# Neural Machine Translation: A Comprehensive Overview

Fouzia Afrin Jui, Palash Ahmed, Anjuman Are Ekra
Department of Computer Science & Engineering
Jahangirnagar University
Savar, Dhaka
Bangladesh

## *Abstract*

Neural Machine Translation (NMT) is a subfield of machine learning that has revolutionized the field of machine translation. It involves the use of neural networks to automatically translate text from one language to another. This research paper provides an overview of NMT, including its architecture, training, evaluation metrics, applications, and challenges. We first provide a brief history of NMT and its importance, followed by a literature review of related work. We then describe the NMT architecture in detail, including its components such as the encoder, decoder, attention mechanism, and optimization algorithm. Next, we discuss the training and optimization of NMT, including preprocessing, training techniques, and optimization methods. We also provide an overview of evaluation metrics used for NMT, including their strengths and weaknesses. The paper then delves into the applications of NMT, including translation of languages, speech recognition, and image captioning. Finally, we conclude by discussing the challenges in NMT, such as handling rare words, low-resource languages, and domain-specific knowledge, and suggest directions for future research.

***Keywords:*** Neural Machine Translation, machine learning, neural networks, architecture, training, evaluation metrics, applications, challenges.

## 1 Introduction

Neural Machine Translation (NMT) is a subfield of machine learning that has gained significant attention in recent years. The field of machine translation has a long history, with the first machine translation system developed in the 1950s. Rule-based methods and statistical approaches were the predominant methods used for machine translation before the advent of deep learning(Banitz et al., 2020). However, early machine translation systems were rule-based, which meant that they relied on hand-crafted rules to translate text. These rule-based systems had limited success due to the complexity of natural languages, and the need for human experts to create the rules.

With the advent of machine learning, statistical machine translation (SMT) systems were developed. SMT systems rely on statistical models that learn the translation probabilities from large amounts of bilingual corpora. Although SMT systems outperformed rule-based systems, they still had limitations, such as difficulty in handling rare words and long sentences.

The emergence of neural networks and deep learning has led to the development of a new approach to machine translation, called Neural Machine Translation (NMT). NMT is based on deep neural networks and has shown remarkable results in translating texts between different languages. NMT involves the use of artificial neural networks to automatically translate text from one language to another. NMT has emerged as a promising alternative to SMT. NMT uses deep neural networks to learn the translation probabilities from bilingual corpora. NMT has been shown to outperform SMT in terms of accuracy and efficiency. However, NMT also has its limitations, such as the need for large amounts of training data and the difficulty in handling rare words.

The use of NMT has revolutionized the field of machine translation by providing a more accurate and efficient way of translating text. In this paper, we provide an overview of NMT, including its architecture, training, evaluation metrics, applications, and challenges.

## 2 Literature Review

Su et al. (2018) provided an NMT model for hierarchy-to-sequence attention in this study. They used an encoder model which uses two layers of RNNs to represent the input sentence as a word-clause-sentence hierarchical structure. Using the NIST Chinese-English and English-German translation tasks, they assessed their model. This model significantly outperforms a number of different baselines. However, they did not examine the model's efficacy in document-level NMT in this paper.

Xie et al. (2022) focused on a novel method to improve the translation quality of named entities for NMT in this work. After introducing the notations, the training strategy and network architecture are discussed where they used Transformer as the backbone of their model. They tested the translation of four languages—English, German, Chinese, and Japanese—which are collectively referred to as En, De, Zh, and Ja. It is cost-free for use in the real world and allows us to utilize heavy, high-quality NER models. But this method is yet to explore how to solve the entity translation disambiguation issue that is important for improving the translation quality.

Johnson et al. (2017) proposed a straightforward approach of utilizing a single NMT model to translate between multiple languages. This paper proposes introducing an artificial token at the beginning of the input sentence to identify the target language the model should translate to in order to make use of multilingual data within a single system. This is a simple modification to the input data that allows for the use of multilingual data. This paper found that zero-shot translation can be effective and that training a model across multiple languages can improve performance at the individual language level.

Sennrich et al. (2015) investigate ways to train with monolingual data without altering the architecture of the neural network in this paper. They used Groundhog as the implementation of the NMT system for all experiments. With additional monolingual data and parallel texts in English, German, and Turkish, they evaluated NMT training. They proposed two simple methods for this purpose.

Li et al. (2022) proposed a strategy to combine a neural-machine translation-based approach and a search-based automatic program repair method. ARJANMT, a novel framework for automatically repairing Java programs, is presented in this work. In order to examine the correctness and repairability of the proposed framework, two sets of controlled experiments are carried out on 410 bugs from two benchmarks. According to the findings of the experiments, combining these two kinds of repair methods—search-based and neural-machine-translation-based—gives better results or fixes bugs that were previously impossible to fix on their own.

Farhan et al. (2020) presented a new approach based on deep learning to convert dialectal sentences into standard language. They also introduced novel deep learning systems that translate dialectal sentences into standard language in supervised and unsupervised settings. In order to deal with the OOV problem, they employed a technique to learn embeddings using character n-gram and BPE. However, their approach did not focus on improving accuracy by utilizing neural networks and NER components to handle proper nouns and locations. Additionally, they did not expand their dataset to include longer sentences or investigate other dialects and languages, such as English.

Klein et al. (2017) have presented an open-source toolkit for neural machine translation (NMT). Their approach involves a conditional language modeling perspective, where the probability of a target sentence is modeled. However, they did not focus on further developing OpenNMT to achieve state-of-the-art MT results, and also did not prioritize providing a stable framework for production use.

## 3 NMT Architecture

NMT is based on neural networks and consists of an encoder and a decoder. The encoder takes the input sentence and converts it into a fixed-length vector, also called a thought vector. The decoder then generates the output sentence from the thought vector. The encoder and decoder are trained jointly using backpropagation to minimize the difference between the predicted output and the ground truth output.

In addition to the encoder and decoder, NMT also includes an attention mechanism. The attention mechanism allows the decoder to focus on specific parts of the input sentence when generating the output sentence. This improves the performance of NMT by allowing the decoder to take into account the context of the input sentence. The optimization algorithm used for training NMT is also important, with the most popular algorithm being Adam.

## 4 Training and Optimization

Training an NMT model involves feeding it pairs of source and target language sentences, and adjusting the parameters of the model to minimize the difference between the predicted translation and the actual translation. Optimization techniques such as stochastic gradient descent (SGD) and Adam are commonly used to update the model parameters during training. Regularization techniques such as dropout and weight decay can be used to prevent the overfitting of the model.

## 5 Evaluation Metrics

Neural Machine Translation (NMT) has become a popular approach to machine translation in recent years. Like any other machine learning algorithm, the performance of NMT can be evaluated using various metrics. In this section, we will discuss some of the most commonly used evaluation metrics for NMT.

### I. BLEU

BLEU (Bilingual Evaluation Understudy) is a widely used evaluation metric for machine translation. It compares the machine-generated output to a reference translation and calculates a score based on the number of matching n-grams (subsequences of words). The score ranges from 0 to 1, with 1 being the best possible score.

While BLEU is a popular metric, it has some limitations. For example, it is not very effective at capturing semantic similarity between the machine-generated output and the reference translation. Additionally, BLEU does not take into account the fluency of the generated output.

### II. METEOR

METEOR (Metric for Evaluation of Translation with Explicit ORdering) is another popular evaluation metric for machine translation. It uses a combination of precision, recall, and alignment between the machine-generated output and the reference translation to calculate a score. METEOR also takes into account the fluency of the generated output and is better at capturing semantic similarity than BLEU.

### III. TER

TER (Translation Error Rate) is a metric that measures the number of edits required to transform the machine-generated output into the reference translation. It is a more fine-grained metric than BLEU and takes into account the word order, grammar, and semantics of the generated output.

### IV. Other Metrics

Other metrics that are used to evaluate the performance of NMT include ROUGE (Recall-Oriented Understudy for Gisting Evaluation), CIDEr (Consensus-based Image Description Evaluation), and WER (Word Error Rate).

While these metrics are widely used in the evaluation of NMT, it is important to note that they have their limitations. No single metric can capture all aspects of machine translation, and the choice of evaluation metric depends on the specific task and the languages involved. It is always a good idea to use multiple metrics to get a more comprehensive evaluation of the performance of NMT.

## 6 Applications of NMT

Neural Machine Translation (NMT) has found applications in various fields, including:

1) **International Business and Trade:** NMT has become an essential tool for businesses involved in international trade, enabling them to easily communicate with partners, customers, and suppliers across the world. It helps to overcome the language barrier and allows businesses to expand their reach and increase sales.
2) **Tourism:** NMT is used extensively in the tourism industry for translating travel guides, menus, and other documents for

tourists. It helps tourists to navigate unfamiliar territory and enhances their overall travel experience.

3) **Healthcare:** NMT is also used in healthcare for translating medical documents, such as patient records, prescriptions, and medical reports. It helps to improve the quality of care by enabling healthcare professionals to communicate effectively with patients and colleagues who speak different languages.
4) **E-Commerce:** NMT is used in e-commerce platforms for translating product descriptions, reviews, and customer support inquiries. It helps to increase customer engagement and boost sales by providing a seamless shopping experience for customers who speak different languages.
5) **Legal:** NMT is used in the legal industry for translating legal documents, such as contracts, patents, and court proceedings. It helps to overcome the language barrier in international legal cases and enables lawyers to work with clients and partners from different countries.
6) **Localization:** NMT is used for localizing software, websites, and mobile apps for different regions and languages. It helps to expand the global reach of businesses by enabling them to provide a localized experience to their customers.
7) **Government:** NMT is used in government agencies for translating official documents, such as treaties, agreements, and laws. It helps to facilitate communication and collaboration between countries and enables governments to work together more effectively.

NMT has become an indispensable tool for various industries and applications, enabling businesses, organizations, and individuals to communicate and collaborate across different languages and cultures.

## 7 Future Directions and Challenges

Neural Machine Translation (NMT) has made significant progress in recent years, but there are still several challenges that need to be addressed to improve its performance and usability. Some of the challenges and future directions for NMT are:

1. **Handling Rare and Low-Resource Languages:** While NMT has shown excellent performance in high-resource languages such as English, French, and Chinese, it still struggles to handle rare and low-resource languages. NMT systems need to be improved to handle the diversity of languages in the world, especially those with limited resources.
2. **Domain Adaptation:** NMT systems often suffer from domain mismatch, where the training data and test data are from different domains. Domain adaptation techniques need to be developed to make NMT systems more adaptable to new domains.
3. **Explainability and Interpretability:** NMT is often considered a black box, making it difficult to understand how it generates translations. Explainability and interpretability techniques need to be developed to make NMT more transparent and trustworthy.
4. **Neural Architecture:** The neural architecture of NMT systems has a significant impact on their performance. Future research needs to focus on developing more efficient and effective neural architectures for NMT.
5. **Multilingual and Multimodal Translation:** The future of NMT lies in its ability to handle multilingual and multimodal translation, where multiple languages and modalities (e.g., text, speech, image) are involved. NMT systems need to be developed that can handle multiple languages and modalities simultaneously.
6. **Human Evaluation:** While automatic evaluation metrics such as BLEU and METEOR are widely used, they do not always align with human judgments of translation quality. Future research needs to focus on developing better human evaluation methods for NMT.

Overall, NMT has a promising future, but there are still many challenges that need to be addressed to improve its performance and usability. Addressing these challenges will require collaboration between researchers, developers, and industry practitioners, and will require a multidisciplinary approach that draws on expertise from linguistics, computer science, and other fields.

## 8 Conclusion

In conclusion, Neural Machine Translation (NMT) has become a rapidly advancing field with significant potential for transforming the way we communicate and collaborate across languages and cultures. NMT systems have shown remarkable progress in recent years, demonstrating their superiority over traditional machine translation methods in terms of accuracy and fluency. The deployment of NMT in various domains, such as international business, tourism, healthcare, e-commerce, and government, has highlighted its potential to enhance global communication and improve the quality of services offered.

However, several challenges need to be addressed to improve NMT performance and usability, including the handling of rare and low-resource languages, domain adaptation, explainability and interpretability, neural architecture, and evaluation metrics. Future research will need to address these challenges and focus on developing new techniques and models that can enhance NMT's robustness and adaptability.

The successful development and deployment of NMT systems will require interdisciplinary collaborations between researchers, developers, and industry practitioners. Addressing the challenges of NMT will require a multidisciplinary approach that leverages the expertise of computer scientists, linguists, and other related fields. Ultimately, NMT's success will depend on its ability to facilitate communication and collaboration across different languages and cultures, enabling individuals, organizations, and governments to work together more effectively in a globalized world.

## *References*

Su, J., Zeng, J., Xiong, D., Liu, Y., Wang, M., & Xie, J. (2018). A hierarchy-to-sequence attentional neural machine translation model. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(3), 623-632.

Xie, S., Xia, Y., Wu, L., Huang, Y., Fan, Y., & Qin, T. (2022). End-to-end entity-aware neural machine translation. *Machine Learning*, 1-23.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., ... & Dean, J. (2017). Google’s multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, *5*, 339-351.

Sennrich, R., Haddow, B., & Birch, A. (2015). Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Li, D., Wong, W. E., Jian, M., Geng, Y., & Chau, M. (2022). Improving search-based automatic program repair with Neural Machine Translation. *IEEE Access*, *10*, 51167-51175.

Farhan, W., Talafha, B., Abuammar, A., Jaikat, R., Al-Ayyoub, M., Tarakji, A. B., & Toma, A. (2020). Unsupervised dialectal neural machine translation. *Information Processing & Management*, *57*(3), 102181.

Klein, G., Kim, Y., Deng, Y., Senellart, J., & Rush, A. M. (2017). Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

Banitz, B. (2020). Machine translation: a critical look at the performance of rule-based and statistical machine translation. *Cadernos de traduçao*, *40*, 54-71.