NADI 2023: The Fourth Nuanced Arabic Dialect Identification Shared Task

Muhammad Abdul-Mageed,[λ,ξ] **AbdelRahim Elmadany,**[λ] **Chiyu Zhang,**[λ]
El Moatez Billah Nagoudi,[λ] **Houda Bouamor,**[δ] **Nizar Habash**[μ]
[λ]The University of British Columbia, Vancouver, Canada; [ξ]MBZUAI, Abu Dhabi, UAE
[δ]Carnegie Mellon University in Qatar, Qatar
[μ]New York University Abu Dhabi, UAE
{muhammad.mageed@,a.elmadany@,chiyuzh@mail,moatez.nagoudi@}.ubc.ca
hbouamor@cmu.edu nizar.habash@nyu.edu

Abstract

We describe the findings of the fourth Nuanced Arabic Dialect Identification Shared Task (NADI 2023). The objective of NADI is to help advance state-of-the-art Arabic NLP by creating opportunities for teams of researchers to collaboratively compete under standardized conditions. It does so with a focus on Arabic dialects, offering novel datasets and defining subtasks that allow for meaningful comparisons between different approaches. NADI 2023 targeted both dialect identification (Subtask 1) and dialect-to-MSA machine translation (Subtask 2 and Subtask 3). A total of 58 unique teams registered for the shared task, of whom 18 teams have participated (with 76 valid submissions during test phase). Among these, 16 teams participated in Subtask 1, 5 participated in Subtask 2, and 3 participated in Subtask 3. The winning teams achieved 87.27 F_1 on Subtask 1, 14.76 Bleu in Subtask 2, and 21.10 Bleu in Subtask 3, respectively. Results show that all three subtasks remain challenging, thereby motivating future work in this area. We describe the methods employed by the participating teams and briefly offer an outlook for NADI.

1 Introduction

Arabic is a term usually used to collectively refer to a host of languages and language varieties, rather than a single language. While most of these languages and varieties are similar to one another, there can be significant differences between some of them. For example, Egyptian Arabic and Moroccan Arabic are not mutually intelligible. Arabic can also be classified into three broad categories, classical, modern standard, and dialectal. Of these, *Classical Arabic (CA)* represents the variety used in old forms of literature such as poetry and the Qur'an, the Holy Book of Islam. Association with religion and literary expression endows CA with prestige, and it continues to be used to date side by side with other varieties. *Modern Standard Arabic (MSA)* (Badawi, 1973; Abdul-Mageed et al., 2020b) is a modern-day variety that is more familiar to native speakers and is usually employed by pan-Arab media organizations, government, and in education. The third category, *Dialectal Arabic (DA)*, is itself a superclass that is collectively assigned to a host of varieties that are sometimes defined regionally (e.g., Gulf, Levantine, Nile Basin, and North African (Habash, 2010; Abdul-Mageed, 2015)), but are increasingly recognized at the more nuanced levels of country or even sub-country (Bouamor et al., 2018; Abdul-Mageed et al., 2020b)). NLP treatment of Arabic dialects has thus far focused more on dialect identification (Abdul-Mageed et al., 2020b; Bouamor et al., 2019; Darwish et al., 2018), machine translation (MT) (Zbib et al., 2012), morphosyntax (Obeid et al., 2020).

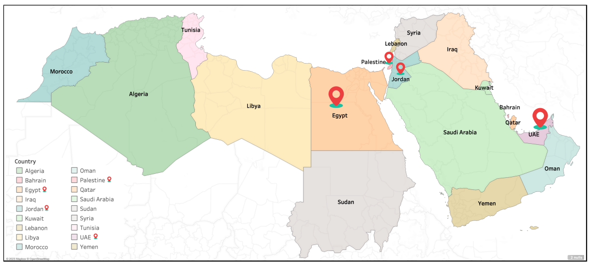

Figure 1: A map of the Arab World showing the 18 countries in the *Subtask 1* dataset and the 4 countries in the *Subtask 2* and *Subtask 3* datasets. Each country is coded in a color different from neighboring countries. Subtasks 2 and 3 countries are coded as red pins.

Dialect identification is the task of automatically detecting the source variety of a given text or speech segment, and is the main focus of the current work where we introduce the findings and results of the fourth Nuanced Arabic Dialect Identification Shared Task (NADI 2024). The main objective of NADI is to encourage research on Ara-

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 600–613
December 7, 2023

bic dialect processing by offering datasets and facilitating diverse modeling opportunities under a common evaluation setup. The first instance of the shared task, NADI 2020 (Abdul-Mageed et al., 2020a), focused on province-level dialects. NADI 2021 (Abdul-Mageed et al., 2021b), the second iteration of NADI, focused on distinguishing both MSA and DA according to their geographical origin at the country level. The third instance, NADI 2022 (Abdul-Mageed et al., 2022), investigated both Arabic dialect identification and dialectal sentiment analysis. NADI 2023, the current edition, continues this tradition of extending to tasks beyond dialect identification. Namely, we propose new subtasks focused at machine translation from Arabic dialects into MSA.

More concretely, NADI 2023 shared task is comprised of three subtasks: **Subtask 1** on dialect identification, while **Subtask 2** and **Subtask 3** are on dialect MT. The difference between Subtask 2 and Subtask 3 is that the former is a *closed track* where participants are allowed to use only our provided training data, whereas the latter is *open track* and so allows participants to train their systems on any additional datasets so long as these additional training datasets are public at the time of submission. While we invited participation in any of the three subtasks, we encouraged teams to submit systems to *all* subtasks. By offering three subtasks, our hope was to receive systems that exploit different methods and architectures. Many of the submitted systems investigated diverse approaches, thus fulfilling our objective. A total of 58 unique teams registered for NADI 2023. Of these, 18 unique teams actually made submissions to our leaderboard (n=76 valid submissions during test phase). We received 14 papers from 14 teams, of which we accepted 13 for publication. Results from participating teams show that both dialect identification at the country level and dialectal MT remain challenging even to complex neural methods. These findings clearly motivate future work on all tasks.

The rest of the paper is organized as follows: Section 2 provides a brief overview of Arabic dialect identification and sentiment analysis. We describe the two subtasks and NADI 2023 restrictions in Section 3. Section 4 introduces shared task datasets and evaluation setup. We present participating teams and shared task results and provide a high-level description of submitted systems in Section 5. We conclude in Section 6.

2 Literature Review

2.1 Arabic Dialects

As stated earlier, Arabic can be broadly categorized into CA, DA, and MSA. While CA and MSA have been examined extensively (Harrell, 1962; Cowell, 1964; Badawi, 1973; Brustad, 2000; Holes, 2004), DA became the center of attention only relatively recently. A significant challenge in studying DA has been the scarcity of resources. This prompted researchers to create new DA datasets, usually targeting a limited number of specific regions or countries (Gadalla et al., 1997; Diab et al., 2010; Al-Sabbagh and Girju, 2012; Sadat et al., 2014; Harrat et al., 2014; Jarrar et al., 2016; Khalifa et al., 2016; Al-Twairesh et al., 2018; Alsarsour et al., 2018; Kwaik et al., 2018; El-Haj, 2020). This was followed by several works that introduced multi-dialectal datasets and models for region-level dialect identification (Zaidan and Callison-Burch, 2011; Elfardy et al., 2014; Bouamor et al., 2014; Meftouh et al., 2015). The initial Arabic dialect identification shared tasks were part of the VarDial workshop series, primarily utilizing transcriptions of speech broadcasts (Malmasi et al., 2016). This was followed by creation of the Multi-Arabic Dialects Application and Resources project (MADAR), which provided finer-grained data and a lexicon (Bouamor et al., 2018). Although MADAR's dataset was used for identifying dialects at both the country and city levels (Salameh et al., 2018; Obeid et al., 2019), the fact that it is commissioned, rather than naturally occurring, makes it not be optimal for dialect identification especially in contexts such as social media.

Subsequently, larger datasets that cover between 10 to 21 countries were introduced (Mubarak and Darwish, 2014; Abdul-Mageed et al., 2018; Zaghouani and Charfi, 2018; Abdelali et al., 2021; Issa et al., 2021; Baimukan et al., 2022; Althobaiti, 2022). The majority of these datasets are compiled from social media posts, especially Twitter. Other works collect data at a more granular level. For instance, Abdul-Mageed et al. (2020b) introduces a Twitter dataset along with several models to identify variations in Arabic dialects at the country, province, and city levels. Althobaiti (2020) provides an overview of computational work on Arabic dialects. More recently, benchmarks such as ORCA (Elmadany et al., 2023) and DOLPHIN (Nagoudi et al., 2023) boast dialectal coverage. The NADI shared task continues to lead efforts on providing datasets and common evaluation settings for identifying Arabic dialects (Abdul-Mageed et al., 2020a, 2021b, 2022).

2.2 Machine Translation of Arabic Dialects

Several studies focus on machine translation of Arabic dialects. For example, Zbib et al. (2012) demonstrate effects of using both MSA and DA data on performance of Dialect/MSA to English MT. Sajjad et al. (2013) employs MSA as an intermediary language for translating Arabic dialects into English. Salloum et al. (2014) examine the impact of sentence-level dialect identification and various linguistic features on Dialect/MSA to English translation. Guellil et al. (2017) propose a neural system for translating Algerian Arabic written in Arabizi and Arabic script into MSA, while Baniata et al. (2018) introduce a system that translates Levantine (Jordanian, Syrian, Palestinian) and Maghrebi (Algerian, Moroccan, Tunisian) into MSA. Sajjad et al. (2020) propose an evaluation benchmark for Dialectal Arabic to English MT, along with several NMT systems using different training setups such as fine-tuning, data augmentation, and back-translation. Farhan et al. (2020) offer an unsupervised dialectal system where the source dialect (zero-shot) is not represented in training data. Nagoudi et al. (2021) propose a transformer-based MT system for translating from code-mixed MSA and Egyptian Arabic into English. More recently, Kadaoui et al. (2023) present a comprehensive evaluation of large language models (LLMs), including Bard and ChatGPT, on the machine translation of ten Arabic varieties. To the best of our knowledge, our work is the first shared task to enable investigating MT in $four$ Arabic dialects, namely *Egyptian*, *Emirati*, *Jordanian*, and *Palestinian*. For our MT subtasks, we also annotate and release a novel dataset and facilitate comparisons in a standardized experimental setting.

2.3 Previous NADI Shared Tasks

NADI 2020 The first NADI shared task, (Abdul-Mageed et al., 2020a) was co-located with the fifth Arabic Natural Language Processing Workshop (WANLP 2020) (Zitouni et al., 2020). NADI 2020 targeted both country- and province-level dialects. It covered a total of 100 provinces from 21 Arab countries, with data collected from Twitter. It was the first shared task to target naturally occurring fine-grained dialectal text at the sub-country level.

NADI 2021 The second edition of the shared task (Abdul-Mageed et al., 2021b) was co-located with WANLP 2021 (Habash et al., 2021). It targeted the same 21 Arab countries and 100 corresponding provinces as NADI 2020, also exploiting Twitter data. NADI 2021 improved over NADI 2020 in that non-Arabic data were removed. In addition, NADI-2021 teased apart the data into MSA and DA and focused on classifying MSA and DA tweets into the countries and provinces from which they are collected. As such, NADI 2021 had four subtasks: MSA-country, DA-country, MSA-province, and DA-province.

NADI 2022 The third edition of the shared task (Abdul-Mageed et al., 2022) was co-located with WANLP 2021.[1] It focused on studying Arabic dialects at the country level as well as dialectal sentiment (i.e., sentiment analysis of data tagged with dialect labels). We discuss NADI 2023 in more detail in the next section.

3 Task Description

In NADI-2023, we place our emphasis on two NLP tasks, both crucial to processing of dialectal Arabic. Dialect identification remains an important step in any pipeline for processing dialects, for which reason NADI-2023 **Subtask 1** maintains the focus on identification of Arabic dialects. In particular, Subtask 1 targets dialect at the country level. Another important NLP task that has not particularly witnessed accelerated progress over the past few years is machine translation of Arabic dialects. For this reason, we take as our second focal point MT of dialects through **Subtask 2** and **Subtask 3**. We now describe each subtask in detail.

3.1 Subtask 1: Dialect Identification

Dialect identification has consistently been central to the NADI shared task over the years (2020a; 2021b; 2022). In NADI-2023, we continue to focus on dialect identification through Subtask 1.

Data For this purpose, we provide a new Twitter dataset (i.e., TWT-2023), encompassing 18 distinct dialects, totaling 23.4K tweets. We also provide access to additional datasets for training. These are NADI-2020 (Abdul-Mageed et al., 2020a), NADI-2021 (Abdul-Mageed et al., 2021b), and MADAR (Bouamor et al., 2018) *training* splits. We

[1] https://sites.google.com/view/wanlp2021

Country	Content
Algeria	مهم جميع واحد انا ثقيل عليه يبلوكيني و ميضليش ينفخلي فيهم بهدره تع تقيي يكونو معاك هاك يشوفوك وليت هاك يولو هاك و يكتلوك بهاك سورتو لا صابوك هاك يولو عليك غي هاك هوما هاك
Iraq	بس ما اعتقد لان هيج وقت كل الاكلات حلوه بو خوشي بهس تهت فيت زيانه ته سبي بيت اوب ههمي دلي هه
Jordan	من لما صحيت حاسه في اشي غريب اشي ناقص لحتي امي سالتني شربتي قهوه اليوم ها ها ها هه ممكن والله كل اشي بهالبلد ممكن يصير
Saudi Arabia	ذيك الايام خشم بخاري بغيبوبه وش اسمه يمدي حرمته تدعي علي هه
Sudan	اكتر من ١٠ سنين شغاله يا استنكرت موضوع يا ذرفت دموع اسه جابت ليها قطع رؤوس #محن_الكيزان في #السودان اي زول في العلاقه عاوز اعمل تقيل بقتل العلاقه والله يعني الطرفين ساكتين لو في طرف م عاوز التاني اصرفها ليهو

Table 1: Random examples from NADI-2023 Subtask-1 training dataset spanning five different countries.

Dialect	Source (Dialect)	Target (MSA)
Egyptain	ايوا حضرتك ده حق ابويا، حقي أنا بقى؟	إي حضرتك هذا حق أبي ، أين حقي أنا إذن ؟
	حتى ابوك نفسه مش حيقدر يشفعلك عنده.	وأبوك نفسه لن يقدر يشفع لك عنده .
	يا عم احنا مش ناقصينك الله يسترك عيب كده اختشي فوق فوق.	يا عم ! نحن فاض بنا الكيل ، الله يسترك ، هذا عيب ! استحِ ! أفِق ! أفِق!
Emirati	زين خبروني شو السالفة ؟	إذا أخبروني، ما القصة؟
	مابا حد يدري في الفريج، إن بو محمد انسرق	لا أريد أن يعلم أحدًا في الحي، أن أبو محمد انسرق
	انزين، بغينا اثنين زنجبيل حار	حسنا، نريد اثنين من الزنجبيل الحار
Jordanian	بس ما بعرف شو، مش عارف شو صارلي	لكن لا أعلم ماذا، لا أعلم ما حدث لي
	كله منك انت السبب ليش ماخليتني ماسك بخناقه	كله بسببك أنت السبب لماذا لم تدعني ممسكاً بعنقه
	بكير من عمرك	بارك الله في عمرك
Palestinian	طيب و هاي الملاعب وين؟	حسناً، وهذه الملاعب أين؟
	شو يا أبو ناصر سمعت صاير تهدد في القتل	ما هذا يا أبو ناصر، سمعت أنك أصبحت تهدد بالقتل
	شاطر، إلك عندي باكيت حلقوم	حصيف، سأكافئك بعلبة كاملة من حلوى الحلقوم

Table 2: Random examples from MT-2023-DEV dataset spanning the four covered dialects.

refer to these datasets as NADI-2020-TWT, NADI-2021-TWT, and MADAR-2018, respectively. We provide further details about these datasets in Section 4.1. Table 1 shows examples from tweets in our NADI-2023 dataset for five countries.

Restrictions It is essential to note that Subtask 1 operates under a ***closed-track*** policy where participants are allowed to use for system training *only* datasets we provide. That is, no external data sources can be used for training purposes in this subtask.

3.2 Subtasks 2 and 3: Machine Translation

In this competition version, we introduce a new theme to NADI centered around machine translation from *four* Arabic dialects to Modern Standard Arabic (MSA) at the sentence level. We present two versions of this competition, one is a closed track (Subtask 2), and the other is an open track (Subtask 3).

Dev and Test Data For both Subtask 2 and Subtask 3, we manually curate new development and test datasets that each cover *four* Arabic dialects: *Egyptian*, *Emirati*, *Jordanian*, and *Palestinian*. We refer to these new datasets as MT-2023-DEV and MT-2023-TEST, respectively. MT-2023-DEV comprises 400 sentences, with 100 sentences representing each of the four dialects; whereas MT-2023-TEST has a total of 2,000 sentences, 500 from each dialect. Table 2 shows example sentences from MT-2023-DEV for each of the four countries. During the competition, we intentionally kept the source domain of these datasets undisclosed. Since we typically keep a live leaderboard for post-competition evaluation, we will not disclose the MT-2023* data domain.

Restrictions For the MT theme, restrictions on use of training datasets depend on the type of track. We offer two tracks, one closed and another open each with its own subtask. We introduce these subtasks now, detailing respective track information.

Subtask 2 – Closed-Track Dialect to MSA MT For Subtask 2 training, we restrict to the MADAR parallel dataset (Bouamor et al., 2019). More precisely, participants were allowed to use only the training split of MADAR parallel corpus for this subtask, and report on the development and test sets we provide. This meant that use of MADAR development and test datasets was not allowed for Subtask 2.

Subtask 3 – Open-Track Dialect to MSA MT For Subtask 3 training, participants were allowed to train their systems on any additional datasets of their choice so long as these additional training datasets are public at the time of submission. For example, participants were allowed to manually create new parallel datasets. For transparency and wider community benefits, we required researchers participating in the open track subtask to submit the datasets they create along with their Test set submissions.

Country	NADI-2020	NADI-2021	MADAR-18
Algeria	$1,491$	$1,809$	$1,600$
Bahrain	210	215	–
Egypt	$4,473$	$4,283$	$4,800$
Iraq	$2,556$	$2,729$	$4,800$
Jordan	426	429	$3,200$
Kuwait	420	429	–
Lebanon	639	644	$1,600$
Libya	$1,070$	$1,286$	$3,200$
Morocco	$1,070$	858	$3,200$
Oman	$1,098$	$1,501$	$1,600$
Palestine	420	428	$1,600$
Qatar	234	215	$1,600$
Saudi Arabia	$2,312$	$2,140$	$3,200$
Sudan	210	215	$1,600$
Syria	$1,070$	$1,287$	$3,200$
Tunisia	750	859	$3,200$
UAE	$1,070$	642	–
Yemen	851	429	$1,600$
Total	$\mathbf{20,370}$	$\mathbf{20,398}$	$\mathbf{40,000}$

Table 3: Distribution of Subtask-1 additional training data. For NADI-2023, we also distribute a total of $18,000$ tweets for Train, $1,800$ for Dev, and $3,600$ for Test (with $1,000$, 100, and 200 from each country for 18 countries listed in the table for Train, Dev, and Test, respectively). For Subtask 2 and Subtask-3, we extract MADAR-4-MT from Egyptian, Emirati, Jordanian, and Palestinian data in MADAR-18 (see Section 4).

4 Shared Task Datasets and Evaluation

In this section, we describe the datasets we make available to participants, introduce the chosen evaluation metrics, and outline the clear instructions we provided for the submission process.

4.1 Datasets

- **TWT-2023**: Abdul-Mageed et al. (2020b) introduce a vast dataset comprising $\sim$6B tweets from 2.7M users. They systematically extract tweets that contain geographic information and subsequently embark on a manual annotation process for each user, classifying their location at the city, state, and country levels. This effort results in the identification of $\sim$ 500M tweets coming from 233K users spread across 319 cities within 21 Arab countries. For Subtask 1, we randomly select from this data $1,000$ training, 100 development, and 200 testing tweets for each of the 18 covered countries. In total, this amounts to $23,400$ tweets that we refer to as TWT-2023. We split TWT-2023 into Train (18K), Dev (1.8K), and Test (3.6K).

- **NADI-202X-TWT**. We also distribute **NADI-2020-TWT** and **NADI-2021-TWT** datasets. These datasets are similarly collected from Twitter. For both of them, we use the Twitter API to crawl data from 21 Arab countries for a period of 10 months (Jan. to Oct., 2019). For each case, we label tweets from each user with the country from which they posted for the whole of the 10 months period, thus exploiting *consistent posting location* as a proxy for *dialect labels*. We use the same training splits as both NADI-2020 and NADI-2021, but only include data that cover the 18 Arab countries we target in the current 2023 edition. It is also noteworthy that we do not provide the NADI-2022 training dataset since it is identical to the training set used in NADI 2021.

- **MADAR-18**: The MADAR corpus is a collection of parallel sentences encompassing the dialects of 25 cities from across the Arab world, along with English, French, and MSA. Since this dataset does not originally have country-level labels, we map the 25 cities to their respective countries. As a result, we acquire a customized version of MADAR that we refer to as MADAR-18. We offer the dialectal side of MADAR-18 for optional use for training systems for Subtask-1.

- **MADAR-4-MT**: We extract parallel dialectal-to-MSA data of four dialects from MADAR-18 for training MT systems for Subtask-2 and Subtask-3. The four pairs involve Egyptian, Emirati, Jordanian, and Palestinian at the dialectal side.

Table 3 present the statistics and characteristics of NADI-2023's Subtask-1 training, development, and test datasets, along with the distribution of our additional resources, i.e, NADI-2020-TWT, NADI-2021-TWT, and MADAR-18.[2]

4.2 Evaluation Metrics

The official evaluation metric for Subtask-1 is the macro-averaged F_1 score. In addition to this metric, we also report system performance in terms of `Precision`, `Recall`, and `Accuracy` for submissions to this Subtask 1. For both Subtask 2 and Subtask 3, we use the `Bleu` score as the official metric. The `Bleu` score is computed separately for each of the four dialects (i.e., Egyptian, Emirati, Jordanian, and Palestinian). We then use the average of these individual Bleu scores to rank the submitted systems for Subtask 2 and Subtask 3.

4.3 Submission Roles

We allowed participant teams to submit up to *five* runs for each test set, for each of the three subtasks. In each case, we only retain the submission with the highest score for each team. While the official results were exclusively based on a blind test set, we also requested participants to include their results on the development datasets (Dev) in their papers.

To facilitate the evaluation of participant systems, we established a CodaLab competition for scoring each subtask (i.e., a total of three Codalabs).[3] Similar to previous NADI editions, we are keeping the CodaLab for each subtask active even after official competition has concluded. This is to encourage researchers interested in training models and assessing systems using the shared task's blind test sets. Consequently, we will not disclose the labels for the test sets of any of the subtasks.

5 Shared Task Teams & Results

5.1 Participating Teams

We received a total of 58 unique team registrations. At the testing phase, a total of 76 valid entries were submitted by 18 unique teams. The breakdown across the subtasks is as follows: 49 submission for Subtask 1 from 16 teams, 16 submissions for Subtask 2 from 5 teams, and 11 submissions for Subtask 3 from 3 teams. Table 4 lists the 18 teams. A total of 14 teams submitted 14 description papers from which we accepted 13 papers for publication. Accepted papers are cited in Table 4.

5.2 Baselines

For comparison, we provide three baselines for each of the three subtasks. For **Subtask 1**, we finetune $\text{MARBERT}_{\text{v2}}$ (Abdul-Mageed et al., 2021a), $\text{AraBERT}_{\text{twitter}}$ (Antoun et al., 2021), and $\text{CAMeLBERT}_{\text{da}}$ (Obeid et al., 2020), on `TWT-2023` training data (see Section 3.1). For **Subtask 2** and **3**,[4] we finetune AraT5_{v2} (Nagoudi et al., 2022), mT5 (Xue et al., 2020), and AraBART (Eddine et al., 2022) on MADAR-4-MT (see Section 3.1). In each subtask, we label these baselines as **Baseline I**, **II**, and **III**, respectively.

For all the baselines in both tasks, we finetune each model using the training data specific to each subtask (i.e., TWT-2023 for Subtask 1 and MADAR-4-MT for Subtask 2 and Subtask 3) for 10 epochs with a learning rate of $2e-5$ and batch size of 32. The maximum length is set to 256 tokens and we set an early stopping patience to 5. Following each epoch, we evaluate each model and select the best the best-performing model on the respective Dev set. Subsequently, we present the performance metrics of this best-performing model on the test datasets.

5.3 Shared Task Results

Table 5 presents the leaderboard of Subtask 1 and is sorted by macro-F_1. As Table 5 shows, for each team, we take their best macro-F_1 score to represent them. `Team NLPeople` (Elkaref et al., 2023)

[2]Recall that MADAR-4-MT is extracted from MADAR-2018

[3]The different CodaLab competitions are available at the following links: `subtask-1`, `subtask-2`, and `subtask-3`.

[4]We use the same baseline models to both Subtask 2 and Subtask 3.

Team	Affiliation	Tasks
AIC	Applied Innovation Center, Egypt	1
ANLP-RG (Derouich et al., 2023)	University of Sfax, Tunisia	2
Arabitools	STEAM Solutions, Palestine	1
Cordyceps	University of Toronto, Canada	1
DialectNLU (Veeramani et al., 2023)	UCLA, USA	1, 2
Exa	Exa, Iran	1
Frank (Azizov et al., 2023)	MBZUAI, UAE	1
Fraunhofer IAIS	Fraunhofer IAIS, Germany	1, 2
Helsinki-NLP (Kuparinen et al., 2023)	University of Helsinki, Finland	2, 3
ISL-AAST (El-sayed and Elmadany, 2023)	Arab Academy for Science and Technology, Egypt	1
IUNADI (Hatekar and Abdo, 2023)	Indiana University Bloomington, USA	1
Mavericks (Deshpande et al., 2023)	Pune Institute of Computer Technology, India	1
NAYEL	Benha University, Egypt	1
NLPeople (Elkaref et al., 2023)	IBM Research Europe, UK	1
rematchka (Abdel-Salam, 2023)	Cairo University, Egypt	1, 2, 3
SANA (Almarwani and Aloufi, 2023)	Taibah University, KSA	1
UniManc (Khered et al., 2023)	University of Manchester, UK	2, 3
UoT (Nwesri et al., 2023)	University of Tripoli, Libya	1
usthb (Lichouri et al., 2023)	USTHB, Alegria	1

Table 4: List of teams that participated in NADI-2023 shared task. Teams with accepted papers are cited.

Rank	Team	F1	Acc.	Pre.	Rec.
1	NLPeople	87.27	87.22	87.37	87.22
2	rematchka	86.18	86.17	86.29	86.17
3	Arabitools	85.86	85.81	86.10	85.81
4	SANA	85.43	85.39	85.60	85.39
5	Frank	84.76	84.75	84.95	84.75
6	ISL-AAST	83.73	83.67	83.87	83.67
7	UoT	82.87	82.86	83.17	82.86
8	AIC	82.37	82.42	82.57	82.42
9	Cordyceps	82.17	82.14	82.57	82.14
Baseline I	MARBERT$_{v2}$	81.44	81.36	81.68	81.36
10	DialectNLU	80.56	80.50	80.92	80.50
Baseline II	AraBERT$_{twitter}$	77.02	76.97	77.54	76.97
11	Mavericks	76.65	76.47	77.43	76.47
Baseline III	CAMeLBERT$_{da}$	74.56	74.47	74.90	74.47
12	exa	70.72	71.03	72.26	71.03
13	IUNADI	70.22	70.78	71.32	70.78
14	NAYEL	63.09	63.39	63.30	63.39
15	usthb	62.51	62.17	63.07	62.17
16	Fraunhofer IAIS	29.91	33.14	38.47	31.39

Table 5: Results of Subtask 1 (Country-Level DA).

Rk	Team	Overall	Egy.	Emi.	Jor.	Pal.
1	UniManc	14.76	16.04	14.30	12.55	13.55
2	Helsinki	14.28	12.22	23.13	11.15	13.42
3	DialectNLU	13.43	11.45	21.59	10.64	12.66
4	rematchka	11.37	11.18	11.99	10.47	10.86
5	ANLP-RG	10.02	10.25	8.50	10.26	9.33
Baseline I	AraT5$_{v2}$	7.70	5.50	10.45	9.51	6.48
6	Fraunhofer IAIS	5.85	8.08	3.90	4.96	6.01
Baseline II	mT5	2.98	4.17	3.66	3.89	3.95
Baseline III	AraBART	2.63	2.44	3.16	1.89	2.60

Table 6: Results of Subtask 2 (Closed AD to MSA MT)

Rank	Team	Overall	Egy.	Emi.	Jor.	Pal.
1	UniManc	21.10	17.65	28.46	22.03	17.29
2	Helsinki-NLP	17.69	16.11	25.81	15.60	15.91
3	rematchka	11.37	11.18	11.99	10.47	10.86
Baseline I	AraT5$_{v2}$	5.41	5.50	5.84	6.06	4.47
Baseline II	mT5	2.98	4.17	3.66	3.89	3.95
Baseline III	AraBART	1.12	0.00	0.00	1.17	1.10

Table 7: Results of Subtask 3 (Open DA to MSA MT).

obtained the best performance on Subtask 1 with 87.27 macro-F_1. We can observe that 9 teams outperform our strongest baseline, MARBAET (i.e, Baseline I). Table 6 and Table 7 show the leaderboard of Subtask 2 and 3, respectively. Both are sorted by their main metrics, the overall BLEU score. Team UniManc (Khered et al., 2023) won both subtasks, achieving the best BLEU scores of 14.76 and 21.10 on Subtask 2 and 3, respectively. We observe that *five* teams outperform our Baseline I on Subtask 2.

5.4 General Description of Submitted Systems

In Tables 8 and 9, we provide a high-level summary of the submitted systems. For each team, we list

Team	# submit	F_1	Features				Techniques								
			N-gram	TF-IDF	Linguistic	Word embeds	Classical ML	Neu. nets	PLM	Ensemble	Adapter	Hie. Cls	Prompting	Contrast. L	Data Aug.
NLPeople	5	87.27					✓	✓	✓	✓					✓
rematchka	3	86.18							✓	✓	✓		✓	✓	
Arabitools	4	85.86			✓				✓	✓					
SANA	2	85.43							✓	✓					
Frank	2	84.76							✓	✓					
ISL-AAST	5	83.73						✓	✓	✓					
UoT	2	82.87					✓		✓						✓
AIC	5	82.37	✓		✓				✓	✓		✓			✓
Cordyceps	4	82.17													
DialectNLU	5	80.56		✓	✓			✓	✓	✓					
IUNADI	1	70.22			✓				✓	✓					
Mavericks	1	76.65							✓	✓					
NAYEL	5	63.09	✓	✓			✓								
usthb	3	62.51	✓	✓	✓	✓	✓								✓

Table 8: Summary of approaches used by participating teams in Subtask 1. Teams are sorted by their performance on the official metric, Macro-F_1 score. Classical machine learning (ML) indicates any non-neural machine learning methods such as naive Bayes and support vector machines. The term "neural nets" refers to any model based on neural networks (e.g., FFNN, RNN, CNN, and Transformer) trained from scratch. PLM refers to neural networks pretrained with unlabeled data such as MARBERT. (Hie. Cls, hierarchical classification approach); (Contrast. L, contrastive learning); (Data Aug., data Augmentation).

Team	# submit	BLUE	Techniques				
			Classic ML	NN	PLM	Ensemble	Aug.
			Subtask 2				
UniManc	5	14.76			✓	✓	
Helsinki	3	14.28	✓	✓	✓		
DialectNLU	5	13.43			✓	✓	
rematchka	1	11.37		✓	✓		
ANLP-RG	1	10.02		✓	✓		✓
			Subtask 3				
UniManc	5	21.10			✓	✓	✓
Helsinki-NLP	5	17.69	✓	✓	✓		✓
rematchka	1	11.37			✓		

Table 9: Summary of approaches used by participating teams in Subtask 2 and 3. Teams are sorted by their performance on BLEU score for both Subtasks. Classical machine learning (ML) indicates any non-neural machine learning methods such as naive Bayes and support vector machines. "NN" refers to any model based on neural networks (e.g., FFNN, RNN, CNN, and Transformer) trained from scratch. PLM refers to neural networks pretrained with unlabeled data such as AraT5. (Aug., data augmentation).

the best score with the main metric of each subtask and the number of submissions made by the team. As shown in these tables, most teams use pretrained language models (PLM), including Transformer encoder-based PLMs (e.g., AraBERT (Antoun et al., 2020) and MARBERT (Abdul-Mageed et al., 2021a)) for Subtask 1 and Transformer encoder-decoder PLMs (e.g., ArabT5 (Nagoudi et al., 2022)) for Subtask 2 and Subtask 3. Ensemble voting is also an effective approach most teams employ in Subtask 1.

The top team of Subtask 1, i.e., `NLPeople` (Elkaref et al., 2023), exploits MARBERT, AraBERT, and AraT5 with different finetuning strategies (e.g., staged finetuning). To enrich the learning context, they use a retrieval method to find similar texts from the training set for a given text and then append the retrieved texts along with corresponding labels as additional input. Their best submission is an ensemble with ten models. Team `rematchka` (Abdel-Salam, 2023), exploits MARBERT, AraBERT, AraELECTRA (Antoun et al., 2021), and CAMeLBERT (Obeid et al., 2020) with different prompting techniques and add linguistic features to their models. They also use supervised contrastive loss (Gunel et al., 2021) to enhance model finetuning. Teams `SANA` (Almarwani and Aloufi, 2023) and `Frank` (Azizov et al., 2023) both finetune PLMs and apply ensemble voting to achieve their best performance.

On Subtask 2 (closed track), the winning team, `Team UniManc` (Khered et al., 2023), finetune three variants of T5 models (i.e., mT5 (Xue et al., 2021), mT0 (Muennighoff et al., 2023), and AraT5) with the officially released dataset. For Subtask 3 (open track), `Team UniManc` collects four additional supervised datasets and uses GPT-3.5-turbo to translate 2,712 samples from Subtask 1. Team `Helsinki-NLP` (Kuparinen et al., 2023) finetune ByT5 (Xue et al., 2022) and AraT5 with the officially released dataset of Subtask 2. For Subtask 3, they collect six monolingual MSA datasets and synthesize a parallel dataset by exploiting character-level statistical machine translation models to translate the MSA to different dialects. They then finetune PLMs with the supervised dataset from Subtask 2 and their synthetic dataset. Similarly, both teams `DialectNLU` and `rematchka` finetune AraT5 with the training data of Subtask 2.

6 Conclusion

We presented findings and results of NADI-2023, the fourth edition of the NADI shared task focused on fine-grained Arabic dialect identification. This edition also introduced two subtasks centered on machine translation from four Arabic dialects into MSA. Results acquired by participant teams show that dialect identification remains a challenging task but that various types of approaches, many of which involve exploiting language models, can be used to handle the task. Similarly, translating Arabic dialects is unsurprisingly very challenging due to lack of training data. In the future, we plan to continue supporting both dialect identification and machine translation through NADI.

7 Limitations

Our work has a number of limitations, as follows:

- Although we strive for widest coverage, this edition of NADI focused on only 18 country-level dialects. This is due to our inability to develop high quality datasets for a few countries such as *Comoros*, *Djibouti*, *Mauritania*, and *Somalia*.

- NADI continues to use short texts for the Arabic dialects. Due to lack of dialectal data from other sources, we depend on short posts from Twitter. Although these data have thus far empowered development of effective dialect identification models, it is desirable to afford data from other domains that have longer texts. This will allow development of more widely applicable models.

- Our MADAR-18 dataset is commissioned and, although useful, should not be used to analyze Arabic dialects as a replacement for naturally occurring data.

- Our machine translation subtasks focus only on four dialects and do not offer sizeable datasets. Modern MT systems need much larger data to perform well. Again, in spite of our best efforts, parallel datasets involving dialects remain limited.

8 Ethical Considerations

The NADI-2023 Subtask 1 dataset is sourced from the public domain (i.e., X former Twitter), with user personal information and identity carefully concealed. Similarly, the NADI-2023 Subtask 2 and Subtask 3 datasets are manually created. Again, we take meticulous measures to remove user identities and personal information from this dataset. As a result, we have minimal concerns about the retrieval of personal information from our data. However, it is crucial to acknowledge that the datasets we collect to construct NADI-2023 Subtask 1 may contain potentially harmful content. Additionally, during model evaluation, there is a possibility of exposure to biases that could unintentionally generate problematic content.

Acknowledgments

We acknowledge support from Canada Research Chairs (CRC), the Natural Sciences and Engineering Research Council of Canada (NSERC; RGPIN-2018-04267), the Social Sciences and Humanities Research Council of Canada (SSHRC; 435-2018-0576; 895-2020-1004; 895-2021-1008), Canadian Foundation for Innovation (CFI; 37771), Digital Research Alliance of Canada,[5] and UBC Advanced Research Computing-Sockeye.[6]

References

Reem Abdel-Salam. 2023. rematchka at NADI 2023 shared task: Parameter Efficient tuning for Dialect Identification and Dialect Machine Translation. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*.

[5] https://alliancecan.ca
[6] https://arc.ubc.ca/ubc-arc-sockeye

Ahmed Abdelali, Hamdy Mubarak, Younes Samih, Sabit Hassan, and Kareem Darwish. 2021. QADI: Arabic dialect identification in the wild. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 1–10, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Muhammad Abdul-Mageed. 2015. *Subjectivity and sentiment analysis of Arabic as a morophologically-rich language*. Ph.D. thesis, Indiana University.

Muhammad Abdul-Mageed, Hassan Alhuzali, and Mohamed Elaraby. 2018. You tweet what you speak: A city-level dataset of Arabic dialects. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021a. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020a. NADI 2020: The first nuanced Arabic dialect identification shared task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 97–110, Barcelona, Spain (Online). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2021b. NADI 2021: The second nuanced Arabic dialect identification shared task. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 244–259, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2022. Nadi 2022: The third nuanced arabic dialect identification shared task. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 85–97.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, and Lyle Ungar. 2020b. Toward micro-dialect identification in diaglossic and code-switched environments. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5855–5876, Online. Association for Computational Linguistics.

Rania Al-Sabbagh and Roxana Girju. 2012. YADAC: Yet another dialectal Arabic corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2882–2889, Istanbul, Turkey. European Language Resources Association (ELRA).

Nora Al-Twairesh, Rawan N. Al-Matham, Nora Madi, Nada Almugren, Al-Hanouf Al-Aljmi, Shahad Alshalan, Raghad Alshalan, Nafla Alrumayyan, Shams Al-Manea, Sumayah Bawazeer, Nourah Al-Mutlaq, Nada Almanea, Waad Bin Huwaymil, Dalal Alqusair, Reem Alotaibi, Suha Al-Senaydi, and Abeer Alfutamani. 2018. SUAR: towards building a corpus for the saudi dialect. In *Fourth International Conference On Arabic Computational Linguistics, ACLING 2018, November 17-19, 2018, Dubai, United Arab Emirates*, volume 142 of *Procedia Computer Science*, pages 72–82. Elsevier.

Nada Almarwani and Samah Aloufi. 2023. SANA at NADI 2023 shared task: Ensemble of Layer-Wise BERT-based models for Dialectal Arabic Identification. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*.

Israa Alsarsour, Esraa Mohamed, Reem Suwaileh, and Tamer Elsayed. 2018. Dart: A large dataset of dialectal arabic tweets. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Maha J Althobaiti. 2020. Automatic Arabic dialect identification systems for written texts: A survey. *arXiv preprint arXiv:2009.12622*.

Maha J. Althobaiti. 2022. Creation of annotated country-level dialectal Arabic resources: An unsupervised approach. *Nat. Lang. Eng.*, 28(5):607–648.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Wissam Antoun, Fady Baly, and Hazem M. Hajj. 2021. Araelectra: Pre-training text discriminators for arabic language understanding. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop, WANLP 2021, Kyiv, Ukraine (Virtual), April 9, 2021*, pages 191–195. Association for Computational Linguistics.

Dilshod Azizov, Jiyong Lee, Siwei Liu, and Shangsong Liang. 2023. Frank at NADI 2023 Shared Task: Trio-Based Ensemble Approach for Arabic Dialect Identification. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*.

MS Badawi. 1973. Levels of contemporary Arabic in Egypt. *Cairo: Dâr al Ma'ârif*.

Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2022. Hierarchical aggregation of dialectal data for Arabic dialect identification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 4586–4596. European Language Resources Association.

Laith H Baniata, Seyoung Park, and Seong-Bae Park. 2018. A neural machine translation model for arabic dialects that utilizes multitask learning (mtl). *Computational intelligence and neuroscience*, 2018.

Houda Bouamor, Nizar Habash, and Kemal Oflazer. 2014. A multidialectal parallel corpus of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1240–1245, Reykjavik, Iceland. European Language Resources Association (ELRA).

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)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Houda Bouamor, Sabit Hassan, and Nizar Habash. 2019. The MADAR shared task on Arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 199–207, Florence, Italy. Association for Computational Linguistics.

Kristen Brustad. 2000. *The Syntax of Spoken Arabic: A Comparative Study of Moroccan, Egyptian, Syrian, and Kuwaiti Dialects*. Georgetown University Press.

Mark W. Cowell. 1964. *A Reference Grammar of Syrian Arabic*. Georgetown University Press, Washington, D.C.

Kareem Darwish, Hamdy Mubarak, Ahmed Abdelali, Mohamed Eldesouki, Younes Samih, Randah Alharbi, Mohammed Attia, Walid Magdy, and Laura Kallmeyer. 2018. Multi-dialect Arabic POS tagging: A CRF approach. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Wiem Derouich, Sameh Kchaou, and Rahma Boujelbane. 2023. ANLP-RG at NADI 2023 shared task: Machine Translation of Arabic Dialects: A Comparative Study of Transformer Models. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*.

Kshitij Deshpande, Yash Patwardhan, Vedant Deshpande, and Sudeep Mangalvedhekar. 2023. Mavericks at NADI 2023 Shared Task: Unravelling Regional Nuances through Dialect Identification using Transformer-based Approach. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*.

Mona Diab, Nizar Habash, Owen Rambow, Mohamed Altantawy, and Yassine Benajiba. 2010. COLABA: Arabic dialect annotation and processing. In *LREC workshop on Semitic language processing*, pages 66–74.

Moussa Kamal Eddine, Nadi Tomeh, Nizar Habash, Joseph Le Roux, and Michalis Vazirgiannis. 2022. Arabart: a pretrained arabic sequence-to-sequence model for abstractive summarization. *arXiv preprint arXiv:2203.10945*.

Mahmoud El-Haj. 2020. Habibi - a multi dialect multi national Arabic song lyrics corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1318–1326, Marseille, France. European Language Resources Association.

Shorouk Adel El-sayed and Noureldin Elmadany. 2023. ISL-AAST at NADI 2023 shared task: Enhancing Arabic Dialect Identification in the Era of Globalization and Technological Progress. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*.

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.

Mohab Elkaref, Movina Moses, Shinnosuke Tanaka, James Barry, and Geeth De Mel. 2023. NLPeople at NADI 2023 Shared Task: Arabic Dialect Identification with Augmented Context and Multi-Stage Tuning. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*.

AbdelRahim Elmadany, ElMoatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. ORCA: A challenging benchmark for Arabic language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9559–9586, Toronto, Canada. Association for Computational Linguistics.

Wael Farhan, Bashar Talafha, Analle Abuammar, Ruba Jaikat, Mahmoud Al-Ayyoub, Ahmad Bisher Tarakji, and Anas Toma. 2020. Unsupervised dialectal neural machine translation. *Information Processing & Management*, 57(3):102181.

Hassan Gadalla, Hanaa Kilany, Howaida Arram, Ashraf Yacoub, Alaa El-Habashi, Amr Shalaby, Krisjanis Karins, Everett Rowson, Robert MacIntyre, Paul Kingsbury, David Graff, and Cynthia McLemore. 1997. CALLHOME Egyptian Arabic transcripts LDC97T19. Web Download. Philadelphia: Linguistic Data Consortium.

Imane Guellil, Faical Azouaou, and Mourad Abbas. 2017. Neural vs statistical translation of algerian arabic dialect written with arabizi and arabic letter. In *The 31st Pacific Asia Conference on Language, Information and Computation PACLIC*, volume 31, page 2017.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2021. Supervised contrastive learning for pre-trained language model fine-tuning. In *9th International Conference on Learning Representations,* *ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Nizar Habash, Houda Bouamor, Hazem Hajj, Walid Magdy, Wajdi Zaghouani, Fethi Bougares, Nadi Tomeh, Ibrahim Abu Farha, and Samia Touileb, editors. 2021. *Proceedings of the Sixth Arabic Natural Language Processing Workshop*. Association for Computational Linguistics, Kyiv, Ukraine (Virtual).

Nizar Y Habash. 2010. *Introduction to Arabic natural language processing*, volume 3. Morgan & Claypool Publishers.

Salima Harrat, Karima Meftouh, Mourad Abbas, and Kamel Smaïli. 2014. Building resources for algerian arabic dialects. In *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, pages 2123–2127. ISCA.

R.S. Harrell. 1962. *A Short Reference Grammar of Moroccan Arabic: With Audio CD*. Georgetown classics in Arabic language and linguistics. Georgetown University Press.

Yash A. Hatekar and Muhammad S. Abdo. 2023. IU-NADI: Country-level Arabic Dialect Classification in Tweets for the Shared Task NADI 2023. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*.

Clive Holes. 2004. *Modern Arabic: Structures, Functions, and Varieties*. Georgetown Classics in Arabic Language and Linguistics. Georgetown University Press.

Elsayed Issa, Mohammed AlShakhori1, Reda Al-Bahrani, and Gus Hahn-Powell. 2021. Country-level Arabic dialect identification using RNNs with and without linguistic features. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 276–281, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Mustafa Jarrar, Nizar Habash, Faeq Alrimawi, Diyam Akra, and Nasser Zalmout. 2016. Curras: an annotated corpus for the Palestinian Arabic dialect. *Language Resources and Evaluation*, pages 1–31.

Karima Kadaoui, Samar M Magdy, Abdul Waheed, Md Tawkat Islam Khondaker, Ahmed Oumar El-Shangiti, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Tarjamat: Evaluation of bard and chatgpt on machine translation of ten arabic varieties. *arXiv preprint arXiv:2308.03051*.

Salam Khalifa, Nizar Habash, Dana Abdulrahim, and Sara Hassan. 2016. A large scale corpus of Gulf Arabic. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4282–4289, Portorož, Slovenia. European Language Resources Association (ELRA).

Abdullah Khered, Ingy Yasser Abdelhalim, Nadine Abdelhalim, Ahmed Soliman, and Riza Batista-Navarro. 2023. UniManc at NADI 2023 Shared Task: A Comparison of Various T5-based Models for Translating Arabic Dialectical Text to Modern Standard Arabic. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*.

Olli Kuparinen, Aleksandra Miletić, and Yves Scherrer. 2023. The Helsinki-NLP Submissions at NADI 2023 Shared Task: Walking the Baseline. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*.

Kathrein Abu Kwaik, Motaz Saad, Stergios Chatzikyriakidis, and Simon Dobnik. 2018. Shami: A corpus of levantine arabic dialects. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Mohamed Lichouri, Khaled Lounnas, Aicha Zitouni, Houda Latrache, and Rachida Djeradi. 2023. USTHB at NADI 2023 shared task: Exploring Preprocessing and Feature Engineering Strategies for Arabic Dialect Identification. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*.

Shervin Malmasi, Marcos Zampieri, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, and Jörg Tiedemann. 2016. Discriminating between similar languages and Arabic dialect identification: A report on the third DSL shared task. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 1–14, Osaka, Japan. The COLING 2016 Organizing Committee.

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.

Hamdy Mubarak and Kareem Darwish. 2014. Using Twitter to collect a multi-dialectal corpus of Arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 1–7, Doha, Qatar. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M. Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15991–16111. Association for Computational Linguistics.

El Moatez Billah Nagoudi, Ahmed El-Shangiti, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2023. Dolphin: A challenging and diverse benchmark for arabic nlg. *arXiv preprint arXiv:2305.14989*.

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.

Abdusalam Nwesri, Nabila Shinbir, and Hasan Hasan. 2023. UoT at NADI 2023 shared task: Automatic Arabic Dialect Identification is Made Possible. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*.

Ossama Obeid, Mohammad Salameh, Houda Bouamor, and Nizar Habash. 2019. ADIDA: Automatic dialect identification for Arabic. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 6–11, Minneapolis, Minnesota. Association for Computational Linguistics.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association.

Fatiha Sadat, Farzindar Kazemi, and Atefeh Farzindar. 2014. Automatic identification of Arabic language varieties and dialects in social media. In *Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP)*, pages 22–27, 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.

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.

Mohammad Salameh, Houda Bouamor, and Nizar Habash. 2018. Fine-grained Arabic dialect identification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1332–1344, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Wael Salloum, Heba Elfardy, Linda Alamir-Salloum, Nizar Habash, and Mona Diab. 2014. Sentence level dialect identification for machine translation system selection. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 772–778.

Hariram Veeramani, Surendrabikram Thapa, and Usman Naseem. 2023. DialectNLU at NADI 2023 Shared Task: Transformer Based MultiTask Approach Jointly Integrating Dialect and Machine Translation Tasks. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

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.

Wajdi Zaghouani and Anis Charfi. 2018. Arap-tweet: A large multi-dialect Twitter corpus for gender, age and language variety identification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Omar F. Zaidan and Chris Callison-Burch. 2011. The Arabic online commentary dataset: an annotated dataset of informal Arabic with high dialectal content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 37–41, Portland, Oregon, USA. Association for Computational Linguistics.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar 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.

Imed Zitouni, Muhammad Abdul-Mageed, Houda Bouamor, Fethi Bougares, Mahmoud El-Haj, Nadi Tomeh, and Wajdi Zaghouani, editors. 2020. *Proceedings of the Fifth Arabic Natural Language Processing Workshop*. Association for Computational Linguistics, Barcelona, Spain (Online).