National Conference on
"Role of Engineers in Nation Building" organized by VIVA Institute of Technology, Mumbai
(6th and 7th March 2020)
International Journal of Engineering Research & Science (IJOER)
ISBN:[978-93-5391-287-1] ISSN: [2395-6992] [Vol-6, Issue-3, March- 2020]

A brief survey of Machine Translation Systems

Raj Vyas[1], Kirti Joshi[2], Hitesh Sutar[3], Tatwadarshi Nagarhalli[4]
[1]Department of Computer Engineering, Mumbai University, Mumbai
rajbvyas131@gmail.com
[2]Department of Computer Engineering, Mumbai University, Mumbai
kirti.mj99@gmail.com
[3]Department of Computer Engineering, Mumbai University, Mumbai
sutarhitesh192@gmail.com,
[4]Department of Computer Engineering, Mumbai University, Mumbai
tatwadarshipn@viva-technology.org

Abstract**—***The use of machine translation has given a huge boost in the field of natural language processing. Machine translation is used to convert a text or speech from one language to another. The earlier stages has seen various types of machine translations which include Statistical Machine Translation (SMT), Rule-based Machine Translation (RBMT) and Hybrid Machine Translation. These types are based on the large volumes of bilingual text, grammatical analysis of source language and combination of both. One of the best type of machine translation Neural Machine Translation (NMT) uses an artificial neural network to predict the likelihood of a sequence of words. In this survey analysis of different approaches of machine translation has been carried out. The analysis of BLEU score is also carried out of these approaches. The survey includes translation on English, Hindi, Myanmar and Punjabi languages. Syntax, Hybrid, ConceptNet and Statistical machine translation approaches have been analyzed in the survey.*

Keywords—Natural language processing, Neural machine translation (NMT)

I. INTRODUCTION

India being a multilingual country, after every 50 kilometers the spoken language changes. Therefore, there is no single universal language.There are 23 official languages in India. Nearly 50% of Indian population speaks Hindi [1]. Being such a diverse country, it becomes extremely difficult for people to learn all the languages.Doing the manual translation in such scenarios become time consuming and costly.Hence, there is a need for having translation system in place to address such issues.

Till now the algorithms used for machine translation are Rule-based Machine Translation (RBMT), Statistical Machine Translation (SMT), and Neural machine translation. RBMT relies on countless built-in linguistic rules and millions of bilingual dictionaries for each language pair. The technique uses the complex rule sets and then transfers the grammatical structure of the source language into the target language. RBMT systems are built on gigantic dictionaries and sophisticated linguistic rules. SMT utilizes statistical translation models generated from the analysis of monolingual and bilingual training data. The translation is selected from the training data using algorithms to select the most frequently occurring words or phrases. SMT technology relies on bilingual corpora such as translation memories and glossaries to train it to learn language pattern and uses monolingual data to improve its fluency. NMT is a type of machine translation that depends on neural network models (based on the human brain) to develop statistical models for the purpose of translation. The primary benefit of NMT is that it provides a single system that can be trained to decipher the source and target text [2].

II. LITERATURE SURVEY

S. Saini, et. al. [3] have proposed a combined approach of eight completely different design combos of NMT (Neural Machine Translation) for English to Hindi translation and compared its results with standard computational linguistics techniques. From this experiment it absolutely was discovered that NMT needs terribly less quantity of knowledge size for coaching. The kind of neural networks utilized in NMT is Recurrent Neural Networks (RNN). There was increasing impact whereas victimisation RNN, results were conjointly inaccurate. Thus, to beat the shortcomings of the RNNs, the

International Journal of Engineering Research & Science (IJOER)
ISBN:[978-93-5391-287-1] ISSN: [2395-6992] [Vol-6, Issue-3, March- 2020]

employment of Long- and Short-Term Memory (LSTM) models for secret writing and coding was done. Once the dataset was pre-processed, the supply and target files were fed into the encoder layer to organize the vectors from the sentences.

P. Vijayalakshmi, et. al. [4] have proposed a S2ST system for travel expressions which might perform translation between English and Hindi and is predicated on applied Statistical Machine Translation approach. During this system speech recognition is meted out mistreatment HMM based mostly acoustic models at word level. The transition between each language happens with the assistance of 3 major modules particularly, speech recognition, computational linguistics and speech synthesis. ASR system is employed that is liable for changing speech signal in language A to corresponding text in same language. ASR system uses HMM to coach the system. Once, the text is obtained the Machine Translation (MT) system comes into image. Here a applied Statistical Machine Translation (SMT) supported Bayesian criteria is employed. This SMT performs the task of changing the text in language A into text in language B. The results of the SMT produces translation for every word within the target language.

P. Kumar, et. al. [5] have proposed a system that aims at translating text from English to Hindi using stepwise procedure. Firstly, the English sentences are stored in file. These sentences are passed to the syntax analyzer for checking their correctness and grouping. The grouping is done on the basis that some words like places; names which do not require translation. These groups are passed to synthesizer. Synthesizer contains Hindi dictionary, encyclopedia and web mining. Hindi dictionary gives the Hindi meaning of English sentence. If any English word which is not in Hindi dictionary are searched into the encyclopedia. If any word which is not in both will be searched through web. The words which are not present in all three are marked as a wrong word. The syntax matcher takes the Hindi words of the respective English word from synthesizer and replaces that English word with Hindi word.

O. Dhariya, et. al. [6] have proposed a system that has been implemented in a sequence of four primary steps: Segmentation, Translation, POS, Tagging and Rearrangement. Segmentation is performed through first finding all possible sub-parts in the sentence belong to parallel Hindi-English database. Later, the remaining components of the sentence would be broken into words. At the top of this stage, output would have set of phrases, straightforward sentences and words. per POS, proper name denotes a selected name used for a personal person, place, or organization. Tagging is that the method to spot the linguistic properties of every individual matter unit. The parallel Hindi-English lexicon contains the tag of every English word. Depending upon the assigned tag, the system finds the proper grammatical structure for the sentence and rearranges the words to construct a grammatically correct sentence. All the segments whether words or partial simple sentences are translated individually. Depending upon the assigned tags to the words, a matching grammar rule is selected for the given input.

N. Raju, et. al. [7] have proposed a system in which a corpus for Statistical Machine Translation system by using the parallel text of Telugu and English languages was developed. In the initial phase, the data is prepared by tokenizing the corpus, Filtering out long sentences and lower casing the corpus. Statistical machine translation mainly consists of Language Model, Translation Model and decoder. The Language Model will determine the probability for the target language. This is most important in SMT because it can achieve the adequacy and fluency for the translated text. It will survive two purposes that is word order and word choice. In word order, it specifies which word would precede the sequence of words. In Word choice, there will be set of words for a translated word from which a single should be selected. Here the probability is decomposed by using Markov chain rule. In word choice, probabilities are calculated for words instead of sentences because no translation model is capable for calculating the probabilities in sentence level. In this a TM called Phrase Based Translation Model is used, it divides a sentence into smaller number of phrases and then each phrase is translated one at a time. Decoder is used to maximize the probability of translation. Here the word for translation is chosen by using maximum likelihood. In the search space it finds the best probability of all possible translations. It finds the most probable translation. It makes use of heuristic search to find the best translation. Thus, the input to the system is a Telugu sentence and the output will be English sentence.

International Journal of Engineering Research & Science (IJOER)
ISBN:[978-93-5391-287-1] ISSN: [2395-6992] [Vol-6, Issue-3, March- 2020]

Y. Ghadage, et. al. [8] have proposed a system that is divided into two phases i.e. training and testing. First in training phase speech utterances of each sentence is recorded. Speech signal is preprocessed and segmented into words. For each word acoustic features are extracted using MFCC method. Such features for each word forming feature vector is stored for reference. In testing phase the speech utterance to be tested is preprocessed, segmented into words and features are extracted for each word. These features are compared with the reference feature vector stored during training phase. This is done by using combination of SVM and Minimum Distance Classifier. The word having minimum difference is given as recognized word. The system is trained with the training database and the recorded speech utterances stored in test database are used to test the system. Minimum distance classifier and support vector machine techniques are used for classification purpose. The trained speech samples are saved as reference models into database. After that each segmented speech sample of test speech signal is passed over reference models and minimum distance is computed. Each word recognition is done by using minimum distance and SVM model. The whole system is implemented and tested in MATLAB software.

M. Faisullah, et. al. [9] have proposed a system where overall work describes regarding speech to speech translation (S2ST) system for English to Hindi over Bluetooth network and more it'll be extended to mobile network. This work is completed by keeping in mind regarding 3 modules, that is, Automatic Speech Translation (ATR), Machine Translation (MT) and Text To Speech (Speech Synthesizer) modules. ATR captures the speech from mobile devices and converts it into text and it uses Bluetooth network for causing. Along with these 3 modules, one more featured module is added for playing translation operation, that is, Bluetooth module. Speech to Speech Translation contains chiefly 3 modules that square measure, Automatic Speech Recognition Module (ASR), artificial intelligence Module (MT) and Text To Speech (TTS) Module. ASR module captures the voice or speech from the mobile device through speaker and identifies the language spoken by the user and converts the speech into text and so the text send to next module. Machine Translation module consists of library for each language and once text is received by this module, it converts the text of 1 language to a different as per user selection and so it sends the translated text to last module.

A. Kaur, et. al. [10] have proposed a system which is started by the pre-processing stage that is a collection of operations, applied on input file so it will become processable for translation engine. Filtering is another major task performed during this part. In filtering, bound special expressions area unit detected and marked like replacement multiple spaces, collocations etc., within the input text. After that, tokenizer split a stream of characters into purposeful units referred to as tokens. The tokenizer takes the generated text as input. Individual words or tokens area unit extracted and processed to generate its equivalent tokens within the target language. The tokenizer, victimization area as delimiter, takes tokens one by one from the text and provides it to translation engine till the complete input text is processed. the interpretation engine is that the elementary part of MT system. It extracts token created by the tokenizer as input and generates the translated token within the target language. These translated tokens area unit concatenated one once another employing a area as delimiter. Then this generated text is passed to the post process part. Post-processing is that the last part of this MT system.

M. Bansal, et. al. [11] have proposed a system where pre-processing of English sentences is done using tokenization, true-casing and cleaning method. It converts the uppercase letters into lowercase letters and reduces length of those sentences which have length more than specified length. Moses tool is used for translating. While translating sentences from English to Hindi some unknown words are left. The extraction of those words is followed by source and target language words enrichment with common-sense knowledge (i.e. ConceptNet). This common sense data is translated using rule-based machine translation system. The translator maps the source word with target language words. The output generated by Rule-based machine translation is then combined with previously generated sentence.

Y. ShweSin, et. al. [12] have proposed a system where a large-scale parallel corpus was created initially. In this corpus, the bilingual sentences from Text Books, on-line Speaking category and native News, as well as BTEC corpus and angular position Corpus, was ready. Preprocessing consists of tokenizing and truecasing. For Word-level Neural artificial intelligence system, Moses's tokenization script was used and UCSY NLP word segmenter to phase English sentences and Asian country sentences severally. Quality analysis metrics such as BLEU (Bilingual analysis Understudy) was used that could be a script from Moses. blue cheese measures the exactness of associate MT system computed through the comparison of the system's output and a group

International Journal of Engineering Research & Science (IJOER)
ISBN:[978-93-5391-287-1] ISSN: [2395-6992] [Vol-6, Issue-3, March- 2020]

of ideally correct and frequently human-generated reference translations. it absolutely was seen that solely the performance of character-level neural machine translation was higher than the Word level neural artificial intelligence system. The performance of Character level neural artificial intelligence system improves up to two BLEU over Word level neural artificial intelligence system. Character level neural artificial intelligence system is capable of achieving compared to baseline systems. As a human translator, character level NMT has an impression in learning a much better translation.

P. Salunkhe, et. al. [13] have proposed a system that consists of Parallel Multi-Engines that method applied statistical and rule-based translation for same input document and turn out a optimized result by playacting applied math over rule based that provide fluent language sense outputs. Mapper algorithmic program is been employed in rule based mostly Translation, with Agriculture corpus, medical and touristy corpus for applied math analysis. Sanskrit wordnet has been enforced to boost lexicon and incorporate higher translation result. Current System has been projected for text Document which might be extended to speech and voice. Comparative analysis for purpose read in one dimension of solely restricted set of Queries is finished with Google Translator. Holding hybrid approach as higher methodology. a scientific survey of solely ten key articles employed in analysis has been done. This analysis article is extension of the previous analysis surveys and partial implementations.

S. Singh, et. al. [14] have proposed a system in which the process of translation is carried out through some stages: Preprocessor will provide an interface for the MT System to the web. This module will collect the input text from the web server which is in the form of HTTP request coming from the user. The preprocessing involves segmentation of the sentences, tokenization and POS tagging of the input. This leads to English rule formation for the sentence. If rule is matched with the rules of database, then corresponding Hindi is retrieved to the user. If rule is not matched then we go for the generation of Hindi rules on the basis of English rule formed and that Hindi rule is reordered to get Hindi. The whole focus in the process of machine translation is on the Target language rule generation and on the pattern of how it is aligned. This arrangement of Hindi in its SOV form is known as reordering. Formation of reordering rules has been manually done. These rules lead to the understanding of pattern of Target language formation, on the basis of which probability of occurring of subject-object, object-verb and subject-verb pairs together was obtained. The output accuracy may vary according to the length of the sentences.

III. ANALYSIS

The analysis of different machine translation approaches and techniques has been discussed in the following table. The BLEU score and accuracy of some papers is shown to get the overview of the technique or the approach.

TABLE 1: ANALYSIS OF MACHINE TRANSLATION SYSTEMS

Sr. No.	Title	Techniques used	Real Time System	BLEU score/Accuracy	Limitations
1.	Neural Machine Translation for English to Hindi [3]	Neural Machine Translation (NMT)	Yes	18.21(BLEU)	Fine-tuning of long and rare sentences using smaller data sets is not done.
2.	Hindi-EnglishSpeech-to-SpeechTranslation System for Travel Expressions [4]	Automatic SpeechTranslation, MachineTranslation and Text To Speech.	Yes	-	In this system,the speechrecognition iscarried out using HMM based acoustic models at the wordlevel.
3.	Syntax DirectedTranslator for Englishto Hindi Language [5]	Syntax Directed Translation.	No	60-70%	Accuracy can be increased.

International Journal of Engineering Research & Science (IJOER)
ISBN:[978-93-5391-287-1] ISSN: [2395-6992] [Vol-6, Issue-3, March- 2020]

4.	A hybrid approach forhindi-english machinetranslation [6]	Combined approach of PSMT, EBMT andRBMT is proposedto develop a novelhybrid data drivenMT system.	Yes	83-90%	Currently it is not giving the good resulton complex and multiple combinesentences it can be extended to be able toperform well for those sentences too.
5.	Statistical MachineTranslation System forIndian Languages [7]	Statistical Machine Translation (SMT).	No	-	The results can be better if the size of thecorpus is larger.
6.	Speech to TextConversion forMultilingualLanguages [8]	Mel-FrequencyCepstralCoefficient (MFCC),Support VectorMachine (SVM)	No	-	Evaluation techniques can be used for better accuracy of the system.
7.	Multilingual Speechto SpeechTranslation Systemin BluetoothEnvironment [9]	Automatic Speech Translation, MachineTranslation and Text to Speech.	Yes	-	The communication for translation totally dependson the availability of theBluetooth environment
8.	A Web BasedPunjabi to HindiStatistical MachineTranslation System [10]	Machine Translationusing SMT.	No	87-97%	It is a one-way TranslationSystem and can beimplemented in reversemode.
9.	Improvement ofEnglish-HindiMachine Translationusing ConceptNet [11]	ConceptNet	No	27.09 (BLEU)	Accuracy can beincreased.
10.	Large ScaleMyanmar to EnglishNeural MachineTranslation System [12]	Neural MachineTranslation.	No	Word Level21.88(BLEU) CharacterLevel23.92(BLEU)	It can be seen that theperformance ofcharacter-level neuralmachine translation isbetter than the Word levelneural machine translationsystem.
11.	Hybrid Machine Translation for English to Marathi [13]	Hybrid: Statistical translation, Rule based translation	No	-	At times system may generate translation withsynonymous words and reference translation do not containthem so system evaluation for automated translation fall invalue which is bug.
12.	Syntax Based Machine Translation using Blended Methodology [14]	Syntax based translation	No	58.85%	For better accuracy the system needs a smaller number of words.

International Journal of Engineering Research & Science (IJOER)
ISBN:[978-93-5391-287-1] ISSN: [2395-6992] [Vol-6, Issue-3, March- 2020]

IV. CONCLUSIONS

In this paper detailed study of several machine translation approaches and their techniques with respect to the BLEU scores has been done. This will help one to decide which approach to choose while translating from English to any Indian language. The accuracy of the model is measured with BLEU score stating how good the model is. With recent developments in Neural networks for Machine Translation one can create a model with faster speed and better accuracy. This can help in having a real time translation system with better translation quality. As digital literacy has been improving in rural areas, the cloud based translation model can also be deployed with NMT which can then deter the language barriers among people in India by having a real time speech to speech translation system.

REFERENCES

[1] https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers_in_India
[2] https://localizeblog.com/types-of-machine-translation/
[3] S. Saini and V. Sahula, "Neural Machine Translation for English to Hindi", IEEE Conference, Malaysia, September 2018, pp. 25-30.
[4] P. Vijayalakshmi, "Hindi-English speech-to-speech translation system for travel expressions", IEEE Conference, Chennai, India, September 2015, pp. 250-255.
[5] P. Kumar, S. Srivastava and M. Joshi, "Syntax Directed Translator for English to Hindi Language", IEEE Conference, Kolkata, India, March 2016, pp. 455-459.
[6] O. Dhariya, S. Malviya and U. Tiwary, "A hybrid approach for Hindi English machine translation", IEEE Conference, Vietnam, April 2017, pp. 389-394.
[7] B.N.V Raju and M. S. V. S. Raju, "Statistical Machine Translation System for Indian Languages", IEEE Conference, Bhimavaram, India, August 2016, pp. 174-177.
[8] Y. Ghadage and S. Shelke, "Speech to Text Conversion for Multilingual Languages", IEEE Conference, Melmaruvathur, India, November 2016, pp. 236-240.
[9] M. Faizullah, A. Shaji, T. SivaKarthick, S. Vivek and A. Aravind, "Multilingual Speech to Speech Translation System in Bluetooth Environment", IEEE Conference, Kanyakumari, India, December 2014, pp. 1055-1058.
[10] A. Kaur and J. Rani, "A Web Based Punjabi to Hindi Statistical Machine Translation System", IEEE Conference, Chandigarh, India, April 2016.
[11] M. Bansal and G. Jain, "Improvement of English-Hindi Machine Translation using ConceptNet", IEEE Conference, Noida, India, May 2018, pp. 198-202.
[12] Y. ShweSin, K. Soe and K. Htwe, "Large Scale Myanmar To English Neural Machine Translation System", IEEE Conference, Nara, Japan, December 2018, pp. 464-465.
[13] P. Salunkhe, A. Kadam, S. Joshi, S. Patil and D. Thakore, "Hybrid machine translation for English to Marathi: A research evaluation in Machine Translation: (Hybrid translator)", IEEE Conference, Chennai India, November 2016
[14] S. Singh, A. Kumar, P. Sahu and P. Verma, "Syntax based machine translation using blended methodology", IEEE Conference, Dehradun, India, March 2017, pp. 242-247.