Evaluation of English-Hindi Machine Translation Output

A thesis submitted in 2023 to the University of Hyderabad
in partial fulfilment of the award of

MASTER OF PHILOSOPHY

in

Applied Linguistics

by

ADITI AGARWAL

20HAHL01

at

Centre for Applied Linguistics and Translation Studies

School of Humanities

University of Hyderabad

हैदराबाद विश्वविद्यालय
University of Hyderabad

प्रतिष्ठित संस्थान
INSTITUTION OF EMINENCE
राष्ट्रीय अपेक्षाएँ, वैश्विक मानक
National Needs, Global Standards

March-2023

CERTIFICATE

This is to certify that the dissertation entitled "**Evaluation of English-Hindi Machine Translation Output**" submitted by **Ms. Aditi Agarwal** bearing Reg. No. **20HAHL01** I partial fulfilment of the requirements for the award of the Master of Philosophy in Applied Linguistics is a bonafide work carried out by her under my supervision and guidance.

The thesis has not been submitted previously in part or in full to this or any other University or Institution for the award of any degree or diploma.

Head of Department/Centre

Signature of the Supervisor

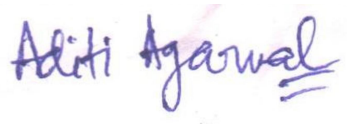
DECLARATION

I, **Aditi Agarwal**, hereby Declare that this Dissertation entitled **"Evaluation of English-Hindi Machine Translation Output"** submitted by me under the guidance and supervision of Professor S. Arulmozi is a bonafide research work. I also declare that it has not been submitted previously in part or in full to this University or any other University or Institution for the award of any degree or diploma.

I also consent to upload this dissertation on the INFLIBNET- Shodhganga repository.

Date: 25.03.2023

Name: Aditi Agarwal



Registration Number: 20HAHL01

Acknowledgements

विद्यां ददाति विनयं,
विनयाद् याति पात्रताम् ।
पात्रत्वात् धनमाप्नोति,
धनात् धर्मं ततः सुखम् ॥

I would like to thank my supervisor **Pr. Arulmozi** for his guidance and supervision in this research undertaking. I would like to thank both- my supervisor and my RAC member Dr. K. Parameshwari for having faith in this research topic, even at times when I had lost confidence. I am grateful for the guidance I have received from them.

I would especially like to thank my supervisor for teaching me about research itself more than anything else. He made me aware of the things that truly matter as a researcher. A supportive guide contributes not only to the research but also to the well-being of the researcher. I would also thank him for instilling in the confidence that I can take an undertaking such as conducting research on my own.

Next, I would like to thank my department CALTS- the professors and the office staff for the institutional support. I thank Mr. P V S R Murthy, Mr. D Mallesh, and Ms. Swathi for answering my queries and processing the paperwork. Their work, which is often invisible, is of great importance to a research scholar.

Next, I would like to thank my language consultants- Ms. Mimansa Sharma, Mr. Paul Marandi, and especially **Mr. Rahul Singh**. Chancing upon a language consultant like Rahul is nothing short of a miracle in the course of this research. In a world where many promises are made and seldom kept, Rahul's proved his to the belong to the minuscule set of promises fulfilled. Just when I had almost

given up and was brimming with frustration and on the verge of pathos and melancholy, Rahul made his angelic presence known and fulfilled his promise. Initially, Rahul appeared to be just another being making the same promise that had remained unfulfilled many a times- the Herculean task of rating my sentences *pro bono.* But he actually did it! Not just that, he did it at an a speed which was so unimaginable that I couldn't have conjured the thought of it even in my dreams! This incident restored my faith in the kindness of humanity and that there still is hope out there. It also encouraged me towards making new friends.

Thank you everybody for rating my sentences as it instilled me with hope with each sentence rated. It is an act of true selflessness and awesomeness that you have done. YOU ARE AWESOME!!!!!!!! I am grateful to the core and humbled by your participation and patience. Not to mention, I am in an awe of you for doing this mind-boggling task for me. If this is not friendship then I don't know what is! The world needs more people like you in it. You didn't have to do it, it was no obligation of yours to carry out this work, you did it despite knowing that there is no official credit for this work. It's of no 'use' to you. Yet you did it anyway, and thus, I acknowledge you over here as I feel indebted.

Next, I would like to thank the C canteen for providing me with food and company during the night owl hours. This was really important for my health both mental and physical. I have had some of the most cherished moments of my university life there. *Viva la* C canteen!

I would like to thank all of my friends and well-wishers for the support and morale boost you gave me from time to time. Thank you for the advice, the listening ear, the consolation and many other features of camaraderie that contributed to the formation of my dissertation. I am not mentioning any names as there are too many of you who have supported me and made an indirect contribution to my dissertation. The list will be too long and any omission out of human error will be unfair. I would although like to mention the name of my friend Dr. Male Shiva Ram Reddy for his direct contribution in my dissertation in the form of moral support and academic advice during the writing

phase. I would especially like to thank Ms. Sumukhi Marupaka for her kindness, generosity, and vehicular support. Life at Hyderabad wouldn't have been possible without your contribution. From my first friend in HCU you have come a long way in giving me wings unconditionally.

Finally, I would like to thank my family- my siblings, parents, grandparents, and relatives and community for valuing my education. I would especially like to acknowledge my parents' sacrifice of sleep- the biggest in my opinion. I am extremely grateful to my parents for waking up every school-day from an alarm just to send me and my siblings to school for **14 years** despite their sleep issues. I am not even capable of imagining such an arduous sacrifice. I thank my grandmothers for their wisdom and concern. I thank Arjun for providing comic relief which is a much needed stress buster during research journey. I thank Aayushi for everything, words cannot describe what she means to me.

Contents

Chapter 1

Introduction

Evaluation of Machine Translation is an important and integral part in the field of computational linguistics. Machine translation evaluation is of two types- Automatic and Human. Of the two, human evaluation of machine translation is considered as the gold standard. This is so because the methodology of human MT evaluation is the closest to the desired objective of achieving outputs in the MT systems matching the translations of a human translator. Even though human evaluation is the gold standard, there is a dearth of studies conducted in the domain of human evaluation of machine translation. This is because the human translation evaluation methodology is a resource intensive one.

Resources such as time and human language consultants who are proficient in language and technology are required to achieve the task, and thus ultimately it is more costly to carry out the resource-intensive task of MT

evaluation by humans. The present study has aimed and achieved to do this resource-intensive task using human evaluators.

1.1 Machine Translation

Machine translation refers to using computers to translate language data from a source language to a target language. Machine translators have come a long way since their inception, yet there is a longer way ahead for the current MTs to be anywhere near human translation. Though not comparable to human translators, the MTs have found their use in the day-to-day translation of the mundane (yet necessary) items of language such as road signs, menus, cooking instructions, or song lyrics. These are instances of information access. Another domain where the MT finds its use is computer-aided translation or CAT. Draft translations produced by the MT are post-edited by human translators to save their time and energy. CAT often finds its use in localization. Current MTs work on encoder-decoder models.

There are three main types of Machine Translation (MT) methods:

1.1.1 Rule-based Systems

These systems rely on numerous linguistic rules based on morphology, syntax, semantics, etc. The rule-based MTs include transfer-based machine translation, Interlingua based machine translation, and dictionary-based machine translation. The advantage of these types of MTs is that they give good

quality domain specific output. The disadvantage of rule-based systems is that the amount of manual work required for the upkeep of MT and the requirement of domain specific knowledge is high.

The Vauquois Triangle is a representation of how the rule-based machine translation model works which uses Interlingua- an artificial language which is an intermediate template of all languages in a particular machine translation system. This triangle represents the process of turning the source text to target text in three different steps. The left side of the triangle characterizes the source language; the right side the target language. The idea is that the word level is the shallowest level where a direct transfer from SL to TL happens. But, as we move further to the more complicated processes of syntactic and semantic transfer, the Interlingua comes into play.

The interlingua-based MT analyses the SL sentence into a language-independent (interlingual) representation of its meaning and then generates the output sentence by converting the meaning representation into the TL.

The interlingual approach of machine translation is not useful in Indian context because our languages tend to have a huge amount of variation and an approach like that of Vauquois triangle is useful only for language cognates.

1.1.2 Statistical Systems

The statistical MTs use parallel bilingual text corpora to learn and generate translations from. It uses statistics to do the same. The advantage is that

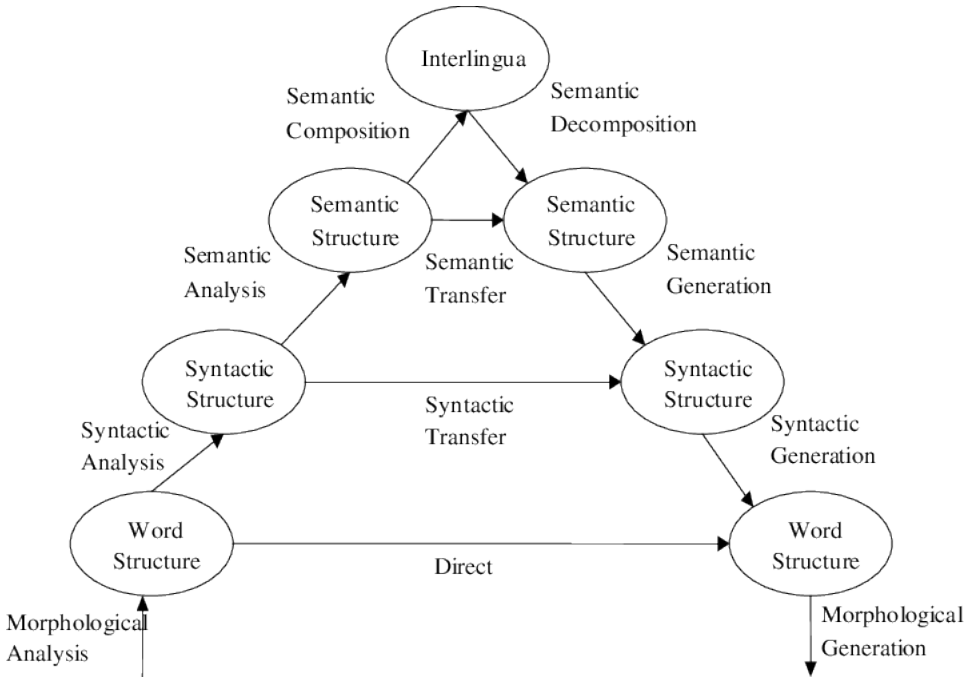


Figure 1.1: **The Vauquois Triangle, Vauquois (1968)**

with large corpora the model can be trained well to generate good quality outputs and the frequency of corpus used improves the results. The disadvantages of statistical systems are that long chains are difficult to observe in corpora, the long sentences might have zero probability. The accuracy is difficult to improve and specific errors are difficult to fix. This type of MT is also unsuitable for languages with differences in word order.

1.1.3 Neural Systems

Artificial neural networks are used to obtain machine translation. It works on an encoder-decoder with an attention model. It requires big data. The advantages include that it can be built with one network instead of a pipeline of many tasks. It uses a fraction of memory used by the statistical MT. The disadvantages include that the speed of training the models is slow in this type of MT; rare words are difficult to translate because of the lack of training.

1.1.4 Hybrid Systems

The hybrid type is a combination of two or more of the above types. These kind of systems are in wide use today as more and more machine translation systems are shifting from the statistical model to the neural model.

1.2 English to Hindi Machine Translation

1.2.1 Differences and Similarities

The first basic difference in English and Hindi is that while English is an SVO language, Hindi is an SOV language. This basic difference in word order often causes chaos in the machine translation output.

While English and Hindi both belong to the Indo-European language family, English is from the Germanic and Hindi is from the Indo-Iranian division. Both the languages have stemmed from the Proto-Indo-European language. This is also one of the reasons that the two languages share a few similarities. The differences are of interest for the purpose of the current study.

1.2.2 Challenges in Human Translation

These same differences pose a challenge in human translation, but these can be overcome by a well-versed translator. The main challenge in human translation comes from words or concepts which do not have an equivalent in the other language. This has been one of the most challenging problem known to mankind in the field of translation.

1.2.3 Challenges in Machine Translation

When it comes to machine translation, the challenges increase manifold. First, it is necessary to familiarise the algorithm with the linguistic differences in the two languages to and from which the translation has to be carried out. These linguistic differences along with the challenges faced by human translators add up to make machine translation the formidable task that it is.

1.2.4 Challenges in Translation in Education Domain

When it comes to education domain, or any other domain specific translation for that matter, the first challenge in translation is the lexical equivalent of the jargon of that language. The jargon might also be used colloquially and the meaning in the colloquial usage and domain usage might differ. It might also be the case that the same word at an etic level might be present in two different domains and have different meanings at an emic level, Pike (1967). This is something a state-of-the-art machine translation system needs to infer from the context of the whole text.

Another problem with jargon is that the Hindi expression of a jargon might be absent or too obsolete that it is not in use by the language speakers practically. In such a case, a human translator uses her discretion to either opt for the English word written in Devanagari, or use the Hindi equivalent expression appropriate in the context.

Sometimes an equivalent technical term might not even exist in Hindi for its English counterparts. In such cases, it is optimum for a machine translation system to use the English term but in Devanagari and not Roman script.

1.3 Machine Translation Evaluation

Translations can be evaluated according to the factors of adequacy and fluency. Adequacy is also known as faithfulness or fidelity. It measures how well the translation captures the exact meaning of the original sentence. Fluency measures how fluent the translation is in its target language along the lines of grammaticality, readability, clarity, and essence.

1.3.1 Computer Evaluation

The most popular automatic metric for machine translation is called BLEU (for BiLingual Evaluation Understudy). It is based on n-gram precision, i.e., the number of words matching in the machine translation and human translation. It is based on the premise that a good machine translation will tend to contain words and phrases that occur in a human translation of the same sentence.

1.3.2 Human Evaluation

Human Evaluation is still considered as the Gold Standard in the domain of MT evaluation even though many advanced techniques have come up for automatic MT evaluation. This is so because translation in itself is a tricky task to achieve even by human translators as there are many aspects to translation-cultural, linguistic, and otherwise- that languages encapsulate in themselves. The translation of these is a feat to achieve.

Even human translators vary in their approach to translation and in the words they use to translate the same body of text. The domain of translation is still widely studied and researched upon. All this is only an indication of the complexity of translation and how Herculean a task translation is for a machine to perform.

Due to this, the human evaluation for any kind of evaluation of machine translation is the best and most intuitive bet to make. The native language speakers possess all unwritten and dynamic rules of a living language within their mind and thus, are an authority on any kind of translation work or sentence judgement in their first language.

Studies of human evaluation are not taken up widely very often as they are resource-intensive- requiring native speakers' time and expertise. In some tasks of machine translation evaluation, the knowledge of both source and target language is required on the language consultant's part. This might be quite difficult to achieve for certain language pairs due to either scarcity of speakers possessing the required proficiency in both the languages or be-

cause of the high market value of skills of certain language pair speakers. Constraints such as these are obstacles in achieving the desired results if resources are lacking, which is the case more-often-than-not.

Human evaluation relies on the language instinct of a native speaker and has various metrics based on how the research study is designed. A native speaker is defined as any person who acquired the said language as a child as an L1 or first language. First language may not necessarily be just one language, a child can acquire more than one language in their critical period of language acquisition.

1.4 Education Domain

1.4.1 New Education Policy

The New Education Policy emphasises on the implementation of mother tongue as the medium of instruction till class 5th, Kalyani (2020). Hindi is the most widely spoken mother tongue in India, according to the census data, census.india.gov.in (2011). NEP also emphasises on instruction in Hindi and exams to be conducted in English and Hindi for higher Education Institutes (HEIs). Most of the study material is in English and the availability of study material in regional languages is a challenging task because of the resources involved in obtaining so.

1.5 The Research Question

My research is aimed at finding out the degree to which a student accessing study material can depend on machine translators for accessing the said material in her desired language. The target language will tend to be the language of the vernacular medium school that the student has attended. The language in concern for this study is Hindi.

My research further delves into observing the patterns and scope for improvement in the said machine translators.

Since the translation domain is targeted at students, it becomes essential to select the state-of-the-art MTs that are free to use. A paywall is not accessible for a common student.

1.5.1 Objectives

• My research aims at evaluating the machine translation output of English-Hindi and Hindi- English machine translation systems at the educational level to understand the error pattern at different sentence levels and complexities.

• This research will highlight the systematic patterns of error in the algorithm to deduce where it needs to improve. It will also serve as a precedence for any future attempts at machine translation and the upcoming systems can account for these structural errors in their systems.

1.5.2 Research Question

• To understand the underlying linguistic discrepancies- syntactic, semantic, etc. present in the current algorithm of the state-of-the-art MT systems using study material at educational level.

• To find out if the current MT scenario is ready to be put to any practical use or not.

• To find a resolution for the errors present in the system.

Chapter 2

Review of Literature

2.1 Machine Translation

Jurafsky and Martin (2009) throw light at machine translation as a whole and at the encoder-decoder models in detail in the chapter "Machine Translation and Encoder-Decoder Models" of their book "Speech and Language Processing". The BLEU metric for MT evaluation is also discussed in detail in the chapter. The chapter also discusses the linguistic component pertaining to typology and language divergences. Phenomena like word order, lexical divergences, morphological typology, referential density, pro-drop etc. have been talked about in the chapter.

The encoder-decoder models are the standard algorithm for machine translation. They are state of the art algorithms for not only machine learning but also for many other tasks where complex mappings between two

sequences are involved like summarization, dialogue, semantic mapping, etc. Encoder-decoder models with RNNs and Transformers have been discussed in detail. The training of such models has been explained. Attention mechanism has also been explained. Beam search, a decoder algorithm, has been explained. Apart from the abovementioned, some practical phenomena and ethical issues have been talked about in the chapter.

The book Speech and Language Processing is meant to be a textbook for computer science students. Thus, the linguistics component is limited to translation errors that might occur during machine translation and just the surface has been touched upon, whereas the mathematical component and the algorithmic component has been described in detail.

2.2 Machine Translation Evaluation

2.2.1 Automatic Evaluation of Machine Translation

Miller and Beebe-Center (1956) lay the foundation for the systematic procedure to evaluate machine translation in their paper titled "Some psychological methods for evaluating the quality of translations". The method of BLEU (BiLingual Evaluation Understudy) which is based on the premise that a good machine translation will tend to contain words and phrases that occur in a human translation of the same sentence. The assessment of machine translation on the basis of the n-gram precision it has with respect to human translation has been evolved from the experiment illustrated in this

seminal work.

Callison-Burch et al. (2006) argue that the sole criterion of BLEU scores in not a good judge of the quality of translation performed by the MT. It shows experimentally that an MT with a lower BLEU score might be superior in quality and hence, the total reliability on BLEU scores is argued against.

The paper proposes the appropriate uses for Bleu to be: tracking broad, incremental changes to a single system; comparing systems which employ similar translation strategies (such as comparing phrase-based statistical machine translation systems with other phrase-based statistical machine translation systems); and using Bleu as an objective function to optimize the values of parameters such as feature weights in log linear translation models, until a better metric has been proposed. Inappropriate uses for Bleu include comparing systems which employ radically different strategies (especially comparing phrase-based statistical machine translation systems against systems that do not employ similar n-gram-based approaches); trying to detect improvements for aspects of translation that are not modeled well by Bleu; and monitoring improvements that occur infrequently within a test corpus.

Implementing BLEU requires standardizing on many details of smoothing and tokenization. Post (2018) in their paper is recommending to use standard implementations like SACREBLEU rather than trying to implement BLEU from scratch. This is because BLEU isn't a single metric but requires a number of different parameters; preprocessing schemes have an impact on scores; and papers vary in hidden parameters and schemes which they might

not necessarily report.

2.2.2 Corpus-based Machine Translation

ParaCrawl: Web-scale acquisition of parallel corpora. ACL. The main motivation behind creating this corpus was improvement in machine translation systems. The corpus is named ParaCrawl and it is the largest language corpora for some language pairs that is publicly available. It is a parallel corpus for European languages, and is heavily dominated (73 percent) by 5 languages: French, German, Spanish, Italian and Portuguese. This corpus employs web crawling open-source software as one of the main methods of data extraction. The corpus is meant to be a training model for future MTs, especially for those belonging to low resource languages, Bañón et al. (2020).

OpenSubtitles is an open-source parallel corpus derived from movie and television subtitles. It comprises 1689 bitexts spanning 2.6 billion sentences comprising 17.2 billion tokens from 60 languages including Indian languages such as Bangla and Hindi. This is a versatile range of data which will include slangs as well as formal language from documentaries and everything in between spanning across genres. Lison and Tiedemann (2016) explain the process of cross-linguistic alignment of the data in detail in the paper.

The United Nations Parallel Corpus is the official corpus composed from the United Nations documents. It consists of manually translated UN documents from 1990 to 2014 spanning 25 years for the six official UN languages: Arabic, Chinese, English, French, Russian, and Spanish. The corpus consists

of pair-wise aligned documents and a fully-aligned six way subcorpus for the six languages. It is under a liberal license and can be downloaded free of cost. The corpus is a Moses-based statistical machine translation system. Baseline BLEU scores for the same have been provided in the paper, Ziemski et al. (2016).

The Europarl corpus is composed of the official proceedings of the European Parliament crawled from the web. On its initial release it comprised 11 official languages of the EU. 10 new languages from the new EU members have been added to the corpus on its later releases. The latest release of this corpus was in 2012 consisting of 21 European languages. This corpus was designed to be used as training data for Statistical Machine Translation. It has around 60 million words per language, Koehn (2005).

2.3 Human Evaluation of Machine Translation

Licht et al. (2022) propose a new scoring metric Cross Lingual Semantic Textual Similarity(XSTS) along the lines of Semantic Text Similarity(STS) Metric. This metric focuses on adequacy rather than fluency. They also introduced a calibration metric for discerning the inter-annotator evaluation. They compared each evaluator's score with the human translations and normalised the score of each evaluator if they were scoring higher or lower than the average score.

Freitag et al. (2021) propose a "Platinum Standard" (better than the already established Gold Standard of the human evaluation, in their definition) for the evaluation of machine translation. However, this so-called "Platinum Standard" has still not been adopted for studies in the discipline of machine translation evaluation.

Chapter 3

Research Methodology

3.1 Corpus Building

The sentences fed to the respective systems are Indian English sentences. The decision to use the Indian English input has been made keeping in mind that the users requiring an EN-HIN translation are more likely to encounter such type of English rather than say British or American English. And thus, it will be more useful to evaluate the MT system on the kind of data for which it will be most likely used.

As the objective of the research is to find out the usefulness and scope for improvement of machine translators with special focus on the domain of Education, I have chosen the lecture transcript of the course titled "Applied Linguistics" from the NPTEL website to use as the corpus for the study.

3.1.1 About the Source

NPTEL stands for National Programme on Technology Enhanced Learning. It is an initiative funded by the Ministry of Education (erstwhile Ministry of Human Resource and Development), Government of India. Along with the premiere institutes of the country which includes seven IITs and the IISc, Banglore, the Ministry of Education aims at bridging the learning gap through this initiative.

NPTEL

NPTEL is the world's largest repository of courses in Engineering, basic sciences, and some humanities and management subjects. It also has a Youtube channel which is the most subscribed educational channel with 1.3 billion views and more than four million subscribers.

It has more than 56,000 hours of video content. The important thing is that all this content is transcribed and subtitled. Of this content, 12,000 hours of English content is translated in regional Indian languages. Transcripts of one such course form the contents of my corpus.

Translation of the English transcripts of NPTEL video courses are being carried out in 11 regional languages. The goal is to help the students coming from local language schooling understand NPTEL course content better.

MoE AND MOOC

Massive Open Online Courses (MOOC) is essentially an asynchronous teaching-learning platform, where the process involves use of pre-recorded lectures, resource video materials, lecture notes, assignments and quizzes, as content and self assessment at regular intervals. The learning, through scheduling of fixed time duration for completion of courses and, therefore, the simultaneous participation of teachers and a large number of students may be termed synchronous and is thus similar to a classroom, albeit on the Internet and being much larger in size. When offered with consideration for students in non-urban and rural areas through supplementary DVDs and mobile delivered content, they enable quality and equitable access to a much larger population of students and can lead to a significant rise in the Gross Enrollment Ratio. These courses are open for anyone to access – free of cost. So anyone who is interested in learning gets access to quality content, which also includes discussion with the course faculty and access to assignments for self testing. The faculty who are currently offering courses are from the IITs or from other reputed institutes such as CMI, IMSc etc.

3.1.2 The Rationale

The final data after running the initial pilot study has been taken from the NPTEL lecture transcripts. This corpus exclusively compiled for this research keeping in mind the research issues that this study aims to explore

and answer. The NPTEL lecture titled "Applied Linguistics" is chosen for the same.

This corpus is made of exclusively Indian English sentences and the content of the corpus is also India-centric as it comprises of the lectures given in the Indian Institute of Technology-Madras. The rationale behind the selection of such data is that it is well suited for a student looking for a Hindi translation of English course material.

Since it is an Introductory course, the language consultant need not have a specialisation in the subject domain and thus is able to evaluate such material based on just the knowledge of her/his mother tongue, i.e., Hindi and not be limited because of the lack of subject knowledge. In other words, only the knowledge of Hindi is required to rate this data of translated sentences from the original source language- English and the consultant need not have prior knowledge in the subject matter.

For example, if Physics would have been the content of the data instead of Applied Linguistics, a language speaker of Hindi would show apprehension in rating such sentences which are Hindi translations if they do not have a background in Physics at the Bachelor degree level. Moreover, finding such language consultants who have proficiency in subject matter in both the languages is a difficult task to achieve, especially given limited resources for the conduct of this research.

This is so because the subject matter is heavily loaded with technical jargon in such subjects in both the languages. Although, one might be well

acquainted with such jargon in one language, they might have difficulties in another language. This makes the probability of finding a Hindi speaker proficient in the technical jargon in both English and Hindi and willing to rate one thousand sentences threefold becomes very slim.

(I have mentioned one thousand sentences being rated threefold because that is the modus operandi of the present study. Three language consultants have rated the same one thousand sentences each for the three different translation systems, i.e., Google, Bing, and Yandex.)

On the other hand, with respect to a subject such as Linguistics, this is a subject which is not taught at school level and only basic introduction is given at the graduate level. The course material which forms the data for this present study is that of post graduate level. So, at the post-graduate level, the subject is taught from the very basic and builds up to advanced levels at par with other subjects.

The question of the study being confined to just the introductory level might arise. This is not the case. As I have demonstrated above, the sentences in our data are both introductory and technical in nature as the subject matter builds up.

The sentences mainly comprise of introducing the jargon which serves two purposes

1. Evaluating the description and examples of a concept, and
2. Introducing the subject jargon.

Since the subject jargon is being introduced and the concepts are also be-

ing explained, it serves as an example set of all types of educational material and not confined to a specific subject.

The jargon which is used in introductory and slightly advanced level is the same in essence. The only difference is that the advanced level material tends to be heavily loaded in jargon per sentence compared to the beginner level material which tends to be on the not-so-heavy side. This means that if an MT system is able to process a particular jargon in beginner level, it is also likely to be able to process it at an advanced level, because the jargon is a word after all. And if an MT system is making mistakes in processing jargon at a beginner level material, it is bound to make similar or more serious mistakes when the sentences become more complex.

There is a scope of study for the above mentioned problem but it is beyond the scope of the current research as we shall find out in the Data and Analysis chapter of this dissertation.

3.2 Evaluation

Three native speakers of Hindi have given the rating to thousand sentences each of English-Hindi machine translation on a grade of 0-4 where each rating corresponds to:

4-Perfect translation; both adequate and fluent

3-Good translation with minor error; adequate and somewhat fluent

2-Understandable translation with major error; somewhat adequate and

somewhat fluent

1-Bad translation; inadequate and somewhat fluent

0-Gibberish; neither adequate nor fluent

The highest score that a sentence can get is 4. Let S be the total number of sentences, then,

Highest Score N = S*4

To calculate adequacy/comprehensibility, only the sentences which are rated 2, 3, and 4 will be considered.

To calculate fluency, only the sentences which are rated 3 and 4 will be considered.

Both will be calculated as:

Fluency=

$$\sum_{i=3}^{4} Si/N$$

Adequacy=

$$\sum_{i=2}^{4} Si/N$$

Since it is a comparative study, three different machine translators were tested on their translation of the same sentence. All the sentences marked below 4 are also marked for the discrepancies in the translation; i.e., whatever linguistic component that the translation seemed to be lacking in are marked. This evaluation can exclusively be done by a trained linguist who is the native speaker of the target language. The overall missing component from the neural network of the MT will thus be highlighted. These shortcomings

can then be rectified to create better MT systems.

The three machine translation systems chosen for the study are Google, Microsoft Bing, and CDAC. The comparison will also give an insight into the functioning and differences in the programming structures of the three.

3.3 Evaluators

Three language consultants have rated the whole data (1000 different sentences each across the three different MT systems). Their details are as follows:

Rahul Singh(M) is pursuing Master of Performance Arts in Theatre Arts degree from the Department of Theatre Arts of the Sarojini Naidu School of Arts and Communication, University of Hyderabad. He grew up in Patna, Bihar.

Mimansa Sharma(F) was pursuing Master of Philosophy degree in History from the Department of History of the School of Social Sciences, University of Hyderabad at the time of the evaluation of sentences. She grew up in Jammu, Jammu and Kashmir.

Paul Marandi(M) was pursuing Master of Arts in German Language, Literature and Culture Studies from the Centre for German Studies of the School of Language, Literature and Culture Studies, Jawaharlal Nehru University at the time of the evaluation of sentences. He grew up in Noida, Uttar Pradesh (part if the National Capital Territory of Delhi).

The three language consultants' first language is Hindi and all three are proficient in English. This was discerned by asking them the question: "What is your first language?". If a prospective language consultant replied "Hindi" only then were they taken into consideration for the evaluation else they were ineligible. All three language consultants are proficient in English. This is discerned by self-assessment and also taking into consideration that they are pursuing higher education in the top-ranked central universities of India where the medium of instruction is English and they have to write long assignments, exams, etc. in English.

It should be noted that there is diversity among the language consultants in the matter of their geographical location and their sex. All three hail from three different states/union territories - Bihar, Jammu and Kashmir, and Nationl Capital Territory of Delhi. This diversity is essential so that no single variety of the Hindi language prevails in this study, else the bias will be high. The difference in sex is also important so that the study is not male-biased as many researches tend to be. The language consultants have provided a rating and not the translation for the sentences. They have contributed pro-bono for the advancement of artificial intelligence, for the benefit of fellow Hindi speakers, and most importantly for the love of their native language.

It is to be noted that all the three language consultants belong to the capital cities of their respective home-states. This is an important factor because capital cities tend to use Hindi in official as well as home domain,

the interiors tend to use Hindi for only official purposes and the local language for informal domains. So, the language consultants belonging to capital cities is a feature rather than a shortcoming.

Chapter 4

Data and Analysis

4.1 Pilot Study

A pilot survey at a smaller scale was conducted to determine which machine translation systems should be used for the final study of MT evaluation. The metric was decided on 75 percent of Adequacy. The MT systems with values lesser than that of 75 percent adequacy were not taken into consideration in the final study as it was found too inadequate for any practical usage of the said system.

100 Indian English sentences have been translated via four prominent machine translation systems. In a spreadsheet the following columns have been made adjacent to each other: the original sentences, the output Hindi translations, the Rating of the translation which the language consultant has given to the machine translated sentence, correction (if the error is minor),

error (explaining the error in detail), error type (to classify the error), and comments (regarding the MT model and suggestions for the same).

Of these seven columns, the first one is the sentence from the Indian English corpus in the domain of Education, the second column is the machine output when the sentence was fed into the MT system, the rating and comments have been given by the language consultant who doesn't have a background in Linguistics, the rest of the columns are filled in by me based on my linguistic training.

Observations: Of the four machine translation systems taken into account, Yandex and Google have fared well in the pilot study, Bing has been moderate with, while TDIL has been incompetent at the translation of the sample sentences. The MT systems ranked from least errors to most are Yandex, Google, Bing, and TDIL with 10, 13, 21, and 84 errors out of 96 sentences respectively.

This gives their respective adequacy to be:

Yandex: 89.33,

Google: 86.46,

Bing: 78.13, and

TDIL: 12.5.

With 12.5 percent adequacy, the TDIL leaves a massive scope for improvement.

I shall discuss the errors of the MT systems in detail in the following sections. It is notable that most of the errors were repetitive intra-MT system

and also inter-MT systems. This repetition of errors forms a pattern which I am to discern, as a linguist, for further improvement in the quality and accuracy of the MT systems.

TDIL: The TDIL MT system has produced the greatest number of unintelligible outputs to a layman's eye. It was observed that the outputs in fact, follow a pattern. The major problems lay in

1. Word-to-word translation instead of word sense disambiguation, and

2. Syntax: the word order was ignored multiple times by the MT while producing the outputs which sometimes resulted in understandable yet not fluent sentences. And the other times, resulted in totally unintelligible sentences. Agreement errors were also observed relating to number, gender, and honorifics. PRO errors with pseudo-AGENT have been observed.

3. Pragmatic: The sense of the words was wrongly interpreted, resulting in absurd dictionary translations which gave way to semantically nonsensical sentences. Many words were interpreted out-of-context.

4. Other errors observed were: spelling mistakes, postpositional errors, and prescriptive grammar errors (using obsolete words).

5. The above mentioned errors either occurred singly or in combination further deteriorating the quality of the output.

Bing: The two major errors observed with the Bing MT system are:

1. Pragmatic: out-of-context translation; wrong choice of words (either obsolete or lacking the context specific nuance); and

2. Semantic: word-by-word translations have been given in many cases

where the whole sentence should have been processed first to give the output. This has also resulted in garden path errors.

Google:

1. Voice: Active voice has been used in the output instead of passive voice, which would have been preferable. Four errors out of 13 have been of this kind for the given sample size of 96 sentences. Making up about one-third proportion of the errors.

2. Second major error class is that of prescriptive grammar. Examples like 'kshay rog' instead of simple 'teebee' for the translation of tuberculosis reflects the prescriptive nature of the translation as a common person would readily identify tuberculosis by its English acronym TB rather than from its Sanskritized version of 'kshay rog'.

3. Other types of errors relate to honorifics, semantics, and pragmatics. Wrong agreement has caused the honorific errors. The semantic and pragmatic errors largely deal with wrong interpretation and translation of word isolates in a context.

4. Only one sentence in this sample size has been wrongly interpreted syntactically. Whether this error is repeated in a pattern or whether it is a standalone error will be determined by a larger sample size.

Yandex: The major shortcoming of the Yandex MT is its lack of interpretation of a long sentence with a punctuation mark such as a hyphen, which resulted in the majority of the errors. Though these errors seemed to be of different types, they followed a similar pattern. For this the Yandex

MT needs to follow a combination of both the top-down and the bottom-up approach while parsing the sentence for translation.

Similarities: The most frequently occurring error across the MTs seems to be the sentence parsing which leads to haphazard interpretations and outputs. This requires the training of neural networks to identify garden path sentences. Other similarities include: agreement, current vocabulary, and pragmatic word sense disambiguation.

4.2 Error Analysis

3,011 sentences were evaluated by three language consultants for the present study. The language consultants were tasked with the evaluation of 1000, 1000, and 1011 sentences each across three different MT systems for the research amounting to a total of 3011 sentences. The following is the table of frequencies of the cumulative ratings of all three language consultants across different MT systems.

The above tables show the frequency of the sentence ratings for the three MT systems, based on which we can calculate the parameters of adequacy and fluency of each MT system and make a comparison.

First I will demonstrate how different MT systems fare on the parameters of adequacy and fluency and then make the comparison and talk about linguistic issues.

Google

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	25	.8	.8	.8
	1	359	11.9	11.9	12.8
	2	1041	34.6	34.6	47.3
	3	985	32.7	32.7	80.0
	4	601	20.0	20.0	100.0
	Total	3011	100.0	100.0	

Figure 4.1: Google

Bing

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	28	.9	.9	.9
	1	354	11.8	11.8	12.7
	2	1354	45.0	45.0	57.7
	3	711	23.6	23.6	81.3
	4	564	18.7	18.7	100.0
	Total	3011	100.0	100.0	

Figure 4.2: Bing

4.2.1 Adequacy

Adequacy is measured as

$$Adequacy = \sum_{i=2}^{4} Si/N$$

Yandex

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	62	2.1	2.1	2.1
	1	297	9.9	9.9	11.9
	2	1288	42.8	42.8	54.7
	3	829	27.5	27.5	82.2
	4	535	17.8	17.8	100.0
	Total	3011	100.0	100.0	

Figure 4.3: Yandex

Google:

For the MT System Google, all the sentences that were rated higher than 1 were used to calculate the adequacy score. There were 1041 sentences that were given the rating of 2 by the language consultants, 985 sentences were given the rating of 3, and a total of 601 sentences were given a perfect rating of 4 by the three language consultants. So, the formula for calculation is as follows:

$$Adequacy_{Google} = \sum_{i=2}^{4} S[1041 + 985 + 601]/3011$$

This can be further simplified as

$$Adequacy_{Google} = 2627/3011$$

A total of 2,627 sentences were deemed adequate by the language consul-

tants which leaves the total adequacy to be

$$Adequacy_{Google} = 0.8724676187313$$

The above figure can be rounded off to three digits as 0.872 as the adequacy for the Google machine translation system.

Bing:

For the MT System Bing, all the sentences that were rated higher than 1 were used to calculate the adequacy score. There were 1354 sentences that were given the rating of 2 by the language consultants, 711 sentences were given the rating of 3, and a total of 564 sentences were given a perfect rating of 4 by the three language consultants. So, the formula for calculation is as follows:

$$Adequacy_{Bing} = \sum_{i=2}^{4} S[1354 + 711 + 564]/3011$$

This can be further simplified as

$$Adequacy_{Bing} = 2629/3011$$

A total of 2,629 sentences were deemed adequate by the language consultants which leaves the total adequacy to be

$$Adequacy_{Bing} = 0.8731318498837$$

The above figure can be rounded off to three digits as 0.873 as the adequacy for the Bing machine translation system.

Yandex:

For the MT System Yandex, all the sentences that were rated higher than 1 were used to calculate the adequacy score. There were 1288 sentences that were given the rating of 2 by the language consultants, 829 sentences were given the rating of 3, and a total of 535 sentences were given a perfect rating of 4 by the three language consultants. So, the formula for calculation is as follows:

$$Adequacy_{Yandex} = \sum_{i=2}^{4} S[1288 + 829 + 535]/3011$$

This can be further simplified as

$$Adequacy_{Yandex} = 2652/3011$$

A total of 2,652 sentences were deemed adequate by the language consultants which leaves the total adequacy to be

$$Adequacy_{Yandex} = 0.8807705081368$$

The above figure can be rounded off to three digits as 0.880 as the adequacy for the Yandex machine translation system.

4.2.2 Fluency

Fluency is measured as

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

Google:

For the MT System Google, all the sentences that were rated higher than 2 were used to calculate the fluency score. There were 985 sentences that were given the rating of 3 and a total of 601 sentences were given a perfect rating of 4 by the three language consultants. So, the formula for calculation is as follows:

$$Fluency_{Google} = \sum_{i=3}^{4} S[985 + 601]/3011$$

This can be further simplified as

$$Fluency_{Google} = 1566/3011$$

A total of 1,566 sentences were deemed fluent by the language consultants which leaves the total fluency to be

$$Fluency_{Google} = 0.5200929923613$$

The above figure can be rounded off to three digits as 0.520 as the fluency

for the Google machine translation system.

Bing:

For the MT System Bing, all the sentences that were rated higher than 2 were used to calculate the fluency score. There were 711 sentences that were given the rating of 3 and a total of 564 sentences were given a perfect rating of 4 by the three language consultants. So, the formula for calculation is as follows:

$$Fluency_{Bing} = \sum_{i=3}^{4} S[711 + 564]/3011$$

This can be further simplified as

$$Fluency_{Bing} = 1275/3011$$

A total of 1,275 sentences were deemed fluent by the language consultants which leaves the total fluency to be

$$Fluency_{Bing} = 0.4234473596811$$

The above figure can be rounded off to three digits as 0.423 as the fluency for the Bing machine translation system.

Yandex:

For the MT System Yandex, all the sentences that were rated higher than 2 were used to calculate the fluency score. There were 829 sentences that were given the rating of 3 and a total of 535 sentences were given a perfect rating of 4 by the three language consultants. So, the formula for calculation is as follows:

$$Fluency_{Yandex} = \sum_{i=3}^{4} S[829 + 535]/3011$$

This can be further simplified as

$$Fluency_{Yandex} = 1364/3011$$

A total of 1,364 sentences were deemed fluent by the language consultants which leaves the total fluency to be

$$Fluency_{Yandex} = 0.4530056459647$$

The above figure can be rounded off to three digits as 0.453 as the fluency for the Yandex machine translation system.

4.3 Comparison

The calculation of the Adequacy and Fluency scores leaves us with the following data:

Adequacy

Google: 0.872

Bing: 0.873

Yandex: 0.880

Fluency

Google: 0.520

Bing: 0.423

Yandex: 0.453

As it can be observed from the above data, the Google and Bing machine translation systems are comparable when it comes to the criterion of Adequacy while the Yandex system slightly outperforms them on the same criterion. Whereas, when it comes to the criterion of Fluency, the Google ML system outperforms the other two by several percentage points (Yandex by 6.7 percentage points and Bing by 9.7 percentage points when rounded off to one decimal point).

Even though the Google MT system is outperforming the others, its performance on the criterion of fluency is not impressive at all. Rather, it leaves room for improvement. It can also be noted that Bing is lagging behind overall on both the criteria despite performing slightly better than Google on the criterion of Adequacy. But, since the difference was minuscule, i.e., of two sentences out of a total of 3,011 this difference is negligible and therefore be ignored.

This will be more apparent from the visual representation of the data in the bar graph format.

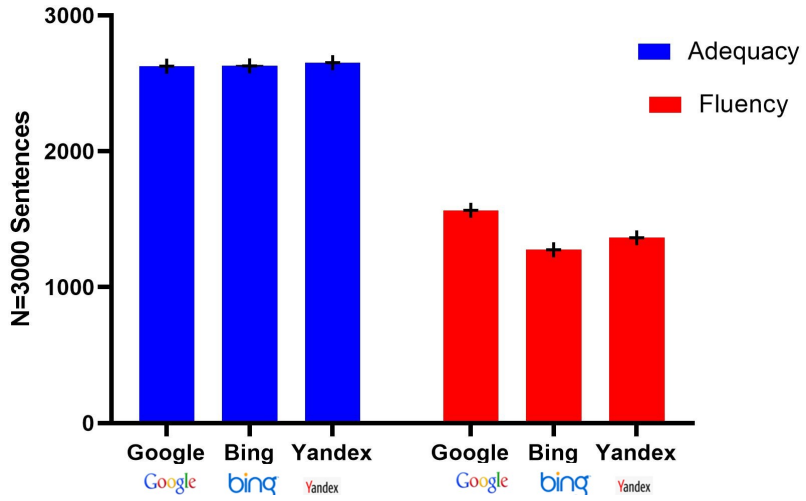


Figure 4.4: Bar Graph Representation of the Results

From the above graph it is apparent that all the three machine translation systems are nearly equal when it comes to the metric of Adequacy but Google performs slightly better and Bing slightly worse on the metric of Fluency. Although all three systems are poor performers on the metric of Fluency.

4.4 Consolidation of All Linguistic Issues and Discussion

4.4.1 Roman Script

The output across all three machine translation systems containing Roman script are marked below the perfect score of 4. These Roman script output amount to either gibberish or not translated at all.

Another pattern observable is that the sounds in the Roman script are translated as their English alphabet equivalent and not as sounds themselves when they are presented in between a sentence. This pattern is observable in all the three machine translation systems used for this study. For example, the sentence:

'Now “fa” is not in abundance in our language.'

is translated as:

‘अब "एफए" हमारी भाषा में बहुतायत में नहीं है।’

across all three systems.

The characters enclosed inside the quotation marks are read as /e-fə-e/ instead of /fɑ/.

This error was not present in short sentences.

Another sentence,

'And “Gha” is a sound which has both more air and vibration, get this thing.'

has output as follows:

Google:

‘और "जीएचए" एक ध्वनि है जिसमें अधिक हवा और कंपन दोनों हैं, यह बात प्राप्त करें।’

Bing:

‘और "घा" एक ध्वनि है जिसमें अधिक हवा और कंपन दोनों होते हैं, यह चीज प्राप्त करें।’

Yandex:

‘और "Gha" एक आवाज है, जो दोनों के और अधिक हवा और कंपन, इस बात मिलता.’

Google translates "Gha" as [ʤi - ɛtʃ - e], Bing translates it as [$g^{h}ɑ$], and Yandex simply reiterates the text in Roman script as it is. Here, the best translation is provided by Bing, Google does translates it alphabetically which is incorrect and Yandex producing a result in Roman is unacceptable. This is an example of how the same word can be translated in different manners and how improvement is required.

The second clause of this sentence is abysmally translated by all the three machine translation systems, but that is beyond the scope of this subsection and will be discussed in an upcoming subsection - 'Word Sense Disambiguation' of this dissertation.

4.4.2 Word Sense Disambiguation

The word लागू /lɑgu/ apperars 52 times in Bing, 63 times in Google, and 50 times in Yandex as a translation of the word "Applied" in "Applied Linguistics". This kind of lexical preference in translation is unacceptable for a state-of-the-art machine translator.

लागू /lɑgu/ refers to applied in the context such as:

“The rules of grammar are applied here.”

but, it is not acceptable in the context of "Applied Linguistics" where the word अनुप्रयुक्त /ənʊprəjʊkʈ/ is more appropriate.

This is just one example of lexical preference the MT systems have displayed that a human translator would not opt for. A language speaker would also disapprove of such a word choice as happened in the current study. All three language consultants marked the sentences with such word choice 1-point lower than they would have otherwise.

Other examples of abrupt lexical preference include:

- The choice of तालिका /ʈalɪkɑ/ for the word 'table'. तालिका /ʈalɪkɑ/ refers to a mathematical table and the sentence related to the physical object - the furniture table. The appropriate word for the said table is मेज़ [mɛʤ]. This error is present in all three machine translation systems.
- Use of प्रयोगशाला /prəjogʃalɑ/ for the word 'labial'. प्रयोगशाला /prəjogʃalɑ/ refers to a laboratory and not lips. All three systems interpreted labial as lab-ial as in belonging to a laboratory and not belonging to lips. This error is borne out of negligence of the context of the translation and taking into account only the lexical interpretation of the word 'labial'. The correct translation of the word 'labial' here should have been ओष्ठ-संबन्धी /oʃʈʰ-səmbənɖʱi/ instead of प्रयोगशाला /prəjogʃalɑ/.

An example of phrasal word sense disambiguation is the sentence:

"You can think, you can read, one can speculate, one can write something else, but can never be a complete answer to a question like this, get it? "

The output is:

Google:

"आप सोच सकते हैं, आप पढ़ सकते हैं, कोई अनुमान लगा सकता है, कोई कुछ और लिख सकता है, लेकिन इस तरह के एक प्रश्न का पूरा जवाब कभी नहीं हो सकता है, इसे प्राप्त करें?"

Bing:

“आप सोच सकते हैं, आप पढ़ सकते हैं, कोई अटकलें लगा सकता है, कोई कुछ और लिख सकता है, लेकिन इस तरह के सवाल का पूरा जवाब कभी नहीं हो सकता, इसे प्राप्त करें? ”

Yandex:

“आप सोच सकते हैं, आप पढ़ सकते हैं, कोई अटकलें लगा सकता है, कोई कुछ और लिख सकता है, लेकिन इस तरह के सवाल का पूरा जवाब कभी नहीं हो सकता, इसे प्राप्त करें?”

All the three translations are good translations except for the last phrase “इसे प्राप्त करें” [ɪse prɑpt̪ kərẽ]. It literally translates to “get it” as in receiving an object and not as its intended meaning of understanding of something- here a concept in context.

इसे	प्राप्त	करें
ɪse	prɑpt̪	kərẽ
this-ACC	get	do PL/HON

“Get it.”

This is a simple error that can be fixed by either a larger corpus of training data or by manually intervening with the machine translation system to make it context dependent.

4.4.3 Agreement Issue

The gender agreement is not up to the mark in all the three machine translation systems. An example of this is the sentence:

“Professor – student conversation starts”

which is translated as:

"प्रोफेसर - छात्र वार्तालाप शुरू होता है"

across all the three translation systems.

प्रोफेसर	-	छात्र	वार्तालाप	शुरू	होता	है
profesər	-	cʰat̪r	vart̪alap	ʃʊru	hot̪a	hɛ
professor	-	student-MSg	conversation	start	be-PRTCont.M	be-PRT

“Professor – student conversation starts”

Here, the noun gender agreement is violated. The sentence is translated correctly but for the the word [hot̪a] which is the masculine form of the be-verb in Hindi. The correct word would be [hot̪i] which is the feminine form of the be-verb in Hindi. This is because nouns are arbitrarily assigned genders in Hindi and 'conversation' being a noun is assigned the feminine gender.

This mistake is repeated several times displaying that the machine translation systems are biased towards the masculine agreement forms. This occurs because of the skewed data which is biased towards the masculine gender agreement forms.

Chapter 6

Conclusion

It can be concluded that Google performed better of all the three machine translation systems chosen for the final study and Bing's performance was the least impressive. Yandex on the other hand fared slightly better on the criterion of adequacy than the other two but not well when it came to the criterion of fluency.

It is imperative to make a note that the TDIL machine translation system is in too nascent a stage to be actually used for any translation work for any practical purpose.

It is also a fair conclusion to say that none of the state-of-the-art machine translation systems is fit to be used solo for any practical purpose by a student looking for a translation of her course material from English to Hindi language. Although, it may be used as a part of computer aided translation where a translator can feed the source text into an MT system and generate

an output which can further be edited. The machine translation system is then useful in reducing the burden in terms of giving a skeleton output which then has to be extensively corrected. Only about half (Google performed the best giving the fluency of 52 percent) the sentences would be useful and the rest would have to be edited.

This process might be helpful or might lead to extra burden and frustration. This is an issue that can further be explored. But, the conclusion still remains that the machine translation systems are not of any practical use to a common student with limited resources. A student will have to rely on professional translation for studying the course material in entirety.

6.1 Significance

Human evaluation of machine translation is an important and integral part of computational linguistics. It is considered as the gold standard for evaluation of machine translation data. This is an area of computational linguistics where truly the linguistic component outshines the computational component. There is a dearth of studies conducted on human evaluation of machine translation in Indian languages as human evaluation is a resource-intensive task.

All the Indian languages are low-resourced, i.e., they lack sufficient enough corpora to make language processing models. Machine translation is one of the instances of natural language processing. Hindi is the Indian language

with the most developed resources owing to its large number of speakers, which means the NLP tools are most developed for Hindi among the living languages of India. Working on human evaluation of machine translation among the language pairs EN-HI and HI-EN will give us an idea of the progress to be made in NLP for other low-resource languages as well.

• This research will pave the way for future MT systems in avoiding the shortcomings present in the current encoder-decoder model. Any attempt at making an MT system for an Indian language would like to take into account the current errors in the state-of-the-art MT systems and would adjust their algorithm so as to not repeat the same mistakes.

• This research will be helpful in finding out the degree of usability of MT systems for educational purposes. In light of the newly imposed New Education Policy, the biggest challenge so far has been the availability of study material in regional languages.

• Most of the research work done on MT evaluation is carried out by computer scientists who tend to ignore the qualitative aspect of research and focus on obtaining values of adequacy and fluency and compare the systems against one another. A linguistic perspective is essential for the progress of technology as it tells the source of the problem and not just the magnanimity of it.

6.2 Limitations of This Study

This study is largely quantitative although it does look into the qualitative aspect of the machine translation systems, but it is limited. This study can be conducted on a larger scale with intentional sentences to diagnose the exact problems lying within the programming of the machine translation systems in order to eliminate them. Such an extensive study and corpus is beyond the scope of this study.

Another limitation of this study is that it is carried out on machine translation systems designed for general purpose and trained on mainly social media data. This training data varies vastly from the Education domain data that I have tested the systems on. This study is still carried out despite this limitation because free-to-use machine translation systems are the only accessible machine translation systems for a typical Hindi-speaking student.

6.3 Scope for Further Research

The corpus for this study is composed of master level course in Applied Linguistics. Further research can be carried out on corpus composed of other levels of education- primary, secondary, high school, intermediate, graduate, postgraduate, and research levels across different subjects and streams to get a better understanding of the multiplicities that are required for creating a machine translation system that is actually usable.

This study is also limited in its language pair. More language pairs can

be added to the current English-Hindi language pair. This research can also be duplicated for other language pairs including Hindi-English.

This study is limited in its domain. Research can also be carried out across domains to ensure the overall usefulness of the machine translation systems. Domains such as health, literature, entertainment, travel, etc. are some of the examples of the domains that similar research projects can be carried out on.

Bibliography

Bañón, M., Chen, P., Haddow, B., Heafield, K., Hoang, H., Esplà-Gomis, M., Forcada, M. L., Kamran, A., Kirefu, F., Koehn, P., Ortiz Rojas, S., Pla Sempere, L., Ramírez-Sánchez, G., Sarrías, E., Strelec, M., Thompson, B., Waites, W., Wiggins, D., and Zaragoza, J. (2020). ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Callison-Burch, C., Osborne, M., and Koehn, P. (2006). Re-evaluating the role of bleu in machine translation research. In *EACL 2006, 11st Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, April 3-7, 2006, Trento, Italy*, pages 249–256. Association for Computational Linguistics. 11th Conference of the European Chapter of the Association for Compuational Linguistics ; Conference date: 03-04-2006 Through 07-04-2006.

census.india.gov.in (2011). Indian census 2011. Technical report, Registrar

General and Census Commision of India.

Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., and Macherey, W. (2021). Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Jurafsky, D. and Martin, J. H. (2009). *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, Upper Saddle River, N.J.

Kalyani, P. (2020). An empirical study on nep 2020 [national education policy] with special reference to the future of indian education system and its effects on the stakeholders. *JMEIT*, 7:ISSN : 2394–8124.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In Hutchins, J., editor, *The Tenth Machine Translation Summit Proceedings of Conference*, pages 79–86. International Association for Machine Translation.

Licht, D., Gao, C., Lam, J., Guzman, F., Diab, M., and Koehn, P. (2022). Consistent human evaluation of machine translation across language pairs.

Lison, P. and Tiedemann, J. (2016). OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*

(LREC'16), pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Miller, G. A. and Beebe-Center, J. G. (1956). Some psychological methods for evaluating the quality of translations. *Mech. Transl. Comput. Linguistics*, 3:73–80.

Pike, K. L. (1967). *Language in Relation to a Unified Theory of the Structure of Human Behavior*. Mouton.

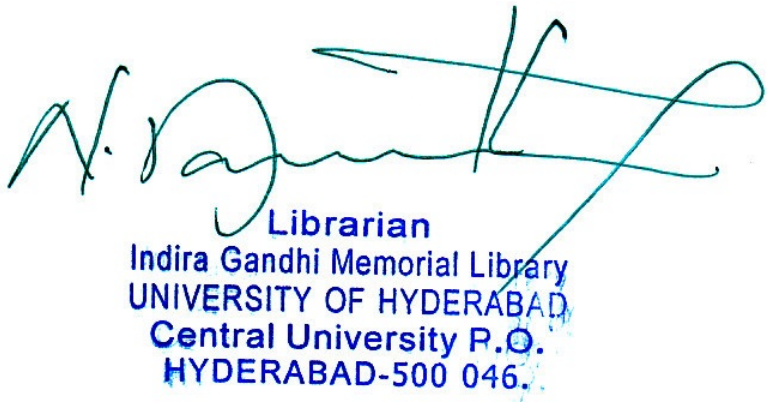
Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Vauquois, B. (1968). A survey of formal grammars and algorithms for recognition and transformation in mechanical translation. In *IFIP Congress*.

Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B. (2016). The united nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*.

Evaluation of English-Hindi Machine Translation Output

by Aditi Agarwal



Submission date: 21-Mar-2023 11:52AM (UTC+0530)
Submission ID: 2042477097
File name: Aditi_Agarwal.pdf (536.7K)
Word count: 9288
Character count: 46491

Evaluation of English-Hindi Machine Translation Output

ORIGINALITY REPORT

7%	6%	3%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	nptel.ac.in Internet Source	2%
2	docplayer.net Internet Source	1%
3	ebin.pub Internet Source	1%
4	Submitted to University of Illinois at Urbana-Champaign Student Paper	<1%
5	riunet.upv.es Internet Source	<1%
6	www.researchgate.net Internet Source	<1%
7	Lecture Notes in Computer Science, 2010. Publication	<1%
8	www.iiits.ac.in Internet Source	<1%
9	Lecture Notes in Computer Science, 2015. Publication	<1%

10	uou.ac.in Internet Source	<1%
11	www.iccs.informatics.ed.ac.uk Internet Source	<1%
12	Sanja Seljan, Nikolina Škof Erdelja, Vlasta Kučiš, Ivan Dunđer, Mirjana Pejić Bach. "chapter 11 Quality Assurance in Computer-Assisted Translation in Business Environments", IGI Global, 2021 Publication	<1%
13	www.textures-archiv.geisteswissenschaften.fu-berlin.de Internet Source	<1%
14	Gupta, Vaishali, Nisheeth Joshi, and Iti Mathur. "Subjective and objective evaluation of English to Urdu Machine translation", 2013 International Conference on Advances in Computing Communications and Informatics (ICACCI), 2013. Publication	<1%
15	mafiadoc.com Internet Source	<1%
16	www.cs.upc.edu Internet Source	<1%
17	www.jnu.ac.in Internet Source	<1%

Exclude quotes	On	Exclude matches	< 14 words
Exclude bibliography	On		