Review of Different Approaches for Machine Translations

B.A. Abdulsalami[#1], B.J. Akinsanya*[2]
[#1]Assistant Lecturer, Department of Mathematics & Computer Science, Fountain University
Osogbo, Osun State, Nigeria

Abstract: *Language is an effective and systematic medium of communication which explicitly represents the ideas, emotions, desires and other expressions of the human. Effort to access information from one language to another leads to the interest in the area of machine translation. Machine Translation (MT) is an automated system that analyzes text or speech from source language and produces an equivalent text or speech in target language ideally without human intervention. It is a subfield of information technology and applied linguistics dealing with the translation of human languages. The availability of MT makes it ideal for translation of Web pages and other materials on the internet. Different approaches have been proposed and each approach has its major merits and demerits. Thus, in this work, the various approaches to machine translations are reviewed, in addition to its future need in order to provide more robust machine translation systems.*

Keywords: *Machine Translation, Dictionary-Based Approach, Rule-Based Approach, Corpus-Based Approach, Hybrid Approach, Knowledge-Based Approach.*

I. INTRODUCTION

Language is an effective and systematic medium of communication which explicitly represents the ideas, emotions, desires and other expressions of the human. According to [8], there are over 6,800 living languages exist in the world. Thus, in order to access information written in another language we need to create tools for translating information from one language to another.

Translation is the transfer of the meaning of a text from one language to another for a new relationship. Translation is not always a straight forward case of substituting word(s) in the source language with the equivalent word(s) in the target language. In business community, language barrier is a factor stifling the growth of international business deal. It is difficult to gain foothold in international trading because it is more convenient for the global audiences to read and communicate in their native languages. Translations were first made by human translators but this human translation was unable to meet demands for translated content. These limitations of human translation led to the discovery of Machine Translation.

According to [1], Machine Translation (MT) is a subfield of computational linguistics that investigates the use of computer software to translate text or speech from one natural language to another. It is the area of information technology and applied linguistics dealing with the translation of human languages[1]. Translation by or with the aid of machines can be faster than manual translation and can also reduce the cost per word of a translation. The use of MT leads to improvements in quality of word, particularly in the use of consistent terminology within a text or for a particular kind of client. Similarly, the availability of MT makes it ideal for translation of Web pages and other materials on the internet.

II. MACHINE TRANSLATION APPROACHES

MT tools performance are being evaluated based on intended use of the translation, accuracy, speed and the translation process. Different approaches have been proposed and each approach has its major merits and demerits. The various approaches that have been proposed to automate the translation process in literatures are:

A. Dictionary Based Machine Translation (DBMT)

This approach of translation is based on entries of a language dictionary. The equivalent word/phrase is used to develop the translated verse. The first generation of machine translation was entirely based on machine-readable or electronic dictionaries. This method is helpful in translation of phrases but not sentences.

B. Corpus Based Machine Translation (CBMT)

Corpus based machine translation (also referred to as data drive machine translation) was invented in 1989 and has been one of the dominating approach due to it high level of accuracy achieved. It uses bilingual parallel corpus to obtain knowledge for new incoming translation. It uses large amount of raw data in the form of parallel corpora to acquire translation knowledge. The accumulation and management of this huge bilingual data corpus is costly. CBMT is further classified into the following sub approaches: Statistical machine translation (SMT), Example based machine translation (EBMT) and Context based machine translation (CBMT).

1) Statistical Based Machine Translation (SBMT)

In this approach, translation is generated on the basis of statistical models. The statistical model parameters are derived from the analysis of bilingual text corpora. The model can be classified into three: Statistical word-based translation model; Statistical phrase-based translation model; Statistical syntax-based model. SMT depends on a language model, a translation model and a decoding algorithm. The translation model ensures that the machine translation system produces target hypothesis corresponding to the source sentence. The language model ensures the grammatically correct output. This approach requires minimal human effort. It can be created for any language pair that has enough training data. However, real word training sets may over ride translations. For example "I took the spoon to the guest room" gets mistranslated as "I took the cutlery to dining room" due to an abundance of "cutlery to dinning" in the training set.

The key problems in statistical MT are: estimating the probability of a translation, and efficiently finding the sentence with the highest probability. High quality aligned bilingual corpora is still expensive and time consuming to create but, once created becomes a valuable asset to any organization implementing SMT technology, with translations benefiting from economies of scale over time.

2) Example Based Machine Translation (EBMT)

This is an approach based on recalling or finding analogous samples (of the language pair). In this approach, the system is given a set of sentences in the source language (the language from which one is translating) and corresponding translations of each sentence in the target language which point to point mapping. These examples are used to translate similar types of sentences of source language to the target language, so if a previously translated sentence occurs again, the same translation is likely to be correct again. It also uses dictionaries to translate text.

3) Context Based Machine Translation (CBMT)

It is an approach that requires extensive monolingual target text corpus, a full-form bilingual dictionary and optionally a smaller monolingual source text corpus to run its algorithm. The monolingual source text corpus is use to further improve the translational quality. A system that uses this approach is able to handle longer strings and can handle word ambiguities. It segregate translated segments as high or low on the basis of level of confidence. It also preserves context translated verse and has the capability to generate alternative phrase in case a suitable match is not found in the target language.

C. Knowledge Based Machine Translation (KBMT)

Knowledge-based machine translation is the process of applying syntactic knowledge of the source language and semantic knowledge relevant to the source text in order to produce recognized language-free meaning representation, which may then be rendered in many different languages. The analysis process of producing a meaningful representation is far more complex than that of using target language knowledge to express the meaningful representation in the target language, because the former is a many-to-one mapping, whereas the latter may be coerced into a one- to-one mapping. KBMT is used in some domains such as stocks and other security negotiations, doctor patient communication, weather forecasts, banking transactions, financial reports, economic analyses, invoices and purchase orders, etc. Thus, KBMT is particularly well-suited for multi-lingual translation in high volume well-defined semantic domains. A times some esoteric domain knowledge necessary for semantic analysis will be lacking, so retaining the possibility of interacting with a human user knowledgeable of the domain (but not of different target languages) is done to clarify any difficulties too complex for the domain semantics to handle.

D Rule Based Machine Translation (RBMT)

This approach has much to do with morphological, syntactic and semantic information about the source and the target language. The linguistics rules are then built over these rules and bilingual dictionaries for this pair are used. RBMT is able to deal with the needs of variety of linguistic phenomena and is extensible and maintainable but requires high investment. RBMT approach deals with the word-order problem and since it uses linguistic knowledge, the produced error can be traced. The benefit of this approach is that it analyzes the input on the syntactic and to extent semantic levels. The downside of RBMT is that it requires a deep linguistics knowledge as well as long time to prepare the rules. RBMT is divided into three sub-approaches namely: Direct, Interlingua and Transfer approach.

1) Direct Approach

This is an approach in which words of source language are translated without passing through intermediary representation. It can be characterized as 'word-for-word' translation with some local word-order adjustment. It gave the kind of translation quality that might be expected

from someone with a very cheap bilingual dictionary and only the most rudimentary knowledge of the grammar of the target language. There can be frequent mistranslations at the lexical level and largely inappropriate syntax structures which mirrored too closely those of the source language.

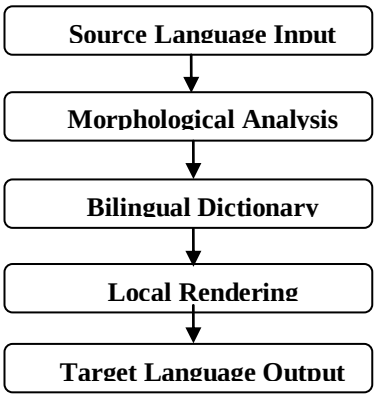

Fig. 1 Steps for Direct Approach

2) Interlingua Approach

This approach is an approach in which the source language is transformed into an intermediary language which is independent on any of the languages involved in the translation. The intermediate representation includes all information necessary for the generation of the target text without 'looking back' to the original text. It is an inherent parts of a branch called inter linguistics and aimed at creating linguistic homogeneity across the globe. This approach has more relevance in multilingual translation machine because it emphasizes on single representation for different languages. This approach requires only two modules which are synthesis and analysis. Each analysis module can be independent, both of all other analysis modules and of all generation modules. Target languages have no effect on any processes of analysis; the aim of analysis is the derivation of an 'interlingua' representation.

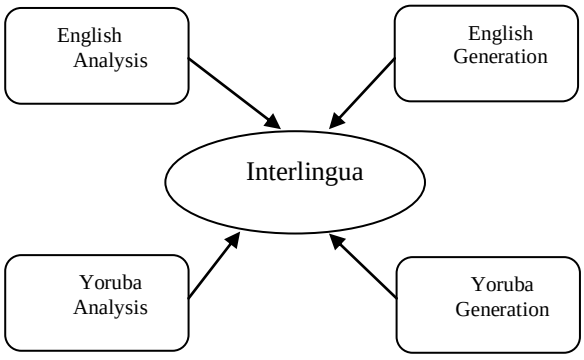

Fig. 2. Interlingua Model with Two Language Pairs

3) Transfer Approach

This is an approach in which the source language is transformed into an abstract, less language-specific representation. This representation is then generated for the target language using bilingual dictionaries and grammatical rules. This approach can be divided into three categories:

1. *Analysis:* This is done based on the linguistic information such as morphology, part-of-speech, syntax, semantics etc. Heuristics as well as algorithm are applied to parse to source language and derive the syntactic structure of the text to be translated or the semantic structure.
2. *Transfer:* The syntactic/semantic of source language is transferred into the syntactic/semantic structure of the target language.
3. *Synthesis:* In this section, it replaces the constituents in the source language to the target language. It depends on the language pair involved in which there are different representations for different languages.

It is possible with this translation approach to obtain fairly high quality translations, with accuracy in the region of 90%.

E. Hybrid Machine Translation

Hybrid approach is the combination of the strengths of statistical and rule based translation approaches. It involves rules post-processed by statistics. The translations are performed using rules based engine and statistics are used in an attempt to adjust and correct the output from the rules engine. The statistics engine is guided by rules and the rules are used to preprocess data. This approach is better than the previous as it is the most widely used MT systems nowadays. It uses the rule-based and the statistical approaches. There have been several research works which combine both approaches and has more power, flexibility, and control in translation.

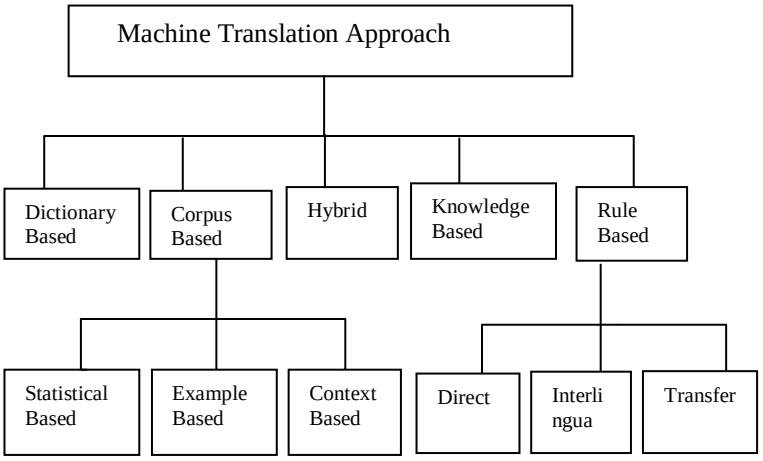

Fig. 3 Different existing Machine Translation Approaches

III. STRENGTH(S) AND WEAKNESS (ES) OF EACH APPROACH

TABLE I
STRENGTH(S) OF EACH APPROACHES.

APPROACH	STRENGTH
Dictionary-Based	It is helpful in translation of Phrases.
Corpus-Based	It can be built in much less time and do not require linguistic experts to apply language rules to the system. It can mimic the style of the training data to generate output based on the frequency of patterns allowing them to produce more fluent output.
Knowledge-Based	It is well-suited for multi-lingual translation in high volume well-defined semantic domains
Rule-Based	It is built with much less data instead uses language rules and dictionaries. It achieves good quality results.
Hybrid	It is more effective, flexible and still the leading approach. It is combination of rule-based and statistical-based approach strengths.

TABLE II
WEAKNESS (ES) OF EACH APPROACH

APPROACH	WEAKNESS
Dictionary-Based	It is not helpful in translation of sentences.
Corpus-Based	Creating High quality bilingual corpora is expensive and time consuming. A time real word training sets may over ride translations.
Knowledge-Based	A time some esoteric domain knowledge necessary for semantic analysis will be lacking.
Rule-Based	Language is constantly changing, which means rules must be managed and updated where necessary. It sometimes lacks fluency.
Hybrid	It is costly.

IV. LITERATURE REVIEW

[11] developed English to Yoruba translation system using rule-based approach to machine translation. The machine translator can translate modify and non-modify simple sentences (subject verb object (SVO)). The English to Yorùbá translation process was modeled using phrase structure grammar and re-write rules. The re-write rules were designed and tested using Natural Language Tool Kits (NLTKs). Parse tree and Automata theory based techniques were used to recognize the computational problem underlining the translation process. Unified Modeling Language (UML) was used for the software design. The Python programming language and PyQt4 tools were used to implement the model. The developed machine translator was tested with simple sentences.

Also, [6] in their work described a bi-lingual Machine Translation (MT) tool in the agriculture domain for translation of Arabic interrogative sentence into English. This MT tool automates the translation of user interfaces of knowledge-based systems. The tool is used in Central Laboratory for Agricultural Expert Systems (CLAES) for developing bilingual (Arabic-to-English) expert systems. The transfer-based MT approach was used in designing the translator. The major design goal of the tool is to be used as a stand-alone tool and be well integrated with general MT systems for Arabic sentence.

Moreover, [12] developed a machine translation of noun phrases from English to Igala using the rule-based approach. The objective of the research was to model language processors that can accepts as input Noun phrases in English language and translate them to Igala language. The model implementation was done using VisualBasic.net programming language as front end and Microsoft Access database as back end. 120 randomly selected English noun phrases was used for testing the application using Bilingual Evaluation Understudy (BLEU) method for evaluating Machine Translation Systems. An accuracy of 90.9 % was obtained.

Another work by [4] is the development of an online machine translation between Chinese and Spanish. Statistical Machine Translation was used in developing the system. The online system allows translating either from a web based interface or from two mobile applications (one for Android and one for iOS). Phrase-based statistical MT approach with Moses (statistical MT system licensed under LGPL) was used. The word-based and the phrase-based models were originally based on the noisy-channel model.

Furthermore, [9] proposed English to Malayam using Hybrid Machine Translation Approach. This

hybrid approach extends the baseline statistical machine translator with a translation memory. The system was implemented and evaluated using BLEU score and precision measure and the hybrid approach is found to improve the performance of the translator. Open source tools such as IRSTLM, GIZA++, Moses decoder etc. were used for implementing the proposed system. The tools were installed in Ubuntu 10.4 operating system environment.

[5] also developed a Rule-based Machine Translation of Noun Phrases fro Punjabi to English. The system is based on transfer approach, with three main components: an analyzer, a transfer component, and a generation component. A Punjabi morph analyzer developed at 'Advanced centre for technical development of Punjabi language' was used for analyzing. The database for the system consists of more than one lakh words from which 63,000 are the inflected nouns which are derived from about 18,000 root nouns. The database contains the grammatical category of each word and also the inflected words it can form. From this database, the tagger gets the information and tags each word of the phrase.

In addition , [10] proposed a Rule Based Hybrid Statistical Machine Translation (RBHSMT). The system was designed as a bilingual system that converts English Sentence to Tamil sentence using particular rules. The process of RBSHMT is divided into Training Stage and Development Stage through which a translated text in Tamil can be obtained for the entered English text. The training stage consists of five modules: Language Module, Pre processing, Phrasal Module, Translation Module and Post Processing. The Development stage consists of two modules: Transliteration and Decoding. System Architecture .

[13] also proposed an efficient Hybrid Machine Translation (approach is proposed which involves the combination of rule based and statistical technique for translating text from source language which is English to the target language which are the Indian languages such as Tamil, Malayalam and Hindi. Rule based approach was used to form linguistics rules of the source language, context free grammars (CFG) were used in generation of the language structures and then the errors in the translated sentences were corrected by applying a statistical technique. It is going to integrate with Automatic Speech Recognition (ASR) and Text To Speech (TTS) systems in order to develop a speech to speech translation system for English to Indian languages. The evaluation tools used for the proposed hybrid machine translation system were BLEU and NIST evaluation metric.

V. CONCLUSION

From the various work that have been reviewed so far, it was found that Machine translation of natural languages is not a simple task, but rather, a very difficult task; Though people perceived it as the simple substitution of words in one natural language for words in another. But this is not so due to the complexity of some natural languages. Researchers in the field of Machine Translation have approached it in various ways. The common approaches are Statistical, Rule-based and Hybrid approach .Statistical Machine Translation approach is growing. The development and increasing availability of cloud-based computing is enhancing the high computer processing power and storage capacity required to run SMT technology effectively, making SMT a game changer for the localization industry. Through the internet, Training data for SMT engines is becoming more widely available and increasing volumes of multilingual content being created by both companies and private internet users.

Also, the two major goals in any translation system development work are accuracy and speed of translations; however most researchers failed to evaluate the accuracy and speed of their systems. Bilingual Evaluation Understudy (BLEU) is the mostly used evaluation measure for machine translation systems. Translation works for language can further be enhanced by researchers if they can combine two or more approaches in order to develop more robust systems. The systems should also be made in mobile application form so as to allow effective usage and dissemination.

REFERENCES

[1] Abiola O.B, Adetunmbi A.O and Oguntimilehin A. (2015) , Department of Computer Science Afe Babalola University, Ado Ekiti, Nigeria and Department of Computer Science, Federal University of Technology Akure, Nigeria: "*A Review of the Various Approaches for Text to Text Machine Translation*", International Journal of Computer Applications:Volume 120-No.18, June 2015.

[2] Bonnie J. Dorr, Eduard H. Hovy, Lori S. Levin(2013) "*Machine Translation: Interlingua Approach*"

[3] Jordi Centelles & Marta R. Costa-Jussa(2013) " *Machine Translation for Chinese-Spanish: Experimenting with online Statistical and Rule-Based Paradigms*" July, 2013.

[4] Jordi Centelles and Marta R. Costa-Jussa (2014) "Chinese-to-Spanish Rule-Based machine Translation System" Proceedings of the 3rd Workshop on Hybrid Approaches to Translation (HyTra)@EACL 2014, pp 82-86.

[5] Kamaljeet Kaur batra & G.S Lehal(2010) "Rule Based Machine Translation of Noun-Phrases from Punjabi to English" International Journal of Computer Science, Vol 7, Issue 5, ISSN 1694-014, pp 409-413. www.ijsci.org, 2010.

[6] Khaleed Shalan, A Hendam and A Rafea (2016), The British University in Dubai, Informatics Dubai International Academic City,Honorary Fellow, School of Informatics, University of Edinburgh, Central Lab. For Agricultural Expert Systems (CLAES), TEUES6 El Nour St., Giza, 1123 Egypt:"*An English-Arabic Bi-Directional Machine Translation Tool in the Agriculture Domain*" September 2016.

[7] M.D Okpor(2014), Department of Computer Science, Delta State Polytechnic, Delta State, Nigeria: "*Machine Translation Approaches: Issues and Challenges*". IJCSI Vol. 11, Issue 5, No. 2, September 2014, ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784.

[8] Masaru Tomita & Jaime G. Carbonell, Carnegie Mellon University " *Knowledge-BasedMachine Translation, The CMU Approach"*, Institute of Software Research, School Computer Science 1987

[9] Nithya B. & Shibily Joseph(2013): *" A hybrid Approach to English to Malayam Machine Translation*" International Journal of Computer Science Application, Vol. 81, No. 8, ISSN 0975-8887, 2013, pp 11-15. www.ijcaonline.org.

[10] S.R.Priyanga, AP & A.AzhaguSindhu(2013) *"Rule Based Statistical Hybrid Machine Translation"* International Journal of Science and Modern Engineering (IJISME) ISSN: 2319-6386, Vol.1, Issue-5, April 2013.

[11] Safiriyu I. Eludiora & Odetunji A. Odejobi(2016), Department of Computer Science and Engineerin Obafemi Awolowo University, Ile- Ife, Nigeria: " *Development of an English to Yoruba Machine Translator*". I.J. Modern Education and Computer Science, 2016

[12] Sani Felix Ayegba, Osuagwu O.E. ,Njoku Dominic Okechukwu(2014), Department of Computer Science, Federal Polytechnic Idah, Kogi State, Nigeria, Department of Computer Science, Imo State University, Owerri, Nigeria and Department of Electrical/Electronics Engineering, Imo State Polytechnic, Umuag, Nigeria: " *Machine Translation of Noun Phrases from English to Igala using the Rule-Based Approcah*" West African Journal of Industrial & Academic Research Vol.11 No.1 June 2014.

[13] Sangeetha. J, S. Jothilakshmi, R. N. Devendra Kumar (2014), "An Efficient Machine Translation System for English to Indian Languages Using Hybrid Mechanism" International Journal of Engineering and Technology (IJET), pp 1909-1919, Vol 6 No 4 Aug-Sep 2 / 3 2014, ISSN: 0975-4024.

[14] Sneha Tripathi & and Juran Krishna Sarkhel(2013), Department of Library and Informatics Science, University of Kalyani and Central Library, Banaras Hindu University: "*Approaches to Machine Translation*". Annals of Library and Information Studies, Vol. 57, December

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

[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] Venkateswara Prasad T. and Mayil Muthukumaran. G (2013) "Telugu to English Translation using Direct Machine Translation Approach" International Journal of Science and Engineering Investigations (IJSEI), pp 25-32, vol2 issue 12 January 2013, ISSN:2251-8843.