Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 8, August 2015, pg.308 – 313

RESEARCH ARTICLE

USING HYBRID APPROACH FOR ENGLISH-TO-YORUBA TEXT TO TEXT MACHINE TRANSLATION SYSTEM (PROPOSED)

Abiola O.B[1], Adetunmbi A.O[2], Oguntimilehin A[3]

[1,3]Department of Computer Science, Afe-Babalola University, Ado-Ekiti, Nigeria

[2]Department of Computer Science, Federal University of Technology, Akure, Nigeria

[1]adeoyetoyin@yahoo.com; [2]bayoadetunmbi@gmail.com; [3]ebenabiodun2@yahoo.com

Abstract—*Machine translations mainly deal with the transformation from one natural language to another. There are only few works on African languages especially the Yoruba language. It is observed that researchers that have attempted translation work in this area have approached it with a single approach to machine translations and they have not been able to achieve much as a result of this. This research has found that a more robust and sensible translation is best achieved when more than one approach is used. Therefore our proposed hybrid model leverages the strengths of statistical and rule-based approaches to achieve a better and more robust translation system for English to Yoruba translations which will be resistant or impervious to failure regardless of user's inputs*

Keywords— *Machine Translation, Hybrid, SMT, RBMT, Yoruba*

I. INTRODUCTION

Machine translation of natural languages has been considered as a very difficult task. It can be perceived as the simple substitution of words in one natural language for words in another. Yet it is not so simple because of the complexity of some natural languages. Many words have various meanings in some languages and so they can be translated in different ways. Also, sentences might be ambiguous and have various meanings. Grammatical relations can vary depending on the languages, and translating sentences from languages having different relations means reformulating them. Besides, problems due to the associated world knowledge may be encountered and these are usually difficult to solve. [7], [13]. Machine translation has been defined as a subfield of computational linguistics that investigates the use of computer software to translate text or speech from one natural language to another. [11]. Different approaches to machine translation have been proposed in history and each of these approaches has its own benefits and drawbacks as reviewed in [1]. Though the various benefits and drawbacks of each was not fully discussed in their work. They identified the following approaches to machine translations: direct-based approach, rule-based approach which is further divided into (transfer- based and Interlingua respectively), corpus-based which also comprises of (statistical-based and example-based) and off course the knowledge-based approach as well.

Our work is based on building a hybrid system for effective and efficient translations between English and Yoruba language by combining two approaches: the rule-based and statistical-based approaches so as to achieve better and more robust translations between these languages. However, there is need to highlight some of the strengths and drawbacks of these two approaches and see a way of retaining and integrating their advantages and get rid of their disadvantages to achieve our objectives. The rule-based machine translation approach involves the application of morphological, syntactic and semantic rules in the analysis of the source language text and the synthesis of the target-language text. It can further be divided into transfer-based and Interlingua machine translations respectively. In rule-based approach, a database of translation rules is used to translate text from source to target language. This approach deals with the word-order problem and since it uses linguistic knowledge, the produced errors can be traced. [14], [10]. RBMT parses the source text and produces an intermediate representation which may be a parse tree or some abstract representation. The target text is generated from the intermediate representation. These systems rely on the specification of rules for morphology, syntax, lexical selection, transfer, semantic analysis and generation. It identifies the relationship between source-language words and their structural representations as affirmed in [1].

There are three major modules or phases in a rule-based system according to [14]. They are:

i. **Analysis Phase:** this is the phase where the source language structure is produced. That is, the source language text is transformed into abstract source language representation. Linguistic information about the sentence structure and each word in the text is obtained at this phase.
ii. **Transfer Phase**: is the phase where the source to target text transfer is carried out with the help of a bilingual dictionary. This dictionary is stored in the transfer module for easy accessibility when the system is running.
iii. **Synthesis or Generation Phase:** this is the phase where target language representation is re-organized and cleaned. This phase is used to generate target language text using target level structure. [14].

The rule-based approach has a modular structure and it easily handles ambiguities that carry over from one language to another. Also, in this approach, no necessary equivalence between source and target language is required and it deals well with word-order problem. However, some of the source text meaning can be lost in the translation with rule-based and it is a must to construct transfer rules for each language pair which is usually costly and sometimes there can be syntactic mis-match, lower coverage and inflexibility to be robust. Above all, rule-based approach fails to achieve satisfactory performance for large scale applications as affirmed by [8].

The statistical machine translation (SMT) is part of corpus based machine translation and it requires less human effort to undertake translation. SMT is a machine translation paradigm where translations are generated on the basis of statistical models. These statistical models parameters are derived from the analysis of bilingual text corpora. [9]. Statistical-based machine translation uses purely statistical-based methods in aligning the words and generation of texts. It is based on the view that every sentence in a language has a possible translation in another language. That is every sentence in the target language is a possible translation of the input sentences. SMT automatically align words and phrases within sentence pairs in a parallel corpus. Probabilities are determined automatically by training a statistical model using the parallel corpus. The key motivation behind this approach is to reduce rate of errors. [14], [9]. The general idea in SMT system is that the translation will be from the most likely translated word. This approach consists of three different models: The language model (LM) which computes the probability of the target language 'T' as probability P (T). The translation model (TM) which computes the conditional probability of target sentences given the source sentence. P (T/S). Decoder maximizes the product of the language model and the translation model probabilities. [11].

Statistical approach has a way of dealing with lexical ambiguities, it can deal with idioms that occur in the training data, it requires minimal human efforts, minimal linguistics knowledge and can be created for any language pair that has enough training data. This approach can prototype a new system quickly at a very low cost and the knowledge acquired is always consistent because all the data in the corpus are jointly considered during acquisition process and also linguistics uncertainty problems is resolved in this approach by a solid mathematical basis. [8]. However the SMT approach does not explicitly deal with syntax, knowledge is represented shallower and quality and fluency are worse than rule-based. Some translations work have been done for English to Yoruba languages using either of these approaches but only little has been achieved with a single approach because of the structural, syntactic and grammatical differences between the two languages. However, hybrid approach leverages the strength of statistical and rule-based translation methodologies and this approach is proposed in this work to improve translations from English to Yoruba. This approach involve either rules post-processed by statistics: in which translations are performed using rules based engine and statistics are used in an attempt to adjust and correct the output from the rules engine or, statistics guided by rules in which rules are used to pre-process data in an attempt to better guide the statistical engine [1].

The motivation for developing hybrid machine translation systems in this work stems from the failure of a single technique to achieve a satisfactory level of accuracy in the translations between the two languages. Many hybrid machine translation systems have been successful in improving the accuracy of translation works for some languages. There are several popular machine translation systems which employ hybrid methods among them are PROMT, SYSTRAN and Asia Online.

TABLE 1:
SUMMARY OF THE ADVANTAGES AND DISADVANTAGES OF RULE-BASED AND STATISTICAL-BASED APPROACHES

Approach	**Advantages**	**Disadvantages**
Rule-Based	• It has modular structure • It easily handles ambiguities • No necessary equivalence between SL and TL is required • Deals with word-order problem • Easy to build an initial system • Based on linguistic theories	• Some source text meaning can be lost in the translation • Must construct transfer rules • Sometimes there is syntactic mismatch • Costly in terms of formulating rules • Lower coverage and inflexible to be robust • Fails to achieve satisfactory performance for large scale application.
Statistical-Based	• Deals with idioms that occur in the training data • Requires minimal human efforts • Can be created for any language pair with enough training data • Language independent • Can prototype a new system quickly at a very low cost • Resolves linguistics uncertainty problems by a solid mathematical basis. • Model is mathematically grounded • Extract knowledge from corpus	• Does not explicitly deal with syntax • Knowledge is represented shallower than in rule-based • Quality and fluency are worse than rule-based • Neither predictable nor consistent

II. STRUCTURE OF THE YORUBA LANGUAGE

Yorùbá, (native name èdè Yorùbá, ‘the Yorùbá language') is a dialect continuum of West Africa with over 50 million speakers. The Yorùbá dialect continuum itself consists of various dialects. The Yor̀ubá people originated from the Western Nigeria and the places where the language is spoken are termed ‘Ilè Yorùbá meaning the Yor̀ubá land. The various Yorùbá dialects in the Yorùbá land of Nigeria can be classified into three major dialect areas which are: North-West Yorùbá (NWY) which include Abéòkúta, Ìbàdàn, ọ̀yọ́, Ògùn and Lagos (Èkó) areas, Central Yor̀ubá (CY) which include Ìgbómìnà, Yàgbà, Ifè, Èkìtì, Àkúré, and Ìjèbú areas and South-East Yorùbá (SEY) which include Òkìtìpupa, Oǹdó, ọ̀wọ̀, Ságámù and parts of Ìjèbú. The language has its origins in the Yor̀ubá people, who are believed to be the descendants of Òdùdụwà’, the son of a powerful god called ‘Olódùmarè’. They refer to themselves as ‘Omo Òdùduwà’, which means Òdùduwà’s children. [3].

Yor̀ubá is a tonal language with three level tones: High (Ohùn òkè), Low (Ohùn Ìsàlẹ̀) and Mid (Ohùn àárín), represented with [´], [`] and [ˉ] respectively. Every syllable must have at least one tone; a syllable containing a long vowel can have two tones. The three level tones determine the meanings that each word has in Yorùbá. For example, a form that has the same form of vowels and consonants can have different meanings depending on the tones that it has. That is the tonality of a word can totally alter the meaning. The following examples present all the three tonalities (i) Òjó ‘personal name’ (ii) Òjò ‘rain’ (iii) Ojo ‘cowardice’. Same word with different meanings which are differentiated with tonal marks respectively.

(i) Igba ‘two hundred’(ii) Igbá ‘calabash’ (iii) Ìgbà ‘time’ (iv) Ìgbá ‘garden egg’ (v) Igb̀a ‘climbing rope’[3]. Same word with different meanings which are differentiated with tonal marks respectively.

Yor̀ubá has 25 alphabets the Latin letters *c, q, v, x, z* are not used. Two types of sounds are identifiable in Yor̀ubá language: they are the vowel sounds and the consonant sounds respectively. The consonant system consists of eighteen (18) phonetic consonants which are: “ b, t, d, k, g, p[kp], gb, f, s, ṣ[ʃ], h, j[dz], m, n. r, l, y[j], w”. Consonants are sounds produced by obstructing the air flow totally or partially at some point in the track. The standard Yor̀ubá vowel system consists of seven oral vowels and five nasal vowels. All Yor̀ubá vowels are voiced because; the vocal cord is vibrated whenever they are pronounced. There are seven oral vowels in Yoruba: a, e, ẹ[ɛ], i, o, ọ[ɔ] and u. The nasal vowels are “in ĩ, ẹn[ɛ̃], an [ã], ọn [ɔ̃] and un [ũ]” as affirmed by [3]. As a result of the differences in the syntactic and grammatical features as well as the vocabularies between English language and Yoruba, there is need for researchers working in this area to have a profound understanding of their structures so as to achieve effectively translations.

III.PROBLEM DEFINITION OR MOTIVATION

Despite decades of efforts on Machine translations, Fully-automatic general purpose high quality machine translation system (FGH-MT) is extremely difficult to build. In fact, there is no system in the world of any pair of languages which qualifies to be called FGH-MT as affirmed by [16]. Machine Translations in European and Asian languages have received a considerable amount of research attention, but works on African languages especially the Yoruba language is rare. In Nigeria, only few people speak more than one out of the three major indigenous languages which brings about communication barrier since none of these languages is being elevated to the status of National language. According to [6], the dominance of the English language in Nigeria is quite overwhelming. This can be seen in practically all domains. Although a language only dies when nobody speaks it any more, Yorùbá is yet to die even though people are still speaking it, but the threat of extinction is still solidly there. There is need to fully address problems of English to Yoruba machine translations**.** Although, automatic translation between languages which are morphologically rich and syntactically different like English and Yoruba is generally regarded as a complex task. But the question is, what is the success rate in previous translation works or attempts on these two languages and what can be done to improve on the existing work in this area. Certain overt and practical measures have been taken to promote the Yorùbá language over the years but none of these measures have yielded good result yet. This research work provides one of the ways out to prevent loss of the Yor̀ubá language in today's flow of globalization by providing added value to the language and also enhance better relationship between visitors, learners and dwellers to communicate effectively and accurately thereby solving major challenges facing the language.

Researchers have tried to use single approach to machine translations in dealing with English to Yoruba and vice-versa but we are motivated in this work by the failure, inconsistency and inflexibility of the rule-based approach to attain satisfactory performance for large scale application, the shallower representation of knowledge and lower quality and fluency of the statistical machine translation approach, also by the failure of Google translate to translate correctly sometimes and therefore the need to have a more robust and sensible system which will be resistant or impervious to failure regardless of user's inputs by combining two approaches in a hybrid model so as to leverage the strengths of statistical and rule-based approaches to achieve a better and more robust translation system for Yoruba language.

IV.REVIEW OF RELATED WORK

Web-Based English to Yoruba Noun-Phrases Machine Translation System was proposed by [3]. She was motivated by the need to reduce the extinction threat of the Yoruba language by providing a global platform for the language and the need to have an automated machine translation system for Yoruba language. Rule-based approach was used and the formulated rules were specified using context-free grammar and finite state automata was also used to formulate computational model and for recognizing the grammar of the language. A web-enabled platform for translating English noun-phrases to Yoruba was formulated. The approach is rule-based and as a result, there is lower coverage and the system is inflexible to be robust and it is also limited to noun-phrases other phrases that make-up a complete sentence were not considered in the work.

English to Yoruba Translation system using rule-based approach to machine translations was developed by [14]. He was motivated with the need to contribute to knowledge in machine translations by experimenting with the Yoruba language, and the need to address problems of English to Yoruba machine translator and make information available to a global audience and the fact that some problems of English to Yoruba machine translator is yet to be fully addressed. Text corpora were collected from home domain, context-free grammars were use to model the two languages, re-write rules and parse tree were also used. Automata theory was used to model the computational problem underlining the translation process. The study provides an exposition to the process underlining English to Yoruba text translation. Digital database was developed from the study which will be useful for further research. The approach is rule-based. The system can only work for simple sentences with one verb. The issue of split-verbs, interrogative numerals and complex sentences were not addressed.

Using statistical machine translation as a language tool for understanding the Yoruba language was by [16]. They were motivated by the fact that a lot of research has been done on machine translations but little or no attention has been paid to local languages of developing countries like Nigeria and the need to provide tools that could tackle the problem of language translation between English and Yoruba, lack of parallel corpus for English to Yoruba and the lack of SMT translator for English to Yoruba. In this work, translations were generated on the basis of statistical models whose parameters were derived from the analysis of bilingual text corpora. Language model and translation model for the system involved the use morphological analyzer. Lexical dictionaries were also used. English to Yoruba Bible was used for corpus and bilingual text aligner was also used. Moses open source toolkit was used as decoder. The system was evaluated with BLEU and NIST metrics. English to Yoruba parallel corpus was created and an SMT translator that performs translations from Yoruba to English and vice versa was developed. The approach is statistical-based hence; there is shallower representation of knowledge and lower quality and fluency.

A morphology lexical analyzer for Yorùbá language was developed by [4]. A rule-based approach to Computational model which uses grammatical rules was used to construct a morphological lexical analyzer which breaks words into tokens of its parts of speech. However, it is limited to the fact that it can only break words into tokens of its parts of speech and a well sophisticated lexicon with part of speech tags is needed for effective implementation of the system. The system cannot handle multi-word expressions like phrase-to-phrase and the fact that it is rule-based, the coverage is low and it fails to achieve satisfactory performance for large scale application

An attempt was made to develop a computational model of Yorùbá Morphology in [5]. Rule-based approach was used for morphological analysis and finite-state automata was used to internally represent morphological corpus. The morphological analysis was performed by parsing of the input word through the finite-state network. However the corpus is restricted to Yorùbá prefixes, infixes, verbs and nouns. It can only give representation of the major ways of Yorùbá word forms and the system is only good for a beginner.

It is obvious that all these works are at their infact stages and the researchers emphasized and encouraged the need for further research on the subject matter. Most of the existing translation works on Yoruba language are either rule-based or statistical in which accuracy and fluency etc are major drawbacks. However, achieving a more robust and efficient system especially for a rich language like Yoruba is essential and that is the motivation for our proposed hybrid approach.

V. THE PROPOSED METHODOLOGY FOR THE SYSTEM

In this paper, a Hybrid Machine Translation (HMT) approach is proposed which is the combination of rule based and statistical technique for translating texts from English to Yoruba language. The methods that will be adopted will include the following among others:

Understanding the differences in the grammatical structures of the two languages which will involve morphological, syntactic and semantic analysis of the two languages. Rule structures for every possible sentence in the language will be developed which will cover all aspects of the source and target languages to ensure efficient translations. The rules will be specified with context free grammar and computational model for the proposed system will be formulated using finite state automata. Parts of speech will be divided into their own subcategories and simplifying and segmenting of the input language text will be used to improve the quality of the system. Parse structure of a sentence will be derived using the Stanford parser and Yoruba text editor will be developed and use to analyse Yoruba words and bilingual lexicon will be expanded and use for effective translations. The system will be splited into two phases:

i. Training phase
ii. Translation phase

During training phase, the following will be done:

- Document collection which is the collection of texts which will form the corpus
- Building the language model for the target language from the monolingual corpus i.e., Pr(e)
- Building the translation model from the target language to the source language i.e., Pr(f|e)

The translation phase will be the decoding phase in which heuristic search procedure will be used to find a good translation of the given source text. Translations will be performed using a rule-based engine. Statistics will then be used in an attempt to adjust and correct the output from the rules engine. Then the errors in the translated sentences will be corrected by applying a statistical technique. A language model will give the probability of a sentence. The probability will be computed using n-gram model**.** Language model can be considered as computation of the probability of single word given all of the words that precede it in a sentence. The goal of SMT is to estimate the probability of a sentence. A sentence will be decomposed into the product of conditional probability by using chain rule.

The probability of a sentence P(S) will be broken down as the probability of individual words P (w)

In order to calculate sentence probability, it will be required to calculate the probability of a word, given the sequence of words preceding it.

$$\begin{aligned} P(S) &= P(w_1\, w_2\, w_3 \ldots w_n) \\ &= P(w_1)\, P(w_2/\, w_1)\, P(w_3/\, w_1\, w_2)\, P(w_4/\, w_1\, w_2\, w_3)\ldots \\ &\quad P(w_n/\, w_1\, w_2 \ldots w_{n-1}) \end{aligned}$$

An n-gram model will w simplifies the task by approximating the probability of a word given all the previous words. An n-gram of size 1 is referred to as a unigram, size 2 as bi-gram or diagram, size 3 as trigram and size five or more as n-gram. Language model will compute probability of target language sentences. Translation model will calculate the probability of target sentences given the source sentence and decoder will maximize the probability of translated text of target language. The system will be evaluated on the parameters of fluency and adequacy respectively. Both manual and automatic evaluation technique will be employed to measure the efficiency of the new approach. Probabilities will be computed using chain rule, Bayes theorem, Hidden Markov

Model.etc The hybrid model will be implemented using Pyton Programming Language. BLEU and NIST will be used for experimental evaluation. The expected contribution of the research to knowledge is to show the effectiveness of the hybrid approach in machine translations of English to Yoruba to achieve a more robust and effective translations as against the single approach of rule-based or statistical based approach respectively

VI. CONCLUSION

Accuracy and speed of translation are two main measures to evaluate the performance of machine translation tools. From reviews, language is evolutionary in nature; hence it is difficult to say that one approach would be sufficient to handle the translation process. Linguistics irregularities, ambiguities, lack in universality of grammar and lexicon are some of the reasons behind the failure of systems to achieve 100% accuracy in machine translations. Various approaches to machine translations have been proposed for the translation of English to Yoruba but each of this approach has its major benefits and drawbacks. In order to enhance translation work in Yoruba language, there is need to integrate the advantages of these approaches and get rid of their disadvantages in a hybrid system by combining at least two approaches to achieve more robust systems for effective usage and dissemination so as to protect the loss of the language in today’s flow of globalization.

REFERENCES

[1] Abiola O.B, Adetunmbi A.O, Oguntimilehin A (2015) “Review of the Various Approaches to Text to Text Machine Translations” International Journal of Computer Applications. Vol120 No 18, June, 2015, pp 7-12. ISSN: 0975-8887.

[2] Abiola O.B, Adetunmbi A.O, Oguntimilehin A (2013) “Computational Model Of English To Yoruba Noun-Phrases Translation System” FUTA Journal of Research in Sciences ISSN: 2315-8239, pp 34-43, Volume 9 No 1.

[3] Adeoye (2012) “Web-Based English to Yoruba Noun-Phrases Machine Translation System” Master’s Thesis, Federal University of Technology (FUTA), Akure.

[4] Aladesote I, Olaseni O.E, Adetunmbi A.O, Akinbohun.F (2011) “A Computational Model of Yoruba Morphology Lexical Analyzer” International Journal of Computational Linguistics (IJCL), Vol 2, Issue 2, pp 37-47.

[5] Awoyele. I.(2008) “Computational Model of Yoruba Morphology” PGD Thesis, Federal University of Technology, Akure, Ondo State.

[6] Felix Abidemi Fabunmi and Akeem Segun Salawu (2005) “Is Yoruba an Endagered Language” Nordic Journal of African Studies. Vol 14, No 3, pp 391-408.

[7] Hany Hassan (2009) “Lexical Syntax for Statistical Machine Translation” Ph.D Thesis

[8] Jorg (2009) “Machine Translation: Rule-Based Machine Translation and Machine Translation Evaluation” Department of Linguistics and Philosophy, Uppsala University.

[9] Latha R.Nair and David Peter (2012) “Machine Translation Systems for Indian Languages” International Journal of Computer Application” Vol 39, No 1 ISSN 0975-8887, pp 25-31. Dublin City University

[10] Michel Simard, Nicola Ueffing, Pierre Isabelle and Roland Kuhn (2007) “ Rule-Based Translation with Statistical Phrase-Based Post-Editing” Proceedings of the Second Workshop on Statistical Machine Translation, pp203–206, Prague, June 2007 Association for Computational Linguistics.

[11] Nakul Sharma (2011) “English to Hindi Statistical Machine Translation System” Masters Thesis, Tharpa University, Indian.

[12] Nithya B. and Shibily Joseph (2013) “A Hybrid Approach to English to Malayalam Machine Translation” International Journal of Computer Application, Vol 81 No 8. ISSN 0975-8887, pp 11-15. www.ijcaonline.org.

[13] Pankaj Kumar and Er.Vinod Kumar “Statistical Machine Based Punjabi to English Transliteration System for Proper Nouns” International Journal of Application or Innovation in Engineering and Management (IJAIEM). Vol 2, Issue 8, ISSN 2319-4847, pp318-321.

[14] Sneha Tripathi and Juran Krishna Sarkhel (2010) “Approaches to Machine Translation” Annals of Library Vol 57, pp 388-393.

[15] Safiriyu Ijiyemi Eludiora (2014) “Development of an English to Yoruba Machine Translation System” PhD thesis, Obafemi Awolowo University, Ile-Ife, Osun State.

[16] Vishal.G. and Gupreet.S.L (2010) “Web-Based Hindi to Punjabi Machine Translation System” Journal of Emerging Technologies in Web Intelligence, Vol.2, No 2, May, 2010. Pp 148-151. www.academypublisher.com/ojs/index.

[17] Yetunde Folajimi and Omonayin Isaac (2012) “ Using Statistical Machine Translation as a Translation Tool for understanding Yoruba Langauge” EIE’s 2nd Conference. Comp, Energy Net Robotics and Telecom./ eieCon 2012.