Published by

British Journal of Contemporary Research

BEXFORD PUBLISHING

Machine Translation for the Assamese Language: Prospects, Challenges, and Future Directions

Original Research Article | Volume 1 | Issue 2 | 2026 | Article Number: 053

Accepted: 03 July 2026 | Published: 10 July 2026 | ISSN: 2979-8582 (Online)

DR. BIJOY KRISHNA DOLEY

Department of Assamese S.B.M.S. College, Sualkuchi, Kamrup, Assam, India

DBKD 0000-0002-6749-6034

Correspondence: DR. BIJOY KRISHNA DOLEY, bijoydu87@gmail.com

Abstract

Machine Translation (MT) is the process of automatically converting text from one human language (the source language) into another human language (the target language) using computer software, without requiring direct human intervention. Owing to the rapid advancement of information and communication technologies, machine translation has become an indispensable tool in contemporary society. For resource-scarce languages such as Assamese, MT has emerged as a significant and timely area of research and development. With the expansion of modern technology, a vast proportion of scientific, technical, and academic knowledge is available primarily in English and other global languages. Translating this enormous body of information into Assamese through traditional human translation is both time-consuming and expensive. Consequently, the need for efficient and scalable machine translation systems has become increasingly important. In recent years, Assamese machine translation has witnessed notable progress through platforms such as Google Translate, Microsoft Translator, and the Government of India's Bhashini (National Language Translation Mission) initiative. Although these systems have improved considerably, they are still far from flawless. Compared with traditional Rule-Based Machine Translation (RBMT) and Statistical Machine Translation (SMT) approaches, recent developments in Neural Machine Translation (NMT) and Large Language Models (LLMs) have demonstrated superior translation quality and contextual accuracy. Despite these advancements, the scarcity of high-quality digital linguistic resources and annotated datasets remains the most significant challenge for Assamese machine translation. This paper examines the current status of machine translation for the Assamese language, explores its prospects and potential applications, analyzes the linguistic and technical challenges arising from the grammatical structure of Assamese, and discusses future directions for research and development in this field.

Keywords: Assamese Language, Machine Translation, Neural Machine Translation, Natural Language Processing (NLP), Parallel Corpus

1.0 Introduction

Machine Translation (MT) (Koehn, 2010) is an important branch of Natural Language Processing (NLP) and Artificial Intelligence (AI). It refers to the use of computer software to automatically translate text from one human language (the source language) into another human language (the target language). In the digital age of the twenty-first century, machine translation is no longer merely an experimental field of computer science; rather, it has emerged as an indispensable and powerful tool for global communication, business, education, and knowledge exchange.

Assamese is an independent and rich language recognized in the Eighth Schedule of the Constitution of India. It is not only the official language of Assam but has also historically served as a lingua franca among various ethnic communities and regions of Northeast India. Assamese possess a rich literary heritage and a profound expressive capacity. However, with the rapid expansion of globalization and technological advancement, a vast amount of new knowledge in fields such as medicine, space science, information technology, law, education, and commerce is predominantly available in English and other major global languages.

The traditional process of human translation is often time-consuming, expensive, and labor-intensive, making it difficult to disseminate this growing body of knowledge among Assamese-speaking communities, students, researchers, and the rural population in the digital era. In this context, machine translation has become increasingly important for democratizing knowledge and bridging linguistic barriers. It enables faster access to information and facilitates communication across different language communities.

Over the past decade, rapid developments in computer science and artificial intelligence have significantly advanced machine translation technologies for the Assamese language. Major technology companies such as Google and Microsoft have incorporated Assamese into their translation platforms. Likewise, the National Language Translation Mission (NLTM), popularly known as Bhashini, launched by the Ministry of Electronics and Information Technology (MeitY), Government of India, has provided new momentum to the digital translation ecosystem of Assamese through the development of various language technologies and translation tools. (Ministry of Electronics and Information Technology [MeitY], 2024).

Despite these developments, the quality of machine translation for Assamese has not yet reached the level achieved by widely resourced languages such as English, Spanish, or Mandarin Chinese. Owing to its unique syntactic structure, complex grammatical features, extensive use of suffixes, and the relative scarcity of high-quality digital linguistic resources, machine translation systems often produce inaccurate, unnatural, or contextually inappropriate translations. These limitations present significant technical and linguistic challenges for the development of robust Assamese machine translation systems.

Therefore, the primary objective of this paper is to critically examine the current status of machine translation in Assamese, analyze the major linguistic and technological challenges involved, and explore its future prospects in the context of emerging advancements in artificial intelligence and language technology.

1.1 Objectives of the Study

- To examine the current status and prospects of machine translation for the Assamese language.

- To identify the technical and linguistic challenges associated with Assamese machine translation.
- To explore future directions for the development of machine translation systems in Assamese.

1.2 Methodology of the Study

A mixed-methods approach has been adopted for this study, combining both qualitative and quantitative research techniques. The entire process of study is divided into the following sequence:

1.2.1 Sources of Data Collection

- **Primary Data:** The three most commonly used translation tools today are Google Translate, 'Bhashini' (government AI tool), and 'ChatGPT' (OpenAI). 50 simple sentences used daily, 25 phrases/Jatua dialects and 25 complex government/legal sentences (English) have been used as sources.
- **Secondary Data:** Previous papers on Natural Language Processing (NLP), Computational Linguistics and Machine Translation and reports of the Ministry of Information Technology (MeitY), Government of India have been reviewed.

1.2.2 Evaluation Metrics

Two methods have been applied to measure the accuracy of the translation:

- **Automated Evaluation:** BLEU score (Papineni et al., 2002) used in previous research.
- **Human Evaluation:** Three aspects have been emphasized when checking the quality of machine translations; These are Fluency, Adequacy and Grammatical Correctness.

2.0 Evolution and Technological Context of Machine Translation

The history and evolution of Machine Translation (MT) are closely linked to advances in Computer Science, Linguistics, and Artificial Intelligence (AI). The earliest experiments in machine translation began in the 1950s, shortly after World War II, when researchers attempted to translate Russian texts into English using techniques inspired by cryptography. Since then, machine translation has evolved significantly and has now entered the era of Large Language Models (LLMs). To understand the challenges and prospects of machine translation for a low-resource language such as Assamese, it is necessary to examine its technological evolution through four major stages.

2.1 Rule-Based Machine Translation (RBMT) (Sathish, n.d.)

Rule-Based Machine Translation (RBMT) is the oldest and most traditional approach in the history of machine translation. This method was widely used from the 1960s to the 1980s. In an RBMT system, linguists and language experts manually develop extensive grammatical rules for both the source and target languages, along with large bilingual dictionaries. The machine then follows these predefined rules and dictionaries to translate words and sentences from one language into another.

Limitations: This approach proved highly complex and labor-intensive for the Assamese language. Assamese grammar contains numerous exceptions, and its morphology and syntactic structures are highly nuanced. Developing and maintaining millions of grammatical rules manually was an extremely difficult task. As a result, machine translation for Assamese did not witness significant progress during this stage.

2.2 Statistical Machine Translation (SMT) (Koehn, 2010)

In the late 1990s, with the growth of computer memory and computational power, Statistical Machine Translation (SMT) emerged as a new paradigm. Unlike RBMT, SMT does not rely primarily on manually crafted grammatical rules. Instead, it utilizes a large parallel corpora consisting of texts that have already been translated by humans.

Using probability theory and statistical models, computers analyze these bilingual corpora to determine the most likely translation of a word or phrase in a given context. Google Translate, launched in 2006, initially relied on SMT technology.

Limitations:

SMT requires massive parallel corpora containing millions of aligned sentence pairs to achieve satisfactory performance. During the early years of SMT development, several projects supported by the Government of India and various universities attempted to build Assamese–English translation systems. However, due to the severe scarcity of digital linguistic resources and parallel corpora for Assamese, these systems produced poor-quality translations. The output was often literal, awkward, and grammatically incorrect.

2.3 Neural Machine Translation (NMT) (Koehn, 2020)

A major breakthrough in machine translation occurred in 2016 with the introduction of Neural Machine Translation (NMT). This approach employs artificial neural networks and deep learning techniques inspired by the functioning of the human brain. Rather than translating a sentence word by word, NMT processes the entire sentence as a unified representation, captures its contextual meaning, and generates a more natural translation in the target language.

Limitations: The adoption of NMT has significantly improved the quality of Assamese translations in recent years. Modern translation systems such as Google Translate and Microsoft Translator use NMT-based architectures, enabling them to handle Assamese sentence structures, including its Subject–Object–Verb (SOV) word order, more effectively. However, because Assamese remains a low-resource language with limited digital data, NMT systems still struggle with complex literary, technical, legal, and culturally nuanced texts. Consequently, translation accuracy may decline in such contexts.

2.4 Transformers and the Era of Large Language Models (LLMs) (Google Developers, n.d.)

The most recent advancement in machine translation is the Transformer architecture, introduced in 2017, and the Large Language Models (LLMs) built upon it, such as ChatGPT, Claude, and Google Gemini. Transformer-based models employ a mechanism known as self-attention, which enables them to understand relationships between words across an entire sentence regardless of their distance from one another. This capability allows the model to capture deeper semantic and contextual information more effectively than previous approaches.

Limitations: Modern LLMs have brought a new dimension to Assamese machine translation. They are capable of producing contextually appropriate and stylistically natural translations rather than merely literal ones. Through techniques such as zero-shot and few-shot learning, these models can transfer knowledge from resource-rich languages to Assamese even when limited Assamese training data are available.

Nevertheless, significant challenges remain. One of the major concerns is hallucination, where the model generates information that appears plausible but is factually incorrect. Additionally, LLMs may occasionally produce inaccurate translations, outdated expressions, or culturally inappropriate interpretations. Therefore, human review and post-editing remain essential to ensure the accuracy and reliability of translated content.

3.0 Potential of Machine Translation for the Assamese Language

Machine translation has created numerous opportunities for Assamese society and the Assamese language. Some of the major possibilities are discussed below.

3.1 Democratization of Education and Knowledge

Educational resources in Assamese, particularly in the fields of higher education, science, medicine, and technology, remain limited. However, recent advancements in machine translation have made it possible to translate global educational resources into Assamese instantly. As a result, students studying in Assamese-medium schools and colleges, especially in rural areas, can gain access to global knowledge and learning materials without being disadvantaged by language barriers.

3.2 E-Governance and Administrative Transparency

Machine translation can facilitate the rapid translation of thousands of pages of government notifications, legal documents, official reports, and websites of both the Central and State Governments into Assamese. This can significantly improve communication between the government and citizens, thereby enhancing administrative transparency, public participation, and access to information.

3.3 Language Preservation and Digital Presence

According to UNESCO, languages that are not actively used in digital media face a greater risk of decline or extinction in the future. Machine translation helps ensure that Assamese remains functional and relevant in the digital age by enabling its use on the Internet, social media platforms, and other global digital networks. In this way, machine translation contributes to the preservation, promotion, and continued growth of the Assamese language in the digital world.

4.0 Major Problems and Challenges of Machine Translation in the Assamese Language

The major problems and challenges of machine translation in Assamese are summarized below:

4.1 Resource Scarcity or Lack of Digital Data (Data Scarcity)

Modern Artificial Intelligence (AI) and Neural Machine Translation (NMT) models rely heavily on large volumes of data. Billions of parallel sentences are available on the Internet for languages such as English and Hindi. However, for Assamese, the availability of such a parallel corpus—a collection of accurately aligned source and target language sentences—is extremely limited. This scarcity of high-quality linguistic data significantly affects the performance and accuracy of machine translation systems.

4.2 Linguistic and Grammatical Complexity

Word Order Differences

The sentence structure of English follows the SVO (Subject–Verb–Object) pattern, whereas Assamese generally follows the SOV (Subject–Object–Verb) pattern. As a result, machine translation systems often struggle to place verbs correctly, particularly in long and complex sentences.

Agglutinative Nature of Assamese

Assamese is an agglutinative language in which words are formed by combining roots with various inflections, suffixes, and postpositions. Consequently, a single word may convey a large amount of grammatical information. Machine translation systems often face difficulties in analyzing and interpreting such complex word formations accurately.

Use of Classifiers and Suffixes

Assamese employs a variety of classifiers and suffixes such as -টো (to), -জনী (joni), -খন (khan), and -দল (dal), which have no direct equivalents in English. As a result, machine translation systems frequently fail to capture their semantic and grammatical functions accurately, leading to unnatural or incorrect translations.

4.3 Idioms, Proverbs, and Cultural Differences

Idiomatic expressions, proverbs, and culturally specific phrases pose significant challenges for machine translation. Machines often translate such expressions literally rather than contextually. For example, Assamese idioms may be translated word-for-word, resulting in expressions such as “broken waist” or “tight hand,” which are semantically inappropriate in the target language. Such literal translations fail to convey the intended meaning and cultural nuances of the original text.

5.0 Comparative Observations

Below is a comparative list of tools, technologies used, strengths and weaknesses of machine translation Suppose-

Translation Tool (MT Tool)	Technology used	Strengths	Weaknesses
Google Translate	Neural MT	Fast and extensive use for simple sentences.	makes grammatical errors in complex sentences; Jatua cannot hold the pattern.
Bhashini (Government of India)	AI & Deep Learning	More suitable for government and administrative translations.	Lack of naturalness in the language of ordinary conversation.
ChatGPT/LLMs	Transformer Models	understands a lot of context; Can translate creatively.	Sometimes they use hallucinations or obsolete Assamese words.

6.0 Solutions and Future Directions

Technological advancement alone is not sufficient to make Assamese machine translation as natural, fluent, and grammatically accurate as human translation. Achieving this goal requires a multifaceted effort involving technological, linguistic, social, and economic initiatives. Some important measures are discussed below.

6.1 Building a Large Parallel Corpus

Modern Neural Machine Translation (NMT) models are highly dependent on the availability of large amounts of data. To move Assamese out of the category of a low-resource language, it is necessary to

develop a digital repository containing millions of parallel sentences. Government support is crucial in this regard. Existing bilingual (English–Assamese) government documents, gazettes, notifications, and official records available across various departments of the State and Central Governments should be consolidated into an open digital database.

At the same time, digital repositories can also be expanded through crowd-sourcing initiatives. University students, researchers, and members of the general public can participate in data collection, translation, and validation activities through digital platforms. Programs such as the Government of India's Bhasha Daan initiative under the Bhashini platform provide a suitable model for such collaborative efforts.

6.2 Collaboration between Linguists and Software Engineers

The direct involvement of linguists is essential for teaching Assamese grammar, dialectal variations, and linguistic nuances to computational systems. Most errors in machine translation arise not from technical limitations alone but from linguistic complexities. The deeper meanings, cultural contexts, and grammatical structures of a language cannot be fully captured through computer science principles alone.

Therefore, there is an urgent need to establish joint research centers where linguists and software engineers can work together. Such collaboration would help machine learning systems better understand complex linguistic features of Assamese, including conjunct consonants, compound words, inflectional forms, and classifier suffixes such as -টো (-to), -জন (-jon), and -খন (-khan).

6.3 Development of Hybrid Machine Translation Models

Data-driven neural networks often generate grammatically incorrect sentence structures, whereas rule-based systems generally preserve grammatical accuracy but may produce unnatural or rigid translations. Future research should therefore focus on developing hybrid machine translation models that combine the strengths of both approaches.

In such a system, Neural Machine Translation (NMT) would be responsible for understanding meaning and context, while rule-based components would ensure grammatical correctness and preserve the Subject–Object–Verb (SOV) sentence structure characteristic of the Assamese language.

6.4 Standardization of Digital Orthography and Language Resources

Variations in spelling across Assamese digital resources often create confusion for machine translation systems. For example, inconsistent spellings of the same word in different online sources can negatively affect model training and translation quality.

To address this issue, organizations such as the Asom Sahitya Sabha and other linguistic institutions should work toward establishing a standardized digital orthography for Assamese. Based on such standardization, digital spell-checking and language-validation tools can be developed to ensure consistency and improve the quality of language resources available for machine learning applications.

6.5 Large Language Models (LLMs) and Context-Aware Training

Large Language Models (LLMs) such as ChatGPT, LLaMA, and Claude possess remarkable capabilities in language generation and understanding. However, these models have largely been trained on English-dominant datasets and therefore often lack a deep understanding of Assamese linguistic and cultural contexts.

This limitation can be addressed by fine-tuning LLMs using extensive Assamese-language resources, including folklore, historical texts, literary works, novels, idioms, and proverbs. Such localized training would enable these models to interpret and translate culturally embedded expressions more accurately rather than relying on literal word-for-word translations.

As a result, future machine translation systems will be better equipped to capture the true contextual meaning of Assamese idiomatic and colloquial expressions, thereby producing translations that are both linguistically accurate and culturally appropriate.

7.0 Conclusion

This study demonstrates that machine translation is not merely a technological innovation but also a powerful instrument for the preservation, development, and globalization of the Assamese language in the twenty-first century. The potential of machine translation in Assamese is vast, particularly in the fields of education, administration, e-governance, and knowledge dissemination.

Despite significant progress, several challenges continue to hinder the development of high-quality Assamese machine translation systems. These include the language's complex grammatical structure, rich system of classifiers and suffixes, idiomatic expressions, and, most importantly, the scarcity of digital linguistic resources and parallel corpora.

Recent advances in Neural Machine Translation (NMT) and Large Language Models (LLMs) have substantially improved translation quality. However, these systems still require human supervision and post-editing to ensure accuracy and contextual appropriateness. Therefore, a collaborative effort involving linguists, computer scientists, software engineers, educational institutions, and government agencies is essential for the continued advancement of Assamese machine translation.

The creation of large-scale digital corpora, the standardization of digital linguistic resources, and sustained governmental and institutional support will play a crucial role in strengthening the digital presence of Assamese. With appropriate research, funding, technological innovation, and public participation, machine translation can significantly contribute to establishing Assamese as a vibrant and globally accessible language in the digital age.

References

1. Google Developers. (n.d.). Transformers. https://developers.google.com/machine-learning/crash-course/llm/transformers

2. Hutchins, W. J., & Somers, H. L. (1992). An introduction to machine translation. Academic Press.

3. Jurafsky, D., & Martin, J. H. (2023). Speech and language processing (3rd ed., draft). Prentice Hall.

4. Koehn, P. (2010). Statistical machine translation. Cambridge University Press.

5. Koehn, P. (2020). Neural machine translation. Cambridge University Press.

6. Ministry of Electronics and Information Technology. (2024). National Language Translation Mission (NLTM) - Bhashini project report. https://bhashini.gov.in

7. Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. Proceedings of ACL.

8. Poibeau, T. (2017). Machine translation. MIT Press.

9. Sathish, K. (n.d.). Rule-based machine translation. Medium. https://medium.com/@keerthanasathish/rule-based-machine-translation-7b0074a91d20

Official report

Government of India. Ministry of Electronics and Information Technology (MeitY). National Lan-guage Translation Mission (NLTM) - Bhashini Project Report. 2024, https://bhashini.gov.in. Accessed 16 June 2026.