English to Yorùbá Machine Translation System using Rule-Based Approach

Agbeayangi, Abayomi O.  
Department of Computer Engng.  
Moshood Abiola Polytechnic  
Abeokuta, Ogun State-Nigeria.  
abayomi_sola@yahoo.co.uk  
(correspondent author)

Eludiora Safiriyu I.  
Dept. of Computer Sci. & Engng  
Obafemi Awolowo University  
Ile-Ife, Osun State-Nigeria  
eludioraomolola@gmail.com

Adenekan, Olujide A.  
Department of Computer Engng.  
Moshood Abiola Polytechnic  
Abeokuta, Ogun State-Nigeria.  
adenekeanolujide@yahoo.com

Abstract - Rule-based approach is a good approach for Machine Translation System used for language with lots of grammar which Yorùbá language is one. In this paper we present Rule-based approach to Yorùbá Machine Translation System. The popularity of Yorùbá language among the three main languages in Nigeria calls for the need to computerise the language. Transfer Rule-Based Machine Translation is use in the development of the System. It was used because it allows us to use manual tagging of the part of speech (POS). Rewrite rules were developed for the two languages (Yorùbá and English). The data was collected from home domain vocabularies. The re-write rule was verified using Natural Language Toolkits (NLTKs) and implement using python programming language.

The system interface gives the user the opportunity to type simple English language sentence and the resulting Yorùbá Translation is displayed. The result shows that the system performance is close to the expert opinion, having considered the scope for which the system is developed. Based on the result gathered there are some issues to address that could be considered in a future work.

Keywords—Machine translation, Yorùbá language, rule-based approach, English language, sentence

I. INTRODUCTION

According to [1], Machine translation (MT) is a sub-field of computational linguistics that investigates the use of computer software to translate text or speech from one natural language to another. At its basic level, MT systems perform simple substitution of words in one natural language for words in another; but this alone usually cannot produce a good translation of a text, because recognition of whole phrases and their closest counterparts in the target language is needed. Solving this problem with statistical techniques is a rapidly growing field that has led to better translations, handling differences in linguistic typology, translation of idioms, and the isolation of anomalies [2].

According to [3], improved translation quality can also be achieved by human intervention: for example, some systems are able to translate more accurately if the user has unambiguously identified which words falls into a specific part of speech (POS). With the assistance of these techniques, MT has proven useful as a tool to assist human translators and in a very limited number of cases, can even produce output that can be used directly without further processing.

The progress and potential of machine translation has been debated much through history. Since the 1950s, a number of scholars have questioned the possibility of achieving fully automatic machine translation of high quality [4]. Some critics claim that there are in-principle obstacles to automatizing the translation process of what has been achieved is the development of programs which can produce ‘raw’ translations of texts in relatively well-defined subject domains, which can be revised to give good-quality translated texts at an economically viable rate or which in their unedited state can be read and understood by specialists in the subject for information purposes. In some cases, with appropriate controls on the language of the input texts, translations can be produced automatically that is of higher quality needing little or no revision [5].

The remaining part of the paper is structured as follows: Section 2 explains machine translation approaches; section 3 gives an overview of English and Yorùbá similarities, differences and characteristics. Section 4 gives the system design and implementation, while Section 5 discusses the results. Section 6 concludes the paper.

II. MT SYSTEM APPROACHES

MT systems can be developed using three approaches depending on data (corpora) availability and type of language. The three approaches are: rule-based (RBMT), statistical and Hybrid. There are two types of rule-based machine translation systems: Transfer Rule-Based Machine Translation and Inter-lingual RBMT Systems [6].

Statistical machine translation is a machine translation paradigm where translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora. The statistical approach contrasts with the rule-based approaches to machine translation as well as with example-based machine translation [7]. Statistical machine translation (SMT) is an approach to MT that is characterized by the use of machine learning methods. In less than two decades, SMT has come to dominate academic MT research and has gained a share of the commercial MT market [8].

Hybrid machine translation (HMT) leverages the strengths of statistical and rule-based translation methodologies [9]. Several MT companies (Asia Online, LinguaSys, Systran, PangeaMT, and UPV) are claiming to have a hybrid approach using both rules-based and statisticals.
Rule-based Machine Translation (RBMT) also known as ‘Knowledge-based Machine Translation’, ‘Classical Approach’ of MT is a general term that denotes machine translation systems based on linguistic information about source and target languages. Basically the linguistic information can be retrieved from (bilingual) dictionaries and grammars covering the main semantic, morphological and syntactic regularities of each language [10][11].

Rule based machine translation system, consists of collection of rules called grammar rules, lexicon and software programs to process the rules [12]. In [13], it is defined as systems that use large collections of rules, manually developed over time by human experts, which map structures from the source to the target language. The rules are written with linguistic knowledge gathered from linguists or other means.

When the sentences are inputted (in source language), a RBMT system generates the output sentences (in target language) on the basis of morphological, syntactic and semantic analyses of both the source and the target languages.

Rules play major role in various stages of the translation: syntactic processing, semantic interpretation, and contextual processing of language. Figure 1 below is the schematic diagram of a rule-based machine translation system.

![Figure 1. Rule-based Machine Translation](source: http://tinyurl.com/p66x5j)

Transfer-based machine translation is a type of machine translation based on the idea of inter-lingua and is currently one of the most widely used methods of machine translation [6]. Both transfer-based and interlingua-based machine translation have the same idea; to make a translation, it is necessary to have an intermediate representation that captures the "meaning" of the original sentence in order to generate the correct translation [14]. In interlingua-based MT this intermediate representation must be independent of the languages in question, whereas in transfer-based MT, it has some dependence on the language pair involved. The way in which transfer-based machine translation systems work varies substantially, but in general they follow the same pattern; they apply sets of linguistic rules which are defined as correspondences between the structure of the source language and that of the target language [6].

A. Stages in Translation

The methods which are chosen and the emphasis depends largely on the design of the system, however, most translation systems include the following stages:

- Morphological analysis
- Lexical categorization
- Lexical transfer
- Structural transfer
- Morphological generation

Transfer Rule-Based Machine Translation was used in this research because it allows us to use manual tagging of the part of speech (POS). The work is also restricted to simple sentences translation.

III. ENGLISH AND YORÙBÁ LANGUAGE

According to [15], English language basically moves from concrete to abstract, while Yorùbá language moves from abstract to concrete. Thus, Yorùbá language can be seen as a complex language to study. It has a lot of cultural entities (Proverb - ewi, oriki, etc) which cannot we adequately represent in English (e.g. iṣẹ̀ ni ọ́gùn îsẹ̀).

There are various differences and similarities between English and Yorùbá Language, some of which are discussed here:

- Yorùbá language borrows English language words for most of the words that does not have a Yorùbá equivalent.
  
  Bíró → Bíró, Bread → Búrédi

- In English Language determinant (e.g the) always come after a noun but in Yorùbá language, determinant always follow noun.
  
  The<N> boy<N> → ọmọkùnrin<N>náà<N>Det>

- Yorùbá language is a tonal language with 3 distinct tones while English is not.

- Most sentences in English language cannot be translated to Yorùbá using word-for-word translation. e.g.

  The boy is coming → Náà ọmọdẹ̀kùnrin n bò
  The correct translation must be; ọmọdẹ̀kùnrin náà n bò.

The main features of both languages are highlighted in the table below:

<table>
<thead>
<tr>
<th>TABLE I. FEATURES OF ENGLISH AND YORÙBÁ LANGUAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Attributes</strong></td>
</tr>
<tr>
<td>Language types</td>
</tr>
<tr>
<td>Timing</td>
</tr>
<tr>
<td>Orthography</td>
</tr>
<tr>
<td>Language</td>
</tr>
</tbody>
</table>
IV. METHODOLOGY

The software is designed as a window application. The grammar of the language is written to follow the re-writes rule developed for the Translation process. Sample of the re-write rules developed for the two languages (English and Yorùbá) are given below:

**English language:**

\[
S ::= <NP> <VP> \\
VP ::= <V> | <NP> <V> | <V> <V> \\
PP ::= <P> <NP> \\
V ::= <NP> <V> \\
NP ::= <DET> <N> | <PRON> <N> | <DET> \\
<ADJ>
\]

**Yorùbá language:**

\[
S ::= <NP> <VP> \\
VP ::= <NP> <V> | <NP> <PP> | <V> | <V> <V> \\
| <PRON> <V> | <V> <NP> | <VP> <PP> \\
PP ::= <P> <NP> \\
NP ::= <N> | <PRON> | <DET> | <ADJ> <PP> | <NP> | <PRON> \\
\]

S stands for the sentence, NP, PP, VP, N, V, ADJ and DET are the non-terminals. NP is Noun Phrase, PP is prepositional phrase, P is preposition, ADJ is Adjective, N is Noun and V is verb. The Left hand side (LHS) is substituted with the Right hand side (RHS) until the terminals are reached.

In figure 2 below the re-write rule for Yorùbá translation is model using Finite State Automaton (FSA). The Automaton is tested with sample input as shown in figure 3.

![Fig. 2. State Diagram for Yorùbá Re-Write Rule](source: [20])

Sample sentences are run in figure 4 and 5 using JFAP to show the pattern of the translation parse tree structure.

![Fig. 3. Yorùbá Re-Write Rule with accepted states](source: [20])

The grammar parse three structure was also model using NLTK parser. NLTK is a very important tool that is used in this research work to easily model the behavior of both language sentence tree structures. Example is show in figure 6 and 7 below.
A. DATA COLLECTION

The data (corpus) for the research was from simple sentence spoken in the home environment. The sentences are further broken down into their part of speech (POS). The different parts of speech are stored in pairs. Figure 8 shows a sample of the stored part of speech.

B. DEVELOPMENT TOOLS

The main tools used in this research are:

- **Python programming language** – this is the core programming environment for the application development.
- **NTLK (Natural Language Toolkit)** – this is a support kit for python programming language. Its features include: support for parsing, Part of Speech (POS) tagging, corpora design and analyses.
- **PyQt** – this is also a support kit for the design of the application GUI.
- **py2exe** – this was used to compile the python codes (.py) to an executable file (.exe).
- **NSIS** – use basically in building the window installer for the application.

C. REQUIREMENT ANALYSIS

The requirements and specifications of the software are as follow:

- to present a user friendly interface to the user;
- to give the user access to enter simple basic sentences in English language provided the sentence is within the domain covered;
- parse the sentence to understand the structure;
- translate and output the equivalent meaning of the sentences entered in standard Yorùbá language; and
- give the user ability to add to the corpus (database)

The system sequence diagram and the use case diagram are shown in figure 9 and 10 below.
V. RESULTS AND DISCUSSION

The system perform is compared with another machine translation system. Google Translator was use since is a standard machine translator that include an English to Yorùbá translation. Figure 11, 12 and 13 below shows the result of the output from our system with output from google translator.

The result from the output from Google Translator shows that our system gives a better translation as can be considered as a standard translation for Yorùbá languages.
VI. CONCLUSION

According to the research conducted in [16], improvement to the translation process can be done only by formalizing our linguistic knowledge and enriching the computer with adequate rules to deal with the linguistic phenomenon. Thus, to give an automated high-quality translations, there is need to generate more robust re-write rules that can accommodate the translation.

There are lots of work that can be done to improve the quality of the system output inorder to increase its usefulness.

Other areas of note for further research include ability to extent the system capability to translate longer sentences, ability to translate paragraph text and the issue of ambiguities in translation.

REFERENCES


