International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 29, No. 3, June, 2024, pp. 831-838
Journal homepage: http://iieta.org/journals/isi

Enhancing Neural Machine Translation with Fine-Tuned mBART50 Pre-Trained Model: An Examination with Low-Resource Translation Pairs

Zhanibek Kozhirbayev

National Laboratory Astana, Nazarbayev University, Astana 010000, Kazakhstan

Corresponding Author Email: zhanibek.kozhirbayev@nu.edu.kz

https://doi.org/10.18280/isi.290304

Received: 3 January 2024
Revised: 20 March 2024
Accepted: 1 April 2024
Available online: 20 June 2024

Keywords:
neural machine translation, pre-trained models, fine-tuning, denoising auto-encoder, low-resource languages, Kazakh-Russian, Russian-Tatar

ABSTRACT

In the realm of natural language processing (NLP), the use of pre-trained models has seen a significant rise in practical applications. These models are initially trained on extensive datasets, encompassing both monolingual and multilingual data, and can be subsequently fine-tuned for target output using a smaller, task-specific dataset. Recent research in multilingual neural machine translation (NMT) has shown potential in creating architectures that can incorporate multiple languages. One such model is mBART50, which was trained on 50 different languages. This paper presents a work on fine-tuning mBART50 for NMT in the absence of high-quality bitext. Adapting a pre-trained multilingual model can be an effective approach to overcome this challenge, but it may not work well when the translation pairs contain languages not seen by the pre-trained model. In this paper, the resilience of the self-supervised multilingual sequence-to-sequence pre-trained model (mBART50) were investigated when fine-tuned with small amounts of high-quality bitext or large amounts of noisy parallel data (Kazakh-Russian). It also shows how mBART improves a neural machine translation system on a low-resource translation pair, where at least one language is unseen by the pre-trained model (Russian-Tatar). The architecture of mBART was employed in this study, adhering to the traditional sequence-to-sequence Transformer design. A Transformer Encoder-Decoder model with Byte Pair Encoding (BPE) was trained in our baseline experiment. The experiments show that fine-tuned mBART models outperform Baseline Transformer-based NMT models in all tested translation pairs, including cases where one language is unseen during mBART pretraining. The results show an increase in the BLEU score of 11.95 when translating from Kazakh to Russian and by 1.17 points in BLEU score when translating from Russian to Tatar. Utilizing pre-trained models like mBART can substantially reduce the data and computational requirements for NMT, leading to improved translation performace for low-resource languages and domains.

1. INTRODUCTION

In the field of NLP, there is a growing trend towards the use of models that have been pre-trained. These models are initially trained on a large amount of data, which can be either monolingual or multilingual, and then fine-tuned with a smaller dataset. The latest studies in the area of multilingual NMT have indicated the possibility of designing architectures that can handle multiple languages. One such model is mBART50, which was trained on 50 different languages. It can be adjusted or "fine-tuned" to perform a wide range of tasks. This flexibility makes them highly valuable in the field of NLP.

Machine translation (MT) is considered one of the main tasks of NLP, the purpose of which is to translate source language sentences into the target language. It has become an important technology in today's globalized world, where communication between people who speak different languages is increasingly common. Using the achievements of linguists, computer scientists, and the process of improving machine translation has grown exponentially over the past couple of years. Machine translation can be categorized into 3 types.

The RBMT relies on translation rules and linguistic knowledge that are manually crafted. Given the inherent complexity of natural languages, it is challenging to account for all language irregularities [1]. The rule-based technique entails using morphological, syntactic, and semantic rules to analyze the text in the source language and generate the target language content. Transfer-based translations and Interlingua machine translations are the two further categories of this approach. This method addresses the word-order issue, and the errors caused can be tracked because it employs linguistic information. RBMT receives the original text and generates an intermediate representation that might be an abstract form. The destination text is derived from the intermediate form. The set of rules for morphology, syntax, lexical choice and transfer, and semantic analysis and production is the foundation of these systems. This strategy, nevertheless, necessitates a lot of work and commitment from people. Systems based on transfers are more adaptable and can be expanded to include language pairings in a multilingual setting. A bilingual dictionary is used to translate sentences that match one of the rules or examples. It leverages the comparative understanding of the source and target languages. Multilingual translation can be done using Interlingua-based systems. Interlingua method transforms words into a universal language in order to translate them into multiple languages.

Statistical approaches that extract linguistic information from data are gaining more and more attention with the advent of parallel corpora [2]. Data-driven method of statistical machine translation leverages parallel dataset and considers translation as an issue of mathematical basis. The method is broken into three different techniques:

-Word based SMT: Word by word translations are performed after chunking down sentences into their smallest constituent parts, or words.

-Phrase based SMT: introduced by Koehn [2] and mostly employs phrases as the basic translational unit.

-Hierarchical phrases based SMT: This model is a fusion of phrase-based SMT and syntax-based translation.

Unlike former approach, SMT learns hidden structures, such as word or phrase alignment, directly from a parallel corpus. NMT is a recent approach to MT that uses artificial neural networks to learn how to translate one language into another [3]. NMT models have shown remarkable improvements in translation quality over the past few years and have become the dominant approach in MT research and applications. Unlike traditional machine translation systems that rely on rule-based approaches and handcrafted features, NMT models are trained end-to-end using large amounts of parallel corpora, which are pairs of sentences in the source language and their translations in the target language. However, such data is not always available, especially for low-resource languages or domains.

Pre-trained models play a crucial role in neural machine translation (NMT). They are initially trained on substantial volumes of data and have learned to recognize patterns and relationships within the data. They can be used as a starting point for NMT, which significantly reduces the amount of time and resources required to train a new model. Pre-trained models can also improve the quality of NMT output, especially for low-resource languages or domains where training data is scarce. By leveraging the knowledge learned from a large amount of data, pre-trained models can help NMT systems generate more accurate and natural translations. Additionally, pre-trained models can be adjusted for targeted domains, further enhancing their utility and improving translation quality.

There are several pre-trained models for NMT that have been developed and made publicly available by researchers and companies. Here are some examples:

mBART (multilingual BERT): A pre-trained model that supports 50 languages and is trained on substantial volumes of monolingual and parallel data [4]. Multilingual translation models benefit from multilingual fine-tuning, where a pretrained model is fine-tuned simultaneously on multiple language directions. This approach extends pretrained models to include additional languages without performance loss. On average, multilingual fine-tuning improves by 1 BLEU over the strongest baselines and shows a significant improvement of 9.3 BLEU on average compared to bilingual baselines built from scratch.

XLM (cross-lingual language model): A pre-trained model that supports over 100 languages and is trained on monolingual and parallel data using a masked language modeling objective [5]. This work extends generative pretraining for English natural language understanding to multiple languages, showcasing the effectiveness of cross-lingual pretraining. Two methods for learning cross-lingual language models (XLMs) are proposed: an unsupervised method using monolingual data and a supervised approach leveraging parallel data. The results demonstrate state-of-the-art performance in cross-lingual classification and significant improvements in unsupervised and supervised machine translation. Notably, on XNLI, there is a 4.9% absolute gain in accuracy. In the field of unsupervised machine translation, the method employed achieves a BLEU score of 34.3 on the WMT'16 German-English task, exceeding the previous state-of-the-art performance by over 9 BLEU points. When it comes to supervised machine translation, a new benchmark is set with a BLEU score of 38.5 on the WMT'16 Romanian-English task, surpassing the previous best method by more than 4 BLEU points.

MASS (masked sequence-to-sequence pre-training): A pre-trained model that is trained on monolingual and parallel data using a masked sequence-to-sequence objective [6]. The MASS method utilizes the encoder-decoder framework to reconstruct a fragment of a sentence given the rest of the sentence. Specifically, the encoder processes a sentence in which a fragment (consisting of several consecutive tokens) has been randomly masked. The decoder then attempts to predict this masked fragment. The joint training approach empowers MASS to cultivate the ability to extract representations and model language. After additional fine-tuning on a variety of language generation tasks with zero or low resources, including neural machine translation, MASS exhibits substantial enhancements over baseline models that lack pre-training or use different pre-training methods. Notably, it achieves state-of-the-art accuracy, with a BLEU score of 37.5, in unsupervised English-French translation, even surpassing the performance of the early attention-based supervised model.

T5 (text-to-text transfer transformer): A pre-trained model that is trained on a variety of natural language processing tasks, including machine translation, using a text-to-text format [7]. T5 is a pre-trained encoder-decoder model designed for a multi-task setting, incorporating a mixture of unsupervised and supervised tasks. Each task is transformed into a text-to-text format. T5 exhibits versatility across various tasks by effectively handling them without task-specific modifications. This adaptability is achieved by adding a task-specific prefix to the input corresponding to each task.

MarianMT: A pre-trained model that supports over 50 languages and is trained on substantial volumes of parallel data using a sequence-to-sequence model [8]. MarianMT is a rapid translation framework developed in C++ and primarily managed by the Microsoft Translator team. It serves as the NMT engine powering Microsoft's NMT service.

The goal of our study is to investigate the impact of pre-trained mBART models on machine translation performance in different scenarios. The fine-tuned model used is a mBART-large-50 multilingual Sequence-to-Sequence model, an extension of the original mBART model that supports multilingual machine translation for 50 languages. The study focuses on the Kazakh-Russian language pair and examines NMT performance when both languages are seen by the pre-trained model, as well as the scenario when one language is not seen. For this purpose, the Russian-Tatar language pair was utilized to adjust the mBART model, even though Tatar was not among the 50 supported languages. The study also investigates how the size of the parallel sentence corpus used for fine-tuning affects model performance. In addition, the researchers compared the performance of the mBART model to that of other machine translation models, such as the transformer-based model. This research provides valuable insights into the effectiveness of pre-trained models like mBART in low-resource settings and the potential for leveraging multilingual models to improve machine translation performance.

The structure of this paper is as follows: Section 2 provides an overview of recent work on the use of pre-trained models for NMT. The goal and objectives of this work are also presented in Section 2. Information about the data set and the characteristics of the Kazakh-Russian and Russian-Tatar language pairs are given in Section 3. It also describes the Baseline Transformer approach and fine-tuning of the pre-trained mBART descriptions. The results derived from the experiments carried out are detailed in Section 4. The summary and conclusions drawn from these experiments, along with potential areas for future research, are presented in Section 5.

2. LITERATURE REVIEW

Neural machine translation has emerged as a favored method for automated translation, given the numerous recent developments in the area. A notable progression is the employment of pre-trained models, which can markedly enhance the performance and effectiveness of NMT systems. In this literature survey, we will delve into multiple research works that have investigated the application of pre-trained models in NMT.

A significant study highlighted the efficacy of pre-trained models in scenarios with limited resources [9]. The investigators trained a model on an extensive parallel corpus and subsequently fine-tuned it on a smaller dataset for a specific language pair. The outcomes indicated substantial enhancements in translation quality, particularly for languages with limited resources. The paper’s primary contributions include showcasing the enhancement of machine translation quality in NMT systems by integrating monolingual target sentences into the training set. The authors examined two strategies for filling the source side of monolingual training instances, utilizing a dummy source sentence and a source sentence acquired using backtranslation, referred to as synthetic.

Another study uses an initially trained word embedding model to initialize the NMT encoder and decoder, which improved the model's ability to handle out-of-vocabulary words and rare words [10]. Two simple and effective attentional mechanisms for neural machine translation are proposed: the global approach, which always considers all source positions, and the local approach, which selectively attends to a subset of source positions at a time. The results showed that using pre-trained embeddings significantly improved the translation quality, especially for low-frequency words. The recent study by Zhu et al. [11] introduces a novel algorithm referred to as the BERT-fused model, which involves utilizing BERT to obtain representations for an input sequence. These representations are then combined with each layer of the encoder and decoder of the neural machine translation (NMT) model through attention mechanisms. The authors conduct experiments across various machine translation tasks, including supervised tasks at both the sentence and document level, semi-supervised tasks, and unsupervised tasks. They argue that the proposed BERT-fused model achieves state-of-the-art performance on seven benchmark datasets. Another research paper [12] delves into the utilization of pre-trained models in NMT and introduces a novel method known as the Concerted Training Approach (CTNMT). This method is specifically designed to maximize the effectiveness of BERT in NMT. The authors put forth three innovative techniques to seamlessly blend the strengths of pre-trained BERT with traditional NMT. These techniques encompass asymptotic distillation, a dynamic mechanism for knowledge fusion, and rate-scheduled updating.

Recently, multilingual BART (mBART) has shown promising results in NMT. mBART is a pre-trained transformer-based model that learns to represent text in a language-agnostic way. However, since it was trained on a mixed-language corpus, it may not be optimized for NMT tasks. To address this issue, a multilingual variant of the BART (Bidirectional and Auto-Regressive Transformer) model, known as mBART, was introduced [4]. In their study, the authors argue that incorporating mBART initialization results in enhanced performance in all but the most well-resourced scenarios, with improvements of up to 12 BLEU points for low-resource machine translation tasks and over 5 BLEU points for several document-level and unsupervised models. Moreover, it facilitates the transfer of knowledge to language pairs lacking parallel text or not included in the pre-training dataset. Another study further explored the effectiveness of mBART in NMT. The researchers compared the performance of mBART to several other pre-trained models, including mBERT and XLM-R. They found that mBART achieved the best performance across multiple language pairs, including low-resource languages [13]. Additionally, the recent paper [14] investigated the use of mBART in domain adaptation for NMT. The researchers evaluate a methodology on five language pairs across three domains with varying levels of data availability. Their results show that models that use an mBART initialization typically exhibit superior performance when compared to those that use a random Transformer initialization.

Analysis of literature data shows that performance outperforms other popular NMT models and is particularly effective in low-resource languages and cross-language transfer scenarios by using the pre-trained mBART model. Additionally, it can be fine-tuned for domain adaptation, further improving its utility in real-world NMT applications. Given its strong performance and versatility, mBART is likely to continue to be a valuable tool for NMT researchers and practitioners.

3. MATERIALS AND METHODS

3.1 Dataset

This part of the section focuses on the datasets used in the NMT experiments for the language pairs KZ-RU and RU-TA. The parallel corpus for each language pair is derived from various sources, albeit limited in availability. The selection of these sources is based on their accessibility, given the scarcity of options for the specified language pairs.

3.1.1 KZ-RU parallel corpus

A parallel corpus for Kazakh and Russian was compiled from three separate sources. The primary source, developed in our laboratory [15, 16], consists of over 890K parallel sentences extracted from 15 online news portals. These portals are affiliated with state agencies, national firms, and other semi-governmental entities that adhere to a policy of bilingual publication in both Kazakh and Russian. Additionally, these portals typically provide page-level alignment, meaning a Russian version of a page directly links to its Kazakh counterpart, and vice versa. This feature simplified document alignment during the web crawling phase.

The Kazakh-Russian parallel corpus was further supplemented by a second source, which was derived from language resources developed by researchers from KazNU [17]. This source added more than 86,000 parallel sentences to the corpus, thereby enriching it and expanding its range. This is a significant contribution to the corpus, enhancing its diversity and comprehensiveness.

The third source comes from TIL Corpus [18, 19], which is a parallel corpus that combines the majority of publicly available datasets for 22 Turkic languages. Authors of TIL Corpus state, that the parallel corpus was compiled by combining open source datasets. The following open data sets were incorporated into TIL Corpus:

- The Tatoeba corpus: This is a comprehensive language resource that includes datasets for more than 500 languages and thousands of translation pairs;
- JW300: This resource focuses on 59 language pairs of interest and contains approximately 5.2 million parallel sentences;
- GoURMET5: it has 7 language pairs.

The Kazakh-Russian parallel sentences were combined from the following domains: UDHR, Bible, Ted Talks and Mozilla.

3.1.2 RU-TA parallel corpus

Under a non-disclosure agreement, the Tatar-Russian bitext dataset [20, 21] from the Institute of Applied Semiotics, Academy of Sciences of the Republic of Tatarstan, was acquired. It contains 360K+ parallel sentences and covers a diverse range of domains. The second source is also taken from the TIL Corpus and has been combined from the following: religious (Bible) and conversational (TED Talks).

Table 1 displays the size of dataset for the KZ-RU and RU-TA pair languages.

3.2 Methods

3.2.1 Baseline Transformer model

This part of the text provides details on the training of baseline NMT models, which are rooted in the Transformers framework. The training was enabled by Joey NMT. It incorporates a variety of prominent NMT functionalities within a neat and user-friendly code base. It delivers performance on standard benchmarks that matches that of more sophisticated toolkits [22].

Encoder. When given an input x_1, ..., x_{l_x}, $E_{SRC}x_i$ was used to find word embeddings.

$$X \in \mathbb{R}^{l_x \times d} \tag{1}$$

where, l_x represents the length of the sentence, d refers to the dimensionality of the embeddings.

The given learnable parameters were delineated in the subsequent manner:

$$A \in \mathbb{R}^{d \times d_a}, B \in \mathbb{R}^{d \times d_a}, C \in \mathbb{R}^{d \times d_o} \tag{2}$$

where, d_a is dimensionality of the attention space, d_o is output dimensionality.

Using these matrices, the input matrix is converted into new word representations H as:

$$H = \underbrace{\text{softmax}(XA\,B^T X^T)}_{\text{self-attention}} XC \tag{3}$$

It achieves multiheaded attention by computing this transformation k times, where each iteration corresponds to a distinct attention head. In each of these computations, diverse sets of parameters denoted as A, B, and C are applied, allowing the model to capture different facets of information simultaneously. We compact the results of computing all k Hs, perform layer normalization, and then add a feed-forward layer:

$$H = [H^{(1)}; \dots; H^{(k)}] \tag{4}$$

$$H' = layer - norm(H) + X \tag{5}$$

$$H^{(enc)} = feed - forward(H') + H' \tag{6}$$

Decoder. The decoder works similarly to the encoder, except it accepts stacked target embeddings $Y \in \mathbb{R}^{l_y \times d}$ as input:

$$H = \underbrace{\text{softmax}(YA\,B^T Y^T)}_{\text{masked self-attention}} YC \tag{7}$$

Multi-headed attention will be calculated once more:

$$Z = \underbrace{\text{softmax}\left(H'A\,B^T H^{(enc)^T}\right)}_{\text{src-trg attention}} H^{(enc)}C \tag{8}$$

$$H^{(dec)} = feed - forward(layer - norm(H' + Z)) \tag{9}$$

where, H' - intermediate decoder representations, $H^{(enc)}$ - final encoder representations.

Target words will be hypotised using $H^{(dec)}W_{out}$.

3.2.2 Fine-tuned mBART model

BART is a pioneering technique that initially trains a comprehensive denoising auto-encoder capable of managing sequence-to-sequence tasks. Its primary training has been conducted on an extensive collection of monolingual data, specifically in English. The foundational architecture behind BART is the Transformer architecture, a key element in its standard sequence-to-sequence implementation. The architecture plays a decisive role in the formation of BART, incorporating a bidirectional auto-encoder and a left-to-right autoregressive decoder. It comprises two architectural types: BERT, equipped with a bidirectional encoder, and GPT, furnished with a left-to-right decoder. The fundamental version consists of 6 layers each for the encoder and decoder, whereas the larger variant contains 12.

mBART is a multilingual adaptation of BART, developed to initial train on several monolingual languages. mBART, like its forerunner, employs a 12-layer structure for both encoding and decoding, with layer-normalization following each layer. The initial variant supported 25 languages, but an improved one now supports up to 50 languages. In our study, we adopted the mBART architecture, which conforms to the traditional sequence-to-sequence Transformer-based design with a 12-layer encoder-decoder and a model dimension of 1024 across 16 heads. For the training of our model, we utilized FairSeq [23].

The performance of the NMT models was evaluated using the BLEU score, a key metric in natural language processing that quantifies the similarity between the generated translations and reference translations. The BLEU score provides a numerical assessment of the model's performance by comparing the overlap of n-grams in the generated output with those in the reference translations. This metric was employed to facilitate a comprehensive evaluation of the NMT models, allowing for a quantitative analysis of their translation quality and alignment with reference translations.

Table 1. Size of parallel dataset

Language Pair	Source	Size of Dataset (Parallel Sentences)
KZ-RU	NU corpus	893 234
	KazNU corpus	86 453
	TIL corpus	3 433 698
	Total	**4 413 385**
RU-TA	TIL corpus	275 462
	Institute of applied semiotics	360 821
	Total	**636 283**

4. RESULTS AND DISCUSSION

This subsection details the experiments carried out on both the Baseline Transformer model and the fine-tuned mBART model. Specifically, it outlines the hyperparameters utilized during the model training process.

4.1 Baseline Transformer model

The Transformer-based NMT is influenced by the quantity of available training data and the set configuration parameters. The model's encoder, used for both KZ-RU and RU-TA language pairs, utilizes a 512-dimensional embedding and hidden layer sizes. The decoder mirrors the encoder's dimensions. The Adam optimizer was employed for training the model, with a learning rate ranging from a maximum of 0.0003 to a minimum of 10^{-8}. Both the encoder and the decoder were set with a dropout probability of 0.3. A significant performance boost was observed when a different tokenization strategy was adopted. While NMT usually tokenizes using words, the use of BPE led to a substantial performance improvement. As a result, BPE was applied with a joint vocabulary size of 32K. Two separate experiments were performed on the KZ-RU language pair to scrutinize the training data. The first experiment utilized the NU and KazNU corpora, while the TIL corpus was the sole basis for the second experiment. The first experiment had a medium-sized training set, and the baseline model's performance was somewhat underwhelming, with a BLEU score of 35.65 on the test set. Conversely, the second experiment demonstrated a significantly higher BLEU score of 59.70 on the test set. The result for the RU-TA language pair was 30.48 BLEU scores. Table 2 displays the BLEU scores of the model on the test set. Additionally, the performance of the baseline models during the training process for the both language pairs is depicted in Figures 1 and 2.

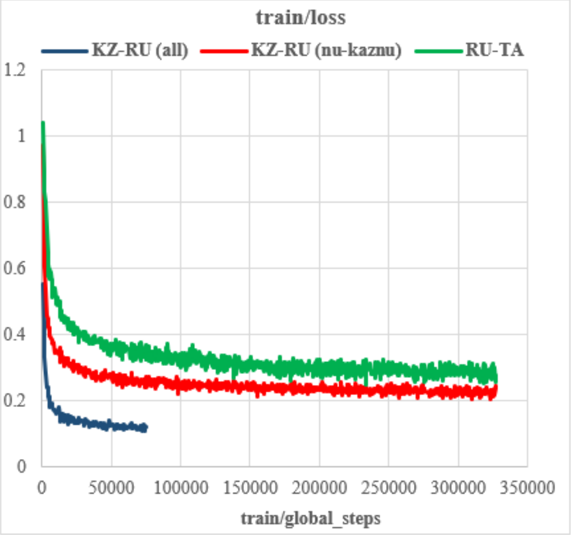

Figure 1. Training Baseline Transformers model: training loss

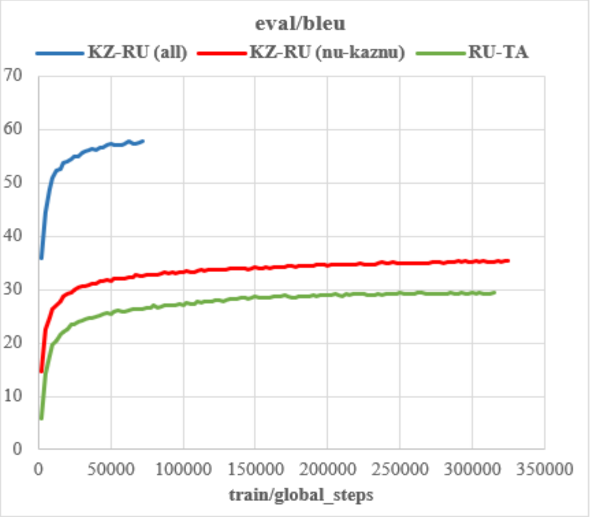

Figure 2. Baseline Transformers model: BLEU score

Table 2. BLEU score on test set

Experiments	Corpus	BLEU Score
Baseline Transformers	KZ-RU (nu-kaznu)	35.65
	KZ-RU (til)	59.70
	RU-TA	30.48
Fine-tuning mBART50	KZ-RU (nu-kaznu)	47.60
	RU-TA	31.65

4.2 Fine-tuning mBART model

In our study, we conducted an experiment to fine-tune the Facebook/mbart-large-50-many-to-many-mmt model. This model, which includes support for both Kazakh and Russian among 50 languages, was utilized. Unfortunately, Tatar language is not supported by this model. To address this issue, a unique token for the “tt_RU” language code was incorporated into the model’s vocabulary. It was discovered that despite the model’s lack of pre-existing knowledge about the Tatar language, it could still be fine-tuned for Tatar. This is because the 50 languages supported by the model could potentially include fragments of Tatar text.

The training process was conducted with bidirectional training, incorporating a dropout of 0.3, label smoothing of 0.2, 2500 warm-up steps, and a maximum learning rate of 3×10^{-5}, as specified by Liu et al. [4]. The model underwent training for up to 100k updates, with the last model selection based on validation likelihood. The performance of the model was evaluated using beam search with a beam size of 5. The results were calculated against the true-target tokenized data, with the scores reported in BLEU.

It is observed that pre-training itself is the most crucial factor. A substantial gain in BLEU is achieved by both language pairs, indicating that the use of pre-trained models can significantly enhance translation accuracy. As illustrated in Figures 3 and 4, the fine-tuned model shows a significant enhancement compared to the baseline assessments across all datasets, particularly a gain of +13.57 points for KZ-RU and +3.67 points for RU-TA on the dev set. The test set results also display a significant improvement, with an increase of +11.95 points for KZ-RU and +1.17 points for RU-TA.

The paired t-tests conducted for both the KZ-RU (NU-KAZNU) and RU-TA language pairs resulted in t-statistic values approaching negative infinity, accompanied by p-values effectively reaching zero. These findings signify a highly significant difference between the Baseline Transformers and Fine-tuned mBART50 models. The negative infinity t-statistic indicates an exceptional level of consistency in the observed differences between pairs. In practical terms, these results strongly support the conclusion that the Fine-tuned mBART50 models exhibit a substantial and statistically significant improvement in translation performance compared to the Baseline Transformers models for both language pairs.

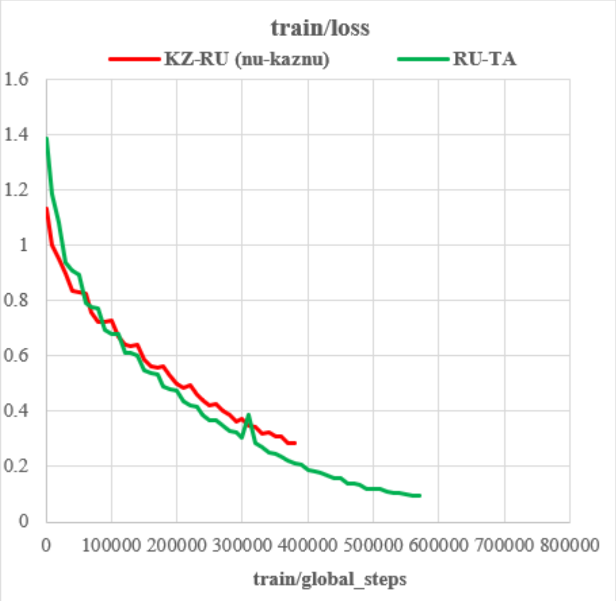

Figure 3. Fine-tuning mBART: train_loss

The improved performance of the fine-tuned mBART model over the baseline can be attributed to effective transfer learning from pre-training on a diverse dataset, allowing the model to grasp nuanced language representations. The bidirectional auto-encoder and left-to-right autoregressive decoder architecture, coupled with multiheaded attention, contribute to enhanced representation learning. The model's notable success in low-resource language pairs underscores its adaptability to challenging scenarios. However, potential limitations include sensitivity to data quality, domain specificity, and hyperparameter choices. Additionally, reliance on BLEU scores may overlook certain aspects of translation quality. Addressing these considerations is crucial for a more comprehensive understanding of the model's effectiveness and potential biases in this study. Continuous exploration and refinement of techniques remain essential for advancing machine translation capabilities.

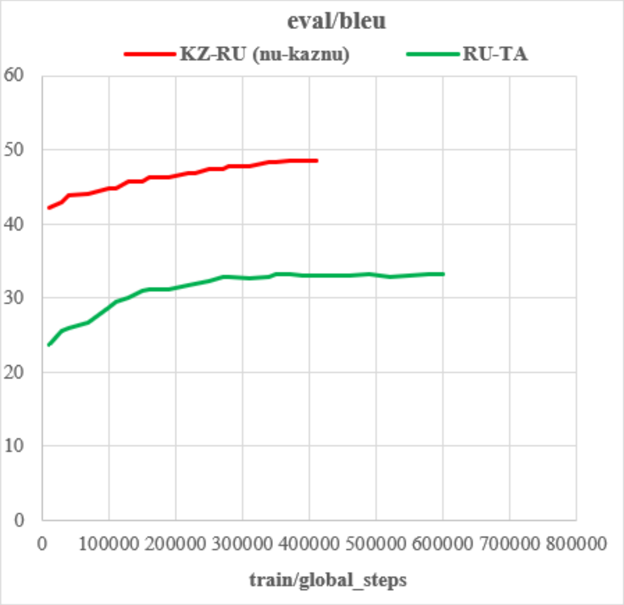

Figure 4. Fine-tuning mBART: BLEU score

5. CONCLUSIONS

This study aimed to train and test different neural machine translation models for the Kazakh-Russian and Russian-Tatar language pairs. The Transformers architecture served as the foundation for the baseline neural machine translation models trained using Joey NMT tool. Additionally, the study fine-tuned the mBART model for both language pairs. The research primarily focused on the Kazakh-Russian language pair and examined the NMT performance when the pre-trained model saw both languages, as well as when it only saw one language. To accomplish this, the Russian-Tatar language pair was utilized to fine-tune the pre-trained model, even though Tatar was not among the 50 supported languages. The study also investigated how the size of the parallel sentence corpus used for fine-tuning influenced model performance.

The outcomes of pre-trained models showed that the trained BART model delivered significantly superior performance compared to the baseline models. An important finding is that multilingual pre-trained models can be adapted to the Tatar language, even without any prior knowledge of Tatar, as was the case with mBART. This approach is particularly useful for low-resource languages or domains where obtaining large amounts of parallel data is challenging and expensive. In addition, the accuracy of translating languages with scarce resources can be enhanced by using pre-trained models. This is achieved by transferring knowledge from language pairs that have abundant resources.

The benefits of utilizing pre-trained models like mBART have been acknowledged, and it is suggested that exploration be conducted into their applicability in domain-specific or specialized contexts in future work. Consideration could be given to tailoring pre-trained models for industries such as healthcare, finance, or legal domains, addressing specific language nuances and requirements within those domains. Furthermore, given the potential of mBART in low-resource language scenarios, attention might be directed towards enhancing its performance through transfer learning from resource-rich languages. Techniques for adapting and refining pre-trained models for diverse linguistic landscapes could be investigated to contribute to the broader accessibility and effectiveness of natural language processing tools.

ACKNOWLEDGMENT

This research has been funded by the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan (Grant No.: AP13068635).

REFERENCES

[1] Barreiro, A., Scott, B., Kasper, W., Kiefer, B. (2011). OpenLogos machine translation: philosophy, model, resources and customization. Machine Translation, 25(2): 107-126. https://doi.org/10.1007/s10590-011-9091-z

[2] Koehn, P. (2010). Statistical Machine Translation. Cambridge University Press.

[3] Stahlberg, F. (2020). Neural machine translation: A review. Journal of Artificial Intelligence Research, 69: 343-418. https://doi.org/10.1613/jair.1.12007

[4] Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics, 8: 726-742. https://doi.org/10.1162/tacl_a_00343

[5] Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H., Stoyanov, V. (2018). XNLI: Evaluating cross-lingual sentence representations. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, pp. 2475-2485. https://doi.org/10.18653/v1/D18-1269

[6] Song, K., Tan, X., Qin, T., Lu, J., Liu, T.Y. (2019). MASS: Masked sequence to sequence pre-training for language generation. In International Conference on Machine Learning, pp. 5926-5936.

[7] Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel, C. (2021). mT5: A massively multilingual pre-trained text-to-text transformer. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, pp. 483-498. https://doi.org/10.18653/v1/2021.naacl-main.41

[8] Soliman, A.S., Hadhoud, M.M., Shaheen, S.I. (2022). MarianCG: a code generation transformer model inspired by machine translation. Journal of Engineering and Applied Science, 69(1): 104. https://doi.org/10.1186/s44147-022-00159-4

[9] Sennrich, R., Haddow, B., Birch, A. (2016). Improving neural machine translation models with monolingual data. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, pp. 86-96. https://doi.org/10.18653/v1/P16-1009

[10] Luong, M.T., Pham, H., Manning, C.D. (2017). Effective approaches to attention-based neural machine translation. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 1412-1421). https://doi.org/10.18653/v1/D15-1166

[11] Zhu, J., Xia, Y., Wu, L., He, D., Qin, T., Zhou, W., Li, H., Liu, T.Y. (2020). Incorporating bert into neural machine translation. arXiv preprint arXiv:2002.06823. https://doi.org/10.48550/arXiv.2002.06823

[12] Yang, J., Wang, M., Zhou, H., Zhao, C., Zhang, W., Yu, Y., Li, L. (2020). Towards making the most of BERT in neural machine translation. Proceedings of the AAAI Conference on Artificial Intelligence, 34(5): 9378-9385. https://doi.org/10.1609/aaai.v34i05.6479

[13] Yuan, B., Li, Y., Chen, K., Lu, H., Yang, M., Cao, H. (2022). An improved multi-task approach to pre-trained model based MT quality estimation. In Machine Translation: 18th China Conference, CCMT 2022, Lhasa, China, pp. 106-116. https://doi.org/10.1007/978-981-19-7960-6_11

[14] Verma, N., Murray, K., Duh, K. (2022). Strategies for adapting multilingual pre-training for domain-specific machine translation. In Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas, Orlando, USA, pp. 31-44.

[15] Makazhanov, A., Myrzakhmetov, B., Kozhirbayev, Z. (2017). On various approaches to machine translation from Russian to Kazakh. In The 5th International Conference on Turkic Languages Processing, Kazan, Tatarstan, pp. 195-209.

[16] Kozhirbayev, Z., Islamgozhayev, T. (2023). Cascade speech translation for the Kazakh language. Applied Sciences, 13(15): 8900. https://doi.org/10.3390/app13158900

[17] Balzhan, A., Akhmadieva, Z., Zholdybekova, S., Tukeyev, U., Rakhimova, D. (2015). Study of the problem of creating structural transfer rules and lexical selection for the Kazakh-Russian machine translation system on Apertium platform. In Proceedings of the International Conference "Turkic Languages Processing" TurkLang-2015, Kazan, Tatarstan, Russia, pp. 5-9.

[18] Mirzakhalov, J., Babu, A., Ataman, D., et al. (2021). A large-scale study of machine translation in the Turkic languages. arXiv preprint arXiv:2109.04593. https://doi.org/10.48550/arXiv.2109.04593

[19] Mirzakhalov, J., Babu, A., Kunafin, A., et al. (2021). Evaluating multiway multilingual NMT in the Turkic languages. arXiv preprint arXiv:2109.06262. https://doi.org/10.48550/arXiv.2109.06262

[20] Khusainov, A., Suleymanov, D., Gilmullin, R. (2020). The influence of different methods on the quality of the Russian-Tatar neural machine translation. In Artificial Intelligence: 18th Russian Conference, Moscow, Russia, pp. 251-261. https://doi.org/10.1007/s10590-011-9090-0

[21] Khusainov, A., Suleymanov, D., Gilmullin, R., Gatiatullin, A. (2018). Building the Tatar-Russian NMT system based on re-translation of multilingual data. In Text, Speech, and Dialogue: 21st International Conference, Brno, Czech Republic, pp. 163-170. https://doi.org/10.1007/978-3-030-00794-2_17

[22] Kreutzer, J., Bastings, J., Riezler, S. (2019). Joey NMT: A minimalist NMT toolkit for novices. arXiv preprint arXiv:1907.12484. https://doi.org/10.48550/arXiv.1907.12484

[23] Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. arXiv preprint arXiv:1904.01038. https://doi.org/10.48550/arXiv.1904.01038