eISSN: 2229-3213 pISSN: 2229-3205
The Journal of Applied Science and Social Science
Volume 16,Issue 06 June, 2026

DEVELOPMENT OF NEW BERT ALGORITHMS FOR NATIONAL TRANSLATION PROGRAMS BASED ON ARTIFICIAL INTELLIGENCE

Safoyev Nodirjon Nematjon o‘g’li
PhD student, Bukhara state technical university

Abstract. This paper explores the development and adaptation of advanced BERT-based algorithms for national machine translation systems. While transformer-based models have significantly improved the quality of neural machine translation, low-resource languages such as Uzbek still face substantial challenges due to limited parallel corpora, rich morphological structures, and domain diversity. The study analyzes recent improvements in BERT architectures, including multilingual pretraining, domain-specific fine-tuning, subword tokenization, and cross-lingual transfer learning. The paper proposes that optimized BERT-based frameworks can significantly enhance translation accuracy, contextual understanding, and semantic consistency in national translation programs. Experimental insights from recent NLP research indicate that hybrid architectures combining encoder-based BERT models with decoder-based transformer systems provide superior performance in low-resource settings.

Key words: BERT, machine translation, NLP, transformer, low-resource languages, Uzbek language, transfer learning, artificial intelligence, subword tokenization, cross-lingual learning.

Introduction, In the era of rapid digital transformation, artificial intelligence (AI) and natural language processing (NLP) have become fundamental technologies driving innovation in language technologies, particularly in machine translation systems. The increasing demand for multilingual communication in government services, education, international trade, and digital media has made machine translation a strategic priority for many countries developing national translation programs. In this context, transformer-based architectures—especially BERT (Bidirectional Encoder Representations from Transformers) and its multilingual variants—have emerged as powerful tools for improving semantic understanding and contextual accuracy in automated translation systems. Traditional machine translation approaches have undergone a long evolutionary path, beginning with rule-based machine translation (RBMT), followed by statistical machine translation (SMT), and finally neural machine translation (NMT). While RBMT relied heavily on manually constructed linguistic rules and bilingual dictionaries, it lacked flexibility and struggled with ambiguity, idiomatic expressions, and contextual variation. SMT introduced probabilistic modeling based on parallel corpora, significantly improving fluency and alignment quality, yet it still suffered from data sparsity and limited contextual awareness, especially for low-resource languages. The introduction of neural machine translation marked a paradigm shift by enabling end-to-end learning from large datasets. However, early neural models based on recurrent neural networks (RNNs) and long short-term memory (LSTM) architectures faced limitations in capturing long-range dependencies and parallel computation efficiency. This challenge was addressed by the Transformer architecture proposed by Vaswani et al. (2017), which replaced recurrence mechanisms with self-attention, enabling more efficient and context-aware language modeling. This innovation laid the foundation for modern pre-trained language models such as BERT, GPT, and XLM-R. BERT, introduced by Devlin et al. (2018), revolutionized NLP by introducing bidirectional pretraining using masked language modeling (MLM) and next sentence prediction (NSP). Unlike unidirectional models, BERT processes text by considering both left and right contextual information simultaneously, enabling a deeper semantic understanding of language structure. Although BERT was originally designed for language understanding tasks rather than generation, its encoder representations have become a core component in many advanced machine translation frameworks through integration with decoder-based transformer models and transfer learning techniques. Despite these advancements, developing high-quality machine translation systems for national languages remains a complex challenge. Low-resource languages such as Uzbek face several structural and computational limitations, including insufficient parallel corpora, rich morphological systems, and high lexical variability. Uzbek, as an agglutinative language, forms words through extensive use of affixes, resulting in a large number of word variants derived from a single root. This morphological richness complicates tokenization, embedding representation, and alignment in neural models. Furthermore, syntactic flexibility and free word order in

Uzbek introduce additional difficulties for translation systems that rely heavily on fixed positional encoding. Idiomatic expressions, culturally specific phrases, and domain-specific terminology further increase the complexity of achieving accurate semantic translation. As a result, standard pre-trained models often fail to fully capture the linguistic nuances required for high-quality translation in national contexts. Another critical limitation lies in the scarcity of large-scale, high-quality parallel corpora for training and fine-tuning modern neural models. While high-resource languages benefit from extensive datasets, low-resource languages often depend on limited and domain-restricted corpora, which restricts model generalization and robustness. In such conditions, techniques such as transfer learning, multilingual pretraining, data augmentation, and cross-lingual representation learning become essential for improving performance. From a technological perspective, integrating BERT-based models into national translation systems offers promising solutions to these challenges. Through multilingual pretraining and fine-tuning on domain-specific datasets, BERT can be adapted to capture language-specific features while maintaining strong cross-lingual semantic alignment. Additionally, subword tokenization methods such as Byte Pair Encoding (BPE) and SentencePiece help mitigate out-of-vocabulary issues by decomposing words into smaller, more manageable units, which is particularly beneficial for morphologically rich languages. In this context, the development of new BERT-based algorithms tailored for national translation programs is not only a technical necessity but also a strategic requirement for ensuring linguistic inclusivity and digital sovereignty. Such systems can significantly enhance the accessibility of digital content, improve public service delivery, and support multilingual communication in both national and international domains. Therefore, the primary objective of this paper is to analyze modern BERT-based NLP approaches in the context of national machine translation systems, investigate their advantages and limitations, and explore advanced algorithmic improvements that can enhance translation quality for low-resource languages. The study aims to contribute to the development of more accurate, efficient, and linguistically adaptive translation models suitable for real-world national applications.

Literature Review. The development of national machine translation systems has been a central research topic in natural language processing (NLP) for several decades. The literature reveals a clear evolutionary trajectory in machine translation approaches, progressing from rule-based systems (RBMT) to statistical machine translation (SMT), then to neural machine translation (NMT), and finally to transformer-based pre-trained language models such as BERT and its multilingual extensions. Each stage has introduced significant improvements in translation quality, while also revealing new limitations that have motivated further research. Early research in machine translation focused on rule-based approaches, where linguistic knowledge was encoded manually through grammatical rules, syntactic structures, and bilingual dictionaries. These systems were highly interpretable but lacked scalability and flexibility. According to early studies in computational linguistics, RBMT systems struggled particularly with ambiguity resolution, idiomatic expressions, and free word order languages. As a result, their performance was limited in real-world multilingual environments, especially for morphologically rich and low-resource languages. In the early 2000s, statistical machine translation (SMT) became the dominant paradigm. Koehn (2010) formalized SMT as a probabilistic framework that learns translation correspondences from large parallel corpora. SMT models, particularly phrase-based systems, significantly improved fluency and alignment compared to RBMT. However, despite their success, SMT systems suffered from several limitations, including data sparsity, poor handling of long-distance dependencies, and difficulty in capturing deep semantic meaning. These issues became more pronounced when dealing with languages with limited parallel corpora, such as Uzbek. The introduction of neural machine translation (NMT) marked a major paradigm shift. Sutskever et al. (2014) proposed sequence-to-sequence (seq2seq) models using recurrent neural networks (RNNs), enabling end-to-end translation learning without explicit feature engineering. Later, attention mechanisms were introduced to improve alignment between source and target sentences, addressing the bottleneck of long sequence processing. However, RNN-based architectures still faced challenges related to sequential computation and difficulty in modeling very long-range dependencies. A significant breakthrough occurred with the introduction of the Transformer architecture by Vaswani et al. (2017). The self-attention mechanism eliminated the need for recurrence and allowed for parallel processing of input sequences. This innovation dramatically improved training efficiency and translation quality. Transformer models became the foundation for most modern NLP systems, including BERT, GPT, and XLM-R. BERT (Devlin et al., 2018) further advanced

NLP by introducing deep bidirectional pretraining using Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). Unlike earlier models, BERT captures contextual information from both left and right directions simultaneously, enabling a more comprehensive understanding of word meaning in context. Although BERT was initially designed for language understanding tasks such as classification and question answering, its contextual embeddings have been widely adopted in machine translation systems as encoder representations. Subsequent research has extended BERT into multilingual and cross-lingual settings. Multilingual BERT (mBERT) and XLM-R (Conneau et al., 2019) were trained on large-scale multilingual corpora, enabling shared semantic representations across languages. These models demonstrated strong transfer learning capabilities, particularly beneficial for low-resource languages. Studies show that cross-lingual pretraining allows knowledge transfer from high-resource languages (e.g., English, Russian) to structurally similar or even distant languages, improving translation quality without requiring large parallel datasets. In addition to multilingual pretraining, subword tokenization techniques have played a crucial role in improving neural translation systems. Byte Pair Encoding (BPE) introduced by Sennrich et al. (2016) and SentencePiece models address the out-of-vocabulary (OOV) problem by decomposing words into smaller units. This is especially important for agglutinative languages such as Uzbek, where a single root word can generate hundreds of morphological variants through affixation. Subword modeling significantly improves vocabulary coverage and generalization performance. Recent literature also emphasizes the importance of domain adaptation and fine-tuning strategies. Pre-trained language models often require adaptation to specific domains such as legal, medical, or governmental text. Fine-tuning BERT on domain-specific corpora has been shown to improve contextual relevance and translation accuracy. Additionally, research in knowledge distillation and model compression has focused on reducing computational complexity while maintaining performance, enabling deployment in real-time translation systems. Another important direction highlighted in the literature is morphological processing for agglutinative languages. Studies focusing on Turkic languages suggest that morphological segmentation and lemmatization improve alignment quality in neural models. For Uzbek in particular, research indicates that incorporating morphological awareness into tokenization and embedding layers can significantly enhance translation performance. Despite these advancements, the literature consistently identifies several unresolved challenges. These include the scarcity of high-quality parallel corpora, difficulty in handling idiomatic and culturally specific expressions, computational resource limitations, and domain generalization issues. Moreover, most existing pre-trained models are optimized for high-resource languages, making direct application to Uzbek and similar languages suboptimal without additional adaptation. Overall, the literature strongly supports the effectiveness of transformer-based and BERT-based approaches in modern machine translation systems. However, it also highlights the necessity of language-specific adaptations, particularly for low-resource national languages. This has led to growing interest in developing specialized BERT-based algorithms tailored for national translation programs, combining multilingual pretraining, morphological processing, and cross-lingual transfer learning to achieve higher translation accuracy and robustness.

Research Results and Discussion. This section presents the analysis of the performance and practical implications of BERT-based approaches in national machine translation systems, particularly in the context of low-resource languages such as Uzbek. The focus is on evaluating how modern transformer-based architectures improve translation quality, contextual understanding, and semantic consistency compared to traditional and earlier neural approaches.

Research Results. The experimental and analytical results reported in recent NLP studies demonstrate that BERT-based models significantly enhance machine translation performance when properly adapted to the target language. The key findings can be summarized as follows: First, multilingual pre-trained models such as mBERT and XLM-R consistently outperform classical SMT and baseline NMT systems in low-resource settings. This improvement is primarily attributed to cross-lingual representation learning, where semantic knowledge is shared across multiple languages. In Uzbek-related translation tasks, these models show better handling of sentence structure and word sense disambiguation compared to non-pretrained neural systems. Second, fine-tuning pre-trained BERT models on domain-specific corpora leads to measurable improvements in translation accuracy. Domain adaptation experiments indicate that even relatively small Uzbek corpora, when used for fine-tuning, can significantly improve BLEU and semantic similarity scores. This confirms that pre-trained models already contain general linguistic knowledge that can be effectively adapted to national languages with limited data. Third, subword tokenization techniques such as Byte Pair Encoding (BPE) and

SentencePiece have been shown to reduce out-of-vocabulary (OOV) issues and improve morphological coverage. For agglutinative languages like Uzbek, where word forms can vary extensively due to suffixation, subword modeling improves alignment between source and target languages and reduces translation errors related to rare or unseen word forms. Fourth, hybrid architectures combining BERT-based encoders with transformer decoders demonstrate superior performance compared to standalone models. In these systems, BERT is used to generate rich contextual embeddings, while the decoder handles sequence generation. This combination improves both semantic accuracy and fluency of translated output. Fifth, qualitative analysis of translated outputs shows that BERT-enhanced systems perform particularly well in resolving lexical ambiguity and maintaining sentence-level coherence. Compared to baseline models, these systems produce fewer grammatical inconsistencies and better preserve the meaning of idiomatic expressions.

Discussion. The findings suggest that BERT-based architectures represent a significant advancement in national machine translation systems, especially for low-resource languages. However, their effectiveness is strongly dependent on adaptation strategies and data availability. One of the most important observations is that pre-trained multilingual models provide a strong foundation for cross-lingual transfer learning. This is particularly valuable for Uzbek, where large-scale parallel corpora are not readily available. The ability of these models to transfer knowledge from high-resource languages such as English and Russian enables meaningful improvements even with limited training data. Another key issue is the morphological complexity of Uzbek. As an agglutinative language, Uzbek generates a large number of word forms from a single root, which poses challenges for word-level tokenization. The adoption of subword-based approaches significantly mitigates this problem, allowing models to generalize better across different morphological variants. However, current tokenization methods are still not fully optimized for Uzbek-specific morphological rules, suggesting the need for more linguistically informed segmentation techniques. The results also highlight the limitation of BERT as a non-generative model. While BERT excels at contextual understanding, it is not inherently designed for sequence generation. Therefore, its integration into encoder-decoder architectures is essential for practical machine translation applications. This hybrid design improves performance but also increases computational complexity, which may limit real-time deployment in resource-constrained environments. Furthermore, domain adaptation plays a critical role in improving translation quality. General-purpose models often struggle with specialized vocabulary in legal, medical, or technical domains. Fine-tuning on domain-specific datasets significantly reduces these errors, indicating that future national translation systems should incorporate modular training pipelines tailored to different application areas. Another important aspect is the balance between model size and computational efficiency. While large transformer models provide higher accuracy, they require substantial computational resources. Techniques such as knowledge distillation and model compression are therefore essential for deploying BERT-based translation systems in mobile and web applications. Overall, the discussion confirms that BERT-based and transformer-based approaches are highly effective for improving national machine translation systems. However, their success depends on several critical factors, including data quality, linguistic adaptation, computational resources, and architectural design choices. In conclusion, while significant progress has been made, further research is needed to develop Uzbek-specific linguistic resources, optimize morphological processing, and design more efficient lightweight transformer models suitable for real-world national translation platforms.

Conclusion. The analysis of modern artificial intelligence approaches in national machine translation systems demonstrates that BERT-based and transformer-based architectures play a crucial role in improving translation quality, contextual understanding, and semantic accuracy. Unlike traditional rule-based and statistical methods, these models are capable of capturing deep bidirectional contextual relationships within text, which significantly enhances their ability to resolve ambiguity and preserve meaning across languages. The study confirms that BERT, while not inherently designed as a generative model, becomes highly effective when integrated into encoder–decoder frameworks or adapted through transfer learning techniques. Multilingual pre-trained models such as mBERT and XLM-R provide an essential foundation for low-resource languages by enabling cross-lingual knowledge transfer from high-resource languages. This is particularly important for national languages such as Uzbek, where large-scale parallel corpora are limited. Another key finding is that linguistic characteristics, especially the

agglutinative morphology of Uzbek, significantly influence translation performance. Subword tokenization methods such as BPE and SentencePiece, along with morphological segmentation techniques, help mitigate out-of-vocabulary problems and improve model generalization. However, current approaches still require further optimization to fully capture language-specific morphological rules. The research also highlights that domain adaptation and fine-tuning are essential for achieving high-quality translation in specialized fields such as legal, educational, and technical texts. Additionally, computational efficiency remains a critical concern, making model compression and knowledge distillation important directions for practical deployment. Overall, BERT-based NLP technologies represent a strong foundation for the development of advanced national translation systems. Future research should focus on building large-scale Uzbek linguistic corpora, improving morphological modeling, and designing lightweight yet efficient transformer architectures suitable for real-time applications.

References

1. Devlin, J., Chang, M.W., Lee, K., Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*.
2. Vaswani, A. et al. (2017). *Attention Is All You Need*. NeurIPS.
3. Conneau, A. et al. (2019). *Unsupervised Cross-lingual Representation Learning at Scale (XLM-R)*.
4. Sutskever, I. et al. (2014). *Sequence to Sequence Learning with Neural Networks*.
5. Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press.
6. Sennrich, R., Haddow, B., Birch, A. (2016). *Neural Machine Translation of Rare Words with Subword Units*.
7. Goodfellow, I., Bengio, Y., Courville, A. (2016). *Deep Learning*. MIT Press.
8. Sutton, R.S., Barto, A.G. (2018). *Reinforcement Learning: An Introduction*. MIT Press.
9. Johnson, M. et al. (2017). *Google's Multilingual Neural Machine Translation System*.