Evaluation and Error Analysis of English-Telugu Neural Machine Translation Output

A dissertation submitted to the University of Hyderabad
for the award of the degree of

Master of Philosophy
in
Applied Linguistics

by

Danaveni Madhukar
Reg. No: 20HAHL02

Supervisor

Dr. K. Parameswari

Centre for Applied Linguistics and Translation Studies
School of Humanities, University of Hyderabad, Hyderabad, INDIA
December, 2022

Center for Applied Linguistics and Translation Studies
School of Humanities
University of Hyderabad

DECLARATION

I hereby declare that the work embodied in this dissertation entitled **"Evaluation and Error Analysis of English-Telugu Neural Machine Translation Output"** is carried out by me under the supervision of Dr.K.Parameswari, Centre for Applied Linguistics and Translation Studies, University of Hyderabad, Hyderabad, and has not been submitted for any degree in part or in full to this university or any other university for the award of any degree or diploma. I hereby agree that my thesis can be deposited in Shodhganga/INFILBNET.

A report of plagiarism statistics from the Indira Gandhi Memorial Library, University of Hyderabad is enclosed.

Danaveni Madhukar
20HAHL02

Dr. K. Parameswari
Supervisor
Centre for Applied Linguistics and Translation Studies
School of Humanities

Centre for Applied Linguistics and Translation Studies
School of Humanities
University of Hyderabad

CERTIFICATE

Dated - 30/12/2022

This is to certify that **Danaveni Madhukar** has carried out the research-work embodied in the present dissertation entitled **"Evaluation and Error Analysis of English-Telugu Neural Machine Translation Output"** at the University of Hyderabad. The dissertation represents his independent work and has not been submitted for any research degree of this university or any other university. The following papers were published during this period:

1. Presented a paper titled "Evaluating English-Telugu Machine Translation Output", in the *43rd International Conference of the Linguistic Society of India* hosted by the Central Institute of Indian Languages, Mysuru from 21-23, December, 2021.

Further, the student has passed the following courses towards the fulfilment of the coursework requirement for M.Phil.

Course Code	**Name**	**Credits**	**Pass/ Fail**
AL701	Research Methodology	4.00	Pass
AL702	Current Trends in Applied Linguistics	4.00	Pass
AL721	Advanced Topics in Applied Linguistics	4.00	Pass
AL724	Research Oriented Readings	4.00	Pass

Dr.K.Parameswari

Supervisor	Head of the Department	Dean
CALTS	CALTS	School of Humanities

Acknowledgement

I would like to express my heartfelt gratitude to my supervisor Dr K Parameswari for her patience and guidance throughout the research period. I could not have succeeded in this task without her expertise and knowledge shared with me. I would say certainly that she is an inspiration for me at every stage of my research.

I am also grateful to Prof. Uma Maheswara Rao. I am thankful to my RAC member Prof. S ArulMozi for his support and insightful comments. I should also extend my thanks to the faculty members of CALTS, Prof. Bhimrao Panda Bhosale, Prof. J. Prabhakara Rao, Prof. K Rajya Rama, Dr. Gracious Mary Temsen Dr. Morey Deepak Tryambak, Dr. Annem Naresh and entire other faculty members. I would thankfully acknowledge the invaluable comments of Prof. Uma Maheshwara Rao during my Pre-submission presentation which enabled me to improve the thesis further.

Here, I would like to acknowledge the office staff of the CALTS, Mr. Murthy, Ms. Swati and Mr. Mallesh for their cooperation. Additionally, I would like to mention the administrative staff, UoH for all the financial support and cooperation.

My Special thanks to Sangeetha Perugu for the editing, proofreading and technical support in completing this dissertation.

Many thanks to Shilpi Harish, P Sreenu, G Arun Prakash and Venu and Naveen for their help during the research tenure. I am also thankful to the research participants and other friends who helped me directly or indirectly in completing this research.

Lastly, I am deeply indebted and grateful to my parents & siblings for their love, care and support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

The modern world has seamless information coming from different parts of the world. Machine Translation (MT) is an automatic translation task in which one natural language is translated into another natural language. MT helps to overcome the language barrier and helps in accessing information in the native language. Language data can be fed to machines in different forms such as text, speech or image and they can be translated into multiple languages output using MT. The use of MT services has been spread across almost all domains. As MT gets popular, research in the area has also become a need of the hour. In a multilingual country like India, there is a huge scope for the development of MTs for Indian languages. Attempts were made to develop MTs for Indian languages. Different types of MTs were developed using various frameworks: Rule-based, Statistical and Neural Machine Translation (NMT). The present study focuses on evaluating the translation of open-source English-Telugu NMT-based systems and understanding their efficacy and failures. This study also presents the nature of errors found in different Neural Machine Translation (NMT) which affect the overall accuracy, fluency and comprehensibility of the output.

1.2 Aims and objectives

The primary aim of the research is to build evaluation criteria for the English-Telugu MTs which can be achieved by following objectives:

1. Exploring the existing open source Neural Machine translation systems for English-Telugu and choosing the MTs for the evaluation process.

2. Reviewing the available human and automatic evaluation methods and formulating the evaluating criteria and scale so as to adopt the study.

3. Evaluating the output of the MTs to find the efficiency of the MTs in terms of adequacy, comprehensibility, fluency and BiLingual Evaluation Understudy (BLEU) score.

4. Finding linguistic errors in the MTs' output and classifying the errors to understand areas of further improvement of the MTs.

1.3 Machine Translation: An Overview

The field of machine translation has its roots in the early 17th century. During this time, Rene Descartes proposed the idea of universal language, i.e., the same meanings can be expressed in different languages by sharing one symbol (Yang et al., 2020). But the remarkable endeavours for the development of MTs could be seen from World war-II. (Weaver, 1952) came up with the proposal for computer-based machine translation based on information theory. From the 1950s to 1990s, Rule Based Machine Translation (RBMT) systems played a dominant role in the development of MT. During this period, research institutes and commercial companies showed their interest in building the RBMT. In 1954, Georgetown University, along with the cooperation of International Business Machines (IBM) computer manufacturers, built a Russian-English MT. After this, a machine translation company called Systran launched a commercial Rule-Based MT and it became the most successful translation system. From 1990 to 2014, Statistical Machine Translation (SMT)systems played a major role in the field of MT. Giza and Giza++[1] are examples for word-based SMT systems. To overcome the shortcomings of word-based MTs, in 2003, phrase-based SMT systems came into the limelight and succeeded in producing better quality output in comparison with earlier SMT systems. The quality output of SMTs attracted researchers to carry out further advanced research in the area by building different SMT models like factored SMT, which make use of morphological information, and syntax-based SMT, using parsing trees (Wang et al., 2021). Since 2014 to till date, the field of MT is predominantly ruled by . Efforts on building RMTs, SMTs and NMTs for western languages are seen in many literatures. However, considerable efforts are exerted in academic research and industries involving languages like Telugu. In this study, we mainly focus on identifying available NMTs for English to Telugu and evaluating their output, thereby presenting their efficiency in real-time usage.

[1] http://www2.statmt.org/moses/giza/GIZA++.html

Along with these, we also classify different types of linguistics errors found in each NMTs to justify the results and suggest the points of improvement in their output.

1.4 Approaches to Machine Translation

MTs can be developed using majorly three approaches: Rule-Based, Statistical-Based, and Neural-Based approaches. In this section, we discuss briefly these approaches followed by other approaches.

1.4.1 Rule Based Machine Translation (RBMT)

The rule-based machine translation is the first traditional approach in the field of MT. It is also referred to as Knowledge-Based machine translations. It is a translation framework, largely, dependent on linguistic rules to assist a machine to translate Source Language (SL) ()tl). As RBMT mostly relies on the linguistic knowledge of humans on any given language, it can be referred to as knowledge-based machine translations. The translation from the SL to TL takes place at different levels such as Analysis, Transfer and Generation as expressed in Vanquois tringle. (Jurafsky, 2000, pg-999) which can be seen below in figure-(1.1).

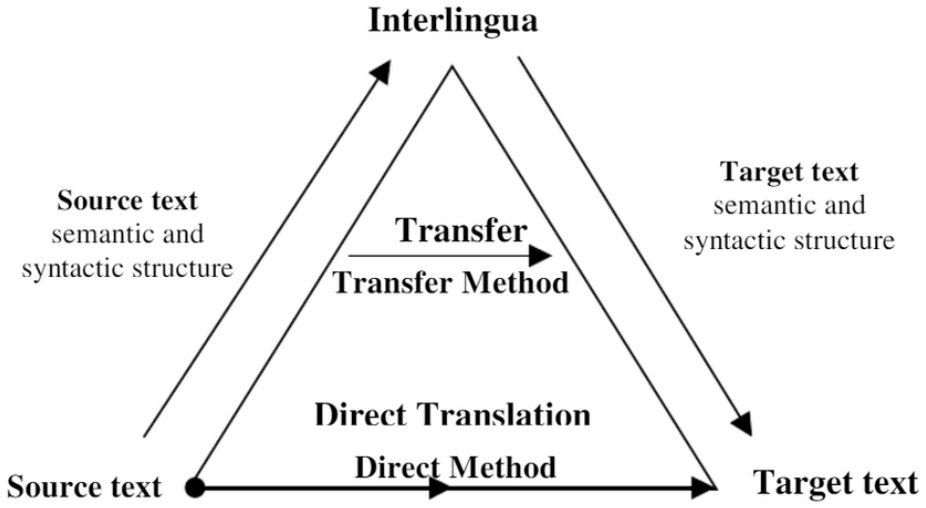


Figure 1.1: Vauquois triangle representing various methods of RBMT approach

1.4.1.1 Direct Method

In the direct method, the source language is directly translated into the target language based on the resources available and there is no intermediate

representation. Bilingual dictionaries are primary resources that MT uses to translate word to word. No information about syntactic or semantics rules are applied in the entire translation process (Maučec and Donaj, 2019). These are mostly first-generation MT systems. Example systems: Anusaaraka MT (Bharati et al., 1997), IBM's Russian to English and English to Russian MTs (1954) (Garje and Kharate, 2013).

Pros and cons:

As it mainly relies on bilingual dictionaries, it only needs these dictionaries as major resources in one-to-one word mapping during the translation which is easy to represent and also easy to implement compared with other methods. The major drawback of the approach is its inability to represent the contextual meaning of a word and it may not be possible to get correct equivalences in every language. Since everything has to be done manually, it needs a lot of human effort and is expensive.

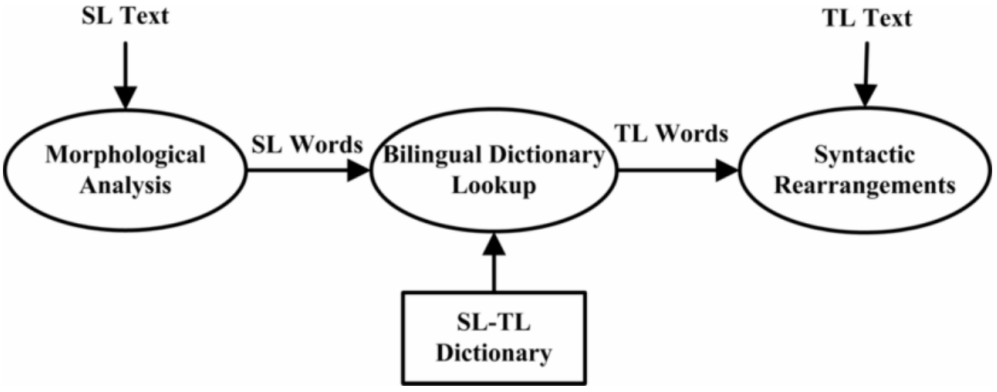


Figure 1.2: Direct Machine Translation Method (Saini and Sahula, 2015)

1.4.1.2 Transfer Method

This method came into operation around the 1960s and was used in the second-generation machine translation system (Tripathi and Sarkhel, 2010). The method is used in the translation of divergent languages. In a transfer-based method, translation takes place in three stages: Analysis, Transfer and Generation. In the analysis, a given source language is analyzed based on its grammatical rules. In the transfer stage, manually drafted transforming linguistic rules made the analyzed source text feasible to convert into the target language. In the generation

stage, the transformed text is translated into the target language according to the grammatical rules of the target language. E.g. (T: - NP V); If the object follows the main verb, the above transforming rule swaps them in order to get the target structure (Bhattacharyya, 2015). The transformation can not go up to the tip of the vauquois triangle but it goes up to certain levels that are syntactic or semantic only. E.g. SHAKTHI IISC Bangalore and MATRA Center for Development of Advanced Computing (CDAC), Pune (Gehlot et al., 2015).

Pros and cons:
since every language has its own grammatical rules, it is very feasible to draw transfer rules for any given language. It works out well for the linguistically closely related languages. But, simultaneously, every natural language has countless grammatical rules and its own colloquial expressions and dialectical variance that are possible only in that particular language, not others. Hence, It is highly difficult to draw linguistic rules for each and every language manually.

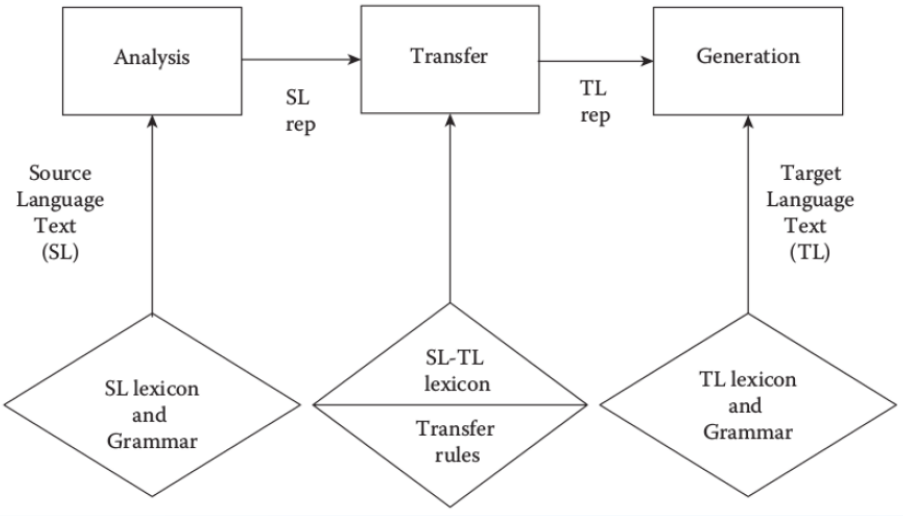


Figure 1.3: Transfer Based Machine Translation Method

1.4.1.3 Interlingual Method

The word "Interlingua" originates from the Latin language. "Inter" means between or intermediary and "lingua" means language. This is an advanced and efficient method when compared to the earlier two methods. It happens at the tip of the vauquois triangle. Here, the meaning of the source text is directly represented using artificial language by taking into consideration lexical, structural and discourse knowledge. Using all this information, the machine tries to disambiguate the meaning representation. And then the meaning representation is generated

into the target language (Bhattacharyya, 2015). E.g. *Anglabharati*[1] MT (1995), KANT MT system (Nyberg et al., 1997).

Pros and cons:
This can be suitable for any language in the world as it aims towards the direct representation of the source text meaning which is the final goal of translation. But it also has a drawback in how to analyze the source language in order to achieve the same meaning to represent it in the target language. And to do that one really needs to understand all human languages which is a highly impossible task.

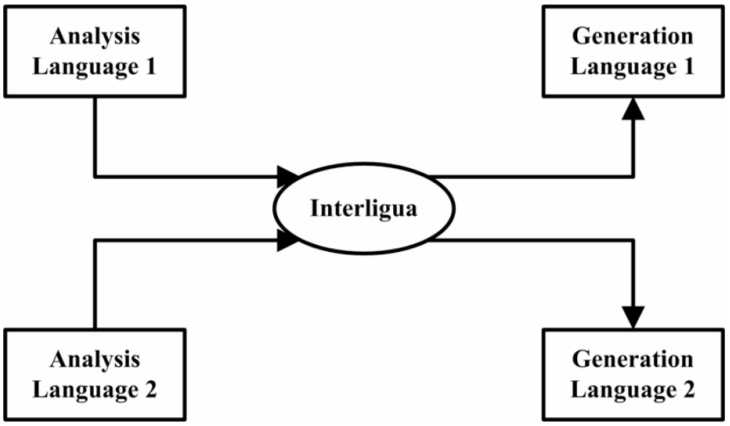


Figure 1.4: Interlingua Machine Translation Method (Saini and Sahula, 2015)

1.4.2 Corpus Based Machine Translation (CBMT)

Corpus-Based Machine Translation(CBMT) is superseded by the RBMT. It became popular because of the translation efficiency and accuracy of the output. It is also referred to as data-driven machine translation. These MTs make use of the available larger set of data in order to produce the MT output, and the required knowledge is drawn from the corpus to translate the source language into the target. Induction of the CBMTs has produced a large number of translated text within less time.

1.4.2.1 Statistical Machine Translation (SMT)

Statistical Machine Translation model is a data-driven model. SMT is a mechanical framework which predominantly works on statistical and probability

[1] https://www.cse.iitk.ac.in/users/rmk/mission/mission.htm

methods. SMTs do not need linguistic rules for translation. SMTs need a large amount of parallel corpus to train them. The source language can be calculated based on the usage of frequency of the phrase occurrence using probabilistic methods. It takes less effort from the linguists because the MTs can acquire suitable information through statistical analysis from bilingual corpora. For that reason, it is also referred to as a type of corpus-based MTs. There are different SMT models available such as word-based, Phrase-Based and tree-based models (Koehn, 2009). Among the SMT-based MTs, phrase-based MT is one of the most successful models. It has a key component in its framework called phrased-based lexicon which allows source text phrases to be translated into the target language (Chéragui, 2012). The most possible translation can be achieved using the below mathematical formula.

$$\hat{T} = argmaxtTP(T|S) = argmaxtTP(S|T)\ P(T)$$

$\hat{T}$ = most possible translation, S = source language, T = target language

P(S|T) = translation model and P(T)= language model

Figure 1.5: SMT Formula

The advantage of the approach reduces the human effort in translating, and extracting information from the corpus itself. The major drawback of SMT is finding a large parallel corpus to train it. No linguistic information is available. And it is very difficult to find a huge parallel corpus for low resourceful languages.

1.4.2.2 Example Based Machine Translation (EBMT)

EBMT uses a matching technique in the translation process where the input is given to the MT as a Source Language. In which the translation will take place from sentence to sentence by mapping word to word. The main feature of the MTs is they are endowed with a huge memory capacity which is loaded with a massive corpus as Examples. When we try to translate the text, the system refers to the existing data which is already available in the memory (Bhattacharyya, 2015). It makes use of the given examples in order to produce the target language. This translation process happens in three phases: matching, alignment and recombination. In the matching stage, the system tries to find out the similar equivalence for the source text in the database. In the alignment stage, once the suitable match is bound in the database

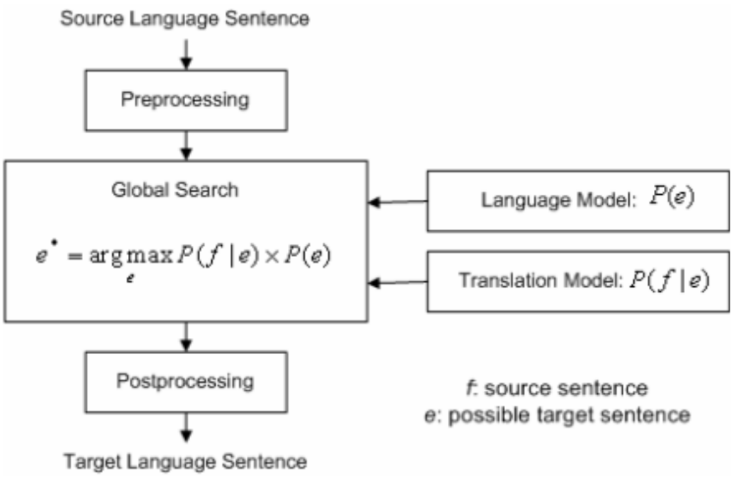


Figure 1.6: Architecture of SMT system (Nguyen and Shimazu, 2006a)

then that identified part of the sentence is aligned in comparison with the other examples. In the recombination stage, the aligned parts are combined according to the target language rules to produce the output (Sanyal and Borgohain, 2013). The main advantage of the approach is it works well with a small amount of dataset and produces the output within a no time delay. And also have disadvantages such as building huge parallel data is so difficult and storage. E.g. *vāsānubada* (Vijayanand et al., 2002), *AnuBharti* (Sinha, 2004).

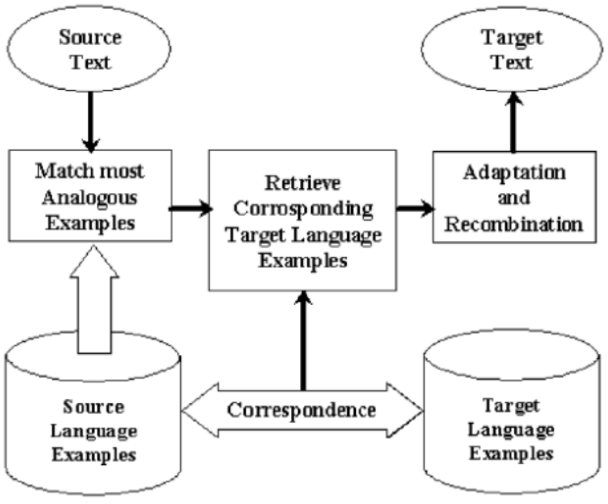


Figure 1.7: the architecture of EBMT system (Sinhal and Chandak, 2012)

1.4.3 Hybrid Machine Translation (HMT)

The main motive behind building Hybrid Machine Translation (HMT) was to improve the performance of MTs by combining rule-based and statistical-based MTs. This includes linguistic-centered approach (RBMT) and non-linguistics approaches (SMT & EBMT). Hybrid MTs could be classified into two types : hybrid systems, which are directed through rule-based systems, can make use of statistical perspective to generate the output. And a system directed by statistical-based MTs can make use of rules in pre and post-processing of the data in order to generate output by compensating each other's drawbacks (Maučec and Donaj, 2019).

1.4.4 Neural Machine Translation (NMT)

Since the research predominantly focuses on the evaluation of NMT approach-based machine translation output, it is important to understand a little elaborately about the architecture of NMT. Broadly, NMT comes under the umbrella of machine translation. This approach is considered to be state-of-the-art in MTs at present. MT, as a whole, is a subsection of Machine Learning. In the 2nd chapter, we will elaborately discuss NMT as this is the primary focus of the study.

1.5 Review of Evaluation of Machine Translation Systems

Once the output is produced by MTs, due to the limitations of MTs there might be issues on output quality. To measure the quality of the output, parameters and different techniques have been drafted since the 1960s. Automatic Language Processing Advisory Committee (ALPAC)(ALPAC, 1966) report discusses the manual evaluation parameters: intelligibility and fidelity along with the quality scale of 9 points on which both parameters could be measured for the first time in the history of MT systems. Thereafter, the Advanced Research Project Agency (ARPA) proposed an evaluation method in 1991, which suggests various methods: comprehension, accuracy and adequacy evaluations. Evaluators were asked to score them on a 1-5 point evaluation scale in order to evaluate the output. Furthermore, TDIL-DeitY had come up with a 5-point scale for evaluating accuracy. And this five scale got edited by (TDIL, 2014). for calculating the comprehensibility and fluency. Language Consortium Data (LDC) also developed a

five-point scale to measure the performance of MTs for National Institute of Standards and Technology (NIST) evaluation Workshop. (Specia et al., 2011) came up with a 4 point-scale to measure adequacy. There are also other evaluation methods like task oriented (White, 1995), Human Translated Error Rate (Snover et al., 2006), Segment Ranking (Callison-Burch et al., 2012) etc. All these human evaluation methods have been developed in measuring the quality of MT outputs in the past. Though human evaluation methods have succeeded in measuring performance, there exist some disadvantages: time-consuming, expensive and finding translation experts. All these provided a basis for building automatic evaluation methods for MTs. Word Error Rate (WER) (Su et al., 1992a) uses the word order of the output sentence to evaluate MTs based on operations of word addition, word deletion and word replacement. But it failed in practice and provided very low results. BLEU (Papineni et al., 2002) is a widely used automatic evaluation metric. This method uses the reference translation produced by humans and output produced by machine based on N-grams overlapping. There are also different automatic metrics like Metric for Evaluation of Translation with Explicit Ordering (METEOR) (Banerjee and Lavie, 2005), Assessment of Text Essential Characteristics (ATEC) etc.

1.5.1 Review of Evaluation of Foreign MTs

(Matsuzaki et al., 2015) conducted a study on evaluation of Eng-Japanese MTs. In which, 40 dialogues from a short conversation were translated into English using Google and Yahoo MTs, and two human translators. The output was given to five Japanese speakers, for whom English is the second language,who were asked to provide a ranking for the human evaluation. And the same output was also evaluated using automatic evaluation methods like BLEU, BLEU+1, Rank-Based Intuitive Bilingual Evaluation Score (RIBES) and Translation Error Rate (TER). The agreement rate between automatic and human evaluation was nearly 90%. The recent studies mainly focused on inter machine translation systems: SMT & NMT based MTs. (Stasimioti and Sosoni, 2020) has conducted a comparative study of performance between English - Greek language GSMT and GNMT systems. The study uses the human method: adequacy and fluency and sixteen postgraduate translation students for error analysis. And the automatics metrics: BLEU, Word Error Rate (WER) and TER were used to evaluate the MTs. the results were NMT outperformed SMT.

1.5.2 Review of Evaluation of Indian Language MTs

(Goyal and Lehal, 2009) performed an evaluation study on Hindi-Punjabi MT which is built on a direct translation of source text to target text. The data set consists of sports, politics, travel etc. For the evaluation, the study uses methods: Intelligibility, accuracy and word error rate. In addition to this, it adopted the 4-point scale to score the quality of the Punjabi output. More than 50 participants participated which included both Hindi and Punjabi language evaluators. The system performed with 95.12 % accuracy. The study does not cover the linguistic error analysis. A diagnostic study was conducted by (Balyan et al., 2013) in 2013 from IIT, Delhi, India. The data set consists of 1000 sentences from the tourism domain. The evaluation was performed on the output of the five English to Hindi MTs. for which, an automatic evaluation tool called DELiC4MT was used to evaluate MTs using linguistic checkpoints as phrase level, Named Entity (NE) and Hjerson's word order, inflection and 18 different checkpoints. Google had outperformed the other four MTs. (Kalyani et al., 2014) has conducted a study on evaluating the performance of the Hindi to English machine translations. For the study, 10,000 sentences were collected and translated from Hindi to English using three different MTs: Google, Bing and Babylon. Manual evaluation has been done using a 5-point scale and parameters. BLEU, TER, METEOR etc. were employed to measure the MTs. But the results were not very impressive and concluded by suggesting that deeper evaluative strategies are required. (Ramesh et al., 2020) has assessed English-Tamil and Hindi-Tamil SMTs and NMT systems. The study employed BLEU method and error analysis. They achieved a very low Bleu score and found different types of errors: word order, finding equivalent domain terms, lexical selection etc.

1.5.3 Review of Evaluation of English to Telugu MTs

(Ojha et al., 2018) conducted a study as part of Workshop on Machine Translation (WMT) 2018 shared task to evaluate English to Indic languages MT systems developed by the RGNLP team. The team conducted a comparative study between the phrase Based statistical Machine Translation (PBSMT) and NMT system. To perform it, they have chosen different English to Indic languages systems. Among which, English to Telugu PBSMT and NMT systems performance were compared by employing the BLEU, RIBES and Adequacy-Fluency Metrics (AMFM) automatic metrics . In which, the English-Telugu PBSMT system produced the second highest BLEU score with 42 , while NMT has produced average results around 15. In other metrics also both MTs had performed on

average. Human evaluation conducted only for English to Hindi and Hindi to English MTs only. To the best of my knowledge, there would be no comparative study conducted on evaluation of English-Telugu NMT systems so far using human and automatic metrics to determine the performance of the systems and error analysis on the output to find out the linguistic errors.

1.6 Methodology

The current research requires a corpus, MTs to evaluate, evaluation techniques and methods for error analysis. They are discussed here.

1. **Corpus collection:**
   The corpus for evaluating MT systems has been collected from () by Technology Development for Indian Languages (TDIL) from the Health domain consisting of 2000 sentences.

2. **Open-source MTs:**
   The English corpus is given as input to the open source English-Telugu MTs such as Google, Bing, Lingvanex, Indian Institute of Information Technology (IIIT-H) MT system and Yandex. Then the output is arranged into different sets of documents using google forms; each set contains 2000 sentences.

3. **Evaluation process:**
   Human and automatic evaluation methods are used for the evaluation of the output. Six volunteer evaluators participated in the human evaluation process; three are monolinguals and the other three are bilinguals. With the help of the evaluators, we carried out two kinds of evaluation: blind and open evaluation. The evaluators marked their responses based on the proposed five-scale parameter(see chapter 3 for more details). For automatic evaluation, BLEU score evaluation is employed to compare the original text with translated text.

4. **Error analysis:**
   Finally, we have carried out the error analysis of the output to classify the errors at various linguistic levels. We have devised a taxonomy of error types based on MT output and evaluation.

1.7 Limitation of the study

The current study is confined to the evaluation of only 5 open sources English-Telugu Neural Machine Translation Systems: Google Translate, Bing Translate, IIIT-H Translate, LingvaNex and Yandex Translate. the MT field is considered to be one of the most dynamic areas as it keeps adopting cutting-edge technologies and feeding incessant language input to the systems. The evaluation results which is done as part of the study might not be the same and the results might be changed and different in the upcoming years. The classified error taxonomy is done based on very less data that is 2000 sentences only. As the input data is very limited, it is difficult to cover all aspects of the language and analyse the errors. These are the major limitations of the current study.

1.8 Chapterization

The chapters are arranged in the dissertation as follows:

Chapter 1 is "Introduction" in which, a broad picture of the Machine translation systems like its approaches, types and its review of literature on English to Telugu Mt systems are provided.

Chapter 2 is "Neural Machine Translation system in English - Telugu." in which, a brief introduction to machine learning, neural network, Neural Machine Translation systems and English-Telugu available translation systems.

Chapter 3 is "Evaluation: Methods and Methodology." in which existing human and automatic evaluation methods and methodology of the current study are discussed elaborately.

Chapter 4 is "Evaluation of Machine Translation Systems" in which evaluated results are discussed.

Chapter 5 is "Error analysis of Neural Machine Translation." in which errors are classified based on linguistic and non-linguistic aspects.

Chapter 6 is the "conclusion". which comprehends the entire study.

Chapter 2

Neural Machine Translation Systems for English-Telugu

2.1 Introduction

This chapter provides a brief review of machine learning and types of learning methods explained along with diagrams. It also discusses Neural Machine Translation (NMT) and different types of architectures of the MTs like Reccurent Neural Network (RNN), Convolutional Neural Network (CNN), Self attention and Transformer based machine translation model. The available open source English-Telugu MTs are introduced concisely. Among which, some of them are developed by research academic institutions, most of them are developed by private technological companies. The top five best performing open source English-Telugu MTs are considered for this study.

2.2 Machine Learning

"Machine Learning (ML) is a programming computer to optimize performance criterion using example data or past experience" (Alpaydin, 2020) i.e A machine learns based on the available data in the database using a model. The input data might be labeled or not be labeled. That learning can be done through using algorithms which are built employing the statistics theories and also mathematical formulas. Algorithms work up on the input data to resolve problems or to draw logical and suitable inferences in order to predict the future or to acquire the required information to produce output.

2.2.1 Types of Machine Learning

Machine learning can happen predominantly in the following four ways: supervised, unsupervised, semi-supervised and reinforcement learning.

2.2.1.1 Supervised Learning

Supervised learning is achieved through feeding the machine with a set of labelled data pairs, viz. , input and output. The machine has to understand core features of the given data-based algorithm provided for learning. Through which that particular algorithm predicts the output. Examples: stock market prediction, face detection etc. KNN means K- nearest neighbor. It is a non-parametric algorithm used in supervised machine learning techniques. It classifies the new data based on existing data in the machine and most similar classes it falls into. where, K is always number which is decided by the programmer. And the nearest neighbor is decided by euclidean distance or Manhattan distance. This is used in supervised machine learning. The algorithm is employed in classifying input data and regression. It resembles the tree-like structure, in which there is a root node and sub nodes. Sub-nodes contain decision and leaf nodes. Decision node contains multiple nodes and it goes on until it ends with a leaf node. Leaf node is the final node which gives output or decision.

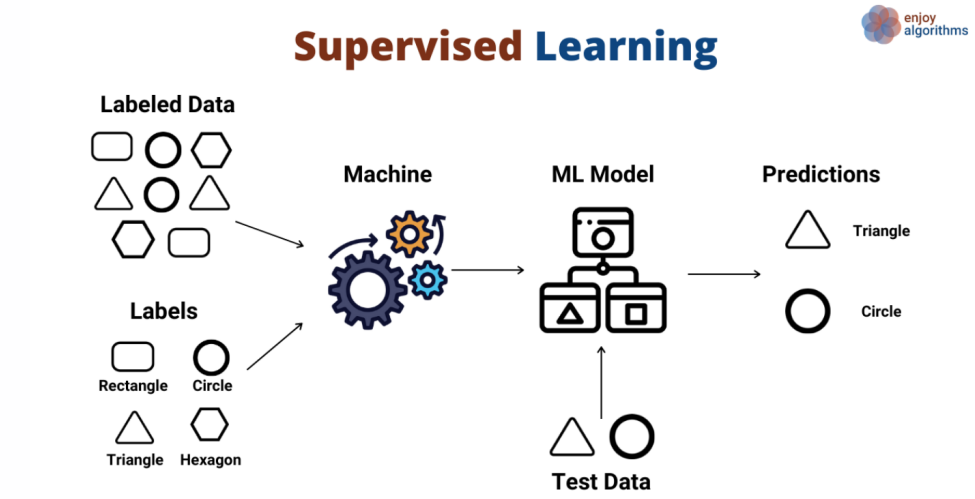


Figure 2.1: Supervised Learning framework

2.2.1.2 Unsupervised Learning

In this model, a set of unlabelled data, i.e. input, would be provided to the machine. The machine tries to identify the similarly existing patterns that mostly occurred generally which is called density estimation by using the algorithm from proved data. The motive behind the modal is to process large amounts of data that is

available in the real world or "big data". Example: product segmentation, customer segmentation etc. K-means is an unsupervised machine learning algorithm. Which is employed to create clusters from given unlabeled data as input to a machine. In which, K denotes the number of clusters to be created and acts as a random center data point. The closest data points can be assigned to the center point to make clusters. The distance is decided using the euclidean distance method.

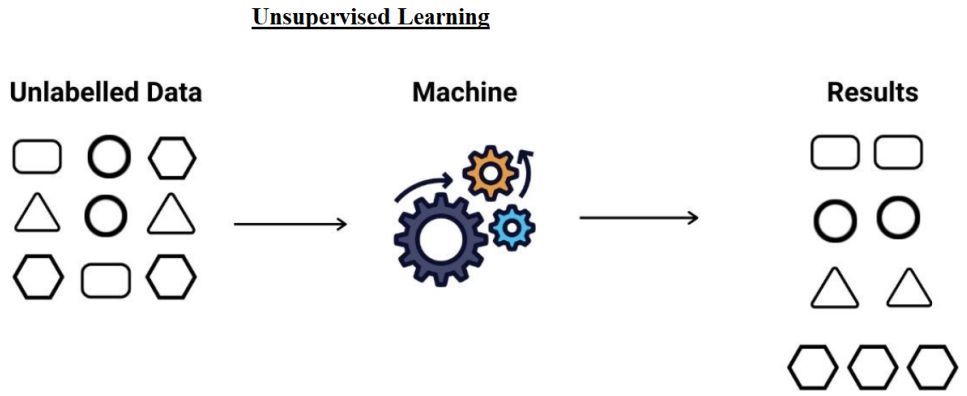


Figure 2.2: Unsupervised Learning framework
(Source;www.enjoyalgorithms.com)

2.2.1.3 Semi-Supervised Learning

This model is a combination of supervised and unsupervised learning models. In which a machine is provided with a labeled data set as well as unlabeled data. The model tries to use the supervised learning algorithm to label the unlabeled data. Examples: studying medical images, speech analysis and Machine translation etc.

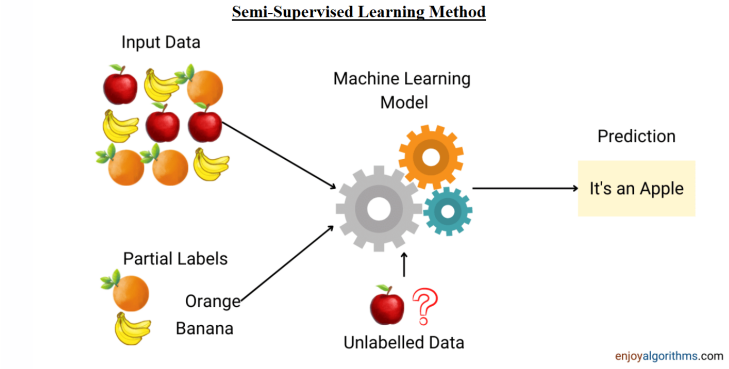


Figure 2.3: Semi-supervised Learning framework
(Source:www.enjoyalgorithms.com)

2.2.1.4 Reinforcement learning

This learning model uses trial and error methods to come up with the highest probable way to reach the solution. This takes place in an uncertain and complex environment where more options are possible. If it comes up with the highest possible way that can be rewarded otherwise a penalty can be imposed. The method is predominantly employed in building gaming applications. Example: Chess and other game applications. Q-learning is a reinforcement learning method. In which Q refers to the quality of the action initiated by an agent by employing the Bellman equation.

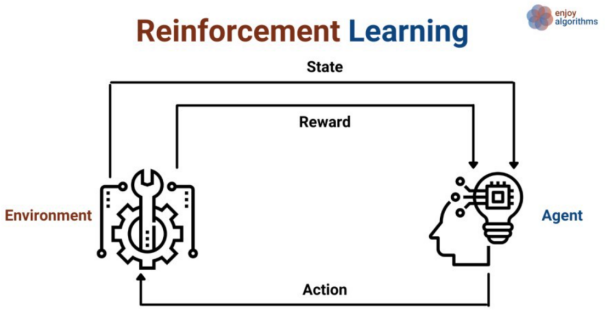


Figure 2.4: Reinforcement supervised Learning framework
(Source: www.enjoyalgorithms.com)

2.3 Neural Network and Neural Machine Translation (NMT)

Neural network is a network of artificial neurons in which every neuron interconnects with the other and creates multiple connections (Rebala et. al, 2019). A basic unit of artificial neuron resembles the biological neuron as shown in the figure below:

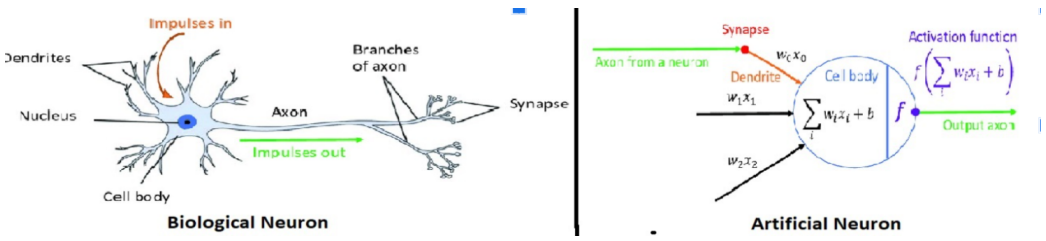


Figure 2.5: Resemblance of biological and artificial neuron
(Roffo, 2017).

Every artificial neuron interconnects with other neurons and forms a complex network called the neural network. Neural networks can be adopted in supervised machine learning for classification of the data and unsupervised machine learning in clustering a given data in the process of training machines. Neural networks are considered to be a part of deep learning. Many deep learning based applications like speech recognition and face recognition adopt the neural network technique for getting the best results and it succeeded in meeting the desired outputs as well. The reason behind the success of deep learning is that it contains neural network architecture which enables it to be capable of dealing with many complex problems processing it through the multi-layer networks (figure:2.6).

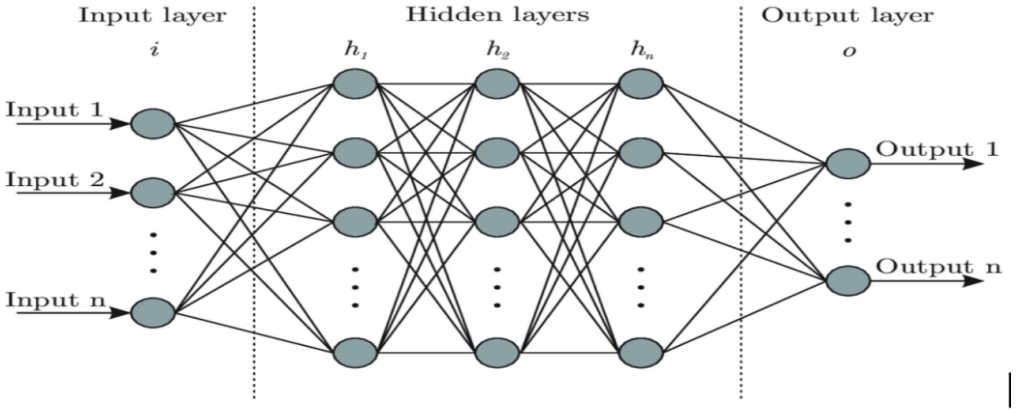


Figure 2.6: Architecture of Neural Network
(Bre et al., 2018)

The multi-layer neural network is also referred to as Deep Neural Network (DNN). And this has the capability of computing large amounts of data with high-speed. The technique had been adopted for various applications. One of the applications that adopted the DNN technique is the machine translation field. Neural Machine Translation approach is a recently developed computational framework. It is considered a superseded version of statistical MTs. NMTs have outperformed aforementioned traditional approaches which predominantly depended on trial and error methods. NMT employs Deep neural network technique. In which, the words are represented in the form vector representation. In comparison with SMTs, there would be no separate components: language model, translation model and reordering model. NMTs are built on a single sequence model which creates a large neural network. Which predicts one word at a time in order to produce the output. The prediction precision of NMT is often so high (Maučec and Donaj, 2019).

2.3.1 Architecture of NMT

Basic architecture of NMT contains two key components: Encoder and Decoder.

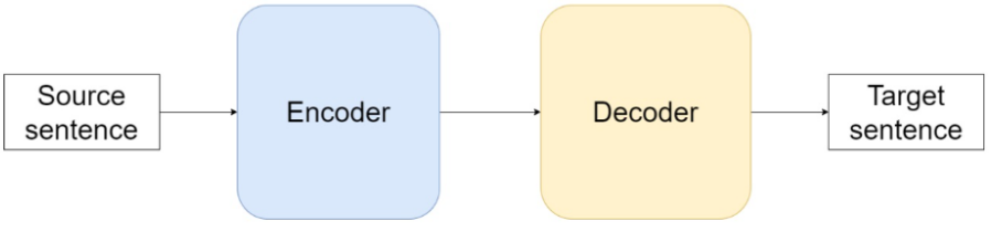


Figure 2.7: NMT Architecture
(Yang et al., 2020)

Every modern NMT contains these two components, which is considered a classic and original structure. (Kalchbrenner et al., 2014) and (Bahdanau et al., 2014). proposed the NMT structure that is inspired by a neural language model. Encoder and decoder take different parts of the translation, i.e. , the encoder initiates the translation then decoder concludes the translation process. Encoder is provided with source language as input. Then the source text reads word by word in sequence in order to convert words into vector values. This conversion takes place in the hidden layers. This whole process that happened in the encoder is referred to as encoding. Then, the encoded vector values are provided to the decoder to initiate the exact reverse process that takes place during the encoding. The vector representations are directly translated into target language. So there will be no visible representations in between the encode and decoder. This process is called end-to-end translation (Yang et al., 2020).

2.3.2 Types of NMT Architectures

In building encoders and decoders, DNN methods: Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), RNN and attention hybrid and self-attention or transformer methods are used. All these methods are employed, in order to make the NMTs more efficient in producing output (Tan et al., 2020). Based on these methods NMTs are categorized into three types: RNN based NMT, CNN based NMT and Self Attention Network or Transformers.

2.3.2.1 RNN based NMT model and Architecture

Sutskever et al. (2014) proposed the first RNN based NMT model in an effort to build the pure deep RNN based model. The model generated output as equal to the state-of-the-art of the SMT system at that time. Since then, remarkable efforts have been put forward by researchers aiming to build the RNN based NMTs. Steadily the NMT took a dominant position and became a state of the art translation model in the field of MT. Google has adopted this as their core model. Consider the figure below:

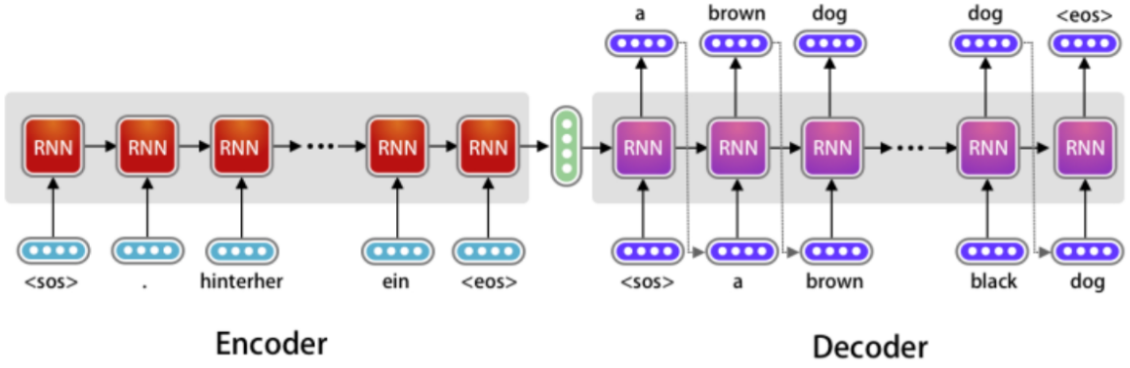


Figure 2.8: RNN based NMT model
(source:http://www.adeveloperdiary.com/).

2.3.2.2 CNN based NMT model and Architecture

As further developments take place in RNN based models, CNN based NMT are also introduced in the MT field. The performance of the NMT was very poor, in an earlier stage, in comparison with RNN based NMTs. To improve the quality of the output, researchers tried to build hybrid models: Kalchbrenner Blunsom's (2014) model contains CNN based encoder and RNN based decoder and Cho et al. tried a similar hybrid model. Kaiser et al. proposed a complete CNN based NMT model. Kalchbrenner et al.(2014) came up with CNN based NMT referred to as ByteNet NMT, which is performed well at character level translation only. RNN encoder NMT, proposed by Gehring et al. (2017) achieved the equivalent performance with then RNN based NMTs. The advantages of the model is that it has the ability to solve more complex problems with high training speed (Yang et al., 2020).

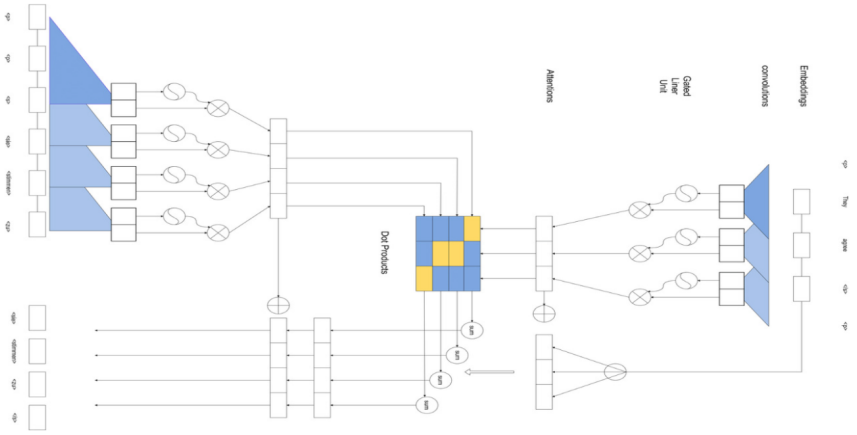


Figure 2.9: CNN based NMT model
(Yang et al., 2020)

2.3.2.3 Self-Attention based NMT or Transformers and Architecture

The proposal for building the self attention base NMT is put forward by viswani et al. (2017). This model is considered as state of the art in existing NMTs. The model is completely dependent on self- attention networks. The design takes advantage of both the RNN and CNN modals. Transformers contain encoder and decoder blocks. Encoder block consists of 6 similar components. Every component contains one multi-head layer and two sub-layers, which are equipped with layer normalization and residual connection. Decoder block consists of 6 components, in which each of them contains three sub-layers. Two of them have a multi-head self-attention layer and the other one is a fully connected network (Yang et al., 2020).

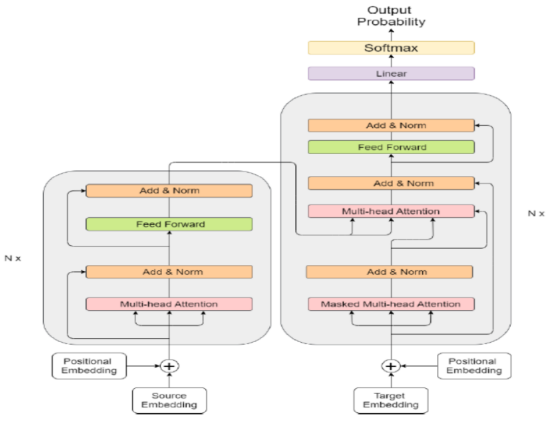


Figure 2.10: Self-attention based NMT
(Yang et al., 2020)

The researchers also believe that the NMT is the future hope for good and

quality translation. The main aim behind this approach is to build a single MT for all natural languages around the world.

2.4 English-Telugu Nerual Machine Translation systems

since last two decades a noticeable progress has been taken in the development of Machine translation. This development lead to build most advanced systems like neural machine translation systems. There are several NMT systems came into existence though all of them are not open-source. For this study, we considered only open-source MTs. Existing open source English-Telugu MTs include: Google, Bing, C-DAC, Devnagri, Yandex,, IIIT-H Mt, Lingvanex etc. A pilot study was made for the selection of the best MTs for English-Telugu. Among them, the top five MTs are considered for evaluation in this study.

2.4.1 Google Translate

Google translate[1] is an open source neural machine translation platform developed by google. It translates a language into another language. It was first launched in 2006 by offering SMT services. It updated its translation system into NMT in 2016. At present, it is providing its translation services to nearly 109 languages in the world. It is one of the most used machine translation systems with active users. As of 2016, there were 500 million users per month and it usually translates more than 100 billion words per day (blog.google, 2016). It offers its services online and offline platforms. Online services are like web applications and mobile applications with an active internet connection. And offline services are offered without internet connection. Currently , Google is using hybrid model translation architecture. The model consists of a transformer encoder and RNN decoder. (ai.google blog, 2020). It can allow translation of 10k characters at one go. It offers translation services in the following formats: Text to Text, Image to Text and speech to speech and text.

2.4.2 Bing Translate

Bing translator[2] is also known as microsoft translator. It is an open source, upto the certain translation volume, neural machine translation developed by Microsoft. It has been offering its translation services as a part of Azure Cognitive Services.

[1] https://translate.google.co.in/
[2] https://www.bing.com/translator

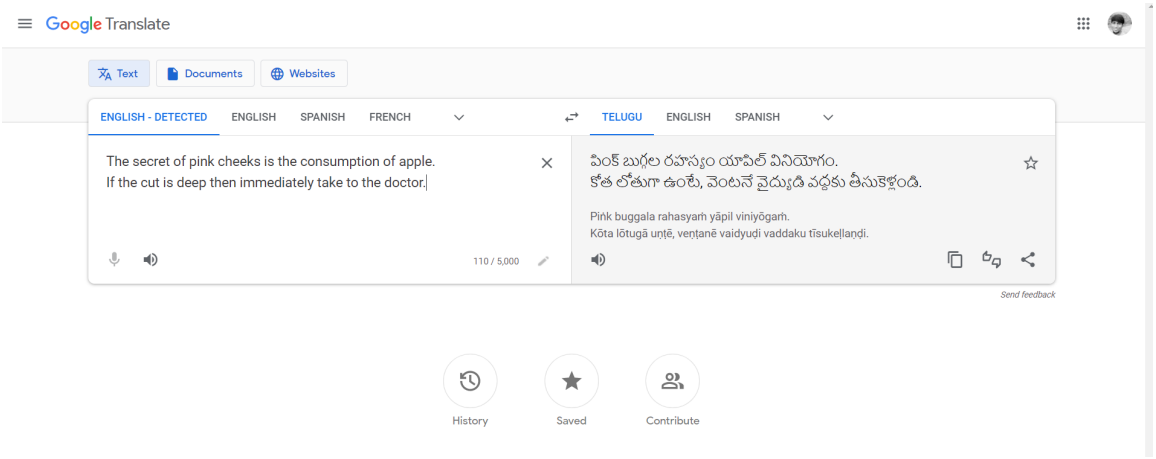


Figure 2.11: user interface of the Google MT

Microsoft launched Bing translation services in 2007(microsoft) . In 2018, Bing upgraded translator as neural machine translation. The services are available in almost 109 languages. It offers its translation services on different platforms: android, IOS and website etc. and also provides online and offline services. Translation can be done in different ways: text to text, speech to text and text to speech etc. Microsoft translator is using a deep RNN and transformer NMT model (microsoft translator, 2019). It can allow translation of 1k characteristics at one go on a website.

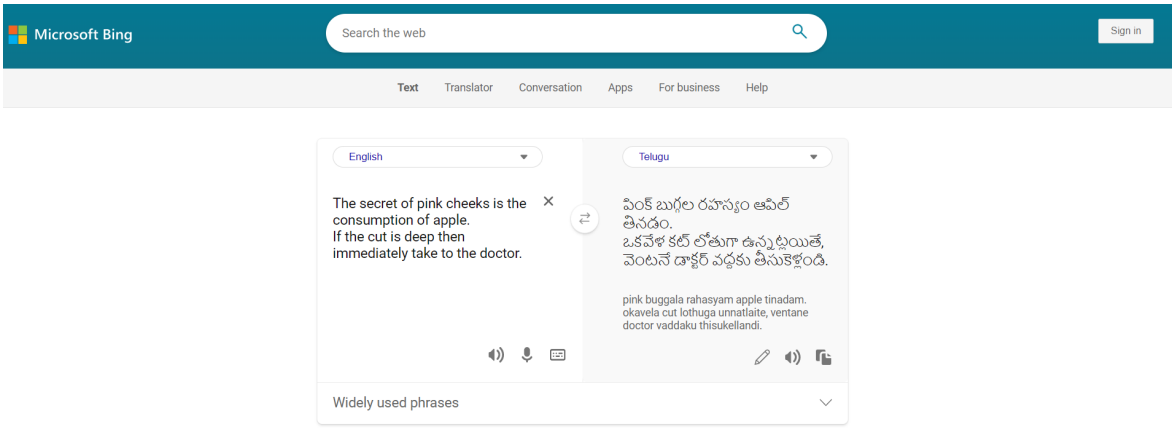


Figure 2.12: User interface of Bing MT

2.4.3 IIIT-H MT

IIIT-Hyderabad[1] is an academic research institution, India. It has developed an English to Telugu neural machine translation using bidirectional Long Short-Term Memory (LSTM) network technique which is a type of RNN sequence to sequence model. (NOTE: As there are no published research papers available, the information is acquired from personal contact.)

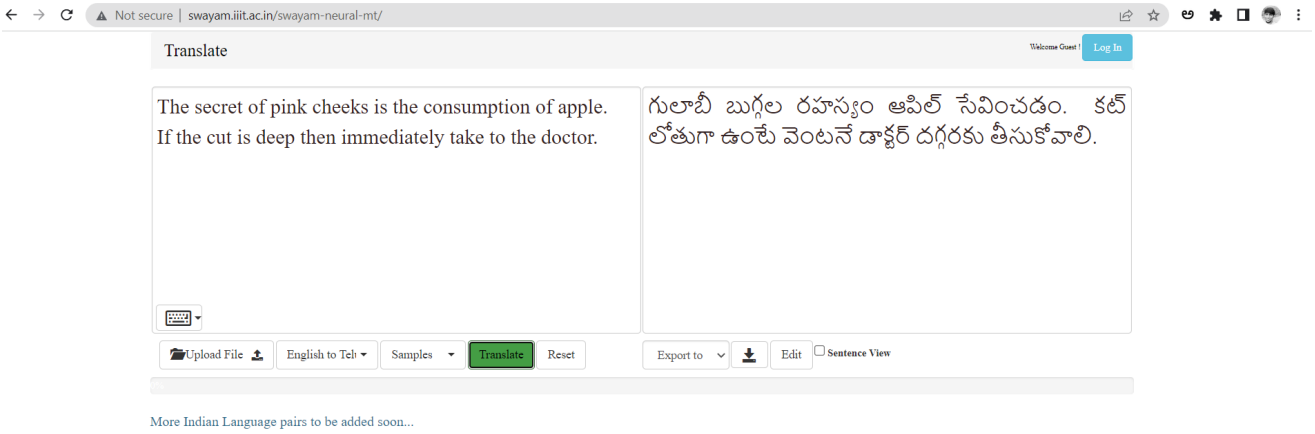


Figure 2.13: User interface of IIIT-H MT

2.4.4 LingvaNex Translate

Lingvanex[2] is a translation system developed by Nordicwise Limited Company. It is also an open resource translation system upto certain low value. It was first launched in 2012. It was updated with state-of-the-art technology neural machine translation. At present, it is offering translation services in almost 108 languages. It offers services at various platforms: mobile applications like android and IOS, web applications and messengers etc. It can translate different input formats: Text, voice, documents and websites. It is feasible to be used offline as well as online. To get access the online translator.

2.4.5 Yandex translate

Yandex[3] translation is a translation system developed by yandex, a russian based technological company. It was first launched in 2011. Initially, it came up with an SMT system. Gradually, it built a hybrid system (2017) by combining SMT

[1] http://swayam.iiit.ac.in/swayam-neural-mt/
[2] https://lingvanex.com/demo/
[3] https://translate.yandex.com/

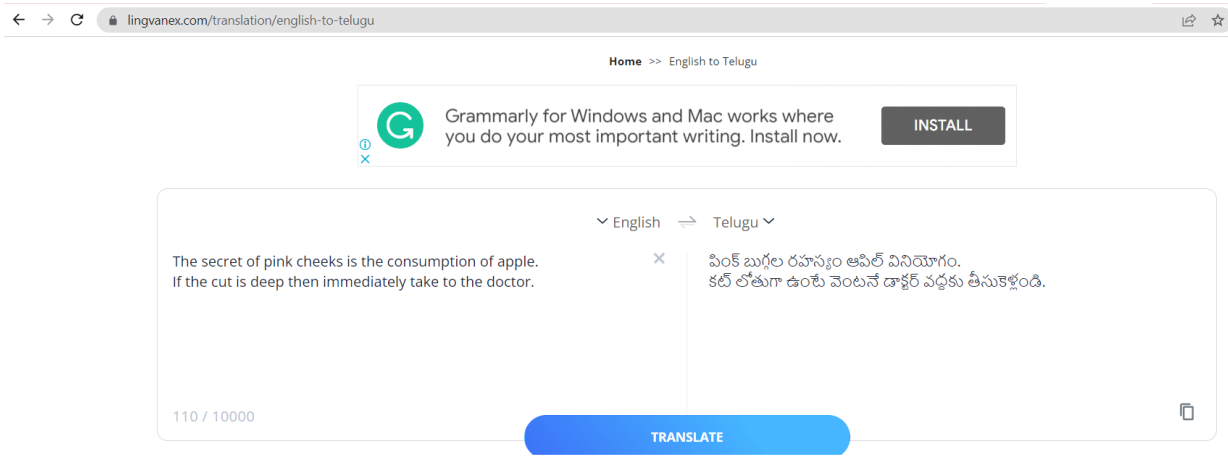


Figure 2.14: User interface of LingvaNex MT

and NMT systems to improve the quality of output in terms of fluency like human (yandex blog, 2017). It is available for translation in nearly 98 languages. It can be accessed through online website and mobile applications: android and iOS. It offers translation services in various forms: text-text, voice-input, document translation, text in image translation and translating websites. It allows users to translate 10k characters at one go.

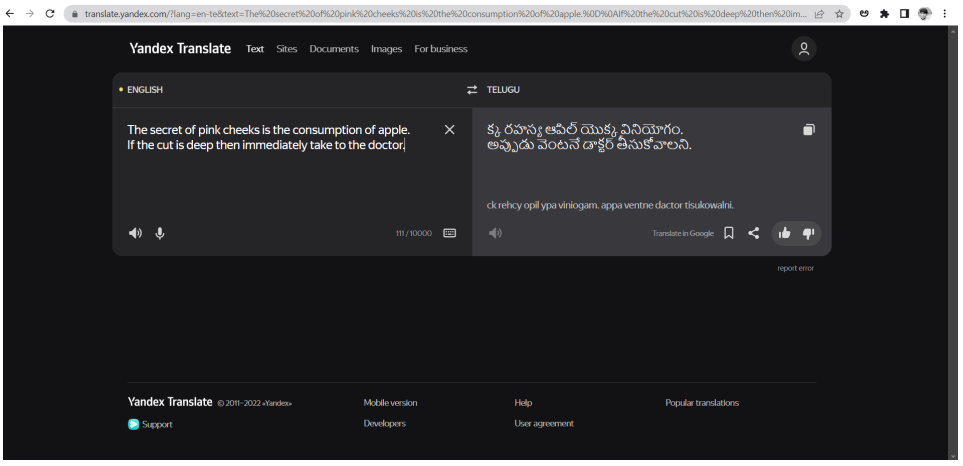


Figure 2.15: User interface of Yandex MT

2.4.6 Devnagri

Devnagri[1] is an online translation service platform. It was started in 2020 in Delhi, India. Devnagri has built an AI based neural machine translation system and also offers manual post editing services to deliver accurate output to clients. It

[1] https://devnagri.com/english-to-telugu-translation/

exclusively offers its translation services into 12 Indian languages from English as source language. The translator can be accessed through the website only as that is not available offline. It is also providing other various services: transliteration, website translation, document translation, OCR and image translation.

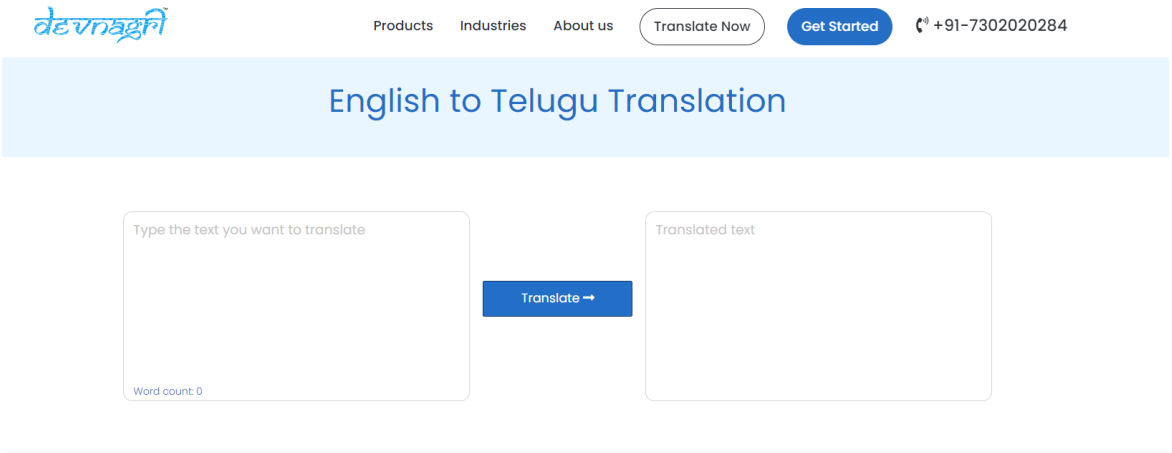


Figure 2.16: User interface of Devnagri MT

2.4.7 MoxWave Translate

MoxWave[1] is a neural machine translation system developed by process 9 technologies private limited company. It was firstly launched in 2009 in Delhi, India. It provides software solutions and translation services to clients. It offers two modes of translation services: machine translation and also manual translation. Currently, MT is available in 14 Indian languages and also manual translation is available in 22 Indian and international languages. The translation services are in different forms like text-text, speech to text, text to speech and speech translation. Any customer can access the translation system online only as it does not extend their services offline. Customers can translate 100 words at one go using MT. The following link can be redirected to the access system.

2.4.8 IBM Watson Translate

Watson[2] language translator is a machine translation system developed by IBM technologies company. It provides its translation services in different formats: speech to text, text to speech, text to text and documents translation. Further it can also translate websites using the URL. Currently, the MT is available in 58 Indian

[1] https://dts.moxwave.com/Translation
[2] https://www.ibm.com/in-en/cloud/watson-language-translator

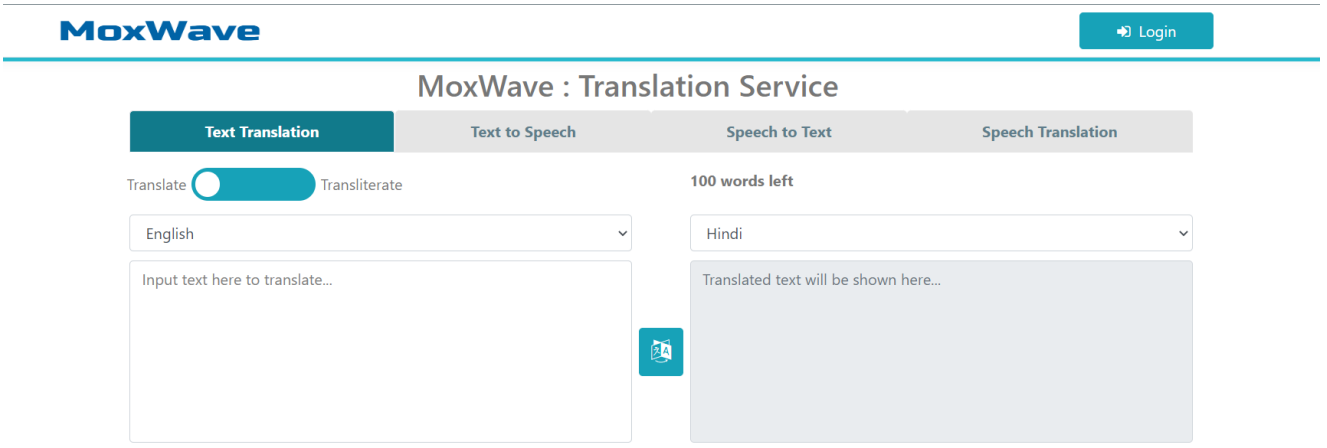
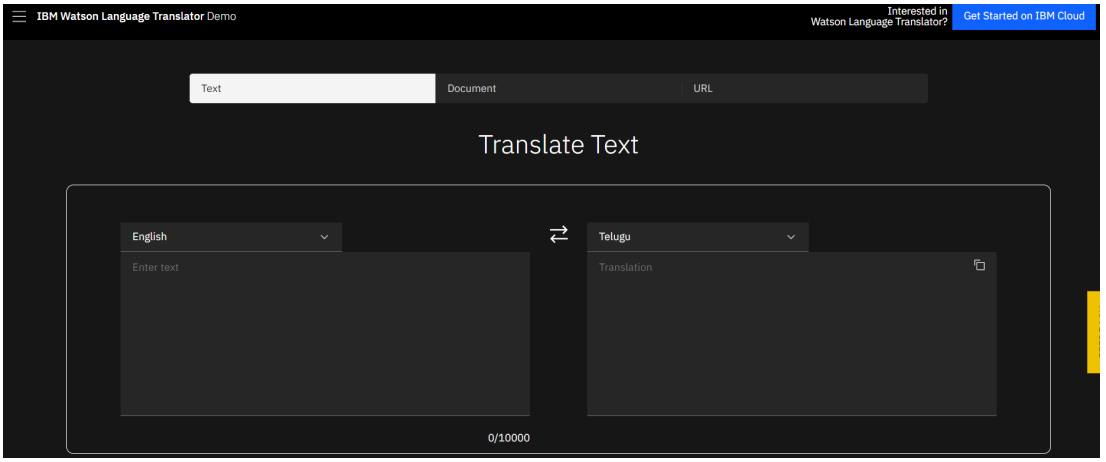


Figure 2.17: User interface of Moxwave MT

and International languages. The MT covers multiple domains related terminology. 10000 characters can be translated at one go. The following link can be used to access the MT system.



2.4.9 Amazon Translate

Amazon[1]translate is provided by amazon web service well known as AWS cloud computing powered by amazon technology company. It builds attention based neural machine translation which is considered as cutting-edge technology in the MT field. It is now available in more than 80 languages. The translation service can be integrated with various other communication based messaging applications like email and other customer based chat bots. Its translation covers multi domain documents like technical, academic etc. It also offers services related to language translation: amazon polly, amazon transcribe, amazon S3 etc.

[1]https://aws.amazon.com/translate/

2.5 Selection of NMTs for the current study

To select the English-Telugu NMts for the current study, a pilot study has been conducted in order to calculate the efficient of existing NMT systems. Among them the top five most efficient translated slystems are selected which are Google, Bing, IIIT-H, LingvaNex and Yandex. These NMT systems are evaluated in using various evaluating methods. Existing evaluation methods are discussed in Chapter 3 and the evaluation of these NMT systems is attempted Chapter 4.

Chapter 3

Evaluation : Methods and Methodology

3.1 Introduction

This chapter discusses the currently available evaluation techniques in the field of MT as evaluation techniques are considered to be one of the main tasks in the assessment of efficiency of the translation systems. Since the mid 1960's, notable efforts have been made by commercial, academic and research organizations and experts for the development of important tools and techniques for the evaluation of the MT system. Subsequently, Various types of systematic approaches are developed in measuring the MT systems. In the literature of the MT field, the evaluation process is classified into two types of paradigms: Glass box and Black box (Dorr et al., 2010). According to Dorr (2010) Glass box evaluation mainly focuses on the internal linguistic components of the MT system and also emphasizes how far a given system is handling the linguistic properties and theories practically. This is subjective in nature. Whereas the Black box evaluation emphasizes on the predetermined dataset chosen for the evaluation and comparison of the various systems' output. This tests the robustness of the MTs and handles the various types of the datasets in terms of the domain specificity, structure and style. This is objective in nature. The later one comprises two types of measuring processes: Intrinsic and Extrinsic. Intrinsic deals with the quality of the output in comparison with the reference text which is said to be of high quality. And the extrinsic measure focuses on the effectiveness of any given MT output based on the specific task. It is also referred to as task-based technique. Furthermore Intrinsic measurement is classified into two types: Human and Automatic measurement techniques. These two methods have been employed for the current study. These two methods are discussed a little more elaborately in the following sections.

3.2 Human Evaluation Methods

Human evaluation techniques use human intervention in assessing the quality of a given output by considering the characteristics like adequacy, fluency and understandability (ALPAC, 1966; Dorr et al., 2010; White, 1995). The human judges are also referred to as annotators who play a vital role in determining the quality of the output. The annotators might be bilingual or monolingual that can be decided based on a type of method chosen for the evaluation process (Chatzikoumi, 2020). Human evaluation is more subjective in nature. Human evaluation is considered as an important aspect in an evaluation process. Because, the output produced by any given MT reaches humans as they will be final consumers of the output. It is also called the gold standard method (Sanders et al., 2011). The method is again classified into two types: Directly Expressed Judgment (DEJ) and Non-Directly Expressed Judgment (NON-DEJ). DEJ contains adequacy, fluency, ranking, quality checking annotation task and direct assessment evaluation methods. Non-DEJ method contains task-based, semi-automated, error classification and analysis and post-editing. The current study also calculates comprehensibility or understandability level of the output(Chatzikoumi, 2020).

3.2.1 Directly Expressed Judgment (DEJ) Evaluation Method

Under this method, the following evaluation methods are discussed briefly.

3.2.1.1 Adequacy

This is a very popular assessment method to check the quality of a given output. The method is employed to assess to what extent the meaning of source text is conveyed into the translated text (ALPAC, 1966; Popović, 2020; White, 1995). This is also referred to as semantic adequacy. In which, bilingual human judges assess the quality in comparison with source text and translation text. To compare the both texts, the evaluators must have good proficiency in both texts in order to gain the reliability. Generally, the evaluators are given a multi-point scale for providing the score. Several multi-point scales are available for the assessment like 4, 5, 7, and 10 multiple-point scales etc (Sanders et al., 2011).

3.2.1.2 Fluency

Through this method the evaluators have been asked to evaluate the fluency of the target text in terms of their grammaticality and readability.(Popović, 2020; White, 1995). In which, the information correctness is not given priority in assessment of a given output. No reference or source text is accessible for evaluators during this task. The evaluation is carried out by employing monolinguals who have good command over the target text as being an native speaker of the language. This task is also carried out in the same manner as followed in adequacy. The score for the fluency test provided using the multi-point scales.

3.2.1.3 Accuracy

This method is also used for measuring the semantic information which is translated into a target language in comparison with a given reference text. This evaluation process is done by employing monolingual judges who are well versed in target language. In which no intervention of source text is involved in providing score to target text (Sanders et al., 2011).the noticeable difference between adequacy and accuracy is that adequacy is measured based on source text and target text for which bilingual human judges are needed but in measuring accuracy, as mentioned above, is based on comparing reference and target language both are in the same language for which monolingual judges are needed.

3.2.1.4 Comprehensibility

It is also referred to as understandability. In which, whether a translated output is able to be understood easily without much cognitive effort is measured (White and O'Connell, 1994; **?**) Which can be determined by bilingual and monolingual judges as it is an understanding of the meaning and grammaticality of a given output.

3.2.1.5 Ranking Method

In this method, evaluators are asked to provide suitable ranks for given candidate translations or system produced translations. The comparison is carried out using a relative-ranking procedure between various machine translation outputs (Bojar et al., 2016a). According to Görög (2014) the maximum number of candidate translations for the comparison should not be more than three. More than this number of translations affect judgment and lead to impaired in the reliability of the results. WMT 16 evaluation campaign set a standard comparison of five MT output at a time (Bojar et al., 2016a). Furthermore the method is classified into

two types: Quick comparison and Ranking translations. Quick compassion is performed in order to choose the best and most accurate translation by an evaluator among the maximum number of three translations. Ranking translation is carried out by ranking from 1 (Best) to 3 (Worst) among segments of the candidate translations; each and every segment in the translation are marked with above ranks (Chatzikoumi, 2020).

3.2.1.6 Direct Assessment (DA)

Direct assessment method is adopted to evaluate the adequacy and fluency of a candidate translation in isolation without in comparison with other translations (Graham et al., 2013)(and (Graham et al., 2017) DA method uses monolingual human judges for measuring adequacy and fluency which is quite different from the above adequacy method where bilinguals participate in providing score to the translations in comparison with the source text. In contrast to this, in the DA method, monolinguals provide scores to candidate translation separately. There will be no other translations available during the assessment. This method can avoid bias in assessment and also make the process simple by measuring the adequacy and fluency at one go (Bojar et al., 2016a).

3.2.2 Non-directly expressed judgment (Non-DEJ) evaluation method

In this method, the translated text is marked by human judges indirectly. Semi Automated, task-based and error classification and analysis methods are explained as follows.

3.2.2.1 Semi Automated Methods

This method is popularly known as the human-in-the-loop evaluation method. As the name suggests the evaluation method is a variation of the automatic method with the intervention of human judges (Sanders et al., 2011). There are a few important semi automated methods available, such as HTER, HBLEU and HMETEOR (Snover et al., 2006).

3.2.2.2 Task-Based Method

In the Task-based evaluation method, annotators are assigned to different types of tasks and requested to perform those particular tasks such as detecting relevant

information, asking questions and answers on the text, providing key terms in the blanks. In this way, the annotators participate indirectly in the evaluation of MT output by performing those particular tasks (Chatzikoumi, 2020).

3.2.2.3 Error Classification and Analysis

This is an important and extensively used method in the evaluation process. In which, the evaluator or researcher may classify based on the nature of the errors occured in MT output then provide analysis in detail. There different types of error typologies are available such as Multidimensional Quality Metrics (MQM) developed by QTLaunchPad and DQF (Chatzikoumi, 2020).

3.2.2.4 Post-editing

According to Lacruz et al. (2014) by using post editing a MT output is transformed into as par with human-like quality and also deliverable translation. Postediting is classified into two types: Light and Full, broadly based on the amount of edits performed by a posteditor. Light post editing is said to be a good enough translation and full post editing is said to be a human-like translation (Massardo et al., 2016) Post-editing also referred as measurement method to evaluate the quality of the output by considering the factors such as temporal and cognitive efforts (Lacruz et al., 2014).

3.3 Automatic Evaluation Methods

In Automatic evaluation Methods, no human intervention is made in the process evaluating the output. Which measures the quality of the output in comparison with the one or more than one reference translations of the source text and also without any reference translations (Han et al., 2012). The automatic methods are classified into following types: lexical similarity and linguistic similarity. Lexical similarity evaluation methods deals with edit distance, precision and recall and word order. Whereas linguistic similarity deals with linguistic aspects such as syntactic, semantic etc.(Han, 2016a). Human evaluation methods have several drawbacks in terms of time-consuming, expensive and subjectivity. To minimize or remove them, automatic evaluation methods were proposed by various researchers in the MT field. The advantages of methods are considered as cost-effective and less time-consuming compared to human methods. In which, a human evaluator will be replaced with a machine to evaluate the quality of the output based on the

automatic method that was adopted for the evaluation. The main strategy for the techniques are comparing the MT output with one or more human translation also known as reference translation so as to find out how far the MT output retained meaning.

3.3.1 Edit Distance Method

The method is employed for the transformation of MT output into the human-like translation, that is reference translation, by making minimum number edits. So, the number of edits are calculated using the following automated metrics. The edit distance task is carried out by adopting a method called Levensthtein's distance (Levenshtein et al., 1966). There are some edit operations available such as addition, substitution and elimination. Using these operations, output is transformed into a reference translation (EuroMatrix, 2007). The most popular methods are explained below.

3.3.1.1 Word Error Rate (WER)

WER is an automatic evaluation technique to measure the quality of Automated Speech Recognition (ASR) system which was introduced by Su et al. (1992) ASR performs speech to text translation. It uses the levenshtein distance algorithm to compare the similarities between output and reference string. WER is computed based on the number of deletion, insertion and substitutions made in the MT output text. The technique is adopted in plagiarism detection, DNA analysis and spell checkers. The formula is as follows:

$$\text{WER} = \frac{\text{substitution+insertion+deletion}}{\text{reference}_{\text{length}}} \tag{1}$$

S = Substitution,
I = Insertion,
D = Deletion,
N = Number of words in original text.

3.3.1.2 Translation Error Rate (TER)

Translation Error Rate (TER) had been proposed by snover et al. (2006) to compensate for the drawback of the WER metric. It is also an edit distance based automatic metric like WER. It computes its quality based on the edits performed by a translator in MT output to match the reference translation. If the number of edits are more, then the provided MT output is very distinctive from reference translation. The minimum number edits can be from 0 to infinity.

$$TER = \frac{SUB+INS+DEL+SHIFT}{\bar{N}} \tag{2}$$

S = Substitution,
I = Insertion,
D = Deletion,
S= shifts,
N = average number of words in original texts.

3.3.1.3 Metric for Evaluation of Translation with Explicit Ordering (METEOR)

METEOR is expanded as Metric for Evaluation of Translation With Explicit ORdering. It was first proposed by S. Benerjee and A. Lavie (2004). It is a precision and recall based technique. As Bleu scores inclined towards precision, It is said to be METEOR also more inclined towards recall than precision. Using this technique high correlation with human judgment were achieved. The calculation takes place based on the number of equivalences found between MT output string and the reference string. Those equivalents can be found by different matching stages: exact, stem and synonymy matching.

$$F_{mean} = \frac{P \cdot R}{\alpha P + (1-\alpha) R} \tag{3}$$

3.3.1.4 Position-Independent Word Error Rate (PER)

The metric compares the word strings produced by machine translation without considering the sequence of the reference translations. This was proposed by Christoph Tillmann (1997). To calculate the following formula can be used:

$$\text{PER} = 1 - \frac{\text{correct} - \max(0, \text{output-length} - \text{reference-length})}{\text{reference-length}} \tag{4}$$

3.3.2 Precision and Recall

Precision is seen by considering the acceptable n-grams in MT output (found in at least one reference translation) ratio with total number n-grams that are present in the same MT output. Using this ratio, the percentage of correct terms in MT output is calculated (Chatzikoumi, 2020). Precision focuses on quality of the output.

$$\text{precision} = \frac{\text{correct}}{\text{output-length}} \tag{5}$$

Ratio of the recall is seen between a MT output and Reference translation by considering the number of n-grams presented in a given MT output against the number of similar n-grams presented in a given reference translation. Recall focuses on the quantity of the output.

$$\text{recall} = \frac{\text{correct}}{\text{reference-length}} \tag{6}$$

The most popular precision and recall based method is BLEU score. Which is discussed below in detail.

3.3.3 BiLingual Evaluation Understudy (BLEU) Score Method

BLEU (Papineni et al., 2002)expanded as a BiLingual Evaluation Understudy which was first proposed in 2001 and also considered as a cost effective and fast method for any language pair (EuroMatrix, 2007). BLUE score metric is an most commonly used automatic method in assessing the performance of MT systems. The major idea behind the method is that "the closer a machine translation is to a professional human translation, the better it is." (Papineni et al., 2002). Which means that how far a machine translator can produce similar meaningful sentences in comparison with the human translation done by a professional. To calculate the familiarity of reference text with target text, n-grams are used, that is 1-5 grams, the evaluation. As 1 grams shows the least correlation, 5 grams marks the highest correlation between the text. Furthermore the method is discussed in detail in the next chapter as current study employed the method for the evaluation purpose.

- **Step 1** calculate modified n-gram precision precision based on the following formula:

$$p_n = \frac{\sum_{C \in \{\text{Candidates}\}} \sum_{\text{n-gram} \in C} \text{Count}_{\text{clip}}(\text{n-gram})}{\sum_{C' \in \{\text{Candidates}\}} \sum_{\text{n-gram}' \in C'} \text{Count}(\text{n-gram}')} \tag{7}$$

  In simple terms,

$$\textbf{Modified precision}(Pn) = \frac{\begin{array}{c}\textit{Sum of the clippe } n-\textit{gram count} \\ \textit{for all the condidate sentences in the reference text}\end{array}}{\textit{The number of candidate } n-\textit{grams}}$$

- **Step 2 :** calculating the Brevity penalty based on the following formula:

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases} \tag{8}$$

  c = words count in a reference text
  r = words count in a MT output

- **Step 3:** based on the above calculations, final BLEU score will be calculated by using the following formula:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{9}$$

BP = Brevity Penalty
N = Number of n-grams 1-gram, 2-gram, 3-gram and 4-gram
N always refers to 4
Wn = Weight for each modified precision
Pn = Modified precision

3.4 Evaluation Methodology Used for This Study

The section discusses the methodology that has been employed for the evaluation of the current study. Any study that involves the evaluation of MT output has to follow a systematic procedure. The current study also follows an evaluation procedure that includes test data collection, selection of MT systems and selection of suitable evaluation methods both: human and automatic. Each and every aspect of the aforementioned evaluation procedure is explained in detail in the following sections.

3.4.1 Test Data-set Collection

A natural language has its own set of grammatical rules which makes language a complex system. It is a very difficult task to make a machine understand in order to generate output. Any grammatical error in test data can diminish the quality of output. Therefore, good test data can minimize errors in output and would be helpful to increase the quality of output. Evaluating a machine translation on less quality test data can severely affect final evaluation results. For these reasons, the selection of test data is one of the important steps in the process of evaluation MTs. For the study, the test data size of 2000 English and Telugu parallel data

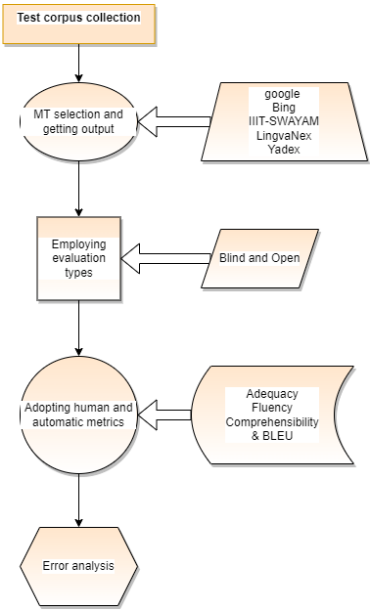


Figure 3.1: flowchart

are collected from the Indian Languages Corpora Initiative (ILCI) pertaining to the health domain. English sentences from the ILCI corpus are used as the source text for the machine translation. And manually translated sentences are used as the reference text to evaluate BLEU score. The test data consists of three types: simple, complex and compound sentences which cover various constructions like declarative, imperative, exemplary and interrogative sentences.

3.4.2 Evaluation Criteria

MT Evaluation is measuring the quality of the output by the MTs using some manual and automatic metrics. Once the data had been collected, then the English sentences were translated into Telugu as an output using the aforementioned website-based English-Telugu machine translation systems. The translated parallel text is divided into smaller sets to make the process furthermore easy and reduce the cognitive load of the evaluator. Each smaller set contains 100 sentences. Using the google spreadsheets and form builder application, the sentences were converted into Google forms to record the responses from evaluators on online mode to assign scores to.

As the study had been conducted for open and blind evaluation, two different sets of google forms were prepared. Open evaluation means, where the evaluator will be accessible to both texts, that is, source text and translated text. For the

evaluation of both open and blind, 5 Multi-point scale is proposed (table 3.1).

Quality of sentence output	**score**
Completely meaningful with perfect translation	4
Mostly meaningful with no grammatical error	3
Moderately meaningful with minor error	2
Slightly meaningful with major errors	1
Poorly meaningful and nonsense	0

Table 3.1: Multi-point scale for adequacy

Blind evaluation means, in which, an evaluator can access the only translated text and no source text is provided for the task (see table 3.2)

Quality of sentence output	score
Completely fluency	4
Mostly fluency	3
Moderately fluency	2
Slightly fluency	1
Poorly fluency	0

Table 3.2: Fluency Scale

The above scale is built based on the likert-type scale, which is a widely used rating scale model to record responses. The left column, in the table, represents the quality of sentence output with five responses. The right column, in the same table, represents the score that is assigned to that particular option.

(a) **Completely meaningful with perfect translation and completely flunecy:** In which, the output will be a native like fluent, completely meaningful and that can be easily understandable and comprehensible. It is assigned with a score of 4. 100 percent meaning is retained.

(b) **Mostly meaningful with no grammatical error and mostly fluency:** it represents the grammatically correct structure with less adequate and less comprehensible. It is awarded with a score of 3. 75 percent is retained.

(c) **Moderately meaningful with minor error and Moderately fluency:** in which, the output contains grammatical errors with less fluency. It can be marked with a score of 2. 50 percent meaning is retained.

(d) **Slightly meaningful with major errors and Slightly fluency:** it represents the output with major grammatical errors which can be low adequacy and low comprehensible. It is awarded with a score of 1. 25 percent meaning is retained.

(e) **Poorly meaningful and nonsense and poorly fluency:** senseless or absurd translations and null translations can be marked with this option. It is assigned with a score of 0. 0 percent is retained.

For the study, two types of evaluation methods have been adopted which are human and automatic evaluation metric.

3.4.3 Human Evaluation Methods

Human evaluation methods also can be referred to as manual evaluation methods. It can be performed by human evaluators who are end users of a machine translator output. As every end user has a different unique pattern in dealing with language. That is the reason, it is considered subjective nature. Hence, it is important to employ human evaluation methods for the evaluation. Current study undertakes three more general and popular human evaluation methods: Adequacy, Fluency and Comprehensibility with newly proposed formulas for the evolution.

(a) **Adequacy**: the method is employed to assess to what extent the meaning of source text is conveyed into the translated text (ALPAC, 1966; Popović, 2020; White, 1995). This evaluation is calculated when the open (both source and target texts are accessible to the evaluators) evaluation is attempted. Adequecy is calculated by following principle.

$$\text{Adequacy} = \sum_{i=3}^{4} \frac{Si}{N}$$

(b) **Fluency**: Through this method the evaluators have been asked to evaluate the fluency of the target text in terms of their grammaticality and readability. This evaluation is calculated when the blind (only the target text is accessible to the evaluators) evaluation is attempted.(Popović, 2020; White, 1995). This can be calculated based on the following principle.

$$\text{Fluency} = \sum_{i=3}^{4} \frac{Si}{N}$$

principle:

(c) **Comprehensibility** : it is also considered as understandability of the meaning of the output. It is calculated on both open and blind evaluation. (White and O'Connell, 1994) and calculated on the below principle.

$$\text{Comprehensibility} = \sum_{i=2}^{4} \frac{Si}{N}$$

3.4.4 Human Evaluators

In the process of evaluation of MTs, evaluators play a key role in deciding the quality of the output. Evaluators, who have a knowledge on linguistics aspects of the both languages: source text and target text, can help to generate most accurate results compared to an evaluator with no linguistics and unknown to that particular language. If the study follows an open evaluation approach, an evaluator, with good linguistic knowledge of both languages that is source and target text, is required. The open evaluation is conducted to calculate the adequacy of the MT outputs. If the study employs a blind evaluation approach then an evaluator, having good linguistic knowledge of target language or target text, is required. The blind evaluation is performed to calculate the fluency of the MT outputs. The current study undertakes three bilinguals, who have a good linguistic knowledge of the languages: English and Telugu, for the open evaluation, and three monolinguals, who have possessed good linguistics knowledge of Telugu, for the blind evaluation. The prepared google forms documents have been distributed through Email among them to record their responses. The google form contains broadly three sections: in the first section, a brief note about the research; in the second section, personal details of the evaluators; in the third section, output data to be scored by the evaluators. It appears as below Each google form, for the open evaluation can be appeared as below:

and for the blind evaluation it is appeared as following

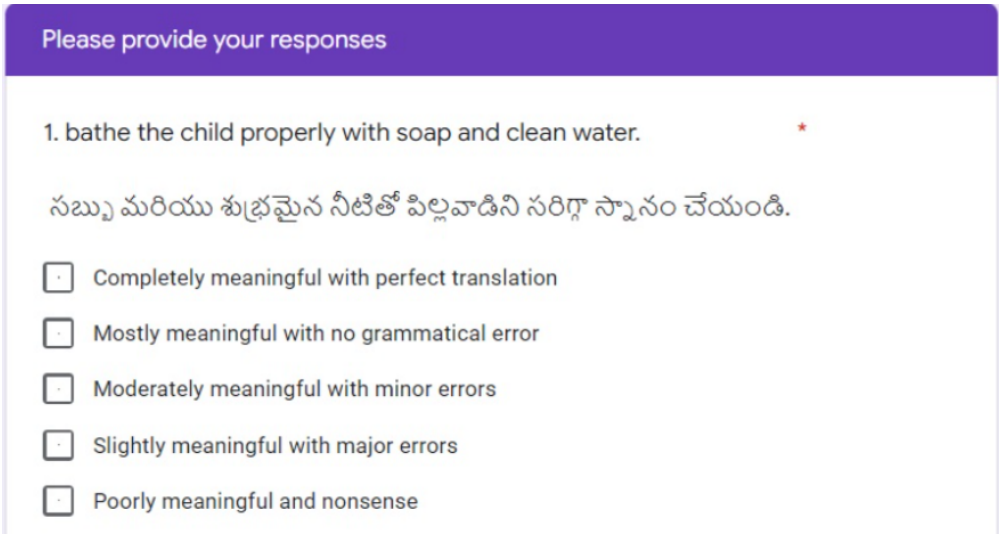


Figure 3.2: Open evaluation sample

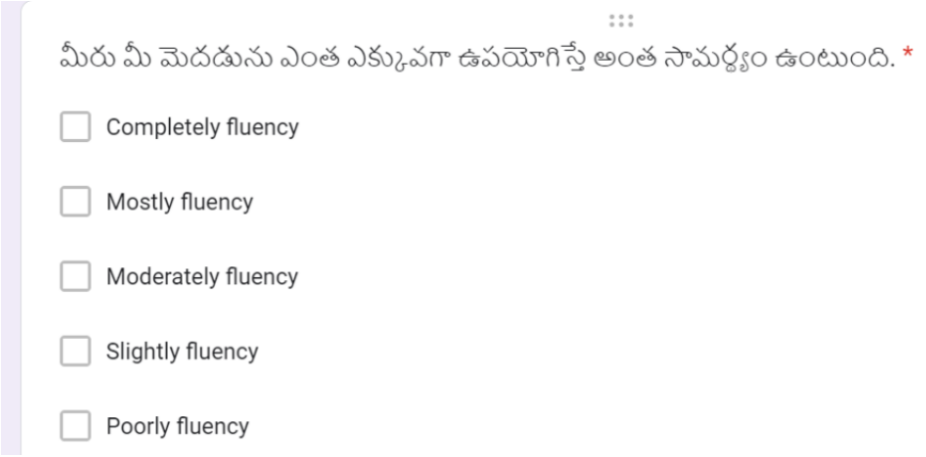


Figure 3.3: Blind Evaluation sample

The evaluators should record their responses based on a 5 multi-point scale (figure.1) in both the tasks. The advantage of using the proposed scale is that, which incorporates the above three human metrics, that is, one response of the evaluator can be used to evaluate for the above three human evaluation methods. There is no need of using separate scales where evaluators should be asked to give multiple responses for each and every method. Since human evaluation methods can be treated as time consuming and high-cost task, the proposed scale can, to some extent, be cost efficient and time efficient.

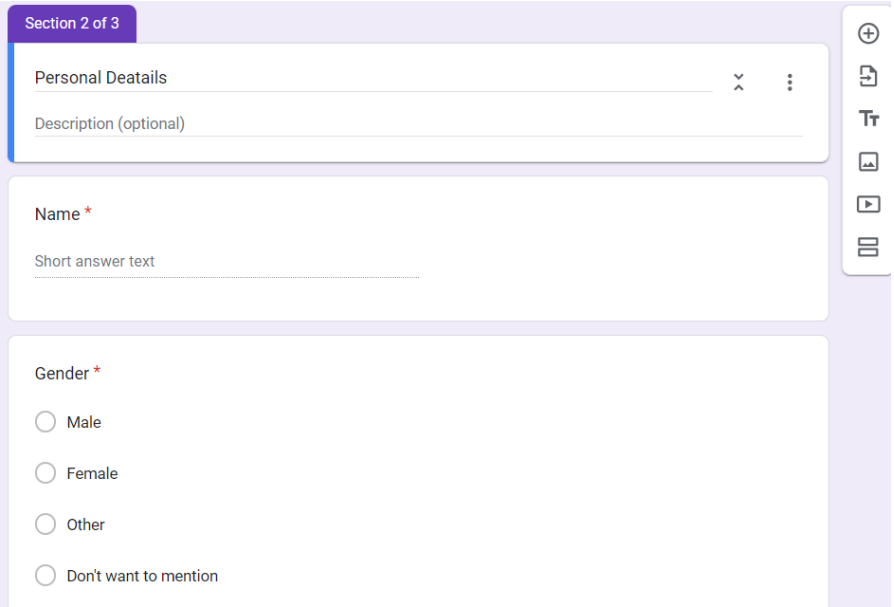


Figure 3.4: Evaluator's personal information sample

3.4.5 Automatic Evaluation Techniques

Human evaluation methods have several drawbacks in terms of time-consuming, expensive and subjectivity. To minimize or remove them, automatic evaluation methods were proposed by various researchers in the MT field. These methods are considered as cost-effective and less time-consuming compared to human methods. In which, a human evaluator will be replaced with a machine to evaluate the quality of the output based on the automatic method that was adopted for the evaluation. For the current study, the BLEU (papineni et al., 2002) automatic metric has been employed to evaluate the MTs.

3.4.5.1 BLEU Method

BLEU (Papineni et al., 2002). expanded as a Bilingual Evaluation Understudy. BLUE score metric is an most commonly used automatic method in assessing the performance of MT systems. The major idea behind the method is that "the closer a machine translation is to a professional human translation, the better it is."(Papineni et al., 2002). This means that how far a machine translator can produce similar meaningful sentences in comparison with the human translation done by a professional. To conduct the BLUE score evaluation, MT output, also referred to as candidate translation, and human translation also referred to as reference translation, are needed. The evaluation of MT output can be carried out with one or more than one reference translation. Basically, the process is

performed by considering the overlapping of n-grams, that is, the sequence of words that are present in both texts in common: MT output and reference translation. Evaluation can be calculated based on uni-gram, bi-gram tri-gram and four-gram. It is calculated without taking into consideration word order, that is, position-independent. The maximum and minimum evaluation score for BLEU is 0 and 1. If an MT output is perfectly matched with the reference text then a score of 1.0 is assigned and considered as a good translation. If the MT output is completely mismatched then a score of 0.0 is assigned.

Step 1 calculate modified n-gram precision based on the following formula:

$$p_n = \frac{\sum_{C \in \{\text{Candidates}\}} \sum_{\text{n-gram} \in C} \text{Count}_{\text{clip}}(\text{n-gram})}{\sum_{C' \in \{\text{Candidates}\}} \sum_{\text{n-gram}' \in C'} \text{Count}(\text{n-gram}')} \tag{7}$$

In simple terms,

$$\textbf{Modified precision}(Pn) = \frac{\textit{Sum of the clippe n-gram count for all the condidate sentences in the reference text}}{\textit{The number of candidate n-grams}}$$

Step 2 : calculating the Brevity penalty based on the following formula:

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases} \tag{8}$$

c = words count in a reference text
r = words count in a MT output

Step 3: based on the above calculations, final BLEU score will be calculated by using the following formula:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{9}$$

BP = Brevity Penalty
N = Number of n-grams 1-gram, 2-gram, 3-gram and 4-gram
N always refers to 4
Wn = Weight for each modified precision
Pn = Modified precision

3.4.6 Inter-Rater Agreement

: Inter-Rater Agreement is a research evaluation instrument. Which measures how far raters agree mutually by assigning the same value using the same scale for each item that is provided for the study. For the current study, Fleiss kappa (Cohen, 1960). method is used to calculate the Inter-rater agreement. The method is feasible and employed for a study when there are more than two raters and the available assigning values are in the nominal. As this study involves more than two and having the nominal values it is considered as the most suitable method for calculating the Inter-rater agreement. Once the calculation is carried then the results are interpreted based on the level of agreement that study has achieved. kappa formula and The level of agreement can be seen as following.

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

K= kappa
Po= observed agreement
Pe= expected agreement if random judgment
results can be interpreted based on the level of agreement that appears in the following table (Landis and Koch, 1977):

By employing all these tools, the selected English-Telugu NMT systems are evaluated and results are discussed in the chapter 4.

Kappa	Level of Agreement
> 0,8	Almost perfect
> 0,6	Substantial
> 0,4	Moderate
> 0,2	Fair
> 0	Slight
< 0	No agreement

Figure 3.5: Kappa level of agreement table
(Landis and Koch, 1977)

Chapter 4

Evaluation of English-Telugu Neural Machine Translation Systems

4.1 Introduction

This chapter provides the results of the MT systems evaluation. The evaluation results of both processes: human and automatic, are elaborately discussed in the current chapter. As part of the human evaluation, the results of the Adequacy, Fluency and comprehensibility are discussed in detail. Adequacy results are presented in terms of average adequacy and aggregate adequacy. Whereas, the fluency results are also presented in terms of average fluency and aggregate fluency. The results of comprehensibility are calculated in monolingual and bilingual perspectives and provided in the following section. As a part of the automatic evaluation, the results of the BLEU score are provided. Furthermore, the inter-rater agreement evaluation results of each MT are presented in detail. As we discussed in the previous chapter, we have employed five-multi point scales for the evaluation of MTs.

4.2 Evaluation Process

Evalaution process also can be referred as methodology for the conducting study. Any methodlogy, in reserach study must follow a systemic procedure. corpus for the study, 2000 English sentences, has been collected form Indian languages corpora Initiative (ILCI). The corpus belong to the health domain. it consist of simple, complex, compound sentences. criteria for the evaluation of the output, 5 multipoint scales are used for assessing the quality of the output. three bilinguals, who have good knowledge about English and Telugu, and three mono-linlguals , who are well versed in their native language Telugu are participated to mark the their responses. Blue score is calculated on the output. futhure more inter rater

agreement also calculated separately for bilingual and monolingual evaluators. The results are enumerated as the following:

4.3 Inter-Rater Agreement

An inter-rater reliability analysis was performed between the dependent samples of Rater 1, Rater 2 and Rater 3. For this purpose, the Fleiss Kappa was calculated, a measure of the agreement between more than two dependent categorical samples. The Fleiss kappa is used to calculate the inter-rater reliability where there are more than 2 data collectors involved in the research study. The following table represents the MT system in the left column and the Fleiss kappa score in the right column.

MT systems	**Kappa score**	**Level of agreement**
Google	0.23	Fair agreement
Bing	0.21	Fair agreement
IIIT-H (EN-TL)	0.11	slight agreement
LingvNex	0.5	slight agreement
Yandex	0.1	slight agreement

Table 4.1: Inter-Rater Agreement results

The table shows the agreement between the Rater 1, Rater 2 and Rater 3 of five MT systems. Google MT has recorded a high Fleiss kappa score of 0.23 which is considered a fair level of agreement according to the interpretation of the Fleiss kappa level of agreements. LingvNex has obtained a low Fleiss Kappa score of 0.05 which is considered to be a slight level of agreement.

4.4 Adequacy

The detail explanation of the adequacy is provided in the third and fourth chapters. The following scale is proposed and used for marking of output quality by bilingual evaluators.

The Figure - 4.2 depicts individual adequacy scores provided by the evaluators for the five MT systems. The information in the table is represented as the average percentage of each MT. From the second column of the table, It is observed that google translate gets the highest 4 scores with an average percentage of 50.68 percentage compared with the other MTs systems. In contrast to this, in the last column, Yandex gets 4 scores with an average percentage of 8.55, which is six

Quality of sentence output	score
Completely meaningful with perfect translation	4
Mostly meaningful with no grammatical error	3
Moderately meaningful with minor error	2
Slightly meaningful with major errors	1
Poorly meaningful and nonsense	0

Table 4.2: 5 Multi-point scale for adequacy

5-point scale	Google	Bing	IIIT-H MT	LingvNex	Yandex
4 score	50.68	40.98	26.11	24.5	8.55
3 score	28.58	36.81	34.53	30.15	5.06
2 score	15.76	18.5	30.63	33.4	31.71
1 score	4.11	3.1	7.36	10.03	19.65
0 score	0.6	0.85	1.35	1.91	35.01

Table 4.3: Average adequacy scores of English-Telugu MTs output.

times lower in comparison with google MT. In continuation, it is also observed that google is marked with the lowest number of 0 scores than other MTs, with an average percentage of 0.6. Whereas, Yandex translate received the highest number of 0 scores with an average percentage of 35.01. The most probable reason for obtaining the high results of google translate would be the vast number of users. And the availability of abundant parallel databases. Furthermore, employing hybrid model architecture known as a transformer-based NMT system which is the combination of transformer encoder and RNN decoder translation model considered as the state of the art technology in the field of machine translation. And also, such failure for Yandex might be the less number of users and lock of databases. It also uses the MT model, combining both SMT and NMT approaches, which is a less advanced translation model than the state-of-the-art translation models.

Also, one can find the numerical representation of the responses of each score in Figure - 4.3. To the average adequacy percentage (see Figure - 4.2) and aggregate adequacy (see Figure - 4.4).

Overall, It is observed that from Figure 4.4 google translate has performed well and stood at the top in the aggregate adequacy rate with the 79.26 percentage.

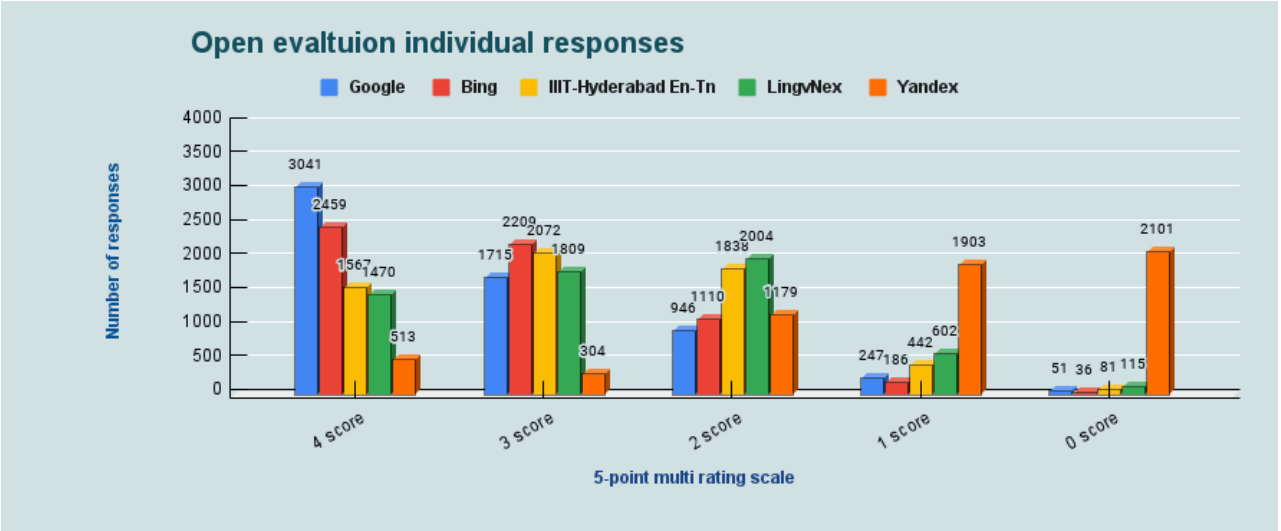


Figure 4.1: Open Evaluation Individual Responses

MT systems	Adequacy(%)
Google	79.26
Bing	77.8
IIIT-H	60.65
LingvNex	54.65
Yandex	13.61

Table 4.4: Overall Adequacy(%)

Whereas, Yandex's translation performance is six times far lower than Google translate with an aggregate Adequacy percentage of 13.61. To calculate the aggregate adequacy, 4 and 3 scores have been considered as most of the information is retained in the MT output from the source text. Suppose we see the efficiency of the systems in terms of aggregate adequacy in comparison with one another, in the first case. In that case, Google and Bing have only a two per cent aggregate adequacy difference between them which is very less than any of the other two MTs. In the second case, IIIT-Hyderabad EN-TL and LingvNex have a 6 per cent difference.

4.5 Fluency

The detail explanation of the fluency is provided in the third and fourth chapters. The following scale is used for marking of output quality by monolingual evaluators.

Quality of sentence output	score
Completely fluency	4
Mostly fluency	3
Moderately fluency	2
Slightly fluency	1
Poorly fluency	0

Table 4.5: Fluency Scale

5-point scale	**Google**	**Bing**	**IIIT-H MT**	**LingvNex**	**Yandex**
4 score	40.58	35.45	33.9	25.23	8.46
3 score	22.51	24.56	23.75	24.63	15.1
2 score	19.63	20.43	21.46	19.6	17.63
1 score	11.33	12.26	12.46	13.5	21.36
0 score	5.93	7.31	8.35	16.53	37.43

Table 4.6: Average fluency scores (%) of English-Telugu MTs

The Figure - 4.6 depicts individual fluency scores provided by the evaluators for the five MT systems. The information in the table is represented as the average percentage of each MT. From the second column of the table, It is observed that google translate gets the highest 4 scores with an average percentage of 40.58 in comparison with the other MTs systems. In contrast to this, in the last column, Yandex gets 4 scores with an average percentage of 8.46 which is five times lower in comparison with google MT. In continuation to this, it is also observed that google is marked with the lowest number of 0 scores than other MTs with an average percentage of 5.93. Whereas, Yandex translate received the highest number of 0 scores with an average percentage of 37.43 which is six times higer than the google MT. The most probable reasons for obtaining high adequacy rate of google translate would be the vast number of users. And the availability of abundant parallel databases. Furthermore, employing hybrid model architecture known as transformer based NMT system which is the combination of transformer encoder and RNN decoder translation model considered as the state of the art technology in the field of machine translation. And also, such failure for the Yandex might be the less number of users and lock of databases. It also uses the MT model having the combination of both SMT and NMT approaches, which is a less advanced translation model than the state of the art translation models.

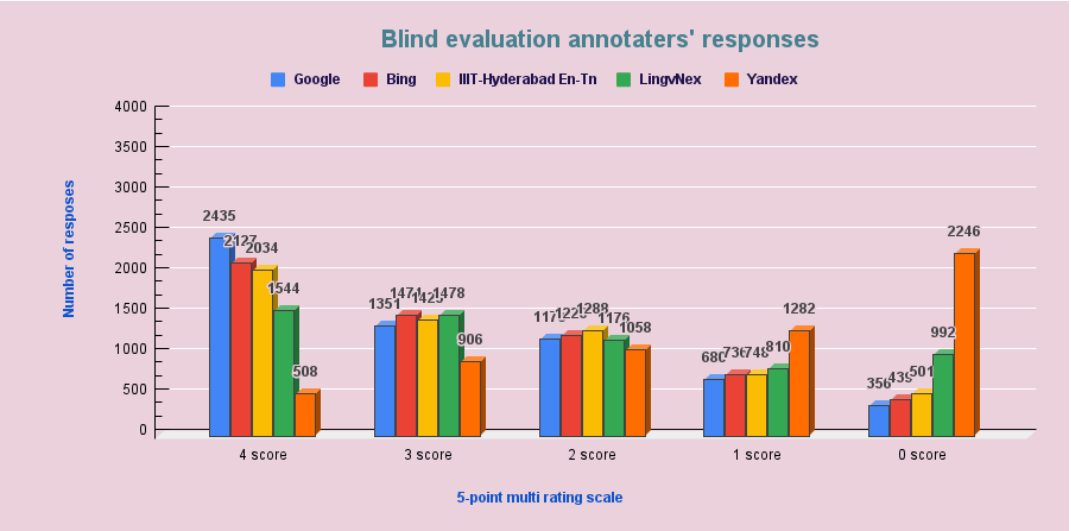


Figure 4.2: Blind Evaluation Individual Responses

MT systems	Fluency (%)
Google	63.1
Bing	60.01
IIIT-H (EN-TL)	57.65
LingvNex	50.36
Yandex	23.56

Table 4.7: Over all Fluency %

Overall, It is observed that from the Figure 4.7 google translate has performed well and stood at the top in the aggregate adequacy rate with the percentage of 63.1. Whereas, Yandex translation performance is alomst three times lower than Google translate with the aggregate fluency percentage of 23.56. To calculate the aggregate adequacy, 4 and 3 scores have been considered for the calculation. If we see the efficiency of the systems in terms of aggregate fluency in comparison with one another, in the first case, Google and Bing have only three percentge of aggregate fluency difference between them which is very less difference between any of the other two MTs. In the second case, IIIT-Hyderabad EN-TL and LingvNex have a 6 percent difference between them.

4.6 Comprehensibility

The detail explanation of the comprehensibility is provided in the third and fourth chapters. The Figure - 4.8 depicts bilinguals' and monolinguals' comprehensibility

MT systems	Bilinguals' comprehensibility	Monolinguals' comprehensibility
Google	96.03	82.73
Bing	95.3	80.45
IIIT-H MT	91.28	79.11
LingvNex	88.05	69.46
Yandex	45.33	41.2

Table 4.8: Average comprehensibility of each MT

of each system provided by them. The information in the table is represented in percentages. Bilinguals' comprehensibility is calculated based on the scores of adequacy provided by bilingual evaluators, and monolingual comprehensibility is calculated based on the score of fluency as the monolingual evaluators respond. To calculate comprehensibility, 4, 3 and 2 level scores have been considered. Google secured the top place in terms of both results: bilingual comprehensibility with a percentage of 96.03 and monolingual comprehensibility with a percentage of 82.73. Yandex occupied at last in the figure-4.8 with a percentage of 45.33 in bilinguals comprehensibility and 41.2 in monolinguals' comprehensibility.

4.7 BiLingual Evaluation Understudy (BLEU)

The detail explanation of the Bleu score method is provided in the third and fourth chapters. For the calculation of output, a interactive Blue score tool [1] devloped by Tilde technologies located in Latvia in Europe.

MT systems	Bleu score results (%)
Google	12.29
Bing	9.03
IIIT-H MT	7.37
LingvNex	6.85
Yandex	3.93

Table 4.9: Bleu score results

The table shows the results of the BLEU score of each individual system. It is observed from the above results that google MT has secured first place in the list with a score of 12.29. Whereas Yandex has been in last place on the list with a

[1] https://www.letsmt.eu/Bleu.aspx

score of 3.93. It is very well known that there can be as many human or reference translations available. Hence, having no particular and standard human translation might be the likely reason for the low scores.
For Instance:

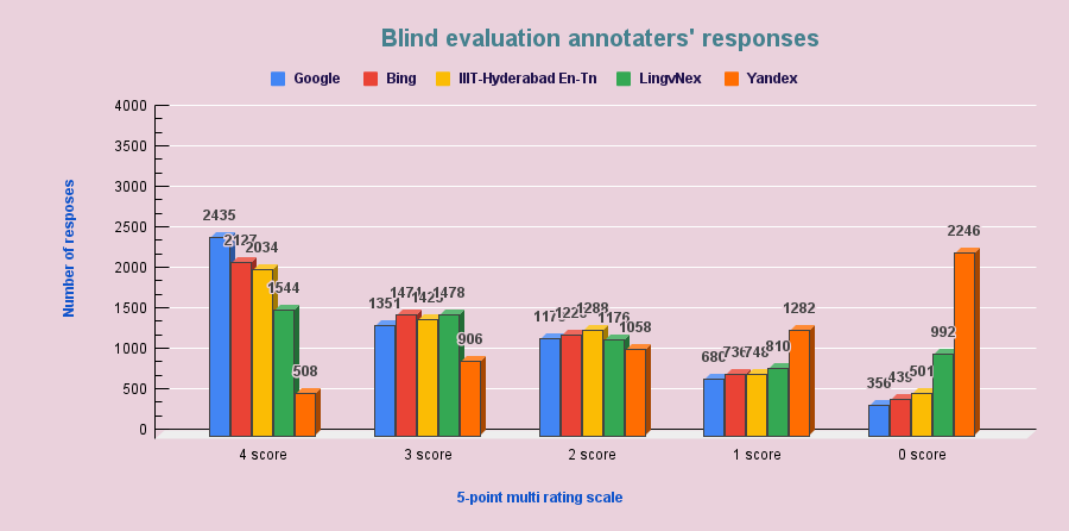


Figure 4.3: Examples for BLUE score Calculation

It is observed from the example that human and machine translations give the same sense that the source text is giving. But in the above example, using Telugu synonyms *lābhaM* and *mēlu* of the English equivalent "beneficial" is the reason for such a low BLEU score.
as the current chapter discusses the results of evaluation. The following chapter provides the error classification of the Neural Machine Translation output.

Chapter 5

Error Analysis of Machine Translation Outputs

5.1 Introduction

In this chapter, machine translation output from various MTs are analyzed in terms of linguistic and non-linguistic appropriateness. Each MT's output is analyzed for linguistic correctness and classified based on the errors. It is important to attempt a analysis of evaluation of MT's output in-order to decide the efficiency of the system as well as to decide further post-editing procedure. It is a known fact that any MT output needs human interference to make the output effective. Here, we broadly classify errors into linguistic and extra-linguistic errors which are further divided into various types. Firstly, a linguistic divergence of English and Telugu is outlined for a better understanding of linguistic processes.

5.2 Linguistic Divergence between English and Telugu

Telugu is a South-Central Dravidian language operated in the states of Telangana and Andhra Pradesh (Krishnamurti, 2003). English belongs to the Indo-Eurpoean language family (Quirk, 2010). Some key linguistic features of Telugu and English are listed below (Krishnamurti and Gwynn, 1985a):

- Telugu is the head-final language branching left with Subject-Object-Verb (SOV) word order. Whereas English is the head-initial language with right-branching and Subject-Verb-Object (SVO) type with strict word order.
- Telugu is an agglutinative language exhibiting inflectional morphology. English is an analytical language whereby inflection is exhibited through words.
- Morphologically, English and Telugu differ in terms of marking case, number, gender, person and agreement. Telugu marks morphological information in

the form of morphemes, especially suffixes whereas English marks them using separate words and a specific position in a sentence. Gender, Number, Person (GNP) information is marked on the verb as agreement in Telugu, whereas in English, only third-person singular is marked in present tense, For ex. *He comes*

- Telugu consists of post-positions whereas English has prepositions.
- Telugu is a pro-drop language that allows subject-less sentences. Pro-drop languages allow pro-drop to an extent that the φ - features (gender, number, person, etc) are reflected on the verb for the local recovery of the dropped arguments (Biberauer, 2008, p.331). However, in English, subject is mandatory and a sentence cannot drop the subject.
- In addition to this, Telugu allows verb-less sentences in normal and adjacent predicate construction which is not the case in English.
- In Telugu, quirky subjects i.e non-nominative subjects constructions are a common phenomenon whereas its not the case with English.
- Causative constructions in Telugu are found as non-periphrastic structures with the suffix [-*iMcu*] added to the verb stem whereas the causative construction is realised as a periphrastic, proto-typically with the use of the verb.
- Coordination in Telugu is realized in two ways. One is using the conjunction *mariyu* and the other is through vowel lengthening of the conjoints.
- English and Telugu also differs in formation of complement clauses, relative clauses and participial constructions. Relative clauses in English often occur with
- English is productive in infinitive constructions whereas infinitive constructions is almost absent in Telugu(except for certain constructions).

5.3 Classification of Machine Translation Output Errors

Classification of MT errors became increasingly important over a decade to evaluate strengths and weakness of the MTs. This enables MT to work on the

errors and identify mechanisms to rectify them in future. Error classification can be done manually or automatically using a machine. As part of this study, manual classification of errors is attempted.

A review on error identification in MT output is discussed below:

(Popovic, 2011) introduced a toolkit named 'Hjerson' [1] that classifies errors of an MT automatically. It is an open-source toolkit programmed in python. The tool requires a reference translation and the MT output of the same corpus. Hjerson classifies errors into inflectional errors, syntactic or re-ordering errors, missing or extra words and lexical errors. The analysis provided by the toolkit is noted to be highly co-related with that of human evaluation. However, it is not always possible to provide the reference translation for every input, which can be a challenge for such automatic toolkits.

(Font-Llitjós and Carbonell, 2004) initiated an MT project called 'AVENUE'[2] at Carneige Millon Universty to develop MT systems for low online resourceful languages. In which English-Spanish transfer-based MT system's output was evaluated by employing bilingual speakers to extract linguistic informativeness and improve the system in terms of accuracy as part of post-editing. In this, the researchers had proposed a hierarchical error classification. The classification has classified errors as follows: wrong word order, wrong word form, wrong sense of the word, wrong agreement, incorrect word and no translation. Using this classification scale, evaluators marked the error types 70% correctly.

(Daems et al., 2017) stresses the impact of classification of MT errors on several post editing tasks. It is considered important to distinguish between the final output of MT and the MT that is fit for post-editing. Error classification is essential to decide if the output of MT is suitable for post-editing or its better to translate from scratch. (Daems et al., 2017) identifies the post-editing effort indicators and how the different error types influence the post-editing efforts.

In this section, we attempted a human classification of errors based on human evaluation which can be automatized for future purposes. Errors are classified broadly into four types: as seen in figure-5.1 the first three are linguistic and the fourth one is non-linguistic errors: (1) Morphological, (2) Syntactic, (3) Semantic and (4) Miscellaneous errors. Further, finer classification of these 4 types is attempted based on the errors encountered in the MT output of the selected MTs..

[1] http://www.dfki.de/~mapo02/hjerson/.
[2] https://www.cs.cmu.edu/~avenue/

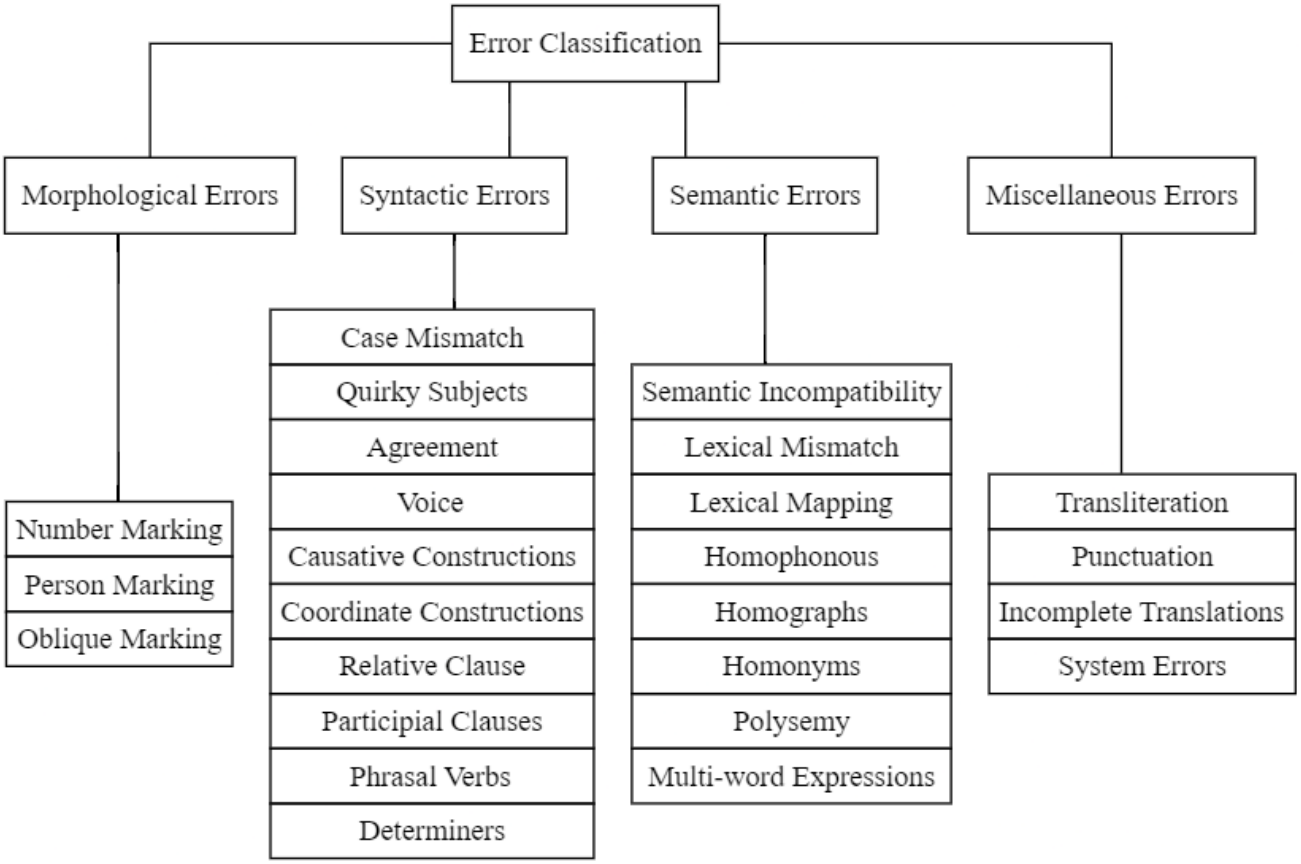


Figure 5.1: Classification of Error Types

5.3.1 Morphological Errors

Morphological errors include errors pertaining to gender, number person, case on nouns and Tense Aspect Mood(TAM) on verbs and other lexical and functional categories including adjectives, adverbs, post-positions etc. Any errors regarding the lexical categories like differences in morphology of Telugu are discussed as part of this section. Every MT system might not encounter all the classified errors, however, all such systems' output that results in erroneous output are discussed here.

5.3.1.1 Number Marking

In nouns, number is expressed in two ways viz., in singular and plural. The default plural marker is *-lu*.

In certain MT systems, it is observed that plural nouns were translated as singular nouns especially when a number modifier precedes the noun. Consider the translations of Bing MT-(5.1).

(5.1) 'Make the child drink this water two spoonfuls twice daily.' **[Eng.]**
pillavāḍu ī nīṭini rōjuki reMḍusārlu reMḍu ceMca **[Tel.]**
child this water-ACC daily-DAT twice two spoon
tāg-aḍāniki cēyaMḍi
drink-for make **[Bing]**

In ex-(5.1), the phrase 'two spoonfuls' is translated as a *reMḍu ceMca* 'two spoon' where as the appropriate translation is *reMḍu ceMcālu*, leading to erroneous output.

(5.2) 'Some injections are also important for children' **[Eng.]**
pillala-ku konni sūdi kūḍā avasaraM **[Tel.]**
children-DAT *some injection* also important **[LingvaNex]**

Similarly, in the example-(5.2), 'some injections' is translated as *konni sūdi* that literally translates to 'some injection' wherein the noun 'injection' is in a singular noun form.
It is observed that these issues are common among the other chosen MT systems.

5.3.1.2 Person Marking

Person information refers to participants/personal pronouns in the text or discourse. These pronouns occur in three persons: first, second, and third. Telugu is varied in its use of personal pronouns and has various information encoded on the pronouns like the honorificty, inclusive vs exclusive, deixis etc.
The first person plural pronoun has both inclusive and exclusive information coded on the personal pronoun i.e 'memu' 'exclude the listener' and 'manam' includes the speaker and listener'. The fact that English does not mark any such information on personal pronouns lead to the difficulty in translating such information into the target language, Telugu. Consider some examples wrongly translated for person:

(5.3) 'We withdraw hand when all of a sudden a pin pierces finger' **[Eng.]**
akasmātu-gā pinnu vēli-ki guccukunn-appuḍu mēmu **[Tel.]**
sudden-ADV pin finger-DAT pierce-then *we-EXC*
cēti-ni upasaMhariMcukuMt-ā-M
hand-ACC withdraw-HAB-1.PL. **[IIIT-H]**

In ex-(5.3), the first person plural exclusive pronoun *mēmu* 'we' is used instead of the inclusive alternative *manamu* 'we-inclusive'. The given source language text is a generic statement for which the inclusive pronoun is ideally used.

Though this error does not change the meaning of the sentence entirely, makes the sentence sound odd.

Likewise, in the example-(5.4), for the demonstrative 'these', a personal pronoun 'they' *vīriki* instead of '*vīṭiki*' is used. Interestingly, this mis-translation changes the meaning of the sentence completely.

(5.4) 'Therefore these have the maximum importance in man' **[Eng.]**
aMduvallanē *vīri-ki* *puruṣulalō* *atyadhika* **[Tel.]**
due to that they.DAT men-LOC maximum
prāmukhyata *unna-di*
importance have-3.SG.N **[Bing error]**

5.3.1.3 Oblique Marking

Oblique case denotes the change in the nominal stem when it accommodates other case relations, for eg. *illu*(direct), *iMṭi* (oblique). Any such errors including wrong interpretation of oblique or unnecessary inclusion of oblique case falls under this section. Consider the following examples:

(5.5) 'According to the point of view of medical sciences, knee is a weak synovial joint' **[Eng.]**
vaidya *śāstrāla* *drokkoṇaM* *prakāraM* *mōkāli* **[Tel.]**
medical-OBL sciences-OBL point of view according knee-OBL
balahīnam-aina *sainōvial* *jāiMṭ*
weak-ADJ synovial joint **[LingvaNex error]**

In the example-(5.5), the word 'mokali' is in the direct case, however it is translated as *mōkāli* 'knee-OBL' instead of the direct case, *mōkālu*.

(5.6) 'In maximum people the straight or right hand is much stronger and dexterous' **[Eng.]**
gariśṭa *vyaktula-lō* *saraḷa* *lēda* *kuḍi* *cēti* *cālā* **[Tel.]**
maxim um people-LOC straight or right-OBL hand much
balaM-gā *mariyu* *nirlakṣyaM-gā* *uM-ṭuM-di*
stronger and dexterous be-HAB-3.SG.N **[LingvaNex]**
As explicated in the example-(5.6), the token 'right hand' is in direct case however, when translated into Telugu, it is marked in oblique form. The correct translation is

cēyi 'hand' in its direct case. As mentioned above, this leads to confusion in the source text interpretation.

5.3.2 Syntactic Errors

Syntactic errors contribute to a majority of errors that affects the grammaticality and in-turn affects the comprehensibility of the text. Syntactic errors include errors pertaining to case-mismatch, errors with specific constructions like non-nominative, participial, coordination, relative, participial phrasal verb and issue of translating determiners. Each type of syntactic error is explicated with illustrations in this section.

5.3.2.1 Case Mismatch

Case-mismatch is a prominent syntactic error that arises when the target languages gets a wrong case marker

(5.7) 'Taking 1 spoon from it make the child drink 4 times a day.' **[Eng.]**
dāni nuMḍi 1 *ceMcā tīsukun-i pillavāḍi-ni rōju-ku* **[Tel.]**
it from 1 spoon taking child-ACC day-DAT
4 *sārlu trāgāli*
4 times drink. [[**Google error**]].

In the sentence-(5.7) translated by the Google MT, it can be observed that the recipient 'pillavāḍu' is marked with the accusative case marker. However, there recipient requires to be the dative case marker. In this case, the TL translation does not affect the comprehensibility to a large extent, however, it affects the grammaticality of the sentence.

(5.8) 'Bathe the child with soap and clean water properly' **[Eng.]**
sabbu mariyu subhramaina nīṭitō pillavāḍini sariggā **[Tel.]**
soap and clean water child properly
snānaM cēyaMḍi
bath do-IMP [**Yandex error**]

In the example-(5.8), considered from the Yandex MT, wherein 'the child' must be in the dative case but translated as *pillavāḍini* accusative case.
Case mismatches contribute to the grammaticality quotient of the translated output.

5.3.2.2 Quirky Subjects

Telugu allows quirky or non-nominative subject construction in which the actual subject or the doer or experiencer is in not in nominative case (usually occurs in dative, locative etc) whereas the theme is in nominative case. These constructions are absent in English. Among non-nominative subject constructions, dative subject is the most common construction in Telugu (Bhaskararao and Subbarao, 2004). It occurs with nouns indicating physiological state, psychological state, possession, cognitive state, etc. However, English encodes all such information in the nominative case on the nouns. Hence, when such constructions have to be translated to Telugu, the system often fails to provide the correct output. Consider the mistranslated sentences from IIIT-H:

(5.9) 'If the child gets cold then bring in use some domestic techniques.' **[Eng.]**
pillavāḍu *jalubu* *cēs-tē*, *konni* *dēśiya* *paddatu-la-nu* **[Tel.]**
child cold get-COND some domestic technique-PL.ACC
upayōgiMcāli
use-3.HOR **[IIIT-H]**

In the above sentence, the noun 'child' functions as an experiencer in Telugu. The verb 'jalubu cēyu' denotes a 'physiological state' which requires the subject to be experiencing this state. Such noun phrases are case-marked with '-ki/ku'. However, the given output marks the noun 'child' with a nominative case which results in an erroneous output. Direct translation of case from English to Telugu is unacceptable and results in such errors. The correct translation must be '*pillavāḍiki jalubu cēstē konni dēśiya paddatulanu upayōgiMcāli*'.

5.3.2.3 Agreement Error

Telugu marks gender, number, person information on the verb. This GNP information on the verb often helps to disambiguate the subject in pro-drop sentences. Issues with wrong GNP marking on the verb, predicate adjective etc are considered part of agreement issues.

(5.10) 'The day he does not wet his bed, praise him and take outside' **[Eng.]**
atanu tana mancham taḍi cēyani rōju ata-nni **[Tel.]**
he his bed wet does not day him-ACC
meccukun-i bayaṭ-ki *tīsukuveḷa-tā-ḍu*
praise-CONJP outside take-HAB-3.SG.M **[LingvaNex]**

The example-(5.10) is an output by LingvaNex. the second person subject is dropped in the clause with the verb 'praise' and 'take'. It is known that the subject of the imperative construction is the second person. However, here, the agreement of the finite verb is in the third person, singular, masculine form which makes the sentence confusing to interpret.

(5.11) 'Earlier, we mentioned 12 cranial nerves out of which 7 had their origin from medulla.' **[Eng.]**
antaku-mundu mēmu 12 *kapāla narāla-nu* **[Tel.]**
then-before we-EXCL 12 cranial-OBL nerve-ACC
prastāviMc-ā-mu viṭilō 7 *vāti mūlānni meḍullā nuMḍi*
mention-PST-1.PL these LOC 7 their origin medulla from
kaligi-uM-di
have-be-3.SG.N **[LingvaNex]**

In the LingvaNex Telugu output-(5.11), the number agreement of the finite verb 'have' is wrongly marked as singular. The correct translation is *kaligi unnāyi*.

5.3.2.4 Voice Error

Voice errors occur when an active voice is translated as passive or vice-versa. Consider the following examples:

(5.12) 'Medical research has already even proved it many years ago.' **[Eng.]**
vaidya pariśōdhana-lu cālā saMvatsarā-la **[Tel.]**
Medical research-PL many year-[PL]
kritamē nirūpiMca-baḍḍ-ā-yi.
ago prove-PASS-PST-3.PL **[Google]**

Here, in the example -(5.12), the input sentence is in active voice. But the google MT translated it into a passive sentence changing the complete meaning of the sentence. 'Medical research' is the subject of the sentence, but as the verb is translated to a passive verb it makes 'medical research' the object of that sentence. This alters the meaning of the sentence entirely. The actual sentence is 'Medical research has already even proved it many years ago' but is altered as 'medical researches were proved many years ago'.

5.3.2.5 Causative Constructions

Causative constructions in Telugu are marked morphologically on the verb. Whereas in English it is represented structurally using the lexical item 'make'. When causative constructions are to be translated into Telugu, the 'iMcu' suffix is added to the verb. It is observed that MT systems often provide mistranslated output for causative constructions. Consider the examples from LingvaNex and Google & IIIT-H MT systems.

(5.13) 'Bathe the child with soap and clean water properly' [Eng.]
sabbu mariyu subhramaina nīṭitō pillavāḍini sariggā **[Tel.]**
soap and clean water child properly
snānaM cēyaMḍi
bath do-IMP **[LingvaNex]**

The verb bathe can be interpreted both as a transitive/intransitive verb i.e. 'to take/give bath'. For example,' I bathed'/I bathed the child'. Here, in example-(5.7), 'bath' refers to a transitive verb as the object 'the child' is present in the sentence. However, in the case of Telugu, the verb 'bath' is a noun-verb compound. An intransitive verb in Telugu can be turned into a transitive verb using causativization (Krishnamurti and Gwynn, 1985b). Hence, to express a transitive sense, causative construction is used. This divergence between English and Telugu is not identified by the system resulting in erroneous output. The correct translation of the given input is '*sabbu mariyu subhramaina nīṭitō pillādiki snānaM cēyiMcaMḍi*'.

(5.14) 'Taking 1 spoon from it make the child drink 4 times a day.' [Eng.]
dāni nuMḍi 1 *ceMcā tīsukuni pillavāḍini rōju-ku* **[Tel.]**
it from 1 spoon taking child-ACC day-DAT
4 *sārlu trāgāli*
4 times drink. **[Google]**

In the example-(5.14), the English input sentence contains a causative verb 'make the child drink', however, causative is absent in the translation, making it merely a transitive verb. As there is an error in the finite verb, the whole can be misinterpreted contributing to a major error.

(5.15) 'Make the child have his breakfast by 7:30 or eight in the morning.' [Eng.]
udayaM 7:30 *lēdā* *enimidi* *gaMṭalaku* *pillala-ku* [Tel.]
morning 7:30 or eight clock child-DAT
alpāhāraM *tīsukoMḍi*
breakfast have [**IIIT-H**]

5.3.2.6 Coordinate Constructions

Coordinate constructions in Telugu are formed using the conjunction *mariyu* 'and' joining the conjoints. Coordination also happens with the vowel lengthening of final vowel of the conjoints. However, it should be noted that conjoints of whichever lexical category they belong to, should be in the same case for nouns/pronouns and same TAM for verbs. In the MT evaluation, it is found that coordinate constructions is another challenge for most of the MT systems. Consider the following examples:

(5.16) 'The question is in fact very interesting and important.' **[Eng.]**
prańa *vāstavāniki* *cālā* *āsaktikaraMgā* **[Tel.]**
question infact very interest-ADV
mariyu *mukhyamainadi*.
and important. **[Google]**

The google output-(5.16), the conjoints are 'interesting and important' which are translated as *āsaktikaraMgā mariyu mukhyamainadi*. The first conjoint is with the adverbial marker whereas the second is the predicative adjective. The correct translation must translate both as the predicative adjectives. The google MT system fails to provide the appropriate translation leading to grammatical and comprehensibility errors.

5.3.2.7 Relative Clause

A simple sentence in Telugu can be changed into a relative clause by replacing its finite verb by a relative participle (or verbal adjective) in the corresponding tense-mode and shifting the noun that it qualifies as head of the construction (Krishnamurti and Gwynn, 1985b). These are one more difficult arena for MT system to translate. Consider the following error with relative clause formation in Google MT system:

(5.17) 'Because of diabetes deep wounds that are non-healing occur.' [**Eng.**]

madhumēham *kāranaMgā* *lōtaina* *gāyā-lu* [**Tel.**]
diabetes reason-ADV deep wound-PL
mānaḍam lēdu
healing be.NEG [**Google**]

In the sample translation from Google-5.17, the actual translation must be *madhumēhaM kāranaMgā mānalēni lōtaina gāytālu ērpaḍatāyi*. But the system failed to provide the relative clause translation of the clause 'deep wounds that are non-healing'.

5.3.2.8 Participial Clauses

Participial clauses are subordinate clauses that modify the matrix clause. Participial clauses include conjunctive participles indicating serial action, manner etc.

(5.18) 'Often in childhood sunburns occur by roaming a lot in sunlight.' [**Eng.**]
taracugā *bālyaM-lō* *sūryakāMti-lō* *cālā* *tirugutū* [**Tel.**]
often childhood-PREP sunlight-PREP alot roam-PROG-PART
sanbarns *sambhavistāyi.*
sunburns occur. [**IIIT-H**]

In the IIIT-H translated output, the gerund 'roaming' in the SL is translated using a participial constructions when it is not required. The correct translation is *taracugā bālyaMlō sūryakāMril¯o cālā tiragaḍaM valana sanbarns sambhavistāyi*. The reason clause 'tiragaḍaM valana' is mistranslated as a participial constructions.

(5.19) 'Virus also enters in the child body by repeatedly kissing him.' [**Eng.**]
vairas *kūḍā* *śiśuvu* *śarīraM-lō* *padēpadē* [**Tel.**]
virus also child body-PREP repeatedly
muddupeṭṭukon-i *pravēśistuMdi*
kiss-CONJP enters. [**LingvaNex**]
LingvaNex also mistranslates the reason clause 'by repeatedly kissing' in the example-(5.19) as a conjunctive participial clause leading to a complete misinterpretation of the SL text.

5.3.2.9 Phrasal verbs

Phrasal verbs like 'take off', 'try on' etc. observed to be mistranslated by the selected MTs in most cases. Consider the following cases.

(5.20) 'Other than this, this virus can be finished off in a temperature of 75-199 degrees.' [**Eng.**]
idi *kākuMḍā* *ī* *vairas* 75 *nuMḍi* 100 [**Tel.**]
this be-NEG this virus 75 from 100
ḍigrīla *uṣṇōgratalō* *pūrti* *cēyavaccu*
degree temperature finish do-can [**Yandex.**]

In the sample translation-(5.20), the phrasal verb 'finished off' which means to 'kill' is mistranslated as 'pūrti cēyu' (to finish(lit.)). However, this meaning does not fit the context and leads to a mistranslation.

5.3.2.10 Determiners

In Telugu, determiners as a separate category do not exist(Krishnamurti, 2003). The specificity that determiners denote is encoded in Telugu using number words viz, oka/okaṭi. Determiners are marked as case marker to indicate specifity by some MTs leading to erroneous output. Consider the example from IIIT-H.

(5.21) 'Eat a chapati less at night so that the stomach stays light.' [**Eng.**]
rātripūṭa *capātini* *takkuvagā* tinaMḍi [**Tel.**]
night chapati less eat
tadvārā *kaḍupu* *tēlikagā* *uMṭuMdi*
so that stomach light stays. [**IIIH-T**]

In the translated output-(5.21), 'a chapati' 'oka capāti' is translated as *capātini* which is grammatically ill-formed.

5.3.3 Semantic Errors

Semantic errors include all errors pertaining to the meaning of a word, phrase or the sentence. Semantic errors contribute largely to the comprehensibility factor. Even if the sentence is grammatical, if the selection of words is incorrect, the whole sentence stands incomprehensible.
Semantic errors are classified into errors with semantic in incompatibility

homophones, homograph, homonym, polysemy, set collocates, Named-entities, technical terms.

5.3.3.1 Semantic Incompatibility

Collocates in a sentence are mutually compatible with each other in meaning to convey a specific sense in a sentence. For example, 'old' in 'old man' and 'old clothes are two different elements in Telugu. The same sense of 'oldness' is encoded differently in clothes and in man. In Telugu, 'old man' translates to '*musali vādu*' whereas 'old clothes' translates to '*pāta battalu*'. Hence, words in a sentence must be semantically compatible with each other.

(5.22) 'In this season stale food should not be eaten.' **[Eng.]**
ī sījaMlō pāta āhāraM tinakūḍadu **[Telu.]**
this season stale food eat **[IIIT-H]**

In the sentence-(5.22), stale is translated as '*pāta*' which literally translates to 'old' which is usually used for inanimate inedible items. In this case, food cannot be compatible with the collocate old. The appropriate word for stale should be '*niluva unna*' .

(5.23) 'The initial 12 years of age are extremely important' **[Eng.]**
prāraMbha 12 *saMvatsarālu vayassu cālā* **[Tel.]**
initial] 12 years age extremely
mukhyamainadi
important **[LingvaNex]**

The word 'initial' can have multiple interpretations based on the context like '*modati, toli, prārambha*' . In the context of sentence-(5.23), 'initial 12 years' should be translated to 'modati 12 samvastrāla' whereas the given system translates it with 'prārambha' which is correct in terms of word-word translation but is not appropriate in the context of this sentence.

(5.24) 'What is to be understood is that only the passage is closed'
artham cēsukōvalas-ina viṣayaM emi-ṭaMṭē, **[Eng.]**
understand do-REF-OBLI-ADJ matter what-QUO prakaraṇaM mātramē
chapter only

mūsi-vēyabaḍu-tuM-di
close-DO-PASS-PST-3.SG.N

In 5.24, the word 'passage' is used in the context of blockage of blood vessels. However, in the translation, the word passage is translated as a 'chapter'. Here, it is contextually wrong and calls out a need for synsets to identify the domain of the corpus.

(5.25) If bone breaks then it takes a long time for it to join
emuka virigitē adi cēr-aḍāniki cālā samayaM paḍu-tuM-di
bone break-COND that join-for very time fall-HAB-3.SG.N

In 5.25, the word 'join' is used in the context of 'bone joining'. Telugu uses several words for 'join' based on the noun, like 'join a class' *klās-lō cēru*, etc. Here, 'join' means 'atakaḍaM' 'to stick' together(lit). However, it is not correctly translated.

5.3.3.2 Lexical Mismatch

Some words often have specific meanings in the language. This information, sometimes, is often used for other terms, which results in erroneous output. All such cases are considered under this section

(5.26) 'For the cleanliness of the scalp long hair should be cut short' [**Eng.**]
juṭṭu yokka parishubhrata kōsaM poḍavāṭi [**Tel.**]
scalp of cleanliness for long
juṭṭunu cinnagā kattirincāli
hair short cut [**Yandex**]

'Scalp' and 'hair' in sentence -(5.26), are both translated as 'juṭṭu' (hair) in the Google output which results in inaccurate translations.

5.3.3.3 Lexical Mapping

In this section, we discuss issues lexical mapping. Certain words and phrases which are part of the basic terminology of the target language are not successfully mapped by the system. Such words are observed to be transliterated instead of translating. Consider sentence:

(5.27) 'The secret of pink cheeks is the consumption of apple.' [**Eng.**]
pink buggala rahasyaM āpil tīsukōvaḍaM [**Tel.**]
pink cheeks secret apple consumption. [**Yandex.**]

The color pink is translated to 'gulābi raMgu' in Telugu which is quite commonly used. It is unnatural to transliterate such basic color terms. However, it is interesting to note that google translate provides a correct translation when the word 'pink' is given in isolation. But it fails to translate when it is part of a sentence.

5.3.3.4 Homophonous

Homophones are words that have same pronunciation and are written differently. Most MT systems seem to be confused with homophones. Consider the error obtained from LingvaNex MT:

(5.28) 'Going out to roam is also a good solution to reduce tension' [**Eng.**]
rōM-ku *veḷḷ-aḍaM* *kūdā* *udrikthathanu* [**Tel.**]
Rome/roam go-GEN also tension
tagginchdaniki *manchi* *pariskāraM*
reduce good solution [**LingvaNex**]

As explicated in (5.28), the word 'roam' is transliterated as 'Rome' which is a blunder and leads to a complete non-sensical construction.

5.3.3.5 Homographs

Homographs are defined as words that are written alike and differ in pronunciation. Like in case of homophone, homograph misinterpretation is also common the systems.

(5.29) 'In diabetes, large and minute complications can be born in the vessels.' [**Eng.**]
diabetis *nāḷālalō* *pedda* *mariyu* [**Tel.**]
diabetes nerves-LOC large and
nimiśāla *smasyalu* *talethutāyi*
minute complications born. [**Google**]

The word 'minute' in 5.29 refers to an 'extremely small' but in the above output, it is translated as *nimiṣāla* 'minutes'(as in time) which is a completely wrong interpretation of the given SL sentence.

5.3.3.6 Homonyms

Homonym refers to words that are written and pronounced in the exactly same way. Some such words are incorrectly translated by the system. Consider the example from the Google MT:

(5.30) 'Every time after motion wash his hands and legs properly.' [**Eng.**]
kadalika/calanaM *taruvata* *pratisāri* *atan-ni* *cētulu* [**Tel.**]
motion after everytime he-ACC hands
mariyu kāllanu sariggā kadagāli
and legs properly wash. [**Google**]

In the example-(5.30), the word 'motion' is translated as 'movement' wherein here, it refers to 'human excreta'.
Errors pertaining to homophones, homographs and homonyms are a serious hurdle to MTs and have to be complemented with synsets to provide correct translation.

5.3.3.7 Polysemy

Polysemy refers to a semantic phenomenon wherein a single word can have multiple variants used based on the context. Sometimes, related words that are irrelevant in the given context can be used that in-turn lead to wrong semantic interpretation. Consider the examples from LingvaNex and IIIT-H in their wrong usage of polysemous words

(5.31) 'Keep in mind that for oily, dry, fine, sensitive skins etc the lotions are also of different types.' [**Eng.**]
gurtuMcukōMdi *nūne* *poḍi* *cakkaṭi* *sunnitham-aina* [**Tel.**]
keep in mind oily dry fine sensitive-ADJ
tokkalu *modalaina* *vāṭi* *kōsaM* *lōṣanlu*
skin that for their for lotions
kūdā *vibhinna* *rakālugā* *uMtāyi.*
also different types are. [**LingvaNex error**]

In the example-(5.31), the word 'skin' here, refers to the skin on human body, however, the system translated it as a 'peel'.

(5.32) 'The small masses of some cells only may be seen somewhere in middle' [**Eng.**]
konni *kośikala* *cinna- cinna* *janasamūhaM* [**Tel.**]
some cells small masses

madhya-lō *ekkadō* *cūdavaccu*
middle-in somewhere seen. [**IIIT-H**]

As explicated in (5.32), the word 'masses' has multiple polysemous words in Telugu. However, 'masses' here, refers to inanimate cells which must be translated just as *samūdāyaM* however it is translated as 'human masses'.

5.3.3.8 Multi-Word Expressions

This section deals with multi-word expression errors. multi-word expressions need at least two words to form. they work as units which makes them very distinctive in nature. Furthermore, the multi-word expressions are devised into three types which are seen following:

1. Set Collocations

Set collocates refer to a pair or group of words that occur together as a set. For example, fit and healthy, etc. It often happens so that set collocates might not have exact equivalents in the target language. Hence, they might look unaligned in the sentence or same words are used for both the words in the set. This leads to mistranslation. Consider the following example from LingvaNex MT.

(5.33) 'Body builders and athletes are stay absolutely fit and healthy.' [**Eng.**]
bādī *bilder-lu* *mariyu* *athelt-lu* *khaccithangā* [**Tel.**]
body bilder-s and athletes absolutely
aārōgyaMgā *mariyu* *aārōgyaMgā* *uMntāru*
fit and healthy are. [**LingvaNex**]

As mentioned earlier, 'fit and healthy' is translated as '*aārōgyaMgā mariyu aārōgyaMgā*' which are both exactly the same words. The probable correct translation would be '*dhru-ḍMgā mariyu aārōgyaMgā*'

(5.34) 'you will have to feed the child with your hand little by little.' [**Eng.**]
mīru *pillavādi-ni* *cēti-tō* *koddiga* [**Tel.**]
you child-DAT hand-INST little by little
tinipiMcālsi *uMtuMdi*
to feed will have. [**IIIT-H**]

Another example includes-(5.34) by IIIT-H output that translates 'little by little' as just 'little' leading to an error.

2. Named Entities

Named entities indicate proper nouns or common nouns of a specific type having a specific translation in each language. Some such words are wrongly translated by some MTs. Consider the translation of 'wood-apple' by IIIT-H:

(5.35) 'In cramp or internal pain, grind the leaves of wood - apple and cook in jaggery.' **[Eng.]**
gajji lēka lōpali noppilō cekka yapil aākulanu nōri **[Tel.]**
cramp or internal pain wood apple leaves grind
bellaMlō uḍikiMcāli
jaggery cook. **[IIIT-H]**

The 'wood-apple' is translated literally as 'cekka yapil' which does not make any sense.

3. Scientific Terms

Scientific terms like the terminology from a specific register like medical terms, computer terms etc., fall under this section. Some technical terms are translated, leading to a very absurd sentence like the following example by Bing:

(5.36) 'If there is a deformity in your feet, like there is corn or hammer-toe in feet.' **[Eng.]**
mī pādālalō vakalyam unn-atl-ayitē pādālalō **[Tel.]**
your feet deformity is-there-if feet
mokkajonna lēdā sutti boṭanavēlu uMtuMdi
corn or hammer toe is. **[Bing]**

In the example-(5.36) the word 'corn or hammer-toe' has specific meaning that is *āne lēdā vankara vēlu but system produced a literal translation 'sutti boṭanavēlu'* which is not correct.

(5.37) 'It will be better that you take the shoes to the podiatrist. [**Eng.**]
mīru *pādarakshala* vaddaku būṭu tīsukeḷḷaḍaM maMcidi. [**Tel.**]
you podiatrist to shoes take better [**Yandex**]

The technical term 'podiatrist' refers to 'a doctor who treats ailments related to feet' is translated as 'pādarakshalu' literally meaning 'feet-protector'. But this word 'pādarakshalu' is used for footwear in Telugu. For terms like this, stringing words together does not make a meaningful compound. Such errors have also been identified.

(5.38) 'night blindness: the first symptom of xerophthalmia.' [**Eng.**]
rātri- aMdhatvaM jirōphtālmiya yokka modaṭi laksnnaM. [**Tel.**].
night blindness xerophthalmia of first symptom

In the (5.38), the term 'nightblindness' is a scientific term and has a specific translation as 'rēcīkaṭi'. However, the system translated it literally into 'night' *rātri blindness* 'aMdatvaM'.

(5.39) 'Do not give water to the child apart from breast-feeding.' [**Eng.**]
stanyam ivvaḍaM kk̄uMdā bidda-ku nīru ivva-vaddu. [**Tel.**]
breast feeding apart from child-DAT water give-not. [**Bing**]

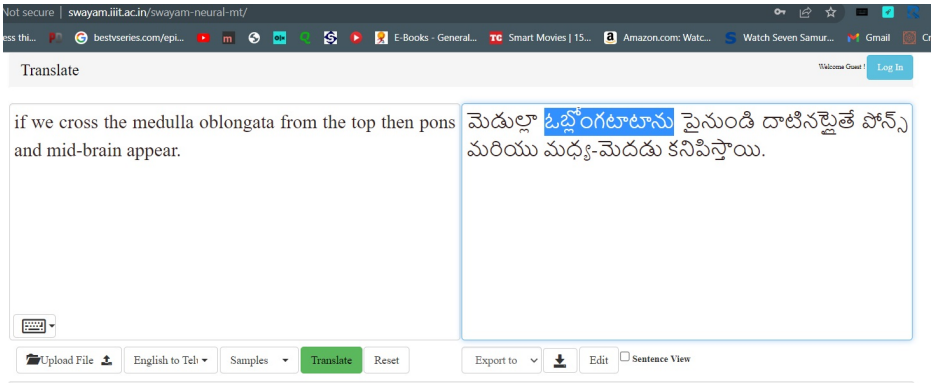
In the example-(5.39 the word 'breast-feeding' has specific meaning that is *p[̄a]lu ivvaṭM*. But system produced a literal translation 'stanyam-ivvaḍaM' which is not correct.

5.3.4 Miscellaneous Errors

Miscellaneous errors include orthographical, incomplete translations, system errors, punctuation errors etc.

5.3.4.1 Transliteration error

Transliteration errors are wrong transliteration in the script of the TL language text. When the Transliterated TL text is not orthographically appropriately represented by the system. we categorize them as transliterated. Consider below errors by NMT system.



(5.40) 'if we cross the medulla oblongata from the top then pons and mid-brain appear.' ' **[Eng.]**

medulla *oblongata-ta-nu* *pai* *nundi* *datinatlayite*' **[Telu.]**

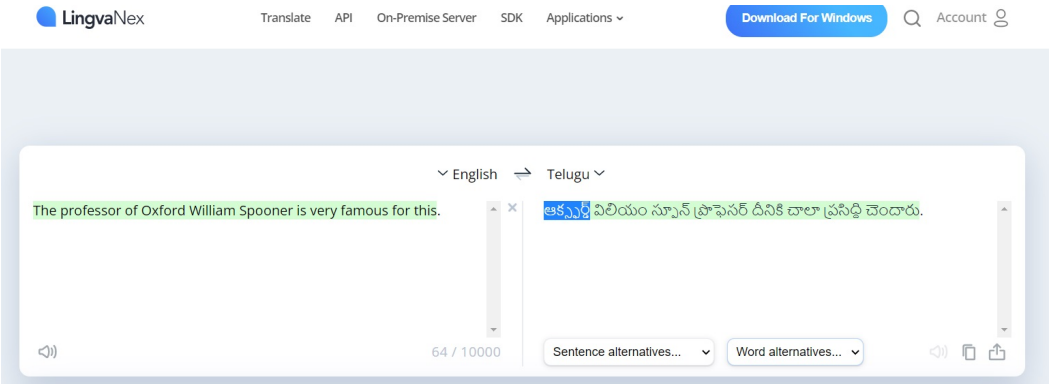
medualla oblongata top from cross

pons *mariyu* *madhya* *medhadu* *kanipistayi*

pons and mid brain appear

' **[IIIT-H]**

In the example-5.40 the English word 'oblongata' is transliterated as 'oblongata-ṭa-nu' which is orthographically incorrect in Telugu. In the transliteration extra 'ta' is added marked in red in the example. which makes the meaning senseless.



(5.41) 'The professor of oxford william spooner is very famous for this.' **[Eng.]**

oksafard *itviliyam* *spūner* *profesar* *diniki* *prasiddi* *cendāru* **[Tel.]**

oxford william spooner is very famous for this. **[IIIT-H]**

In the example-(5.42) the English word 'oxford' is transliterated as 'aks-a-fard' which is orthographically incorrect in Telugu. In the transliteration extra vowel 'a' is added

marked in red in the example. The coreect one would be 'aksfard'

5.3.4.2 Punctuation

Punctuation errors are unique kind of errors in which a punctuation mark like a comma or exclamation brings in a change in the translation. Consider the sample translation by google:

(5.42) 'Like a skilful computer cerebellum does two very important jobs.' [**Eng.**]
naipuṇya kaligina kampyūṭar serebellam lāgā [**Tel.**]
skilful have computer cerebellum like
reṇḍu mukhyamaina panulu cēstun-di.
two importance jobs do-3.SL.. [**Google**]

In the sample translation-(5.42), There is no punctuation is used and the translated output is correctly interpreted as 'like a skilful computer cerebellum does two important jobs'. However, when a 'comma' is inserted after computer in the input sentence, the meaning of the sentence completely differs from the original input sentence and translates literally to 'A skilful computer cerebellum, does two very important jobs'

5.3.4.3 Incomplete Sentence

This section deals with lexical mismatch from source to target text. It is observed that often phrases or clauses from the source text are missing in the target language translation. This increases the risk of missing out on important information from the source language. Some such sentences which are not completely translated are listed below:

(5.43) 'Feed nutritious food to the child by making it tasty.' [**Eng.**]

rucikaramaina āhārānni pillalaki tinipincaṇḍi [**Tel.**]
tasty food child feed. [**IIIT-H**]

In the above example (5.43), the noun phrase nutritious food is not translated in the output sentence. The actual output translation should be as following

pōṣaka āhārānni rucikar-aMgā vaMḍi pillala-ku tinipiMcaMḍi
nutritious-ADJ food tasty make child-ACC feed

(5.44) 'These pills provide cent per cent protection to the woman.' [**Eng.**]
ī mātralu strīki śāta rakṣaṇa kalpistāyi. [**Tel.**]
These pills woman percent protection provide. [**google**]

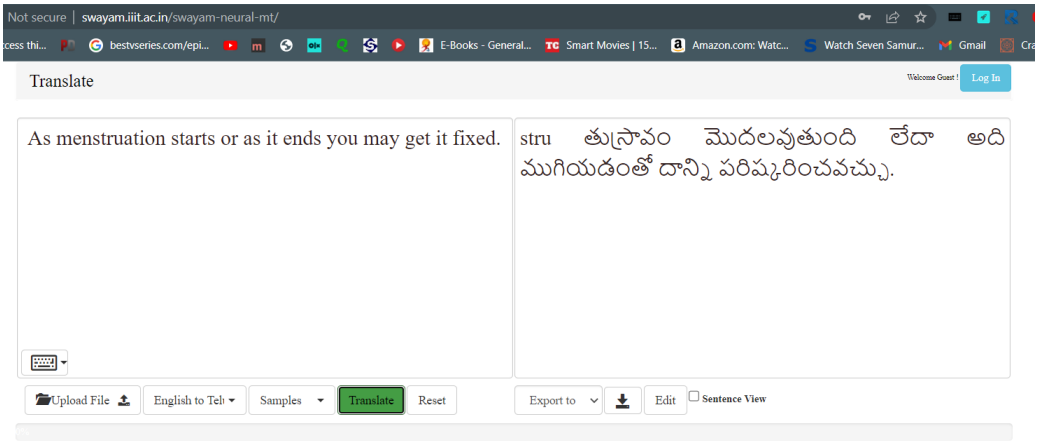
In the example-(5.44), only the word *per cent* is translated and not 'cent' leading to a confusion in interpretation.

5.3.4.4 System Error

System errors are the conditions in Which, some extra random text appears between the translated text which has no meaning. This random text can be predicted as some technical error in MT.

Orthoraphic Mapping

Orthographic errors are errors pertaining to the script of the TL language text. When the TL text is not Orthographically appropriately represented by the system , we categorize them as orthographical errors. consider the below errors by NMT systems.



Consider the script issues that one encounters in the IIIT-H MT system.

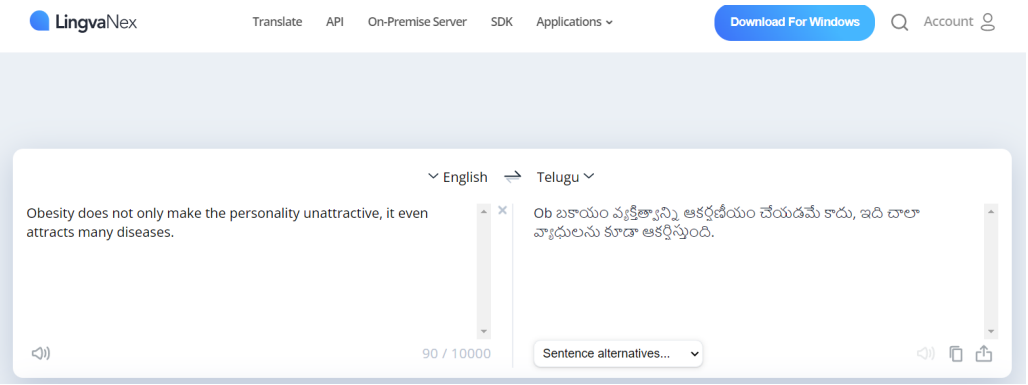
(5.45) 'As menstruation starts or as it ends you may get fixed.' [**Eng.**]
stru tsrāvaM modalavṭṭumdi lēdā [**Tel.**]
menstruation start or
adi mugiyadamtō daanni pariskrimchavachu.
it end it fixed . [**IIIT-H**]

In ex-(5.45) the English word 'menstruation' is translated as '*stru-tusravaM*' which is orthographically incorrect in Telugu. One can observe them in the sentence marked in red.

(5.46) 'In cancer it is fruitful like medicine.' **[Eng.]**
kancer lo idi medicineshadham lantidi **[Tel.]**
cancer in it medicine fruitfuklike **[LingvaNex]**

In the example-(5.46), the english term medicine is translated as medicine-shadham in Telugu.

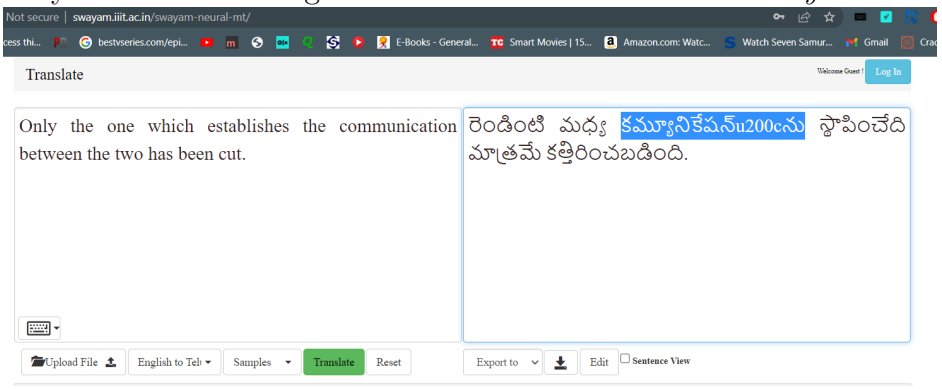
In the example-(5.46) the English word 'medicine' is translated as 'medicine-shadham' which is lexically incorrect in Telugu. One can observe them in the sentence marked in red. The Telugu equivalent is 'aushadham'



(5.47) 'obesity does not only make the personality unattractive, it even attracts many diseases.' **[Eng.]**
Ob-bakayam vyakthitvaani akarshaniyam cheyadame **[Tel.]**

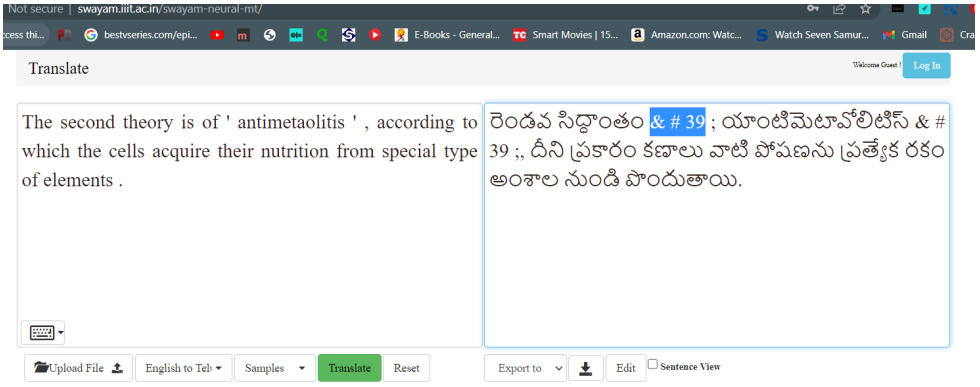
obesity peronality unatarctive does-only
kadu chaalaa vyadulanu kuda akarshistundi.
not it many diseases kuda atract

In ex (5.47) the english word 'obesity' is translated as 'Ob-bakayam', which is lexically incorrect in Telugu. The correct lexical item is *ubakāyaM*.



(5.48) 'Only the one which establishes the communication between the two has been cut.' [**Eng.**]
reMdiMti madhya kamyunikeshan-u200cnu sthapinchedi [**Tel.**]
two between communication establishes
matrame kattimca-baḍindi.
only cut-been

In ex-(5.48) the English term 'communication' is translated as *'kamyunikeshan-u200cnu'*. in which, the correct equivalent term is represented in the target text. but along with this there is also some random and unnecessary text that is 'u200cnu' is added in the end of the translated text marked in the red.



(5.49) 'The second theory is of 'antimetaolitis' according to which the cells acquire their nutrition from special type of elements. [**Eng.**]

rendava *siddantam* *& # 39 ;* *yāMtimetāvōlitis* *& # 39 ;*
dīni [**Tel.**]
second theory error antimetaolitis error which
prakāram *kanālu* *vāti* *pōshaṇanu* *pratyēka* *rakam*
according cells their nutrition special type
aMshāla nundi pondutāyi
elements from acquire.

In ex-(5.49) one can observe that there is random special characters and numerical number that is ‘ # 39 ;’ in Telugu translation which is absent in the English text.

5.4 Error Statistics

In this section, statistics of errors encountered by each system are presented.

Types of Errors	**GMT**	**BMT**	**IIIT-H MT**	**LMT**	**YMT**
Morphological	102	88	105	108	32
Syntactic	130	188	244	375	414
Semantic	366	485	480	525	567
Miscellaneous	316	313	388	415	906
Total Errors	**914**	**1074**	**1217**	**1423**	**1919**

Table 5.1: Over all Error Statistics of Each System

The table (5.1) depicts overall errors from each system. It is observed from the table that the category of semantic errors are high in each and every system and also category morphological errors are less in all systems. So, It can be concluded that the area of semantic data incorporation into the MTs needs to be focused on in order to improve the NMTs translation ability.

In tables 5.2, 5.3, 5.4, 5.5 and 5.6, statistics of various errors has been calculated for NMT systems Google, Bing, IIIT-H, LingvaNex and Yandex respectively. It is observed from the statistics is that google has stood at the top with low number of errors. and Yandex is stood in the last as it has more error rate. All MTs are showing the errors classified in our study and the semantic incompatibility issues are found to be the major issue.

Type of errors	Number	Percentage
Morphological errors	102	11.1
Number Marking	0	0
Person Marking	60	6.5
Oblique Marking	42	4.2
Syntactic errors	130	14.2
Case mismatch	12	1.3
Quirky subject	18	1.9
Determiners	0	0
Agreement	5	0.5
Voice	6	0.5
Causative Construction	25	2.7
Coordinate construction	15	1.6
Relative clause	22	2.4
participial clause	19	2.0
Phrasal Verbs	8	0.8
Semantic errors	366	40
Semantic Incompatibility	193	21.1
Lexical Mismatch	85	9.2
Lexical Mapping	62	6.7
Homophone	0	0
Homographs	2	0.2
Homonyms	1	0.1
Polysemy	6	0.6
Multi-word expressions		
Set collocates	3	0.3
Named Entities	5	0.5
scientific terms	9	1
Miscellaneous	316	34.5
Transliteration	117	12.8
punctuation	10	1
Incomplete sentences	43	4.7
System errors	146	15.9
Total number of errors	**914**	

Table 5.2: Google MT errors statistics

Type of errors	Number	Percentage
Morphological errors	88	8.19
Number Marking	22	2.04
Person Marking	41	3.81
Oblique Marking	25	2.32
Syntactic errors	188	17.5
Case mismatch	15	1.39
Quirky subject	23	2.14
Determiners	2	0.18
Agreement	16	1.48
Voice	12	1.11
Causative Construction	26	2.42
Coordinate construction	11	1.02
Relative clause	39	3.63
participial clause	25	2.32
Phrasal Verbs	19	1.76
Semantic errors	485	45.15
Semantic Incompatibility	209	19.45
Lexical Mismatch	166	15.45
Lexical Mapping	95	8.84
Homophone	0	0
Homographs	1	0.09
Homonyms	0	0
Polysemy	2	0.18
Multi-word expressions		
Set collocates	4	0.37
Named Entities	3	0.27
scientific terms	5	0.46
Miscellaneous	313	29.14
Transliteration	727	6.70
punctuation	15	1.39
Incomplete sentences	97	9.03
System errors	129	12.01
Total number of errors	**1074**	

Table 5.3: Bing MT errors statistics

Type of errors	Number	Percentage
Morphological errors	105	8.62
Number Marking	14	1.15
Person Marking	56	4.60
Oblique Marking	35	2.87
Syntactic errors	244	20.04
Case mismatch	24	1.97
Quirky subject	16	1.31
Determiners	15	1.23
Agreement	30	2.46
Voice	41	3.36
Causative Construction	20	1.64
Coordinate construction	42	3.45
Relative clause	12	0.98
participial clause	31	2.54
Phrasal Verbs	13	1.06
Semantic errors	480	39.44
Semantic Incompatibility	225	18.48
Lexical Mismatch	82	6.37
Lexical Mapping	139	11.42
Homophone	2	0.164
Homographs	0	0
Homonyms	4	0.32
Polysemy	5	0.41
Multi-word expressions		
Set collocates	3	0.24
Named Entities	8	0.65
scientific terms	12	0.98
Miscellaneous	388	31.88
Transliteration	75	6.16
punctuation	25	2.05
Incomplete sentences	115	9.44
System errors	173	14.21
Total number of errors	**1217**	

Table 5.4: IIIT-H MT errors statistics

Type of errors	Number	Percentage
Morphological errors	108	7.58
Number Marking	29	2.03
Person Marking	47	3.30
Oblique Marking	32	2.24
Syntactic errors	375	26.35
Case mismatch	30	2.10
Quirky subject	46	3.23
Determiners	25	1.75
Agreement	20	1.40
Voice	16	1.12
Causative Construction	68	4.77
Coordinate construction	54	3.79
Relative clause	42	2.95
participial clause	31	2.17
Phrasal Verbs	43	3.02
Semantic errors	525	36.89
Semantic Incompatibility	309	21.71
Lexical Mismatch	82	5.76
Lexical Mapping	102	7.16
Homophone	2	0.14
Homographs	3	0.21
Homonyms	6	0.42
Polysemy	11	0.77
Multi-word expressions		
Set collocates	3	0.21
Named Entities	2	0.14
scientific terms	5	0.35
Miscellaneous	415	29.16
Transliteration	64	4.49
punctuation	32	2.24
Incomplete sentences	124	8.71
System errors	195	13.70
Total number of errors	**1423**	

Table 5.5: LingvaNex MT errors statistics

Type of errors	Number	Percentage
Morphological errors	32	1.61
Number Marking	10	0.50
Person Marking	13	0.65
Oblique Marking	9	0.45
Syntactic errors	414	20.86
Case mismatch	31	1.56
Quirky subject	48	2.41
Determiners	20	1.00
Agreement	52	2.62
Voice	30	1.51
Causative Construction	49	2.46
Coordinate construction	56	2.82
Relative clause	71	3.57
participial clause	21	1.05
Phrasal Verbs	36	1.81
Semantic errors	567	28.57
Semantic Incompatibility	382	19.25
Lexical Mismatch	50	2.52
Lexical Mapping	81	4.08
Homophone	3	0.15
Homographs	5	0.25
Homonyms	16	0.80
Polysemy	6	0.30
Multi-word expressions		
Set collocates	12	0.60
Named Entities	3	0.15
scientific terms	9	0.45
Miscellaneous	906	48.94
Transliteration	350	20.66
punctuation	0	0
Incomplete sentences	355	17.94
System errors	201	10.33
Total number of errors	**1984**	

Table 5.6: Yandex MT errors statistics

Chapter 6

Conclusion

The aim of the current study is to evaluate the nmt systems between English-Telugu to assess the translation efficiency of the NMT systems. As a result of the evaluation, errors are classified into multiple linguistics and non-linguistic errors. The modern world has been flooded with information day by day but most of that information is available in some languages only. To make the information accessible to the native languages, which are less resourceful languages, building Machine Translation (MT) for such languages is important. MT is an automatic translation task in which one natural language is translated into another natural language. MT helps to overcome the language barrier and helps in accessing information in the native language. Language data can be fed to machines in different forms such as text, speech or image and they can be translated into multiple languages output using MT. The use of MT services has been spread across almost all domains. As MT gets popular, research in the area has also become a need of the hour. In this study, an attempt is made to evaluate the current nmt systems for their accuracy, comprehensibility and Fluency using an evaluation scale with the help of human evaluation.

The chapter organization of the dissertation is discussed here. The first chapter provides an introduction of MT systems. Different MT systems such as RMT, SMT and NMT systems are briefed along with their architectures. In RMT systems direct, transfer and interlingua methods are discussed based on the vanquos triangle. Then, corpus-based systems like EBMT, SMT and HMTs have been provided along with their architectures. The architecture of NMTs is also discussed. Short introduction of the NMTs are given. The review of literature has been done on the evaluation of foreign, Indian and English to Telugu MT systems. In the end of this chapter, the methodolgy of the study, limitation and chapterizations are discussed comprehensively.

The second chapter provides brief review of the machine learning methods like supervised, unsupervised, reinforcement and semi-supervised learning. To understand better the architecture flow diagrams of the each learning is provided.

This chapter also deals with Neural Network and its types like RNN, CNN and self-Attention and Transformer models. It also introduces the avialble Enlgish to Telugu MT systems: Google, Bing, LingvaNex, IIIT-H, Devnagri, Yandex and etc.

The third chapter states evaluation methods used in MT output. Two types of are followed viz. human and automated evaluation. Under human evaluation methods: directly expressed judgment (DEJ) and Non-directly expressed judgement (NoN-DEJ) evaluation methods are reviewed and comprehensively. The DEJ method includes adequacy, accuracy, comprehensibility, Ranking method and direct assessment. The non-DEJ method includes semi automated, task-based, error claddification and alalysis and post editing method are reviewd comprehensively. Furthermore automatic evalution methods included: word order rate, translation error rate meteor position-independent word error rate and BLEU methods. Each of the automatic method are provided along with their formulas. Then, the adpoted evaluation methodology for the current study is discussed here. this section gives complete picture of all steps: test data collection, evaluation criteria, human evaluation methods, human evaluators, automatic evaluation methods and Inter-Rater Agreement.

Fourth chapter provides the results of MT evaluation of the selected 5 MT systems which are Google, Bing, IIIT-H, LingvaNex and Yandex. The results are calculated in terms of adequacy, fluency, comprehensibility, BLEU score and inter-rater agreement by using different measurement scales. Google has performed well and scored however various types of errors found with respect to English-Telugu MT. high and stood at the top. yandex perfomed poorly and scored very less in comparision with other selected systems and occupied in the last place. the evaluation process is very important for the further development of the MTsystems.

Fifth chapter the core part of this study is attempted the classification of errors and analysis of the output. This chapter includes the discussion of errors such as linguistic errors such as, morphological, syntactic and semantic and miscellaneous errors of NMT systems. These errors are discussed and explained in detail with suitable examples. Classifying errors of the each system provides advantages and shortcomings by providing the complete picture of where it needs to be worked on to improve the translation efficiency of any given NMT. As very limited study had been conducted so far in this area, which gives huge scope to conduct a research for

the development the current NMT systems, considering the error types discussed here. Though various linguistic and non-linguistic errors occurred. Predominantly semantic incompatibility issues are high and all NMTs are facing this issue.

Future work:

The current study evaluated the only 5 open source English-Telugu NMT systems based on their performance in the pilot study. Also this area several other open source NMT systems available which needs to be studied further.

Limitation the study includes, the present error classification is proposed on analysing the very less input data, 2000 sentences, pertaining to health domain only. Thus there is a scope to extend this study by collecting the corpus from various other domains and taking large amount of corpus, which may help in improving the current error classification in our future work.

As the Neural MT systems adopt cutting edge technology and updating themselves time to time, the results of the current study is treated as transient. Thus, it is likely that a comparative study of evaluation of NMT systems can be attempted. by conducting the same study in near future to track how well the current systems will improved in terms of their translation efficiency. NMTs being block-box systems, it is highly difficult to suggest the improvement in terms of their algorithm , however based on the evaluation and the error typed discussed in this research can be used to develop bench-mark dataset for development, training and testing NMT systems.

Bibliography

ALPAC, Languages. 1966. Machines: Computers in translation and linguistics, report by the automatic language processing advisory committee, division of behavioral sciences. *National Academy of Sciences, National Research Council, Publication*, 1416. 9, 30, 41

Alpaydin, Ethem. 2020. *Introduction to machine learning*. MIT press. 14

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. 19

Balyan, Renu, Sudip Kumar Naskar, Antonio Toral, and Niladri Chatterjee. 2013. A diagnostic evaluation approach for english to hindi mt using linguistic checkpoints and error rates. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 285–296. Springer. 11

Banerjee, Satanjeev and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72. 10

Bharati, Akshar, Vineet Chaitanya, Amba P Kulkarni, and Rajeev Sangal. 1997. Anusaaraka: Machine translation in stages. *VIVEK-BOMBAY-*, 10:22–25. 4

Bhaskararao, Peri and Karumuri Venkata Subbarao. 2004. *Non-nominative subjects*, volume 1. John Benjamins Publishing. 63

Bhattacharyya, P. 2015. *Machine Translation*. CRC Press. 5, 6, 7

Biberauer, Theresa. 2008. *The limits of syntactic variation*, volume 132. John Benjamins Publishing. 57

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara

Logacheva, Christof Monz, et al. 2016a. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198. 31, 32

Bojar, Ondřej, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016b. Results of the wmt16 metrics shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 199–231.

Bre, Facundo, Juan M Gimenez, and Víctor D Fachinotti. Prediction of wind pressure coefficients on building surfaces using artificial neural networks. *Energy and Buildings*, 158.

Bre, Facundo, Juan M. Gimenez, and Víctor D. Fachinotti. 2018. Prediction of wind pressure coefficients on building surfaces using artificial neural networks. *Energy and Buildings*, 158:1429–1441. 18

Callison-Burch, Chris, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada. Association for Computational Linguistics. 10

Chatzikoumi, Eirini. 2020. How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering*, 26(2):137–161. 30, 32, 33, 36

Chéragui, Mohamed Amine. 2012. Theoretical overview of machine translation. In *ICWIT*, pages 160–169. Citeseer. 7

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46. 46

Daems, Joke, Sonia Vandepitte, Robert J Hartsuiker, and Lieve Macken. 2017. Identifying the machine translation error types with the greatest impact on post-editing effort. *Frontiers in psychology*, 8:1282. 58

Dorr, B., Matthew G. Snover, and Nitin Madnani. 2010. Part 5: Machine translation evaluation chapter 5.1 introduction. 29, 30

EuroMatrix, P. 2007. 1.3: Survey of machine translation evaluation. *EuroMatrix Project Report, Statistical and Hybrid MT between All European Languages, co-ordinator: Prof. Hans Uszkoreit*. 34, 37

Font-Llitjós, Ariadna and Jaime G Carbonell. 2004. The translation correction tool: English-spanish user studies. 58

Garje, Goraksh V and GK Kharate. 2013. Survey of machine translation systems in india. *International Journal on Natural Language Computing*, 2(4):47–65. 4

Gehlot, Akanksha, Vaishali Sharma, Shashi Pal Singh, and Ajai Kumar. 2015. Hindi to english transfer based machine translation system. *arXiv preprint arXiv:1507.02012*. 5

Gehring, Jonas, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252. PMLR.

Görög, A. 2014. Quality evaluation today: the dynamic quality framework. In *Proceedings of Translating and the Computer 36*.

Goyal, Vishal and Gurpreet Singh Lehal. 2009. Evaluation of hindi to punjabi machine translation system. *arXiv preprint arXiv:0910.1868*. 11

Graham, Yvette, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41. 32

Graham, Yvette, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30. 32

Han, Aaron LF, Derek F Wong, and Lidia S Chao. 2012. Lepor: A robust evaluation metric for machine translation with augmented factors. In *Proceedings of COLING 2012: Posters*, pages 441–450. 33

Han, Lifeng. 2016a. Machine translation evaluation resources and methods: A survey. *arXiv preprint arXiv:1605.04515*. 33

Han, Lifeng. 2016b. Machine translation evaluation resources and methods: A survey.

Hassan, Hany, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. 2018. Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*.

Hutchins, W John. 2001. Machine translation over fifty years. *Histoire epistémologie langage*, 23(1):7–31.

Hutchins, William John. 1986. *Machine translation: past, present, future*. Ellis Horwood Chichester.

Jurafsky, Dan. 2000. *Speech & language processing*. Pearson Education India. 3

Kalchbrenner, Nal, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. 19

Kalyani, Aditi, Hemant Kumud, Shashi Pal Singh, and Ajai Kumar. 2014. Assessing the quality of mt systems for hindi to english translation. *arXiv preprint arXiv:1404.3992*. 11

Koehn, Philipp. 2009. *Evaluation*, page 217–246. Cambridge University Press. 7

Krishnamurti, Bhadriraju. 2003. *The dravidian languages*. Cambridge University Press. 56, 68

Krishnamurti, Bhadriraju and John Peter Lucius Gwynn. 1985a. *A grammar of modern Telugu*. Oxford University Press, USA. 56

Krishnamurti, Bhadriraju and John Peter Lucius Gwynn. 1985b. *A grammar of modern Telugu*. Oxford University Press, USA. 65, 66

Lacruz, Isabel, Michael Denkowski, and Alon Lavie. 2014. Cognitive demand and cognitive effort in post-editing. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas*, pages 73–84. 33

Landis, J Richard and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174. 46, 47

Läubli, Samuel, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. A set of recommendations for assessing human–machine parity in language translation. *Journal of Artificial Intelligence Research*, 67:653–672.

Levenshtein, Vladimir I et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union. 34

Massardo, Isabella, Jaap van der Meer, Sharon O'Brien, Fred Hollowood, Nora Aranberri, and Katrin Drescher. 2016. Mt post-editing guidelines. *The Netherlands: TAUS Signature Editions*. 33

Matsuzaki, Takuya, Akira Fujita, Naoya Todo, and Noriko H. Arai. 2015. Evaluating machine translation systems with second language proficiency tests. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 145–149, Beijing, China. Association for Computational Linguistics. 10

Maučec, Mirjam Sepesy and Gregor Donaj. 2019. Machine translation and the evaluation of its quality. *Recent Trends in Computational Intelligence*, page 143. 4, 9, 18

Nguyen, Phuong-Thai and Akira Shimazu. 2006a. Improving phrase-based smt with morpho-syntactic analysis and transformation. viii, 8

Nguyen, Thai Phuong and Akira Shimazu. 2006b. Improving phrase-based statistical machine translation with morpho-syntactic analysis and transformation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 138–147.

Nyberg, Eric, Teruko Mitamura, and Jaime G Carbonell. 1997. The kant machine translation system: from r&d to initial deployment. 6

Ojha, Atul, Koel Chowdhury, Chao-Hong Liu, and Karan Saxena. 2018. The rgnlp machine translation systems for wat 2018. 11

Olive, Joseph, Caitlin Christianson, and John McCary. 2011. *Handbook of natural language processing and machine translation: DARPA global autonomous language exploitation*. Springer Science & Business Media.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics. 10, 37, 44

Popovic, Maja. 2011. Hjerson: An open source tool for automatic error classification of machine translation output. *Prague Bull. Math. Linguistics*, 96:59–68. 58

Popović, Maja. 2020. Informative manual evaluation of machine translation output. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5059–5069. 30, 31, 41

Quirk, Randolph. 2010. *A comprehensive grammar of the English language*. Pearson Education India. 56

Ramesh, Akshai, Venkatesh Balavadhani Parthasarathy, Rejwanul Haque, and Andy Way. 2020. An error-based investigation of statistical and neural machine translation performance on hindi-to-tamil and english-to-tamil. Association for Computational Linguistics (ACL). 11

Rebala, Gopinath, Ajay Ravi, and Sanjay Churiwala. 2019a. *An Introduction to Machine Learning*, 1st edition. Springer Publishing Company, Incorporated.

Rebala, Gopinath, Ajay Ravi, and Sanjay Churiwala. 2019b. *An Introduction to Machine Learning*.

Roffo, Giorgio. 2017. Ranking to learn and learning to rank: On the role of ranking in pattern recognition applications. *arXiv preprint arXiv:1706.05933*. 17

Saini, Sandeep and Vineet Sahula. 2015. A survey of machine translation techniques and systems for indian languages. In *2015 IEEE International Conference on Computational Intelligence & Communication Technology*, pages 676–681. IEEE. viii, 4, 6

Sanders, Gregory, Mark Przybocki, Nitin Madnani, and Matthew Snover. 2011. Human subjective judgments. *Handbook of Natural Language Processing and Machine Translation*, pages 750–759. 30, 31, 32

Sanyal, Sugata and Rajdeep Borgohain. 2013. Machine translation systems in india. *arXiv preprint arXiv:1304.7728*. 8

Sinha, R Mahesh K. 2004. An engineering perspective of machine translation: anglabharti-ii and anubharti-ii architectures. In *Proceedings of international symposium on machine translation, NLP and translation support system (iSTRANS-2004)*, pages 10–17. 8

Sinhal, RA and MB Chandak. 2012. Divergence: A challenge in example based machine translation. In *Proceedings of the International Conference on Soft Computing for Problem Solving (SocProS 2011) December 20-22, 2011*, pages 805–812. Springer. viii, 8

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231. 10, 32

Specia, Lucia, Najeh Hajlaoui, Catalina Hallett, and Wilker Aziz. 2011. Predicting machine translation adequacy. In *Proceedings of Machine Translation Summit XIII: Papers*, Xiamen, China. 10

Stasimioti, Maria and Vilelmini Sosoni. 2020. Translation vs post-editing of nmt output: Insights from the english-greek language pair. In *Proceedings of 1st Workshop on Post-Editing in Modern-Day Translation*, pages 109–124. 10

Su, Keh-Yih, Ming-Wen Wu, and Jing-Shin Chang. 1992a. A new quantitative quality measure for machine translation systems. In *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*. 10

Su, Keh-Yih, Ming-Wen Wu, and Jing-Shin Chang. 1992b. A new quantitative quality measure for machine translation systems. In *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*.

Sutskever, Ilya, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27. 20

Tan, Zhixing, Shuo Wang, Zonghan Yang, Gang Chen, Xuancheng Huang, Maosong Sun, and Yang Liu. 2020. Neural machine translation: A review of methods, resources, and tools. *AI Open*, 1:5–21. 19

TDIL. 2014. Machine translation evaluation. pages 1–44. 9

Tripathi, Sneha and Juran Krishna Sarkhel. 2010. Approaches to machine translation. 4

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Vijayanand, Kommaluri, S.I. Choudhury, and P. Ratna. 2002. Vaasaanubaada: automatic machine translation of bilingual bengali-assamese news texts. *Language Engineering Conference, 2002. Proceedings*, pages 183–188. 8

Wang, Haifeng, Hua Wu, Zhongjun He, Liang Huang, and Kenneth Ward Church. 2021. Progress in machine translation. *Engineering*. 2

Weaver, Warren. 1952. Translation. In *Proceedings of the Conference on Mechanical Translation*. 2

White, John S. 1995. Approaches to black box mt evaluation. In *Proceedings of Machine Translation Summit V*. 10, 30, 31, 41

White, John S and Theresa A O'Connell. 1994. Evaluation in the arpa machine translation program: 1993 methodology. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*. 31, 42

Yang, Shuoheng, Yuxin Wang, and Xiaowen Chu. 2020. A survey of deep learning techniques for neural machine translation. *arXiv preprint arXiv:2002.07526*. 2, 19, 20, 21

Acronyms

ALPAC Automatic Language Processing Advisory Committee. 9

AMFM Adequacy-Fluency Metrics. 11

ARPA Advanced Research Project Agency. 9

ASR Automated Speech Recognition. 34

ATEC Assessment of Text Essential Characteristics. 10

BLEU BiLingual Evaluation Understudy. vi, 2, 10, 37, 54, 55

CBMT Corpus Based Machine Translation. iv, 6

CDAC Center for Development of Advanced Computing. 5

CNN Convolutional Neural Network. 14, 19

DEJ Directly Expressed Judgment. 30

DNN Deep Neural Network. 18

EBMT Example Based Machine Translation. iv, 7

HMT Hybrid Machine Translation. iv, 9

IBM International Business Machines. 2

IIIT-H Indian Institute of Information Technology. 12

KNN K-Nearest Neighbour. 15

METEOR Metric for Evaluation of Translation with Explicit Ordering. vi, 10, 35

ML Machine Learning. 14

MQM Multidimensional Quality Metrics. 33

MT Machine Translation. 1, 12

NE Named Entity. 11

NIST National Institute of Standards and Technology. 10

NMT Neural Machine Translation. iv, v, 1, 9, 14, 17–21

NON-DEJ Non-Directly Expressed Judgment. 30

PBSMT phrase Based statistical Machine Translation. 11

RBMT Rule Based Machine Translation. iv, 2, 3

RIBES Rank-Based Intuitive Bilingual Evaluation Score. 10

RNN Reccurent Neural Network. 14

SL Source Language. 3

SMT Statistical Machine Translation. 2

SOV Subject-Object-Verb. 56

SVO Subject-Verb-Object. 56

TDIL Technology Development for Indian Languages. 12

TER Translation Error Rate. vi, 10, 35

WER Word Error Rate. vi, 10, 34

WMT Workshop on Machine Translation. 11

भारतीय भाषा संस्थान
शिक्षा मंत्रालय, भारत सरकार,
मानसगंगोत्री, मैसूरु

CENTRAL INSTITUTE OF INDIAN LANGUAGES
Ministry of Education, Government of India
Manasagangothri, Mysuru – 570006

ICOLSI-43/57/2021

&

LINGUISTIC SOCIETY OF INDIA
Deccan College Post Graduate & Research Institute, Pune.

लिंग्विस्टिक सोसाइटी ऑफ इंडिया
डेक्कन कॉलेज स्नातकोत्तर एवं शोध संस्थान, पुणे

Certificate

This is to certify that Mr/Ms/Dr/Prof **Danaveni Madhukar** has participated in ***43rd International Conference of the Linguistic Society of India*** online hosted by the Central Institute of Indian Languages, Mysuru from 21-23 December 2021. The title of his/her presentation is

Evaluating English-Telugu Machine Translation Output

(Sujoy Sarkar)
Coordinator, ICOLSI-43

(G. Umamaheshwara Rao)
President, LSI

(Shailendra Mohan)
Director, CIIL

Evaluation and Error Analysis of English-Telugu Neural Machine Translation Output

by Danaveni Madhukar

Submission date: 29-Dec-2022 05:04PM (UTC+0530)
Submission ID: 1987275519
File name: madhu_draft.pdf (3.92M)
Word count: 20939
Character count: 111488

Evaluation and Error Analysis of English-Telugu Neural Machine Translation Output

ORIGINALITY REPORT

4%	2%	3%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	"Translation Quality Assessment", Springer Science and Business Media LLC, 2018 Publication	<1%
2	Submitted to University of Colombo Student Paper	<1%
3	icon2021.nits.ac.in Internet Source	<1%
4	www.tdx.cat Internet Source	<1%
5	aclanthology.org Internet Source	<1%
6	Eirini Chatzikoumi. "How to evaluate machine translation: A review of automated and human metrics", Natural Language Engineering, 2019 Publication	<1%
7	C. K. Quah. "Translation and Technology", Springer Science and Business Media LLC, 2006	<1%

	Publication	
8	"Natural Language Processing and Chinese Computing", Springer Science and Business Media LLC, 2018 Publication	<1 %
9	"Data Management, Analytics and Innovation", Springer Science and Business Media LLC, 2023 Publication	<1 %
10	etheses.whiterose.ac.uk Internet Source	<1 %
11	"Natural Language Processing and Chinese Computing", Springer Science and Business Media LLC, 2019 Publication	<1 %
12	Lecture Notes in Computer Science, 2015. Publication	<1 %
13	tesisenred.net Internet Source	<1 %
14	Submitted to International American University Student Paper	<1 %
15	Submitted to Associatie K.U.Leuven Student Paper	<1 %
16	Maheen Akhter, Sahar Noor, Muhammad Ramzan, Hikmat Ullah. "Evaluating Urdu to	<1 %

Arabic Machine Translation Tools", International Journal of Advanced Computer Science and Applications, 2017
Publication

17	Nizar Habash, Joseph Olive, Caitlin Christianson, John McCary. "Chapter 2 Machine Translation from Text", Springer Science and Business Media LLC, 2011 Publication	<1 %
18	"Recent Advances in Example-Based Machine Translation", Springer Science and Business Media LLC, 2003 Publication	<1 %
19	Submitted to Florida Polytechnic University Student Paper	<1 %
20	Sahinur Rahman Laskar, Abinash Gogoi, Samudranil Dutta, Prottay Kumar Adhikary et al. "Investigation of negation effect for English–Assamese machine translation", Sādhanā, 2022 Publication	<1 %
21	archive-ouverte.unige.ch Internet Source	<1 %
22	"Natural Language Processing and Chinese Computing", Springer Science and Business Media LLC, 2020 Publication	<1 %

23	www.um.edu.mt Internet Source	<1 %
24	"Computational Linguistics", Springer Science and Business Media LLC, 2013 Publication	<1 %
25	"Intelligent Systems and Applications", Springer Science and Business Media LLC, 2021 Publication	<1 %
26	Lecture Notes in Computer Science, 2005. Publication	<1 %
27	dokumen.pub Internet Source	<1 %
28	www.yumpu.com Internet Source	<1 %
29	"Explorations in Empirical Translation Process Research", Springer Science and Business Media LLC, 2021 Publication	<1 %
30	Bonnie Dorr, Joseph Olive, John McCary, Caitlin Christianson. "Chapter 5 Machine Translation Evaluation and Optimization", Springer Science and Business Media LLC, 2011 Publication	<1 %

31	Irene Rivera-Trigueros, María-Dolores Olvera-Lobo, Juncal Gutiérrez-Artacho. "chapter 60 Overview of Machine Translation Development", IGI Global, 2021 Publication	<1 %
32	Joke Daems, Sonia Vandepitte, Robert J. Hartsuiker, Lieve Macken. "Identifying the Machine Translation Error Types with the Greatest Impact on Post-editing Effort", Frontiers in Psychology, 2017 Publication	<1 %
33	Paolo Lorusso. "Lexical Parametrization and Early Subjects in L1 Italian", International Journal of Linguistics, 2018 Publication	<1 %
34	Amit Khaparde, Akshat Kumar, Mohit Kumar, Suyash Gupta. "Multi-lingual code-mixed machine translation system", AIP Publishing, 2022 Publication	<1 %
35	Maja Popović, Hermann Ney. "Towards Automatic Error Analysis of Machine Translation Output", Computational Linguistics, 2011 Publication	<1 %

Exclude quotes	On	Exclude matches	< 14 words
Exclude bibliography	On		