ISSN: 0067-2904

Training Data Generation for Libyan Dialect–MSA Translation Using Word Replacement and Fine-tuning Strategies

Husien Alhammi[1]*, Kais Haddar [2]
[1]*Faculty of Economics and Management, University of Sfax, Sfax, Tunisia*
[2]*Laboratory MIRACL, University of Sfax, Sfax, Tunisia*

Received: 27/1/2025 Accepted: 21/12/2025 Published: 30/10/2026

Abstract

Translating Arabic dialects into modern standard Arabic is a difficult task due to the significant differences in grammar and vocabulary between them. Using neural machine translation approaches to translate Arabic dialects faces a significant challenge, which is the lack of parallel data. Unavailability of parallel data is a common problem in low-resource languages such as the Libyan dialect. To solve this problem, this paper presents a novel method to address this problem. This method uses both word replacement and fine-tuning techniques to build Libyan dialect-modern standard Arabic parallel data for neural machine translation. To build parallel data from monolingual modern standard Arabic sentences, it uses a word replacement to exchange modern standard Arabic words in the modern standard Arabic text with Libyan dialect equivalents. It also uses fine-tuning techniques to enhance the overall translation of the neural machine translation model, which is already pretrained on parallel data that is generated by word replacement. The method is evaluated using a test dataset of 1,000 parallel sentences. The results show that the method proposed in this paper achieves promising results compared to the baseline technique. It obtained a score of 1.37 for BLEU, a score of 12.87 for chrF, and 87.34 for TER. In contrast, the baseline model achieved a BLEU score of 0.12, a chrF score of 6.89, and a higher TER of 100.02.

Keywords: machine translation, MSA monolingual data, training data, word replacement, Libyan dialect, NMT model.

1. Introduction

Modern Standard Arabic (MSA) is the formal language in all Arab countries, which is used for official aspects such as education and news, while Arabic dialects are used in everyday conversation by Arabic people. However, the Libyan Dialect (LD) is one of the Maghrebi dialect family that is spoken in North Africa. The Arabic dialects can be classified into Gulf, Levantine, Egyptian, and Maghrebi Arabic. Although the LD and MSA have many similarities, they also have many differences in vocabulary, grammar, and phonology. For instance, the LD phrase "شبحت القطوسة", which means "I saw the cat" in English, can be translated to "رأيت القطة" in MSA. In this example, it can be noted that the vocabulary in LD and MSA phrases is different. There are many advantages to translating the LD into MSA. It enables people who speak other Arabic dialects to understand the LD, as well as enhancing the performance of various Natural Language Processing (NLP) tasks, such as cross-language

*Email: * h1974hami@gmail.com

Open Access Articles. This articles is licensed under

retrieval. Translating Arabic dialects into MSA presents a challenge because Arabic dialects have an informal nature and lack standardized and large-scale parallel datasets. Parallel datasets include two places for text: one for the source language and the second for the target language. Neural Machine Translation (NMT) models can give good results due to training them on big data, but they can give bad results due to training them on small data. Building parallel datasets by humans can be costly and a waste of time because it requires a lot of human work.

Researchers use Data Augmentation (DA) techniques to increase their data size for addressing the lack of parallel datasets in low-resource languages. The Word Replacement (WR) [1] [2] and Back-Translation (BT) approach [3] [4] are the DA techniques that are used for increasing data size. In order to modify or augment the text, the WR is used for this. Parallel pseudo-training data can be generated by the BT by translating source languages to target languages and using both languages to generate parallel data. Furthermore, transfer learning, such as fine-tuning techniques, is also used to solve the data scarcity problem for NMT in low-resource languages. This enables leveraging information gained from pre-trained models. Fine-tuning refers to the process in machine learning that takes a model that has already been pre-trained on a broad dataset and further trains it on a smaller, domain-specific dataset to adapt it to a particular task, such as medical translation or text classification. In this paper, a method is introduced to address the data scarcity challenge through WR and fine-tuning techniques. First, a large number of monolingual MSA sentences were collected for the target side of the parallel dataset. The same MSA sentences were also used to represent the source side. The MSA sentences on the source side are automatically replaced with their LD equivalents to create an LD-MSA parallel dataset. This large LD-MSA parallel dataset is used to pre-train the NMT model, and then the model is fine-tuned on a small genuine LD-MSA parallel dataset for translating LD into MSA.

It is expected that an NMT model, initially pre-trained on a large modified parallel dataset and then fine-tuned on a smaller parallel dataset, will demonstrate significant improvements over the same model trained solely on a small baseline dataset. The proposed method in this paper presents a novel idea to address the lack of parallel data in low-resource languages by combining word replacement and fine-tuning strategies in one process. The conventional NMT models use available extensive parallel data for translation, which are not available in low-resource languages. The proposed method offers an effective solution for translating dialects into their standard languages by generating such datasets. Several experiments were done to demonstrate that the method can be used for addressing the scarcity of training parallel data in low-resource dialects.

2. Related Work

Low-resource languages lack training parallel data that can be used for NMT systems. The scarcity of this data hinders the development of NMT systems. NMT models require large amounts of parallel data to achieve high translation quality. To overcome the issue, much research has used DA techniques to generate training data. Several significant studies that were conducted on DA and Arabic dialect machine translation are listed below.

In 2024, Alhammi and Haddar [5] used a rule-based approach to create a machine translation system for translating LD dual forms into MSA dual forms, which is the first machine translation system for LD. Their translation system consists of three stages, which include morphological analysis and disambiguation, lexical and structural transfer, and morphological generation. Their system achieved reasonable results. The scores were 40.3% for BLEU, 41.37% for TER, and 66.9% for ChrF.

Moukafih et al. in 2021 [6] used Multi-Task Learning (MTL) frameworks to overcome data scarcity in Arabic dialects by leveraging the similarities between them in order to improve

overall translation quality. In their experiments, they used a shared encoder for multiple source dialects and specific decoders for target languages (many-to-one) on the small PADIC dataset. The results indicate that their MTL approach, particularly the many-to-one strategy, significantly outperforms traditional statistical machine translation methods across the majority of dialect translation pairs tested, achieving superior performance in 88% of cases. This demonstrates the effectiveness of MTL for low-resource dialect translation.

Sennrich et al. in 2016 [3] generated synthetic parallel data by BT monolingual target sentences into the source language for training a standard NMT framework, significantly boosting translation quality for both English-German and Turkish-English. The work also demonstrates that BT data enhances fluency, reduces overfitting, and improves domain adaptation, especially in low-resource languages such as Turkish-English. Their BLEU scores increased by approximately 2.8 to 3.7 points for English-to-German translation and improved by about 2.1 to 3.4 points for Turkish-to-English translation.
Poncelas et al. [7] evaluated how synthetic data created by BT monolingual target-language text affects the quality of NMT models. The authors train systems using authentic parallel data only, synthetic data only, and hybrid combinations of both. They found that adding BT data improves performance in low-resource settings, and hybrid systems often outperform those trained on authentic data alone, especially when data is limited. However, too much synthetic data can begin to harm model quality. Surprisingly, synthetic-only models perform nearly as well as authentic-only models in low-resource settings, though their scalability is limited. They conducted many experiments for German-to-English translation. For example, when the model was trained solely on one million authentic parallel sentences, it achieved a BLEU score of 0.2296. Training the model only on one million synthetic parallel sentences resulted in a slightly lower BLEU score of 0.2290. However, combining one million sentences for each authentic and synthetic in a hybrid dataset led to a notable improvement, with the BLEU score increasing to 0.2421.

Omer Nacar et al. [8], in 2024, described a robust translation approach to translate five major Arabic dialects (Gulf, Egyptian, Levantine, Iraqi, and Maghrebi) to MSA. The approach used generative models (GPT-3.5 and GPT-4) to expand the small dataset of dialectal Arabic with synthetic MSA translations, followed by fine-tuning the AraT5 V2 model on this data. They conducted their experiments on large external datasets. The BLEU score in their result was higher. It was 85.5% for the validation set and 22.6% for the blind test set. The system performance on the blind test set was lower due to the small dataset size. In their study, the AraT5 V2 shows the ability to capture the linguistic features of Arabic dialects and the efficiency of GPT-4 for enhancing translation accuracy at lexical, syntactic, semantic, and orthographic levels.

In 2025, Fuxue Li et al. [9] proposed a DA method at the clause-based level to improve NMT for low-resource languages. They used syntactic parsing to split the original training corpus sentences into meaningful clauses. To generate pseudo-parallel data, extracted clauses were translated by a backward NMT model to create synthetic bilingual pairs for pseudo-parallel data. In their work, two DA strategies were used, which include Direct Mixture Augmentation (DMA) and Concatenate Mixture Augmentation (CMA). The DMA was used to combine original and clause-based synthetic data, while the CMA was used to join many clauses together to form longer pseudo-sentences. Their results showed that the BLEU score improved for Arabic-English and other low-resource tasks by 1.88 compared to baselines. It also showed the results of the CMA outperforming DMA for longer sentences. Their work has limitations, including the dependency on the accuracy of syntactic parsing and the limitation of clause-level augmentation (e.g., word or phrase).

The studies in this section show that rule-based MT is effective for morphologically rich languages such as Arabic dialects. Nevertheless, it relies heavily on linguistic expertise and handcrafted rules, and it struggles with scaling larger vocabularies; its translations usually have low accuracy and fluency. Furthermore, many studies demonstrated that despite the lack of training data in low-resource languages, using transfer learning, pretrained multilingual models such as mBART and mT5, with data augmentation techniques, can significantly enhance NMT quality. This research aims to address the problem of insufficient training data in low-resource dialects through employing strategies such as data augmentation (WR) and transfer learning (fine-tuning).

3. The Proposed Method

The methodology involves combining WR and fine-tuning techniques to address the lack of LD-MSA training data that can be used for training the NMT model. A large number of monolingual MSA sentences were collected. These monolingual MSA sentences were used to create a parallel MSA-MSA dataset for the source and target sides. The MSA words that represent the source side in this parallel dataset were replaced with LD equivalents in order to generate a pseudo-parallel LD-MSA dataset. The pseudo-parallel LD-MSA dataset was used to pre-train the NMT model. The model was then retrained on a small genuine parallel LD-MSA dataset to fine-tune it to LD-MSA. However, the WR strategy is typically used to augment existing datasets. Nevertheless, in this work, it is used for modifying existing datasets. Figure 1 illustrates the main steps of the proposal method.

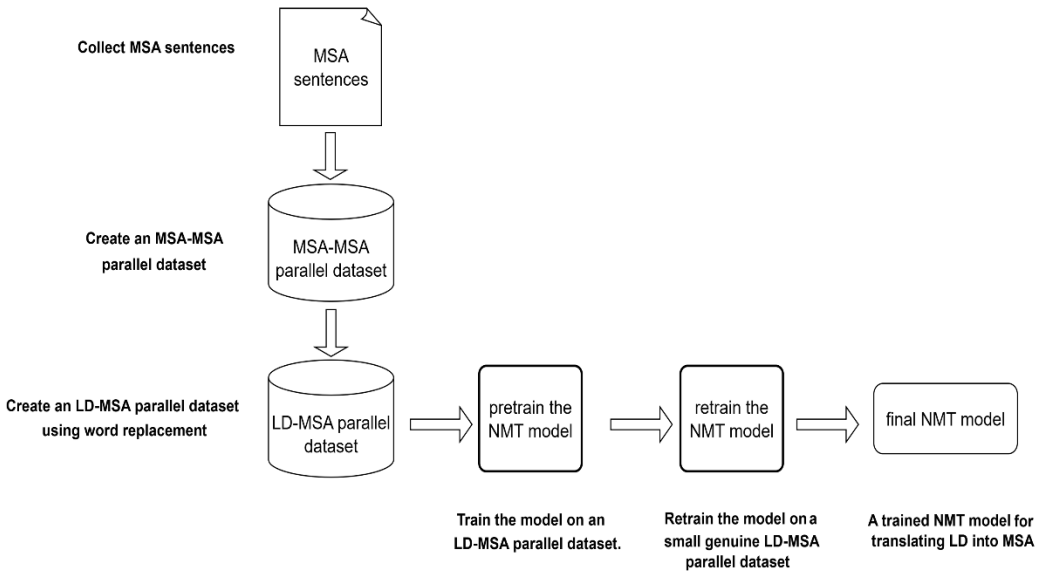

Figure 1: The main steps of the proposed method.

3.1 Datasets Collection

Two groups of data were collected in this work. The first group contains approximately 47k monolingual MSA sentences from the Tatoeba-v2023-04-12 corpus [10]. The size of the data was increased to 60k by adding additional data from various Hugging Face datasets, such as xnli2.0_arabic [11]. These monolingual MSA sentences were used to create the pseudo-parallel LD-MSA dataset. Additionally, the second group contains 9k parallel LD-MSA sentences. It includes about 4k parallel LD-MSA sentences from the MADAR corpus [12], almost 2k of 2-gram parallel sentences from the LD-MSA corpus [5], as well as 3k of parallel sentences, which were generated manually by the authors.

3.2 Lexical Replacement Strategy

A WR strategy was applied to modify MSA sentences on the source side to LD sentences, thereby producing an LD-MSA parallel dataset. This approach makes use of the linguistic similarities between LD and MSA. Both languages have a significant overlap in vocabulary, and they use similar sentence structures, which are verb-subject-object (VSO) and subject-verb-object (SVO) structures.

Translating between American and British English is a good example of explaining the notion of WR. For instance, the British English sentence "The lorry is parked near the pavement" can be translated to "The truck is parked near the sidewalk" in American English. In this translation, the British words "lorry" and "pavement" are replaced with their equivalents, "truck" and "sidewalk," in American English, respectively. Obviously, the other words ("the," "is," "parked," and "near") remain unchanged because they are shared words. Likewise, a WR strategy can be utilized to convert MSA sentences to LD by replacing MSA words with their LD equivalents. For example, the MSA sentence "أين أقرب مخبز", which translates to "where is the nearest bakery" in English, can be converted to LD by replacing the words "أين" and "مخبز" with their LD equivalents "وين" and "كوشة" respectively, resulting in the LD sentence "وين اقرب كوشة". In this study, a tool for WR was developed to automatically replace MSA words with their LD equivalents. A bilingual dictionary of LD-MSA [13] was utilized for one-to-one word substitutions. For instance, the MSA word "قطة", which means "a cat" in English, can be replaced with its LD equivalent "قطوسة". Each entry in the bilingual dictionary contains LD words and their MSA equivalents, with their linguistic information, such as Part of Speech (POS), gender, number, and verb tense. Additionally, a 2-gram parallel corpus [5] was employed to replace pairs of MSA words with their corresponding LD words. For example, the MSA phrase "غير صحيح", meaning "incorrect", can be replaced with its LD equivalent "مش مزبوط". Moreover, word affixes such as the definite article "ال", meaning "the", are considered in the process of WR. For instance, despite the bilingual dictionary including the MSA word "مخبز" (without the definite article "ال") and its corresponding LD term "كوشة", both meaning "a bakery", the MSA word "المخبز" (with the definite article "ال"), indicating "the bakery", can be replaced with the LD equivalent "الكوشة". Table 1 presents examples of sentences from both the source and target sides of the parallel LD-MSA after the WR process, along with the English and reference translations.

Table 1: Examples of the source and target sides from the parallel LD-MS dataset after the WR process

LD Sentences (human translation)	English Translation	LD Sentences (after WR process) (source side)	MSA Sentences (target side)
يحب يهدرز على الجيش	He likes to talk about the army	هو يحب ان يهذرز عن الجيش	هو يحب ان يتحدث عن الجيش
طلع الكناسة	take out the trash	طلع الكناسة	أخرج القمامة
اعطاني بوي هلبة فلوس	My father gave me a lot of money.	أعطاني بوي هلبة من الفلوس	أعطاني والدي الكثير من المال
شري توم زوز قوانطيات موطو جلد	Tom bought a pair of leather motorcycle gloves.	شري توم راجل قوانطيات موطو جلدية	اشترى توم زوج قفازات دراجة نارية جلدية
نطلب منك تحترم قراري	I ask you to respect my decision.	أطلب منك أن تحترم قراري	أطلب منك أن تحترم قراري
العنوان مش مزبوط	The address is not correct	العنوان مش مزبوط	العنوان غير صحيح
الجو سمح هلبة هلبة اليوم	The weather is very nice today.	إن الجو سمح هلبة هلبة اليوم	إن الطقس جميل جدا جدا اليوم

3.2.1 Word Replacement Challenges

Although WR facilitates the generation of extensive LD-MSA parallel data from monolingual MSA sentences, it also presents several limitations. These challenges are not addressed in the study, as they necessitate specialized processing.

1- Contextual Appropriateness: A word-for-word replacement may not always be contextually appropriate. When directly replacing MSA words with their LD equivalents, MSA words could be replaced with unsuitable words, resulting in semantic inaccuracies. For example, the MSA word "جيد", meaning "good", can be replaced by two LD words, "باهي" or "كويس", in the MSA sentence "شعور جيد", meaning "good feeling". Two phrases would be generated: "شعور باهي" and "شعور كويس", respectively. The phrase "شعور كويس" is better than the phrase "شعور باهي". In this case, the process of choosing the correct word to replace it is very important.

2- Grammatical Variations: LD grammatical structures are different from MSA. Focusing on lexical differences and ignoring syntactic differences between LD and MSA can reduce replacement accuracy. For example, the MSA sentence "سوف أذهب إلى السوق", meaning "I will go to the market" and its translation in LD "بنمشي لي السوق", both languages have different grammatical structures of the future tense, which can clearly make WR a challenging task.

3- Polysemy Handling: MSA words that can have many meanings are called polysemy. Selecting appropriate words for replacing these words in MSA by LD words requires a good understanding of the meaning of these polysemes. In this case, a WR strategy may struggle to select an appropriate word for polysemy. For example, the MSA word "جميل", meaning "beautiful", has three LD equivalents, which are "باهي", "كويس", and "قنين". Selecting an appropriate LD equivalent can be a challenge.

4- Dual and Negation Forms: In the domain of linguistics, dual forms in MSA differ from those in LD. For example, the dual form "سيارتان" in MSA, meaning "two cars", is expressed as "زوز كراهب" in LD. Moreover, the negation forms in MSA are not the same as in LD. For example, the negation phrase "لم أذهب" in MSA, meaning "I did not go", can be " ما مشيتش" in LD. Obviously, the differences in dual and negation forms between LD and MSA pose a challenge and necessitate further effort to tackle.

Despite these limitations, the WR strategy can provide a significant boost to NMT model training, particularly in low-resource languages where genuine parallel data is limited.

3.3 NMT Model Architecture

A well-established sequence-to-sequence model with an attention mechanism was implemented using the TensorFlow and Keras libraries. It was used solely as a baseline to evaluate the proposed method. The model has many different components. These components are mentioned below.

- Data Preprocessing: Involves normalization and tokenization, sequence padding, and the division of data into training and validation datasets.
- Model Architecture: Uses a bidirectional LSTM encoder and an LSTM decoder with Bahdanau attention. Where hyperparameters are embedding size=256, encoder/decoder layers=1, and hidden units=512.
- Training: Supports training with early stopping and checkpointing, using teacher forcing.
- Inference: It uses beam search for generating translation.
- Evaluation: It calculates BLEU, chrF, and TER scores for translation results.

3.4 Fine-tuning Approach

A fine-tuning approach was used to enhance the translation quality of the NMT model for translating LD into MSA. First, an NMT model was pretrained on a pseudo-parallel LD-to-MSA dataset that was already generated by WR, which enables the NMT model to learn general information between LD and MSA, such as general patterns and lexical mappings. After that, the model was fine-tuned on a smaller and high-quality parallel dataset, which was built for LD-to-MSA translation. The fine-tuning process was used to enhance translation quality by benefiting from the information that was obtained by the NMT model from the pretraining process. This allows the NMT model to generate accurate translations in MSA.

4. Experiments

This research used Google Colab Pro++ to conduct three experiments for translating LD into MSA in order to evaluate the translation quality of the NML model. Two LD-MSA datasets generated by WR were used for pretraining the NML model. In addition, one high-quality LD-MSA dataset was used for fine-tuning the NML model.

- Dataset 1: It includes 8k LD-MSA parallel sentence pairs (group 1).
- Dataset 2: It has 30k LD-MSA parallel sentence pairs generated by WR (group 2).
- Dataset 3: It has 60k LD-MSA parallel sentence pairs produced by WR (group 2).
- Dataset 4: Unseen parallel LD-MSA dataset of approximately 1k for testing the NMT model (group 1).
- Experiment 1: It serves as the baseline, where the NMT model was trained from scratch using Dataset 1.
- Experiment 2: The NMT model was first pre-trained on Dataset 2 and subsequently fine-tuned on Dataset 1.
- Experiment 3: The NMT model was initially pre-trained on Dataset 3 and then fine-tuned on Dataset 1.

The translation quality of the NMT model on dataset 4 was evaluated by using the BLEU [14], chrF [15], and TER [16] metrics. Figure 2 shows how the loss and accuracy of the NMT model were changed when pretraining it on dataset 3.

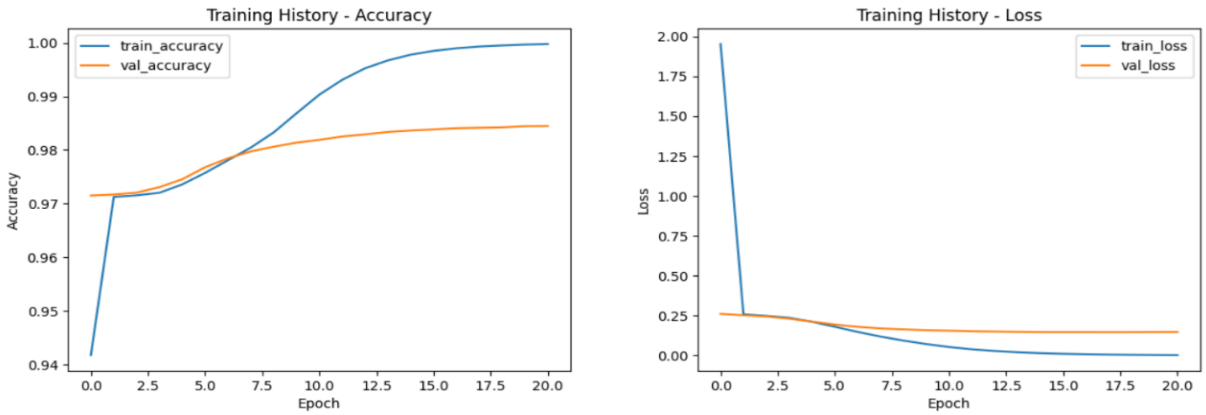

Figure 2: illustrates model behavior during pretraining on a dataset of 60k.

5. Results and Discussion

Three experiments were conducted in this research. The results of these experiments were used to evaluate the efficiency of the method for generating pre-trained data for translating LD into MSA. In Experiment 1, the NMT model was trained from scratch on Dataset 1, which presents the baseline of this study. The translation scores of the Experiment 1 results were used as a baseline for evaluating the translation quality of the NMT models. The

translation scores of this experiment were very low. The scores were 0.12 for BLEU, 6.89 for chrF, and a high TER of 100.02. In contrast, Experiment 2 used two steps for training. The model was first pretrained on Dataset 2 and then fine-tuned on Dataset 1, which showed a slight improvement in translation accuracy compared to Experiment 1, had a score of 0.90 for BLEU, a score of 10.69 for chrF, and a reduced TER of 92.82. The results of Experiment 3 showed noticeable improvements compared to previous experiments. First, the model was pretrained on Dataset 3, and then Dataset 1 was used to fine-tune it. The results of this experiment were the best results among the experiments. The translation scores of this experiment were a BLEU score of 1.37, a chrF score of 12.87, and a TER of 87.34. The results of experiments clearly showed that the performance of the NMT model in Experiment 2 and Experiment 3 was better than the baseline NMT model in Experiment 1. In Experiment 2, the NMT model was pretrained on the larger dataset of 30k before fine-tuning on Dataset 1, achieving a good enhancement compared to the baseline. This improvement was significantly increased in Experiment 3, where the model was pretrained on the larger dataset of 60k before fine-tuning on Dataset 1, resulting in the best performance between all experiments. The results of the experiments clearly showed that increasing the pre-training dataset size that was used to pre-train the NMT model enhances the translation quality. Table 2 illustrates the results of three experiments.

The evidence of this work showed that the WR is very important for creating larger datasets that can be used for pre-training NMT models. Fine-tuning these pre-trained NMT models on a smaller and high-quality dataset can improve translation quality compared to the baseline. Additionally, the experiments clearly indicated that enlarging the size of the pre-trained dataset resulted in significant enhancements in translation quality. In other words, increasing the size of the pre-trained dataset will contribute to improved translation quality. Consequently, WR can be used to enlarge training data by exploiting the existing monolingual data of standard languages for translating low-resource dialects into their standard languages. For example, this method can also be used to translate the Tunisian dialect into MSA.

Table 2: The translation scores of the three experiments.

Experiments	Evaluation Metrics			
	BLEU	chrF	TER	Model Description
Experiment 1	0.12	6.98	100.02	baseline model: train on dataset 1 (without fine-tuning)
Experiment 2	0.90	10.69	92.82	fine-tuned model: Pre-trained on Dataset 2, then fine-tuned on Dataset 1
Experiment 3	1.37	12.87	87.34	fine-tuned model: Pre-trained on Dataset 3, then fine-tuned on Dataset 1

Finally, it is suggested that MSA can be effectively utilized as a means to facilitate the translation of Arabic dialects into other languages. For example, to translate LD to English, WR is used to replace MSA words in a large-scale MSA-English dataset with LD equivalents. This modified dataset is used to pre-train the NML model, and then the model is retrained on a small, high-quality LD-English dataset.

6. Conclusion and Future Work

In this study, an NMT method is proposed for translating LD into MSA. This method uses WR to generate training datasets and uses fine-tuning techniques to enhance translation quality. The results of the three experiments clearly show that the proposed method achieved

significant performance compared to the baseline. To evaluate the NMT model for all experiments, a test dataset of 1k unseen parallel sentences was used. In Experiment 1, the baseline model was trained from scratch by using a small and high-quality dataset of 8k, achieving very low performance because of the small size of the dataset that was used for training it, which is the 8k dataset. The translation scores of Experiment 1 were 0.12 for BLEU, 6.98 for chrF, and 100.02 for TER. In contrast, Experiment 2 and Experiment 3 have improved translation performance compared to Experiment 1 based on the size of pre-trained data that was used to pretrain the NMT model. The model was first pre-trained on larger datasets of 30k, then fine-tuned on a high-quality dataset of 8k. The model yielded slight results, where scores were 0.90 for BLEU, 10.69 for chrF, and 92.82 for TER. Likewise, the model was pretrained on a 60k dataset, and then a high-quality dataset of 8k was used to fine-tune it. The scores were 1.37 BLEU, 12.87 chrF, and 87.34 TER. The results also showed that the translation quality would improve if the size of the pretrained dataset increased. Although the experiments show noticeable improvement, the results remain low because the datasets that were used to train the NMT model were too small to achieve good results.

To achieve higher translation quality, we suggest increasing the pre-training dataset to at least one million and 100k high-quality datasets for fine-tuning, as well as using the transformer model architecture. The positive side of this research shows the importance of WR and fine-tuning in enhancing translation quality, offering an effective approach for mitigating the problem of scarcity of parallel training data in low-resource dialects.
Future work will focus on developing more advanced WR that takes into account the context and grammatical structure of sentences for enhancing the quality of pre-training data. Additionally, the results derived from this data will be compared to those obtained from fine-tuning the AraT5 and mBART50 models.

References

[1] Y. Xiao, L. Wu, J. Guo, J. Li, M. Zhang, T. Qin, and T.-y. Liu, "A survey on non-autoregressive generation for neural machine translation and beyond," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 10, pp. 11 407–11 427, 2023.

[2] F. Gao, J. Zhu, L. Wu, Y. Xia, T. Qin, X. Cheng, W. Zhou, and T.-Y. Liu, "Soft contextual data augmentation for neural machine translation," in Proceedings of the 57th annual meeting of the association for computational linguistics, 2019, pp. 5539–5544.

[3] R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," arXiv preprint arXiv:1511.06709, 2015.

[4] S. Edunov, M. Ott, M. Auli, and D. Grangier, "Understanding back-translation at scale," arXiv preprint arXiv:1808.09381, 2018.

[5] H. Alhammi and K. Haddar, "A rule-based system for translating Libyan dialect dual forms to modern standard Arabic," in International Conference on Language Processing and Knowledge Management. Springer, 2024, pp. 279–290.

[6] Y. Moukafih, N. Sbihi, M. Ghogho, and K. Smaïli, "Improving machine translation of Arabic dialects through multi-task learning," in International Conference of the Italian Association for Artificial Intelligence. Springer, 2021, pp. 580–590.

[7] A. Poncelas, D. Shterionov, and A. Way, "Gm d. b. wenniger, and p. passban,"investigating backtranslation in neural machine translation,"," arXiv preprint arXiv:1804.06189, 2018.

[8] O. Nacar, A. Alharbi, S. Sibaee, S. Ahmed, L. Ghouti, and A. Koubaa, "Asos at osact6 shared task: Investigation of data augmentation in Arabic dialect-msa translation," in Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation@ LREC-COLING 2024, 2024, pp. 104–111.

[9] F. Li, M. Shao, H. Yan, and C. Chi, "A clause-based data augmentation method for low-resource neural machine translation," IEEE Access, 2025.

[10] J. Tiedemann, Tatoeba Translation Challenge, "Tatoeba-v2023-04-12, opus, 2023," https://opus.nlpl.eu, accessed: 2025-01-13.

[11] A. Conneau, G. Lample, R. Rinott, A. Williams, S. R. Bowman, H. Schwenk, and V. Stoyanov, "Xnli: Evaluating cross-lingual sentence representations," arXiv preprint arXiv:1809.05053, 2018.

[12] H. Bouamor, N. Habash, M. Salameh, W. Zaghouani, O. Rambow, D. Abdulrahim, O. Obeid, S. Khalifa, F. Eryani, A. Erdmann, and K. Oflazer, "The MADAR Arabic dialect corpus and lexicon," in Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan: European Language Resources Association (ELRA), May 2018. [Online]. Available: https://aclanthology.org/L18-1535/

[13] H. Alhammi and K. Haddar, "Toward building a bilingual dictionary for libyan dialect-modern standard Arabic machine translation." in KDIR, 2022, pp. 75–81.

[14] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[15] M. Popovi´c, "chrf: character n-gram f-score for automatic mt evaluation," in Proceedings of the tenth workshop on statistical machine translation, 2015, pp. 392–395.

[16] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers, 2006, pp. 223–231.