Comparative Analysis of Machine Translation Models for Bangla Language: Evaluating Performance and Potential

Md. Tariqul Islam
dept. of Computer Science and Engineering
Rajshahi University of Engineering and Technology
Rajshahi, Bangladesh
tariquli241@gmail.com

Md Khatami
dept. of Computer Science and Engineering
Rajshahi University of Engineering and Technology
Rajshahi, Bangladesh
khatamionik@gmail.com

Dr. Md. Nazrul Islam Mondal
dept. of Computer Science and Engineering
Rajshahi University of Engineering and Technology
Rajshahi, Bangladesh
mondal@cse.ruet.ac.bd

Nahin Ul Sadad
dept. of Computer Science and Engineering
Rajshahi University of Engineering and Technology
Rajshahi, Bangladesh
nahin@cse.ruet.ac.bd

***Abstract*—The fifth most common language in the world is Bangla. Worldwide, 237 million people speak Bangla. When attempting to connect with members of other groups, this enormous population encounters many challenges. One crucial difficulty that divides them is language. An effective machine translation model is quite useful in removing this barrier. finding a translation model that works well. Four different translation models—Encoder-Decoder, Transformer, GRU, and Marian-based models—are compared in this work. We investigate the performance evaluation of these models using BLEU score and accuracy measures, driven by the necessity to evaluate the effectiveness of various architectures in capturing translation nuances. The comparison of four translation models, namely Encoder-Decoder, Transformer, GRU, and Marian-based models sheds light on their respective performance metrics. The accuracy rates for the first three models stand at 65%, 82%, and 90%, respectively. Correspondingly, their BLEU scores are recorded as 9.61, 11.82, 11.63, and 12.53 respectively. Our comparison reveals that the Transformer and GRU architectures exhibit remarkable proficiency in linguistic processing, outperforming the traditional Encoder-Decoder model and even surpassing the Marian-based model in both accuracy and BLEU score. These results highlight the revolutionary potential of advanced machine translation models such as Transformers and GRU. Our work contributes to a better understanding of these models' capabilities and guides future advancements in machine translation research and applications by offering a thorough examination of their performance.**

***Index Terms*—Transformer, GRU, BLEU Score, Encoder-Decoder, Translation,**

I. Introduction

The global village concept [1] refers to the increasing demand for language translators due to the growing communication between different communities using different languages. However, the number of available translators is insufficient for all translations. This has led to the development of machine translation models, such as the Encoder-decoder-based, transformer-based, and LLM, which have become increasingly popular in professional environments, such as online meetings. The rapid development of communication technology has transformed our world into a global village. However, linguistic diversity poses a significant challenge to effective communication. The demand for language translators has surged, yet available resources are inadequate [2]. This has led to a growing interest in improving machine translation technology [2]. Machine translation, which originated in the 1960s [3], initially struggled due to limited computational capabilities. However, advancements in computing power over the past two decades have propelled the performance of machine translation models. Today, machine translation plays a vital role in various professional settings, facilitating communication in online meetings and beyond. With Bangla being the fifth most spoken language globally [4], its speakers, numbering around 237 million [4], face significant communication barriers when interacting with other communities. Human translators are not always available, and their numbers are insufficient to meet the demand. Therefore, the need for effective machine translation models is paramount. Such models would provide continuous translation services comparable to human translators, thereby easing communication for Bangladeshi people and bridging linguistic gaps.

II. Related Work

Isidora Stevanović and Luka Radičević [5] The authors provided a concise comparative analysis of the two most widely used translation models in their research. Their objective was to pinpoint the key advantages and disadvantages of statistical and neural machine translation, offering a fresh perspective on the field. They explored four primary approaches to machine translation: direct, rule-based, corpus-based, and knowledge-based. Subsequently, they conducted a study on statistical and neural machine translation systems, revealing approximately 200 errors in SMT compared to 150 errors in neural machine translation. This investigation was based on 196 sentences from a well-balanced evaluation set. To round off their theoretical examination of machine translation systems, they evaluated and compared four distinct machine translators: GUAT, Amebis Presis, Microsoft Bing Translator, and Google Translator. Ultimately, they concluded that neural machine translation surpasses its statistical predecessor in sophistication. Depending on the specific translation needs, any of the systems evaluated could be a suitable choice.

Research by Maria Stasimioti Vilelmini, Sosoni Despoina, Mouratidis, and Katia Kermanidis [6] compared three machine translation systems: a customized neural machine translation system, a generic statistical machine translation system, and a tailored-NMT system. The study focuses on the English-to-Greek language pair. They showed that the tailored NMT system surpassed both the statistical machine translation and the neural machine translation systems. The total number of errors in SMT is 699, 550 in NMT, and 414 in tailored NMT systems. They also show that SMT has 40% grammatical errors, NMT has 31%, and tailored NMT has 30% grammatical errors. Regarding the variations between the SMT and NMT outputs, their research demonstrates that the NMT systems provide translations of a better quality, which supports the conclusions of earlier studies on a range of language pairings. Specifically, the study shows that both the tailored NMT and the generic NMT outputs score higher when it comes to human and automatic assessment criteria; the tailored NMT output does even better than the generic NMT output. Furthermore, the tailored NMT output was evaluated better for both adequacy and fluency, making it the top output.

Mentioning the importance of quality estimation of machine translation Sugyeong Eo, Chanjun Park, Hyeonseok Moon, Jaehyung Seo, and Heuiseok Lim [7] published their work, where they undertake comparison tests and analyses using cross-lingual language models (XLMs), multilingual BERT, and XLM-RoBERTa, and show QE utilizing the as-yet-unutilized multilingual BART model. The experiment results allowed us to demonstrate that the XLM-TLM model outperformed the other models on both sub-tasks and that pre-training's induction of language alignment learning had a beneficial effect.

Using morphosyntactic divergence as a lens, Jiaming Luo, Colin Cherry, and George Foster [8] did a fine-grained comparative study of machine translations (MTs) against human translations (HTs). They determine that MT is more conservative than HT, with less morphosyntactic variation, more convergent patterns, and more one-to-one alignments, through assessments on both the aggregate level and the individual pattern level. We also see that for the less common source patterns, MT tends to be less comparable to HT.

Adam Roberts, Noam Shazeer, and Colin Raffel A team of Google researchers led by Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu conducted ground-breaking research [9]. They begin by introducing the T5 model, also referred to as the text-to-text Transfer Transformer model. Their work opens many roads of NLP tasks such as translation, Sentence completion, Paraphrasing, Sentiment analysis, and so on. Their introduced T5 model is pre-trained on large amounts of data. Which later was fine-tuned on an interesting downstream assignment.

III. Methodology

Our work is mainly concerned with the careful comparison and assessment of different translation models. We have invested significant time and energy into developing three separate models, each precisely designed to capture certain language subtleties, in addition to integrating a popular pre-trained model as a baseline for comparison. In order to optimize performance and facilitate the training process, we have carefully selected a dataset that includes a wide 25 variety of language pairings and contextual settings relevant to our research [10]. The dataset was carefully vectorized and subjected to difficult preprocessing before the training phase began. This thorough preparation made sure the data was free of noise and irregularities and converted it into a format that would work well for training our models. Furthermore, the textual information was carefully encoded into numerical representations using the vectorization process, which helped in speedy calculation and improved the models' capacity to identify significant patterns in the data. We wanted to provide a strong basis for our comparative research by using these strict methods, which would help us understand the advantages and disadvantages of each model we were looking at.

IV. Result & Analysis

The differences between English and Bangla offer a complex tapestry of opportunities and prob- lems in language and translation studies. The need for precise and effective translation models grows as globalization speeds up the flow of ideas and information across linguistic bound- aries. Within this framework, the thesis explores the subtleties of several English-to-Bangla translation models intending to conduct a thorough comparative study that clarifies their ad- vantages, disadvantages, and general effectiveness. This chapter thoroughly reviews the study, the methodology used, and the reasoning behind the approaches used. It acts as a doorway to the subsequent examination of data and analysis. By combining theoretical frameworks, empir- ical studies, and computational approaches, this research aims to decipher the nuances present in English to Bangla translation problems. Through a critical examination of several translation models—from neural machine translation architectures to rule-based systems—this study aims to reveal the subtleties involved in both language

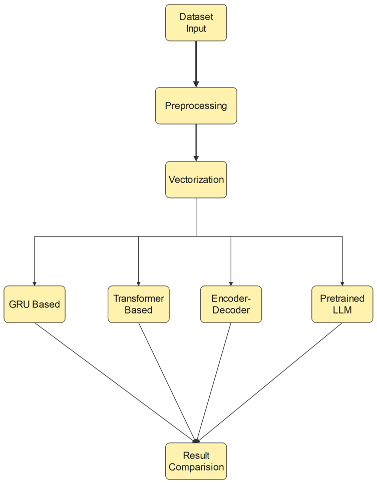


Fig. 1: Methodology

transfer and cultural adaptation. Furthermore, by comparing these models to standards for precision, fluidity, and cultural authenticity, this research hopes to offer insightful information about how machine translation technology is developing. As we set out on this academic adventure, we must recognize the wider ramifications of this research project, both in terms of developing the theoretical discourse in translation studies. As a result, this introduction lays the groundwork for a thorough investigation of English-to-Bangla translation models and a thorough analysis that aims to significantly contribute to academics and industry.

A. Evaluation Metric

It's important to take into consideration several measures that reflect various facets of translation quality when assessing English-to-Bangla translation models. The following list of assessment measures is typical in the field.

- BLEU (Bilingual Evaluation Understudy) BLEU is a widely used measure for comparing candidate translations to reference translations. It compares n-grams between the two, with a score ranging from 0 to 100. A score near 100 indicates better similarity between the reference texts and the applicant's translation. However, a score of 100 is rare and sometimes impractical. Machine translation systems typically use BLEU values between 0 to 30 or 40, with higher scores indicating higher translation quality. BLEU ratings should be read alongside other evaluation metrics and qualitative assessments. Variables such as text difficulty, language pair, and access to excellent reference translations can affect BLEU ratings.

$$P_n = \frac{\sum_{c\varepsilon candidates}\sum_{n-gram\varepsilon c} count_{clip}(n-gram)}{\sum_{c'\varepsilon candidates}\sum_{n-gram\varepsilon c'} count_{clip}(n-gram')} \quad (1)$$

where candidates denote the candidate translation, c denotes each phrase in the candidate translation, and P_n denote the matching degree of the nth order, which is typically the fourth order, Candidates denotes the translation of the candidate, and c denotes each sentence in the candidate translation. Lastly, the BLEU computation technique is:

$$BLEU = BP \bullet exp(\sum_{N}^{n=1} logP_n) \quad (2)$$

$$BP = \left\{ \begin{array}{l} 1, c > r \\ e^{1-\frac{r}{c}}, c \leq r \end{array} \right. \quad (3)$$

where w_n is the weight coefficient, r, and c are the lengths of the reference and candidate translations, and BP is the length penalty factor.

B. Result

Table 1 summarizes the performance of three developed models. From the table, it is shown that GRU based model has higher accuracy. Thought only based on accuracy we can't assess a model. So, next, we compare their BLUE score and their human assessment.

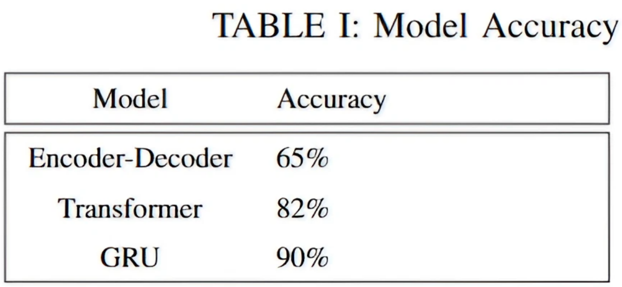


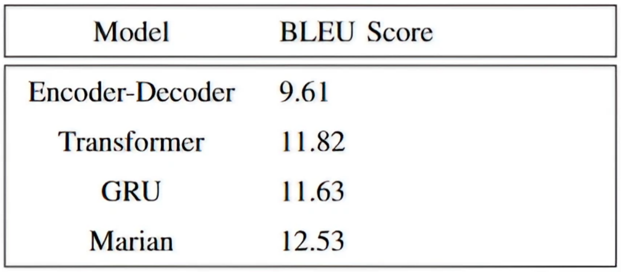
TABLE I: Model Accuracy

Model	Accuracy
Encoder-Decoder	65%
Transformer	82%
GRU	90%

Table 2 summarizes the performance of three developed models and compares them with the pre-trained Marian large language model. From the table, it is shown that the GRU and the transformer-based model have a higher BLEU Score.

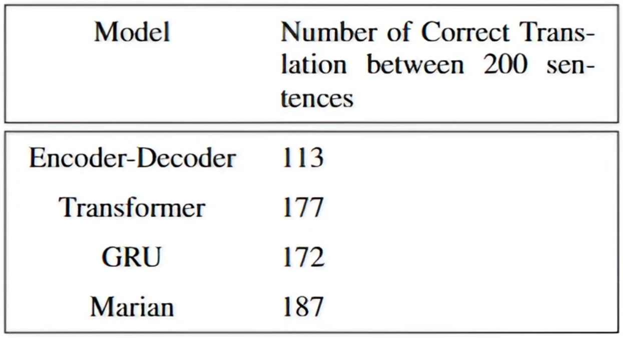
Table 3 summarizes the correctness of three developed models and compares them with the pre-trained Marian large language

TABLE II: Model BLEU Score



Model	BLEU Score
Encoder-Decoder	9.61
Transformer	11.82
GRU	11.63
Marian	12.53

TABLE III: Model Correctness



Model	Number of Correct Translation between 200 sentences
Encoder-Decoder	113
Transformer	177
GRU	172
Marian	187

model. The table shows that the GRU and the transformer-based model translate nearly the same number of sentences as the Marian Large Language model.

C. Analysis

The comparison of four translation models, namely Encoder-Decoder, Transformer, GRU, and Marian-based models, sheds light on their respective performance metrics. The accuracy rates for the first three models stand at 65%, 82%, and 90%, respectively. Correspondingly, their BLEU scores are recorded as 9.61, 11.82, and 11.63, respectively. Interestingly, these figures underscore the Transformer and GRU models' substantial efficacy in capturing translation nuances compared to the Encoder-Decoder model. Notably, the Transformer model stands out with an accuracy of 82% and a commendable BLEU score of 11.82, showcasing its superior capability in linguistic processing. The GRU model, though slightly behind the Transformer, still surpasses the Marian-based model in both accuracy and BLEU score, with an accuracy rate of 90% and a BLEU score of 11.63. This observation is particularly noteworthy, considering that the Marian model is a large-scale language model known for its robustness. The competitive performance of the Transformer and GRU models against the Marian-based model highlights their efficacy in capturing semantic nuances and linguistic intricacies. These findings suggest that while large language models like Marian offer substantial linguistic capabilities, models like Transformer and GRU demonstrate a commendable balance between performance and computational efficiency, making them promising candidates for various translation tasks

V. Conclusion

According to the study, the Transformer model translated Bangla to English with the highest accuracy and BLEU ratings. In terms of translation adequacy and fluency, it fared better than other models. Significance for the Bangla Community: This development might greatly benefit the Bangla-speaking population by reducing obstacles to communication and improving access to worldwide knowledge. Effect on Machine Translation: The Transformer model's performance highlights how machine learning can transform language translation technology by providing more precise and effective solutions. These results support continuing initiatives to enhance machine translation and promote improved interlanguage comprehension. After a thorough comparison of these four models for translation from Bangla to English, the study offers important new information on which model works best for this kind of translation. Through analysis, these models' strengths and weaknesses have been clarified, offering a more complex picture of how well they translate from Bangla to English. While the specifics of the results are limited to the thesis's comparative analysis, it is evident that the study provides a thorough assessment to determine the best model for precise and effective translation from Bangla to English. These findings have significant ramifications for the Bangla-speaking population as well as the larger machine translation technology community. The discovery of a more efficient translation model is a step forward for the Bangla-speaking population in terms of removing language barriers, promoting more seamless communication, and gaining access to information in English, the universal language. This might provide Bangla speakers with greater educational, cultural, and economic possibilities by improving the accessibility and understandability of a wide range of English information.

References

[1] D. J. O'Byrne and A. Hensby, *Globalization: The Global Village*. 1 2011.

[2] "Machine Translation Market Size, share Trends analysis Report by application (Automotive, military Defense, electronics, IT, healthcare, others), by technology, by region, and segment forecasts, 2023 - 2030," 4 2024.

[3] "A (Brief) History of Machine Translation — Smartling."

[4] W. contributors, "List of languages by number of native speakers," 4 2024.

[5] I. Stevanović and L. Radičević, "Comparative analysis of machine translation systems," *ResearchGate*, 11 2020.

[6] M. Stasimioti, V. Sosoni, K. L. Kermanidis, and D. Mouratidis, "Machine Translation Quality: A comparative evaluation of SMT, NMT and tailored-NMT outputs," 11 2020.

[7] S. Eo, C. Park, H. Moon, J. Seo, and H. Lim, "Comparative analysis of current approaches to quality estimation for neural machine translation," *Applied sciences*, vol. 11, p. 6584, 7 2021.

[8] J. Luo, C. Cherry, and G. Foster, "To Diverge or Not to Diverge: A Morphosyntactic Perspective on Machine Translation vs Human Translation," *arXiv (Cornell University)*, 1 2024.

[9] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *arXiv (Cornell University)*, 1 2019.

[10] "Tab-delimited Bilingual Sentence Pairs from the Tatoeba Project (Good for Anki and Similar Flashcard Applications)."