-387-7

[2] United Nations General Assembly, (2012). *Study on the role of languages and culture in the promotion and protection of the rights and identity of indigenous peoples,* Expert Mechanism on the Rights of Indigenous Peoples, Fifth Session, A/HRC/EMRIP/2012/3.

[3] Odunola A. H. & Kolade A. (2012). Empowering national development in Nigeria through appropriate national communication policy. Kuwait Chapter of Arabian Journal of Business and Management Review, 2(3).

[4] Awofolu, O. & Malita, M. (2002). "The making of a Yoruba-English machine translator", *Journal of Computing Sciences in Colleges*, 17(6), 236–237.

[5] Folajimi, Y. O. & Isaac, O. (2012). *Using Statistical Machine Translation (SMT) as a Language Translation Tool for Understanding Yoruba Language*. EIE's 2nd Intl' Conf.Comp., Energy, Net., Robotics and Telecom.| eieCon2012 (pp. 86–91).

[6] Afolabi, A., Omidiora, E., Arulogun, T. (2013). "Development of Text to Speech System for Yoruba Language", *Innovative Systems Design and Engineering*, 4(9), 1–7.

[7] Osunade, O., Dawodu, D., Phillips, O. F. (2015). "A Simple Data Driven Yoruba Language Dictionary", *Journal of Literature, Languages and Linguistics*, 10, 96–102.

[8] Tripathi, S., & Sarkhel, J. K. (2010). "Approaches to machine translation", *Annals of Library and Information Studies*, 57, 388–393.

[9] Dorr, B. J., Hovy, E. and Levin, L. (2004). "Machine Translation: Interlingual Methods", *Encyclopedia of*

Language and Linguistics 2nd edition ms. 939, Brown, Keith (ed.).

[10] Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014, June). The Stanford CoreNLP Natural Language Processing Toolkit. In ACL (System Demonstrations), 55–60.

AUTHOR DETAILS:

Corresponding Author:

Iwara ARIKPO[1]: Department of Computer Science, University of Calabar, Nigeria. Email: iwara.arikpo@unical.edu.ng; iiarikpo@gmail.com

Iniobong DICKSON[2]: Department of Computer Science, University of Calabar, Nigeria.