NLP_DI at NADI Shared Task Subtask-1: Sub-word Level Convolutional Neural Models and Pre-trained Binary Classifiers for Dialect Identification

Vani Kanjirangat
IDSIA-USI/SUPSI, Switzerland
vanik@idsia.ch

Tanja Samardzic
URPP Language and Space, UZH
tanja.samardzic@uzh.ch

Ljiljana Dolamic
armasuisse S+T, Switzerland
Ljiljana.Dolamic@armasuisse.ch

Fabio Rinaldi
IDSIA-USI/SUPSI, Switzerland
fabio.rinaldi@idsia.ch

Abstract

In this paper, we describe our systems submitted to the NADI Subtask 1: country-wise dialect classifications. We designed two types of solutions. The first type is convolutional neural network CNN) classifiers trained on subword segments of optimized lengths. The second type is fine-tuned classifiers with BERT-based language specific pre-trained models. To deal with the missing dialects in one of the test sets, we experimented with binary classifiers, analyzing the predicted probability distribution patterns and comparing them with the development set patterns. The better performing approach on the development set was fine-tuning language specific pre-trained model (best F-score 26.59%). On the test set, on the other hand, we obtained the best performance with the CNN model trained on subword tokens obtained with a Unigram model (the best F-score 26.12%). Re-training models on samples of training data simulating missing dialects gave the maximum performance on the test set version with a number of dialects lesser than the training set (F-score 16.44%)

1 Introduction

Arabic Natural Language Processing (NLP) is traditionally faced with the problem of dialect identification. Although Arabic is spoken by a large community of about 400 million people, this community is distributed around different countries and extremely diverse in term of regional linguistic varieties, often called dialects. Modern Standard Arabic (MSA), which is the official language in many Arabic speaking countries is highly formal language used in books and official communication, but newspapers and online writing already show considerable diversification, which is greatly increased in the spoken language of everyday communication. MSA differs from regional varieties lexically, syntactically and phonetically (Zaidan and Callison-Burch, 2014).

In the long history of Arabic Dialect Identification (ADI), multiple datasets have been developed. Some of the most popular datasets include: The ADI VarDial dataset (Zampieri et al., 2017, 2018), which includes Arabic text that is both speech transcribed and transliterated (Malmasi et al., 2016; Ali et al., 2016). Arabic Online Commentary (AOC) is another dataset, which includes a large-scale repository of Arabic dialects obtained from reader commentary of online Arabic newspapers (Zaidan and Callison-Burch, 2011). Multi Arabic Dialect Applications and Resources (MADAR) corpus constitutes parallel sentences written in different Arabic city dialects from travel domain (Bouamor et al., 2019).

Classification methods tried out on these datasets range from feature-based machine learning approaches (Touileb, 2020; Younes et al., 2020; AlShenaifi and Azmi, 2020; Harrat et al., 2019), n-gram based language models (Çöltekin et al., 2018; Butnaru and Ionescu, 2018) and ensemble models El Mekki et al. (2020) to neural and pre-trained models (AlKhamissi et al., 2021; El Mekki et al., 2021; Elaraby and Abdul-Mageed, 2018; Ali, 2018).

In this paper, we describe the solutions submitted by our team to the Nuanced Arabic Dialect Identification (NADI) shared task 2022 (Abdul-Mageed et al., 2022), subtask-1, which targets a more fine-grained classification than in previous tasks. The NADI shared task focuses on the study and analysis of Arabic dialects at country-level, province-level and city-level. NADI 2020 (Abdul-Mageed et al., 2020) and 2021 (Abdul-Mageed et al., 2021) tasks focused on dialects across 21 Arab countries and 100 provinces.

This paper is organized as follows: The data

Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP), pages 468 - 473
December 8, 2022 ©2022 Association for Computational Linguistics

Models	Fscore(%)	Accuracy(%)
Unigram_CNN	17.06	32.45
BPE_CNN	17.17	33.97
AraBERT	21.38	37.54
Multi-dialect-Arabic-BERT	26.59	42.61

Table 1: Evaluation results on development set

	Average Positive Probabilities				
Dialect	**TEST-B**	**DEV1**	**DEV2**	**DEV3**	**DEV4**
Bahrain	0.8995	0.8907	0.8905	0.8965	0.8973
Jordan	0.8888	0.9053	0.9041	0.9146	0.9097
Lebanon	0.8557	0.8588	0.8622	0.8576	0.8605
Qatar	0.8984	0.8798	0.8788	0.8811	0.8837
UAE	0.9244	0.9009	0.9023	0.9019	0.9040
Oman	0.9203	0.9219	0.9219	0.9194	0.9228
Algeria	0.7978	0.8806	0.8588	0.8825	0.8836
Egypt	0.9432	0.9447	0.9456	0.9076	0.9496
Libya	0.8973	0.9185	0.9168	0.9215	0.9105
Palestine	0.8990	0.9086	0.9080	0.9227	0.9072
Tunisia	0.8589	0.9162	0.9141	0.9080	0.9185
Syria	0.8840	0.8969	0.8944	0.9020	0.8973
Morocco	0.8417	0.8735	0.8626	0.8767	0.8751
KSA	0.9408	0.916	0.9166	0.9205	0.9227
Yemen	0.8793	0.8899	0.8889	0.8991	0.8918
Kuwait	0.9459	0.9297	0.9296	0.9329	0.9299
Iraq	0.8652	0.8896	0.8857	0.8899	0.8623
Sudan	0.8276	0.8931	0.8990	0.9128	0.8975

Table 2: Comparing the average positive predicted probabilities for each binary classifier on simulated development set and TEST-B. The possible missing dialects identified by our approach are bolded.

statistics is described in Section 2, methods used are discussed in Section 3, experimental results are reported in Section 5, followed by conclusions in Section 6.

2 Data

The subtask 1 of NADI 2022 provides training and development sets with 18 country dialects. The training set constitutes 20,398 instances and development set 4871 instances. In the evaluation phase, two test sets were provided, TEST-A with 4871 instances and TEST-B with 1474 instances. TEST-A had all the 18 dialects as in the training set, while TEST-B had *k* missing dialects, where *k* < 18.

3 Models and Methods

We tried two kinds of solutions described in the following subsections.

3.1 Approach 1: Sub-word Level Convolution Neural Network

In our first solution, we train from scratch a Convolution Neural Network (CNN) on subword tokens produced with different algorithms. The CNN is an adapted version of the architecture proposed by Kim et al. (2016). This architecture is originally used for building a neural language model (NLM). To use this architecture for dialect classification, we take the CNN encoder part substitute the decoder part with dense and softmax layers. We used the CNN filter sizes as proposed by Kim et al. (2016). In general, the filter size can be seen as the length of n-grams and hence using different filters helps to capture text units of different spans.

To decide the optimal splits for input subword tokenization, we tune on the development set the vocabulary size (vocab_size) of two subword tokenization algorithms from the SentencePiece[1] library: the Unigram model and Byte Pair Encoding (BPE). We experimented with gradually increasing vocab_size, ranging from the character vocab_size to $0.4 * |V|$ following Mielke et al. (2019), where $|V|$ is the word-level vocabulary size, and kept the one which gave the best performance on the development set. The optimal vocabulary size turned out to be 20,045 for Unigram model and 7,045 for BPE.

3.2 Approach 2: Pre-trained Models

Our second solution makes use of pre-trained models, specifically BERT-based (Devlin et al., 2019) language-specific models. We used AraBERT (Antoun et al.)[2] and Multi-dialect-Arabic-BERT (Talafha et al., 2020)[3] models for our experiments. AraBERT is a BERT-based model, pre-trained additionally with Arabic articles from Wikipedia, OSCAR[4] and OSIAN corpus (Zeroual et al., 2019). Multi-dialect-Arabic-BERT model is initialized with the weights of Arabic-BERT model[5] and further trained on the 10M unlabelled tweets provided by NADI shared task. For loading and fine-tuning the pretrained models, we used the HuggingFace[6]

[1] https://github.com/google/sentencepiece

[2] https://huggingface.co/aubmindlab/bert-base-arabert

[3] https://huggingface.co/bashar-talafha/multi-dialect-bert-base-arabic

[4] https://oscar-corpus.com/

[5] https://huggingface.co/asafaya/bert-base-arabic

[6] https://huggingface.co/

transformer library and followed BERT single sentence classification pipeline.

For TEST-A, we used the fine-tuned AraBERT and Multi-dialect-Arabic-BERT models directly for the predictions. In TEST-B, we did additional adaptations, specifically to deal with the unknown or missing dialect(s) (described in Subsections 4.1 and 4.2).

4 Adaptation to Unknown Set of Dialects (TEST-B)

To deal with the missing dialects in TEST-B, we apply two additional techniques to the Multi-dialect-Arabic-BERT model as the baseline. These techniques are described in the remainder of this subsection.

4.1 Label Smoothing

Label smoothing helps to alleviate overfitting problem (Müller et al., 2019) and is used as an effective regularization technique in neural models. We used label smoothing (LS) with a specific α (hyperparameter) for fine-tuning the pre-trained model.

4.2 Binary Classifiers

In order to identify the possible missing dialects, we train binary classifiers, one for each dialect in the training set. Given an input sentence, we pass it through each of the 18 classifiers to identify whether the sentence belongs to the particular dialect class/not. For instance, if the classifier is for dialect *Egypt*, then it predicts whether the sentence dialect is Egypt/Not.

Uneven distribution of training data across dialects has a strong impact on models in such binary classification set-up causing strong preferences for some classes. To deal with this issue, we sample balanced datasets for each dialect class. For this, we label all the instances belonging to the particular dialect class as 1 and sample equal number of instances from the remaining classes in the training set without replacement and label it as 0. This helped in boosting the performance for some classes.

In an ideal situation, we expect that for a particular sentence input, only one of the 18 classifiers predicts 1, which means the sentence belongs to the respective dialect class. Further in the ideal scenario, for any sentence input, the missing dialects (in TEST-B) should not be predicted. But, since these country dialects are closely related and overlapping, misclassifications can occur quite often. To tackle this, we need to devise some approach to decide a threshold or some pattern that can help us in deciding the possible missing dialects.

To set the threshold for missing dialects, we simulate TEST-B conditions on the development set. We randomly removed some dialect classes from the development set and performed the evaluations. We performed multiple simulations and recorded the average correct prediction probabilities for each dialect class. We repeated the same for TEST-B. We then analyzed the probability distribution patterns and compared the average probabilities of each dialect from TEST-B with the simulated development sets. Further, we observed the change/ difference in probabilities and identified those dialects with an evident drop in average probabilities. The probabilities for four simulated development sets are tabulated in Table 2 with the missing dialects as: *DEV1: {'palestine', 'yemen', 'lebanon'}*, *DEV2: {'yemen', 'algeria', 'syria'}*, *DEV3: {'egypt', 'tunisia', 'morocco'}* and *DEV4: {'sudan', 'libya', 'iraq'}*. Based on these observations, we selected five dialects: {'Algeria', 'Tunisia', 'Morocco', 'Iraq' and 'Sudan'} as the missing dialects and retrained the Multi-dialect-Arabic-BERT model by removing these five dialects from the training set.

5 Experimental Settings and Results

The results obtained on the development set are reported in Table 1. The F-scores obtained with pretrained models (AraBERT 21.38% and Multi-dialect-Arabic-BERT 26.59%) is considerably higher than those obtained with the CNN models (Unigram_CNN 17.06% and BPE_CNN 17.17%).

Table 3 shows the official evaluation of our models on two test sets provided by the organizers. In TEST-A (with all the 18 dialects), we used the four models: *Unigram_CNN*, *BPE_CNN*, *AraBERT* and *Multi- dialect-Arabic-BERT*. In TEST-B (with missing dialects), we submitted five models: *Unigram_CNN*, *BPE_CNN*, *Multi- dialect-Arabic-BERT*, *Multi- dialect-Arabic-BERT_LS* (Multi- dialect-Arabic-BERT with Label Smoothing with α = 0.1) and *Binary classifiers + Multi- dialect-Arabic-BERT* (Binary classifiers with Multi- dialect-Arabic-BERT). In Binary classifiers + Multi- dialect-Arabic-BERT, we use the binary classifier approach as discussed in Section

Test Set	Models	Fscore (%)	Accuracy (%)
TEST-A	Unigram_CNN	16.18	31.39
	BPE_CNN	16.66	33.50
	AraBERT	19.99	36.65
	Multi-dialect-Arabic-BERT	26.12	42.07
TEST-B	Unigram_CNN	8.71	18.59
	BPE_CNN	7.58	19.34
	Multi-dialect-Arabic-BERT	13.47	27.88
	Multi-dialect-Arabic-BERT_LS	13.75	27.88
	Binary classifier + Multi-dialect-Arabic-BERT	16.44	27.68

Table 3: Official evaluation results on test set

4.2 for identifying the possible missing dialects and further retraining the model.

It can be observed that the best performance on TEST-A was achieved with Multi-dialect-Arabic-BERT model. On TEST-B, pretrained models work better with the best result achieved in the last setting (Binary classifiers + Multi- dialect-Arabic-BERT model).

Now, we discuss briefly the different outcomes on the two test sets. In both test sets, the best results are obtained by language specific pre-trained models. In TEST-B, all the scores are higher and the results with pretrained models are much better. We believe that this difference can be attributed to two factors. First, the smaller number of classes seems to make the task easier for all the models. Second, our adaptation techniques are better suited to the setting with pretrained models. Label smoothing helped in improving the performance slightly (Multi- dialect-Arabic-BERT_LS) and binary classifiers with model retraining brings additional improvement.

Overall, based on the official results, we achieved a F-score of 21.28%.

6 Conclusion

In this paper, we described and discussed two kinds of solutions for the NADI shared task, subtask 2: automatic country-wise identification of Arabic dialects. Among the solutions that we submitted, the language specific pre-traiend models gave the best performance in both TEST-A and TEST-B. Label smoothing and simulating the missing dialect scenario with binary classifiers were our techniques for TEST-B with unknown set of labels. These techniques improve the performance compared to the baseline setting. In TEST-B, adaptation techniques enable better performance on this set, but there is still a lot of room for improving the performance.

In future work, we aim to pursue the development of CNN architectures for fine-grained discrimination. We plan to investigate self-attention mechanisms with CNN and unsupervised deep embedding clustering.

References

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*, pages 97–110.

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*, pages 244–259.

Muhammad Abdul-Mageed, Chiyu Zhang, Abdel-Rahim Elmadany, Houda Bouamor, and Nizar Habash. 2022. NADI 2022: The Third Nuanced Arabic Dialect Identification Shared Task. In *Proceedings of the Seven Arabic Natural Language Processing Workshop (WANLP 2022)*.

Ahmed Ali, Najim Dehak, Patrick Cardinal, Sameer Khurana, Sree Harsha Yella, James Glass, Peter Bell, and Steve Renals. 2016. Automatic dialect detection in arabic broadcast speech.

Mohamed Ali. 2018. Character level convolutional neural network for arabic dialect identification. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 122–127.

Badr AlKhamissi, Mohamed Gabr, Muhammad El-Nokrashy, and Khaled Essam. 2021. Adapting marbert for improved arabic dialect identification: Submission to the nadi 2021 shared task. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 260–264.

Nouf AlShenaifi and Aqil Azmi. 2020. Faheem at nadi shared task: Identifying the dialect of arabic tweet. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 282–287.

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.

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.

Andrei Butnaru and Radu Tudor Ionescu. 2018. Unibuckernel reloaded: First place in arabic dialect identification for the second year in a row. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 77–87.

Çağrı Çöltekin, Taraka Rama, and Verena Blaschke. 2018. Tübingen-oslo team at the vardial 2018 evaluation campaign: An analysis of n-gram features in language variety identification. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 55–65.

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.

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.

Abdellah El Mekki, Abdelkader El Mahdaouy, Kabil Essefar, Nabil El Mamoun, Ismail Berrada, and Ahmed Khoumsi. 2021. Bert-based multi-task model for country and province level msa and dialectal arabic identification. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 271–275.

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.

Salima Harrat, Karima Meftouh, Karima Abidi, and Kamel Smaïli. 2019. Automatic identification methods on a corpus of twenty five fine-grained arabic dialects. In *International Conference on Arabic Language Processing*, pages 79–92. Springer.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. 2016. Character-aware neural language models. In *Thirtieth AAAI conference on artificial intelligence*.

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.

Sabrina J Mielke, Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner. 2019. What kind of language is hard to language-model? *arXiv preprint arXiv:1906.04726*.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help? *Advances in Neural Information Processing Systems*, 32:4694–4703.

Bashar Talafha, Mohammad Ali, Muhy Eddin Za'ter, Haitham Seelawi, Ibraheem Tuffaha, Mostafa Samir, Wael Farhan, and Hussein Al-Natsheh. 2020. Multi-dialect arabic bert for country-level dialect identification. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 111–118.

Samia Touileb. 2020. Ltg-st at nadi shared task 1: Arabic dialect identification using a stacking classifier. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 313–319.

Mutaz Younes, Nour Al-Khdour, and AL-Smadi Mohammad. 2020. Team alexa at nadi shared task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 237–242.

Omar 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.

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

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the vardial evaluation campaign 2017. In *Proceedings of the fourth workshop on NLP for similar languages, varieties and dialects*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Ahmed Ali, Suwon Shon, James Glass, Yves Scherrer, Tanja Samardžić, Nikola Ljubešić, Jörg Tiedemann, et al. 2018. Language identification and morphosyntactic tagging. the second vardial evaluation campaign.

Imad Zeroual, Dirk Goldhahn, Thomas Eckart, and Abdelhak Lakhouaja. 2019. Osian: Open source international arabic news corpus-preparation and integration into the clarin-infrastructure. In *Proceedings of the fourth arabic natural language processing workshop*, pages 175–182.