


Article

A Reverse Positional Encoding Multi-Head Attention-Based Neural Machine Translation Model for Arabic Dialects

Laith H. Baniata ^{1,*}, Sangwoo Kang ^{1,*}  and Isaac. K. E. Ampomah ²¹ School of Computing, Gachon University, Seongnam 13120, Korea² Shell Center, 2 York Road, London SE1 7LZ, UK

* Correspondence: laith@gachon.ac.kr (L.H.B.); swkang@gachon.ac.kr (S.K.)

Abstract: Languages with a grammatical structure that have a free order for words, such as Arabic dialects, are considered a challenge for neural machine translation (NMT) models because of the attached suffixes, affixes, and out-of-vocabulary words. This paper presents a new reverse positional encoding mechanism for a multi-head attention (MHA) neural machine translation (MT) model to translate from right-to-left texts such as Arabic dialects (ADs) to modern standard Arabic (MSA). The proposed model depends on an MHA mechanism that has been suggested recently. The utilization of the new reverse positional encoding (RPE) mechanism and the use of sub-word units as an input to the self-attention layer improve this sublayer for the proposed model's encoder by capturing all dependencies between the words in right-to-left texts, such as AD input sentences. Experiments were conducted on Maghrebi Arabic to MSA, Levantine Arabic to MSA, Nile Basin Arabic to MSA, Gulf Arabic to MSA, and Iraqi Arabic to MSA. Experimental analysis proved that the proposed reverse positional encoding MHA NMT model was efficiently able to handle the open grammatical structure issue of Arabic dialect sentences, and the proposed RPE MHA NMT model enhanced the translation quality for right-to-left texts such as Arabic dialects.



Citation: Baniata, L.H.; Kang, S.; Ampomah, I.K.E. A Reverse Positional Encoding Multi-Head Attention-Based Neural Machine Translation Model for Arabic Dialects. *Mathematics* **2022**, *10*, 3666. <https://doi.org/10.3390/math10193666>

Academic Editors: Nebojsa Bacanin and Catalin Stoean

Received: 12 August 2022

Accepted: 21 September 2022

Published: 6 October 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: reverse positional encoding (RPE); multi-head attention; neural machine translation (NMT); Arabic dialects; MSA

MSC: 68T07

1. Introduction

Neural encoder-decoder models have been successfully applied to different natural language processing tasks, such as question answering, chatbots, and machine translation (MT) [1]. For several years, machine translation MT researchers have attempted to provide high-quality translations of common languages. Languages with limited resources, such as Arabic dialects, which are spoken versions of MSA in all countries in the Arab world, have been disregarded by the industrial sector and the scientific research community. Traditional machine translation approaches that perform translation tasks from ADs to MSA generate incompatible sentences. These traditional MT approaches perform translation twice for parts of a source sentence and generate target sentences [2]. Furthermore, a linguistic concept called diglossia occurs in the Arabic language, in which language speakers use local Arabic dialects in informal settings and the standard Arabic language in formal settings. For instance, local communities in Algeria use MSA and the Algerian dialects based on the situation and context. The Algerian dialect reflects its history, identity, culture, and lived experience. There are different types of Arabic dialects; they vary by region, such as Levantine (Palestine, Syria, Jordan, Lebanon), Maghrebi (Algeria, Morocco, Tunisia, and Libya), Iraqi, Nile Basin dialects (Sudan and Egypt), and Arabian Gulf dialects (Saudi Arabia, Yemeni, UAE, Oman, Kuwait, Qatar, and Bahrain). Machine translation is divided into four types: NMT, hybrid MT, rule-based MT, and statistical MT (SMT). Conventional translation models, such as SMT, require powerful computing devices. Word order is considered one

of the syntactic problems in Arabic dialects. To investigate the word order problem, we must first determine where the verb, object, and subject are located in a sentence. As mentioned in the literature review, languages are classified as verb-object-subject such as Arabic, subject-object-verb such as Korean, subject-verb-object such as English, and other language support sentences with a free order of words such as Arabic dialect. The free-word order feature in Arabic dialect sentences conveys information about the object, subject, and different types of information. These variations present a difficulty for SMT approaches because as Arabic dialects' sentences grow longer, they will carry information about more than the verb, subject, and object but also other different complicated contextual information. In NMT approaches, the encoder adds the input sequence to a single vector representation, as noted in the encoder-decoder structure, where the decoder uses this vector representation to generate the output sequences. Moreover, one disadvantage of this structure is that the input sequence data are lost and the translation performance decreases as the length of the input sequence increases. In addition, the absence of structured orthographies for ADs is considered one of the difficulties in building effective NMT approaches for these dialects. This absence includes morphological dissimilarities for these dialects, which are obvious in the usage of suffixes and affixes that do not exist in modern standard Arabic. Moreover, to train NMT models, a large amount of training data is required, which is not available in the case of Arabic dialects. The translation quality declines with a decrease in the training data for AD. Rare-word translation in colloquial Arabic is an open issue. The translation of Arabic dialect texts requires fewer word-level approaches. For example, in NMT methods that operate on a word level, the translation of out-of-vocabulary words has been studied using back-off to dictionary lookup [1,2]. In reality, these methods make incorrect assumptions. For example, as a reason for the differences in the morphological synthesis between MSA and colloquial Arabic, one-to-one translation between a source word and a target word does not occur. In addition, NMT approaches that work at the word level are considered inefficient for translating unseen words.

In recent years, intensive research has been conducted on NMT, from designing new architectures such as the plain sequence-to-sequence model [3] to presenting the attention mechanism approach [4] and models that use only self-attention instead of recurrent neural networks (RNN) [5]. From these architectures, transformers have emerged as the dominant NMT paradigm. The transformer model is a popular architecture that achieves state-of-the-art performance for various learning tasks. It outperformed the convolutional neural networks (CNN) and RNN models in different translation tasks. The former uses the self-attention approach to measure the relationship between two different words in a sentence. Furthermore, the transformer contributes significantly to increasing the quality of the machine translation and different natural language processing (NLP) tasks. When the transformer architecture is applied to direct text translation between different languages, it emerges as the highest-performing option for several datasets. By using the self-attention approach in every encoder and decoder, the transformer model attends to words that exist in the same sentence, whether it is a target sentence or a source sentence. Furthermore, the transformer model performs encoding of the positional information for every word, such as the word order, as a positional encoding (PE) in which the RNN and CNN are not utilized. The transformer architecture depends on MHA layers for managing sequences of different lengths.

MHA layers provide little positional information. The values and keys in multi-head attention are managed as sets without ordering, and re-ordering the queries simply results in a re-ordered output. While the decoder might obtain certain positional information from the left-to-right masking scheme, the encoder does not have access to any positional information for right-to-left text, such as Arabic dialects. This problem can be solved by reversing the positional encoding, such that the encoder can have access to any positional information. One of these methods is to copy unidentified words into the target text, as applied in [1,2]. This mechanism is considered applicable to names, but it requires morphological changes and transliteration when characters are different.

The methods and algorithms that deal with the translation of colloquial Arabic are under investigation and research. To the best of our knowledge, no previous research has investigated the impact of reversing positional encoding for right-to-left text translation, such as ADs, in the multi-head attention-based NMT model. In summary, our contributions are as follows:

- This research project presented a multi-head attention NMT model based on reversing positional encoding for the translation of right-to-left texts. Moreover, this project used an approach called word-piece to create a sub-word unit for ADs. The bilingual evaluation understudy (BLEU) results of the practical experiments showed that reversing positional encoding for the multi-head attention NMT approach enhanced the translation quality of ADs to MSA. The proposed NMT model, which employs reverse positional encoding, achieved higher quality and efficiency for the translation of rare words when compared with models that use a large vocabulary.
- Furthermore, the research introduced a new mechanism that shows the positive impact of reversing positional encoding for right-to-left text, such as Arabic dialects, and compares it with models that use ordinary absolute positional encoding for right-to-left text. In addition, the project discussed the effect of employing sub-word units in Arabic dialect translation.
- This study examined the effect of training the suggested model with various numbers of decoders and encoders, as well as multiple attention heads (AH) in the decoders and encoder's MHA sub-layer. In addition, the proposed model was trained in various dimensions for sub-word embedding.

The rest of the paper is organized as follows. The literature review is presented in Section 2. Section 3 presents and discusses the proposed model in detail. Results, experiments and analysis are presented in Section 4. Finally, we discuss the conclusion of the research in Section 5.

2. Literature Review

The majority of MT scientific research focused decades ago on translation for high-resource languages such as Spanish–English, German English, and French–English; the parallel corpora for these high-resource languages are freely available. Research on low-resource language translation has increased over the past four years. Research on translation between written varieties has been applied to SMT models such as Slovenian, Serbian, Croatian [6], colloquial Arabic [7], and Urdu and Hindi [8,9], compared the translation of Catalan to Spanish by applying rule-based MT (RBMT), phrase-based MT, and NMT models. In the evaluation of the in-domain test set, the NMT approach outperformed the other models in terms of performance and quality. The results of practical experiments for the out-of-domain test dataset revealed that the RB model from Spanish to Catalan and PB approaches from Catalan to Spanish produced better translation quality [10], proposed a neural MT model that was trained to perform translation between closely related languages. The authors conducted practical experiments with European Portuguese and Brazilian Portuguese languages as well as a corpus of subtitles for NMT training. The authors obtained an extra 0.9 BLEU points for performing translation from European Portuguese to Brazilian Portuguese when the authors did a comparison to the SMT model that had been trained with the same data. Furthermore, the NMT model gained an extra 0.2 BLEU points when the translation was performed in the reverse direction. These results prove that the NMT model provides a more reliable translation quality than the SMT model using the BLEU score and human evaluation metric. Most studies on Arabic dialect translation have used SMT and rule-based approaches. For example, [11] presented a multidialectal Arabic corpus (PADIC). The PADIC corpus contains MSA and various dialects such as the Levantine dialects (Syria and Palestine) and Maghrebi dialects (Tunisia and Algeria). Different practical experiments have been conducted for several SMT approaches for all ADs and MSA pairs. By experimenting with different smoothing techniques, the authors investigated the impact of the language model (LM) on machine translation by combining

them with a significantly larger one. The most accurate translation results were obtained in the Algerian dialect, which is expected given the lack of similarity between the Algerian dialect and MSA. Consequently, the statistical MT model was unable to capture the semantic and syntactic characteristics of the Algerian dialect. However, the SMT model achieved high performance for the Syrian and Palestinian dialects. The most reliable machine translation results were obtained from the Palestinian dialect and MSA.

All of the aforementioned approaches focused on RBMT and SMT models. To develop an RBMT and SMT model, certain shortcomings must be addressed. For example, considerable time is required to design and develop this model. In addition, it is important to modify the rules to increase the RB machine translation performance, which requires an excellent level of linguistic comprehension. In the case of SMT models, powerful computing devices are required, and SMT cannot deal with syntax problems in the Arabic dialect, which is the word order. Few studies have investigated translation between closely related languages, such as ADs and MSA, using the NMT approach [12], presented the first NMT model that performs the translation from ADs to MSA. The authors introduced a multitasking-based NMT model for the translation from ADs to MSA. The proposed model is based on multi-task learning, where each source language has its encoder, and a single decoder is shared by all language pairs. The results of actual experiments demonstrate that the suggested multi-task-based NMT model can produce high-quality MSA phrases and learn the predictive data of various targets at the same time by using a limited amount of training data. Among many other approaches for translating colloquial Arabic, there is a need for the incorporation of external knowledge, such as part-of-speech tags, into NMT models for ADs [13], introducing a multi-task NMT model that shares an encoder for two different tasks: translating ADs into MSA and POS at the segment level. The proposed model shares two layers between translation tasks: the shared layer and the invariant layer. By allowing the proposed model to perform alternative training among the POS and translation tasks, the proposed model exploits the exceptional information and produces better translation quality from AD to MSA. Experiments were conducted for the translation of Levantine to MSA and from Maghrebi to MSA. A few publications have employed sub-word units in machine translation for Arabic texts [14] and used the romanization approach to convert Arabic text into sub-word units. The researchers discussed the influence of this approach on NMT in different segmentation settings. In addition, the researchers measured the findings to approach the trained MSA. Furthermore, the researchers made the romanized Arabic text as an input for the Arabic-sourced NMT in comparison to well-known components such as lemma, POS tags, and morph characters. The results of practical experiments conducted on Arabic-Chinese translation demonstrated that recommended approaches handled the unidentified word issue and increased the quality of translation for the Arabic-Chinese task. Before conducting the practical experiments, the authors performed a preprocessing phase for the text. Furthermore, Chinese-to-Arabic experiments were conducted using the NMT model.

Among the several approaches for the translation of Arabic dialects, exploiting the word-piece model to generate sub-word units and use them as input features for the transformer model [15], presented the first transformer-based NMT model that uses sub-word units to translate from AD to MSA. The researchers used sub-word units and the shared vocabulary for the translation from AD to MSA to improve the behavior of the self-attention sub-layers for the encoder by obtaining the general dependencies among words of the AD input sentence. Practical experiments were conducted on five ADs: the Levantine, Maghrebi, Iraqi, Nile, and Gulf dialects. Experimental results confirm that the proposed model addresses the unknown word problem and improves translation performance [5] presenting a transformer model that used absolute positional encoding based on sinusoidal functions. The PE is summed with the input embedding at the bottom of the decoder and encoder [16] presented an approach in which multi-head attention was modified to effectively calculate the representation of relative positions or distances among tokens. Information about the relative position is included by adding vectors

that represent the pairwise relationships among the current positions and other positions. Researchers have performed experiments based on this approach [17] and mentioned in their research that absolute position encodings do not contain information to recognize positional differences that lead to performance loss. To solve this problem, the researchers performed a derivation different from that in [16] and tested their approach on a language modeling task. Transformer models are Seq-to-Seq structures that are able to perform mapping for the speech inputs to translations. This mechanism for modeling positions in this model was tailored for text modeling, so it is less ideal for acoustic inputs [18], adapted the relative position encoding to the speech transformer, where the key addition is the relative distance between input states in the self-attention network. The model was able to utilize the synthetic data better than the transformer and adopts better variable sentence segmentation quality for speech translation.

3. The Proposed Reverse Positional Encoding Multi-Head Attention NMT Model

This study developed an MHA-based NMT model to execute the translation from ADs to MSA. The proposed model is based on a new mechanism, which is the reversal of absolute positional encoding. Furthermore, the proposed model uses ADs and MSA sub-word units. The proposed model was developed based on a transformer model recently introduced by [5]. As shown in Figure 1, the proposed model is a model with an encoder and decoder, and this model has a remarkable structure depending on MHA, reversed positional encodings, and sub-word units [19,20]. For the proposed model, as depicted in Figure 1, the decoder and the encoder are made up of a bundle of various layers. Each layer consists of two distinct sublayers: a multi-head attention sublayer and a position-wise feed-forward sublayer (FNN). The decoder and encoder in the presented framework used the self-attention sublayer and the feed-forward sublayer to generate variable-length sequences without the need for CNN or RNN units. The proposed model encodes Arabic dialect source sentences as an intermediate representation using MHA and decodes them using the MHA and encoder-decoder attention. The encoder in the proposed model converts the input sequence (x_1, \dots, x_n) into a vector representation sequence, $Z = (z_1, \dots, z_n)$. Considering Z , the decoder in the proposed model generated an output sequence (y_1, \dots, y_b) . The embedding layer in the decoder and encoder transforms the input tokens (source tokens in the encoder and target tokens in the decoder) into vectors of dimension d_{model} . As the information required for proximity among tokens is viewed in the MHA sublayer, the information for the token's position is embedded using a new reverse positional encoding (RPE) mechanism. In particular, RPE provides a matrix representing the position information of tokens in a right-to-left sentence, and the proposed model adds RPE to the embedding matrix of the input tokens. Every component of the RPE is calculated using sine and cosine functions of varying frequencies.

$$\text{RPE}(pos, 2i) = \sin\left(pos/10,000^{2i/d_{model}}\right),$$

$$\text{RPE}(pos, 2i + 1) = \cos\left(pos/10,000^{2i/d_{model}}\right),$$

where pos denotes the position of every input token, i represents the dimension of every element, and d_{model} denotes the embedding dimensions of the input tokens. The embedding matrix with RPE is the input to the first layer of the encoder or decoder. The encoder layer is divided into two sublayers: a self-attention sublayer and a wise, fully connected feed-forward network sublayer (FFN). The decoder layer is divided into three sublayers: masked MHA, encoder-decoder MHA, and FNN. The residual connection suggested in [21] was implemented in the sublayers, followed by layer normalization [22]. The output for each sublayer is $\text{LayerNorm}(x + \text{Sublayer}(x))$, where $\text{Sublayer}(x)$ is the output of the original sublayer. The encoder decoder uses the MHA mechanism. The MHA calculates h dot-product attention after mapping three input vectors linearly,

$q, k, v \in R^{d_{model} \times d_k}$ ($i = 1, \dots, h$), where d_{model} denotes the dimension of the input vectors and $d_k = d_{model} / h$. Every dot-product attention is referred to as the head (H_i ($i = 1, \dots, h$))

$$H_i = \text{Attention}(q', k', v'), \tag{1}$$

$$\text{Attention}(q', k', v') = \text{Softmax}\left(\frac{q'k'^T}{\sqrt{d_k}}\right)v', \tag{2}$$

$$q' = qW_i^Q, k' = kW_i^K, v' = vW_i^V. \tag{3}$$

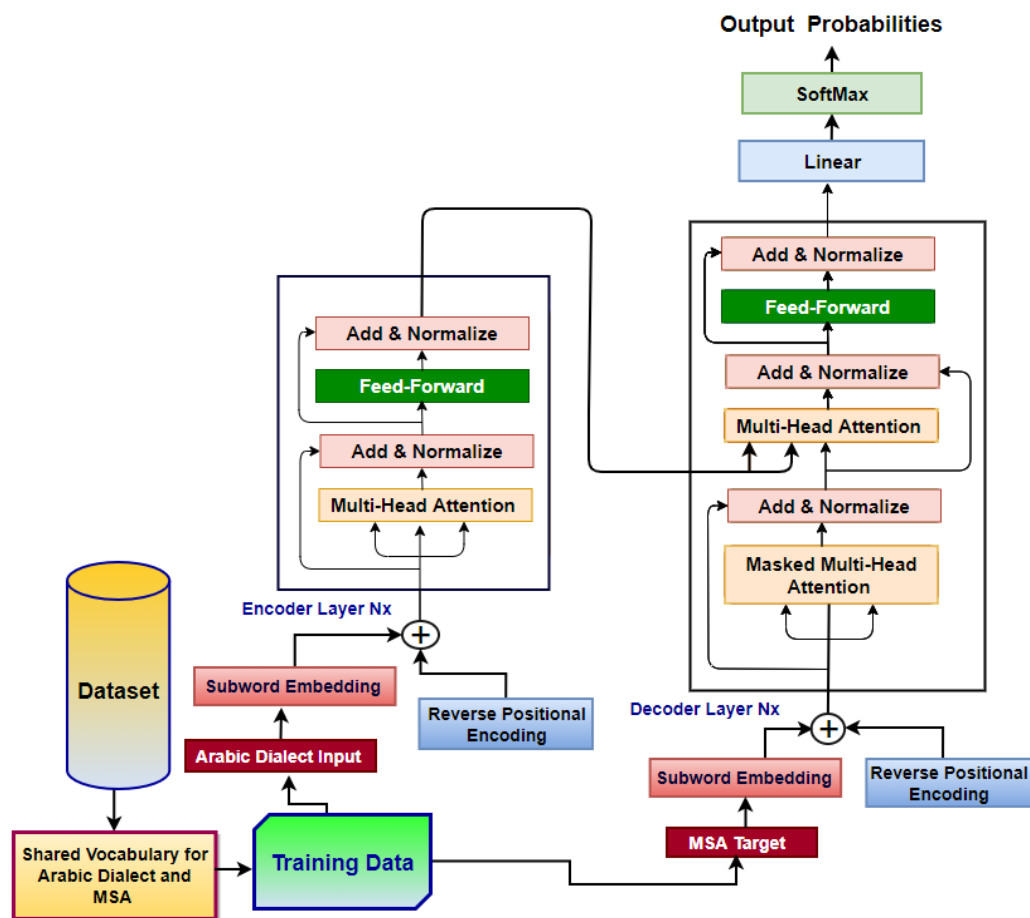


Figure 1. The proposed reverse positional encoding multi-head attention-based NMT model for Arabic dialects.

Therefore, the MHA linearly performs mapping to the concatenated heads with a parameter matrix, $W^o \in R^{d_{model} \times d_k}$

$$\text{MultiHead}(q, k, v) = \text{Concat}(H_1, \dots, H_h)W^o. \tag{4}$$

Equation (4) is computed using the encoders' MHA by substituting the encoder's intermediate states, x_1, \dots, x_n , for q, k, v . Particularly, every head calculates the weighted sum below.

$$z_i = \sum_{j=1}^n \alpha_{ij}x_jW^V, \tag{5}$$

where z_1, \dots, z_n denote the outputs of the MHA. Every coefficient α_{ij} , was calculated using softmax:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^n \exp(e_{ik})}, \tag{6}$$

where e_{ij} is computed as follows:

$$e_{ij} = \frac{(x_i W^Q)(x_j W^K)^T}{\sqrt{d_z}} \quad (7)$$

where d_z denotes the dimensions of z_i . The decoder's MHA calculates Equation (4) by adding the decoder's intermediate states q, k, v . It is unlikely that the decoder will acquire knowledge about the words that will be formed later when forecasting a word during the inference phase; only the intermediate state of the sub-sequence that was created can be used for the MHA. Thus, the masked MHA is presented to the decoder's MHA to avoid calculating the MHA between the forecasted word and the next words. The masked MHA is calculated by modifying Equation (7) as follows:

$$e_{ij} = \begin{cases} \frac{(x_i W^Q)(x_j W^K)^T}{\sqrt{d_z}} & (i \geq j), \\ -\infty & (\text{otherwise}). \end{cases} \quad (8)$$

The coefficient indicating the strength of the connection of both a specific word and the word positioned behind it ($i < j$) is zero, and it can be controlled such that the relation is not considered. Consequently, Equation (6) was modified.

$$\alpha_{ij} = \begin{cases} \frac{\exp(e_{ij})}{\sum_{k=1}^n \exp(e_{ik})} & (i \geq j), \\ 0 & (\text{otherwise}). \end{cases} \quad (9)$$

For the encoder-decoder attention, the intermediate state of the encoder was utilized for q , and the outputs of the encoder were utilized for k, v . The FNN for input x is calculated as follows:

$$\text{FNN}(x) = \max(0, xW_1 + b_1)W_2 + b_2, \quad (10)$$

where $W_1 \in R^{d_{model} \times d_{ff}}$, $W_2 \in R^{d_{ff} \times d_{model}}$ are parameter matrices, and b_1, b_2 are biases.

3.1. Reverse Positional Encoding (RPE)

The essential building blocks of any language are word positioning and order. If the words are re-ordered, the meaning of the whole sentence is changed. To implement natural language processing answers, the built-in function in the RNN can handle the order of sequences. The sentence is parsed sequentially, word by word. This mechanism combines the order of words with the backbone of RNNs. Nonetheless, the transformer model does not utilize the RNN or CNN in preference for self-attention and manages each data point independently. The removal of the RNNs layer will accelerate the training time, and long dependencies in the sentence will be captured. For each word in a sentence concurrently passing through the encoder/decoder of the transformer architecture, the model has no idea of the location or order of the words. Accordingly, there is still a need for a mechanism to include the order of the words in the model; one solution is to provide the model with some insight into order by adding a piece of information to each word about its location in the sentence. We name this "piece of information" the reverse positional encoding (RPE). RPE is a method through which information regarding the order of objects in a sequence is maintained. The RPE represents the location of an entity in a sequence such that each position is given a unique representation. Our proposed MHA-based NMT model introduces a new mechanism in which reverse positional encoding is used to encode order dependencies between words in an AD sentence, as shown in Figure 2. Arabic dialect text is considered a right-to-left text, in which the first word starts on the right side of the sentence and the last word ends on the left side of the sentence. Given the embedding

sequence of the AD source sentence with length J , $X = \{x_1, \dots, x_J\}$, where x_1 represents the embedding of the last word in the AD sentence, and x_J represents the first word in the AD source sentence. The reverse optional embedding is calculated depending on the position of each word using Equation (11).

$$\text{Rpe}(j, 2i) = \sin\left(j/10000^{2i/d_{\text{model}}}\right), \text{Rpe}(j, 2i + 1) = \cos\left(j/10000^{2i/d_{\text{model}}}\right), \quad (11)$$

where j stands for the position index of the word in the sentence and i stands for the dimension number for the position index. Thus, a reverse positional embedding (RPE) exists.

$$\text{RPE} = \{\text{Rpe}_j, \dots, \text{Rpe}_1\}. \quad (12)$$

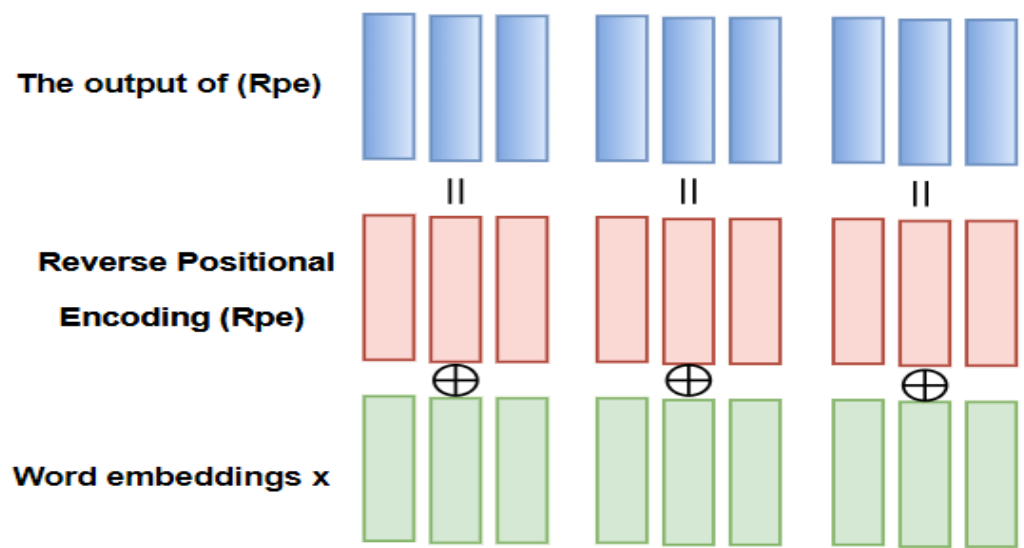


Figure 2. Reverse positional encoding mechanism.

Each Rpe_j is summed to the corresponding word embedding x_j as united embedding v_j :

$$v_j = x_j + \text{Rpe}_j, \quad (13)$$

Finally, a sequence of embeddings $\{v_1, \dots, v_J\}$ is sentence representation that will be initialized H^0 . Later, H^0 will be fed into the self-attention sublayer to learn the representation of the sentence. For humans to translate a sentence, they rearrange word orders depending on the original sentence’s overall semantics and context to obtain a single synonymous sentence that is easier to understand and translate. Note that reverse positional encoding in the proposed model needs to consider the information for the target language and source language, and it is applied to both the encoder and decoder. A detailed description of RPE mechanism is shown in Figure 3.

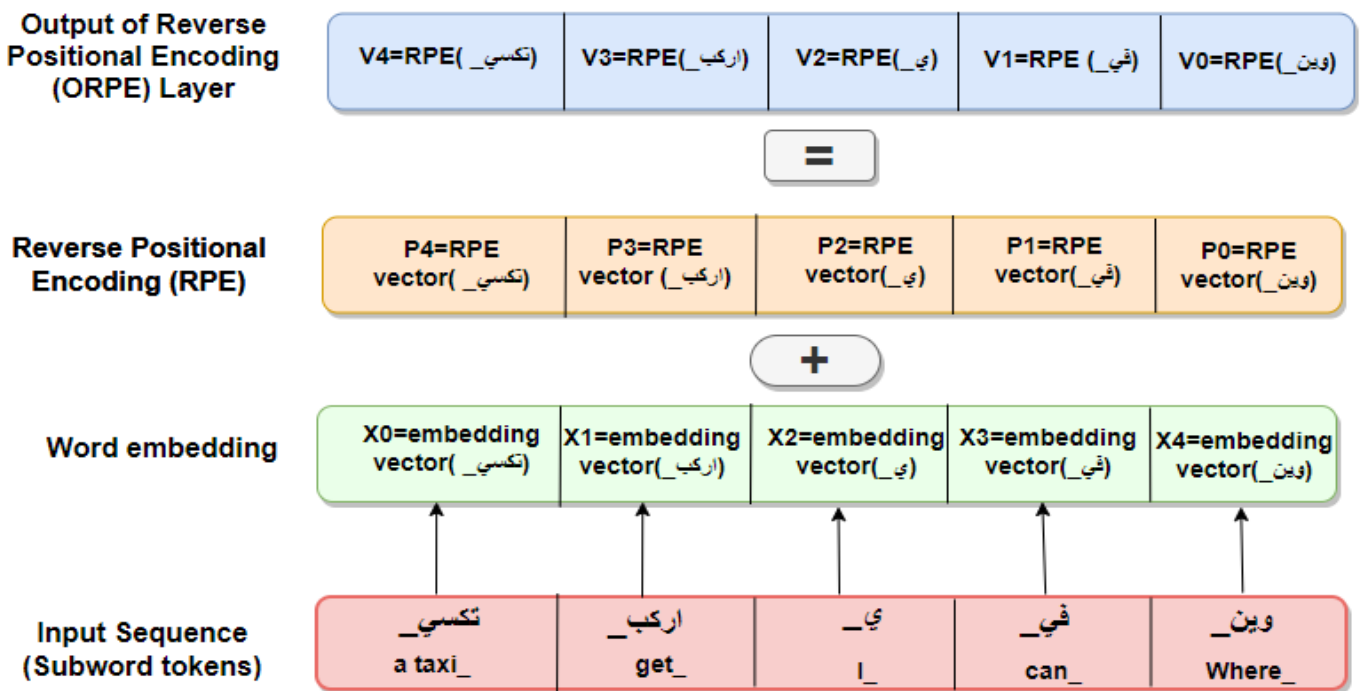


Figure 3. A detailed structure of the reverse positional encoding mechanism.

3.2. Segmentation of Arabic Dialect Words

Segmenting Arabic dialect words into sub-word units is considered the best method among several methods for solving translation challenges from Arabic dialects to MSA. This study used the word-piece model (WPM) implementation. This model ensures that all Arabic dialect sentences are segmented using a standardized method. The process begins by dividing the words into word-pieces using a trained WPM model. Before training the word-piece model, word boundary marks were added to ensure that the original word sequence was extracted without the vagueness of the word-piece sequence. This model produces a word-piece sequence, which is then re-shaped to an identical word sequence in the decoding stage. The following example depicts the word and equal word-piece sequence for a sentence in the Iraqi Arabic dialect.

Sentence: "احس ببروده ومعدتي تاذيني كلش"

Sentence Translation: "I feel cold and my stomach hurts so much "

Word-pieces: "احس_ ببرود_ه_ ومعدت_ي_ تاذين_ي_ كلش"

As demonstrated in the preceding example, the Iraqi Arabic word "ومعدتي" "my stomach" is broken down into two-word chunks "ومعدت" a noun-derived particle from "ومعدتي" and "ي" a suffix-derived particle from "ومعدتي". The word "ببروده" "cold" is broken down into two-word chunks "ببرود" a noun-derived particle from "ببروده" and "ه" a suffix-derived particle from "ببروده". The Iraqi word "تاذهيني" "hurts me" is broken down into two-word chunks "تاذهين" a noun-derived particle from "تاذهيني" and "ي" a suffix-derived particle from "تاذهيني". The remaining words were saved as single-word chunks. The word-piece strategy is accomplished using a data-driven technique that maximizes the language model likelihood for training data and provides a word depiction. Given a set of tokens S and a training corpus, the optimization is completed by assigning S-word chunks in a manner in which the final corpus includes the fewest word chunks when segmented by the word-piece mechanism. Instead of two ends, a token was added at the beginning of the words. Depending on the data, the number of primary characters was reduced to a variable number. In addition, the remaining characters were assigned to a specific anonymous alphabet to

avoid joining the word-piece vocabulary with separate characters. An outstanding BLEU score result will be obtained along with a rapid decoding phase for all language pairs that have been assessed when utilizing a vocabulary between 24,000 and 100,000-word chunks. In translation, it is effective to directly copy infrequent terms or digits from the source language to the target language. To facilitate this type of direct copying, we utilized a shared word-piece strategy for AD and MSA. By employing this mechanism, the exact string is segmented in the same manner in the source and target sentences, making it easier for the model to copy these tokens. Word chunks achieve a balance between the efficiency of words and the adjustability of alphabets. This inspiration can be summarized in two ways. The first fact demonstrates that processing the AD and MSA is achieved by employing a shared vocabulary procedure. The vocabulary was shared between the encoder and decoder of the proposed RPE MHA-based NMT model. By exploiting the shared vocabulary procedure between both the decoder and encoder, the RPE MHA-based NMT model replaces the terms in the input source sentence with the translation terms in the target sentence. The second fact states that the shared vocabulary improves the alignment of the embedding vectors. The RPE MHA-based NMT model gained outstanding BLEU results utilizing word chunks, which is due to the suggested model's capability to handle unlimited vocabulary without depending on characters.

4. Experiments

Different practical experiments have been conducted to assess the suggested MHA-based NMT model that utilizes the reverse positional encoding mechanism for various translation tasks. The proposed reverse-positional encoding-based MHA NMT model was evaluated to measure its translation performance from ADs to MSA. Practical experiments were applied to five different ADs: Maghrebi, Levantine, Gulf, Nile Basin, and Iraqi. The Maghrebi dialect is widely used in Algeria, Morocco, Tunisia, and Libya. The Levantine dialect is commonly used in Jordan, Lebanon, Syria, and Palestine. The Nile Basin Arabic is widely used in Egypt, South Sudan, and Sudan. Gulf Arabic is widely used in United Arab Emirates, Saudi Arabia, Kuwait, Oman, Qatar, Bahrain, and Yemen. Iraqi Arabic is popular in Iraq.

4.1. The Data

Regarding the translation tasks, this study used the same dataset applied by Baniata et al. [15]. This dataset has five parallel corpora: the PMM-LEV corpus for Levantine Arabic-MSA, PMM-MAG corpus for Maghrebi Arabic-MSA, MADAR_Nile dataset for Nile Basin Arabic-MSA, MADAR-Gulf dataset for Gulf Arabic-MSA, and MADAR-Iraqi corpus for Iraqi Arabic MSA. The multi-head attention-based NMT model that uses the RPE mechanism was trained on 54,736 sentence pairs for the Maghrebi dialect, 36,850 sentence pairs for the Levantine dialect, 18,000 sentence pairs for the Nile Basin dialect, 18,000 sentence pairs for the Gulf dialect, and 5000 sentence pairs for the Iraqi dialect. Corpora content has been collected from different sources, such as social media, movies, and dramas. For the test dataset, the proposed RPE MHA-based NMT model was tested on 3000 sentence pairs for the Levantine dialect, 3000 sentence pairs for the Maghrebi dialect, 2000 sentence pairs for the Nile Basin dialect, 2000 sentence pairs for the Gulf dialect, and 1000 sentence pairs for the Iraqi dialect. Furthermore, the proposed model was trained with 17,736 sentence pairs for the Maghrebi dialect and 13,805 sentence pairs for the Levantine dialect on different dataset used by Baniata et al. [13]. The proposed model was tested on 2000 sentence pairs for the Maghrebi dialect and 2000 sentence pairs for the Levantine dialect on the dataset used by Baniata et al. [13]. The corpus for each AD was split into two sets: 80% for the training dataset, and 20% for the test dataset. In addition, every AD test dataset was obtained from a familiar domain. The datasets used in this project consisted of unprocessed information, which might affect the quality of the suggested model. Further all ADs and MSA sentences were pre-processed. In addition, English alphabets, punctuation, hashtags, and diacritics were eliminated in all ADs and MSA texts. The orthographic normalization

process was applied to all AD and MSA texts. MSA sentences are shorter than AD sentences, and MSA sentences have more diverse tokens than AD sentences. To calculate the expected success rate of a predictor, three mechanisms can be utilized: cross-validation, independent dataset test, and jackknife test. In this study, the K-fold cross-validation approach was used, where K was set to 2 to create a train/test split to assess the RPE MHA NMT model. To prevent model overfitting, we applied an early stopping option with a patience parameter set to three epochs, and the model checkpoint was utilized to save the best weights when evaluating the suggested RPE MHA NMT model.

4.2. The Suggested Model Setup

The suggested RPE multi-head attention NMT model was built using TensorFlow and Keras libraries. The practical experiments for all Arabic dialect translation tasks were executed using cutting-edge configurations, where the sub-word embedding dimension had six values: 1024, 520, 512, 564, 530, and 400. The hidden state has two values, 1024 and 512, and the attention heads have two values, 8 and 4. The position-wise FNN had a filter with 1024 and 564 dimensions. The proposed model was trained on Maghrebi-MSA, Levantine-MSA, Gulf-MSA, Nile-MSA, and Iraqi-MSA tasks. The introduced model consists of an encoder subnetwork with 12, 8, and 4-layers and it consists of a decoder subnetwork with 12, 8, and 4-layers.

4.3. Training and Inference of the Suggested Model

The proposed RPE multi-head attention NMT model was trained for 13,000 iterations where the batch size was 2024 and the maximum length of the sequence was 70-sub-word tokens for the MAG-MSA task. In addition, the suggested model was trained for 13,000 iterations where the batch size was 1164 and the maximum length of the sequence was 70 for the LEV-MSA task. The suggested model was trained such that the batch size was 700, and the maximum length of the sequence was 70 for the Nile-MSA task. In the case of the GULF-MSA translation task, the suggested model was trained such that the batch size was 700, and the maximum length of the sequence was 70. Regarding the IRQ-MSA task, the suggested model was trained such that the batch size was 400, and the maximum length of the sequence was 65. In this research project, the Adam optimizer [23] was applied with ($\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 1 \times 10^{-9}$). Additionally, in accordance with the study by So et al. [1], a single cosine cycle with warm-up was used as the learning rate schedule algorithm. A beam search was used to produce target sentences during the inference stage. A beam size of 6 and a length penalty of 1.1 are applied for the translation tasks: Maghrebi-MSA, Levantine-MSA, Nile_Basin-MSA, Gulf-MSA, Iraqi-MSA, Levantine-MSA (Baniata et al. [13] dataset), and Maghrebi-MSA (Baniata et al. [13] dataset). Furthermore, shared vocabulary for the source and target languages was applied in the proposed model. This research project used 21,000 sub-word vocabularies for the Levantine-MSA, Maghrebi-MSA, NB-MSA, and Gulf-MSA translation tasks. A total of 9235 sub-words were used for the IRQ-MSA translation task. The dataset utilized by Baniata et al. [13] for the Levantine-MSA and Maghrebi-MSA translation tasks contained 29,500 sub-word vocabularies. The suggested model is trained, where the attention dropout and the ReLU dropout values are set to 0.1. The proposed RPE multi-head attention NMT model is fast, requiring 557 s per epoch for the MAG-MSA translation task, 134 s per epoch for the Nile_Basin-MSA task, 134 s per epoch for the GULF-MSA task, 399 s per epoch for the LEV-MSA task, 304 s per epoch for the IRQ-MSA task, 256 s per epoch for the MAG-MSA task (Baniata et al. [13]) and 253 s per epoch for the LEV-MSA task (Baniata et al. [13]). The suggested model was trained to minimize the cross-entropy loss for each translation task. In addition, the proposed model was tested on the Maghrebi-MSA and Levantine-MSA datasets used by Baniata et al. [13].

4.4. Results by Automatic Metric

Several experiments have been conducted for the suggested model using reverse positional encoding, a shared vocabulary, and sub-word units for Arabic dialects. The

proposed RPE MHA-based NMT model was trained with various numbers of attention heads in the multi-head attention sublayer and with various numbers of decoders and encoders to find the most efficient configuration for the proposed model. In addition, RPE MHA-based NMT model was trained using different sub-word-embedding dimensions. The translation performance of the proposed model is assessed using the sacreBLEU automatic metric. This section discusses the assessment of the presented RPE MHA NMT model on five AD translation tasks. The results of the practical experiments for the LEV-MSA and MAG-MSA translation tasks are shown in Tables 1 and 2. The findings of the Gulf-MSA, IRQ-MSA, and Nile-MSA translation tasks are summarized in Tables 3–5. Table 1 shows the results of the proposed RPE MHA NMT model with various parameters on the test dataset for the Levantine-MSA translation task. As seen in Table 1, the RPE MHA-based NMT model obtained an excellent 65.77 BLEU score, where the number of attention heads was eight, the number of encoder and decoder layers was 12, and the sub-word embedding dimension was 512. This high BLEU score was achieved as a result of the positive impact of applying the reverse positional encoding mechanism for the right-to-left text and the proximity between the Levantine dialect and MSA. Furthermore, these excellent results were obtained because Levantine Arabic and MSA share many terms and words. It can be noticed from Tables 1, 6, 7 and 8 that the proposed RPE MHA-based NMT model with reverse positional encoding outperformed the transformer-based model with absolute positional encoding, as suggested by Baniata et al. [15], by +2.06 BLEU score points. As illustrated in Table 1, practical experiments with low dimensions of sub-word embedding obtained better BLEU results when compared to experiments with high dimensions of sub-word embedding.

Table 2 presents the results of the proposed RPE MHA-based NMT model on the test dataset for the MAG-MSA translation task. As shown in Table 2, the proposed RPE MHA-based NMT model can correctly translate Maghrebi Arabic sentences to MSA with a BLEU score of 66.71. Maghrebi Arabic contains vocabularies from many different languages, such as Turkish, Old Arabic vocabulary, Amazigh language, and certain new vocabularies borrowed from French, Italian, and Spanish. Thus, the proposed RPE MHA-based NMT model captures the semantics and syntactic features of the Maghrebi Arabic dialect. The proposed model enhanced the translation quality of the MAG-MSA translation, although there is no proximity between MSA and the Maghrebi Arabic dialect in terms of expressions. Traditional NMT models [12,13] cannot generate fluent MSA sentences for the Maghrebi Arabic dialect because Maghrebi Arabic has many vocabularies from many other languages. By utilizing the RPE mechanism and using sub-word units as an input to the encoder, there will be information sharing between sub-word units and word units, and the proposed model will generate fluent MSA sentences. Tables 2 and 8 show that for the MAG-MSA translation task, the proposed model outperformed the transformer sub-word model proposed by Baniata et al. [15] by +1.05 BLEU score points. This is because of the reverse position encoding mechanism for right-to-left text, such as Arabic dialects in the proposed model. Table 3 presents the results of the proposed RPE MHA-based NMT model with different configurations on the test dataset for the Gulf-MSA translation task. As illustrated in Table 3, the proposed model obtained a BLEU score of 47.77 and the proposed model translated the Gulf Arabic sentences to MSA efficiently. Tables 4 and 5 show the results of the Iraqi-MSA and Nile MSA translation tasks. The proposed RPE MHA-based NMT model generated an outstanding translation quality for Iraqi Arabic sentences with a BLEU score of 58.53 and 49.87 on the Nile-MSA translation tasks. As noted in Tables 3–8, the proposed RPE MHA-based NMT model outperforms the transformer-based NMT model with absolute positional encoding proposed by Baniata et al. [15] for the Gulf-MSA, Iraqi-MSA, and Nile-MSA translation tasks. This led to the conclusion that the reverse positional encoding mechanism introduced in this model is powerful and improves the translation quality for right-to-left texts, such as Arabic dialects. Further, the proposed RPE MHA-based NMT model was applied to the Levantine-MSA and Maghrebi-MSA corpora utilized by Baniata et al. [13]. As shown in Tables 6 and 7, the proposed RPE

MHA-based NMT model obtained BLEU scores of 61.94 and 66.87 on Levantine-MSA and Maghrebi-MSA translation tasks, respectively. Therefore, it can be concluded, as seen in Tables 6, 7 and 9, that the proposed RPE MHA-based NMT model outperforms the multi-task NMT with part of speech (POS) tags model that was proposed by Baniata et al. [13] and the transformer-based NMT model with an absolute positional encoding mode that was proposed by Baniata et al. [15] on the Levantine-MSA and Maghrebi-MSA corpora utilized by Baniata et al. [13]. The results show the efficiency and power of applying the reverse positional encoding (RPE) mechanism and the impact of exploiting the sub-word units and shared vocabulary between AD and MSA in the proposed RPE MHA-based NMT model. Using sub-word units of AD as an input feature to the self-attention sub-layer (multi-head attention) in the proposed model is useful for solving the word order issue and for assisting in producing fluent MSA sentences. Moreover, the proposed model could solve grammatical issues and capture semantic and syntactic features from the AD source language. This is a result of applying the RPE mechanism, using sub-word units as an input feature to the MHA sublayer, and using the vocabulary shared between AD and MSA.

Table 1. Results for RPE MHA NMT model on Baniata et al. [15] Corpus for the Levantine-MSA translation task, where SWED is the sub-word embedding dimension, F is the filter size, E is the encoder layer, D is the decoder layer, and AH is the number of attention heads.

SWED	F	E	D	AH	BLEU
520	1024	4	4	8	63.99
520	1024	4	4	4	62.57
512	1024	8	8	4	64.35
564	564	4	4	4	61.20
1024	1024	4	4	8	60.39
512	1024	12	12	8	65.77

Table 2. Results for RPE MHA NMT model on Baniata et al. [15] Corpus for the Maghrebi-MSA translation task, where SWED is the sub-word embedding dimension, F is the filter size, E is the encoder layer, D is the decoder layer, and AH is the number of attention heads.

SWED	F	E	D	AH	BLEU
520	1024	4	4	8	61.07
520	1024	4	4	4	60.23
520	1024	8	8	4	64.99
564	564	4	4	4	58.56
1024	1024	4	4	8	62.06
512	1024	12	12	8	66.71

Table 3. Results for RPE MHA NMT model on Baniata et al. [15] Corpus for the Gulf-MSA translation task, where SWED is the sub-word embedding dimension, F is the filter size, E is the encoder layer, D is the decoder layer, and AH is the attention heads number.

SWED	F	E	D	AH	BLEU
520	1024	4	4	8	47.21
520	1024	4	4	4	47.49
400	1024	8	8	4	47.77
564	564	4	4	4	47.16
1024	1024	4	4	8	43.88
400	1024	12	12	8	46.42

Table 4. Results for RPE MHA NMT model on Baniata et al. [15] Corpus for the Iraqi-MSA translation task, where SWED is the sub-word embedding dimension, F is the filter size, E is the encoder layer, D is the decoder layer, and AH is the attention heads number.

SWED	F	E	D	AH	BLEU
512	1024	4	4	8	58.53
512	1024	4	4	4	53.00
512	1024	8	8	4	52.75
564	564	4	4	4	53.09
1024	1024	4	4	8	37.83
512	1024	12	12	8	19.09

Table 5. Results for RPE MHA NMT model on Baniata et al. [15] Corpus for the Nile-MSA translation task where SWED is the sub-word embedding dimension, F is the filter size, E is the encoder layer, D is the decoder layer, and AH is the number of attention heads.

SWED	F	E	D	AH	BLEU
520	1024	4	4	8	47.15
520	1024	4	4	4	48.56
400	1024	8	8	4	49.87
564	564	4	4	4	44.61
1024	1024	4	4	8	43.31
400	1024	12	12	8	47.51

Table 6. Results for RPE MHA NMT model on Corpus used by Baniata et al. [13] for Levantine-MSA translation task, where SWED is the sub-word embedding dimension, F is the filter size, E is the encoder layer, D is the decoder layer, and AH is the attention heads number.

SW-E-D	F	E	D	AH	BLEU
520	1024	4	4	8	60.79
520	1024	4	4	4	55.70
400	1024	8	8	4	57.25
564	564	4	4	4	59.19
1024	1024	4	4	8	39.51
400	1024	12	12	8	61.94

Table 7. Results for RPE MHA Neural-MT model on Corpus used by Baniata et al. [13] for the Maghrebi-MSA translation task, SWED is the sub-word embedding dimension, F is the filter size, E is the encoder layer, D is the decoder layer, and AH is the attention heads number.

SWED	F	E	D	AH	BLEU
520	1024	4	4	8	66.13
520	1024	4	4	4	65.51
400	1024	8	8	4	66.87
564	564	4	4	4	63.36
1024	1024	4	4	8	55.64
400	1024	12	12	8	66.65

Table 8. Results of the transformer-based NMT model that was suggested by Baniata et al. [15] for different translation tasks, where SWED is the sub-word embedding dimension, F is the filter size, E is the encoder layer, D is the decoder layer and AH is the number of attention heads.

Pairs	SWED	F	E	D	AH	BLEU
LEV-MSA	512	1024	12	12	4	63.71
MAG-MSA	512	1024	12	12	4	65.66
NILE-MSA	512	1024	8	8	4	48.19
GULF-MSA	512	564	4	4	4	47.26
IRQ-MSA	512	1024	4	4	4	56.50

Table 9. Results of the multi-task NMT model with part-of-speech (POS) tags that was suggested by Baniata et al. [13] and the results of the transformer-based NMT model that was suggested by Baniata et al. [15] for the dataset used by Baniata et al. [13] for the Levantine-MSA and Maghrebi-MSA translation tasks.

Model	Pairs	BLEU
NMT+POS-LEV [13]	LEV-MSA	43.00
NMT+POS-MAG [13]	MAG-MSA	34.00
Transformer-NMT-Sub-word [15]	LEV-MSA	57.92
Transformer-NMT-Sub-word [15]	MAG-MSA	57.85

4.5. Results by Human Evaluation

Human assessment experiments proved the findings obtained through an automatic metric assessment. In this research project, pilot rating experiments (PRE) [15] were conducted. The participants assessed translations from 1 to 7 on a Likert scale. The performance of translation was evaluated for the MAG-MSA, LEV-MSA, Gulf-MSA, IRQ-MSA, and Nile-MSA translation tasks. We requested seven speakers who understood MSA and every AD to rate the sentences that were produced from the suggested RPE MHA NMT model. Several texts in AD, such as LEV, Gulf, IRQ, Nile, MAG, and one translation in MSA for each AD were given to the speakers. We chose 100 texts randomly and separated them into five subsets of 20 texts. We offered every annotator a subset and requested them to assess the interpretations on a Likert scale of 1 to 7, taking fluency into account. The results generated by the three different models applying PRE are shown in Tables 10 and 11. The average results showed that the seven native speakers had a real view of the translations produced by the proposed RPE MHA-based NMT model. As listed in Table 10, the average scores for the LEV-MSA, MAG-MSA, Gulf-MSA, Nile-MSA, and IRQ-MSA translation tasks obtained by the proposed RPE MHA-based NMT model were 6.46, 6.40, 5.95, 5.90, and 6.39, respectively. As seen in Table 11, the average score on the LEV-MSA and MAG-MSA translation tasks (Baniata et al. [13] dataset) obtained by the multi-task NMT POS model suggested by Baniata et al. [13] was 5.9 and 4.4, respectively. The average score for the LEV-MSA and MAG-MSA translation tasks (Baniata et al. [13] dataset) obtained by the transformer –NMT sub-word model was 6.0 and 6.2, respectively. The average score on the LEV-MSA and MAG-MSA translation tasks ((Baniata et al. [13] dataset) obtained by the proposed RPE MHA-based NMT model was 6.2 and 6.5, respectively. Furthermore, as shown in Table 10, the average scores for the LEV-MSA, MAG-MSA, Gulf-MSA, Nile-MSA, and IRQ-MSA translation tasks obtained by the transformer-NMT sub-word model were 6.35, 6.3, 5.85, 5.8, and 6.10, respectively. The results of PRE prove that the suggested RPE MHA NMT model produces translation performance significantly better than the transformer-NMT sub-word model, and better than the multi-task NMT with POS model for all translation tasks.

Table 10. Human Evaluation Score: PRE.

Model	Pairs	Average Score
Transformer-NMT-Sub-word [15]	LEV-MSA	6.35
Transformer-NMT-Sub-word [15]	MAG-MSA	6.30
Transformer-NMT-Sub-word [15]	Gulf-MSA	5.85
Transformer-NMT-Sub-word [15]	Nile-MSA	5.80
Transformer-NMT-Sub-word [15]	IRQ-MSA	6.10
Proposed REP-MHA-NMT	LEV-MSA	6.46
Proposed REP-MHA-NMT	MAG-MSA	6.40
Proposed REP-MHA-NMT	Gulf-MSA	5.95
Proposed REP-MHA-NMT	Nile-MSA	5.90
Proposed REP-MHA-NMT	IRQ-MSA	6.39

Table 11. Human Evaluation Score: PRE. LEV-MSA and MAG-MSA.

Model	Pairs	Average Score
Transformer-NMT-sub-word [15]	LEV-MSA	6.0
Transformer-NMT-sub-word [15]	MAG-MSA	6.2
MTL-NMT [13]	LEV-MSA	1.4
MTL-NMT [13]	MAG-MSA	1.3
MTL-NMT + POS [13]	LEV-MSA	5.9
MTL-NMT + POS [13]	MAG-MSA	4.4
Proposed REP-MHA-NMT	LEV-MSA	6.2
Proposed REP-MHA-NMT	MAG-MSA	6.5

4.6. Discussion and Analysis

This section shows the positive impact of using reverse positional encoding in the Arabic dialect on MSA translation tasks. Table 12 presents a sample of translations from the proposed RPE MHA NMT model for MAG-MSA, Gulf-MSA, LEV-MSA, Nile MSA, and IRQ-MSA. Because AD lacks standardization, traditional NMT models do not perform translation for certain parts of the AD input source sentences correctly. In addition, clitics and affixes in AD text cannot be captured without using reverse positional encoding or exploiting sub-word units. As shown in Table 12, the suggested RPE MHA NMT model successfully translated the Maghrebi Arabic sentences to MSA. For instance, the suggested model aligned the MAG AD words “انا متكرهش الجزائر” “I don’t hate Algeria” to the MSA words “انا لا اكره الجزائر” accurately. For LEV Arabic sentences, the suggested RPE MHA NMT model properly translated the source sentence without any mistakes. Furthermore, the proposed RPE MHA NMT model generated the best translation performance for the Gulf Arabic and Iraqi Arabic sentences. For Nile Basin Arabic, the proposed RPE MHA NMT model translated 99% of the Nile-Basin AD sentence fluently with the same meaning even adding the term “تناول” “eating” to the generated sentence which is absent in the reference MSA sentence. In addition, the proposed RPE MHA-based NMT model improved the translation quality and performance when compared to the transformer-based NMT sub-word model suggested by (Baniata et al. [15]) and experimented on the same dataset, as shown in Table 12. As noted in Table 12, the suggested RPE MHA NMT model showed outstanding translation quality for the Levantine and Maghrebi dialects to MSA. The positive impact of applying reverse positional encoding, the shared vocabulary between MSA and AD, and the exploitation of sub-word units are obvious. The proposed RPE MHA NMT model is suitable for dealing with right-to-left text and free word order text, such as Arabic dialect, and to generate fluent MSA sentences, as shown in Tables 11 and 12. The suggested RPE MHA NMT model was tested on the same dataset that Baniata et al. [15] experimented with, for Levantine AD, Maghrebi AD, Gulf AD, Nile-Basin AD, and Iraqi AD. As shown in Figure 4, compared to the transformer-based NMT model that has been trained on the same parallel corpora, the suggested RPE MHA NMT model scored

66.71 BLEU results for performing translation from Maghrebi to MSA, 65.77 results for performing translation from Levantine to MSA, 49.87 for translation from Nile Arabic to MSA, 47.77 for translation gulf Arabic to MSA, and 58.53 for translation from Iraqi Arabic to MSA. The results illustrate that the suggested RPE MHA NMT model generates better translation performance than the transformer NMT model by assessing the models using the BLEU score and pilot rating experiment. The proposed RPE MHA-based NMT model was assessed on the same dataset that was experimented by Baniata et al. [13] for the Levantine-MSA and Maghrebi-MSA translation tasks. The proposed RPE MHA-based NMT model obtained outstanding BLEU scores when compared to different NMT models, as shown in Figures 5 and 6. The proposed RPE MHA NMT model can generate correct MSA sentences carrying information about verbs, subjects, and objects in Arabic dialects. This section provides a practical analysis to demonstrate the influence of reverse positional encoding on the suggested MHA NMT model. This analysis includes (1) the effect of using various numbers of encoder layers on the translation performance, (2) the influence of using different source lengths on the translation quality, (3) the impact of using varying beam sizes on translation quality, (4) the influence of MHA of the encoder, and (5) quantitative analysis of RPE MHA NMT. This analysis was conducted on MAG-MSA tasks owing to the size of the dataset and the number of layers that were utilized to train the proposed model.

Table 12. Samples of translations generated by the proposed RPE MHA-based NMT model.

Source Language: MAG (Maghrebi)	انا متنكرهش الجزائر ولكن تنكره الفساد لي كين فيها
English Translation (MAG)	I don't hate Algeria but I hate the corruption that is inside it
Target Language: MSA	انا لا اكره الجزائر ولكن اكره الفساد الذي يملأها
RPE MHA-NMT Model	انا لا اكره الجزائر ولكن اكره الفساد الذي يملأها
English translation for the output of the RPE MHA-NMT Model	I don't hate Algeria but I hate the corruption that fills it
Source Language: LEV (Levantine)	زي ما نقول بس ما اعتقدش انك تلاقي شغل في مجال تخصصك
English Translation (LEV)	As we say but I don't think you gonna find a job in your major area
Target Language: MSA	كما نقول لكن لا اعتقد انك ستجدين عملا في مجال تخصصك
RPE MHA-NMT Model	كما نقول لكن لا اعتقد انك ستجدين عملا في مجال تخصصك
English translation for the output of the RPE MHA-NMT Model	As we say, I do not think that you will find a job in your field of specialization
Source Language: GULF	عندك نسبة تخفيض حق الاعضاء ؟
English Translation (GULF)	Do you have discounts for members?
Target Language: MSA	هل ثمة تخفيضات للاعضاء المشاركين ؟
RPE MHA-NMT Model	هل ثمة تخفيضات للاعضاء المشاركين ؟
English translation for the output of the RPE MHA-NMT Model	Are there discounts for participating members?
Source Language: Nile (Egypt, Sudan)	با فضل الاكل الايطالي
English Translation (Nile)	I like Italian food
Target Language: MSA	افضل الطعام الايطالي
RPE MHA-NMT Model	افضل تناول الطعام الايطالي
English translation for the output of the RPE MHA-NMT Model	I like eating Italian food
Source Language: IRQ (Iraq)	احس بضيق وصدري ياذيني كلش
English Translation (IRQ)	I feel tightness and my chest is hurting me a lot
Target Language: MSA	اشعر بضيق والم شديد في الصدر
RPE MHA-NMT Model	اشعر بضيق والم شديد في الصدر
English translation for the output of the RPE MHA-NMT Model	I feel tightness and severe pain in the chest

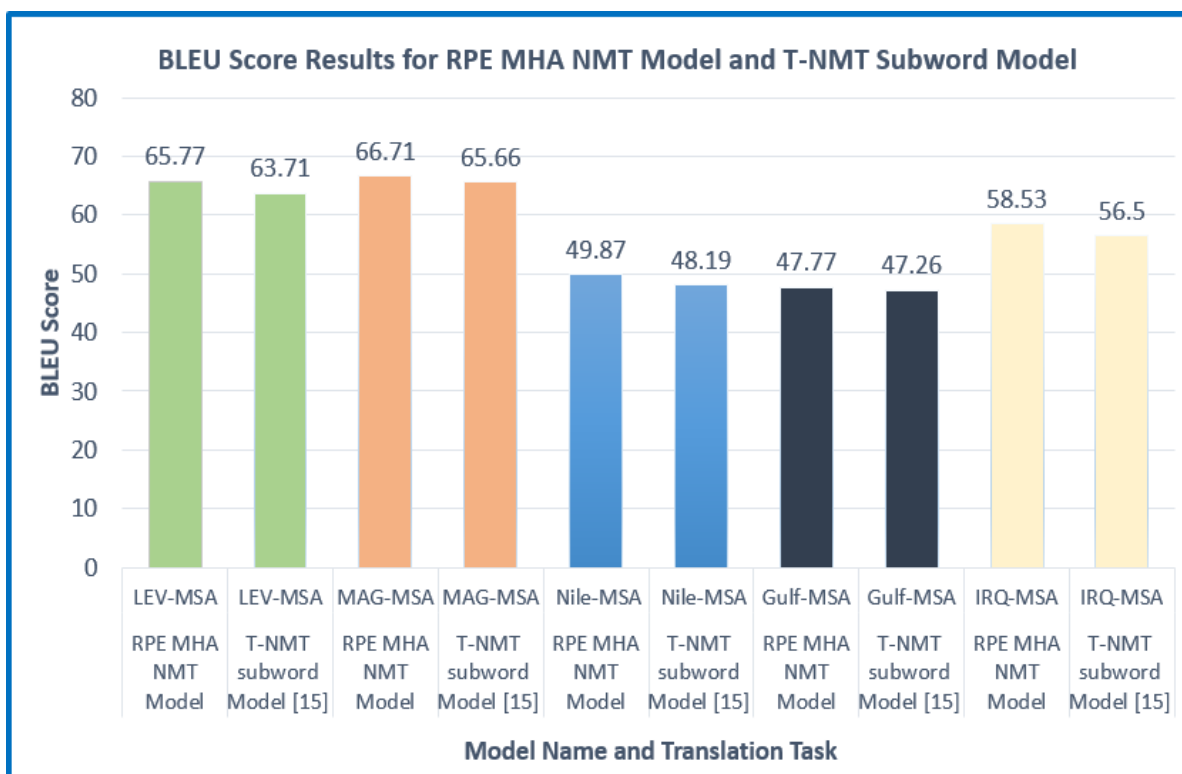


Figure 4. Comparison of BLEU score results for different Arabic dialect pairs.

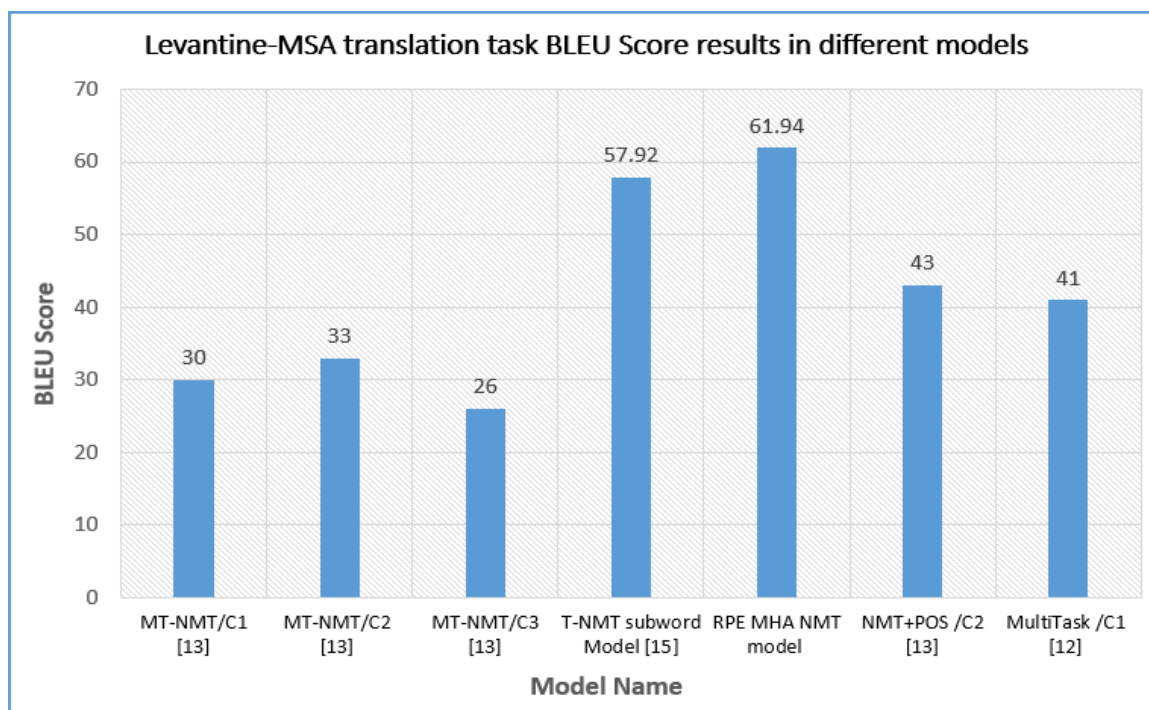


Figure 5. Comparison of BLEU score results of different models for the Levantine-MSA translation task, C1 random embedding, C2 pre-trained/Fast-text, C3 pre-trained/polyglot.

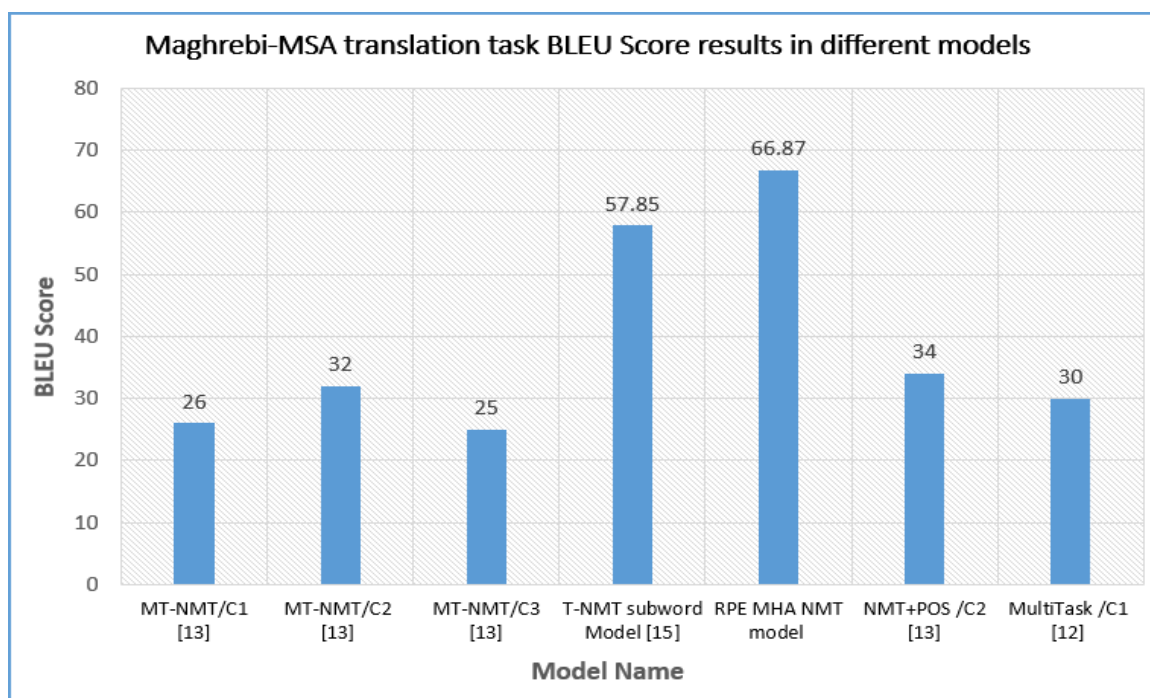


Figure 6. Comparison of BLEU score results of different models for the Maghrebi-MSA translation task, C1 random embedding, C2 pre-trained/Fast-text, C3 pre-trained/polyglot.

4.7. Impact of Encoding Layers V

As observed in Tables 1–7, the effectiveness of the suggested NMT model with reverse positional encoding across different source representations captured from different encoding layers improved the effectiveness of the proposed model for all Arabic dialects, where V is the number of layers in the encoder in the RPE MHA NMT model. The proposed model was trained using three different values of n : 4, 8, and 12. When V is set to 12, it denotes a large model configuration, $V = 8$ denotes a medium-sized model configuration, and $V = 4$ denotes a small-size model setup. As illustrated in Table 1, in the LEV-MSA translation task, there was a noticeable change in the BLEU results as V varied.

4.8. Length of Source Sentence

Capturing long-distance dependencies and contextual information among the tokens in the source sentence improves the translation effectiveness of long sentences. Sentences with the same lengths (in terms of source tokens) have been categorized together, as by Luong et al. [24]. The MAG-MSA translation task was selected for investigating the translation performance for long sentences because of the large size of the Maghrebi Arabic MSA dataset. The evaluation in this section was based on the following lengths: >50, 40–50, 30–40, 20–30, 10–20, and <10. The BLEU score metric was computed for the output of the suggested RPE MHA NMT model. As illustrated in Table 13, the effectiveness of the suggested RPE MHA-based NMT model increases as the input sentence length increases, especially for lengths between 30 and 40 sub-word tokens, lengths between 40 and 50 sub-word tokens, and lengths larger than 50 sub-word tokens with BLEU scores of 65.73, 65.95, and 66.71, respectively. By applying reverse position encoding to a new mechanism, exploiting sub-word units as an input feature to the self-attention sub-layer, and the use of shared vocabulary, the proposed model captures contextually relevant information and dependencies within tokens regardless of their position in the AD sentence input. The effectiveness of the proposed model decreased for short sentences with lengths ranging from 10 to 20 sub-word tokens and lengths less than 10 sub-word tokens. Furthermore, the effectiveness of the proposed model was poor for sentences with fewer than 10 sub-word tokens and obtained a low BLEU score of 24.49. For this reason, short AD sentences have

sub-word tokens, which are morphemes, suffixes, and affixes. These types of sub-word tokens are not easily aligned with the corresponding words in the target language. The outstanding effectiveness of the suggested RPE MHA NMT model obtained for various sentence lengths shows that employing reverse positional encoding and utilizing the sub-word units as an input feature enhances the encoder's MHA sub-layer performance in obtaining word relationships in the AD input phrase.

Table 13. BLEU score on MAG-MSA test dataset with different Sentences' Lengths

Sentence Length	BLEU
<10	24.49
(10–20)	42.89
(20–30)	61.69
(30–40)	65.73
(40–50)	65.95
>50	66.71

4.9. Beam Size Evaluation

The effectiveness of the RPE MHA NMT model was evaluated by varying the beam size. Practical experiments were conducted using the proposed model for all Arabic dialects. We used the word-piece model for sub-word tokenization. Beam size has a positive effect on the decoding speed and translation quality in BLEU [25]. According to Table 14, the outcomes of the experiments on the MAG-MSA translation problem demonstrate that the RPE MHA NMT model has the best performance when the beam size is 7.

Table 14. BLEU score on MAG-MSA test dataset with different Beam sizes

Beam Size	BLEU
1	55.69
2	60.87
3	62.67
4	63.18
5	63.88
6	64.99
7	65.10
8	63.93
9	64.64
10	64.59

4.10. The Impact of the Encoder Self-Attention

The performance of the encoder layers is determined by the ability of the heads in the MHA sublayer inside every layer to obtain contextual information. Attention heads capture contextual information to different degrees. Some heads in the MHA sublayer maintain long-distance relationships between the input tokens. In addition, there are other heads in the MHA sublayer that maintain a short-distance relationship between the input tokens as pointed out by Raganato et al. [26] and Vig et al. [27]. Therefore, the proposed RPE MHA-based NMT model captures the semantic and syntactic characteristics of Arabic dialect source sentences [26]. The utilization of the new reverse positional encoding mechanism affects how information in the source language is managed by encoder layers. According to Vig et al. [27], the RPE mechanism is evaluated by calculating the attention entropy and computing the attention distance spanned through different attention heads with each MHA sublayer. The mean distance \overline{D}_h^l which is spanned by the attention head h for

encoding layer l is calculated as the weighted average distance within pairs of every sentence of a given corpus X . Therefore,

$$\bar{D}_h^l = \frac{\sum_{x \in X} \sum_{i=1}^{|x|} \sum_{y=1}^i w_{i,j}^h \cdot (i - j)}{\sum_{x \in X} \sum_{i=1}^{|x|} \sum_{y=1}^i w_{i,j}^h} \quad (14)$$

where $w_{i,j}^h$ is the attention weight from input token x_i to x_j for the attention head h . i and j signify the tokens' places x_i and x_j in the source sentences. By computing aggregation for the attention distance for every head, the mean attention distance spanned \bar{D}_h^l with reference to the encoding layer l is calculated as follows:

$$\bar{D}_h^l = \frac{1}{N_h} \cdot \sum_{h=1}^{N_h} \bar{D}_h^l \quad (15)$$

where N_h represents the number of AHs used within the layer. The mean attention distance does not reveal anything about how the attention weight is distributed across input tokens for a specific attention head. An AH with a greater mean attention distance may focus on separated sequences of the same tokens [27,28]. To estimate the concentration pattern for the attention head h inside layer l for the input token x_i , the entropy of attention distribution [28], $E_h^l(x_i)$, for attention head h is calculated as

$$E_h^l(x_i) = - \sum_{j=1}^i w_{i,j}^h \log w_{i,j}^h \quad (16)$$

The mean entropy of the attention distribution for encoding layer l is computed similar to the attention distance spanned, as follows:

$$E^l(x_i) = \frac{1}{N_h} \sum_{h=1}^{N_h} E_h^l(x_i) \quad (17)$$

Attention heads with more increased entropy have a more distributed attention pattern, whereas those with lower entropy have a more concentrated attention weight distribution. Attention distance and attention entropy are computed depending on the attention weights generated for arbitrary 2000 sentences from the LEV-MSA task's test split (Baniata et al. [15] dataset). Figure 6 shows the mean attention distance span and the mean entropy of attention distribution for each attention head (AH) for the encoding layer of the proposed RPE MHA MT model for ADs. As illustrated, some focus on short-distance relations between input tokens, whereas other MHA heads capture long-distance relations among input tokens. Furthermore, the entropy of the attention distribution varies across the layers. Additionally, the entropy of the attention distribution varies for AH in the same layer. As reported in Figure 7, for the proposed model, the majority of the AHs that have an increased mean attention span and much more stable values of attention distribution are located in Layer 4. Nonetheless, a large mean attention distance does not mean stable attention distribution. The preceding layers have multiple AHs with a large distance span, but a significantly less consistent attention weight distribution. For instance, in Layer 3, AH five had the largest mean attention span (4.11) but the lowest mean entropy score (0.38). AHs with an increased mean attention distance span concentrate their attention on a word in duplicated sentences that appear at various spots in the source sentence as demonstrated by Vig et al. [27]. This assists in describing the lower entropy of the weight distribution over the sentence of the input tokens. AH that have stable or less stable weight distribution and a short attention distance span particularly focus on nearby tokens. The AH with variable mean attention distance and variable entropy permits the proposed model for ADs to efficiently learn variable structure information across its layers. For the proposed model, the reverse

positional encoding mechanism has a positive impact on the MHA sublayer in the encoder layer. Figure 8 represents the mean average attention distance and entropy for all heads in the MHA through the encoder layers. As illustrated in Figures 7 and 8, revealing the encoder layers to the decoder network permits the encoder to understand the source knowledge more reliably. Figure 8 shows the variation in the average mean attention distance span and entropy of the attention weight distribution for various (AH) over various encoder layers. As illustrated in Figure 8, the proposed RPE MHA NMT model concentrated on the (AH) with a shorter span over layer $2 \leq L \leq$ These layers are used to learn the way to manage contextual and local knowledge in the nearby area of the input source tokens. Layers 1 and 4 were responsible for learning the long-distance interaction between input source tokens. Commonly, applying the RPE describes how the source information of AD is acquired over several attention heads and layers in the encoder, as demonstrated by the entropy of the attention weight distribution and attention distance. This enriches the effectiveness of the proposed RPE MHA NMT model in comprehending the source information required to enhance the quality of translation. This reveals the superiority of the proposed RPE MHA NMT model over sequence-to-sequence architectures, such as the transformer-NMT model for AD [15].

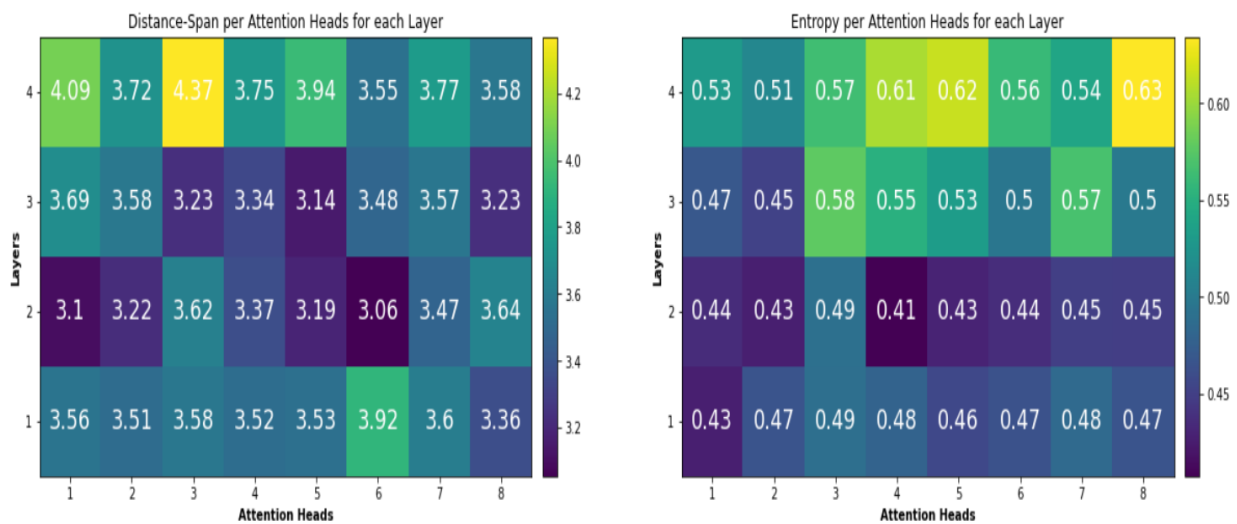


Figure 7. Mean attention distance span and attention distribution entropy concerning each encoder layer and attention heads for the proposed RPE MHA-based NMT model.

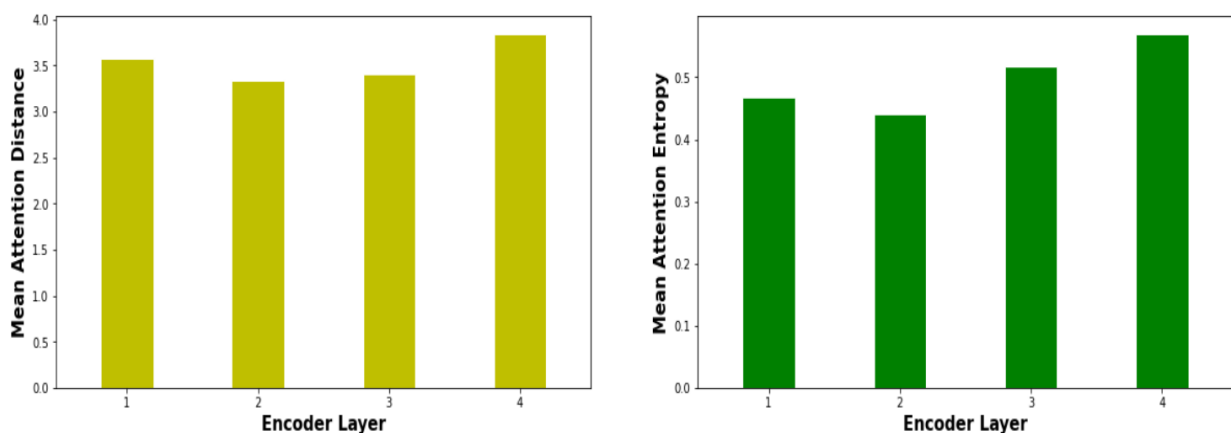


Figure 8. Mean attention distance and mean attention entropy concerning each encoder layer for the proposed RPE MHA-based NMT model.

4.11. Attention Analysis

The suggested RPE MHA NMT model presents a strategy for analyzing the alignment of terms in the resulting translation with terms in the source sentence. This strategy is illustrated in Figure 9 by visualizing the annotation weights [29,30]. Every row of the matrix displays the weights associated with annotations, with the *x*-axis representing the AD sentence and the *y*-axis representing the resulting sentence in the MSA. This explains which parts of the source text were prioritized when the target text was created. As illustrated in Figure 9, the alignment of terms between LEV Arabic and MSA is homogeneous. Strong weights were observed in the diagonals of each matrix. In addition, nonmonotonic and nontrivial alignments were observed. Usually, adjectives, verbs, and nouns are ordered in different ways in Levantine Arabic and MSA, as shown in the lower-right section of Figure 9. We can notice that the suggested model fluently translates a Levantine Arabic phrase (هي عشرين ورجع لي دولارين فراطه اذا سمحت) “This is twenty and give me an exchange of two dollars please” into MSA sentence (هاك عشرين دولار واعطيني فكة دولارين من فضلك) “Here’s twenty dollars and give me two dollars change please”. The suggested RPE MHA-based Neural MT model aligned the term (هي عشرين) “This is twenty” with (هاك عشرين دولار) “Here’s twenty dollars”. The suggested model captured hidden semantics in a context such as (دولار) “dollar” from the Levantine AD source sentence and added it to the translated MSA sentence correctly. Furthermore, the proposed RPE MHA NMT model handled and managed source and target phrases of different lengths.

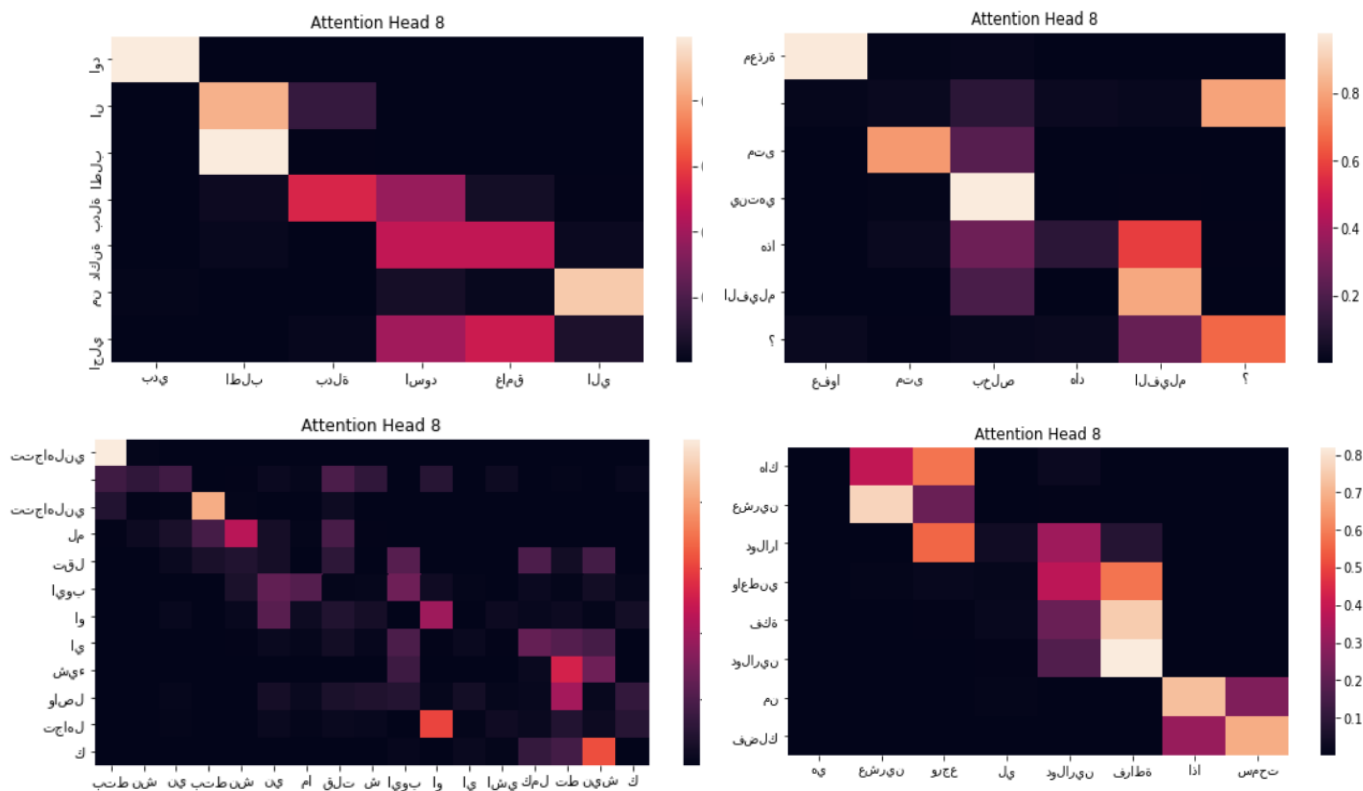


Figure 9. Visualization of Annotation Weights.

5. Conclusions

This research project presented a reverse positional encoding multi-head attention NMT model to translate AD sentences into MSA. The performance and efficiency of the proposed model were optimized by learning it on machine translation tasks from various ADs to the MSA. Exploiting the benefits of the new reverse-positional encoding

mechanism improves the quality of the proposed neural MT model. The practical results of this project prove important features for the RPE MHA NMT model, which exploits the new RPE mechanism and optimizes the BLEU results for the LEV-MSA, MAG-MSA, Nile Basin MSA, Gulf-MSA, and IRQ-MSA translation tasks. The use of the RPE mechanism and sub-word units as an input features to the self-attention layer showed that these two methods are important for low-resource language translation, such as Arabic dialects. Furthermore, training the suggested approach with various settings, such as using multiple heads in the multi-head attention layer, and experimenting with multiple encoders and decoders, improved the translation efficiency of the proposed model. The results of practical experiments on the LEV-MSA, MAG-MSA, Nile Basin MSA, Gulf-MSA, and IRQ-MSA tasks illustrated that the proposed model enhanced the BLEU results compared with other Arabic dialects NMT models. The analysis of experiments and results showed that performance relied on the use of the RPE mechanism and the dimension of sub-word embedding. Experimental analysis clarified that using the RPE method is beneficial in that the global and local semantic information in the context is acquired using the MHA sub-layer in every encoder layer. Furthermore, the proposed RPE MHA-based NMT model can handle the issue of training data rarity. In addition, the suggested model addressed the ADs syntactic issue, which is the free structure of the AD sentence. The suggested RPE MHA-based NMT model, with the RPE mechanism and sub-word units as input features to the self-attention layer, was able to deliver a high-quality translation for Arabic dialects.

Author Contributions: L.H.B., S.K. and I.K.E.A. conceived and designed the methodology and experiments; L.H.B. performed the experiments; I.K.E.A. performed the visualization; L.H.B., S.K. and I.K.E.A. analyzed the data; L.H.B. wrote the paper. S.K. reviewed the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Science and ICT under Grant NRF-2022R1A2C1005316.

Institutional Review Board Statement: Not Applicable.

Informed Consent Statement: This study did not involve humans.

Data Availability Statement: The dataset generated during the current study is available in the [RPE_MHA_NMT_AD] repository (<https://github.com/laith85>) (accessed on 10 August 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jean, S.; Cho, K.; Memisevic, R.; Bengio, Y. On using very large target vocabulary for neural machine translation. In Proceedings of the 53rd Annual Meeting of the Association for the Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, 26–31 July 2015; Volume 1, pp. 1–10.
2. Luong, M.T.; Sutskever, I.; Le, Q.V.; Vinyals, O.; Zaremba, W. Addressing the rare word problem in neural machine translation. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint conference on Natural Language processing, Beijing, China, 26–31 July 2015; pp. 11–19.
3. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. In Proceedings of the 27th International Conference on Neural Information Systems; Montreal, QC, Canada, 8–13 December 2014, Volume 2, pp. 3104–3112.
4. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. In Proceedings of the 33rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
5. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the 3rd Conference on Neural Information Processing system (NIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–9008.
6. Popović, M.; Arcan, M.; Klubička, F. Language related issues for machine translation between closely related south Slavic languages. In Proceedings of the 3rd Workshop on NLP for Similar Languages varieties and Dialects (VarDial3), Osaka, Japan, 12 December 2016; pp. 43–52.
7. Harrat, S.; Meftouh, K.; Smaili, K. Machine Translation for Arabic dialects. *Info. Process. Manag.* **2019**, *56*, 22–273.

8. Durrani, N.; Sajjad, H.; Fraser, A.; Schmid, H. Hindi-to-Urdu machine translation through translation through transliteration. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 11–16 July 2010; pp. 465–474.
9. Costa-Jussà, M.R. why Catalan-Spanish Neural Machine Translation? In Analysis, Comparison and Combination with standard rule and phrase-based technologies. In Proceedings of the Fourth Workshop on NLP for similar Languages, Varieties and Dialects, Valencia, Spain, 3 April 2017; pp. 55–62.
10. Costa-Jussà, M.R.; Zampieri, M.; Pal, S. A neural approach to language variety translation. In Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects, Sana Fe, NM, USA, 20 August 2018; pp. 275–282.
11. Meftouh, K.; Harrat, S.; Jamoussi, S.; Abbas, M.; Smali, K. Machine translation experiments on padic: A parallel Arabic dialect corpus. In Proceedings of the 29th Pacific Asia Conference on Language, information and Computation, Shanghai, China, 30 October–1 November 2015.
12. Baniata, L.H.; Park, S.; Park, S.-B. A Neural Machine Translation Model for Arabic Dialects That Utilizes Multitask Learning (MTL). *Computational Intel. Neuosci.* **2018**, *2018*, 7534712. [[CrossRef](#)] [[PubMed](#)]
13. Baniata, L.H.; Park, S.; Park, S.-B. A Multitask-Based Neural Machine Translation Model with Part-of-Speech Tags Integration for Arabic Dialects. *Appl. Sci.* **2018**, *8*, 2502. [[CrossRef](#)]
14. Aqlan, F.; Fan, X.; Alqwbani, A.; Al-Mansoub, A. Arabic Chines Neural Machine Translation: Romanized Arabic as subword unit for Arabic-Sourced Translation. *IEEE Access* **2019**, *7*, 133122–133135. [[CrossRef](#)]
15. Baniata, L.H.; Ampomah, I.K.E.; Park, S. A Transformer-Based Neural Machine Translation Model for Arabic Dialects that Utilizes Subword Units. *Sensors* **2021**, *21*, 6509. [[CrossRef](#)] [[PubMed](#)]
16. Shaw, P.; Uszkoreit, J.; Vaswani, A. Self-Attention with relative position representation. In Proceedings of the 2018 Conference of North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, LA, USA, 1–6 June 2018; Volume 2, pp. 464–468.
17. Dai, Z.; Yang, Z.; Yang, Y.; Carbonell, J.; Le, Q.; Salakhatdinov, R. Transformer-XL: Attentive language models beyond a fixed-length context. In Proceedings of the 57th Annual Meeting of the Association for the Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 2978–2988.
18. Pham, N.-Q.; Ha, T.-L.; Nguyen, T.-N.; Nguyen, T.-S.; Salesky, E.; Stueker, S.; Niehues, J.; Waibel, A. Relative positional encoding for speech recognition and direct translation. In Proceedings of the 2019 Conference of North American Chapter of the Association for Computational linguistics: Human Language Technologies, Minneapolis, MN, USA, 3–5 June 2019; pp. 3999–4004.
19. Casas, N.; Costa-Jussa, M.R.; Fonollosa, J.A.R. Combining subword representations into word-level representations in the transformer architecture. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, Online, 5–10 July 2020; pp. 66–71.
20. Libovicky, J.; Fraser, A. Towards reasonably-sized character-level transformer NMT by finetuning subword systems. In Proceeding of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–18 November 2020; pp. 2572–2579.
21. HE, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, Nevada, USA, 27–30 June 2016; pp. 770–778.
22. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer normalization. In Proceedings of the Advances in NIPS 2016 Deep Learning Symposium, Barcelona, Spain, 5–10 December 2016.
23. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May; pp. 5886–8577.
24. Luong, M.-T.; Pham, H.; Manning, C.D. Effective Approaches to attention-based neural machine translation. In Proceedings of the 2018 Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 1412–1421.
25. Park, C.; Yang, Y.; Park, K.; Lim, H. Decoding strategies for improving low-resource machine translation. *Electronics* **2020**, *9*, 1562. [[CrossRef](#)]
26. Raganato, A.; Tiedemann, J. An analysis of encoder representations in transformer-based machine translation. In Proceedings of the 2018 Empirical Methods in Natural Language Processing Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Brussels, Belgium, 1 November 2018; pp. 287–297.
27. Vig, J.; Belinkov, Y. Analyzing the structure of attention in a transformer language model. In Proceedings of the Second BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, Florence, Italy, 1 August 2019; pp. 63–76.
28. Ghader, H.; Monz, C. What does attention in neural machine translation pay attention to? In Proceedings of the 8th IJCNLP, Taipei, Taiwan, 27 November–1 December 2017; pp. 30–39.
29. Ampomah, I.K.E.; McClean, S.; Hawe, G. Dual contextual module for neural machine translation. *Mach. Transl.* **2021**, *35*, 571–593. [[CrossRef](#)]
30. Ampomah, I.K.E.; McClean, S.; Lin, Z.; Hawe, G. Every layer counts: Multi-layer multi-head attention for neural machine translation. *Prague Bull. Math. Linguist.* **2020**, *115*, 51–82. [[CrossRef](#)]