Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
Pfaffenwaldring 5B
D-70569 Stuttgart

Master Thesis

Neural Machine Translation of Dialectal-Dialectal Arabic

Malak Rassem

Studiengang:	M.Sc. Computational Linguistics
Prüfer*innen:	Prof. Dr. Sebastian Padó
	Prof. Dr. Sabine Schulte im Walde
Betreuer:	Dr. Dmitry Nikolaev
Beginn der Arbeit:	15.06.2023
Ende der Arbeit:	15.12.2023

Erklärung (Statement of Authorship)

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig verfasst habe und dabei keine andere als die angegebene Literatur verwendet habe. Alle Zitate und sinngemäßen Entlehnungen sind als solche unter genauer Angabe der Quelle gekennzeichnet. Die eingereichte Arbeit ist weder vollständig noch in wesentlichen Teilen Gegenstand eines anderen Prüfungsverfahrens gewesen. Sie ist weder vollständig noch in Teilen bereits veröffentlicht. Die beigefügte elektronische Version stimmt mit dem Druckexemplar überein. [1]

(Malak Rassem)

[1] Non-binding translation for convenience: This thesis is the result of my own independent work, and any material from work of others which is used either verbatim or indirectly in the text is credited to the author including details about the exact source in the text. This work has not been part of any other previous examination, neither completely nor in parts. It has neither completely nor partially been published before. The submitted electronic version is identical to this print version.

Abstract

This thesis addresses the challenging task of neural machine translation (NMT) between various Arabic dialects, an area that has received limited focus in the field of natural language processing. The primary aim is to explore and compare different approaches to dialect-dialect translation, including models trained from scratch, fine-tuning pre-trained monolingual models, and fine-tuning pre-trained multilingual models.

A comprehensive analysis was conducted to evaluate the effectiveness of an "Everything-to-Everything" model compared to models specifically trained for each translation direction. Additionally, the impact of systematically introducing additional data during the training phase, such as various dialects and Modern Standard Arabic (MSA), was examined. The performance of these models was evaluated using a range of automated metrics (such as BLEU and chrF++) and human evaluation of translation quality for a single target dialect. The study also investigates the correlation between machine translation performance and the mutual intelligibility among Arabic dialects based on a range of linguistic distance measures.

The research reveals that fine-tuning a pre-trained monolingual model, AraT5, yields superior performance compared to other approaches, challenging common beliefs about multilingual models in low-resource scenarios. Furthermore, it was found that single-direction models outperform both the everything-to-everything model and the models that incorporated additional data. Moreover, lexical overlap on the type-level achieved higher correlation with the translation quality scores compared to other distance measures.

Through human evaluation, the study validates the effectiveness of the developed models. The findings contribute significant insights into the intricacies of NMT between Arabic dialects, providing a foundation for future research in this field.

Contents

1 Introduction

1.1 Motivation

Arabic dialects are used by millions of people worldwide, and there is a growing demand for automated translation systems that can facilitate communication and understanding among the different dialects.

Unfortunately, all popular translation systems, such as Google Translate, SYSTRAN and Microsoft Translator, use Modern Standard Arabic (MSA) exclusively. Moreover, most NLP research is in MSA; due to its standardisation, since Arabic dialects are not standardised. In addition to that, most research on Arabic dialects in NLP aims to normalise it to MSA. However, these dialects are not just variants of language; they are repositories of rich cultural nuances, historical legacies, and regional identities.

While MSA serves its purpose in formal contexts, its use can often strip communication of the warmth and personal touch that dialects bring. It enables the transmission of stories, traditions, and emotions that might be lost in the standardized form (Wardhaugh and Fuller, 2015). By exclusively relying on MSA, we risk diluting the richness of the Arabic language and culture, making interactions more formal and less genuine. By preserving the flavour and authenticity inherent in Arabic dialects, dialect-dialect translation ensures that the essence of communication remains intact, allowing for meaningful and personal interactions, thereby promoting linguistic diversity and inclusion. Thus, developing methods to deal with this variation is an important area of research in NLP.

1.2 Research Questions

In this thesis project, we aim to explore neural machine translation between Arabic dialects and attempt to answer the following research questions:

1. How does the performance of models trained from scratch, fine-tuned pre-trained monolingual models, and fine-tuned pre-trained multilingual

models compare against each other in the context of dialect-dialect translation?

This can help us gauge how much transfer learning can benefit dialect-dialect translation as well as comparing monolingual models and multilingual models. The fine-tuned pre-trained monolingual model will be represented by AraT5 (Nagoudi et al., 2022), and the fine-tuned pre-trained multilingual model will be represented by mT5 (Xue et al., 2021).

2. How does the performance of a comprehensive "Everything-to-Everything" model compare against models specifically trained for each translation direction?

To further explore multilinguality, we will create separate models, one for each translation direction we will deal with. This approach will help us understand how each model performs when translating between specific dialect pairs. Additionally, we will build another model that includes all dialects we are dealing with at once. This step allows us to explore how well a single model can handle multiple dialects simultaneously, adding an extra layer to our research. It can help us see how multilingual models cope with the complexities of Arabic dialects when they are all in one model, given how closely related they are.

3. What is the impact of systematically introducing additional data during the training phase on the performance of the models?

a) By using additional dialects.

b) By using MSA

This research question is divided into two parts. In the first part, we will investigate the effect of introducing additional dialects. Our approach involves systematically incorporating different dialects to determine which dialects, when added, contribute positively to the training process. The second part involves introducing MSA during training to measure the influence exerted by MSA on the translation models. Detailed information can be found in Section 5.3.3.

4. Can the translation quality scores reflect the mutual intelligibility between Arabic dialects?

This research question seeks to explore whether the quality of the machine-generated translations correlates with mutual intelligibility between different Arabic dialects based on linguistic measures such as vocabulary overlap.

By addressing these research questions, our primary objective is to bridge significant gaps in the current body of research on Arabic dialect translation within the field of NLP.

2 Background

2.1 Arabic and Arabic Dialects

The Arabic language is a member of the Semitic language group, which is part of the larger Afro-Asiatic language family. It is the fifth most spoken language in the world, with over 400 million speakers across the globe and a rich literary history spanning over a millennium. However, the situation in Arabic-speaking countries is a diglossia, which means that there are two varieties of the language used in different contexts: the formal written language, known as Modern Standard Arabic (MSA), and Arabic vernaculars, which include numerous dialects (Versteegh, 2014). MSA is used in formal contexts, such as formal writing, literature, news, and education. Arabic vernaculars pre-social media were mainly spoken; in day-to-day communication, TV programs and movies. Post-social media, written Arabic vernaculars boomed (Zaidan and Callison-Burch, 2014). Books written in Arabic vernaculars are also on the rise.

The diglossic nature of Arabic presents a challenge for natural language processing applications, such as machine translation, as the formal written language and the vernaculars differ significantly phonologically, lexically, morphologically and syntactically. While they simplify certain written Arabic rules, they also introduce new rules.

Furthermore, the Arabic vernaculars are not homogeneous. They do not only differ from MSA but also exhibit variations among themselves, even within the borders of the same country (Owens, 2001). Numerous competing approaches to classifying Arabic dialects have been proposed. Versteegh (2014) classified them geographically into five regions: Arabian Peninsular, Mesopotamian, Levantine, Egyptian and Maghrebi. Although dialects are commonly categorized by country or region, such as Egyptian, Nilo-Egyptian or Mashriqi, it is important to note that each Arabic-speaking country has multiple varieties of dialects with specific linguistic features, thus researchers often note the city variety of the dialect e.g. Cairene. In the 20th century, we saw a trend where the dialect of a political or cultural capital became the de facto national dialect, as observed in Morocco and Egypt, though less so in countries like Syria, Iraq, or Jordan (Ratcliffe, 2021).

There is also a distinction between urban, rural, and nomadic Arabic dialects. Modern speakers of these dialects are not necessarily confined to the traditional lifestyles (e.g., farming or Bedouin life) (Ratcliffe, 2021).

The dialectal variation is a result of the geographical, historical, and social factors that have influenced the evolution of the Arabic language over time. Language contact, a key aspect of these influences, has resulted in numerous variations influenced by ancient local tongues and other languages, such as European languages (Versteegh, 2001; Lucas and Manfredi, 2019). Other than Peninsular dialects, Arabic dialects are thought to be influenced by substrate languages, such as Berber in the Maghreb, Coptic in Egypt, and Aramaic in the Levant and Iraq (Ratcliffe, 2021), making many of these dialects mutually unintelligible and the degree of intelligibility varies from one dialect pair to the other (Kaye and Rosenhouse, 1997).

In this thesis, we will be concerned with the following variants; Lebanese, Egyptian, Gulf, Moroccan, and Tunisian. We will present a short description of each of them.

- Lebanese Arabic: Lebanese is a variant of North Levantine Arabic that is predominantly spoken in Lebanon. In everyday conversation, Lebanese people often code-switch between Arabic, French, and English (Bassam, 2022). Within

Lebanese Arabic, there are several regional varieties. We will focus on the dialect of the capital city of Beirut, which will be abbreviated as BEI from now on.

- Egyptian Arabic: The Egyptian dialect continuum is spoken by more than 100 million people in Egypt, making it the spoken variety with the biggest number of speakers. Due to Egyptian influence throughout the area, particularly through Egyptian music and movies, Egyptian Arabic is also intelligible in the majority of Arabic-speaking countries (Versteegh, 2014). The Coptic language had an impact on the dialect's phonetics, vocabulary and grammatical structure (Bishai, 1962). Its vocabulary is additionally influenced by Turkish, French, Italian, Greek, and English (Hinds and Badawi, 1986). Cairene Arabic, the dialect from the capital city, Cairo is the most prominent dialect of Egyptian Arabic. This Cairene variety will be our focus and will be referred to as CAI from here on.

- Gulf Arabic: Gulf Arabic is a dialect continuum made up of several closely related and somewhat mutually intelligible dialects, native to the Arabian Peninsula. The dialects spoken in the Arabian Peninsula are closer to MSA than elsewhere in the Arab world (Al-Kahtany, 1997). We will focus on a Qatari variant from the capital city of Doha, referenced as DOH from now on.

- Moroccan Arabic: Moroccan Arabic is a variant within the Maghrebi Arabic dialect continuum, with many regional dialects and accents. The majority of other regional accents are overshadowed by the prominent dialect, which is the one spoken in major cities and also dominates the media. Moroccan has a significant amount of Berber loanwords as well as French and Spanish loanwords (Ennaji, 2005). We will be focusing on the dialect spoken in Rabat, the capital, referred to as RAB going forward.

- Tunisian Arabic: Tunisian is also another variant of Maghrebi Arabic, making to a small extent mutually intelligible with Moroccan but only slightly intelligible, if at all, with Egyptian, Levantine, or Gulf Arabic. There are loanwords

Figure 1: World Map Section with Pinpointed Cities: This section of the world map distinctly marks the geographical positions of BEI (1), CAI (2), DOH (3), RAB (4), and TUN (5), each represented by a numbered pinpoint for easy identification.

from French, Berber, Turkish, and Italian in Tunisian Arabic. Tunisians also frequently code-switch, primarily to French (S'hiri, 2013). We will focus on the dialect of the Tunisian capital Tunis, henceforth referred to as TUN.

All the dialects we are dealing with are urban sedentary dialects. No Bedouin varieties are included. Figure 1 shows the geographic location of these dialects.

According to the findings presented in Ratcliffe (2021), there exists a varying degree of similarity between different dialects and MSA. Table 1 partly reproduces Table 1 from this study, showcasing data relevant to the dialects we are examining. While the original comparison in Ratcliffe (2021) does not encompass all the exact cities in our study (with the exception of Cairo), it still offers a framework for understanding the relative proximity of each dialect to MSA. We can observe that, as previously mentioned, the Gulf variant is the closest to MSA, followed by Levantine (represented in this case by Damascene), Maghrebi, and then Egyptian with almost the same score as Maghrebi.

Table 2, also partly reproduces Table 3 from Ratcliffe (2021), which provides a detailed lexical overlap comparison between Arabic dialects that are relevant to this

Dialect	Lex	Phon	Morph	Synx	Avg	Rank
Gulf	82	88	87	84	85.25	1
Damascus	85	79	72	84	80	5
Morocco	77	79	68	86	77.5	7
Cairo	84	79	75	70	77	8

Table 1: Excerpt from Ratcliffe (2021) ranking of Arabic Dialects in terms of conservation to MSA across lexical, phonological, morphological, and syntactic features

	Mor	Cai	Dms	Glf
Mor	-	72	76	76
Cai	72	-	83	79
Dms	76	83	-	83
Glf	76	79	83	-

Table 2: Excerpt from Ratcliffe (2021) representing the lexical overlap comparison between Moroccan, Cairene, Damascene, and Gulf variants

study. The lexical overlap was based on the Swadesh list of each dialect. The data reveals that the highest degree of lexical overlap is observed in two pairs: between the Damascene and Cairo dialects, and between the Gulf and Damascene dialects. This suggests a significant level of lexical similarity within these pairings. Conversely, the table highlights the lowest lexical overlap between the Cairene and Moroccan dialects, indicating a more distinct lexical divergence between these two regional language variants.

2.2 Machine Translation

2.2.1 Basic Approaches to Machine Translation

Machine Translation (MT) is a computer application that enables the translation of texts or speech from one natural language (known as the source) to another (known as the target). The goal of MT is to generate a sentence in the target language that conveys the meaning of the source sentence (Koehn, 2010).

There are different approaches to MT, and one traditional approach is Rule-Based MT. This approach relies on a set of linguistic rules and bilingual dictionaries. These rules are meticulously crafted by language experts to ensure accurate translation between the source and target languages. The process involves analyzing the source text, breaking it down into its syntactic and semantic components, and then generating the target language text based on the corresponding rules and lexicon entries (Nirenburg, 1989). While the rule-based approach has its merits, especially in preserving linguistic nuances, it requires extensive manual effort in rule creation and maintenance.

On the other hand, Statistical Machine Translation (SMT) relies on statistical models that are trained on sentence-aligned parallel corpora. These models learn the probability of a word or phrase in the source language being translated into a word or phrase in the target language. The translation process involves searching for the most probable translation given the source sentence and the trained model (Koehn, 2010).

Neural Machine Translation (NMT) models, which employ neural networks, have gained prominence and replaced SMT as the mainstream approach to MT (Stahlberg, 2020; Koehn, 2020; Zakraoui et al., 2021). The evolution of NMT models began with the Recurrent Neural Network (RNN) architecture. In this approach, the source language sentence is fed into the RNN encoder, which encodes it into a fixed-length vector representation at each hidden state. The decoder then utilizes this vector to generate the target language sentence word by word.

However, RNNs have limitations when it comes to translating long sentences. One crucial drawback is their restricted ability to represent features effectively with fixed-length vectors (Cho et al., 2014). Furthermore, RNNs face challenges in capturing long-term dependencies within a sequence due to the vanishing and exploding gradient problem.

To address these limitations, Long Short-Term Memory (LSTM) networks, a special type of RNN, were used in NMT (Sutskever et al., 2014). LSTMs incorporate memory cells that enable them to capture long-term dependencies more effectively.

To provide additional word alignment information for translating long sentences and overcome the shortcomings of RNNs, the attention mechanism was introduced by Bahdanau et al. (2014). This mechanism dynamically determines word alignment and allows the decoder to focus on different parts of the input sentence based on the current context.

Nevertheless, both RNNs and LSTMs process input sequences sequentially, which can be computationally expensive for long sequences. In contrast, Convolutional Neural Networks (CNNs) were suggested as an alternative, as they can process the input sequence in parallel, resulting in faster computation (Gehring et al., 2017).

A significant shift in NMT architectures occurred with the advent of the Transformer model. The Transformer is a sequence-to-sequence model that uses self-attention mechanisms in both its encoder and decoder to weigh the importance of each word in context. Furthermore, the decoder employs a cross-attention mechanism that relates the input and output sentences during the translation process (Vaswani et al., 2017). This architecture has become increasingly popular in NMT. It has been adopted by current state-of-the-art NMT models as it not only improves the performance of NMT but also speeds up the inference process. (Stahlberg, 2020)

2.2.2 Decoding Strategies

Decoding in machine translation is the process of determining which translation has the highest score. Since there are exponentially many options at each step, even for an input sentence of moderate length, it is computationally too expensive to go through every potential translation, score it, and then select the best one. In fact, decoding for machine translation models has been proven to be NP-complete. Instead, heuristic search strategies are employed, which enable searching for the optimal translation more effectively. While finding the best translation using these strategies is not guaranteed every time, we do expect to find it frequently enough, or at the very least, a translation that is really similar Koehn (2010). The decoding process is pivotal in shaping the output text therefore, the selection of a decoding strategy in machine translation is a decision that significantly impacts the transla-

tion's quality. Among the various strategies available, greedy decoding, beam search, top-k sampling, and top-p (nucleus) sampling are prominent for their distinct approaches and resulting translations. We will briefly highlight the differences between them.

The simplest decoding strategy is the greedy decoding strategy. This strategy simply predicts the highest probability token at each position in the sequence. Although this method is fast and computationally efficient because it makes a decision at each step, it is not able to look ahead in the sequence and thus can not reevaluate its choices later on. This means it can miss more contextually appropriate or more coherent translations.

Beam search builds upon the basic idea of greedy search by considering several translation possibilities at each step. Instead of selecting the single most probable token at each step, beam search keeps track of a predetermined number of options known as the beam width, and eventually chooses the sequence with the highest overall probability. Although this approach requires greater computational resources, it often produces translations that are more accurate, since it assesses several translation options simultaneously, it provides an opportunity to rectify past token choices.

In top-K sampling, which was introduced by Fan et al. (2018), the K most likely next words are filtered and the probability mass is redistributed among them and then the next predicted word will be sampled from these K words only. One problem with top-K sampling is that, because K is fixed, if the probability distribution of the next word is extremely sharp, very unlikely words could be selected among these K words.

To mitigate top-K's issue, top-p sampling was introduced by Holtzman et al. (2020). Rather than selecting the K most likely tokens, top-p sampling selects the smallest set of tokens whose cumulative probability exceeds the probability p. As a result, the size of the set of tokens to sample from can vary according to the next token's probability distribution.

2.3 Low-Resource Languages

The performance of NMT significantly relies on large training data with millions of training examples. NMT systems demonstrate a direct correlation between the amount of data and translation accuracy (Ji et al., 2020). Thus, NMT's performance varies between high-resource and low-resource languages. High-resource languages, with ample available data, tend to yield better results compared to low-resource languages (Koehn and Knowles, 2017). It is worth noting that even languages typically categorized as high-resource can have low-resource domains, where supplementary language resources, such as lexicons, can be utilized to increase effectiveness. However, such resources do not exist for many low-resource languages (Ranathunga et al., 2023).

Although there is no universally agreed definition for low-resource languages, Magueresse et al. (2020) characterizes them as languages that are less studied, resource-scarce, less computerized, less commonly taught, or low-density. NLP researchers consider data availability and the presence of NLP tools as criteria for distinguishing low-resource languages (Hedderich et al., 2021). Classification initiatives have attempted to categorize languages based on these criteria. For example, Joshi et al. (2020) classified over 2,000 languages into six groups based on the availability of raw and annotated datasets per language.

MT stands apart from other NLP tasks due to its bilingual nature. The data availability of a language pair in MT is primarily determined by the availability of parallel sentences between the two languages. Therefore, even if a particular language has a substantial number of monolingual corpora, if it possesses only a limited parallel corpus with another language, that language pair is still considered low-resource. However, there is no standard corpus size for classifying language pairs as low-resource or extremely low-resource (Ranathunga et al., 2023). Early research considered 1 million parallel sentences as indicative of low-resource status (Zoph et al., 2016), while recent studies consider a language pair as low-resource if the available parallel corpora consist of fewer than 0.5 million sentences and as extremely low-resource if it contains fewer than 0.1 million sentences (Lakew et al., 2019;

Zaremoodi et al., 2018).

One of the significant technical challenges in Arabic machine translation arises from the aforementioned issue, the limited availability of datasets and lexical resources that can serve as standardised benchmarks for conducting comprehensive experiments (Zakraoui et al., 2021). Consequently, researchers often collect domain-specific datasets and address Arabic's linguistic complexities based on custom datasets. This challenge becomes even more pronounced when dealing with Arabic dialects, which are considered extremely low-resource languages (Sajjad et al., 2020).

To tackle the low-resource nature of languages, several techniques have been studied. These techniques include leveraging monolingual data, pivoting, multilingual NMT (Lakew et al., 2018b), and transfer learning and finetuning (Ji et al., 2020; Gu et al., 2018). These approaches aim to mitigate resource limitations and enhance the quality of translation for low-resource languages. We will now proceed to present an overview of these techniques.

2.3.1 Leveraging Monolingual Language Models

Language Models (LMs) can be employed to initialize NMT models. This can be done by initializing only the encoder with source embeddings as seen in Abdou et al. (2017)'s study or by initializing both the encoder and decoder with the respective LMs (Ramachandran et al., 2017). Expanding onto that, Zhu et al. (2020) incorporate BERT finetuning for NMT.

2.3.2 Pivoting

Pivoting involves breaking down the translation process of a source-target language pair (X-Z) into two stages: source-pivot (X-Y) and pivot-target (Y-Z). This approach requires training two independent high-resource models: X-Y and Y-Z. Initially, the source sentence is translated using the X-Y model, and then the output is further translated using the Y-Z model to obtain the target sentence (Ranathunga et al., 2023).

While pivot-based models have long been considered a solution for low-resource and zero-shot NMT, recent advancements in multilingual NMT models have surpassed the performance of pivot-based approaches (Arivazhagan et al., 2019).

2.3.3 Multilingual NMT

Multilingual NMT (MNMT) enables the use of a single model for translating between multiple language pairs (Ha et al., 2017) based on either the recurrent model with attention (Johnson et al., 2017) or the Transformer-based model (Vaswani et al., 2018), and the latter has been shown to be superior (Lakew et al., 2018c). Originally introduced to eliminate the need for individual separate bilingual translation models, MNMT models show great promise in translating low-resource language pairs (Ranathunga et al., 2023).

Multiple studies have demonstrated the superiority of multilingual models over their bilingual counterparts, particularly when dealing with a small number of related languages (Lakew et al., 2018c; Tan et al., 2019). However, MNMT systems face various challenges, such as the diverse characteristics of different languages, noise in parallel data, data imbalance across languages, and the curse of multilinguality (Ranathunga et al., 2023; Conneau et al., 2020).

According to Ranathunga et al. (2023) supervised MNMT architectures can be broadly categorized into four paradigms. Firstly, the single encoder-decoder approach, where all source sentences are processed by a shared encoder, regardless of the language, and the decoder can translate into any target language. Secondly, the per-language encoder-decoder paradigm, where each source language has its own encoder and each target language has its own decoder. Lastly, the shared encoder/ decoder at one side with per-language decoder/encoder on the other side.

The current state-of-the-art approach for large-scale MNMT implementations is the adoption of a single encoder-decoder model for all languages (Arivazhagan et al., 2019). This universal model offers advantages in terms of lower complexity and parameter count compared to per-language encoder-decoder models (Ranathunga et al., 2023). Furthermore, it has demonstrated the ability to learn an interlingua

representation (Johnson et al., 2017). One major challenge in employing this architecture is enabling the decoder to correctly identify the target language. Common practices include adding a language identification tag to the source sentence or incorporating the language name as an input feature (Ranathunga et al., 2023).

Extensive research has been conducted to explore the advantages and drawbacks of different architectures. Research by Hokamp et al. (2019) demonstrated that employing a unique decoder for each target language outperforms models with fully shared decoder parameters. Additionally, Sachan and Neubig (2018) found that partial parameter sharing in the Transformer model (Vaswani et al., 2018), yields superior results compared to the full-parameter sharing recurrent model (Johnson et al., 2017). However, the choice of the most suitable model depends on the specific task requirements. In scenarios where hundreds of languages need to be accommodated, maximum parameter sharing, as in Johnson et al. (2017), is preferred to reduce model complexity Ranathunga et al. (2023).

2.3.4 Transfer Learning

Transfer Learning (TL) is a widely utilized technique in low-resource NLP. TL involves leveraging an NMT model trained on a high-resource language pair to initialize a low-resource child model, thereby reducing training time, and improving performance compared to training the child model from scratch, reducing the size requirement on child training data. This approach aims to solve one task from another different, yet related task. TL has demonstrated remarkable efficacy in the context of MNMT, particularly for translating in or between low-resource language pairs (Pan and Yang, 2010; Zoph et al., 2016; Ranathunga et al., 2023). TL consistently outperforms training the child model from scratch when translating low-resource language pairs, even for extremely low-resource children (Lakew et al., 2018a).

A notable case of MNMT-based TL involves finetuning large-scale multilingual language models like mBART (Liu et al., 2020), using limited amounts of parallel data (Cooper Stickland et al., 2021). The success of TL depends on various factors, with the relationship between the languages employed in the parent and child models

being of utmost importance (Dabre et al., 2017; Nguyen and Chiang, 2017; Zoph et al., 2016). Greater relatedness between languages ensures a higher vocabulary overlap when utilizing the surface form as input, resulting in more meaningful cross-lingual embeddings (Ranathunga et al., 2023). This allows for the exploitation of the proximity between an under-resourced language and the closest related resourced language. This is particularly relevant in the context of standard languages and their dialects. (Harrat et al., 2019).

3 Related Work

3.1 Pre-neural MT Work on Arabic Dialects

MT work on Arabic dialects has been limited. Most of the research in Arabic dialects MT primarily focuses on translating dialects into either MSA or English. Very few studies explore translating to a dialect as the target language (Harrat et al., 2019). In fact, to the best of my knowledge, there has only been *one* study (Moukafih et al., 2022) conducted on translating between different dialects in both the source and target languages.

Rule-based approaches with morphological analysis and transfer rules to normalize dialectal words into their MSA equivalent are widely used for translating between MSA and dialects. On the other hand, the dominant methodology in the context of translating Arabic dialects and English involves combining rule-based and statistical approaches and typically depends on MSA as a pivot language. Other approaches employ domain adaptation techniques, treating dialects as a domain adaptation problem.

Abo-Bakr et al. (2008) proposed a system that combined both rule-based and statistical approaches that employed morphological analysis on the input and an Egyptian-MSA lexicon to map Egyptian Arabic to MSA.

Sawaf (2010) also proposed a hybrid MT system to normalize dialectal words. The normalization process involved mapping the dialectal words at the character-

and morpheme-level. The normalized input was then translated to English using either a hybrid or a statistical MT system using MSA as a pivot language.

Salloum and Habash (2011) also mapped dialectal Arabic to MSA specifically to reduce out-of-vocabulary (out-of-vocabulary) words when translating between Arabic and English using an Analyzer for Dialectal Arabic Morphology which they called ADAM.

Furthermore, The Elissa system, also designed by Salloum and Habash (2012), attempts to translate Arabic dialects (Levantine, Egyptian, Iraqi, Gulf) into MSA. The system starts by identifying dialectal words within the source sentence. It then utilizes the ADAM (Salloum and Habash, 2011), morphological transfer rules, and dialect-MSA dictionaries to generate MSA paraphrases, which are then used to construct an MSA lattice. The constructed lattice is subsequently subjected to n-best decoding and selection using a language model, with the goal of identifying the most suitable MSA translations.

Zbib et al. (2012) attempted to translate from Dialectal Arabic (Egyptian and Levantine) to English directly, using a phrase-based hierarchical model while also experimenting with pivoting on MSA. An interesting finding in the study was incorporating a 150M-word MSA corpus when using 200k words of dialectal data the performance improves significantly. However, when the available dialectal data exceeds 400k words, adding MSA training data no longer enhances performance; instead, it has a negative impact.

Al-Gaphari and Al-Yadoumi (2010) used a rule-based approach to convert the Sanaani dialect from Yemen to MSA. Similarly, another rule-based approach, which was devised by Mohamed et al. (2012), focuses on the lesser-studied direction; producing Egyptian Arabic from MSA.

Hamdi et al. (2013) introduced a translation system between MSA and Tunisian dialect verbal forms. Their approach is based on a deep morphological representation based on A Morphological Analyzer and Generator for the Arabic Dialects (MAGEAD) which was introduced by Habash and Rambow (2006).

Sajjad et al. (2013) presented an SMT system designed for translating Egyptian

Arabic to English. Their approach involved converting Egyptian Arabic to MSA through a character-level transformational model that encompassed various linguistic aspects such as morphology, phonology, and spelling. This model was trained using a collection of Egyptian-MSA word pairs.

Tachicart and Bouzoubaa (2014) employed a rule-based approach to translate the Moroccan dialect to MSA. Their system utilized morphological analysis performed with the Alkhalil morphological analyzer (Boudlal et al., 2010), in which they incorporated Moroccan affixes and a bilingual dictionary, collected from television productions scenarios and the web. However, no evaluation of this work was provided.

A framework for translating Tunisian dialect text from social media into MSA was presented by Sadat et al. (2014). Their word-based approach relied on a bilingual lexicon and grammatical mapping rules, with disambiguation performed using an MSA language model.

Jeblee et al. (2014) introduced an MT system that, like the one by Mohamed et al. (2012), delves into the lesser explored direction by translating from English to a dialect through the use of MSA as a pivot. The system is built upon an MT model trained on an English-MSA parallel corpus. The output is then further translated into Egyptian by employing both dialect and domain adaptation techniques. The study's main finding highlights the potential for improving machine translation quality by leveraging domain adaptation between MSA and the Egyptian dialect.

Al-Mannai et al. (2014) put forward an unsupervised approach of morphological segmentation for Arabic dialects, to enhance the quality of SMT from Qatari Arabic to English.

Durrani et al. (2014) focused on improving the translation quality from Egyptian to English by addressing out-of-vocabulary words. They employed a large monolingual language model to score MSA candidates for Egyptian out-of-vocabulary words. The candidates were generated through spelling correction and synonym suggestions based on the context. The MSA results were then translated into English using an SMT system.

Aminian et al. (2014) also addressed the challenge of out-of-vocabulary words in Dialectal-to-English SMT. They approached this issue by normalizing dialectal words to their MSA equivalents using AIDA and MADAMIRA, which are tools developed by Elfardy et al. (2014) and Pasha et al. (2014) respectively.

3.2 NMT Work on Arabic Dialects

There is very limited research for NMT for Arabic dialects. In Shapiro and Duh (2019) study, the authors examine the advantages of performing dialect identification for Arabic to English NMT using Transformers (Vaswani et al., 2017) as opposed to using a general system that covers all dialects. They also investigate the impact of the quality of dialect identification by introducing random noise to reduce language identification accuracy. The results indicate that there is a cross-over point where the error rate of dialect identification is less than 20%. At this point, the pipelined approach outperforms the integrated, multilingual approach in terms of BLEU scores.

Tawfik et al. (2019) explore the role of word segmentation in NMT for Arabic dialects (Egyptian, Levantine & Gulf) to English. They focus specifically on comparing morphology-aware dialectal Arabic word segmentation with other approaches such as Byte Pair Encoding (BPE) and Sub-word Regularization (SR). The results demonstrate that incorporating sufficiently accurate morphology-aware segmentation in combination with BPE or SR yields superior performance. This concludes that morphology-aware word segmentation offers advantages over language-agnostic methods, particularly when leveraging parallel data from a resource-rich language to enhance the machine translation of a related low-resource language.

Al-Ibrahim and Duwairi (2020) presents a framework for translating Jordanian into MSA using an RNN encoder-decoder model. The experiments were divided into word-level and sentence-level translations.

In the paper by Nagoudi et al. (2021), the focus is on translating code-mixed text, specifically a combination of MSA and Egyptian, which they call MSAEA, into English. The authors develop models under different conditions, employing both

standard seq-to-seq Transformers (Vaswani et al., 2017) trained from scratch and pre-trained seq-to-seq language models. The models achieve reasonable performance using only MSA-EN parallel data with the models trained from scratch. Additionally, finetuning LMs on data from various Arabic dialects proves beneficial for the MSAEA-EN translation task. The study concludes that training models on MSA data proves useful for the MSAEA-to-English task in the zero-shot Egyptian Arabic setting. It also highlights the utility of pre-trained language models like mT5 (Xue et al., 2021) and mBART (Liu et al., 2020) in code-mixing tasks.

The previously mentioned study by Moukafih et al. (2022) adopted a different approach, leveraging the linguistic proximity between Arabic dialects. They implemented a neural multi-task learning framework to enable simultaneous translations of multiple dialect pairs.

Slim et al. (2022) present a study on a transductive transfer learning approach for low-resource NMT applied to the Algerian dialect. This approach relies on a finetuning transfer learning strategy that transfers knowledge from a parent model to a child model, aiming to overcome the learning problem associated with limited parallel corpora. The study tests this approach on a sequence-to-sequence model with and without the Attention mechanism (Vaswani et al., 2017). Initially, the models are trained on the parallel multi-dialectal Arabic corpus MADAR (Bouamor et al., 2018) and subsequently switched to a low-resource dataset, PADIC (Meftouh et al., 2015), which includes the Algerian dialect. The paper also explores the impact of transductive transfer learning on Algerian dialect translation from various perspectives. The findings demonstrate that the transductive transfer learning strategy improves the translation performance of the NMT model, irrespective of whether the attention mechanism is employed or not.

4 Datasets

4.1 MADAR

The primary dataset for this research was the Multilingual Arabic Dialect Applications and Resources (MADAR) dataset (Bouamor et al., 2018). The creation of this corpus involved the translation of a selection of sentences from the Basic Traveling Expression Corpus (BTEC) (Takezawa et al., 2007), which is a multilingual spoken language corpus containing tourism-related phrases commonly found in travellers' phrasebooks into various Arabic dialects by their native speakers. BTEC is particularly applicable due to its conversational style that features dialogues between tourists and guides. This aspect makes it especially relevant, as it mirrors the primary contexts in which these dialects are typically used.

MADAR is widely recognized as a comprehensive resource for Arabic dialects, offering a 26-way parallel data structure that includes MSA and 25 city dialects. The breadth of this dataset makes it an ideal choice for our experiments. However, there is a significant variation in the volume of data available for each dialect within the MADAR dataset. Five dialects - Cairo (Egyptian), Beirut (Lebanese), Doha (Qatari), Rabat (Moroccan), and Tunis (Tunisian) - have 12,000 sentences each, while the remaining dialects only have 2,000 sentences each. Given this disparity, our experiments will primarily focus on the five dialects with larger datasets to ensure robustness in our models' training and finetuning processes

A recent paper by Facebook (Team et al., 2022) also utilized the MADAR dataset, mapping 16 of its dialects to the 8 Arabic dialects in their NLLB-200 multilingual model. However, the specifics of this mapping were not detailed in the paper. Indeed, many of the dialects can be collapsed into coarser macro-dialects which may generalize the smaller differences across cities. However, given the potential for larger city dialects to overpower smaller ones in the model, and the risk of adding noise through the inclusion of essentially identical sentences from different cities, we have decided to use the MADAR city-dialect variant with the cities that have 12,000 sentences without collapsing other cities into them.

Lang	**Total Types**	**Total Tokens**
Doha	9967	64096
Cairo	12724	69695
Tunis	12911	66005
Beirut	12970	63854
MSA	13296	81783
Rabat	13653	72175

Table 3: Total types and tokens for the relevant Arabic Dialects and MSA, sorted ascendingly by the total types

	TUN	RAB	CAI	BEI	DOH	MSA
TUN	-	0.1904	0.1716	0.1604	0.1731	0.1490
RAB	0.1904	-	0.1703	0.1572	0.1696	0.1675
CAI	0.1716	0.1703	-	0.2143	0.2365	0.2080
BEI	0.1604	0.1572	0.2143	-	0.2156	0.1662
DOH	0.1731	0.1696	0.2365	0.2156	-	0.2122
MSA	0.1490	0.1675	0.2080	0.1662	0.2122	-

Table 4: Jaccard similarity of lexical types between pairs of dialects and MSA.

The dataset identifies the split for each sentence, allowing for division into training, validation, and testing sets. It includes 9,000 sentences intended for training, 1,000 for validation, and 2,000 designated for testing purposes.

Table 3 describes how many types and tokens there are for each of our relevant dialects in the entire corpus. To dive further into corpora statistics, we also measure lexical overlap between the dialects using both Jaccard similarity and Jenssen-Shannon divergence.

Jaccard similarity measures similarity between finite sample sets and is defined as the size of the intersection divided by the size of the union of the sample sets, illustrated in Equation 1. In our case, the set is the types in the dialect. The results are shown in Table 4.

	TUN	RAB	CAI	BEI	DOH	MSA
TUN	-	0.3951	0.4175	0.4358	0.4295	0.4588
RAB	0.3951	-	0.4176	0.4484	0.4223	0.4486
CAI	0.4175	0.4176	-	0.3666	0.3283	0.3990
BEI	0.4358	0.4484	0.3666	-	0.3854	0.4548
DOH	0.4295	0.4223	0.3283	0.3854	-	0.4103
MSA	0.4588	0.4486	0.3990	0.4548	0.4103	-

Table 5: Jenssen-Shannon divergence of lexical types between pairs of dialects and MSA.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

Jensen-Shannon Divergence is used to measure the similarity between two probability distributions. It is based on the Kullback-Leibler divergence, with the key difference that it is symmetric. For two probability distributions P and Q, the Jensen-Shannon divergence is defined as shown in Equation 2, where M is the mean of P and Q, and KL is the Kullback-Leibler divergence as defined in Equation 3. In our case, P and Q were relative frequencies of tokens in different dialects. The results are shown in Table 5.

$$\mathrm{JSD}(P, Q) = \frac{1}{2}\mathrm{KL}(P \parallel M) + \frac{1}{2}\mathrm{KL}(Q \parallel M) \tag{2}$$

$$\mathrm{KL}(P \parallel Q) = \sum_i P(i) \log\left(\frac{P(i)}{Q(i)}\right) \tag{3}$$

We employed two distinct approaches for analyzing lexical similarities since one is a type-level method and the other is a token-level method. Both methods yielded

relatively consistent results, concurring that CAI and DOH exhibit the greatest similarity, whereas MSA and TUN are the most dissimilar.

Echoing the findings of Ratcliffe (2021), our results in Table 4 similarly suggest a high degree of similarity between DOH and MSA. However, some of our other findings do not completely align with those reported in Ratcliffe (2021) as we saw in Table 2. This divergence is not unexpected and can be attributed to the differences in the data sets and methodologies used in our respective studies. Such variance highlights the inherent complexity in dialectal analysis and emphasizes how methodological choices can significantly influence research outcomes. We will compare the different matrices against the empirical data from our translation experiments, in order to identify the matrix that offers the best explanation and alignment with the experimental results.

4.2 PADIC

Another dataset called PADIC (Meftouh et al., 2015) is also relevant for translation between Arabic dialects. The PADIC dataset is much smaller than the MADAR dataset and comprises approximately 6000 parallel sentences for MSA and the following dialects: Algiers, Annaba, Syrian, Palestinian, Sfax, and Moroccan. While these dialects do not directly align with our intended translation directions, the dataset offers valuable opportunities for conducting additional zero-shot testing. To effectively conduct zero-shot evaluations, viable scenarios are limited to those where MSA is the target language, as this provides access to a gold standard reference. Our approach will include a zero-shot assessment of the "everything-to-everything" model, alongside the evaluation of the single-direction model where the source is the dialect most closely related for each dialect in PADIC.

5 Methodology and Experimental Setup

5.1 Overview

We present an overview of the experiments we hold in Figure 2.

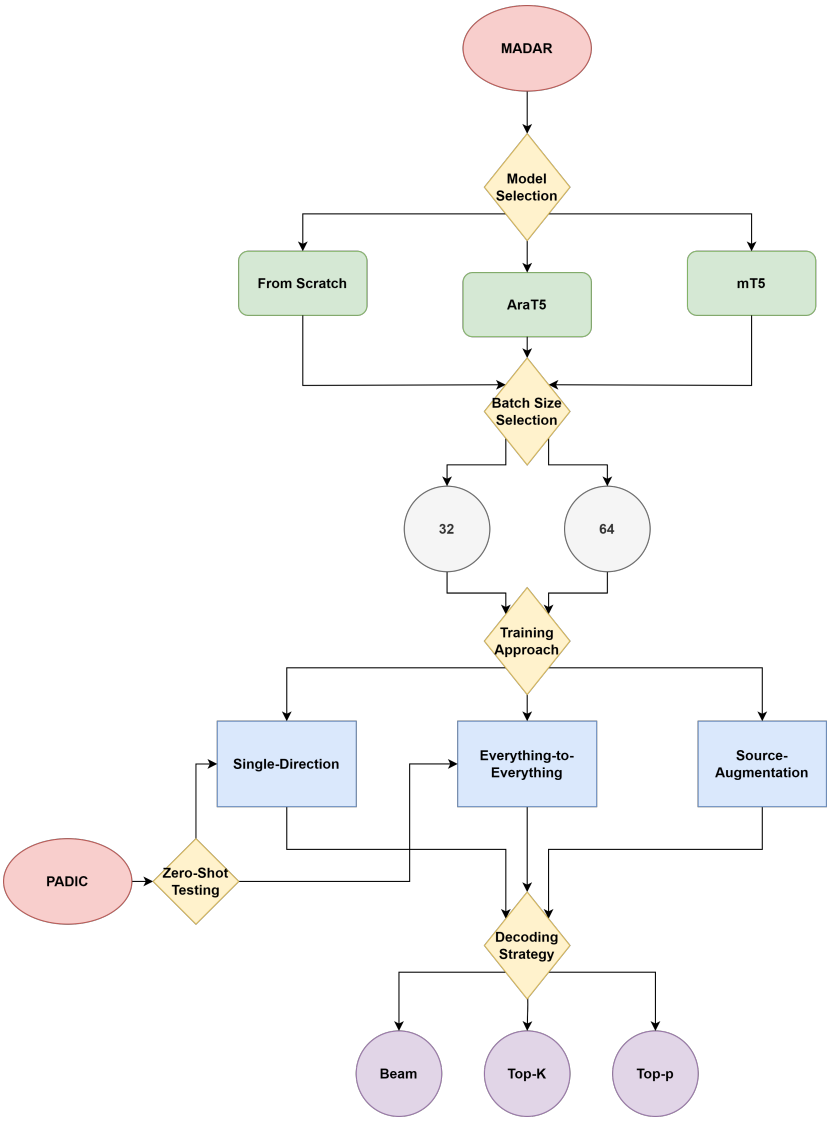

Figure 2: Diagram showcasing our experiments

We describe these steps in the following subsections.

5.2 Model Selection

5.2.1 Training from scratch

Despite the limited training data of 9,000 sentence pairs, training from scratch may serve as a baseline for the transfer learning approach to evaluate how well NMT models can learn to translate between Arabic dialects, which can be thought of as closely-related low-resource languages, without leveraging models pretrained on a vast amount of MSA data. This can also be used to measure how much the performance will improve when implementing more elaborate methods.

Using PyTorch's dynamic features and comprehensive documentation, we created a vanilla Transformer model from scratch, adhering to Vaswani et al. (2017)'s framework. The process of building from scratch provided us the flexibility to implement custom features, therefore, in an effort to enhance the model's comprehension of Arabic syntax and semantics, we took a step further and included a morphological tokenizer to handle Arabic dialects more effectively.

However, unfortunately, as we had anticipated, the lack of sufficient sentence pairs in our dataset presented the model with difficulties, leading to repetition in the outputs and a BLEU score of 0.

5.2.2 Finetuning Models

The next step would be employing transfer learning by finetuning both the AraT5 model Nagoudi et al. (2022) and the mT5 model Xue et al. (2021) for each translation direction. AraT5 is a monolingual model while mT5 is a multilingual model. Both models were pretrained on Arabic data, which is primarily MSA mixed with some dialectal data. Finetuning offers an opportunity to leverage the pre-existing knowledge of MSA in these models, adapting the models to individual dialects, to improve the performance of translating between Arabic dialects.

The finetuning process will be carried out using the transformers Python library provided by HuggingFace. Following the recommendation of AraT5's authors, We

will be finetuning the model for 22 epochs, as that is indeed when the model converges and the training loss stabilizes. Throughout this process, the default Adam optimizer will be employed.

The selection of the optimal model will be based on achieving the highest BLEU score on the validation set, ensuring the best possible performance in terms of language translation accuracy. This approach entails that the decoding strategy employed during validation in the training stage will directly influence the end model's performance. Although greedy is the default setting, beam search has proven to be the best performing in the context of machine translation (Wiher et al., 2022). Consequently, we opted to use a beam width of 4. This decision is based on the observation that while wider beam widths offer marginal improvements in performance, they also lead to a disproportionate increase in computational time, thereby making a beam width of 4 an optimal balance between efficiency and translation accuracy. To thoroughly evaluate the influence of different decoding strategies, we will conduct additional testing on the saved models using top-K and top-p sampling methods.

In the initial phase of our experimentation, we systematically evaluated the model's performance across various batch sizes. Our investigation began with smaller batch sizes, specifically 8, and incrementally increased to 16, 32, and finally 64. The progression beyond a batch size of 64 was not feasible due to memory limitations inherent in our setup. Based on these constraints and the performance outcomes observed, we decided to proceed further in our experiments using the two most optimal batch sizes identified: 32 and 64.

We will also incorporate mixed precision training, which strategically combines different numerical precisions within the training process. This approach is designed to optimize computational efficiency while maintaining model accuracy by utilizing the speed benefits of lower-precision arithmetic and counterbalancing it with higher-precision formats, where necessary to ensure stability and accuracy in model performance.

In the preliminary stages of our experiments, we initially trained models using the fp16 precision setting, widely regarded as a reasonable default value. However, this approach resulted in unstable results for T5-based models, a finding not explicitly documented in the available literature or guidelines. Further experimentation, supported by discussions on the Hugging Face forum, revealed that bf16 precision aligns more closely with the original training protocols of T5 models and therefore is more stable and better performing. We therefore use bf16 throughout our experiments.

The initial results from the monolingual model AraT5 were quite promising. However, when we transitioned to the multilingual model mT5, we observed significant differences. It is important to note that mT5 was not only also pretrained on Arabic but it was pretrained on a much larger Arabic dataset than AraT5, encompassing 57 billion tokens, in contrast to AraT5's 29 billion tokens. This means that mT5 was trained on 96.55% more tokens than AraT5.

Despite this substantial data advantage, the data we finetuned it with was not sufficient to overturn the pretraining, as it continued to generate non-Arabic tokens, and it yielded a disheartening BLEU score of zero. We attributed these shortcomings to the fact that mT5's pretraining relied solely on the mC4 corpus without any supervised training.

Consequently, it became evident that mT5 required further fine-tuning before it could be effectively employed in downstream tasks. One potential solution considered was to continue pretraining it with additional Arabic data. However, this approach might have nudged it towards the characteristics of AraT5. We ultimately decided to proceed with AraT5 for our experimental work.

5.3 Training Approaches

We aim to investigate the impact of different training approaches and strategies on the quality of translations between Arabic dialects. We consider three approaches to further understand how these dialects interact:

1. Single-Direction Translation
2. Everything-to-Everything Translation
3. Augmented-Source Translation

Our first approach is a monolingual approach where each direction has its own monolingual encoder and decoder. The second approach follows the first supervised MNMT paradigm mentioned above in Section 2.3.3, where there is a single shared encoder-decoder. Finally, the third approach follows the third paradigm with a shared encoder for source languages and a per-language decoder for target languages. These approaches are more deeply described in the following subsections.

5.3.1 Single-Direction Translation

In this approach, we will create NMT models for each translation direction. For example, if we have dialects A, B, and C, we will train separate models for A-B, B-A, A-C, C-A, B-C, and C-B. Given that our research encompasses five distinct dialects in addition to MSA this creates a total of 30 unique translation directions, and thus 30 different models.

5.3.2 Everything-to-Everything Translation

In this approach, a single NMT model will be trained to translate between all dialects. This can help us gauge how Arabic dialects benefit from multilinguality given how closely related the languages are. The model will be fed with parallel sentences from all possible translation pairs and trained to learn the translations. The source sentence will have a tag prepended to it, which represents the target dialect the model should translate to. These dialect tags are special tokens that are initially added to the tokenizer's vocabulary and then the size of the token embeddings in the pre-trained model is adjusted to align with the new length of the tokenizer's vocabulary.

We will execute our approach in two distinct phases to assess the impact of including MSA in our translation models. Initially, we will conduct experiments with all five dialects without incorporating MSA. Subsequently, we will repeat the process, this time adding MSA to the mix of source dialects. This methodology is relevant considering that the majority of linguistic resources, especially parallel corpora, are in MSA. By comparing the outcomes of these two phases, we aim to determine whether the inclusion of MSA leads to an improvement in translation quality. This will not only shed light on the practical utility of MSA in multilingual translation scenarios but also help in optimizing the use of available linguistic resources.

5.3.3 Augmented-Source Translation

To further delve into multilinguality, our methodological approach involves training separate models for each combination of source dialects, targeting a specific dialect. For each target dialect, we will create a unique model for every possible combination of the four remaining source vernaculars. This results in a total number of 11 models per target dialect calculated using the n-choose-k formula: $\binom{4}{2} + \binom{4}{3} + \binom{4}{4}$. This approach was chosen to gain a comprehensive understanding of how individual and combined contributions of each source dialect influence the translation process.

To eliminate ordering bias, we have decided against the incremental training models with different source dialects one at a time. Instead, we will train each combination independently. This method aims to eliminate any potential biases that might arise from the sequence in which dialects are introduced during the training phase since the combined source sentences will be pre-shuffled before training.

To further illustrate the approach; a practical example of it is training a model for the combination A and B as sources, targeting dialect C. Once trained, this model will be tested and assessed on two distinct translation directions: A to C and B to C. These tests will reveal the model's performance on each language pair, providing valuable insights into its capabilities.

By implementing this process for various dialect combinations, we aim to deeply understand the complex dynamics inherent in multilingual machine translation, par-

ticularly when involving Arabic dialects. This approach will help identify which specific dialects or combinations that most significantly enhance translation performance for each direction. Ultimately, our goal is to determine the effect of each dialect's addition on the quality of the translations produced.

Additionally, mirroring the approach we adopted with the "everything-to-everything" variant, we plan to explore the potential of adding MSA alongside the source and evaluate its impact on the model's performance to gauge how it affects the translation direction, thereby providing valuable insights into its role as a linguistic intermediary.

5.4 Data Prepration

In the MADAR dataset, each city is represented by a TSV file, with the following structure: ID, split, language, and sentence. Initially, we divide each city file into three separate files for the train, validation, and test splits. We then align sentences by their unique identifiers for each translation direction, resulting in three files for each direction. Subsequently, we convert these files into JSON format, with a single JSON array; each element within this array consists of a JSON object that represents parallel sentences where the keys are language codes, and the values are the corresponding sentences.

In the case of single-direction translations, we prepare the data so that each element contains two dialects. However, for the everything-to-everything variant, we include all five dialects and MSA in each element, iterating over the combinations of input and output sentences during training.

Regarding the augmentation process, we create JSON split files for every possible combination of source dialects to each target dialect. Since the source language varies during training, the key for the input sentences is designated as a string representing the combinatory source, while the key for the output remains the target language's code. The key names used in the training data are indeed irrelevant to the actual training process and are purely a design choice.

The PADIC dataset comprises a single XML file containing 6411 "sentence" elements. Each "sentence" element includes sub-elements for different dialects, each representing the same sentence in a specific dialect. The dialects in PADIC, namely; Algiers, Annaba, Syrian, Palestinian, Sfax, and Moroccan are coded as ALG, ANB, SYR, PAL, TUN and MAR respectively. We had to change the code for Sfax to SFX to avoid confusion with the TUN dialect from Tunis in the MADAR dataset. The dialects will be henceforth, referred to by their code. To unify structures across datasets we took the same approach as we did with MADAR. We isolated sentences of each dialect into one text file and then aligned them for our intended directions. Since we are only utilizing this dataset for testing purposes we created a singular JSON file for each direction.

6 Evaluation

6.1 Evaluation Datasets

We will evaluate the effectiveness of our proposed approaches using several datasets. The MADAR dataset Bouamor et al. (2018), which includes separate train, development, and test splits, will serve as our primary dataset for assessing the performance of our models. Specifically, the test set within the MADAR dataset will provide a rigorous evaluation of our models' performance.

As previously outlined in Section 4.2, our approach entails conducting zero-shot testing across the entire PADIC dataset, encompassing both the "everything to everything" model and the single-direction models. This comprehensive evaluation enables us to assess our model's zero-shot capabilities and its capacity to generalize effectively across previously unseen dialects.

6.2 Automatic Evaluation

The performance of our models can be assessed through manual or automatic methods. One of the most widely used automatic methods is the Bilingual Evaluation

Understudy (BLEU) metric, introduced by Papineni et al. (2002).

BLEU is a precision-oriented metric that measures the closeness of the candidate output of a machine translation system to one or more reference translations. It does this by comparing n-grams (contiguous sequences of n items from a given sample of text or speech) in the machine-generated translations to the n-grams in the reference translations, and counting the number of matches. The scores range from 0 to 1, with 1 being a perfect match to the reference translation.

While BLEU is language-independent and has been widely adopted due to its simplicity and correlation with human judgment, it has limitations, especially when dealing with languages with rich morphology like Arabic. To address this, the AL-BLEU metric Bouamor et al. (2014) was introduced, which extends the standard BLEU to better handle the rich morphology of the Arabic language. AL-BLEU incorporates morphological analysis into the evaluation process, providing a better assessment of translation quality for Arabic. Despite the availability of these adapted metrics, many researchers still use the standard BLEU. Only a few have utilized metrics specifically adapted for the Arabic language. We will report both values for comparability.

6.3 Human Evaluation

In order to further assess our translation models' reliability, we plan to conduct human evaluation experiments. For this purpose, we enlisted the help of seven volunteers from Cairo to review the translations in the Cairene dialect. The decision to use Cairene for evaluation was influenced solely by the ready availability of native speakers. Our methodology involved randomly sampling 200 sentences from the 2000 sentences in the MADAR test split, and pairing each group of 40 sentences with one of the source languages. Corresponding translations into Cairene were generated using the single-direction model that was finetuned for that specific language pair.

We constructed four questions for each sentence, resulting in a total of 800 questions. Our evaluation process was designed to thoroughly compare the quality of generated translations with their respective reference translations. To achieve this,

we formulated a set of questions, each targeting a specific aspect of the translation. This included one question to assess the fluency of the generated translation and another to evaluate the fluency of the reference translation. Similarly, we included a question to measure the accuracy of the generated translation and another to judge the accuracy of the reference translation. The purpose of this dual assessment was to enable a comprehensive analysis of both the quality of the translations produced by our model and the quality of the dataset itself, facilitating a direct comparison between the two.

To ensure an unbiased evaluation, the questions were randomized such that no consecutive questions were derived from the same source language or sentence. Each question was presented on a separate page to minimize any potential bias from previous questions. For the fluency evaluation, participants were shown a single sentence and asked to rate its fluency on a Likert scale ranging from 1 (not fluent) to 5 (very fluent). Similarly, for accuracy assessment, participants were presented with both the original source sentence along with its source language and the translated Cairene sentence. They were then asked to rate the accuracy of the translation on a Likert scale from 1 (not accurate) to 5 (very accurate).

Acknowledging that the source sentences may not always be easily understood by our participants, we provided them with the MSA equivalents as references. However, since MADAR was originally translated from English or French to all dialects, rather than directly from MSA, there might be some inconsistencies between the MSA equivalent and the original source sentences. Therefore, the participants were asked to read the source in the original language first, even if they do not fully understand it and if there are clear discrepancies between the original source sentence and the MSA equivalent, then the original source sentence takes precedence in evaluating the output. Furthermore, the participants were also asked not to be guided by their knowledge of MSA grammar and spelling in order to not bias their evaluation and to take acceptable spelling variations into account.

We utilized Google Forms for this evaluation due to its efficient features like, creating questions programmatically and the ability to import questions from spreadsheets using Google Apps Script. This approach streamlined the process of setting

up the evaluation. However, due to limitations of Google Forms, we had to segment the questions, grouping every 100 into a separate form.

7 Results

7.1 Single-Direction

Table 6 shows the results of the translation quality for each translation direction as indicated by BLEU using the setup with a batch size of 32. The scores varied significantly depending on the source and target dialect pairs. At first glance, it is obvious that DOH as a target dialect achieved remarkably higher scores than its counterparts. The highest BLEU score was observed when translating from BEI to DOH, scoring 29.66, followed closely by CAI to DOH with a score of 29.01. The lowest scores were found in translations from RAB to TUN, and from MSA to RAB, with scores of 8.35 and 10.58, respectively. The AL-BLEU score for MSA output was consistently lower than its corresponding BLEU score. In another light, we also evaluated using the ChrF++ metric, a character-based evaluation method that should theoretically be more sensitive to the morphological aspects of the language. However, with this metric, we did not observe any consistent pattern or advantage over using BLEU. One example is shown in Table 7.

Moving on to the setup with a batch size of 64, training with a batch size of 64 achieved better BLEU scores as shown in Table 8 with the highest score being 30.05 for the CAI-DOH direction followed by 29.64 for BEI-DOH, and the lowest score being 10.99 for the MSA-TUN direction, followed by 11.30 for the RAB-TUN direction. Most language directions saw an increase in BLEU scores. The average BLEU score in the first setup is 16.28 while it is 17.20 for the second setup. This means there is an average increase of 0.92 BLEU points per direction when switching from a batch size of 32 to a batch size of 64. This trend was maintained throughout all experiments therefore we will only report the variant with a batch size of 64 for the rest of the experiments. All results with a batch size of 32 can be found in the appendix.

Source \ Target	TUN	RAB	CAI	BEI	DOH	MSA BLEU	MSA AL-BLEU
TUN		10.75	11.34	14.96	22.39	16.13	11.18
RAB	8.35		11.70	14.98	24.06	15.87	11.10
CAI	10.80	10.54		18.84	29.01	19.41	14.49
BEI	13.62	13.39	17.23		29.66	18.10	13.12
DOH	12.40	13.53	14.52	16.54		19.78	14.52
MSA	10.58	11.43	14.38	15.55	28.56		

Table 6: Results of the single-direction models with a batch size of 32 and beam search decoding strategy

Source \ Target	TUN	RAB	CAI	BEI	DOH	MSA
TUN		33.24	33.95	37.91	43.93	35.97
RAB	29.03		35.52	38.49	45.37	36.33
CAI	33.31	33.21		42.87	51.20	40.69
BEI	37.65	37.28	41.24		51.35	39.48
DOH	36.49	36.95	39.44	41.11		41.53
MSA	34.45	33.86	38.54	40.09	51.18	

Table 7: Results of the single-direction models with a batch size of 32 and beam search decoding strategy using the ChrF++ metric

Source \ Target	TUN	RAB	CAI	BEI	DOH	MSA BLEU	MSA AL-BLEU
TUN		11.95	13.49	15.05	24.02	13.27	8.90
RAB	11.30		12.78	15.56	26.43	17.04	11.42
CAI	13.31	13.65		19.67	30.05	20.77	15.10
BEI	12.28	14.45	17.74		29.64	17.80	12.78
DOH	12.54	14.09	16.31	16.57		21.10	15.80
MSA	10.99	14.39	15.21	15.69	28.71		

Table 8: Results of the single-direction models with a batch size of 64 and beam search decoding strategy

Next, we present the results of different decoding strategies. Table 9 and Table 10 show the results of the top-K and top-p decoding strategies respectively. The top-p strategy outperformed top-K; with an average BLEU score of 15.17, compared to 14.51 for top-K. However, it still performs much worse than the original setup with beam search, with a drop of 2.03 BLEU score. This trend was also reinforced throughout all our experiments. Thus, we will only report the original beam search variant for the remainder of the experiments. The rest of the decoding strategies results can be found in the appendix.

Source \ Target	TUN	RAB	CAI	BEI	DOH	MSA BLEU	MSA AL-BLEU
TUN		10.35	12.68	13.76	22.46	12.63	8.21
RAB	9.86		7.83	14.68	24.96	11.68	7.29
CAI	11.65	12.26		18.70	28.56	14.21	9.88
BEI	10.78	12.68	13.93		27.71	11.69	7.63
DOH	11.31	13.06	11.41	13.34		14.29	9.47
MSA	8.22	9.63	8.94	15.01	27.08		

Table 9: Results of the single-direction models with a batch size of 64 and top-K decoding strategy

Source \ Target	TUN	RAB	CAI	BEI	DOH	MSA BLEU	MSA AL-BLEU
TUN		10.83	13.16	14.22	22.84	12.78	8.69
RAB	10.45		8.72	14.83	25.48	12.88	8.39
CAI	12.23	12.74		19.09	29.19	15.46	10.52
BEI	11.27	12.99	14.91		28.77	12.69	8.79
DOH	11.60	13.28	12.01	14.24		16.09	11.10
MSA	8.88	10.92	9.40	15.40	27.77		

Table 10: Results of the single-direction models with a batch size of 64 and top-p decoding strategy

7.2 Everything-to-Everything

Regarding the Everything-to-Everything model, the results are shown in Table 11. This table is concerned with the variant that did not include MSA in its languages. Similar to the single-direction model, the highest score belongs to the CAI-DOH direction while the lowest score belongs to the RAB-TUN direction, 25.44 and 11.13 respectively. The average BLEU score across all directions is 15.09.

Likewise, the results in Table 12 show the Everything-to-Everything model variant that incorporates MSA. It also closely follows the single-direction models with the highest score also belonging to CAI-DOH at 25.78 BLEU points and the lowest scoring direction is MSA-TUN at 10.78. When not considering MSA, the average score of all the other directions is 14.84. This is 0.25 BLEU points lower than the variant that excludes MSA. The only directions that benefit from adding MSA to the Everything-to-Everything setup are TUN-BEI, RAB-BEI, DOH-BEI, CAI-DOH and BEI-DOH. Only directions with BEI or DOH as a target language seem to show some improvement, indicating that these specific dialects are more receptive to enhancements involving MSA.

Source \ Target	TUN	RAB	CAI	BEI	DOH
TUN		12.36	11.42	12.41	22.21
RAB	11.13		11.71	13.05	23.14
CAI	11.67	13.57		14.77	25.44
BEI	11.70	13.48	14.34		25.23
DOH	11.94	14.24	13.66	14.41	

Table 11: Results of the Everything-to-Everything model without MSA, a batch size of 64 and beam search decoding strategy

Source \ Target	TUN	RAB	CAI	BEI	DOH	MSA BLEU	MSA AL-BLEU
TUN		11.98	11.18	13.07	21.82	15.89	10.98
RAB	10.98		10.97	13.66	22.22	16.26	11.24
CAI	11.09	12.92		14.57	25.78	18.60	13.36
BEI	11.09	13.05	13.76		25.37	17.79	12.49
DOH	11.36	13.62	13.23	15.08		18.77	13.57
MSA	10.78	13.18	12.95	13.84	25.19		

Table 12: Results of the Everything-to-Everything model with MSA, a batch size of 64 and beam search decoding strategy

7.3 Augmentation

We now shift our focus to the augmentation setup. The data is organized such that there is a dedicated table presenting the results for each target dialect to specifically highlight the impact of source augmentation on different target dialects. Table 13 shows the results of source augmentation on the target dialect TUN. We can see that there are various improvements across different evaluation directions. Notably the highest improvement is an increase of 0.26 BLEU points for the DOH-TUN evaluation direction when augmenting DOH with BEI in the source.

In contrast, only one direction showed improvement in the target dialect RAB,

	Evaluation Direction	RAB-TUN	CAI-TUN	BEI-TUN	DOH-TUN
	Reference Score	11.30	13.31	12.28	12.54
Combination Source	BEI+CAI		12.18	**12.48**	
	BEI+DOH			11.84	**12.8**
	BEI+RAB	7.44		8.38	
	CAI+DOH		11.44		11.17
	CAI+RAB	**11.5**	11.91		
	DOH+RAB	9.26			9.35
	BEI+CAI+DOH		11.17	11.24	12.09
	BEI+CAI+RAB	3.09	3.24	3.53	
	BEI+DOH+RAB	**11.44**		11.18	11.23
	CAI+DOH+RAB	**11.54**	11.06		12.04
	BEI+CAI+DOH+RAB	10.37	10.57	10.41	10.33

Table 13: Results of the augmentation models for TUN as a target language with a batch size of 64 and beam search decoding strategy

namely the TUN-RAB direction, as shown in Table 14. The largest improvement occurred when augmenting TUN with DOH in the source, attaining an increase of 0.41 BLEU points. However, the augmentation approach does not yield improvements across all target dialects. Specifically, when the target dialects are CAI, BEI, or DOH, there were no improvements observed in any evaluation direction. This is evident from the data presented in Tables 15, 16, and 17, respectively.

Advancing to the next approach, where we augment MSA to the source of every translation direction; the results are exhibited in Table 18. When compared with the results in Table 8 we can see that the results are worsened for almost all translation directions. The only translation direction that improved was RAB-TUN by only 0.08 BLEU points and this can be considered negligible.

	Evaluation Direction	TUN-RAB	CAI-RAB	BEI-RAB	DOH-RAB
	Reference Score	11.95	13.65	14.45	14.09
Combination Source	BEI+CAI		11.94	11.45	
	BEI+DOH			12.67	13.26
	BEI+TUN	**12.15**		12.48	
	CAI+DOH		13.04		14.04
	CAI+TUN	11.68	12.63		
	DOH+TUN	**12.36**			13.45
	BEI+CAI+DOH		12.76	13.21	13.71
	BEI+CAI+TUN	**12.21**	12.28	12.88	
	BEI+DOH+TUN	**12.11**		12.4	13.49
	CAI+DOH+TUN	**12.13**	12.81		12.95
	BEI+CAI+DOH+TUN	7.66	8.35	8.15	8.31

Table 14: Results of the augmentation models for RAB as a target language with a batch size of 64 and beam search decoding strategy

	Evaluation Direction	TUN-CAI	RAB-CAI	BEI-CAI	DOH-CAI
	Reference Score	13.49	12.78	17.74	16.31
Combination Source	BEI+DOH			15.26	14.91
	BEI+RAB		11.75	15.41	
	BEI+TUN	11.29		15.56	
	DOH+RAB		11.3		14.46
	DOH+TUN	10.8			13.04
	RAB+TUN	11.89	11.74		
	BEI+DOH+RAB		9.92	13.19	12.02
	BEI+DOH+TUN	10.93		14.41	14.37
	BEI+RAB+TUN	10.36	10.29	12.83	
	DOH+RAB+TUN	11.29	11.66		14.18
	BEI+DOH+RAB+TUN	5.84	5.89	7.11	7.08

Table 15: Results of the augmentation models for CAI as a target language with a batch size of 64 and beam search decoding strategy

	Evaluation Direction	TUN-BEI	RAB-BEI	CAI-BEI	DOH-BEI
	Reference Score	15.05	15.56	19.67	16.57
Combination Source	CAI+DOH			15.93	15.18
	CAI+RAB		14.74	16.69	
	CAI+TUN	13.03		15.87	
	DOH+RAB		12.31		12.45
	DOH+TUN	11.47			13.52
	RAB+TUN	13.08	13.92		
	CAI+DOH+RAB		12.39	14.53	14.17
	CAI+DOH+TUN	11.98		15.04	14.56
	CAI+RAB+TUN	10.69	11.85	12.89	
	DOH+RAB+TUN	11.32	11.61		13.29
	CAI+DOH+RAB+TUN	10.43	10.26	12.44	12.53

Table 16: Results of the augmentation models for BEI as a target language with a batch size of 64 and beam search decoding strategy

	Evaluation Direction	TUN-DOH	RAB-DOH	CAI-DOH	BEI-DOH
	Reference Score	24.02	26.43	30.05	29.64
Combination Source	BEI+CAI			28.64	28.39
	BEI+RAB		24.19		26.12
	BEI+TUN	23.21			27.31
	CAI+RAB		24.14	26.99	
	CAI+TUN	23.53		27.88	
	RAB+TUN	22.44	22.76		
	BEI+CAI+RAB		22.37	26.48	26.97
	BEI+CAI+TUN	21.96		24.59	25.1
	BEI+RAB+TUN	21.49	21.63		24.1
	CAI+RAB+TUN	22.1	22.78	26.82	
	BEI+CAI+RAB+TUN	20.28	21.43	23.61	23.26

Table 17: Results of the augmentation models for DOH as a target language with a batch size of 64 and beam search decoding strategy

Source \ Target	TUN	RAB	CAI	BEI	DOH
TUN+MSA		11.8	11.76	13.33	23.34
RAB+MSA	11.38		10.65	13.96	23.54
CAI+MSA	10.74	12.69		15.79	28.36
BEI+MSA	11.53	13.61	15.31		27.6
DOH+MSA	11.18	12.51	15.01	15.31	

Table 18: Results of adding MSA to the source language for each direction with a batch size of 64 and beam search decoding strategy

7.4 Zero-Shot Testing

We conducted zero-shot testing on the comprehensive everything-to-everything model and have presented the resulting findings in Table 19. Among the tested languages, the highest BLEU score was achieved by PAL, reaching 2.55 BLEU points, while the lowest score was observed for ANB, with a BLEU score of 0.72.

Initially, our plan was to exclusively perform testing on the single-direction model whose source language is most related to the unseen source dialect. However, to gain a broader understanding and uncover potential unexpected trends, we decided to expand our scope to include all single-direction models. The outcomes of this expanded analysis can be found in Table 20.

Interestingly, the CAI-MSA model consistently outperformed others, regardless of the source language used as input. For instance, when testing with the CAI-MSA model, PAL yielded the highest BLEU score at 9.55 points, whereas the lowest score was a 0.08 BLEU score, attributed to the MAR language tested with the TUN-MSA model.

The CAI-MSA model's strong performance across various source languages contradicts our initial expectations. We initially assumed that, for instance, the TUN-MSA model would excel with SFX input, given that they come from two Tunisian cities, implying high similarity in input. However, our findings revealed a different trend.

Direction	**BLEU**
ALG-MSA	0.81
ANB-MSA	0.72
MAR-MSA	1.58
PAL-MSA	2.55
SFX-MSA	1.54
SYR-MSA	1.97

Table 19: Results of zero-shot testing on the Everything-to-Everything model

		Unseen Source					
		ALG	ANB	MAR	PAL	SFX	SYR
Testing Direction	TUN-MSA	0.72	0.58	0.08	1.79	1.39	1.16
	RAB-MSA	1.50	1.42	3.26	5.18	2.37	3.20
	CAI-MSA	**2.29**	**1.78**	**3.53**	**9.55**	**3.62**	**5.57**
	BEI-MSA	0.95	0.69	1.30	3.82	1.43	2.99
	DOH-MSA	1.87	1.52	2.74	8.41	2.85	5.10

Table 20: Results of zero-shot testing on our single-direction models

7.5 Human Evaluation

As we mentioned above, the single-direction models with a batch size of 64 and beam search decoding strategy performed the best and thus were chosen for human evaluation. Only the models where CAI was the target language were evaluated. The outcomes of the human evaluation of the generated Cairene translations are presented in Table 21. The accuracy rankings of the generated outputs are as follows: TUN, RAB, BEI, MSA, DOH. Notably, these rankings align perfectly with the BLEU scores obtained by our models. While the reference translations undeniably achieve higher accuracy on average compared to the generated translations, it is important to observe that the actual accuracy of these machine-generated translations is more reliable than what is suggested by their corresponding BLEU scores.

	Source				
	TUN	RAB	BEI	DOH	MSA
Fluency of Generated	4.77	4.88	4.80	4.80	4.72
Fluency of Reference	4.86	4.86	4.94	4.91	4.89
Accuracy of Generated	3.97	4.05	4.34	4.44	4.36
Accuracy of Reference	4.66	4.75	4.79	4.76	4.67

Table 21: Average human evaluated scores of fluency and accuracy of the generated and reference outputs for each source language

8 Discussion

In this chapter, we further analyse the results and attempt to answer our research questions which were:

1. How does the performance of models trained from scratch, fine-tuned pre-trained monolingual models, and fine-tuned pre-trained multilingual models compare against each other in the context of dialect-dialect translation?

2. How does the performance of a comprehensive "Everything-to-Everything" model compare against models specifically trained for each translation direction?

3. What is the impact of systematically introducing additional data during the training phase on the performance of the models?

   a) By using additional dialects.

   b) By using MSA

4. Can the translation quality scores reflect the mutual intelligibility between Arabic dialects?

8.1 Model Selection

In addressing our first research question, we discovered that: fine-tuning a pre-trained monolingual model, specifically AraT5, yielded better performance than either training a model from scratch or fine-tuning mT5, a pre-trained multilingual model. This outcome defied our initial expectations. We had anticipated that mT5 would be superior, primarily because it was pre-trained on a considerably larger Arabic corpus, approximately 96.6% more tokens than that used for AraT5, and because typically multilingual models are believed to perform better in low-resource language scenarios.

Although our results contradict these assumptions, they aligned with the findings from the study conducted by Nagoudi et al. (2022). Their research, which also compared AraT5 and mT5 across various tasks, predominantly favoured AraT5. This surprising superiority of AraT5 suggests that there might be specific challenges associated with mT5, possibly linked to tokenization, given that mT5's vocabulary encompasses a significantly larger number of languages compared to AraT5, suggesting the need for further exploration with diverse models and architectural approaches.

Furthermore, when considering the outcomes in more depth, it becomes evident that while the overall scores may appear modest at first glance, the findings presented by Nagoudi et al. (2022) when translating from MSA to English, two significantly more resource-rich languages, and evaluating on the MADAR dataset, the mT5 model achieved 11.84 BLEU points, while the AraT5 model achieved 10.57 BLEU points. These results demonstrate that the outcomes obtained in our current research fall well within the expected range of performance, reaffirming the validity of our findings.

8.2 AL-BLEU

As we mentioned before, we observed that the AL-BLEU score for MSA output was consistently lower than its corresponding BLEU score. This finding is intriguing, especially considering the theoretical underpinnings and expectations surrounding

the AL-BLEU metric.

AL-BLEU is an adaptation of the standard BLEU metric, which is designed to address some of the limitations of BLEU, particularly in the context of languages with complex morphology like Arabic, where it tends to heavily penalize translations. Unlike BLEU, which only focuses on exact matches, AL-BLEU provides partial credit for translations that are close in terms of morphological and syntactic features, as well as stem-matching.

Theoretically, this means that AL-BLEU scores should be higher than BLEU scores for the same set of sentences, as AL-BLEU is more lenient. This expectation is demonstrated by the data presented in Table 2 in Bouamor et al. (2014), partially reproduced here in Table 22. However, our results diverged from this expectation, consistently showing lower AL-BLEU scores compared to BLEU for the same MSA outputs.

One potential explanation for this discrepancy is that AL-BLEU may be more sensitive to word order. As shown in Table 22, the first hypothesis's AL-BLEU score is significantly higher than the rest, a trend not observed with the standard BLEU metric.

Source: France plans to attend ASEAN emergency summit. Reference: وتعتزم فرنسا لحضور قمة الآسيان الطارئة And-intends France to-attend summit the-ASEAN the-emergency			
Hypothesis	**Rank**	**BLEU**	**AL-BLEU**
وتخطط فرنسا لحضور قمة الآسيان الطارئة And-plans France to-attend summit the-ASEAN the-emergency	2	0.0047	0.4816
وتخطط فرنسا لحضور قمة الأسيان And-plans France to-attend summit the-ASEAN	3	0.0037	0.0840
فرنسا تخطط لحضور القمة الطارئة للأسيان France plans to-attend the-summit the-emergency for-the-ASEAN	1	0.0043	0.0940
خطط فرنسا لحضور قمة آسيان الطوارئ Plans France to-attend summit ASEAN the-emergencies	5	0.0043	0.0604
فرنسا لحضور قمة الاسيان خطط الطوارئ France to-attend summit the-ASEAN plans the-emergencies	4	0.0178	0.0826

Table 22: Partly reproduced table from Table 2 in Bouamor et al. (2014) showing an example of MT output evaluated by BLEU and AL-BLEU

8.3 Dialects as Sources and Targets

The examination of the results of the single-direction models, illustrated in Tables 6 and 8, revealed an intriguing observation. We can see that the results from translation directions are not at all symmetric. This asymmetry means that the quality of translation in one direction, for instance from RAB to DOH, does not mirror the quality of translation in the reverse direction, from DOH to RAB. This non-symmetric characteristic in translation quality indicates that the performance of our translation models is significantly influenced by two independent variables: the source language and the target language.

As we previously mentioned, it is obvious at first glance that DOH, as a target language, outperforms others significantly. In addition to that there seems to be a consistent pattern in the performance of both source and target languages across various experiments, for a certain batch size. We will first examine the effect of

Langs	Src Avg (32)	Src Avg (64)	Diff	Tgt Avg (32)	Tgt Avg (64)	Diff
TUN	15.11	15.56	0.44	11.15	12.08	0.93
RAB	14.99	16.62	1.63	11.93	13.71	1.78
CAI	17.72	19.49	1.77	13.83	15.11	1.27
BEI	18.40	18.38	-0.02	16.17	16.51	0.33
DOH	15.35	16.12	0.77	26.74	27.77	1.03
MSA	16.10	17.00	0.90	17.86	18.00	0.14
AVG	16.28	17.20	0.91	16.28	17.20	0.92

Table 23: Impact of increasing the batch size on the languages as source and as targets

increasing batch sizes on these languages, both as sources and targets. The results are outlined in Table 23. This table illustrates that CAI, as a source language, experiences the most substantial improvement when we increase the batch size from 32 to 64, showing an average gain of 1.77 BLEU points. Conversely, RAB sees the most notable enhancement as a target language due to the batch size increment, with an average increase of 1.78 BLEU points. These improvements are notably higher than the average gains of 0.91 and 0.92 for source languages and target languages respectively.

The performance of target languages, ranked from lowest to highest, follows a consistent sequence across batch sizes: TUN, RAB, CAI, BEI, MSA, and DOH. However, the ranking for source languages shows variability with batch size changes. Initially, for a batch size of 32, the order is TUN, RAB, MSA, CAI, DOH, and BEI. This shifts to TUN, RAB, MSA, DOH, BEI, and CAI when the batch size is increased, primarily due to CAI's significant performance enhancement as a source language. Additionally, DOH and BEI consistently exhibit comparable performances, often tying in their rankings.

It is crucial to note that these rankings are not derived merely by ordering the average scores from Table 23. Such an approach would unfairly disadvantage DOH as a source language. This is because there is no direct DOH-DOH translation, and DOH's performance as a target language significantly surpasses others, which would

Source \ Target	TUN	RAB	CAI	BEI	DOH	MSA	AVG
TUN		2.00	1.00	1.00	1.00	2.00	1.40
RAB	1.00		2.00	2.00	2.00	1.00	1.60
CAI	3.00	1.00		5.00	4.00	4.00	3.40
BEI	5.00	4.00	5.00		5.00	3.00	4.40
DOH	4.00	5.00	4.00	4.00		5.00	4.40
MSA	2.00	3.00	3.00	3.00	3.00		2.80

Table 24: Ranking of languages as sources in the single-direction models with a batch size of 32, where higher is better

Source \ Target	TUN	RAB	CAI	BEI	DOH	MSA	AVG
TUN		1.00	2.00	1.00	1.00	1.00	1.20
RAB	2.00		1.00	2.00	2.00	2.00	1.80
CAI	5.00	2.00		5.00	5.00	4.00	4.20
BEI	3.00	5.00	5.00		4.00	3.00	4.00
DOH	4.00	3.00	4.00	4.00		5.00	4.00
MSA	1.00	4.00	3.00	3.00	3.00		2.80

Table 25: Ranking of languages as sources in the single-direction models with a batch size of 64, where higher is better

skew its average score and not accurately reflect its true performance. To address this, the ranking is based on the performance of each source language for every target direction, and then the average rank is calculated, as illustrated in Tables 24 and 25.

When comparing target languages, the differences in performance are less pronounced. Thus, simply ordering the averages provides a reliable representation of their relative performance. This yields consistent results with those obtained by ranking each performance before averaging, as demonstrated in Tables 26 and 27.

Source \ Target	TUN	RAB	CAI	BEI	DOH	MSA
TUN		1.00	2.00	3.00	5.00	4.00
RAB	1.00		2.00	3.00	5.00	4.00
CAI	2.00	1.00		3.00	5.00	4.00
BEI	2.00	1.00	3.00		5.00	4.00
DOH	1.00	2.00	3.00	4.00		5.00
MSA	1.00	2.00	3.00	4.00	5.00	
AVG	1.40	1.40	2.60	3.40	5.00	4.20

Table 26: Ranking of languages as targets in the single-direction models with a batch size of 32, where higher is better

Source \ Target	TUN	RAB	CAI	BEI	DOH	MSA
TUN		1.00	3.00	4.00	5.00	2.00
RAB	1.00		2.00	3.00	5.00	4.00
CAI	1.00	2.00		3.00	5.00	4.00
BEI	1.00	2.00	3.00		5.00	4.00
DOH	1.00	2.00	3.00	4.00		5.00
MSA	1.00	2.00	3.00	4.00	5.00	
AVG	1.00	1.80	2.80	3.60	5.00	3.80

Table 27: Ranking of languages as targets in the single-direction models with a batch size of 64, where higher is better

8.4 Everything-to-Everything Model

In our study, the everything-to-everything model consistently underperformed the single-direction models, providing further evidence of the challenges associated with multilinguality in machine translation for the dialects relevant to us. This difference in performance is evident in the average BLEU scores: the single direction models achieve an average of 17.04, excluding MSA and 17.2 when including MSA. On the other hand, the everything-to-everything model's average BLEU scores are lower, at 15.09 excluding MSA and 15.34 including MSA, resulting in a decrease of 1.95 and 1.86 BLEU points, respectively. It is noteworthy that only two translation directions showed improvement with the everything-to-everything model. When MSA is excluded, the TUN to RAB direction experienced an increase of 0.41 BLEU points, and the DOH to RAB direction improved by 0.15 BLEU points. When MSA is included, the TUN to MSA direction exhibited the largest gain, with an increase of 2.62 BLEU points, followed by a slight increase of 0.03 BLEU points in the TUN-RAB direction. The TUN-RAB direction benefited from the everything-to-everything model in both scenarios. These findings answer our second research question comprehensively, shedding light on the specific contexts where an everything-to-everything model can offer advantages, though limited, compared to the generally more robust performance of single-direction models in most translation directions.

8.5 Augmentation

Addressing our third research question, we explored the impact of augmenting source dialectal data on the translation performance. Our findings suggest a clear pattern: the lower the initial reference score of a translation direction, the more likely it is to benefit from dialect augmentation. Specifically, all translation directions with initial reference scores up to and including 12.54 exhibited improvement at least once when augmented with additional dialects. Among these, TUN to RAB direction stood out, showing the most significant improvement both in terms of frequency and degree. This observation is consistent with our previous results where the TUN-RAB direction was also one of those that benefited from the everything-to-everything

model scenario.

However, our research indicates that adding more source languages does not necessarily lead to better performance. In fact, the most effective results were achieved when adding only one or two languages. This trend suggests diminishing returns with increasing language augmentation and provides insight into why the everything-to-everything model showed limited effectiveness.

Another important aspect to consider is that the improvement in performance through language augmentation does not necessarily correlate with the genealogical or geographical proximity of the augmented language to the evaluation source. Rather, the improvement is often attributable to the augmentation of languages that are inherently better performing as source languages in our experiments, such as DOH, BEI, and CAI. Expanding upon the previously identified trend, our subsequent experiment focused on adding MSA. Notably, MSA is recognized as a relatively weak source language in our research. Consistent with this characterization, we observed that adding MSA did not yield significant improvements in translation quality.

We earlier observed that in the everything-to-everything model variant incorporating MSA, certain translation directions showed improvements, particularly where BEI and DOH were the target languages. However, the absence of similar improvements in scenarios where MSA was augmented to the source leads us to hypothesize that the enhancements noted in the everything-to-everything model may be a unique interaction.

8.6 Human Evaluation

While analysing the human evaluation results, we noticed that the accuracy of the CAI references when TUN and MSA were the source languages was lower compared to when other languages were the source languages. This observation suggests a potential incongruence in the parallel dataset, where some sentences were translated one way in some languages and another way in other languages, particularly between MSA and CAI, and TUN and CAI. Evidence of these inconsistencies is visible in

MSA:	هالو ، هل هذه تعود إلى روي ؟	CAI:	أهلا، هو ده روي؟
English:	Hello, does this belong to Roy?	English:	Hello, Is this Roy?
MSA:	مهزوز ، وليس مضروب.	CAI:	ده مخلوط بالهز، مش بمعلقة.
English:	Shaken, and not blended	English	It is mixed by shaking, not with a spoon.
TUN:	هاذي اخر محطة؟	CAI:	ديه أخر مره نقف فيها؟
English:	Is this the last station?	English:	Is this the last time we stop?
TUN:	البرنامج الخامس أهوكة غادي.	CAI:	رصيف خمسة هناك.
English:	The fifth program is over there.	English:	Platform five is over there.

Table 28: Sample of discrepancies between reference sentences in MADAR.

several examples listed in Table 28, and similar instances are found across all dialect pairs.

A likely explanation for these inconsistencies stems from the translation methodology used in the MADAR dataset. Unlike typical approaches that might use MSA as a primary source for translation into various dialects, MADAR translations were directly derived from English or French. This approach, as explained in Bouamor et al. (2018), was intentionally chosen to avoid the potential bias that might arise from using MSA as a starting point. While this decision had its rationale in maintaining the authenticity and diversity of dialects, it appears to have introduced these inconsistencies that pose challenges for effective machine translation across Arabic dialects.

As we already mentioned, the generated CAI translations were judged to be more reliable than how the BLEU reflects it to be. Typically, BLEU scores are judged to be satisfactory when they exceed 30%. Scores below this threshold are generally not associated with high-quality translations. However, in the case we are discussing, despite the BLEU scores being significantly lower than 30%, the translations are receiving high human evaluation accuracy scores, specifically in the range of 4 and 5, in contrast to the 1 or 2 that would normally be expected for such low BLEU scores.

In Table 29, we highlight examples that achieved high accuracy scores by our

participants but a relatively low BLEU score. In the first sentence, both "حنروح" (ḥnrūḥ) and "احنا رايحين" (eḥnā raiḥīn) translate to "we are going" using different syntactic structures, both of which are correct and natural. Moreover, the word for 'bus' is spelt differently: once as "الاتوبيس" (elotōbīs) and once as "الاوتوبيس" (elōtōbīs). Since dialects are not standardised and this is a loan word, both spellings can be considered correct. However, this sentence suffers the most in BLEU score since there is no exact match.

In the second sentence, both the generated and reference texts convey the same overall meaning. The reference skips the noun adjunct 'reservation', but the meaning is implied in the context. The reference also adds 'and'; however, it is commonly used as an interjection, so it does not affect the meaning. The word "ليا" (leya) is used in the reference sentence to imply possession, in contrast to using "عندي" (ʕndī) that was generated by the model, which is in fact more common. The generated output dropped the subject 'I'. This is common not just in the vernaculars but also in MSA. In fact, the source sentence, which is in MSA, also dropped the subject.

In sentence number three, the reference included a detail (snowstorm) that was not originally in the source sentence. Furthermore, the reference sentence implies the future tense rather than the past tense by using the word "حتوصل" (ḥtwṣl) rather than the word "وصلت"(wṣlt).

There is a clear mismatch between the reference and the source in the fourth sentence; the verb "looking for" was swapped with 'wear' which was correctly translated by the model. The word "شيك" (shīk), however, is more natural in Cairene than "راقية" (raqya), both meaning elegant/chic.

The fifth sentence shows us a case of synonymous words/phrases that are penalised by BLEU since there is only one reference sentence per source sentence. Both

No.	Src	Source Sentence	CAI Reference	Score	Generated CAI	Score	BLEU
1	BEI	رايحين بالباص؟	حنروح بالاوتوبيس؟	5	احنا رايحين بالاتوبيس؟	5	0
		Are we going with the bus?	Are we going with the bus?		Are we going with the bus?		
2	MSA	لدي حجز . هذا هو رقم تأكيد الحجز .	أنا ليا حجز. وده رقم التأكيد.	5	عندي حجز. ده رقم تأكيد الحجز.	5	13.13
		I have a reservation. This is the reservation confirmation number.	I have a reservation. And this is the confirmation number		I have a reservation. This is the reservation confirmation number.		
3	BEI	الطيارة وصلت مأخرة كرمال التلج القوي.	الطيارة حتوصل متأخرة بسبب عاصفة ثلجية.	4.43	الطيارة وصلت متأخر عشان التلج جامد.	5	7.81
		The plane arrived late because of the heavy snow.	The plane will arrive late due to a snow storm		The plane arrived late because the snow is heavy		
4	DOH	ادور شي راقي.	عايزة البس حاجة شيك.	4.26	أنا بادور على حاجة راقية.	5	9.65
		I'm looking for something elegant.	I want to wear something chic.		I am looking for something elegant.		
5	DOH	بغيت تفاح، لو سمحت.	عايز تفاح، من فضلك.	5	عايز تفاح، لو سمحت.	5	23.64
		I want apples, please.	I want apples, please		I want apples, please		
6	RAB	شنو يمكن لي نشوف فهاد الجولة؟	ايه اللي ممكن اشوفة في الرحلة؟	4.43	إيه اللي ممكن أشوفه في الرحلة دي؟	5	14.54
		What can I see in this tour?	What can I see in the tour?		What can I see in this tour?		
7	TUN	واحد شيزبرغر و واحد عصير، يعيشك.	واحد هامبرج بالجنبة و عصير برتقان، إذا سمحت.	4.71	واحد شيزبرجر و عصير، من فضلك.	5	6.50
		One cheeseburger and one juice, please.	One hamburg(er)* with cheese and an orange juice, please		One cheeseburger and a juice, please		

Table 29: An example of generated sentences that were highly scored on average by the annotators but had a low BLEU score.

“لو سمحت” (lw samḥt) and “من فضلك” (mn fḍlk) are phrases that mean “please”. This specific synonymous phrase pair occurs very frequently in both directions and is penalised when one is generated rather than the other.

The reference and the generated are very similar in the sixth sentence, with the generated version adding “دي” (dī) which translates to ‘this’ making it “in this tour” instead of just “in the tour”. A few orthographic differences occur between them. The first word “إيه” (ēh) in the generated sentence has a hamza “ء” underneath the first (right-most) letter “ا” (alif). The fourth word “أشوفه” (ashūfū) also has a hamza above its alif. This represents a glottal stop. Although essential in MSA, in Cairene it is omitted in certain cases, as done in the reference. However, if it exists, it is not considered wrong. It is important to note that it is clear that it is not the intention of the original translator of MADAR to always omit unnecessary hamzas, as we still see it in the reference in “أنا” (anā) in the first sentence and in “إذا” (eza) in the seventh sentence. Unfortunately, we cannot normalise all alifs to be without hamzas since they are still needed in some cases. Another orthographic difference is also in the word “اشوفة” (ashūfū) as written in the reference and “أشوفه” (ashūfū) as written in the generated sentence. We already discussed the hamza; we now discuss the difference in the last (left-most) letter “ـه” vs. “ـة”. Essentially these are two different letters, the haa and the tied-taa, respectively. If this was written in MSA, it would be considered wrong to exchange the two letters; in our case, the reference would be wrong and the generated text would be right. However, a word that ends in a tied-taa is pronounced like a haa with an H-sound if the speaker stops at the word, and it is pronounced with a T-sound if the speaker connects the word to the following word. So in some cases, these two letters are uttered in the same manner. Therefore, it is a common mistake to swap the two letters. However, it remains

understandable. With dialectal orthography not set in stone, it is hard to judge whether the reference is wrong or just different; it will depend on each person's point of view. These orthographic differences prohibit an exact match, and thus the BLEU score becomes lower.

Other than the already highlighted spelling error in the reference, the last sentence has yet another synonymous "please" issue. We can also see a mismatch where the reference introduces an extra detail, 'orange', but the source did not imply what type of juice it was.

Most of these issues would not be solved when using other metrics such as AL-BLEU. The core of the matter doesn't depend solely on BLEU; rather, the complexity arises from the inherent challenge of automatically evaluating translations in languages with non-standard orthographies and spelling.

8.7 Zero-Shot Testing

In our study, we also conducted a comparative analysis of zero-shot translation results between our model and the NLLB-200 model. We see the results of the zero-shot testing on the NLLB-200 model in Table 30. While the NLLB-200 model exhibited superior scores, the margin of improvement was not overwhelmingly large. This similarity in performance suggests that both models face challenges in generalizing effectively to a broader range of contexts. In our case, the limited generalizability can be attributed to two main factors. Firstly, the amount of data available for training our model was relatively small. Secondly, the MADAR dataset, which formed the basis of our training material, is somewhat domain-specific, focusing primarily on touristic sentences. This specificity likely restricts the models' ability to adapt to a wider variety of linguistic contexts.

A notable limitation in our comparative analysis was the inability to use our test set for a direct comparison with NLLB-200. This constraint arises from the likelihood that NLLB-200 had been trained on segments of our MADAR test split. Consequently, using this test data could potentially compromise the validity of the comparison due to the potential data leak in the NLLB-200 model.

Testing Direction	Unseen Source					
	ALG	ANB	MAR	PAL	SFX	SYR
Mesopotamian-MSA	2.92	1.91	5.50	10.37	5.18	7.54
Ta'izzi-Adeni-MSA	2.90	2.16	5.27	10.19	5.06	7.12
Tunisian-MSA	2.65	1.75	4.50	8.42	4.52	5.80
South Levantine-MSA	2.72	2.16	5.23	10.01	4.88	7.34
North Levantine-MSA	2.86	1.92	4.95	10.12	4.95	7.13
Najdi-MSA	2.92	2.00	5.38	10.58	5.16	7.47
Moroccan-MSA	2.84	2.17	5.14	9.06	4.82	6.11
Egyptian-MSA	**4.65**	**3.21**	**7.97**	**16.92**	**7.91**	**11.22**

Table 30: Results of zero-shot testing on the NLLB-200 model

Interestingly, the NLLB-200 model displayed a consistent pattern in its zero-shot translation results that aligned well with the zero-shot evaluation of our models: the Egyptian source direction consistently outperformed others, irrespective of the input source language. This trend aligns with our findings, where CAI emerged as the best overall source language in our evaluations. Thus, it is not surprising that models trained with CAI/Egyptian as a source language yield the most effective results in zero-shot scenarios.

8.8 Interpretation

To interpret our research findings, we initiated a comparative analysis of three distinct matrices representing relationships between the varieties. The first matrix is based on data from Ratcliffe (2021). The second matrix calculates the Jenson-Shannon divergence, which measures the variation in relative token frequencies of each dialect within the MADAR corpora. The third matrix employs the Jaccard similarity index to quantify the lexical overlap of each dialect represented in MADAR.

Among these, the Jaccard Similarity matrix demonstrated the highest correlation with our results, achieving an overall absolute correlation of 0.696. The other two matrices recorded lower correlations of 0.318 and 0.550, respectively. This finding

Lang	Correlation Coefficient
TUN	0.158
RAB	-0.939
CAI	0.846
BEI	0.782
DOH	0.910
MSA	0.951

Table 31: Correlation coefficient of scores of languages as targets with their lexical overlap

is significant as it suggests that, while the data from Ratcliffe (2021) might reflect phylogenetic linguistic relationships, the trends within our specific dataset, MADAR, play a crucial role in shaping the outcomes of our translation tasks.

Moreover, we observed a strong correlation between the language that shares the highest lexical overlap with all other languages and the average ranking of that language as a source in our translations. This correlation was quantified with a coefficient of 0.937, indicating a robust relationship between these two aspects.

Addressing our final research question, we explored whether translation scores correlate with mutual intelligibility, as indicated by lexical overlap. We calculated the correlation coefficient across languages, focusing on their roles as translation targets. The findings, detailed in Table 31, reveal that the correlation between translation scores and lexical overlap is particularly strong only for specific languages, namely: MSA, DOH, CAI and BEI. There is a strong negative correlation for RAB. This is likely anomalous since its scores are very close together.

The pretraining data of AraT5 has also significantly influenced our results. Figure 3, reproduced from Nagoudi et al. (2022), reveals the dialect distribution that AraT5 was trained on. A huge portion, approximately 60.74%, of the dialectal data comes from Gulf countries such as Saudi Arabia, Bahrain, Kuwait, Qatar, the United Arab Emirates, and Oman. This heavy representation of Gulf dialects, which closely align with DOH, offers a plausible explanation for DOH's exceptional performance

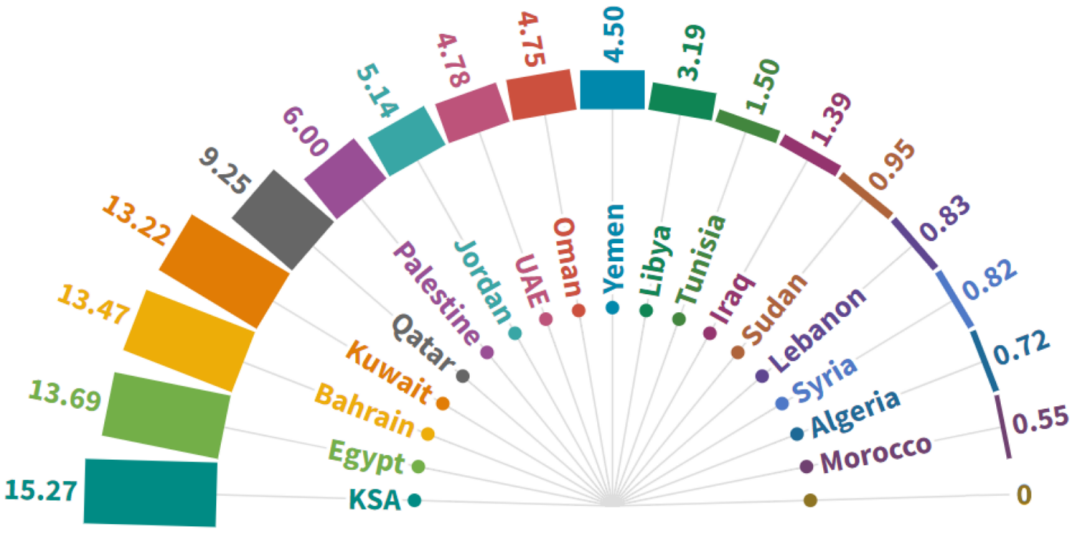

Figure 3: Excerpt from Nagoudi et al. (2022) showing the geographic distribution of the dialectal part of their training data

as a target language in our experiments.

Additionally, the considerable representation of data from Palestine in AraT5's training corpus may account for the superior performance of the Palestinian dialect as a source language in our zero-shot testing.

Conversely, the relatively small percentage of North African languages, including Moroccan, Algerian, and Tunisian dialects, within the pretraining data, sheds light on why our models exhibited challenges when dealing with these specific dialects. The underrepresentation of these dialects in the training corpus likely led to less effective learning outcomes for these languages, resulting in poorer translation performance.

9 Conclusion

This thesis has embarked on an exploratory journey into the domain of NMT between Arabic dialects. Through comprehensive experiments and analyses, it has uncovered valuable insights into the dynamics of dialect-dialect translation, paving the way for future advancements in the field.

Our exploration began with an evaluation of various models trained from scratch, alongside fine-tuned pre-trained monolingual and multilingual models. The results clearly indicated that fine-tuning a pre-trained monolingual model, particularly AraT5, outperformed other methods. This finding emphasizes the potential of monolingual models in handling Arabic dialects, which can be considered low-resource languages, over multilingual models.

The investigation into the "Everything-to-Everything" model versus specific translation direction models revealed that the latter generally provided more accurate translations. This suggests that while multilingual models offer the convenience of handling multiple dialects, they may not always capture the subtle linguistic features as effectively as models dedicated to specific translation pairs.

Our study also delved into the impact of systematically introducing additional data during the training phase. We discovered that the inclusion of more dialects or MSA did not consistently enhance translation performance.

Intriguingly, one of the most compelling findings of this thesis was the identification of a consistent order of effectiveness for both source and target languages in translations. This pattern underscores the fact that certain dialects inherently perform better as either source or target languages in the context of machine translation. This consistency in language performance order was likely due to how close that dialect is to the majority of the rest of the dialects based on lexical overlap. Alternatively, this could result from the disparity in the amount of pre-training data for each language.

Another aspect of this research was the exploration of the correlation between machine translation quality scores and mutual intelligibility among Arabic dialects.

The findings suggest that while there is a correlation, it varies significantly across different dialect pairs. This implies that linguistic similarities and differences play a role in determining the effectiveness of NMT models for Arabic dialects.

The human evaluation of our models further validated their effectiveness in terms of fluency and accuracy. However, discrepancies in reference translations need to be first resolved to ensure the reliability of NMT systems.

In conclusion, this thesis has made significant strides in understanding the complexities of NMT between Arabic dialects. It lays a solid foundation for future research, which could explore more sophisticated model architectures and advanced training techniques. The goal of achieving fluent and accurate dialect-dialect translation in Arabic remains a challenging yet attainable endeavour.

References

Mostafa Abdou, Vladan Glončák, and Ondřej Bojar. 2017. Variable mini-batch sizing and pre-trained embeddings. In *Proceedings of the Second Conference on Machine Translation*, pages 680–686, Copenhagen, Denmark. Association for Computational Linguistics.

Hitham Abo-Bakr, Khaled Shaalan, and Ibrahim Ziedan. 2008. A hybrid approach for converting written Egyptian colloquial dialect into diacritized Arabic. In *The 6th International Conference on Informatics and Systems (INFOS2008)*. Faculty of Computers and Information, Cairo University.

G. H. Al-Gaphari and M. Al-Yadoumi. 2010. A method to convert Sana'ani accent to Modern Standard Arabic. *International Journal of Information Science and Management (IJISM)*, 8(1):39–49.

Roqayah Al-Ibrahim and Rehab M. Duwairi. 2020. Neural machine translation from Jordanian dialect to Modern Standard Arabic. In *2020 11th International Conference on Information and Communication Systems (ICICS)*, page 173–178. IEEE.

Abdallah Hady Al-Kahtany. 1997. The 'problem' of diglossia in the arab world: An attitudinal study of modern standard arabic and the arabic dialects. *al-'Arabiyya*, 30:1–30.

Kamla Al-Mannai, Hassan Sajjad, Alaa Khader, Fahad Al Obaidli, Preslav Nakov, and Stephan Vogel. 2014. Unsupervised word segmentation improves dialectal Arabic to English machine translation. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 207–216, Doha, Qatar. Association for Computational Linguistics.

Maryam Aminian, Mahmoud Ghoneim, and Mona Diab. 2014. Handling OOV words in dialectal Arabic to English machine translation. In *Proceedings of the EMNLP'2014 Workshop on Language Technology for Closely Related Languages*

and Language Variants, pages 99–108, Doha, Qatar. Association for Computational Linguistics.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George F. Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *CoRR*, abs/1907.05019.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Lubna Bassam. 2022. Multilingualism in Lebanon. In *Language and Identity in the Arab World*, pages 174–189. Routledge.

Wilson B. Bishai. 1962. Coptic grammatical influence on Egyptian Arabic. *Journal of the American Oriental Society*, 82(3):285.

Houda Bouamor, Hanan Alshikhabobakr, Behrang Mohit, and Kemal Oflazer. 2014. A human judgement corpus and a metric for Arabic MT evaluation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 207–213, Doha, Qatar. Association for Computational Linguistics.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3387–3396, Miyazaki, Japan. European Language Resources Association (ELRA).

Abderrahim Boudlal, Abdelhak Lakhouaja, Azzeddine Mazroui, Abdelouafi Meziane, MOAO Bebah, and Mostafa Shoul. 2010. Alkhalil morpho sys1: A mor-

phosyntactic analysis system for Arabic texts. In *International Arab Conference on Information Technology*, pages 1–6. Elsevier Science Inc New York, NY.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 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)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Asa Cooper Stickland, Xian Li, and Marjan Ghazvininejad. 2021. Recipes for adapting pre-trained monolingual and multilingual models to machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3440–3453, Online. Association for Computational Linguistics.

Raj Dabre, Tetsuji Nakagawa, and Hideto Kazawa. 2017. An empirical study of language relatedness for transfer learning in neural machine translation. In *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*, pages 282–286. The National University (Phillippines).

Nadir Durrani, Yaser Al-Onaizan, and Abraham Ittycheriah. 2014. Improving Egyptian-to-English SMT by mapping Egyptian into MSA. In *Computational Linguistics and Intelligent Text Processing: 15th International Conference, CICLing 2014, Kathmandu, Nepal, April 6-12, 2014, Proceedings, Part II 15*, pages 271–282. Springer.

Heba Elfardy, Mohamed Al-Badrashiny, and Mona Diab. 2014. AIDA: Identifying code switching in informal Arabic text. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 94–101, Doha, Qatar. Association for Computational Linguistics.

Moha Ennaji. 2005. *Multilingualism, Cultural Identity, and Education in Morocco*. Springer-Verlag.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, and Yann Dauphin. 2017. A convolutional encoder model for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 123–135, Vancouver, Canada. Association for Computational Linguistics.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 344–354, New Orleans, Louisiana. Association for Computational Linguistics.

Thanh-Le Ha, Jan Niehues, and Alexander Waibel. 2017. Effective strategies in zero-shot neural machine translation. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 105–112, Tokyo, Japan. International Workshop on Spoken Language Translation.

Nizar Habash and Owen Rambow. 2006. MAGEAD: A morphological analyzer and generator for the Arabic dialects. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 681–688, Sydney, Australia. Association for Computational Linguistics.

Ahmed Hamdi, Rahma Boujelbane, Nizar Habash, and Alexis Nasr. 2013. The effects of factorizing root and pattern mapping in bidirectional Tunisian - Standard Arabic machine translation. In *Proceedings of Machine Translation Summit XIV: Papers*, Nice, France.

Salima Harrat, Karima Meftouh, and Kamel Smaili. 2019. Machine translation for Arabic dialects (survey). *Information Processing & Management*, 56(2):262–273. Advance Arabic Natural Language Processing (ANLP) and its Applications.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.

Martin Hinds and El-Said Badawi. 1986. *A Dictionary of Egyptian Arabic: Arabic-English*. Librairie du Liban, Kesrouwan, Lebanon.

Chris Hokamp, John Glover, and Demian Gholipour Ghalandari. 2019. Evaluating the supervised and zero-shot performance of multi-lingual translation models. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 209–217, Florence, Italy. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Serena Jeblee, Weston Feely, Houda Bouamor, Alon Lavie, Nizar Habash, and Kemal Oflazer. 2014. Domain and dialect adaptation for machine translation into Egyptian Arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 196–206, Doha, Qatar. Association for Computational Linguistics.

Baijun Ji, Zhirui Zhang, Xiangyu Duan, Min Zhang, Boxing Chen, and Weihua Luo. 2020. Cross-lingual pre-training based transfer for zero-shot neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 115–122.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Alan S. Kaye and Judith Rosenhouse. 1997. Arabic dialects and Maltese. In Robert Hetzron, editor, *The Semitic Languages*, chapter 14, pages 263–311. Routledge.

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press, USA.

Philipp Koehn. 2020. *Neural Machine Translation*. Cambridge University Press.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Surafel M. Lakew, Aliia Erofeeva, Matteo Negri, Marcello Federico, and Marco Turchi. 2018a. Transfer learning in multilingual neural machine translation with dynamic vocabulary. In *Proceedings of the 15th International Conference on Spoken Language Translation*, pages 54–61, Brussels. International Conference on Spoken Language Translation.

Surafel M Lakew, Marcello Federico, Matteo Negri, and Marco Turchi. 2018b. Multilingual neural machine translation for low-resource languages. *IJCoL. Italian Journal of Computational Linguistics*, 4(4-1):11–25.

Surafel M. Lakew, Alina Karakanta, Marcello Federico, Matteo Negri, and Marco Turchi. 2019. Adapting multilingual neural machine translation to unseen languages. In *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong. Association for Computational Linguistics.

Surafel Melaku Lakew, Mauro Cettolo, and Marcello Federico. 2018c. A comparison of transformer and recurrent neural networks on multilingual neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 641–652, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Christopher Lucas and Stefano Manfredi, editors. 2019. *Arabic and contact-induced change*. Number 1 in Contact and Multilingualism. Language Science Press, Berlin.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *arXiv:2006.07264*.

Karima Meftouh, Salima Harrat, Salma Jamoussi, Mourad Abbas, and Kamel Smaili. 2015. Machine translation experiments on PADIC: A parallel Arabic DIalect corpus. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 26–34, Shanghai, China.

Emad Mohamed, Behrang Mohit, and Kemal Oflazer. 2012. Transforming Standard Arabic to colloquial Arabic. In *Proceedings of the 50th Annual Meeting of the*

Association for Computational Linguistics (Volume 2: Short Papers), pages 176–180, Jeju Island, Korea. Association for Computational Linguistics.

Youness Moukafih, Nada Sbihi, Mounir Ghogho, and Kamel Smaili. 2022. Improving machine translation of Arabic dialects through multi-task learning. In *AIxIA 2021 – Advances in Artificial Intelligence*, pages 580–590, Cham. Springer International Publishing.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2021. Investigating code-mixed Modern Standard Arabic-Egyptian to English machine translation. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 56–64, Online. Association for Computational Linguistics.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022. AraT5: Text-to-text transformers for Arabic language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 628–647, Dublin, Ireland. Association for Computational Linguistics.

Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Sergei Nirenburg. 1989. Knowledge-based machine translation. *Machine Translation*, 4(1):5–24.

Jonathan Owens. 2001. Arabic sociolinguistics. *Arabica*, 48(4):419–469.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th*

Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Arfath Pasha, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. MADAMIRA: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1094–1101, Reykjavik, Iceland. European Language Resources Association (ELRA).

Prajit Ramachandran, Peter Liu, and Quoc Le. 2017. Unsupervised pretraining for sequence to sequence learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 383–391, Copenhagen, Denmark. Association for Computational Linguistics.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37.

Robert R. Ratcliffe. 2021. The glottometrics of Arabic: Quantifying linguistic diversity and correlating it with diachronic change. *Language Dynamics and Change*, 11(1):1 – 29.

Devendra Sachan and Graham Neubig. 2018. Parameter sharing methods for multilingual self-attentional translation models. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 261–271, Brussels, Belgium. Association for Computational Linguistics.

Fatiha Sadat, Fatma Mallek, Mohamed Boudabous, Rahma Sellami, and Atefeh Farzindar. 2014. Collaboratively constructed linguistic resources for language variants and their exploitation in NLP application – the case of Tunisian Arabic and the social media. In *Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing*, pages 102–110, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Hassan Sajjad, Ahmed Abdelali, Nadir Durrani, and Fahim Dalvi. 2020. AraBench: Benchmarking dialectal Arabic-English machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5094–5107, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Hassan Sajjad, Kareem Darwish, and Yonatan Belinkov. 2013. Translating dialectal Arabic to English. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–6, Sofia, Bulgaria. Association for Computational Linguistics.

Wael Salloum and Nizar Habash. 2011. Dialectal to Standard Arabic paraphrasing to improve Arabic-English statistical machine translation. In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, pages 10–21, Edinburgh, Scotland. Association for Computational Linguistics.

Wael Salloum and Nizar Habash. 2012. Elissa: A dialectal to Standard Arabic machine translation system. In *Proceedings of COLING 2012: Demonstration Papers*, pages 385–392, Mumbai, India. The COLING 2012 Organizing Committee.

Hassan Sawaf. 2010. Arabic dialect handling in hybrid machine translation. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Research Papers*, Denver, Colorado, USA. Association for Machine Translation in the Americas.

Pamela Shapiro and Kevin Duh. 2019. Comparing pipelined and integrated approaches to dialectal Arabic neural machine translation. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 214–222, Ann Arbor, Michigan. Association for Computational Linguistics.

Sonia S'hiri. 2013. Speak arabic please!: Tunisian arabic speakers' linguistic accommodation to middle easterners. In *Language Contact and Language Conflict in Arabic*, pages 167–192. Routledge.

Amel Slim, Ahlem Melouah, Usef Faghihi, and Khouloud Sahib. 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.

Felix Stahlberg. 2020. Neural machine translation: A review. *Journal of Artificial Intelligence Research*, 69:343–418.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA. MIT Press.

Ridouane Tachicart and Karim Bouzoubaa. 2014. A hybrid approach to translate moroccan Arabic dialect. *2014 9th International Conference on Intelligent Systems: Theories and Applications (SITA-14)*, pages 1–5.

Toshiyuki Takezawa, Genichiro Kikui, Masahide Mizushima, and Eiichiro Sumita. 2007. Multilingual spoken language corpus development for communication research. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 12, Number 3, September 2007: Special Issue on Invited Papers from ISCSLP 2006*, pages 303–324.

Xu Tan, Yi Ren, Di He, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2019. Multilingual neural machine translation with knowledge distillation. *CoRR*, abs/1902.10461.

Ahmed Tawfik, Mahitab Emam, Khaled Essam, Robert Nabil, and Hany Hassan. 2019. Morphology-aware word-segmentation in dialectal Arabic adaptation of neural machine translation. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 11–17, Florence, Italy. Association for Computational Linguistics.

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2Tensor for neural

machine translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 193–199, Boston, MA. Association for Machine Translation in the Americas.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Kees Versteegh. 2001. Linguistic contacts between arabic and other languages. *Arabica*, 48(4):470–508.

Kees Versteegh. 2014. *The Arabic Language*. Edinburgh University Press.

Ronald Wardhaugh and Janet M. Fuller. 2015. *An Introduction to Sociolinguistics*, seventh edition edition. Blackwell textbooks in linguistics. John Wiley & Sons West Sussex, England, West Sussex, England.

Gian Wiher, Clara Meister, and Ryan Cotterell. 2022. On Decoding Strategies for Neural Text Generators. *Transactions of the Association for Computational Linguistics*, 10:997–1012.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Omar F. Zaidan and Chris Callison-Burch. 2014. Arabic dialect identification. *Computational Linguistics*, 40(1):171–202.

Jezia Zakraoui, Moutaz Saleh, Somaya Al-Maadeed, and Jihad Mohamed Alja'am. 2021. Arabic machine translation: A survey with challenges and future directions. *IEEE Access*, 9:161445–161468.

Poorya Zaremoodi, Wray Buntine, and Gholamreza Haffari. 2018. Adaptive knowledge sharing in multi-task learning: Improving low-resource neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 656–661, Melbourne, Australia. Association for Computational Linguistics.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan, and Chris Callison-Burch. 2012. Machine translation of Arabic dialects. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 49–59, Montréal, Canada. Association for Computational Linguistics.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tieyan Liu. 2020. Incorporating bert into neural machine translation. In *International Conference on Learning Representations*.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

Appendix

Source \ Target	TUN	RAB	CAI	BEI	DOH
TUN		10.99	11.13	11.12	21.46
RAB	10.10		11.33	13.03	21.79
CAI	10.79	12.16		13.16	24.21
BEI	11.11	12.77	13.74		24.53
DOH	11.01	12.71	13.51	13.62	

Table 32: Results of the Everything-to-Everything model without MSA, a batch size of 32 and beam search decoding strategy

Source \ Target	TUN	RAB	CAI	BEI	DOH	MSA BLEU	MSA AL-BLEU
TUN		11.05	10.29	11.81	21.99	14.90	9.87
RAB	10.57		10.70	12.60	22.18	15.71	10.58
CAI	10.52	11.86		13.26	25.66	17.62	12.27
BEI	11.03	12.31	12.53		24.74	17.08	11.46
DOH	10.88	12.91	12.50	14.03		17.63	12.51
MSA	10.19	12.01	11.54	13.02	24.15		

Table 33: Results of the Everything-to-Everything model with MSA, a batch size of 32 and beam search decoding strategy

	Evaluation Direction	RAB-TUN	CAI-TUN	BEI-TUN	DOH-TUN
	Reference Score	8.35	10.80	13.62	12.40
Combination Source	BEI+CAI		**11.23**	11.15	
	BEI+DOH			9.84	10.20
	BEI+RAB	**10.19**		10.96	
	CAI+DOH		**11.24**		11.73
	CAI+RAB	**9.67**	9.76		
	DOH+RAB	**10.10**			10.22
	BEI+CAI+DOH		10.28	10.54	10.97
	BEI+CAI+RAB	**8.87**	9.19	9.06	
	BEI+DOH+RAB	8.18		8.02	8.13
	CAI+DOH+RAB	**9.64**	9.24		10.50
	BEI+CAI+DOH+RAB	**9.88**	10.03	9.42	9.71

Table 34: Results of the augmentation models for TUN as a target language with a batch size of 32 and beam search decoding strategy

	Evaluation Direction	TUN-RAB	CAI-RAB	BEI-RAB	DOH-RAB
	Reference Score	10.75	10.54	13.39	13.53
Combination Source	BEI+CAI		**12.06**	11.29	
	BEI+DOH			12.14	12.71
	BEI+TUN	**10.85**		11.44	
	CAI+DOH		**11.71**		11.86
	CAI+TUN	**11.14**	**12.26**		
	DOH+TUN	**11.25**			12.35
	BEI+CAI+DOH		**11.40**	10.79	12.51
	BEI+CAI+TUN	10.52	**11.51**	11.51	
	BEI+DOH+TUN	**11.07**		12.06	12.84
	CAI+DOH+TUN	**11.25**	**11.59**		12.72
	BEI+CAI+DOH+TUN	9.55	10.34	10.41	10.24

Table 35: Results of the augmentation models for RAB as a target language with a batch size of 32 and beam search decoding strategy

	Evaluation Direction	TUN-CAI	RAB-CAI	BEI-CAI	DOH-CAI
	Reference Score	11.34	11.70	17.23	14.52
Combination Source	BEI+DOH			14.79	14.37
	BEI+RAB		11.14	14.92	
	BEI+TUN	11.21		13.21	
	DOH+RAB		11.23		13.69
	DOH+TUN	10.66			14.00
	RAB+TUN	9.80	9.49		
	BEI+DOH+RAB		9.80	13.78	13.37
	BEI+DOH+TUN	10.1		13.51	12.67
	BEI+RAB+TUN	8.45	8.39	10.48	
	DOH+RAB+TUN	10.65	10.56		13.03
	BEI+DOH+RAB+TUN	9.26	8.68	11.42	11.17

Table 36: Results of the augmentation models for CAI as a target language with a batch size of 32 and beam search decoding strategy

	Evaluation Direction	TUN-BEI	RAB-BEI	CAI-BEI	DOH-BEI
	Reference Score	14.96	14.98	18.84	16.54
Combination Source	CAI+DOH			15.32	14.11
	CAI+RAB		13.43	15.73	
	CAI+TUN	13.34		16.09	
	DOH+RAB		11.98		13.09
	DOH+TUN	11.80			12.75
	RAB+TUN	11.53	12.13		
	CAI+DOH+RAB		11.98	13.25	13.29
	CAI+DOH+TUN	11.23		13.89	13.88
	CAI+RAB+TUN	11.21	11.38	12.80	
	DOH+RAB+TUN	10.87	11.86		13.42
	CAI+DOH+RAB+TUN	11.00	11.19	13.28	13.74

Table 37: Results of the augmentation models for BEI as a target language with a batch size of 32 and beam search decoding strategy

	Evaluation Direction	TUN-DOH	RAB-DOH	CAI-DOH	BEI-DOH
	Reference Score	22.39	24.06	29.01	29.66
Combination Source	BEI+CAI			26.93	26.41
	BEI+RAB		22.33		24.60
	BEI+TUN	21.69			24.75
	CAI+RAB		22.48	25.56	
	CAI+TUN	22.26		25.59	
	RAB+TUN	21.10	22.21		
	BEI+CAI+RAB		20.31	24.22	24.16
	BEI+CAI+TUN	20.37		23.72	23.05
	BEI+RAB+TUN	20.33	21.06		23.94
	CAI+RAB+TUN	18.67	18.10	20.55	
	BEI+CAI+RAB+TUN	18.93	19.77	23.25	21.98

Table 38: Results of the augmentation models for DOH as a target language with a batch size of 32 and beam search decoding strategy

		Target				
		TUN	RAB	CAI	BEI	DOH
Source	TUN+MSA		11.13	10.80	12.00	22.17
	RAB+MSA	9.31		9.65	12.99	21.80
	CAI+MSA	10.87	11.91		13.72	26.95
	BEI+MSA	11.01	12.15	14.75		24.81
	DOH+MSA	10.40	11.28	13.76	14.44	

Table 39: Results of adding MSA to the source language for each direction with a batch size of 32 and beam search decoding strategy

Source \ Target	TUN	RAB	CAI	BEI	DOH
TUN		6.70	7.09	7.65	14.89
RAB	6.40		7.19	8.08	15.62
CAI	6.17	7.00		8.65	17.49
BEI	6.34	7.36	8.49		17.05
DOH	5.91	8.12	7.98	8.53	

Table 40: Results of the Everything-to-Everything model without MSA, a batch size of 64 and top-K decoding strategy

Source \ Target	TUN	RAB	CAI	BEI	DOH	MSA BLEU	MSA AL-BLEU
TUN		6.95	6.84	8.35	14.85	9.78	6.13
RAB	6.35		7.30	8.72	15.58	11.11	7.07
CAI	6.28	7.44		9.26	16.83	12.04	7.60
BEI	5.87	7.17	7.97		17.30	11.55	7.53
DOH	6.53	7.55	8.83	9.72		12.10	7.85
MSA	6.13	7.26	7.73	8.77	16.91		

Table 41: Results of the Everything-to-Everything model with MSA, a batch size of 64 and top-K decoding strategy

	Evaluation Direction	RAB-TUN	CAI-TUN	BEI-TUN	DOH-TUN
	Reference Score	9.86	11.65	10.78	11.31
Combination Source	BEI+CAI		7.93	8.10	
	BEI+DOH			6.63	7.24
	BEI+RAB	6.90		7.81	
	CAI+DOH		10.46		10.04
	CAI+RAB	**10.95**	10.90		
	DOH+RAB	8.21			8.17
	BEI+CAI+DOH		7.01	7.29	7.67
	BEI+CAI+RAB	2.89	2.72	3.02	
	BEI+DOH+RAB	6.61		7.02	6.96
	CAI+DOH+RAB	7.51	7.85		7.45
	BEI+CAI+DOH+RAB	5.82	5.86	5.99	6.21

Table 42: Results of the augmentation models for TUN as a target language with a batch size of 64 and top-K decoding strategy

	Evaluation Direction	TUN-RAB	CAI-RAB	BEI-RAB	DOH-RAB
	Reference Score	10.35	12.26	12.68	13.06
Combination Source	BEI+CAI		5.72	5.78	
	BEI+DOH			7.75	8.78
	BEI+TUN	**11.09**		11.92	
	CAI+DOH		7.83		8.93
	CAI+TUN	6.75	7.66		
	DOH+TUN	7.31			8.57
	BEI+CAI+DOH		8.72	8.72	9.00
	BEI+CAI+TUN	7.68	8.20	8.14	
	BEI+DOH+TUN	7.38		7.88	8.37
	CAI+DOH+TUN	7.74	8.08		8.64
	BEI+CAI+DOH+TUN	7.06	7.09	7.67	8.15

Table 43: Results of the augmentation models for RAB as a target language with a batch size of 64 and top-K decoding strategy

	Evaluation Direction	TUN-CAI	RAB-CAI	BEI-CAI	DOH-CAI
	Reference Score	12.68	7.83	13.93	11.41
Combination Source	BEI+DOH			10.21	9.96
	BEI+RAB		7.07	8.68	
	BEI+TUN	7.56		9.72	
	DOH+RAB		6.48		9.01
	DOH+TUN	7.62			9.95
	RAB+TUN	7.07	6.07		
	BEI+DOH+RAB		7.28	9.55	8.82
	BEI+DOH+TUN	6.00		7.48	8.53
	BEI+RAB+TUN	6.71	5.97	8.11	
	DOH+RAB+TUN	6.87	6.53		9.38
	BEI+DOH+RAB+TUN	6.04	5.81	7.33	6.88

Table 44: Results of the augmentation models for CAI as a target language with a batch size of 64 and top-K decoding strategy

	Evaluation Direction	TUN-BEI	RAB-BEI	CAI-BEI	DOH-BEI
	Reference Score	13.76	14.68	18.70	13.34
Combination Source	CAI+DOH			11.53	11.12
	CAI+RAB		10.30	12.76	
	CAI+TUN	11.25		13.66	
	DOH+RAB		11.27		11.44
	DOH+TUN	7.74			8.82
	RAB+TUN	9.61	9.89		
	CAI+DOH+RAB		8.82	10.07	10.61
	CAI+DOH+TUN	8.50		9.05	9.73
	CAI+RAB+TUN	7.36	8.20	8.96	
	DOH+RAB+TUN	7.60	7.57		8.30
	CAI+DOH+RAB+TUN	7.16	7.18	8.31	8.49

Table 45: Results of the augmentation models for BEI as a target language with a batch size of 64 and top-K decoding strategy

	Evaluation Direction	TUN-DOH	RAB-DOH	CAI-DOH	BEI-DOH
	Reference Score	22.46	24.96	28.56	27.71
Combination Source	BEI+CAI			27.42	27.16
	BEI+RAB		21.30		23.76
	BEI+TUN	21.72			26.26
	CAI+RAB		23.51	25.69	
	CAI+TUN	**22.61**		27.01	
	RAB+TUN	20.82	21.03		
	BEI+CAI+RAB		16.49	19.83	19.59
	BEI+CAI+TUN	20.70		24.01	24.17
	BEI+RAB+TUN	15.72	15.66		17.65
	CAI+RAB+TUN	15.82	16.62	19.96	
	BEI+CAI+RAB+TUN	15.66	16.26	17.75	16.58

Table 46: Results of the augmentation models for DOH as a target language with a batch size of 64 and top-K decoding strategy

		Target				
		TUN	RAB	CAI	BEI	DOH
Source	TUN+MSA		7.51	6.32	12.49	22.29
	RAB+MSA	10.67		7.53	10.06	17.06
	CAI+MSA	6.37	7.78		12.18	26.90
	BEI+MSA	6.84	8.94	9.72		27.03
	DOH+MSA	7.89	9.71	9.43	14.61	

Table 47: Results of adding MSA to the source language for each direction with a batch size of 64 and top-K decoding strategy

Source \ Target	TUN	RAB	CAI	BEI	DOH
TUN		6.64	7.80	8.06	16.49
RAB	6.84		8.16	9.17	16.95
CAI	6.80	7.75		10.52	18.66
BEI	7.25	8.15	9.01		18.74
DOH	7.68	8.54	8.68	9.23	

Table 48: Results of the Everything-to-Everything model without MSA, a batch size of 64 and top-p decoding strategy

Source \ Target	TUN	RAB	CAI	BEI	DOH	MSA BLEU	MSA AL-BLEU
TUN		8.13	7.93	9.02	16.48	11.41	6.99
RAB	6.90		7.49	8.74	16.65	11.42	7.19
CAI	6.44	7.74		9.57	18.69	12.91	8.42
BEI	6.48	8.48	8.48		19.10	12.41	8.09
DOH	7.22	8.82	9.07	10.72		13.48	9.04
MSA	6.26	8.25	9.02	9.58	19.27		

Table 49: Results of the Everything-to-Everything model with MSA, a batch size of 64 and top-p decoding strategy

	Evaluation Direction	RAB-TUN	CAI-TUN	BEI-TUN	DOH-TUN
	Reference Score	10.45	12.23	11.27	11.60
Combination Source	BEI+CAI		8.58	8.66	
	BEI+DOH			7.54	7.61
	BEI+RAB	7.45		7.65	
	CAI+DOH		10.69		10.64
	CAI+RAB	**11.00**	11.21		
	DOH+RAB	8.43			8.94
	BEI+CAI+DOH		7.82	7.26	8.18
	BEI+CAI+RAB	3.30	2.95	3.08	
	BEI+DOH+RAB	7.22		7.74	7.65
	CAI+DOH+RAB	7.85	8.71		8.43
	BEI+CAI+DOH+RAB	6.05	6.73	6.37	6.35

Table 50: Results of the augmentation models for TUN as a target language with a batch size of 64 and top-p decoding strategy

	Evaluation Direction	TUN-RAB	CAI-RAB	BEI-RAB	DOH-RAB
	Reference Score	10.83	12.74	12.99	13.28
Combination Source	BEI+CAI		6.41	6.21	
	BEI+DOH			8.87	8.74
	BEI+TUN	**11.51**		12.13	
	CAI+DOH		8.31		9.39
	CAI+TUN	7.34	8.19		
	DOH+TUN	8.20			9.54
	BEI+CAI+DOH		8.95	9.29	9.51
	BEI+CAI+TUN	8.43	8.38	8.43	
	BEI+DOH+TUN	8.01		8.31	9.24
	CAI+DOH+TUN	8.74	8.80		9.37
	BEI+CAI+DOH+TUN	7.20	7.46	8.02	8.20

Table 51: Results of the augmentation models for RAB as a target language with a batch size of 64 and top-p decoding strategy

	Evaluation Direction	TUN-CAI	RAB-CAI	BEI-CAI	DOH-CAI
	Reference Score	13.16	8.72	14.91	12.01
Combination Source	BEI+DOH			11.01	10.87
	BEI+RAB		7.40	9.66	
	BEI+TUN	8.22		11.27	
	DOH+RAB		6.92		9.76
	DOH+TUN	8.59			10.66
	RAB+TUN	7.35	6.98		
	BEI+DOH+RAB		7.70	10.42	9.89
	BEI+DOH+TUN	6.08		7.94	9.09
	BEI+RAB+TUN	7.27	7.03	8.88	
	DOH+RAB+TUN	7.46	7.63		10.03
	BEI+DOH+RAB+TUN	5.85	5.92	6.99	7.21

Table 52: Results of the augmentation models for CAI as a target language with a batch size of 64 and top-p decoding strategy

	Evaluation Direction	TUN-BEI	RAB-BEI	CAI-BEI	DOH-BEI
	Reference Score	14.22	14.83	19.09	14.24
Combination Source	CAI+DOH			12.64	12.14
	CAI+RAB		11.41	13.51	
	CAI+TUN	11.92		14.24	
	DOH+RAB		11.25		11.26
	DOH+TUN	7.51			10.12
	RAB+TUN	10.16	10.85		
	CAI+DOH+RAB		9.30	9.86	9.96
	CAI+DOH+TUN	8.68		10.56	10.88
	CAI+RAB+TUN	7.94	8.71	8.67	
	DOH+RAB+TUN	7.80	8.38		9.30
	CAI+DOH+RAB+TUN	7.91	7.72	9.25	9.60

Table 53: Results of the augmentation models for BEI as a target language with a batch size of 64 and top-p decoding strategy

	Evaluation Direction	TUN-DOH	RAB-DOH	CAI-DOH	BEI-DOH
	Reference Score	22.84	25.48	29.19	28.77
Combination Source	BEI+CAI			27.33	27.59
	BEI+RAB		21.74		24.49
	BEI+TUN	21.91			26.30
	CAI+RAB		23.83	26.63	
	CAI+TUN	**23.12**		27.07	
	RAB+TUN	20.69	22.19		
	BEI+CAI+RAB		18.27	21.40	21.38
	BEI+CAI+TUN	21.24		24.13	23.41
	BEI+RAB+TUN	16.34	16.89		19.24
	CAI+RAB+TUN	17.33	17.55	20.98	
	BEI+CAI+RAB+TUN	16.19	17.35	18.95	18.70

Table 54: Results of the augmentation models for DOH as a target language with a batch size of 64 and top-p decoding strategy

		Target				
		TUN	RAB	CAI	BEI	DOH
Source	TUN+MSA		8.02	7.43	12.70	22.45
	RAB+MSA	10.83		8.05	10.70	17.91
	CAI+MSA	6.69	8.49		12.45	27.85
	BEI+MSA	7.06	9.87	10.53		27.18
	DOH+MSA	8.05	10.03	10.51	14.47	

Table 55: Results of adding MSA to the source language for each direction with a batch size of 64 and top-p decoding strategy