ISSN 2007-9737

A Deep Neural Network Machine Translation Approach // from a Low-Resource Language to English

Sameh Kchaou*, Rahma Boujelbane, Lamia Hadrich Belguith

ANLP Research group, University of Sfax, MIRACL Lab,
Tunisia

{samehkchaou4, rahma.boujelbane}@gmail.com

Abstract. In the digital age, the proliferation of social media platforms such as Facebook, YouTube, and Twitter has led to an unprecedented surge in content produced in both standard languages and dialects. Translating dialects, such as Tunisian Dialect (TD), presents unique challenges for automatic translation systems due to their informal and regionally specific nature. This study aims to address these challenges by developing a robust translation model for Tunisian Dialect into English, a globally recognized formal language. Tunisian Dialect is often written using a mixture of Latin script and Arabic, while frequently incorporating foreign words, adding complexity to the translation task. Moreover, the scarcity of parallel Tunisian dialect to English corpora has hindered the development of accurate translation systems. To overcome these limitations, we created a parallel corpus between Tunisian Dialect and English by leveraging existing data between Tunisian Dialect and Modern Standard Arabic, applying data augmentation techniques to build a substantial trilingual dataset. Our research explores two translation approaches: one using Modern Standard Arabic as a pivot language, and the other translating directly from Tunisian dialect to English. The best model achieved a result of 67.89%, demonstrating significant improvements in translation accuracy and offering valuable insights for tackling dialectal translation challenges.

Keywords. Neural translation, poor languages, social networks, Tunisian dialect, transformer models.

1 Introduction

In today's digital era, the widespread use of communication technologies and social media platforms such as Facebook, YouTube, and Twitter has generated an unprecedented volume and diversity of information. These platforms have become essential tools for global communication, and users produce large amounts of content each day. This content, which includes reviews, comments, and recommendations, is often expressed either in standard languages or dialects. However, translating dialects poses significant challenges for automatic translation systems due to their informal and regionally specific characteristics.

Our research focuses on translating the Tunisian dialect (TD), a widely understood Arabic dialect, into English (EN), a globally recognized formal language. This task is particularly complex due to the informal nature of TD, which is often written using a mix of Latin script and Arabic ("Arabizi") and frequently includes foreign language borrowings, particularly in social media contexts. Additionally, the lack of parallel corpora for TD and EN further complicates the development of effective translation systems.

To address these issues, our work makes several important contributions. Firstly, we developed a parallel TD-EN corpus by leveraging existing TD-MSA (Modern Standard Arabic) data, and applied data augmentation techniques to build a comprehensive trilingual TD-MSA-EN corpus. Given the diglossic nature of Arabic, where MSA serves as the formal standard, we explored two translation strategies: one using MSA as a pivot language, and another translating directly from TD to EN without an intermediate step.
We then fine-tuned existing Neural Machine Translation (NMT) models, including OpenNMT and Fairseq, optimizing them for this unique

trilingual corpus. Our findings demonstrate the value of using MSA as a pivot language for Arabic dialect translation, significantly improving translation quality compared to direct translation methods.

2 Related Work

The development of automatic translation systems for Arabic dialects has become a prominent focus within the computational linguistics community. A major challenge in this area is the scarcity of resources, particularly parallel corpora. Early research in Arabic dialect machine translation primarily relied on rule-based linguistic methods that do not require large datasets. For example, [8] utilized MAGEAD to translate Tunisian Dialect (TD) verbs by incorporating specific prefixes and suffixes, achieving an accuracy of 75%.

[21] advanced this research by developing a translator from TD to MSA. They analyzed various TD corpora to identify morphological features at both the sentence and word levels, using a parallel dictionary and local grammars that were later transformed into finite-state transducers via the NooJ platform. Building on this work, [20] used the same platform to propose a linguistic approach to translate TD into MSA. Their method involved creating inflectional, morphological, and syntactic grammars using finite-state transducers developed from specialized dictionaries. However, improving the accuracy of rule-based systems requires continuous refinement of the rules using linguistic expertise, which is both resource-intensive and time-consuming.

To address these limitations, researchers shifted towards statistical approaches that leverage parallel corpora. One significant contribution in this area is the ”Parallel Arabic Dialectal Corpus: PADIC,” developed by [14], which includes Maghrebi (Algerian, Tunisian, and Moroccan) and Levantine (Palestinian and Syrian) dialects alongside MSA. This corpus has supported the development of several translation systems for various language pairs. Multilingual machine translation experiments were conducted using PADIC, and the authors explored the impact of incorporating large monolingual Arabic corpora on dialect-to-MSA translation quality. Additionally, (Harrat et al., 2019) focused on translating TD into French using statistical methods, achieving an 86% BLEU score with a dataset of 39,000 Tunisian words. Despite progress in statistical methods, research employing deep learning (DL) techniques remains limited due to the lack of resources for under-resourced languages like Arabic dialects. One notable study by [1] applied a recurrent neural network (RNN) encoder-decoder model to translate the Jordanian Arabic dialect into MSA using a small manually created dataset. Likewise, [4] used a combination of data from the PADIC corpus and the multi-dialect parallel corpus of Arabic (MPCA) to collect around 20,000 parallel sentences from Levantine dialects (Jordanian, Syrian, and Palestinian). They proposed a multi-task learning model where each source language had its own encoder, but all languages shared a common decoder. This approach demonstrated promise in improving dialect translation performance.

More recently, researchers have explored data augmentation techniques to enhance dialect translation. For instance, [9] studied Egyptian dialect-to-English translation using lexical replacements within parallel corpora to improve machine translation, language modeling, and automatic speech recognition tasks. Their work contributed to the creation of the first Arabic-English parallel corpus, which addressed resource shortages. Additionally, [17] applied transductive transfer learning for neural machine translation of the Algerian dialect. By testing sequence-to-sequence models with and without attention mechanisms, they demonstrated substantial improvements in BLEU scores, with an increase from 0.3 to over 34 for Seq2Seq models and from less than 17 to over 35 for models with attention mechanisms.

Further research by [2] examined dialect-to-MSA translation using a combination of fine-tuning the AraT5 model with the MADAR dataset, inference using the NLLB model, and prompting with GPT-3.5. Their results showed that GPT-3.5 significantly outperformed the other models, achieving a BLEU score of 29.61 compared to AraT5 and NLLB, which scored 10.41 and 11.96, respectively. This suggests that GPT-3.5 is particularly effective in handling dialectal variations, underscoring its

potential for enhancing Arabic dialect machine translation systems.

Moreover, advancements in deep learning for resource-scarce languages can be seen in work such as [2], which focused on translating static gestures of Thai Sign Language into spoken words using a deep learning model. This research aimed to develop a compact, efficient solution for mobile devices, and experimental results demonstrated superior performance over existing models.

This growing body of research reflects the increasing attention being given to Arabic dialect translation, highlighting the need for continued innovation in both linguistic and statistical methods to overcome resource limitations and improve translation quality across dialects.

Table 1. Examples of written forms in the Tunisian dialect

Sources	Formes
Tunisian constitution	تونس دولة مدنية (It is a civil state) و المؤاطنّات يعني دولة يحكموها المؤاطنون (It means a state that is governed by its citizens.) عن طريق نّاس ينتخبوهم بكل حرية (Through a process where people elect them with complete freedom.)
Social networks	wlh 7el9a graaave :'((I swear, the situation is serious) stop cv pas e5iiii (Stop, no way, brooo) اله يلطف بنّا و ربي يهدي و برّا (May Allah be kind to us, may my Lord guide and forgive)

3 Methodology

3.1 Tunisian Arabic Dialect

The Tunisian Arabic dialect (TD) originated from the fragmentation of 7th-century Arabic and was further enriched through centuries of linguistic influences, shaped by military conquests and population migrations. Reflect elements of South Arabian, Berber, African languages and others that have contributed to its distinctive features. Historically, TD was predominantly an oral language, with noticeable regional variations. However, advancements in technology and socio-political changes have progressively transformed TD into a more versatile medium, now extending beyond spoken communication. Today, TD manifests in several forms, which can be broadly categorized into three groups: original texts written in TD, transcriptions of oral speech, and specialized dictionaries or thesauri dedicated to the dialect.

3.1.1 Written Form

The written form of TD can be divided into two subcategories. The first subcategory encompasses colloquial language written with proper orthography, such as that found in official documents like the Tunisian constitution or in the scripts of Tunisian television series and films. The second subcategory is represented by content on blogs and social media, which is much more diverse and heterogeneous in nature. This form often exhibits inconsistencies in spelling and writing conventions. Table 1 provides examples of each of these written categories.

Additionally, a significant written form of Tunisian Dialect (TD) emerges from the transcription of spoken language. Transcription corpora have been developed to support automatic speech processing tasks, including speech recognition and parsing. However, these transcriptions face challenges beyond the lack of standardized orthography. They often contain imperfections that stem from natural speech disfluencies, such as hesitations, repetitions, and overlapping speech segments, which can further complicate their processing and interpretation. These issues introduce variability into the corpora, affecting the accuracy of automatic systems trained on them. Figure 1 illustrates an example of transcribed Tunisian dialect, highlighting the complexity introduced by such disfluencies.

3.1.2 Other Forms

In addition to corpora, TD has also been provided with resources such as dictionaries [6], ontologies [7], syntactic trees, etc. The most well-known dictionary in TD is the fig dictionary [1] : كرموس

[1] https://www.scribd.com/doc/36699120/Dico-karmous

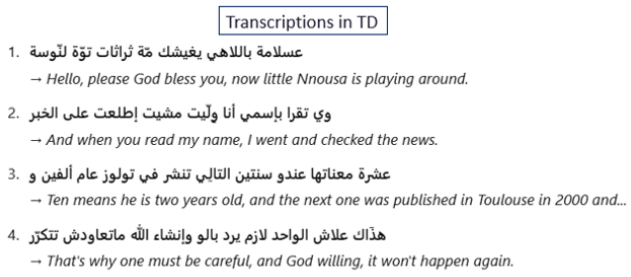

Fig. 1. Extract of transcription form of the Tunisian dialect

Lettre Tad - حرف ط

tampon		*tabe' (tabe3)*	طابع
un morceau de sucre		*tabe' sokkor (tabe3 sokker)*	طابع سكّر
capuchon		*tabou*	طابو
tabouret		*tabouriya*	طابورية
- pain rond traditionnel - four en terre cuite		*tabouna*	طابونة
haie naturelle, rangée de cactus bordant une terre agricole	hedge	*tabya*	طابية
plat entre l'omelette et la quiche, sans pâte et cuit au four		*tajin*	طاجين
tomber		*tah (ta7)*	طاح
voler		*tar*	طار

Fig. 2. A fig dictionary excerpt

krmws. An excerpt from this dictionary is shown in Figure 2.

3.2 Overview of the Machine Translation Models Used

In this paper, we will test the effectiveness of different neural architectures such as standard and transformers architectures.

3.2.1 Standard Models

To evaluate the neural translation of the Standard model, we employed the same architectures presented in [13], specifically Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN). We first implemented an encoder-decoder model, also known as a sequence-to-sequence (seq2seq) model, using the RNN architecture. This model was constructed with a three-layer encoder-decoder based on two commonly used designs: one utilizing Long Short-Term Memory (LSTM) cells and the other using Gated Recurrent Units (GRU). The performance of both LSTM and GRU cells was tested under identical hyper-parameters for comparative analysis. Given that the encoder-decoder architecture can be enhanced by incorporating an attention mechanism between the encoder and decoder [3], we introduced an attention mechanism into the decoder for both models. The attentional model consists of three dense layers, with the trained decoder model comprising an attention layer, an integration layer, a GRU/LSTM layer, and a dense layer.

For the RNN models equipped with attention mechanisms, we utilized the Adam optimizer with beta1 = 0.9, beta2 = 0.98, and epsilon = 10e-9. In contrast, for RNN models without attention mechanisms, we applied the RMSprop optimizer with a constant learning rate of 0.001 to minimize error values. Dropout regularization was used on each layer, with a probability of 0.2 for hidden layers and 0.1 for both input and output layers.

In the second part of our experiment, we replaced the recurrent encoder with a convolutional encoder that includes residual connections, combined with LSTM layers for sequence prediction. CNN approaches, known for their ability to capture long-term dependencies in sequential data, require multiple layers to effectively process these dependencies. We implemented five layers with small parameter values (e.g., larger integration dimensions and longer sentence lengths) to create a sufficiently deep architecture for the dataset used. More details on these models and their parameters can be found in [13].

3.2.2 Fine-Tuning of Transformer Models

Since 2017, the introduction of the ”Transformer” architecture has revolutionized neural network models [22]. This architecture eliminates the need for both convolutional and recurrent layers, replacing them entirely with self-attention mechanisms to capture dependencies between inputs and outputs. This innovation enables the Transformer to more efficiently compute representations, significantly accelerating the learning process.

Several Transformer variants have been developed, each offering distinct configurations of input representation and multi-head attention mechanisms. To identify the most effective

model for translating Tunisian Dialect (TD) to English, we fine-tuned multiple Transformer-based models across different frameworks, including Basic Transformer, Fairseq, OpenNMT, and Joey NMT. The primary distinction between these models lies in their ability to process the entire input sequence simultaneously, rather than incorporating it sequentially, word by word:

— **Basic Transformer:** This is the original Transformer architecture, which relies on a multi-headed self-attention mechanism to capture latent spatial representations of the input and output. We fine-tuned this model using the tensor2tensor [16] implementation, following the specifications in [22]. The model was trained using a batch data generator, grouping sentences effectively. We prepared the dataset with the "tf.data" library and employed the "batch-ON-slices" function to manage the input and output sequences.

— **Fairseq:** Fairseq, developed by Facebook, provides implementations of translation models along with language templates and scripts for custom training. We fine-tuned the integrated Transformer model in Fairseq, adapting it to better handle the TD-to-English translation task.

— **OpenNMT:** OpenNMT is known for its effectiveness in translating low-resource languages by allowing the transfer of parameters from high-resource language models (parent models) to low-resource languages (child models) [23]. To tailor OpenNMT for TD translation, we fine-tuned both its standard and Transformer-based architectures. The standard OpenNMT architecture was configured with a two-layer LSTM, each with 512 hidden units for the encoder and decoder. The Transformer-based OpenNMT model was fine-tuned with 6 layers, each with a hidden size of 512, following the default Transformer settings [22]. Both models were trained for 20,000 epochs to optimize performance.

— **JoeyNMT:** JoeyNMT has also shown strong results in translating low-resource languages [19]. We fine-tuned the JoeyNMT model following the setup described in [13]. The model utilized an encoder-decoder architecture with 4 attention-based layers. We trained this model for 50 epochs, applying a dropout rate of 0.2, a hidden dimension of 1024, and a batch size of 512.

Across all Transformer models, the ADAM optimizer was employed with beta1=0.9, beta2=0.98, and epsilon=10e-9. We applied a dropout rate of 0.3 across all experiments to reduce the likelihood of overfitting and to stabilize training.

3.3 Machine Translation of Tunisian Dialect

- Data Preparation:

Deep Learning (DL) methods have demonstrated promising results on large bilingual corpora within the field of Natural Language Processing (NLP). To assess the effectiveness of these methods for translating Tunisian Dialect (TD) into English (ENG), a large parallel TD-ENG corpus is essential. However, such a resource is scarcely available, and only a few studies have attempted to compile parallel TD-ENG data. The Multi-Dialect Arabic Parallel Corpus (MPCA) [5] is one of the few available resources, consisting of approximately 2,000 parallel TD-MSA-ENG sentences, sourced from social networks, alongside other Arabic dialects.

To address this limitation, we propose a semi-automatic approach to create a parallel TD-ENG corpus. Since Modern Standard Arabic (MSA) is the parent language of TD, we use MSA as a pivot to generate ENG sentences from TD sentences. As illustrated in Figure 3, our method involves first creating a parallel TD-MSA corpus by utilizing existing resources and applying data augmentation techniques. We then apply an MSA-ENG translation model to translate the MSA sentences into English, resulting in a TD-MSA-ENG corpus.

- Existing TD-MSA corpus: The proposed method to build the training corpus consist to collect, firstly, the existing corpus. The first parallel TD-MSA corpus is the one presented in [13]. In this work, authors are collected 11.8k sentences TD-MSA from the existing parallel

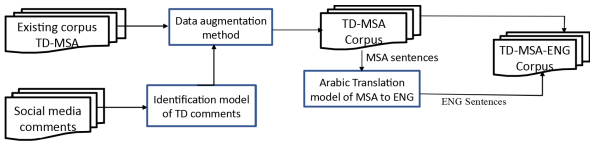

Fig. 3. The proposed method to create the training corpus

corpora described in [12] such as the MADAR corpus [10], PADIC [14], Tunisian Constitution and scraped comments from social networks [12]. Then, they are proposed a series of textual and linguistic augmentation methods to obtain 170k parallel sentences TD-MSA. In order to assess these augmentation techniques' efficacy, they are used a statistical translation model. The ensuing data increase enhances the latter's performance.
The second existing resources is the parallel TD-MSA corpus presented in [11]. Authors of the latter are manually translated 3k TD comments scraped from social networks into MSA. They also proposed a data augmentation method based on words substitution of the existing corpus [12] to get a 15k new parallel sentences.
As a results, we obtain 185k parallel sentences TD-MSA from the existing corpora.

- Data Augmentation Method: Using the collected corpus, which comprises over 180,000 parallel TD-MSA sentences, we propose an augmentation method to generate additional TD-MSA parallel comments, incorporating specific characteristics of Tunisian text, such as code-switching and the use of Arabizi. To achieve this, we employ a Tunisian Dialect identification model to select TD comments from social networks. The model used is the Tunisian Arabic Dialect Identification (TADID) model, trained with the MULTI-DIALECT-ARABIC-BERT classifier [11], which achieved an accuracy of 93.18% on data annotated with three categories: Tunisian (TUN), MSA, and OTHER for sentences in other languages. Using this model, we identified 3,000 TD comments from a set of 5,000 comments scraped from Facebook and YouTube. These selected comments were then translated using the TD-MSA translation models tested in [13], specifically the JoeyNMT model, which achieved the highest BLEU score (60%). As this model was trained on data written in Arabic script, we utilized the "3aransia" Arabic dialect transliteration library[2] to convert Latin-scripted comments into Arabic script, enabling their translation with the same model.

Subsequently, we added the original comments identified by the TADID model, along with their MSA translations produced by JoeyNMT, to the existing corpus. This resulted in an enhanced version of the corpus, now containing comments in code-switched formats with both Arabic and Latin scripts. As a result, the corpus has been expanded to include 188,000 parallel TD-MSA sentences. This corpus will be made publicly available on GitHub[3] to facilitate further research and processing of both Tunisian Dialect and MSA.

- TD-MSA-ENG Corpus: Given the absence of a direct TD-ENG parallel corpus, Modern Standard Arabic (MSA) was utilized as an intermediary language to facilitate the translation between Tunisian Dialect (TD) and English (ENG). The availability of extensive parallel resources and pre-trained models for MSA-ENG translation provides a viable approach for generating English sentences from the TD-MSA corpus.

We evaluated several pre-trained models to determine the most effective tool for translating MSA to English. The models tested include DL Translate, TextBlob, mBART, Google Translate, and PyTranslate. Each of these models leverages different underlying architectures and training data, as summarized below:

— **DL Translate**: A machine translation library based on Hugging Face's Transformer, utilizing models developed by Facebook AI Research. It is built upon multilingual BART [18] and supports translation across more than 100 languages.

— **TextBlob**: A Python library for NLP that leverages the Google Translate API for translation tasks.

[2] https://github.com/3aransia/3aransia
[3] https://github.com/sk-cmd/ressources-parallele-DT-ASM

Table 2. BLEU scores of pre-trained models for MSA to English translation

Model	BLEU Score (%)
DL Translate	**70.01**
TextBlob	57.00
mBART	66.80
Google Translate	68.10
PyTranslate	64.95

Table 3. Statistics of the TD-MSA-ENG corpus

Corpus	#Lines	#TD Vocabularies	#MSA Vocabularies	#ENG Vocabularies
Train corpus	188k	35.6k	33.7k	32k
Test corpus	2k	10.7k	9.8k	7.9k

— **mBART**: A multilingual encoder-decoder model designed for machine translation, trained on large monolingual corpora across various languages [18].

— **Google Translate**: A widely-used translation tool accessed via a free Python library that implements the Google Translate API.

— **PyTranslate**: A Python package offering translation functionality with support for multiple languages.

To assess the translation quality of these models, we used a test set derived from the MSA-ENG segment of the MPCA corpus. The BLEU scores obtained for each model are presented in Table 2.

As indicated in Table 2, the DL Translate model achieved the highest BLEU score (70.01). Consequently, this model was selected for generating the English translations in the TD-MSA-ENG corpus.

- Data Preprocessing: Using this approach, we generated a total of 188k English sentences. Table 3 provides additional statistics on the created TD-MSA-ENG corpus. Of this, 16k sentences were set aside for the development set (DEV), with the remaining sentences used for training. The MPCA corpus was employed as a consistent test set, including 1k TD comments, which were converted to Latin script using the ”aaransia” library to capture a broader representation of TD.

Before feeding the data into translation models, we applied Byte Pair Encoding (BPE) to compress the data by iteratively replacing frequent byte pairs with new tokens, allowing us to split words into subwords based on frequency. We used the subword_nmt library[4] for this segmentation process. Additionally, the multi-dialect-bert-base-arabic tokenizer was employed to encode the data into sequences of word IDs for input into the Transformer models.

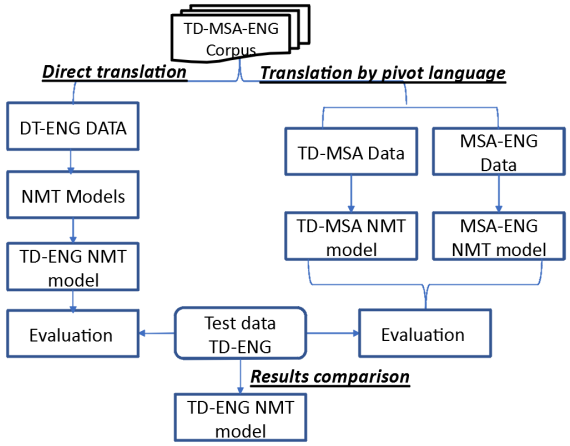

Fig. 4. Proposed method for TD-ENG translation model

4 Machine Translation Experiments and Results

The aim of this work is to identify a familiar language on social networks, in this case the TD, and translate it into a standard foreign language: English. We examine the effectiveness of two methods for building a translation model for TD-ENG using the TD-MSA-ENG corpus. The first method entails developing model using MSA as the pivot language while the second method involves translating TD to English without MSA language. The most effective TD-ENG model is then determined by comparing the output of the different proposed models. Figure 4 summarizes the proposed approach.

[4] https://github.com/rsennrich/subword-nmt

Table 4. BLEU scores of TD-ENG models

Models Names	BLEU scores on Dev set	BLEU scores on Test set
(1) RNN + LSTM	22.00	20.90
(2) RNN + GRU	20.00	18.80
(3) RNN + LSTM + attention	30.00	29.00
(4) RNN + GRU + attention	28.90	26.70
(5) CNN + LSTM + attention	31.52	30.58
(6) Basic OpenNMT	33.50	31.80
(7) Transformer OpenNMT	36.82	36.01
(8) Transformer Fairseq	41.98	**40.15**
(9) Transformer Joey NMT	39.45	39.11

Table 5. BLEU scores of TD-MSA Translation Models

Models Names	% Scores (Dev set)	% Scores (Test set)
(1) RNN avec LSTM	24.80	23.96
(2) RNN avec GRU	22.28	20.25
(3) Attention + RNN avec LSTM	34.82	32.27
(4) Attention + RNN avec GRU	23.26	20.12
(5) Attention + CNN avec LSTM	44.02	43.94
(6) Transformer de Base	44.98	42.95
(7) Basic OpenNMT	48.50	46.70
(8) Transformer OpenNMT	57.46	56.23
(9) Fairseq	62.09	61.22
(10) Joey NMT	**66.25**	**65.92**

Table 6. BLEU scores of Transformers models MSA-ENG

Models names	BLEU scores on Dev set	BLEU scores on Test set
(1) Transformer Fairseq	**51.5**	**50.9**
(2) Transformer Joey NMT	50.09	49.98

4.1 Translation with a Direct Transition from TD to ENG

To fix a translation model from TD to ENG, we use the TD-ENG parallel data of our created TD-MSA-ENG corpus to train the different translation models described above. The table 4 presents the BLEU scores obtained for each model on the TD-ENG data of the MPCA test corpus.

As shown in Table 4, the Transformer models outperformed the standard models (RNN and CNN). Notably, the basic Transformer model yielded the lowest BLEU scores among the Transformer architectures, likely due to the absence of BPE pre-processing. Other Transformer models, which utilized BPE-coded input, demonstrated the effectiveness of data compression through BPE in translating Tunisian Dialect, particularly for handling out-of-vocabulary words in NMT models. Furthermore, within the OpenNMT models, the 'Transformer' architecture consistently outperformed the basic architecture, corroborating the superior performance of Transformer over standard architectures, as previously asserted by the GoogleAI team[5]. Among the models tested, the Fairseq and Joey NMT scores were very close, with Fairseq's Transformer delivering the best result (40.15). Thus, we consider Transformer Fairseq as the model for our TD-ENG translation.

4.2 Translation with the MSA pivot Language

In this section, we train the same models of the TD-ENG translation using the TD-MSA and MSA-ENG data to define a TD-MSA and an MSA-ENG model in order to analyze the translation from TD to ENG by switching from MSA as a pivot language.

- TD-MSA Translation Model: Table 5 shows the performance of standard and Transformers models for TD-MSA translation. The BLEU scores obtained by standard models are consistent with the scores obtained in [13]. Indeed, the power of the Encoder-Decoder architecture has been improved by adding an attention mechanism. Thus, the CNN model formed by an attention mechanism and which uses the LSTM memory is the most adequate model to translate the TD to the MSA. He was able to achieve a BLUE score equal to 54.97% on the entire Test. For the Transformers models, we notice that the Joey NMT model is the model that shows the correct translation. This model was able to achieve 65.92% as BLEU score. Following these results, we choose to use Transformer Joey NMT as a translation model from TD to MSA.

- MSA-ENG Translation Model: Based on previous experiments, with MSA-ENG data, we test Transformers Joey NMT, Fairseq models. These models are the best models that gave the best BLEU score in the context of this work. We adopt two orientations given the richness of the MSA and ENG languages in terms of resources. The first orientation consists in training the Fairseq and Joey NMT models by the MSA-ENG data of our created corpus. Table 6 describes the obtained results.

Fairseq and Joey NMT models give close BLEU scores. The highest score is obtained with Fairseq (50.9%). This supports the fact that it is the most used model in the literature for the translation of poorly endowed languages.

[5]https://medium.com/analytics-vidhya/Transformer-vs-rnn-and-cnn-18eeefa3602b

Table 7. BLEU scores of Transformer MSA-ENG models with corpus enrichment

Models names	BLEU scores on Dev set	BLEU scores on Test set
(1) Transformer Fairseq	**70.15**	**69.95**
(2) Transformer Joey NMT	69.8	68.90

Table 8. BLEU scores of Fairseq model for TD-ENG translation

Fairseq	%BLEU scores on Test set
With direct passage	41.78
With MSA as a pivot language	**67.89**

The second orientation consists in enriching the MSA-ENG data of the created corpus with 200k parallel sentences MSA-ENG existing in the corpus of tweets [6], the United Nations corpus [7] and the TED parallel corpus[8] . Then, we re-train the Fairseq and Joey NMT models. Table 7 describes the obtained BLEU scores. For each of these models, the BLEU scores increased. Moreover, the best result is obtained with the Fairseq model. Therefore, we consider that the Fairseq model trained by the MSA-ENG enriched corpus is that of the MSA-ENG translation model.

Following results of these orientations, we evaluate the Fairseq model for TD-ENG translation through MSA as a pivot language. We project this model for the MSA-ENG translation on the MSA sentences resulting from the translation of the TD sentences with JoeyNMT model of the TD-MSA translation. Table 8 describes the obtained results. TD-ENG Direct Translation Model BLUE score is improved by 40.15% up to 67.89%. Therefore, the model using MSA as a pivot language is the most efficient model for translating TD into a foreign language.

5 Discussion of Results

This study investigates neural machine translation from Tunisian Dialect (TD) to English (ENG), evaluating various configurations of translation models. The most successful approach was achieved by using Modern Standard Arabic (MSA) as a pivot language, implemented through the Fairseq model, which resulted in a BLEU score of 67.89%. This strategy involved two main modules: the first translating TD into MSA, which attained a BLEU score of 66.22%, and the second translating MSA into ENG, with the MSA-ENG module's initial BLEU score of 51.42% rising to 70.2% after data enrichment. These findings suggest that adding more parallel MSA-ENG data could further enhance translation quality.

For instance, the sentence "¡h_dA brnAmj tAfh¿" (It's a stupid program) was initially mistranslated as "This is an empty agenda" by some models. After refinement and additional data, the translation was corrected, highlighting how expanding data resources and fine-tuning can resolve challenges, particularly with ambiguous or complex phrases typical of TD, a dialect known for its variability.

Our results are consistent with existing literature on TD translation. The performance of our models is promising compared to earlier works. For example, the highest reported BLEU score for TD-to-French translation reached 86% (Harrat et al., 2018), while research focusing on TD-to-ENG translation remains extremely limited. The only related study involved a multi-dialect English translation model that included TD [15]. Additionally, most previous studies have concentrated on translating TD to MSA, with a rule-based method achieving a BLEU score of 84% for TD verbs [8]. Comparatively, in other Arabic dialects, the highest BLEU score for Jordanian dialect to MSA translation using a neural approach reached 48% [4].

In conclusion, using MSA as a pivot language significantly enhanced translation accuracy and robustness, proving to be an effective strategy for addressing the challenges of TD-to-ENG translation. Future research should aim to further expand the TD-ENG corpus and explore advanced translation techniques to continue improving performance.

6 Conclusion

This research contributes to the field of machine translation for low-resource languages, with a focus on translating comments in Tunisian Dialect

[6]https://alt.qcri.org/resources/bilingual_corpus_of_parallel_tweets/
[7]https://conferences.unite.un.org/uncorpus
[8]https://www.clarin.eu/resource-families/parallel-corpora

(TD) from social networks into a standardized foreign language. We began by conducting a linguistic analysis of TD, followed by a comprehensive review of machine translation methodologies for low-resource languages. This analysis informed the development of an automatic translation approach from TD to English (ENG) using deep learning techniques.

The foundation of our approach involved constructing a parallel TD-MSA (Modern Standard Arabic) corpus to leverage the extensive resources available for MSA and standardize TD. To complement the existing TD-MSA parallel resources, we introduced a data augmentation strategy to expand the corpus. This augmented MSA data was then used alongside pre-trained models to produce a trilingual TD-MSA-ENG parallel corpus.

Our focus was on developing neural translation models for TD to MSA and subsequently from TD to ENG. We trained three types of neural models—a convolutional model, a recurrent model, and a Transformer model—across various corpus configurations, taking into account the linguistic nuances of social network data. Experimental results revealed that the optimal configuration was achieved using the Joey NMT Transformer model, which reached a BLEU score of 50.92% for TD-MSA translation.

For the TD-to-ENG translation task, employing MSA as a pivot language yielded superior results, with a BLEU score of 67.89%. This outcome underscores the effectiveness of the pivot language approach in addressing the challenges of low-resource dialect translation.

All resources developed in this study are available on GitHub[9]. In the short term, we plan to apply further data augmentation techniques, such as GPT and TF/IDF-based transformations, and inject common social media errors and ambiguities to further expand the corpus. Additionally, we aim to develop a multi-dialectal translation model to create a more generalized system capable of adapting to various Arabic dialects.

[9]https://github.com/sk-cmd/ressources-parallele-DT-ASM

References

1. **Al-Ibrahim, R., Duwairi, R. M. (2020).** Neural machine translation from Jordanian dialect to modern standard Arabic. 2020 11th International Conference on Information and Communication Systems (ICICS).
2. **Atwany, H., Rabih, N., Mohammed, I., Waheed, A., Raj, B. (2024).** OSACT 2024 task 2: Arabic dialect to MSA translation. **Al-Khalifa, H., Darwish, K., Mubarak, H., Ali, M., Elsayed, T.**, editors, 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, ELRA and ICCL, Torino, Italia, pp. 98–103.
3. **Bahdanau, D., Cho, K., Bengio, Y. (2016).** Neural machine translation by jointly learning to align and translate.
4. **Baniata, L., Park, S., Park, S.-B. (2018).** A neural machine translation model for Arabic dialects that utilizes multitask learning (mtl). Computational Intelligence and Neuroscience.
5. **Bouamor, H., Habash, N., Oflazer, K. (2014).** A multidialectal parallel corpus of Arabic. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14).
6. **Boujelbane, R., Ellouze Khemekhem, M., Belguith, L. H. (2013).** Mapping rules for building a Tunisian dialect lexicon and generating corpora. Proceedings of the Sixth International Joint Conference on Natural Language Processing, Asian Federation of Natural Language Processing, Nagoya, Japan.
7. **Graja, M., Jaoua, M., Belguith, L. H. (2011).** Building ontologies to understand spoken Tunisian dialect. CoRR.
8. **Hamdi, A., Boujelbane, R., Habash, N., Nasr, A. (2013).** The effects of factorizing root and pattern mapping in bidirectional Tunisian - Standard Arabic machine translation. Proceedings of the MT Summit, Nice, France.

9. **Hamed, Injy and Habash, Nizar and Abdennadher, Slim and Vu, Ngoc Thang (2022).** Investigating lexical replacements for Arabic-English code-switched data augmentation.
10. **Houda, B., Habash, N., Salameh, M., Zaghouani, W., Rambow, O., Abdulrahim, D., Obeid, O., Khalifa, S., Eryani, F., Erdmann, A., Oflazer, K. (2018).** The MADAR Arabic dialect corpus and lexicon. Proceedings of the 11th Language Resources and Evaluation Conference.
11. **Kchaou, S., Boujelbane, R., Fsih, E., Hadrich-Belguith, L. (2022).** Standardisation of dialect comments in social networks in view of sentiment analysis: Case of Tunisian dialect. Proceedings of The 13th Language Resources and Evaluation Conference (LREC), European Language Resources Association, Marseille, France.
12. **Kchaou, S., Boujelbane, R., Hadrich-Belguith, L. (2020).** Parallel resources for Tunisian Arabic dialect translation. Proceedings of the Fifth Arabic Natural Language Processing Workshop, Association for Computational Linguistics, Barcelona, Spain (Online), pp. 200–206.
13. **Kchaou, S., Boujelbane, R., Hadrich-Belguith, L. (2022).** Hybrid pipeline for building Arabic Tunisian dialect-standard Arabic neural machine translation model from scratch. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP).
14. **Meftouh, K., Harrat, S., Jamoussi, S., Abbas, M., Smaili, K. (2015).** Machine translation experiments on padic: A parallel Arabic dialect corpus. Proceedings of 29th Paclic Asia Conference on Language, Information and Computation.
15. **Sawaf, H. (2010).** Arabic dialect handling in hybrid machine translation. Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA), Denver, Colorado.
16. **Shaw, P., Uszkoreit, J., Vaswani, A. (2018).** Self-attention with relative position representations. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.
17. **Slim, A., A.and Melouah, Faghihi, U., al. (2022).** Improving neural machine translation for low resource Algerian dialect by transductive transfer learning strategy.
18. **Tang, Y., Tran, C., Li, X., Chen, P.-J., Goyal, N., Chaudhary, V., Gu, J., Fan, A. (2020).** Multilingual translation with extensible multilingual pretraining and finetuning. arXiv preprint arXiv:2008.00401.
19. **Tapo, A. A., Coulibaly, B., Diarra, S., Homan, C., Kreutzer, J., Luger, S., Nagashima, A., Zampieri, M., Leventhal, M. (2020).** Neural machine translation for extremely low-resource african languages: A case study on bambara. CoRR.
20. **Torjmen, R., Haddar, K. (2021).** Translation system from Tunisian dialect to modern standard Arabic. Concurrency and Computation: Practice and Experience.
21. **Torjmen, R., Hammouda, N. G., Haddar, K. (2019).** A NooJ Tunisian dialect translator. NooJ.
22. **Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I. (2017).** Attention is all you need. Advances in neural information processing systems.
23. **Zoph, B., Yuret, D., May, J., Knight, K. (2016).** Transfer learning for low-resource neural machine translation. arXiv preprint arXiv:1604.02201.

Article received on 19/09/2023; accepted on 15/02/2025.
**Corresponding author is Sameh Kchaou.*