Adapting MARBERT for Improved Arabic Dialect Identification: Submission to the NADI 2021 Shared Task

Badr AlKhamissi*
Independent
badr [at] khamissi.com

Mohamed Gabr*
Microsoft EGDC
mohamed.gabr [at] microsoft.com

Muhammed ElNokrashy
Microsoft EGDC
muelnokr [at] microsoft.com

Khaled Essam
Mendel.ai
khaled.essam [at] mendel.ai

Abstract

In this paper, we tackle the Nuanced Arabic Dialect Identification (NADI) shared task (Abdul-Mageed et al., 2021) and demonstrate state-of-the-art results on all of its four subtasks. Tasks are to identify the geographic origin of short Dialectal (DA) and Modern Standard Arabic (MSA) utterances at the levels of both country and province. Our final model is an ensemble of variants built on top of MARBERT that achieves an F1-score of 34.03% for DA at the country-level development set—an improvement of 7.63% from previous work.

1 Introduction

The Arab World is a vast geographical region that covers North Africa and Southwest Asia, boasting a population of around 400M that speak different derivatives of a common language. However, by virtue of its size and cultural variety, there exists a dialect continuum across the region wherein the language varieties of neighboring peoples may differ slightly, but distant regions can become mutually unintelligible. This continuum is referred to as Dialectal Arabic (DA) and is the "Low" variety of modern Arabic *diglossia*. On the other hand, the "High" variety is referred to as Modern Standard Arabic (MSA) and is used in formal settings such as academia, mass media, and legislation—and is taught through the formal education system in most Arab countries. This standard variant emerged gradually, but most notably with the advent of the printing press in the 19th century. It diverged from Classical Arabic (CA) into a more simple version that is now used across the Arab World.

The modern vernacular dialects (DA) differ along several dimensions, including pronunciation, syntax, morphology, vocabulary, and even orthography. Dialects may be heavily influenced by previously dominant local languages. For example, Egyptian variants are influenced by the Coptic language, while Sudanese variants are influenced by the Nubian language.

In this paper, we study the classification of such variants and describe our model that achieves state-of-the-art results on all of the four Nuanced Arabic Dialect Identification (NADI) subtasks (Abdul-Mageed et al., 2021). The task focuses on distinguishing both MSA and DA by their geographical origin at both the country and province levels. The data is a collection of tweets covering 100 provinces from 21 Arab countries. The code has been made open-source and available on GitHub[1].

2 Related Work

The first efforts to collect and label dialectal Arabic data goes back to 1997 (Gadalla, 1997). However, studying DA in NLP started to gain traction in recent years as more digital Arabic data became available, especially with the rise of online chatting platforms that place no restrictions on the syntax, style, or formality of the writing. Zaidan and Callison-Burch (2014) labelled the Arabic Online Commentary Dataset (AOC) through crowd-sourcing, then built a model to classify even more crawled data. Abdelali et al. (2020) provided the QADI dataset by automatically collecting dialectal Arabic tweets and labelling them based on descriptions of the author accounts, while trying to reduce the recall of written MSA and inappropriate tweets. Abu Farha and Magdy (2020) provided ArSarcasm: a dataset of labelled tweets. Originally for sarcasm detection, it also contains labels for sentiment and dialect detection. Labelled dialectal Arabic challenges

* Equal contribution.

[1] https://github.com/mohamedgabr96/NeuralDialectDetector

Proceedings of the Sixth Arabic Natural Language Processing Workshop, pages 260–264
Kyiv, Ukraine (Virtual), April 19, 2021.

such as MADAR (Bouamor et al., 2019) and NADI (Abdul-Mageed et al., 2020b, 2021) (which started in 2020) shed light on the underlying challenges of the task. Both comprise labelled Arabic tweets but with class sets of different granularities.

Talafha et al. (2020) presents a solution that won the 2020 NADI shared task (Subtask 1.2) by continuing training AraBERT (Antoun et al., 2020) using Masked Language Modelling (MLM) on 10M unlabelled tweets, then fine-tuning on the dialect identification task. On the other hand, the solution that won Subtask 2.2 uses a hierarchical classifier that takes as input a weighted combination of TF-IDF and AraBERT features to first classify the country, then invokes an ArabBERT-based, country-specific, province-level classifier to detect the province (El Mekki et al., 2020).

The large size of pretrained Transformer models hinders their applicability in many use cases. The usual number of parameters in such a model lies between 150M and 300M and needs between 500MB and 1GB of space to be stored. A novel technique to address these issues is proposed by Houlsby et al. (2019)—bottleneck layers ("adapters") are added to each transformer layer as an alternative to fine-tuning the entire pre-trained model when optimizing for downstream tasks (Houlsby et al., 2019; Bapna and Firat, 2019; Pfeiffer et al., 2021). Only these additional parameters (which can be 1% or less of the size of the main model) need to be stored per downstream task, given that they are the only layers changing. Besides being light-weight and scalable, adapters offer several advantages over traditional approaches of transfer-learning: (1) They learn modular representations that are compatible with other layers of the transformer, (2) They avoid interfering with pre-trained knowledge, mitigating catastrophic forgetting and catastrophic inference—two common downsides of multi-task learning (Pfeiffer et al., 2020).

3 Datasets & Subtasks

Dataset	Country	Province
MSA	Subtask 1.1	Subtask 2.1
DA	Subtask 1.2	Subtask 2.2

Table 1: Subtask ID per Dataset and Granularity Level

The second NADI shared task consists of four subtasks on two datasets (see Table 1). Each consists of 21k tweets for training, 5k for development and 5k for testing collected from a disjoint set of users (Abdul-Mageed et al., 2021). Both datasets are labelled at two levels of granularity: country, and province. The NADI task is the first to focus on sub-country level dialects. However, the data is extremely unbalanced, even at the country-level (see Figure 1), with the most frequent class being Egypt (4283 instances), and the least common class being Somalia (172 instances). The data comes preprocessed with URLs replaced with the token 'URL' and Twitter mentions replaced with the token 'USER'. One of the main challenges of Arabic Dialect Identification is the high similarity of dialects in short utterances; many short phrases are commonly used in all dialects. Since the tweets are collected "in the wild" and DA is not formally defined (the same word can be written in a variety of ways), this makes it even more challenging to capture the meaning of a logical word unit across its different forms.

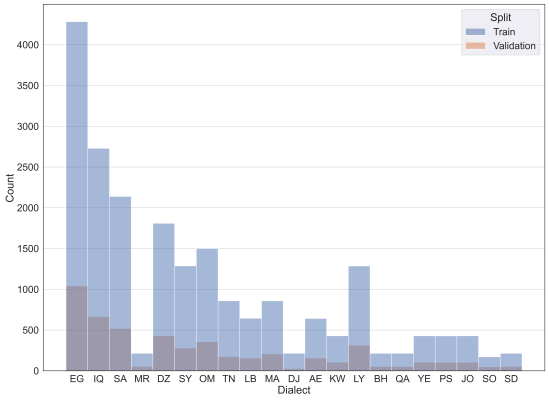

Figure 1: Train/Dev Corpora Sizes per Country (DA)

4 System Description

Our model builds on MARBERT—a publicly available transformer model trained on 1B multi-dialectal Arabic tweets (Abdul-Mageed et al., 2020a). It follows the $\text{BERT}_{\text{BASE}}$ architecture (Devlin et al., 2019) with 163M parameters and similarly uses WordPiece tokenization (Wu et al., 2016). To optimize it for the task at hand, we introduce changes to the architecture and training regimen as described below. Note that, due to time and compute constraints, all hyperparameters were optimized on the development set of Subtask 1.2 then applied as-is to the other three subtasks. We report the result of the best ensemble for each subtask.

General The experiments described below all use the following configuration: The classification

head is a softmax layer over the CLS vector of the last layer $\mathbf{z}_L$ (with a dropout rate of 30% during training). The base learning rate of the classification head is set higher ($1e{-}2$) than the rest of the trainable parameters ($5e{-}6$). The LR schedule is warmed up linearly over 250 steps, then decayed once every 10 gradient updates following an inverse-square-root scheme to the minimum $0.01 \cdot LR_{\text{base}}$. We use the Adam optimizer with decoupled weight regularization (Loshchilov and Hutter, 2019). During training, we evaluate the model every 100 mini-batches of size 32, and halt if the dev macro-F1 score does not improve for 10 consecutive evaluations. The maximum sequence length is 90 for the DA dataset and 110 for the MSA dataset.

Fine-tuning We fine-tune the full MARBERT transformer using the base configuration.

Adapters Here we embed two additional layers at each transformer block (one after the Multi-Head Attention module and one after the FFN module) following the Houlsby et al. (2019) architecture. This allows us to preserve the pre-trained embedded knowledge in the MARBERT layers, which are trained on a rich corpus with less bias towards specific dialects. The final architecture of a transformer block is illustrated in Figure 5.

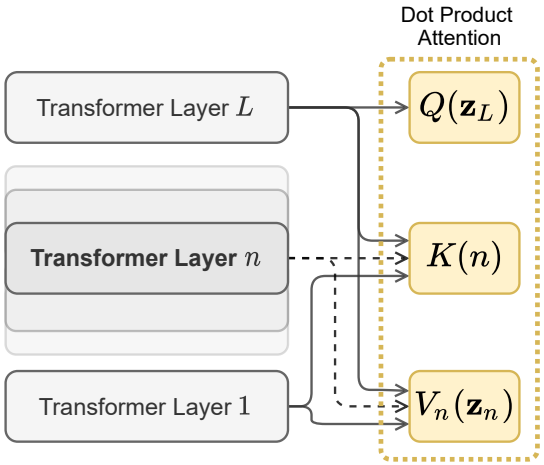

Figure 2: Vertical Attention

Vertical Attention (VAtt) The MARBERT model has $L = 12$ transformer layers. For each $n \in L$, let $\mathbf{z}_n$ be the CLS token at level n (after the adapter). Let k_n be a static learned positional embedding for level n. Apply a scaled dot-product based attention module where: query is $Q(\mathbf{z}_L)$, keys are $\{K(k_n)\}_n^L$, and values are $\{V_n(\mathbf{z}_n)\}_n^L$. This attends over the layers' sentence representations by content-to-level (depth) addressing. We introduce this to allow choice of the abstraction level. See Figure 2.

Ensembling We create an ensemble of multiple models on combinations of the following architectural variables:

- Whether Vertical Attention is used.
- Training adapters or fine-tuning full model.

The soft-max outputs of the models are aggregated together by doing an element-wise multiplication. The best ensemble provide 1.13% F1 boost over the best solo model (see Table 2).

5 Results

Models	DEV		TEST	
	Acc.	F1	Acc.	F1
Adapters	***52.48***	32.10	50.62	30.78
+VAtt	52.28	31.73	***51.08***	30.09
Fine-tuning	51.02	32.23	50.28	30.41
+VAtt	50.07	***32.90***	49.42	***31.30***
Ensemble	**53.42**	**34.03**	**51.66**	**32.26**

Table 2: Ablation study (Subtask 1.2)

The results table shows that the best F1 score is obtained by ensembling the following list of model configurations (all of which use a maximum sequence length of 90): (1) Fine-tuning, (2) Adapters + VAtt, (3) Fine-tuning + VAtt, and (4) Fine-tuning using a linear learning rate schedule instead of the inverse-square-root scheme. Among solo models, the best performer is the fully fine-tuned variant with Vertical Attention, with an F1 score of 32.90 on the development set.

6 Discussion

The confusion matrix of the Ensemble in Subtask 1.2 (Table 2) shows over-prediction of Egyptian and Saudi Arabian, reflecting their over-representation in training (Figure 1). The matrix suggests that the dialects most confused together are often from geographically close countries. For example: Tunisia and Libya, and Qatar and Bahrain. Gulf countries also show high confusion.

Considering the sequence length in words for correct and wrong predictions, we find that correctly predicted sentences tend to be longer than

Subtask	Models	DEV		TEST	
		Acc.	F1	Acc.	F1
1.1	**Ours**	**39.06**	**23.52**	**35.72**	**22.38**
1.2	MARBERT	48.86	26.40	48.40	29.14
	Ours	**53.42**	**34.03**	**51.66**	**32.26**
2.1	**Ours**	**7.04**	**6.73**	**6.66**	**6.43**
2.2	MARBERT	7.91	5.23	8.48	6.28
	Ours	**10.74**	**10.02**	**9.46**	**8.60**

Table 3: Results compared to previous SOTA. MARBERT taken from Abdul-Mageed et al. (2020a).

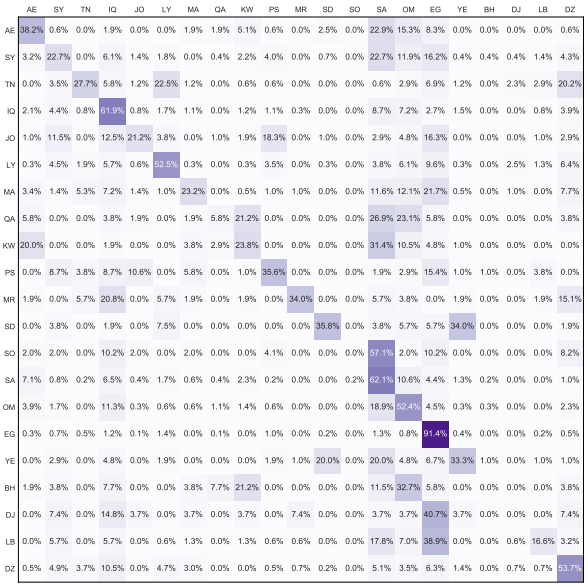

Figure 3: Confusion matrix of the predictions of the best performing model in subtask 1.2.

	Mean/Variance (λ)	Median
Correct	7.16	5.45
Wrong	4.22	3.40
All	4.10	3.26

Table 4: Sequence length (fitted to an Erlang distribution)

the means of all and of wrongly predicted sentences (fit as an Erlang distribution). The numbers were computed after removing the special USER and URL tokens.

7 Conclusion

In this paper, we present a solution that provides a new state-of-the-art on all of the NADI subtasks. Inserting adapters at each of the MARBERT transformer layers preserved the original pre-trained knowledge, stemming from a rich corpus of tweets, while still embedding task knowledge. To further improve the model's performance, we vertically attend on all of the CLS token hidden-states coming out of each of the transformer layers instead of just using the top layer's output. We ensemble models on two design variables: Whether fine-tuning is full or through adapters, and whether Vertical Attention is used. The bias towards dominating classes in the task dataset remains a significant issue that is still not mitigated. As a future work, we would like to employ adapter fusion (Pfeiffer et al., 2021) to attenuate this bias.

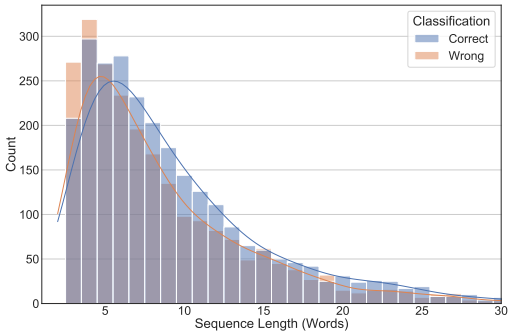

Figure 4: Distribution of length for correct and wrong classifications in subtask 1.2 trimmed at the tail after length > 30.

References

Ahmed Abdelali, Hamdy Mubarak, Younes Samih, Sabit Hassan, and Kareem Darwish. 2020. Arabic dialect identification in the wild.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2020a. Arbert & marbert: Deep bidirectional transformers for arabic. *arXiv preprint arXiv:2101.01785*.

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020b. NADI 2020: The First Nuanced Arabic Dialect Identification Shared Task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop (WANLP 2020)*, Barcelona, Spain.

Muhammad Abdul-Mageed, Chiyu Zhang, Abdel-Rahim Elmadany, Houda Bouamor, and Nizar Habash. 2021. NADI 2021: The Second Nuanced Arabic Dialect Identification Shared Task. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop (WANLP 2021)*.

Ibrahim Abu Farha and Walid Magdy. 2020. From Arabic sentiment analysis to sarcasm detection: The ArSarcasm dataset. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 32–39, Marseille, France. European Language Resource Association.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1538–1548, Hong Kong, China. Association for Computational Linguistics.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Abdellah El Mekki, Ahmed Alami, Hamza Alami, Ahmed Khoumsi, and Ismail Berrada. 2020. Weighted combination of BERT and n-GRAM features for nuanced Arabic dialect identification. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 268–274, Barcelona, Spain (Online). Association for Computational Linguistics.

Hassan and Gadalla. 1997. Callhome egyptian arabic transcripts ldc97t19.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. Adapterfusion: Non-destructive task composition for transfer learning.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. AdapterHub: A Framework for Adapting Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020): Systems Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Bashar Talafha, Mohammad Ali, Muhy Eddin Za'ter, Haitham Seelawi, Ibraheem Tuffaha, Mostafa Samir, Wael Farhan, and Hussein T. Al-Natsheh. 2020. Multi-dialect arabic bert for country-level dialect identification.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, ukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation.

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

A Appendices

A.1 System Architecture

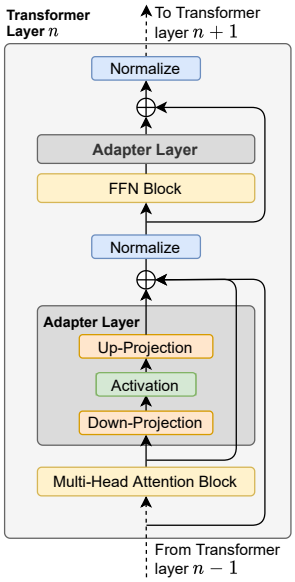

Figure 5: Details of the used model's architecture, specifically looking at one transformer layer. All layers are of the same architecture.[2]

[2]Diagrams generated using `diagrams.net` (draw.io).