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

Badr AlKhamissi, Mohamed Gabr, Muhammad ElNokrashy, Khaled Essam

American University in Cairo

Microsoft EGDC

Mendel.ai

NADI Shared Task

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.

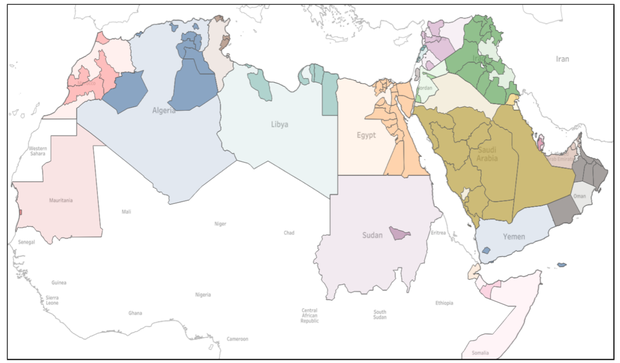
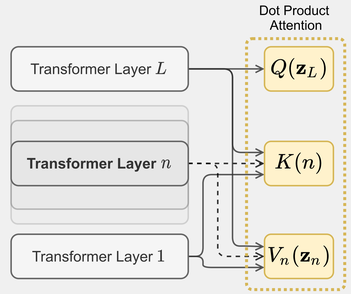


Figure: Taken from [Abdul-Mageed et al., 2021]

The second NADI shared task consists of four subtasks on two datasets: MSA and DA. Each dataset is a collection of tweets covering 100 provinces from 21 Arab countries.

Vertical Attention



Model Architecture

- **Fine-tuning MARBERT** [Abdul-Mageed et al., 2020]
- **Adapters**: adding two additional layers at each transformer block following the [Houlsby et al., 2019].
- **Vertical Attention**: attending on the CLS tokens from all layers then feeding the output to the classifier.
- **Ensembling** on combinations of various architectural variables.

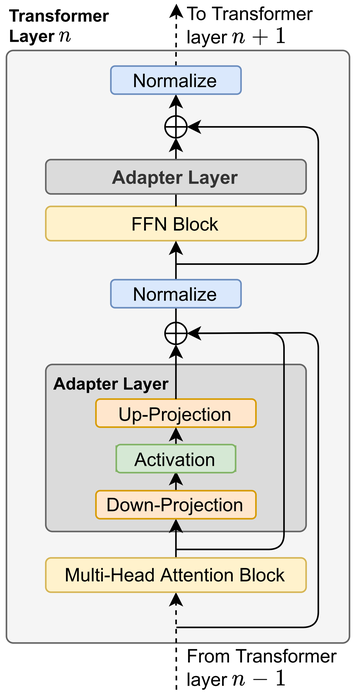


Figure: Details of the used model's architecture, specifically looking at one transformer layer.

Results

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

Analysis

The results obtained show an over-prediction of Egyptian and Saudi Arabian, reflecting their over-representation in training. The dialects that are most confused together are often from geographically close countries.

Correctly predicted sentences tend to be longer than the means of all and of wrongly predicted sentences.

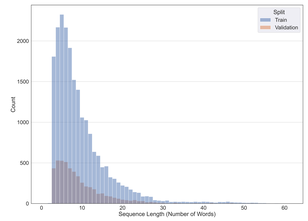


Figure: Distribution of length for correct and wrong classifications in subtask 1.2 trimmed at the tail after length > 30

References

[Abdul-Mageed et al., 2020] Abdul-Mageed, M., Elmadany, A., and Nagoudi, E. M. B. (2020). Arbert & marbert: Deep bidirectional transformers for arabic. *arXiv preprint arXiv:2101.01785*.

[Abdul-Mageed et al., 2021] Abdul-Mageed, M., Zhang, C., Elmadany, A., Bouamor, H., and Habash, N. (2021). NADI 2021: The Second Nuanced Arabic Dialect Identification Shared Task. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop (WANLP 2021)*.

[Houlsby et al., 2019] Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. (2019). Parameter-efficient transfer learning for nlp.