Dialect Identification in Nuanced Arabic Tweets Using Farasa Segmentation and AraBERT

Anshul Wadhawan
Flipkart Private Limited
`anshul.wadhwan@flipkart.com`

Abstract

This paper presents our approach to address the EACL WANLP-2021 Shared Task 1: Nuanced Arabic Dialect Identification (NADI). The task is aimed at developing a system that identifies the geographical location(country/province) from where an Arabic tweet in the form of modern standard Arabic or dialect comes from. We solve the task in two parts. The first part involves pre-processing the provided dataset by cleaning, adding and segmenting various parts of the text. This is followed by carrying out experiments with different versions of two Transformer based models, AraBERT and AraELECTRA. Our final approach achieved macro F1-scores of 0.216, 0.235, 0.054, and 0.043 in the four subtasks, and we were ranked second in MSA identification subtasks and fourth in DA identification subtasks.

1 Introduction

Spoken by about 500 million people around the world, Arabic is the biggest part of the Semitic language family. Being the official language of almost 22 countries belonging to the Middle-East North Africa (MENA) region, it is not only an integral member of the six official UN languages, but also fourth most used language on the Internet (Guellil et al., 2019). Middle East contributes to 164 million internet users and North Africa contributes to 121 million internet users. Comparing with other languages, Arabic language has received little attention in modern computational linguistics, despite its religious, political and cultural significance. However, with rapid development of tools and techniques delivering state-of-the-art performance in many language processing tasks, this negligence is being taken care of.

The presence of various dialects and complex morphology are some of the distinguishing factors prominent in the Arabic language. Also, the informal nature of conversations on social media and the differences in Modern Standard Arabic (MSA) and Dialectical Arabic (DA), both significantly increase this complexity. While DA is used for informal daily communication, MSA is used for formal writing. Social media is the home for both of these forms, with the former being the most common form. Lack of data is the primary reason why many of the Arabic dialects remain understudied. With the availability of diverse data belonging to 21 Arab countries, this bottleneck can be diminished. The Nuanced Arabic Dialect Identification (NADI), with this goal, is the task of automatic detection of the source variety of a given text or speech segment.

Previously, on the lines of Arabic dialect identification, there have been approaches focusing on coarse-grained regional varieties such as Levantine or Gulf (Elaraby and Abdul-Mageed, 2018; Zaidan and Callison-Burch, 2014; Elfardy and Diab, 2013) or country level varieties (Bouamor et al., 2019; Zhang and Abdul-Mageed, 2019). There have been tasks that involved city level classification on human translated data (Salameh et al., 2018). Some tasks have focused on country and province level classification simultaneously (Abdul-Mageed et al., 2020).

In this paper, we present our process to tackle the WANLP-2021 Shared Task 1. The paper is organised in the following way: Section 2 presents the problem statement and details of the provided dataset. Section 3 describes a modularised process that we inculcate as part of methodology. Section 4 describes the experiments that were conducted, with detailed statistics about the dataset, system settings and results of these experiments. A brief conclusion of the paper with the potential prospects of our study are presented in Section 5.

2 Task Definition

The WANLP-2021 Shared Task 1 (Abdul-Mageed et al., 2021) is based on a multi-class classification problem where the aim is to recognize which country or province an Arabic tweet in the form of modern standard Arabic or dialect belongs to. The task targets dialects at the province-level, and also focuses on naturally-occurring fine-grained dialects at the sub-country level. The NADI 2021 task promotes efforts made towards distinguishing both modern standard Arabic (MSA) and dialects (DA) according to their geographical origin, focusing on fine-grained dialects with new datasets. The provided data comes from the domain of Twitter and covers 100 provinces from 21 Arab countries. The task is divided into 4 subtasks as described below:

Subtask 1.1: Country-level MSA identification
Subtask 1.2: Country-level DA identification
Subtask 2.1: Province-level MSA identification
Subtask 2.2: Province-level DA identification

The training dataset has a total of 21,000 tweet, validation and test datasets have 5,000 tweets each. Every example belongs to one of 100 provinces of 21 Arab countries. Additional 10M unlabeled tweets are provided that can be used in developing the systems for either or both of the tasks. F-score, Accuracy, Precision and Recall are the evaluation metrics. However, the official metric of evaluation is the Macro Averaged F-score.

3 Methodology

We present our methodology in two parts. The first part in the methodology is data preprocessing. This is followed by experimenting with different transformer based models for the task at hand. Both these parts have been described in detail in the following sub sections.

3.1 Data Pre-Processing

Transformer based models, that we plan to fine tune on our dataset, are pre-trained on processed rather than raw data. Owing to the variations in expression of opinions among users belonging to different parts of the world, the tweets fetched from the website are a clear representation of these variations. We find these variations on randomly checking the given examples in different forms. It is common for users to use slang words on the Twitter platform, and post non-ascii characters like emojis. Also,

spelling errors, user mentions and URLs are prominent in tweets of most users. These parts within the tweets do not contribute to being informative towards deciding the geographical location of the tweet as they correspond to noise. Thus, the given dataset is cleaned in the following ways, so that the data used for fine tuning has a similar distribution to that used for the pre-training process:

1. Perform Farasa segmentation (for select models only) (Abdelali et al., 2016).

2. Replace all URLs with [رابط], emails with [بريد], mentions with [مستخدم].

3. Remove HTML line breaks and markup, unwanted characters like emoticons, repeated characters (> 2) and extra spaces.

4. Insert whitespace before and after all non Arabic digits or English Digits and Alphabet and the 2 brackets, and between words and numbers or numbers and words.

3.2 Transformer Based Models

The domains of speech recognition (Graves et al., 2013) and computer vision (Krizhevsky et al., 2012) have largely utilised different deep learning techniques and produced significant improvements over the traditional machine learning techniques. In the domain of natural language processing, most deep learning based techniques until now utilised word vector representations (Bengio et al., 2003; Yih et al., 2011; Mikolov et al., 2013) for different classification tasks. Lately, transformer based approaches have shown significant progress towards many NLP benchmarks (Vaswani et al., 2017), including text classification (Chang et al., 2020), owing to their ability to build proficient language models. As an output of the pre-training process, embeddings are produced which are utilised for finer tasks.

3.2.1 AraBERT

AraBERT is an Arabic pretrained language model based on Google's BERT architecture (Antoun et al.). There are six versions of the model: AraBERTv0.1-base, AraBERTv0.2-base, AraBERTv0.2-large, AraBERTv1-base, AraBERTv2-base and AraBERTv2-large. For these variations, the model parameters with respect to the pre-training process have been depicted in Table 1.

Model	Size		Pre-Segmentation	Dataset		
	MB	Params		#Sentences	Size	#Words
AraBERTv0.2-base	543MB	136M	No	200M	77GB	8.6B
AraBERTv0.2-large	1.38G	371M	No	200M	77GB	8.6B
AraBERTv2-base	543MB	136M	Yes	200M	77GB	8.6B
AraBERTv2-large	1.38G	371M	Yes	200M	77GB	8.6B
AraBERTv0.1-base	543MB	136M	No	77M	23GB	2.7B
AraBERTv1-base	543MB	136M	Yes	77M	23GB	2.7B

Table 1: Model Pre-training Parameters

3.2.2 AraELECTRA

Being a method for self-supervised language representation learning, ELECTRA has the ability of making use of lesser computations for the task of pre-training transformers (Antoun et al., 2020). Similar to the objective of discriminator of a Generative Adversarial Network, ELECTRA models are trained with the goal of distinguishing fake input tokens from the real ones. On the Arabic QA dataset, AraELECTRA achieves state-of-the-art results.

For all new AraBERT and AraELECTRA models, the same pretraining data is used. The dataset that is used for pre-training, before the application of Farasa Segmentation, has a total of 82,232,988,358 characters or 8,655,948,860 words or 200,095,961 lines, and has a size of 77GB. Initially, several websites like OSCAR unshuffled and filtered, Assafir news articles, Arabic Wikipedia dump from 2020/09/01, The OSIAN Corpus and The 1.5B words Arabic Corpus, were crawled to create the pre-training dataset. Later, unshuffled OSCAR corpus, after thorough filtering, was added to the previous dataset used in AraBERTv1 without including the data from the above mentioned crawled websites to create the new dataset.

Country	DA		MSA	
	Train	Dev	Train	Dev
Algeria	1809	430	1899	427
Bahrain	215	52	211	51
Djibouti	215	27	211	52
Egypt	4283	1041	4220	1032
Iraq	2729	664	2719	671
Jordan	429	104	422	103
Kuwait	429	105	422	103
Lebanon	644	157	633	155
Libya	1286	314	1266	310
Mauritania	215	53	211	52
Morocco	858	207	844	207
Oman	1501	355	1477	341
Palestine	428	104	422	102
Qatar	215	52	211	52
Saudi_Arabia	2140	520	2110	510
Somalia	172	49	346	63
Sudan	215	53	211	48
Syria	1287	278	1266	309
Tunisia	859	173	844	170
UAE	642	157	633	154
Yemen	429	105	422	88

Table 2: Country Level Data Distribution

4 Experiments

We experiment with eight transformer based models using the given training and validation sets. We calculate the final test predictions by fine tuning the most efficient model, which is decided by the scores produced above, with the concatenated labeled training and validation splits. This is followed by evaluating the test set on this fine tuned model. This section presents the Country-level dataset distribution, system settings, results of our research followed by a descriptive analysis of our system.

Parameter	Value
Learning Rate	1e-5
Epsilon (Adam optimizer)	1e-8
Maximum Sequence Length	256
Batch Size (for base models)	40
Batch Size (for large models)	4
#Epochs	5

Table 3: Parameter Values

Model	Subtask 1.1		Subtask 1.2		Subtask 2.1		Subtask 2.2	
	F1	A	F1	A	F1	A	F1	A
AraBERTv0.1-base	0.283	0.324	0.338	0.390	0.024	0.028	0.025	0.037
AraBERTv0.2-base	0.300	0.344	0.382	0.427	0.038	0.042	0.035	0.051
AraBERTv0.2-large	0.304	0.343	0.362	0.413	0.022	0.030	0.029	0.041
AraBERTv1-base	0.281	0.318	0.306	0.377	0.032	0.040	0.019	0.033
AraBERTv2-base	0.309	0.347	0.389	0.432	0.029	0.038	0.034	0.048
AraBERTv2-large	0.315	0.346	0.416	0.450	0.001	0.010	0.001	0.010
AraELECTRA-base-generator	0.106	0.231	0.165	0.285	0.005	0.018	0.006	0.022
AraELECTRA-base-discriminator	0.192	0.281	0.280	0.375	0.007	0.020	0.006	0.026

Table 4: Validation Set Results

	M-F1	A	P	R
Subtask 1.1	0.216	0.317	0.321	0.189
Subtask 1.2	0.235	0.433	0.280	0.233
Subtask 2.1	0.054	0.060	0.061	0.060
Subtask 2.2	0.043	0.053	0.044	0.051

Table 5: Test Set Results

4.1 Dataset

The country-wise distribution of the provided training and validation splits, for both the tasks of MSA and DA, are shown in Table 2.

4.2 System Settings

We make use of pre-trained AraBERT and Ara-ELECTRA models, with the names of bert-base-arabert, bert-base-arabertv01, bert-large-arabertv2, bert-base-arabertv2, bert-large-arabertv02, bert-base-arabertv02, araelectra-base-generator and araelectra-base-discriminator for fine-tuning the transformer based models. We use hugging-face[1] API to fetch the pre-trained transformer based models, and then fine tuned the same on our dataset. The hyper parameters used for fine tuning these models have been specified in Table 3.

4.3 Results and Analysis

For all subtasks, the performance results of proposed models on the provided validation set with reference to accuracy(A) and weighted F1 scores(F1) are shown in Table 4.

From Table 4, we conclude that:

1. For most of the subtasks, one of the base models performs almost as good as the best performing large model.

[1] https://huggingface.co/transformers/

2. AraELECTRA models seem to perform worse than all AraBERT models, possibly due to their specialization in handling GAN related tasks, which are different from classification based tasks.

3. AraBERTv2-large out performs all other models for subtasks 1.1 and 1.2. For subtasks 2.1 and 2.2, AraBERTv0.2-base produces the best results on the validation set.

From the above results, we choose AraBERTv2-large for subtasks 1.1, 1.2 and AraBERTv0.2-base for subtasks 2.1, 2.2 to be the primary models to fine tune on the concatenated training and validation set as well as carry out inferences on the unseen dataset. The final test set results in terms of Macro F1 Score(M-F1), Recall(R), Accuracy(A) and Precision(P) are specified in Table 5.

5 Conclusion and Future Work

In this paper, we present a comprehensive overview of the approach that we employed to solve the EACL WANLP-2021 Shared Task 1. We tackle the given problem in two parts. The first part involves pre processing the given data by modifying various parts of the text. The second part involves experimenting with different versions of two Transformer based networks, AraBERT and AraELECTRA, all pre-trained on Arabic text. Our final submissions for the four subtasks are based on the best performing version of AraBERT model. With Macro Averaged F1-Score as the final evaluation criteria, our approach fetches a private leaderboard rank of 2 for MSA identification and 4 for DA identification. In the future, we aim to utilise other features relevant for classification tasks like URLs, emoticons, and experiment with ensembles of transformer based and word vector based input representations.

References

Ahmed Abdelali, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak. 2016. Farasa: A fast and furious segmenter for Arabic. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 11–16, San Diego, California. Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020. 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)*.

Wissam Antoun, Fady Baly, and Hazem Hajj. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Araelectra: Pre-training text discriminators for arabic language understanding.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3(null):1137–1155.

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.

Wei-Cheng Chang, Hsiang-Fu Yu, Kai Zhong, Yiming Yang, and Inderjit Dhillon. 2020. Taming pretrained transformers for extreme multi-label text classification.

Mohamed Elaraby and Muhammad Abdul-Mageed. 2018. Deep models for Arabic dialect identification on benchmarked data. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 263–274, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Heba Elfardy and Mona Diab. 2013. Sentence level dialect identification in Arabic. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 456–461, Sofia, Bulgaria. Association for Computational Linguistics.

Alex Graves, Abdel rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks.

Imane Guellil, Houda Saâdane, Faical Azouaou, Billel Gueni, and Damien Nouvel. 2019. Arabic natural language processing: An overview. *Journal of King Saud University - Computer and Information Sciences*.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, page 1097–1105, Red Hook, NY, USA. Curran Associates Inc.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Wen-tau Yih, Kristina Toutanova, John C. Platt, and Christopher Meek. 2011. Learning discriminative projections for text similarity measures. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 247–256, Portland, Oregon, USA. Association for Computational Linguistics.

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

Chiyu Zhang and Muhammad Abdul-Mageed. 2019. No army, no navy: BERT semi-supervised learning of Arabic dialects. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 279–284, Florence, Italy. Association for Computational Linguistics.