Natural Language Processing (2026), **32**, pp. 204–223
doi:10.1017/nlp.2026.10018

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Maghrebi dialects – Arabic bidirectional translation: an improved transformer with transfer learning

Jihad R'baiti[1], Youssef Hmamouche[1] and Amal El Fallah Seghrouchni[1,2]

[1]International Artificial Intelligence Center of Morocco, University Mohammed VI Polytechnic, Rabat, Morocco and [2]LIP6 - UMR 7606 CNRS, Sorbonne University, Paris, France
Corresponding author: Jihad R'baiti; Email jihad.rbaiti@um6p.ma

(Received 3 January 2024; revised 2 January 2026; accepted 12 February 2026; first published online 10 March 2026)

Abstract

Neural Machine Translation (NMT), a subfield of Natural Language Processing, has seen significant advancements with the emergence of transformer architectures and generative artificial intelligence, demonstrating remarkable performance in various languages. However, translating Arabic dialects remains a notable challenge that becomes very pronounced primarily due to their morphological complexity and divergence from standardised grammatical rules. In this paper, we present a hybrid approach for translating the Maghrebi dialects into/from Modern Standard Arabic (MSA). The approach takes advantage of the strengths of the transformer architecture and the BERT language model for transfer learning of representations. To achieve this, we incorporated BERT embeddings into the encoder and decoder stacks of the transformer architecture. The BERT architecture, which we utilised, was trained in a self-supervised manner on Maghrebi dialects and Arabic corpora. The resulting BLEU/BERTScore/ChrF/METEOR scores for the approach were 14.148/79.414/28.885/28.428 and 8.961/20.994/19.465 (BLEU/ChrF/METEOR) for the translation in both directions using the raw data, demonstrating competitive performance compared to ChatGPT and Gemini Large Language Models (LLMs). Furthermore, we evaluated the approach using an ablation study with fine-tuned NLLB-200 and against three combinations of tokeniser techniques used in conjunction with the transformer architecture: Byte-Pair Encoding (BPE) tokeniser, WordPiece tokeniser, and BERT tokeniser. Both evaluations, including human evaluation, confirm the efficacy of our method.

Keywords: Machine translation; Arabic dialects; transformers; transfer learning; low-resource languages

1. Introduction

Artificial intelligence is a powerful domain leveraging the potential of computer science and mathematics to create intelligent systems that enhance various aspects of our modern life, including the economy, education, industry, and social fields. Machine translation (MT) systems have already been proposed on many online platforms. However, there is a notable deficiency in the translation of Arabic dialects, resulting in several limitations, such as hindering progress in the industry and impeding the integration of the Arab world.

Machine translation is a system developed to bridge the communication gap between people from different regions and linguistic backgrounds. It employs various approaches, including rule-based MT (Shiwen and Xiaojing 2014), example-based MT (Somers 1999), statistical MT (Zens, Och, and Ney 2002; Lopez 2008), and NMT. NMT relies on neural networks to learn complex patterns and alignments between words in different languages. It has proven highly effective, outperforming previous approaches in a wide range of language pairs and domains (Luong *et al* 2015; Almahairi *et al.* 2016; Wu *et al.* 2016; Stahlberg 2020). The neural system uses two stacks in its internal architecture: an encoder and a decoder. These components work together to transform

a sequence of words into new word sequences representing the translation output. Despite the vast amount of existing literature on NMT, the majority of studies primarily focus on Latin languages due to the availability of resources and data (Bahdanau, Cho, and Bengio 2014; Sutskever, Vinyals, and Le 2014; Vaswani *et al.* 2017). Translating from Standard Arabic into Maghrebi (North African) dialects and vice versa presents a challenging but under-researched topic. The Maghrebi region covers a significant population, and addressing this research gap would greatly facilitate communication among its inhabitants. Moreover, advancing translation would play a crucial role in analysing texts written in Maghrebi dialects, particularly from social media platforms. Translating to and from Maghrebi dialects poses several challenges. Firstly, the Maghrebi dialects present significant distinctions in various linguistic aspects, including vocabulary, pronunciation, and syntactic structures (Younes *et al.* 2020). For instance, the Moroccan dialect features unique lexical items (Skiredj *et al.* 2024) that set it apart from Algerian dialects, which, in turn, differ in syntax and vocabulary from the Tunisian and Libyan dialects. These differences highlight the linguistic diversity within the region, posing significant challenges for translation between the dialects and Standard Arabic. Additionally, these dialects have evolved as distinct languages over time. They are particularly influenced by many languages, notably French, English, and Spanish. In addition, these dialects are generally spoken languages, with written documents being either rare or poorly structured. As a result, developing stable and consistent linguistic resources, such as dictionaries and corpora, becomes challenging. It is important to mention that until now, existing translation tools, such as those offered by Google, Microsoft, and DeepL, do not include these dialects. Very powerful commercial LLMs such as ChatGPT and Gemini are able to process these languages, yet achieving precise translation remains a challenge. Therefore, the contributions of this work are as follows:

- We propose BERT-TransDial,[a] a modified Transformer architecture designed to analyse the impact of integrating transfer learning from BERT-based embedding layers for multi-dialect translation.
- We evaluated the performance of BERT-TransDial by comparing it with the vanilla transformer using various tokenisation methods and with the pre-trained NLLB-200-distilled-350 M model. Additionally, we examine how preprocessing techniques influence the experimental results.
- We conducted an ablation study to analyse the effects of modifying architectural components and parameter freezing on model performance.
- A human evaluation was conducted to assess the qualitative results of BERT-TransDial and to compare it with Gemini and ChatGPT.

The rest of this paper is structured as follows: Section 2 provides an overview of related works. In Section 3, we state the problem. Section 4 is devoted to the proposed method, and Section 5 covers the experiments, including data, implementation details, and experimental methods. Sections 3 and 7 present the results and human evaluation study. Finally, in the last two sections, we discuss and conclude our work.

2. Related works

In this section, we present related work on NMT, organised into two subsections. The first one addresses the challenges of translating low-resource Arabic dialects, while the second explores the incorporation of language models into NMT architectures.

[a] Link to source code: https://rb.gy/x8gzm5

https://doi.org/10.1017/nlp.2026.10018 Published online by Cambridge University Press

2.1 Low-resource language

Several studies aimed to enhance MT performance for low-resource languages, beginning with exploring preprocessing techniques on parallel corpora. Muischnek and Müürisep (2019) introduced a set of filtering methods designed to exclude problematic sentences during the training process. These techniques proved to be particularly effective for translating morphologically rich languages, such as Estonian and Latvian. Standard Arabic also possesses rich morphology. To address this, Almahairi *et al.* (2016) employed orthographic normalisation for Arabic, along with morphology-aware tokenisation (Tok + Norm + ATB). They presented a first neural approach for Arabic, resulting in a BLEU improvement of more than 4.98 over the baseline. In another study, Al-Ibrahim and Duwairi (2020) introduced an RNN encoder-decoder approach, conducting experiments on both word-level and sentence-level datasets. To translate from the Jordanian dialect to MSA, the results show an accuracy of 91% using the word-level experiment and 63% using the Seq2Seq dataset. In a study by Farhan *et al.* (2020), a comparison was conducted on two systems of D2SLT (Dialectal to Standard Language Translation) employing attentional sequence-to-sequence models (Luong *et al.* 2015; Sutskever *et al.* 2014) and Google's Neural Machine Translation (GNMT) (Wu *et al.* 2016). Additionally, the authors introduced an unsupervised learning approach to translating from Jordanian to MSA. They trained each system on parallel corpora from Saudi-MSA or Egyptian-MSA and tested the models using the Jordanian-MSA parallel corpus. The highest result in the unsupervised setting is 32.14 BLEU metric and 48.25 in the supervised setting. By extending the focus to low-resource languages, Baniata *et al.* (2021) proposed a neural approach based on a transformer architecture (Vaswani *et al.* 2017). They used the WordPiece tokeniser as a subword unit (Sennrich, Haddow, and Birch 2015) to decompose each word into subwords, thereby enhancing the representation of the attention matrix. This approach has been highly effective in translating Arabic dialects to MSA. Baniata et al. have also proposed a new Reverse Positional Encoding (Baniata, Kang, and Ampomah 2022) layer for the same context, aiming to capture the positions of right-to-left texts and enhance translation quality for Arabic dialects. Furthermore, Slim *et al.* (2022) conducted a comparative study between the sequence-to-sequence approaches using recurrent neural networks with and without attention mechanisms (Luong *et al.* 2015). Given the limited size of the available parallel corpus, their contribution involved employing transductive transfer learning on the aforementioned architectures. This approach aimed to enhance translation quality by transferring the final model weights of the parent model to the child one. Thus, resulting in a remarkable increase in the BLEU score. In Slim and Melouah (2024), they incrementally transfer knowledge from a large set of Arabic dialects – MSA (Grandparent Model) – to Maghrebi dialects – MSA (Parent Model) – and then to the specified low-resource dialects (Child Model), such as Algerian, Tunisian, and Moroccan, using the Transformer and attention-based sequence-to-sequence models. The results obtained using Transformer showed improvements over traditional transfer learning methods of 80%, 62%, and 58% for Algerian, Tunisian, and Moroccan dialects, respectively. While Kchaou *et al.* (2023) focused on enhancing the translation of Tunisian Dialects-MSA by employing various data augmentation techniques, including the Back-Translation Augmentation Method (BTAM), Contextual Data Augmentation, and Semi-supervised Data Augmentation. Their approach demonstrated an improved BLEU score after augmenting the training data. The evaluation was conducted on a NMT system, specifically the transformers, which achieved the optimal BLEU score, reaching 60. Furthermore, Faheem *et al.* (2024) investigated semi-supervised NMT for translating Egyptian dialects into MSA. They employed two distinct datasets: one parallel dataset containing sentences in both dialects and a monolingual dataset where the source is not connected to the target language during training. They conducted three translation systems: attention-based seq2seq, which learns embeddings using a shared vocabulary between the two languages; the unsupervised transformer model, which depends only on monolingual data; and the last system, which uses a parallel dataset for the supervised learning phase and then adjusts the

https://doi.org/10.1017/nlp.2026.10018 Published online by Cambridge University Press

monolingual data during the learning process. The results show that the semi-supervised learning approach outperformed both the supervised and unsupervised approaches, achieving a BLEU score of 29.5.

2.2 Transfer learning

The study by Clinchant *et al.* (2019) investigated various approaches to integrate BERT, a large pre-trained model, into an NMT system for the English-German translation task. The study explored several approaches, which included substituting the traditional embeddings with BERT parameters, initialising the model with BERT, and subsequently freezing it before proceeding with fine-tuning. These adaptations resulted in a notable increase in the BLEU score when contrasted with the baseline performance. However, it is worth mentioning that this study did not provide a clear conclusion on which approach improves the robustness of target sentence translation.

By incorporating a pretrained language model (PLM) into a sequence-to-sequence model, the authors in Guo *et al.* (2021) introduced a novel architecture that utilises adapter modules to fine-tune the BERT model. During this fine-tuning process, a latent variable 'z' was introduced to the model. This variable enables the automatic selection of the appropriate adapter, thus enhancing its capabilities. Including our previously discussed approaches, Guo and Le Nguyen (2020) conducted a study that leveraged BERT to capture contextual information for document-level translation in both the encoder and decoder of the Transformer NMT model. This involved considering the preceding and following sentences around the current input sentence x_t to generate the target sentence y_t. This method has demonstrated its effectiveness on TED and News datasets. In addition, Kadaoui *et al.* (2023) presented a comparative study between LLMs (Bard, ChatGPT, GPT-4, NLLB, and other commercial systems) in performing MT tasks across ten diverse varieties of Arabic. The results of this study show that GPT-4 and ChatGPT generally achieve better results compared to the others in all experiments. Yang *et al.* (2020) proposed the framework CT_{NMT} to address problems encountered when fine-tuning BERT and GPT2 language models to a transformer architecture for rich-resource languages. They introduced asymptotic distillation, dynamic switching, and rate-scheduled learning techniques to preserve previous pre-trained knowledge, prevent catastrophic forgetting of language model knowledge, and adjust the learning spaces according to a scheduled policy, respectively. The fusion of all these techniques gives the highest scores, with BLEU scores of 30.1 for En-De, 42.3 for En-Fr, and 38.9 for En-Zh corpora.

3. Problem statement

NMT systems transform a sentence into its translated form in another language. The first deep learning models proposed to tackle this issue were introduced in 2013 and primarily relied on the RNN encoder-decoder architecture (Kalchbrenner *et al.* 2013; Cho *et al.* 2014), which converts the input sentence $X = \{x_1, x_2, \ldots, x_n\}$ into a more concise representation denoted as 'c', represented as a fixed-length vector. To generate the next word y_i using a decoder stack, the system considers the input sentence X and the previously generated words $\{y_{t-1}, y_{t-2}, \ldots, y_1\}$, along with the context vector 'c'.

This architecture poses a potential issue when translating long sentences. As proposed by Bahdanau *et al.* (2014), an attention mechanism was introduced to address this concern. It is calculated as a function of each hidden state of the encoder and the previous hidden state of the decoder. While this architecture gives good translation results, it comes at the cost of increased computational power and time consumption due to the use of recurrent neural networks in its internal structure. In contrast, the Transformer, introduced by Vaswani *et al.* (2017), is an innovative architecture primarily built upon attention mechanisms, feed-forward layers, layer normalisation, and residual connections. It excels at parallel processing of sequential data,

https://doi.org/10.1017/nlp.2026.10018 Published online by Cambridge University Press

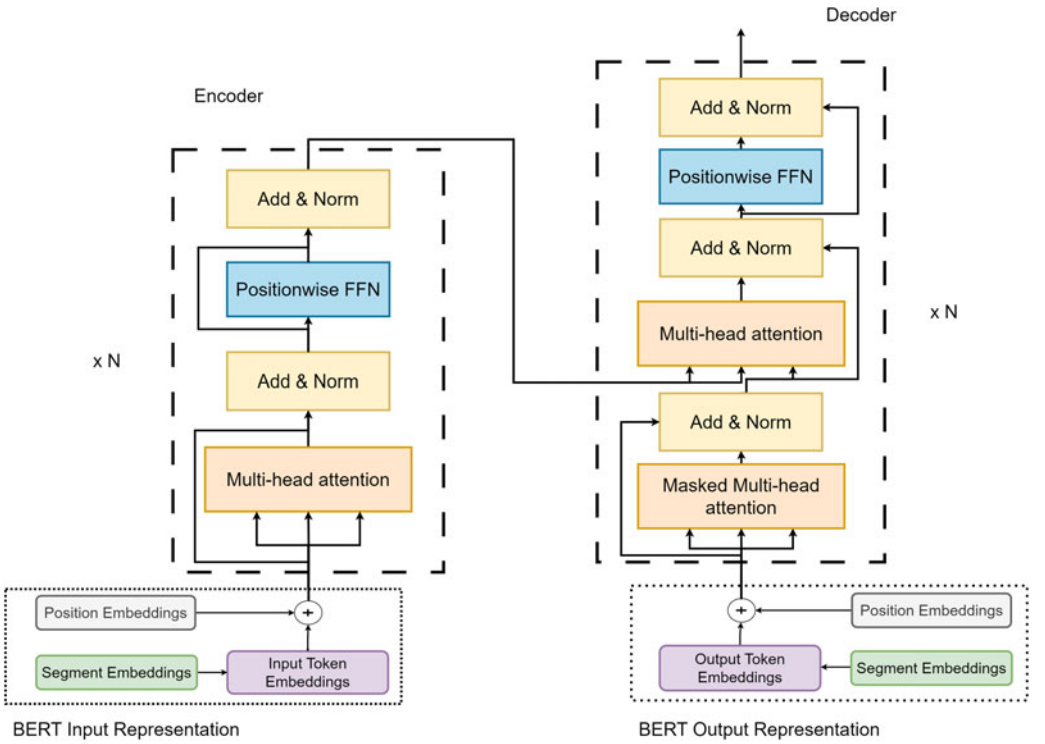

Figure 1. The BERT-TransDial architecture: integration of a fine-tuned BERT-based embedding layer into a Transformer model, replacing the default token embedding and positional encoding components.

resulting in faster training times compared to other models. The NMT problem can be formulated as the following conditional distribution:

$$P(y_t|y_{1,\ldots,t-1}; X; \Theta) \tag{1}$$

Where $X = \{x_1, x_2, \ldots, x_n\}$ represents the input sentence, $\{y_1, y_2, \ldots, y_{t-1}\}$ are previously generated words with $1 < t < m$, where m and n indices are the lengths of Y and X, respectively. Θ represents the model parameters optimised during the training process by minimising the cross-entropy loss function.

4. Approach

In this section, we present the proposed method by describing how the BERT input and output representations are integrated before the original encoder and decoder stacks of the transformer architecture, as depicted in Figure 1.

4.1 BERT input and output representation

In our approach, we employed bidirectional encoder representations from transformers (BERT) (Devlin *et al.* 2018), a pre-trained model in natural language processing. BERT leverages the encoder stacks of the Transformer architecture and is trained using two primary tasks: Next Sentence Prediction (NSP) and Masked Language Modelling (MLM). NSP helps BERT understand the relationships between sentences, while MLM trains the model to predict masked or missing words within a sentence. BERT's training uses a massive corpus from the web, making it

https://doi.org/10.1017/nlp.2026.10018 Published online by Cambridge University Press

capable of capturing a wide range of linguistic patterns and semantic information. For our specific translation tasks, we leverage BERT's capabilities to translate the Maghrebi dialects to/from MSA. Specifically, we enhance the encoder and decoder blocks by integrating a pre-trained embedding layer from the BERT language model. This layer, depicted in Figure 1 as the BERT input and output representation blocks, includes token embedding, segment embedding, and positional encoding. By integrating these components, we effectively substitute the original embedding and positional encoding methods of the Transformer architecture with BERT's embedding features.

In practice, we use DarijaBERT's (Gaanoun *et al.* 2023) embedding layer for the BERT input representation and ArabicBERT (Safaya, Abdullatif, and Yuret 2020) for the BERT output representation when translating from Maghrebi dialects to MSA. Conversely, when translating from MSA to Maghrebi dialects, we reverse this configuration, employing ArabicBERT for the input representation and DarijaBERT for the output representation.

DarijaBERT was primarily trained on Moroccan dialect text, with additional data from Algerian, Tunisian, and Libyan dialects (which were considered noise due to the difficulty of isolating dialects during the data collection process). It utilised a total of three million sentences for training, including data from tweets, YouTube comments, and story datasets, resulting in a model size of 691 MB. It employed the WordPiece technique on the dataset, generating an 80,000-token vocabulary, and was trained with 147 million parameters.

ArabicBERT was trained on an unshuffled version of the OSCAR Arabic classic dataset and incorporates Arabic data from Wikipedia, amounting to a total of 8.2 billion words. The model employs WordPiece tokenisation, resulting in a vocabulary size of 32,000 tokens.

5. Experiments

In this section, we describe the experimental setup, including the dataset, implementation details, and the methods used to compare the BERT-TransDial architecture against existing approaches.

5.1 Data

In this paper, we use the same dataset investigated by Baniata *et al.* (2021). This dataset, known as PMM-MAG, comprises PADIC (Meftouh *et al.* 2015), MPCA (Bouamor *et al.* 2014), and the MADAR corpus (Bouamor *et al.* 2018), including Moroccan, Algerian, Tunisian, and Libyan dialects. The corpora include some unprocessed data. To assess its impact on NMT, we applied a set of preprocessing techniques as referred to in (Muischnek *et al.* 2019; Baniata *et al.* 2021). These techniques involve removing diacritics, duplicate values with different translations in the source and target, non-Arabic words, special characters, numbers, multiple whitespaces, elongations, and normalising Arabic text (by removing repeated letters and normalising إأآ into ا sentence pairs).

5.2 Implementation details

The dataset used throughout our experiments, incorporating the specified corpora, comprises a total of 57,736 sentence pairs for MSA and Maghrebi dialects. The dataset was randomly divided into three parts: 20% for testing and 80% for training, with 20% of the training data reserved for validation. In all experiments, we used a fixed batch size of 128 and employed 8 parallel attention layers or heads with a d_{model} dimension of 768. Both the encoder and decoder blocks consisted of 6 layers. The position-wise feed-forward network had a dimension of 1024, and we maintained a consistent maximum sequence length of 30 for all datasets (training, test, validation). Additionally, the Adam optimisation algorithm (Kingma and Ba 2014) is employed with an initial learning rate of 10^{-6}, and the training process was conducted over 150 epochs using a machine accelerated by four NVIDIA A100-SXM4-80 GB GPUs (Kissami *et al.* 2025).

https://doi.org/10.1017/nlp.2026.10018 Published online by Cambridge University Press

5.3 Experimental methods

Five experiments, including ours, were conducted to assess BERT-TransDial's effectiveness in bidirectional translation between MSA and Maghrebi dialects. These experiments aim to address the research question: *How does our modified transformer architecture with BERT affect model performance compared to existing methods?*

Experiment 1: We employed the BPE algorithm (Shibata *et al.* 1999), a method designed to assign new characters to frequently repeated substrings. This tokenisation process was applied separately for the input and output languages of our dataset using the 'tokenizers' Python library.[b] This resulted in vocabulary sizes of 27,695 for the Maghrebi dialects and 26,404 for MSA. With the data tokenised using BPE, we then proceeded to train the transformer model.

Experiment 2: As a second experiment, we used the WordPiece tokeniser (Schuster *et al.* 2012; Wu *et al.* 2016) for both input and output data within the transformer system. This data-driven approach effectively manages rare words. Its efficiency lies in its capability to handle unknown words seamlessly, as it accurately segments entire words into meaningful units. We used BERT tokenisers to perform this task, which already employs the WordPiece segmentation technique, using specific models for each language: ArabicBERT for MSA and DarijaBERT for Maghrebi dialects.

Experiment 3: We trained the transformer architecture (referred to as the Baseline Transformer), this time employing the 'batch_encode_plus' method, an indexing technique from the pre-trained BERT model. With this method, we converted each sentence in our parallel dataset into a vector of indices, which were then fed into the embedding layer of both the encoder and decoder stacks of the original transformer architecture. Specifically, we used separate models, ArabicBERT and DarijaBERT, to handle MSA and Maghrebi dialects, respectively.

Experiment 4: We benchmarked our approach with the pre-trained open-source transformer model NLLB-200 (No Language Left Behind [Costa-jussà *et al.* 2022]) for translating Maghrebi dialects to and from MSA. To ensure a fair comparison with our experiments, we fine-tuned the 'facebook/nllb-200-distilled-600M' model with a reduced architecture, using only 3 encoder and decoder layers instead of the original 12. This modification reduced the model size to approximately 350 M parameters. We fine-tuned this adjusted model with specific hyperparameters of 100 epochs and a batch size of 16. Since the entire pretrained model was used without excluding any layers of its components, it required more memory and computational time compared to other experiments. These hyperparameters were chosen to optimise resource usage while leveraging its multilingual capabilities for our MT task.

Experiment 5: In BERT-TransDial, we modified the Transformer architecture to incorporate transfer learning from the pre-trained language model BERT. This adaptation includes substituting the token embedding and positional encoding layers with the embedding layer of BERT, as this pre-trained layer already includes them. We implemented this adaptive method using separate BERT models for the source and target languages, specifically DarijaBERT for Maghrebi dialects and ArabicBERT for MSA.

Throughout all these experiments, we assessed both preprocessed and raw data to evaluate their respective impacts on the NMT system.

6. Results

In this section, we analyse the results of our study in three parts. First, we present the quantitative results in the *Experimental Results* subsection, including the BLEU score (Papineni *et al.* 2002), the BERTScore Precision, denoted as P_{BERT}[c] (Zhang *et al.* 2019), ChrF (Popović 2015), and METEOR

[b] https://pypi.org/project/tokenizers/

[c] The P_{BERT} values for MSA → Maghrebi dialects were not reported because Maghrebi dialects are not among the supported languages in the BERTScore model.

https://doi.org/10.1017/nlp.2026.10018 Published online by Cambridge University Press

Table 1. Quantitative results on the test set with preprocessed data using BLEU,P_{BERT}, ChrF, and METEOR metrics for both translation directions

	Maghrebi dialects to MSA				MSA to Maghrebi dialects		
	BLEU ↑	P_{BERT} ↑	ChrF ↑	METEOR ↑	BLEU ↑	ChrF ↑	METEOR ↑
Transformer using BPE	3.473	66.858	10.078	2.819	0.430	6.762	1.228
Transformer using WordPiece	2.624	67.579	9.410	3.018	0.386	6.387	0.992
Baseline Transformer	5.731	70.466	10.510	9.603	4.144	7.773	7.351
NLLB-200	10.120	**79.852**	**29.981**	**29.820**	3.838	**26.217**	**24.157**
BERT-TransDial∗	**12.731**	78.453	26.982	26.475	**8.474**	18.806	18.011

Table 2. Quantitative results on the test set with raw data using BLEU,P_{BERT}, ChrF, and METEOR metrics for both translation directions

	Maghrebi dialects to MSA				MSA to Maghrebi dialects		
	BLEU ↑	P_{BERT} ↑	ChrF ↑	METEOR ↑	BLEU ↑	ChrF ↑	METEOR ↑
Transformer using BPE	3.205	69.625	9.842	4.043	0.531	7.351	1.504
Transformer using WordPiece	2.805	68.473	11.125	3.870	0.642	6.816	1.090
Baseline Transformer	6.015	71.521	11.327	10.172	3.407	8.379	3.385
NLLB-200	12.478	**80.922**	**30.907**	**32.412**	4.187	**26.809**	**26.219**
BERT-TransDial∗	**14.148**	79.414	28.885	28.428	**8.861**	20.994	19.465

(Banerjee *et al.* 2005) to ensure a comprehensive evaluation of BERT-TransDial's performance (Kocmi *et al.* 2021). Following this, the *Ablation Study* investigates specific changes within BERT-TransDial. Finally, the q*ualitative results* explore the translation output using BERT-TransDial on the test set, offering qualitative insights into its effectiveness beyond quantitative metrics.

6.1 Experimental results

The translation results of our experiments on the Maghrebi dialects and MSA in both directions are presented in Tables 1 and 2.

– Table 1 summarises the results of translation obtained using the beam search decoding strategy with a beam width of 5 for both directions: Maghrebi dialects to MSA and MSA to Maghrebi dialects, using preprocessed data. The results indicate that the third experiment, 'Baseline Transformer', outperformed the first and second experiments (using BPE and WordPiece tokenisers) for Maghrebi dialects to MSA translation, achieving higher BLEU, P_{BERT}, ChrF, and METEOR scores. Furthermore, in the fifth experiment, 'BERT-TransDial', the model demonstrated superior performance across all metrics compared to the 'Baseline Transformer'. For the reverse direction (MSA to Maghrebi dialects), similar results were observed: the 'BERT-TransDial' model achieved the best results overall, surpassing the 'Baseline Transformer', which in turn outperformed the 'BPE and WordPiece' models. The fourth experiment, involving fine-tuning the 'NLLB-200' model, demonstrated enhanced translation quality in both directions based on P_{BERT}, ChrF, and

https://doi.org/10.1017/nlp.2026.10018 Published online by Cambridge University Press

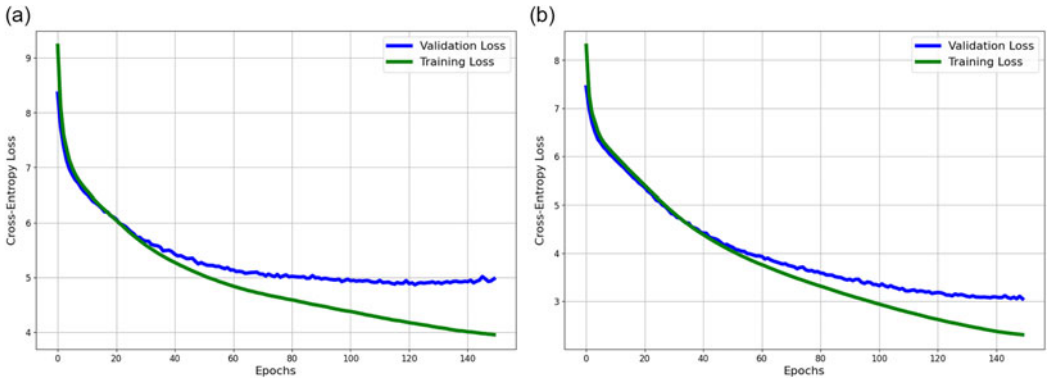

Figure 2. Training and validation losses for MSA to Maghrebi dialects (a) and Maghrebi dialects to MSA (b) on raw data.

METEOR scores. However, the 'BERT-TransDial' model achieved the highest BLEU score in both translation directions, further highlighting its overall effectiveness.

– Table 2 presents the results obtained using raw parallel data. Experiment 5, 'BERT-TransDial', emerged as the most effective approach, surpassing the baseline transformer for Maghrebi dialects to MSA and MSA to Maghrebi dialects, respectively. In contrast, Experiments 1 and 2, which used 'BPE' and 'wordPiece', displayed the poorest results, with lower BLEU, ChrF, and METEOR scores, indicating inferior translation quality compared to the other experiments. As reflected by the higher BLEU score, BERT-TransDial surpasses the fine-tuned 'NLLB-200' model in translating to and from MSA. However, when considering other metrics (P_{BERT}, ChrF, and METEOR), the NLLB-200 model performs better in translation quality.

Curves (a) and (b) in Figure 2 represent the training and validation losses for the translation tasks from MSA to Maghrebi dialects and from Maghrebi dialects to MSA on raw data, respectively. Both curves display decreasing values over 150 epochs. The small difference between the training and validation losses indicates that the model has effectively learned meaningful patterns from the data, enabling strong generalisation to unseen examples for both tasks. During the inference phase, the model with the lowest validation loss is employed to generate translations. The quantitative results of these translations are presented in Tables 1 and 2.

6.2 Ablation study

In this section, we investigate the impact of BERT's embedding position and parameter freezing on BERT-TransDial's performance. First, we analyse the integration of BERT's embedding in the encoder or the decoder. Next, we compare the performance of BERT-TransDial with and without parameter freezing in translating Maghrebi dialects to/from MSA.

6.2.1 BERT's embedding position

In this section, we present the results of an ablation study to assess the effectiveness of the BERT-TransDial architecture. This exploration involves preserving the original token embedding and positional encoding from the vanilla Transformer in the encoder while replacing them in the decoder with the embedding layer derived from the pre-trained BERT model (referred to as 'BERT's embedding in the decoder'). Additionally, we conducted a second experiment by replacing the Transformer's embedding layer with the BERT embedding layer, but this time in the

https://doi.org/10.1017/nlp.2026.10018 Published online by Cambridge University Press

Table 3. Results of the ablation study (BERT's Embedding Position) on the test set for preprocessed data

	Maghrebi dialects to MSA				MSA to Maghrebi dialects		
	BLEU ↑	P_{BERT} ↑	ChrF ↑	METEOR ↑	BLEU ↑	ChrF ↑	METEOR ↑
BERT embedding in the Encoder	8.776	76.008	17.225	19.269	6.139	13.689	14.383
BERT embedding in the decoder	5.437	71.484	12.106	9.969	5.177	10.165	8.394
BERT-TransDial∗	**12.731**	**78.453**	**26.982**	**26.475**	**8.474**	**18.806**	**18.011**

Table 4. Results of the ablation study (BERT's Embedding Position) on the test set for raw data

	Maghrebi dialects to MSA				MSA to Maghrebi dialects		
	BLEU ↑	P_{BERT} ↑	ChrF ↑	METEOR ↑	BLEU ↑	ChrF ↑	METEOR ↑
BERT embedding in the Encoder	9.997	76.353	18.779	20.356	6.418	15.115	14.822
BERT embedding in the decoder	5.303	72.392	11.927	9.804	5.240	10.456	8.656
BERT-TransDial∗	**14.148**	**79.414**	**28.885**	**28.428**	**8.861**	**20.994**	**19.465**

Table 5. Results of the ablation study (with and without parameter freezing) on the test set for raw data

	Maghrebi dialects to MSA				MSA to Maghrebi dialects		
	BLEU ↑	P_{BERT} ↑	ChrF ↑	METEOR↑	BLEU ↑	ChrF ↑	METEOR ↑
BERT-TransDial with parameter freezing	9.560	74.221	19.509	19.758	7.483	17.103	16.697
BERT-TransDial without parameter freezing	**14.148**	**79.414**	**28.885**	**28.428**	**8.861**	**20.994**	**19.465**

encoder (referred to as 'BERT's embedding in the encoder'). These experiments aim to discern the optimal integration point for BERT's embedding within our architecture.

The incorporation of BERT's embedding in the encoder stack in both Tables 3 and 4 resulted in improved performance compared to when employing it in the decoder. As shown in Table 3 (for preprocessed data), the results demonstrate consistently significant improvement in BLEU, P_{BERT}, ChrF, and METEOR scores. Moreover, BERT-TransDial achieves superior performance in translating Maghrebi dialects to and from MSA. Table 4 confirms these findings, showcasing the superiority of 'BERT's embedding in the encoder' over the decoder across all the evaluation metrics. These results underscore the efficacy of integrating BERT embeddings during the encoding phase and provide insight into the performance of the BERT-TransDial across both raw and preprocessed data cases.

6.2.2 With and without parameter freezing

Table 5 compares the performance of BERT-TransDial with and without parameter freezing for translating Maghrebi dialects and MSA, using BLEU, P_{BERT}, ChrF, and METEOR as evaluation metrics. For the translation of Maghrebi dialects to/from MSA, the approach without parameter freezing significantly outperforms the one with freezing, achieving better BLEU, P_{BERT}, ChrF, and METEOR scores. Similarly, for translating from MSA to Maghrebi dialects, the approach without parameter freezing shows superior performance across all metrics. These results indicate that

https://doi.org/10.1017/nlp.2026.10018 Published online by Cambridge University Press

the proposed approach without parameter freezing provides significantly better translation quality and accuracy, evidenced by higher evaluation scores, suggesting that parameter freezing may hinder the model's performance in this context.

6.3 Qualitative results

Table 6 presents the qualitative results of the 'BERT-TransDial', 'BERT in the Encoder', and 'BERT in the Decoder' in translating from Maghrebi dialects into MSA. The 'BERT-TransDial' effectively preserves the intended meaning of sentences across varying lengths, showing particular strength in translating shorter sentences. For instance, in the first example, the source sentence 'كيفكيف' is translated to 'نفس النوع' 'the same type' in English. This aligns contextually with 'نفس الشيء' and is considered correct, whereas the other systems produced incorrect translations. Similarly, for the source sentence 'انجم نشوف حاجة ارخص ؟', BERT-TransDial translates it to 'هل يمكنني ان ارى ارخص ؟', 'Can I see cheaper?' in English. While this translation omits the word 'something', it remains contextually valid according to the reference 'Can I see something cheaper?' On the other hand, 'BERT in the Encoder' generates 'هل يمكن ان اري شيء ؟' missing out the word 'ارخص' ('cheaper'), which captures the core message. Meanwhile, BERT in the decoder produces a completely incorrect translation. Out of the eight sentences presented in Table 6, BERT-TransDial successfully translates the majority, demonstrating its ability to provide reasonable and contextually meaningful translations of Maghrebi dialects into MSA.

In the reverse direction, the BERT-TransDial system demonstrates acceptable translation quality, outperforming the 'BERT in the Encoder' and 'BERT in the Decoder' models. As shown in Table 7, BERT-TransDial generates translations that are reasonably correct, though some mistranslations occur. For instance, in row 4, the source sentence 'كيف حالك سليمة ؟' is translated as 'كيف سليمة ؟' 'How is Salima?' While it omits the word 'doing', the translation remains contextually accurate. Similarly, in row 5, the input sentence 'صدقيني ان الزهرة لا تعلم' is translated as 'والله والله ما تعرف' missing the noun 'Zahra', which reduces the clarity of the output. Nevertheless, it still maintains meaningful aspects of the original sentence, whereas the outputs from the other systems are entirely incorrect. In conclusion, our approach provides reasonably accurate translations for both tasks, with significant improvements when translating into MSA (Table 6), effectively maintaining meaning and detail. This contrasts with the reverse translation task, where the output captures the broader idea but lacks specificity. Among the other systems, 'BERT in the Encoder' produces relatively better results compared to 'BERT in the Decoder' in both Tables 6 and 7.

To analyse systematic errors causing the low BLEU scores, a representative subset of the data was selected. We detected several systematic errors, namely semantic drift, which occurs when the model produces fluent but meaning-changed translations (e.g., 'نخدم في شركة تجارة' → 'في شركة الطيران أعمل'). Dialectal lexical errors, which misinterpret totally Maghrebi vocabulary 'هل لي في بعض القهوة؟' → 'نخو شوية فلفل؟ نجم'. Structural divergence with partial adequacy, which captures cases where intent is preserved but syntactic realisation is non-standard 'إنها دقائق أخرى' → 'شوية دقايق زيد'. Finally, idiomaticity, a major challenge in dialects, occurs when idiomatic expressions are rendered in grammatically correct but pragmatically wrong forms; for instance, the source sentence 'دخل سوق راسك', which means 'Mind your business', was literally translated to 'ادخل الى محل راسك', which means 'Enter in the market of your head'.

https://doi.org/10.1017/nlp.2026.10018 Published online by Cambridge University Press

https://doi.org/10.1017/nlp.2026.10018 Published online by Cambridge University Press

Table 6. Qualitative translation outputs from Maghrebi dialects to MSA generated by the proposed BERT-TransDial model, BERT in the encoder, and BERT in the decoder. The table includes source inputs, reference translations, and corresponding outputs for model comparison

Size	Parallel sentences		Predicted sentences with :		
	Source	Reference	BERT-TransDial	BERT in the encoder	BERT in the decoder
1	كيفكيف	نفس الشيء	نفس النوع	لا الجزاير	لا
		The same thing	The same type	No, Algeria	No
2	احنا نستناو	نحن ننتظر	نحن ننتظر	نحن نحن	ماذا ؟
		We are waiting	We are waiting	We are we	What?
3	لازمني نقلك بالسلامة.	علي ان اقول وداعا	يجب ان اقول لك مع السلامة.	يجب ان اذهب على اذهب.	من فضلك اعطني هذه.
		I have to say goodbye	I must say goodbye to you	I must go on go.	Please give me this.
4	انجم نشوف الكرهبة ؟	هل لي ان ارى السيارة ؟	هل يمكنني ان ارى السيارة ؟	هل يمكن ان اريها ؟	هل يمكن ان تخبرني عن طريق البريد ؟
		May I see the car?	Can I see the car?	Can I see it?	Can you tell me through mail?
5	انجم نشوف حاجة ارخص ؟	هل يمكن ان ارى شيبا ارخص.	هل يمكنني ان ارى ارخص ؟	هل يمكن ان ارى شيء ؟	هل يمكن ان تخبرني عن طريق البريد ؟
		Can I see something cheaper?	Can I see cheaper?	Can I see something?	Can you tell me through mail?
6	طاولة لغير المدخنين ، من فضلك.	ماية غير المدخنين من فضلك.	ماية غير المدخنين من فضلك.	اريد مايدة بالقرب من فضلك.	لقد سرقت حقيبتي.
		A non-smoking table, please.	A non-smoking table, please.	I want a table nearby, please.	My bag has been stolen.
6	تقدر تقول ليا شحال المبلغ ؟	هل لك ان تخبرني قيمة المبلغ ؟	هل يمكنك ان تخبرني كم الاجرة ؟	هل يمكنك ان تخبرني ؟	هل يمكن ان تخبرني عن طريق البريد ؟
		Can you tell me the value of the amount?	Can you tell me how much the fare is?	Can you tell me?	Can you tell me through mail?
9	بس عندي موعد عشان نشوفها الساعة اثنين بالضبط.	لكني لدى موعد لرويتها في الساعة الثانية.	ولكن لدي موعد في الساعة الثانية.	عندي معي في موعد في الساعة الساعة الساعة الساعة.	لقد سرقت حقيبتي.
		But I have an appointment to see her at 2 o'clock.	But I have an appointment at 2 o'clock.	- Translation is not understandable -	My bag has been stolen.

https://doi.org/10.1017/nlp.2026.10018 Published online by Cambridge University Press

Table 7. Qualitative translation outputs from MSA to Maghrebi dialects generated by the proposed BERT-TransDial model, BERT in the encoder, and BERT in the decoder. The table includes source inputs, reference translations, and corresponding outputs for model comparison

Size	Parallel sentences		Predicted sentences with :		
	Source	Reference	BERT-TransDial	BERT in the encoder	BERT in the decoder
1	رايع	يا حلاالوة.	مزيان	مزيان	
		Great	Good	Good	
2	هاهي فاتورتك.	هاهي الفاتورة ديالك.	هذي الفاتورة ديالك.	هادي ديالك.	عندي حجز.
		Here is your invoice	This is your invoice	This is yours	I have a reservation
4	هذا مكلف نوعا ما.	هاد الشي غالي	غالي شوية.	هذا بزاف.	عندي حجز.
		This is expensive	A bit expensive	This is too much	I have a reservation
4	كيف حالك سليمة ؟	كي دايرة سليمة ؟	كيف سليمة ؟	سليمة سليمة	فين كاين شي حاجة ؟
		How is Salima doing?	How is Salima?	Salima Salima	Where is something?
5	صدقيني ان الزهرة لا تعلم	صدقني الزهرة ماتعرفش	والله والله ما تعرف	والله والله	ما نعرفش ما نعرفش
		Believe me, Zahra doesn't know	I swear, I swear, she doesn't know.	I swear, I swear	I don't know, I don't know
5	اعتذر، عندي مشكل مع حاسوبي	سامحوني ، عندي مشكلة في اورديناتوري	اسمح لي عندي مشكل مع عندي مشكل	عندي ، عندي عندي عندي عندي مع عنديي	ما فماش حتى شي حاجة اخرى
		Forgive me, I have a problem with my computer	Excuse me, I have a problem with 'I have a problem	-Translation is not understandable-	-Translation is not understandable-
6	الا تحضر لي بعض القهوة ؟	تنجم تجيبلي قهوة ؟	تنجم تعطيني شوية قهوة ؟	ممكن شوية شوية ؟	قداش من هوني ؟
		Can you bring me a coffee?	Can you give me some coffee?	Is it possible to have just a little?	How much from here?
11	انه فنان غالي الثمن للغاية. و لذلك، خذ حذرك من فضلك.	راه فنان مخادع. داك الشي علاش ، خاصك ترد بالك عافاك.	هو غالي برشا. غالي برشا ، يعيشك.	هذا بزاف ، يعيشك. هذا.	واخا ، عافاك. غادي تمشي لليابان.
		He is a very expensive artist. Therefore, please be cautious.	He is very expensive. Very expensive, please.	This is too much, please. This.	Alright, please. I'm going to Japan.

Table 8. Average results of pilot rating experiment for BERT-TransDial

Models		Average score
Raw Data	MSA to Maghrebi dialects	4.26
	Maghrebi dialects to MSA	4.82
Preprocessed Data	MSA to Maghrebi dialects	4.62
	Maghrebi dialects to MSA	4.75

7. Human evaluation

In this section, (*i*) we present human evaluation (Schuff *et al.* 2023) metrics employed to assess the quality of translations generated by the BERT-TransDial architecture, and (*ii*) compare the translations produced by BERT-TransDial with those of the LLMs Gemini and ChatGPT.

7.1 Pilot rating experiment

In this experiment, we employed a human evaluation metric through a pilot rating experiment. Seven native speakers of Maghrebi dialects and Standard Arabic were selected for this purpose. The objective was to confirm how well the quantitative results align with human-generated translations. During the testing phase, we randomly chose 20 sentences for each case from the generated translations. Specifically, we selected 20 sentences for predicted Maghrebi dialects and 20 for predicted MSA translations using raw data. The same process was repeated for BERT-TransDial with preprocessed data. In total, 80 sentences were evaluated, and ratings were assigned on a Likert scale ranging from 1 to 7 (7 = 'Flawless Translation', 6 = 'Very Good', 5 = 'Good', 4 = 'Acceptable', 3 = 'Poor', 2 = 'Very Poor', 1 = 'Incomprehensible'), as described by Marta *et al.* (2018).

Table 8 presents the average results from the pilot rating experiment using BERT-TransDial. For the Maghrebi dialects to MSA direction, the average score with raw data is 4.82, slightly decreasing to 4.75 with preprocessed data, both considered in the moderate to good range on a 7-point scale. In the reverse direction (MSA to Maghrebi dialects), the scores are 4.26 for raw data and 4.62 for preprocessed data. Notably, 33.21% of sentences translated into Maghrebi dialects and 42.85% translated into MSA received ratings of 6 or higher, indicating that a significant portion of the translations were rated positively. Overall, these results suggest a favourable perception of the model's performance, particularly in translating Maghrebi dialects to MSA. Figure 3 provides a visual representation of the ratings from each participant, offering insights into the consistency of the evaluations.

7.2 Ranking experiment

In the second human evaluation experiment, we compared the results of BERT-TransDial with those of LLMs. The seven participants were tasked with ranking the translation outputs generated by BERT-TransDial, ChatGPT (GPT-4o), and Gemini (1.5 Flash). Using zero-shot prompting, we employed the template prompt suggested by Kadaoui *et al.* (2023), where several LLMs were compared and found to provide better performance than other candidate prompts for MT tasks.[d] Subsequently, participants were instructed to vote for the correct translations without being informed about which system produced each translation. In this task, we used a total of eighty

[d] The prompts used were: 'Translate the following Maghrebi dialects into Modern Standard Arabic:' and 'Translate the following Modern Standard Arabic into Maghrebi dialects:'.

https://doi.org/10.1017/nlp.2026.10018 Published online by Cambridge University Press

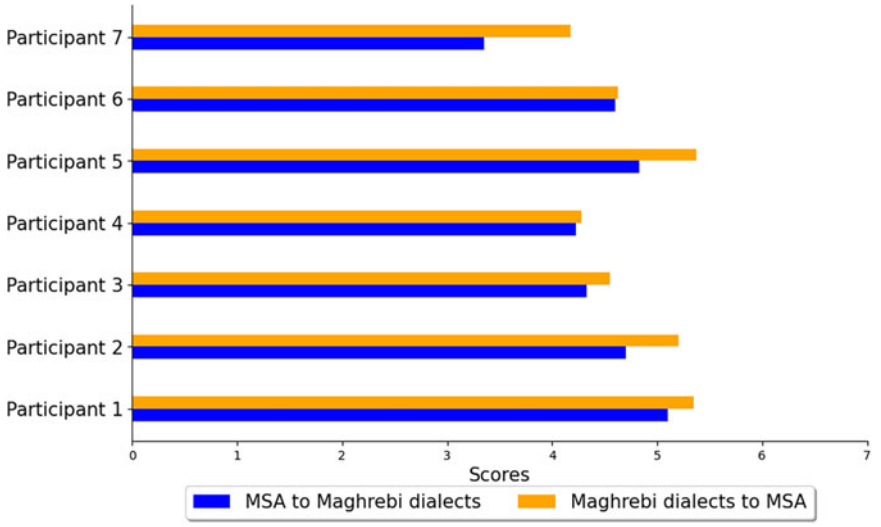

Figure 3. Visualisation displaying the pilot rating experiment, where each bar plot displays a rating ranging from 1 to 7 for each participant, assessing the translation quality produced by BERT-TransDial for Maghrebi dialects and MSA translation generations.

sentences: 40 segments were designated for translating into MSA, and 40 for translating into Maghrebi dialects.

Table 9 summarises the percentage results of the ranking experiment, with inter-annotation agreements between participants of 0.809 and 0.788 for translating into MSA and Maghrebi dialects, respectively, as measured by Average Pairwise Kendall's Tau. The analysis indicates that ChatGPT consistently produces high-quality translations in both cases compared to Gemini and BERT-TransDial. Gemini ranked second, while BERT-TransDial was least preferred. Despite this, it is important to note that BERT-TransDial is not directly comparable to ChatGPT and Gemini in terms of trainable parameters and model size. Nevertheless, BERT-TransDial still gained a notable portion of the preferences, demonstrating its potential as an efficient alternative for translation tasks.

8. Discussion

BERT-TransDial outperforms other experiments in terms of translation quality. This improvement is attributed to the use of a pre-trained embedding layer, which represents input and output sentences using a specific BERT-based language model. In contrast, other experiments give lower BLEU, P_{BERT}, ChrF, and METEOR scores due to their dependence on indexing methods 'Wordpiece and BPE' for segmented Maghrebi dialects and MSA words. These methods require large-scale datasets to achieve high precision in output. Additionally, such models are trained with randomly initialised embedding layers, which limits their performance. Similar limitations are observed in the baseline model, which employs indexing methods from the tokenisers of ArabicBERT and DarijaBERT. The NLLB-200 shows comparable results to those of BERT-TransDial. Although NLLB-200 is a pre-trained model with approximately 350 M trainable parameters, the approach outperforms it in BLEU scores for both translation tasks. This indicates that BERT-TransDial generates translations with higher n-gram precision compared to NLLB-200, predicting each word in a lexically and syntactically coherent manner than NLLB-200, closely aligned with expected patterns in the unseen test data than NLLB-200. While this

https://doi.org/10.1017/nlp.2026.10018 Published online by Cambridge University Press

Table 9. Human evaluation scores – Raking experiment

Models		Percentage of cases
BERT-TransDial preferred	MSA to Maghrebi dialects	17.35%
	Maghrebi dialects to MSA	16.99%
Gemini preferred	MSA to Maghrebi dialects	33.79%
	Maghrebi dialects to MSA	38.69%
ChatGPT preferred	MSA to Maghrebi dialects	43.94%
	Maghrebi dialects to MSA	42.83%

latter achieves slightly better results in other metrics (P_{BERT}, ChrF, and METEOR), these differences reflect its focus on semantic accuracy. Nevertheless, BERT-TransDial remains comparable to NLLB-200 across all metrics, demonstrating its effectiveness. Note that we used a reduced version of NLLB-200-distilled-600 M to match BERT-TransDial's size, which may not fully reflect its original performance.

BERT-TransDial achieved results of 12.731 in BLEU, 78.453 in P_{BERT}, 26.982 in ChrF, and 26.475 in METEOR for preprocessed data, and 14.148 in BLEU, 79.414 in P_{BERT}, 28.885 in ChrF, and 28.428 in METEOR for raw data when translating MSA to Maghrebi dialects and Maghrebi dialects to MSA, respectively. Generating MSA is inherently more accurate than Maghrebi dialects, as the latter includes a wider variety of dialects, also affected by the lack of standardised rules and the rich morphology of Arabic dialects. Notably, both preprocessed and raw data produce approximately similar results with our approach, though performance is slightly better with raw data. We attribute this to the design of BERT models, which were pre-trained on diverse web and book raw corpora that contain a mix of structured and unstructured text. Preprocessing steps such as removing diacritics or punctuation can eliminate important linguistic features, making the text harder to interpret. This may lead to word ambiguities, reduced clarity in sentence structure, and affect the capacity of the model to capture overall context. In contrast, raw data preserves crucial features and contextual patterns, allowing it to better exploit the richness present in the input. Furthermore, in the ablation study, we observe that integrating the BERT's embedding layer into the encoder achieves superior results compared to its use in the decoder. This is due to the fact that incorporating the BERT embedding weights into the encoder results in a better representation of the input sentence, which allows the decoder to clearly align the input sentence with the target language, making it easier to generate a relevant and accurate translation in the target language. Moreover, BERT-TransDial, which involves integrating BERT's embeddings into both the encoder and the decoder within the vanilla transformer, outperforms using BERT embedding in only the encoder or the decoder, achieving higher BLEU, ChrF, and METEOR scores. Next, BERT-TransDial performs significantly better without parameter freezing across all metrics (BLEU, P_{BERT}, METEOR, ChrF) for both translation directions (Table 6). This suggests that allowing the model's parameters to update weights during training improves its ability to learn and adapt to an adequate representation for both language pairs, leading to more accurate and correct translations. Freezing parameters limits the model's capacity to fully capture the BERT's embedding knowledge, whereas not freezing them allows for more effective fine-tuning and optimisation.

Additionally, the qualitative results presented in Tables 6 and 7 demonstrate the effectiveness of BERT-TransDial. The model successfully translated the majority of sentences in both directions. Nonetheless, we observed that certain translation errors could be attributed to linguistic features such as irregular morphology, semantic drift, dialect-specific vocabulary, lexical errors, and the absence of standardised written form – factors that complicate consistent translation

https://doi.org/10.1017/nlp.2026.10018 Published online by Cambridge University Press

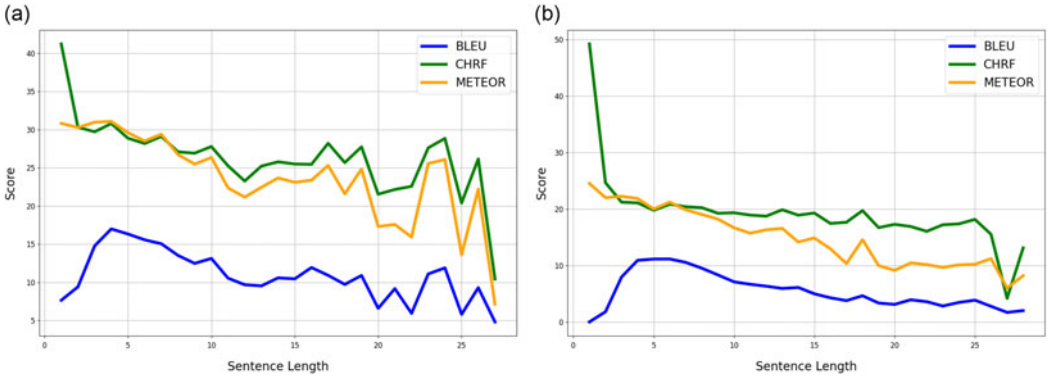

Figure 4. Quantitative metrics based on sentence length were analysed for (a) translation from Maghrebi dialects to MSA and (b) translation from MSA to Maghrebi dialects, using raw data.

across dialects. These challenges, previously discussed in Sections 1 and 6.3, reflect the linguistic variability found in Maghrebi dialects and help explain some of the difficulties the model encountered. Notably, human evaluation during pilot rating experiments provided positive feedback, with scores ranging from 4 to nearly 5, highlighting the consistency of the proposed method. However, our approach achieved the lowest percentages compared to Gemini and ChatGPT, primarily due to its limited performance on lengthy sentences. As reflected in Figure 4, the BLEU score peaks at sentence lengths of 5 to 10 tokens but declines significantly for longer in both directions. The ChrF score in Plot (a) shows a more gradual decline, remaining relatively stable for longer sentences, while in Plot (b), ChrF experiences a significant drop with increased sentence length. Similarly, METEOR fluctuates for shorter sentences but follows a consistent decrease as sentence length increases, reflecting the model's sensitivity to translating complex and lengthy sentences. The Transformer architecture is generally effective at handling long sequences, thanks to its use of multi-head attention. However, in our case, the limitation likely stems from the training data: first, it does not include explicit dialect labels, which restricts the possibility of conducting a detailed cross-dialect performance analysis; second, it contains relatively few long sentences. Specifically, we verified that the average sentence length is 7.66 words for MSA and 5.99 words for Maghrebi dialects, indicating that longer sequences are relatively infrequent. In this context, analysing the impact of data augmentation techniques (Li *et al.* 2019; Lamar *et al.* 2023; Diab *et al.* 2024) and leveraging them to enrich the training set, particularly with longer and more diverse sentence structures, could help mitigate these limitations.

9. Conclusion

In summary, in this work, we tackled the problem of translating low-resource languages, a challenging research topic in NMT. These languages deviate from standard grammatical rules, leading to a lack of accurate preprocessing methods that directly impact their vector representation. In this paper, we introduced an NMT approach for translation between standard Arabic and North African dialects. We employed a neural architecture that integrates transfer learning in a hybrid manner. It incorporates pre-trained embedding blocks from ArabicBERT and DarijaBERT within the encoder and the decoder of a vanilla transformer architecture. This study provides an empirical analysis of the performance of this hybrid design, aiming to better understand its contribution to translation quality. The approach was evaluated using BLEU, P_{BERT}, ChrF, and METEOR metrics, as well as human evaluation through rating and ranking experiments. Our future work will focus on incorporating more Arabic dialects where per-dialect performances are analysed to gain

https://doi.org/10.1017/nlp.2026.10018 Published online by Cambridge University Press

a deeper understanding of these dialects, as well as improving the model's ability to handle lengthy and complex inputs to improve its overall performance and applicability. This includes extending the dataset with longer examples and applying data augmentation techniques to increase the diversity and complexity of input sentences, while also exploring the integration of smaller LLMs as a future direction, where model size and computational cost remain important constraints.

Acknowledgments. Experiments presented in this paper were carried out using the supercomputer Toubkal, supported by Mohammed VI Polytechnic University (https://www.um6p.ma), and the African Supercomputing Center (ASCC).

Competing interests. The author(s) declare none.

References

Shiwen Y. and Xiaojing B. (2014). Rule-based machine translation. In *Routledge Encyclopedia of Translation Technology*. Routledge, pp. 186–200.

Somers H. (1999). Example-based machine translation. *Machine Translation* **14**(2), 113–157. https://doi.org/10.1023/A:1008109312730

Lopez A. (2008). Statistical machine translation. *ACM Computing Surveys (CSUR)* **40**(3), 1–49. https://doi.org/10.1145/1380584.1380586

Zens R., Och F.J. and Ney H. (2002). Phrase-based statistical machine translation. In KI 2002: *Advances in Artificial Intelligence: 25th Annual German Conference on AI,* KI 2002 Aachen, Germany, September 16–20, 2002 Proceedings 25. Springer, pp. 18–32. https://doi.org/10.1007/3-540-45751-8

Almahairi A., Cho K., Habash N. and Courville A. (2016). First result on Arabic neural machine translation. *arXiv preprint arXiv:* 1606.02680.

Wu Y., Schuster M., Chen Z., Le Q.V., Norouzi M., Macherey W. and Dean J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144.*

Luong M.T., Pham H. and Manning C.D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:* 1508.04025.

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

Bahdanau D., Cho K. and Bengio Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:* 1409.0473.

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N. and Polosukhin I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* 30. https://doi.org/10.48550/arXiv.1706.03762

Sutskever I., Vinyals O. and Le Q.V. (2014). Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems* 27. https://doi.org/10.48550/arXiv.1409.3215

Muischnek K. and Müürisep K. (2019). Impact of corpora quality on neural machine translation. In *Human Language Technologies–The Baltic Perspective: Proc. of the Eighth International Conference Baltic HLT* 2018 Vol. **307**, pp. 126. IOS Press. https://doi.org/10.3233/978-1-61499-912-6-126

Baniata L.H., Ampomah I.K. and Park S. (2021). A transformer-based neural machine translation model for Arabic dialects that utilizes subword units. *Sensors* **21**(19), 6509. https://doi.org/10.3390/s21196509

Sennrich R., Haddow B. and Birch A. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:* 1508.07909.

Baniata L.H., Kang S. and Ampomah I.K. (2022). A reverse positional encoding multi-head attention-based neural machine translation model for Arabic dialects. *Mathematics* **10**(19), 3666. https://doi.org/10.3390/math10193666

Slim A., Melouah A., Faghihi U. and Sahib K. (2022). Improving neural machine translation for low resource algerian dialect by transductive transfer learning strategy. *Arabian Journal for Science and Engineering* **47**(8), 10411–10418. https://doi.org/10.1007/s13369-022-06588-w

Clinchant S., Jung K.W. and Nikoulina V. (2019). On the use of BERT for neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pp. 108–117. https://doi.org/10.18653/v1/D19-5611

Guo J., Zhang Z., Xu L., Chen B. and Chen E. (2021). Adaptive adapters: An efficient way to incorporate BERT into neural machine translation. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. **29**, pp. 1740–1751. https://doi.org/10.1109/TASLP.2021.3076863

Guo Z. and **Le Nguyen M.** (2020). Document-level neural machine translation using bert as context encoder. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop*, pp. 101–107. https://doi.org/10.18653/v1/2020.aacl-srw.15

https://doi.org/10.1017/nlp.2026.10018 Published online by Cambridge University Press

Cho K., Van Merriënboer B., Gulcehre C., Bahdanau D., Bougares F., Schwenk H. and Bengio Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. https://doi.org/10.3115/v1/D14-1179

Kalchbrenner N. and Blunsom P. (2013). Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1700–1709.

Devlin J., Chang M.W., Lee K. and Toutanova K. (2018). *BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding*. North American Chapter of the Association for Computational Linguistics.

Gaanoun K., Naira A.M., Allak A. and Benelallam I. (2023). DarijaBERT: A step forward in NLP for the written Moroccan dialect. *International Journal of Data Science and Analytics* **20**(2), 1–13. https://doi.org/10.1007/s41060-023-00498-2

Safaya A., Abdullatif M. and Yuret D. (2020). KUISAIL at SemEval-2020 Task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pp. 2054–2059. https://doi.org/10.18653/v1/2020.semeval-1.271

Meftouh K., Harrat S., Jamoussi S., Abbas M. and Smaili K. (2015). Machine translation experiments on PADIC: A parallel Arabic dialect corpus. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pp. 26–34.

Bouamor H., Habash N. and Oflazer K. (2014). A multidialectal parallel corpus of Arabic. In *LREC*, pp. 1240–1245.

Bouamor H., Habash N., Salameh M., Zaghouani W., Rambow O., Abdulrahim D. and Oflazer K. (2018). The madar arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Kingma D.P. and Ba J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:*1412.6980.

Shibata Y., Kida T., Fukamachi S., Takeda M., Shinohara A., Shinohara T. and Arikawa S. (1999). Byte Pair encoding: A text compression scheme that accelerates pattern matching.

Papineni K., Roukos S., Ward T. and Zhu W.J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318. https://doi.org/10.3115/1073083.1073135

Costa-jussà M.R., Zampieri M. and Pal S. (2018). A neural approach to language variety translation. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 275–282.

Kadaoui K., Magdy S.M., Waheed A., Khondaker M.T.I., El-Shangiti A.O., Nagoudi E.M.B. and Abdul-Mageed M. (2023). TARJAMAT: Evaluation of bard and ChatGPT on machine translation of ten Arabic varieties. In *Proceedings of ArabicNLP 2023*, pp. 52–75. https://doi.org/10.18653/v1/2023.arabicnlp-1.6

Kchaou S., Boujelbane R. and Hadrich L. (2023). 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* **22**(3), 1–21. https://doi.org/10.1145/3568674.

Veeramani H., Thapa S. and Naseem U. (2023). DialectNLU at NADI 2023 shared task: Transformer based multitask approach jointly integrating dialect and machine translation tasks in Arabic. In *Proceedings of ArabicNLP 2023*, pp. 614–619. https://doi.org/10.18653/v1/2023.arabicnlp-1.63

Al-Ibrahim R. and Duwairi R.M. (2020). Neural machine translation from Jordanian Dialect to modern standard Arabic. In *2020 11th International Conference on Information and Communication Systems (ICICS)*, pp. 173–178. https://doi.org/10.1109/ICICS49469.2020.239505

Farhan W., Talafha B., Abuammar A., Jaikat R., Al-Ayyoub M., Tarakji A.B. and Toma A. (2020). Unsupervised dialectal neural machine translation. *Information Processing & Management* **57**(3), 102181. https://doi.org/10.1016/j.ipm.2019.102181.

Sutskever I., Vinyals O. and Le Q. (2014). Sequence to sequence learning with neural networks. In *Neural Information Processing Systems*, pp. 3104–3112.

Yang J., Wang M., Zhou H., Zhao C., Zhang W., Yu Y. and Li L. (2020). Towards making the most of BERT in neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. **34**, No. 05, pp. 9378–9385. https://doi.org/10.1609/aaai.v34i05.6479

Schuster M. and Nakajima K. (2012). Japanese and Korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5149–5152. https://doi.org/10.1109/ICASSP.2012.6289079

Slim A. and Melouah A. (2024). Low resource arabic dialects transformer neural machine translation improvement through incremental transfer of shared linguistic features. *Arabian Journal for Science and Engineering* **49**(7), 1–17. Springer. https://doi.org/10.1007/s13369-023-08543-9

Faheem M.A., Wassif K.T., Bayomi H. and Abdou S.M. (2024). Improving neural machine translation for low resource languages through non-parallel corpora: A case study of Egyptian dialect to modern standard Arabic translation. *Scientific Reports* **14**(1), 2265. https://doi.org/10.1038/s41598-023-51090-4

Kocmi T., Federmann C., Grundkiewicz R., Junczys-Dowmunt M., Matsushita H. and Menezes A. (2021). To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pp. 478–494.

https://doi.org/10.1017/nlp.2026.10018 Published online by Cambridge University Press

Snover M., **Dorr B.**, **Schwartz R.**, **Micciulla L. and Makhoul J.** (2006). 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*, pp. 223–231.

Popović M. (2015). chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop On Statistical Machine Translation*, pp. 392–395. https://doi.org/10.18653/v1/W15-3049

Costa-jussà M.R., **Cross J.**, **Çelebi O.**, **Elbayad M.**, **Heafield K.**, **Heffernan K. and** NLLB Team. (2022). No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:*2207.04672.

Schuff H., **Vanderlyn L.**, **Adel H. and Vu N.T.** (2023). How to do human evaluation: A brief introduction to user studies in NLP. *Natural Language Engineering* **29**(5), 1199–1222. https://doi.org/10.1017/S1351324922000535

Banerjee S. and Lavie A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of The ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation And/or Summarization*, pp. 65–72.

Zhang T., **Kishore V.**, **Wu F.**, **Weinberger K.Q. and Artzi Y.** (2019). BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.1904.09675

Younes J., **Souissi E.**, **Achour H. and Ferchichi A.** (2020). Language resources for Maghrebi Arabic dialects' NLP: A survey. *Language Resources and Evaluation* **54**(4), 1079–1142. https://doi.org/10.1007/s10579-020-09490-9

Skiredj A., **Azhari F.**, **Berrada I. and Ezzini S.** (2024). DarijaBanking: A new resource for overcoming language barriers in banking intent detection for Moroccan Arabic speakers. *Natural Language Processing* **31**(5), 1–31. https://doi.org/10.1017/nlp.2024.55

Kissami I., **Basmadjian R.**, **Chakir O. and Abid M.R.** (2025). TOUBKAL: A high-performance supercomputer powering scientific research in Africa. *The Journal of Supercomputing* **81**(15), 1–45.

Diab N., **Sadat F. and Semmar N.** (2024). Towards guided back-translation for low-resource languages-A case study on Kabyle-French. In *2024 16th International Conference on Human System Interaction (HSI)*. IEEE, pp. 1–4. https://doi.org/10.1109/HSI61632.2024.10613577

Lamar A. and Kaya Z. (2023). Measuring the impact of data augmentation methods for extremely low-resource NMT. In *Proceedings of the Sixth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2023)*, pp. 101–109. https://doi.org/10.18653/v1/2023.loresmt-1.8

Li G., **Liu L.**, **Huang G.**, **Zhu C. and Zhao T.** (2019). Understanding data augmentation in neural machine translation: Two perspectives towards generalization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5689–5695. https://doi.org/10.18653/v1/D19-1570

Cite this article: R'baiti J, Hmamouche Y and El Fallah Seghrouchni A (2026). Maghrebi dialects – Arabic bidirectional translation: an improved transformer with transfer learning. *Natural Language Processing* **32**, 204–223. https://doi.org/10.1017/nlp.2026.10018