Team JUST at the MADAR Shared Task on Arabic Fine-Grained Dialect Identification

Bashar Talafha
Jordan University of Science and Technology, Jordan
talafha@live.com

Ali Fadel
Jordan University of Science and Technology, Jordan
aliosm1997@gmail.com

Mahmoud Al-Ayyoub
Jordan University of Science and Technology, Jordan
malayyoub@gmail.com

Yaser Jararweh
Duquesne University, USA
jararwehy@duq.edu

Mohammad AL-Smadi
Jordan University of Science and Technology, Jordan
maalsmadi9@just.edu.jo

Patrick Juola
Duquesne University, USA
juola@mathcs.duq.edu

Abstract

In this paper, we describe our team's effort on the MADAR Shared Task on Arabic Fine-Grained Dialect Identification. The task requires building a system capable of differentiating between 25 different Arabic dialects in addition to MSA. Our approach is simple. After preprocessing the data, we use Data Augmentation (DA) to enlarge the training data six times. We then build a language model and extract n-gram word-level and character-level TF-IDF features and feed them into an MNB classifier. Despite its simplicity, the resulting model performs really well producing the 4th highest F-measure and region-level accuracy and the 5th highest precision, recall, city-level accuracy and country-level accuracy among the participating teams.

1 Introduction

Give a piece of text, the Dialect Identification (DI) is concerned with automatically determining the dialect in which it is written. This is a very important problem in many languages including Arabic. Unlike previous works on Arabic DI (ADI), which take a coarse-grained approach by considering regional-level (Zaidan and Callison-Burch, 2014; Elfardy and Diab, 2013; Zampieri et al., 2018) or country-level (Sadat et al., 2014) dialects, a new task has been proposed for the fine-grained ADI focusing on a large number of city-/country-level dialects (Bouamor et al., 2019).

This task is quite challenging as it covers 25 different dialects in addition to Modern Standard Arabic (MSA). Some of these dialects are very close to each other as we observe in our analysis of the training data (see Section 2). Also, due to the relatively small size of the dataset, cutting-edge techniques for document/sentence classification, which are based on word embeddings and deep learning models, perform poorly on it. In fact, according to (Bouamor et al., 2019), the top performing systems for this task as well as the previously published baseline (Salameh et al., 2018) all use traditional (non-neural) machine learning approaches. This is very surprising if one takes into account that the use of Deep Learning in Arabic NLP is still at its early stages (Al-Ayyoub et al., 2018).

In this paper, we describe our team's effort to tackle this task. After preprocessing the data, we use Data Augmentation (DA) to enlarge the training data six times. We then build a language model and extract n-gram word-level and character-level TF-IDF features and feed them into a Multinomial Naive Bayes (MNB) classifier. Despite its simplicity, the resulting model performs really well producing the 4th highest Macro-F1 measure (66.33%) and Region-level Accuracy (84.54%) and the 5th highest Macro-Precision (66.56%), Macro-Recall (66.42%), City-level Accuracy (66.42%) and Country-level Accuracy (74.71%) among the participating teams. Unfortunately, due to a problem with our submission file, the official results for our system were extremely poor, which placed our team at the bottom of the official ranking.

The rest of this paper is organized as follows. In Section 2, we discuss the task at hand while analyzing the provided data. In Section 3, we describe our system and its details while, in Section 4, we present and analyze its results and performance. Finally, the paper is concluded in Section 5.

Proceedings of the Fourth Arabic Natural Language Processing Workshop, pages 285–289
Florence, Italy, August 1, 2019.

2 MADAR Task, Dataset and Metrics

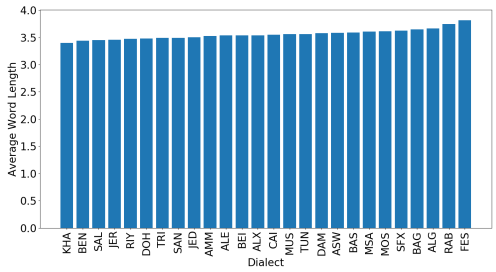
The shared task at hand comprises of two subtasks. The first one is the Travel Domain ADI, whose data are taken from Multi-Arabic Dialect Applications and Resources (MADAR) project (Bouamor et al., 2018). Our team only focused on this subtask. The second subtask is the Twitter User ADI and it is outside the scope of this work.

For the subtask at hand, the organizers provide three sets: train (stored in a file called `MADAR-Corpus-26-train` and we refer to it as Corpus-26), development (dev) and test. The train, dev and test sets consist of 41,600, 5,200 and 5,200 parallel sentences, respectively, written in MSA as well as the local dialect of 25 cities: Rabat (RAB), Fes (FES), Algiers (ALG), Tunis (TUN), Sfax (SFX), Tripoli (TRI), Benghazi (BEN), Cairo (CAI), Alexandria (ALX), Aswan (ASW), Khartoum (KHA), Jerusalem (JER), Amman (AMM), Salt (SAL), Beirut (BEI), Damascus (DAM), Aleppo (ALE), Mosul (MOS), Baghdad (BAG), Basra (BAS), Doha (DOH), Muscat (MUS), Riyadh (RIY), Jeddah (JED) and Sana'a (SAN).

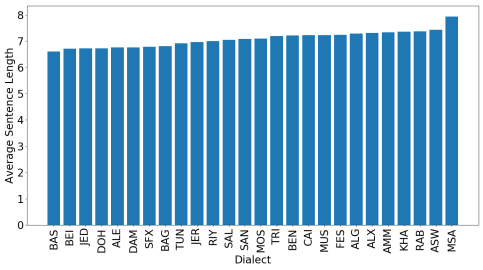
To aid in the training and model building processes, the organizers also provide additional train & dev data sets consisting of 54,000 and 6,000 parallel sentences covering only six dialects: BEI, CAI, DOH, MSA, RAB and TUN. The additional train set is stored in a file called `MADAR-Corpus-6-train` and we refer to it as Corpus-6.

Before we go into the details of our system, we present a simple analysis of the provided data. Figure 1 shows that the sentences of the dialects do not differ much in terms of average word/sentence lengths per dialect (Figures 1(a) and 1(b)) or the number of unique words per dialect (Figures 1(c)). Our analysis shows that while there are 27,501 unique words in all dialects, there is a small number of words (specifically, 84 words) common in all dialects. Examples of such words include: اليابانية، الشارع، المجرم، جوليا، فرانسيسكو، شهر، جولة، البريد، مفتاح، بعيد.

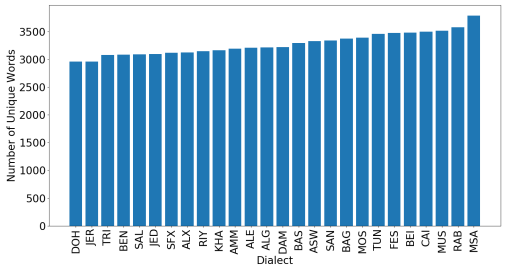
Now, the most interesting part in our analysis is the varying similarity between the different dialects pairs under consideration. Overall, there are 7,280 common sentences between dialects pairs and the average number of common sentences between dialects pairs, on average there is 22.4 common sentences between any dialects pairs. Another relevant observation is the repetition of sentences across different dialects pairs, which is not limited to the dialects from the same country or region. For example, the dialect pairs with 100 or more common sentences are: AMM-JER, DAM-ALE, JER-SAL, AMM-SAL, DAM-JER & AMM-DAM, whereas, the pairs with less than 5 common sentences are: BEI-FES, MSA-BEI, MSA-MOS, MSA-SFX, MSA-TRI, MSA-TUN, RAB-ASW, RAB-KHA, RAB-RIY, RAB-SAN, RAB-BAS & RAB-MOS. Below, we list all dialects under consideration grouped per country and per region. We also list in the parentheses the average number of common sentences within each country (with more than one dialect) and each region.



(a) Average word length per dialect



(b) Average sentence length per dialect



(c) Number of unique words per dialect

Figure 1: Corpus-26 statistics.

1. Maghreb (18.29): Morroco: RAB & FES (50); Algeria: ALG; Tunisia: TUN & SFX (52); Libya: TRI & BEN (66).

2. Nile Basin (42.67): Egypt: CAI, ALX & ASW (67); Sudan: KHA.

3. Levant (88.4): Palestine: JER; Jordan: AMM & SAL (101); Lebanon: BEI; Syria: DAM & ALE (129).

4. Gulf (42.52): Iraq: MOS, BAG & BAS (54.33); Qatar: DOH; Oman: MUS; Saudi: RIY & JED (72.0);

5. Gulf of Aden: Yemen: SAN.

6. MSA.

This list shows that Levant dialects are the most similar while the Maghrib ones are the least similar.

Finally, to evaluate the participating systems, the subtask organizers use Accuracy (on the city, country and region levels denoted here by Acc_{cty}, Acc_{cntr} and Acc_{rgn}, respectively) in addition to Macro-averaged Precision, Recall and F1 measure (denoted here by Pre, Rec and F1, respectively).

3 System

In this section. we describe the system that produces the highest accuracy on the dev set starting from the preprocessing stage all the way up to the final classification stage.

Preprocessing and Data Augmentation (DA). Our system starts with a couple of preprocessing steps. The first one is a very simple one in which quotation marks, Arabic quotation marks, commas, Arabic commas, question marks, Arabic question marks and emoticons are replaced with spaces.

Another preprocessing step the system performs is DA. While DA has been shown to be very effective for image processing tasks (Chatfield et al., 2014; He et al., 2016; Chollet, 2016; Ebrahim et al., 2018), it use in text processing tasks is still limited (Fadaee et al., 2017; Kafle et al., 2017). Since the training data is small, a data augmentation step is performed on Corpus-26 by applying random shuffling on Corpus. In Corpus-26, there are 1,600 sentences for each dialect, while, in Corpus-6, there are 9,000 sentences for each of the six dialects in this corpus: BEI, CAI, DOH, MSA, RAB and TUN. The system takes 8,000 sentences (instead of 9,000) for each dialects in order to balance them with the other dialects (shuffled). Therefore, overall, we have 8,000 sentences (from Corpus-6) + 1,600 sentences (from Corpus-26) = 9,600 sentences for each of these six dialects. For the remaining dialects, and since the order of words is not necessary to identify the dialect, we apply a random shuffling to generate five new sentences from each sentence by using different random seed for each generated sentence. So, for each of these 20 dialects, we have $1,600 \times 6 = 9,600$ sentences. To sum up, the training data has a total of 249,600 sentences; 9,600 sentences for each of the 26 dialects under consideration.

Features Extraction. For each dialect, a language model is extracted using Kenlm[1] with its default parameters using the training data (Corpus-26). For each sentence, we extract a vector of size 26 that represents a language model probability for each dialect. We also extract a word-level Term Frequency-Inverse Document Frequency (TF-IDF) features ranging from unigram to 6-gram in addition to character-level n-grams TF-IDF features where n ranges from 1-gram to 5-grams.

Classifier. An MNB classifier with $\alpha = 0.5$ is applied using the One-vs-the-rest strategy. It is worth mentioning that we experiment with several deep learning-based classifiers such as Convolutional Neural Networks (CNN) (Kim, 2014), Recurrent Neural Networks (RNN) with Long Short-Term Memory (LSTM) cells,[2] Separable Convolutional Network (sepCNN) (Chollet, 2017), Doc2Vec-FFNN,[3] Transformer (Vaswani et al., 2017) and Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018). However, none of them performed well on the validation set. So, we did not submit their results.

4 Results and Discussion

In this section, we present and analyze the results and performance of our best model. Nothing is mentioned about the other models with which we experimented. The results of the model on the test set are presented in Table 1. The table shows that,

[1] https://github.com/kpu/kenlm

[2] https://bit.ly/2K3lNFM

[3] We train a Doc2Vec model (Le and Mikolov, 2014) and extract the feature vectors from it for each sentence. We then feed these vectors into a feed-forward neural network (FFNN) to classify the sentence as one of 26 classes.

	Ours	Top System	Base-line	Overall Comparison
F1	66.33	67.32	67.89	4th highest
Pre	66.56	67.73	68.41	5th highest (tied)
Rec	66.42	67.33	67.75	5th highest
Acc_{cty}	66.42	67.33	67.75	5th highest
Acc_{cntr}	74.71	75.69	76.44	5th highest
Acc_{rgn}	84.54	85.13	85.96	4th highest

Table 1: The results of our model on the test set compared with the other models.

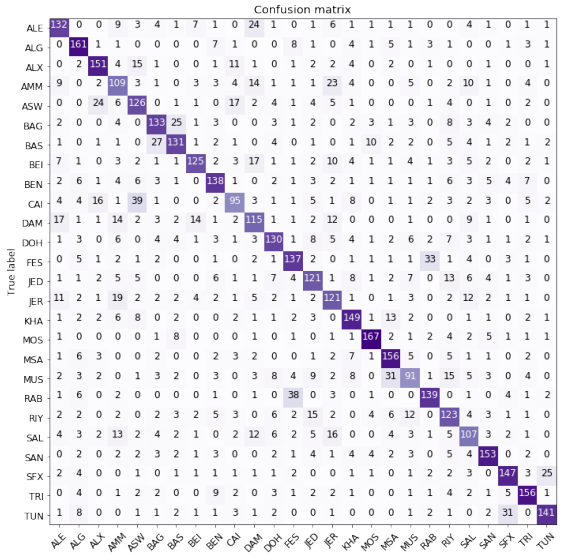


Figure 2: Our model's confusion matrix for the test set.

despite our models' simplicity, its results (which range between 4th highest and 5th highest numbers) are surprisingly good. It differs only by a small number from the top system.

To understand the strengths and weaknesses of our model, we analyze the confusion matrix for the test set (shown in Figure 2). The figure shows that the model suffers while trying to differentiate between similar dialects. For example, 39 test samples from CAI are labeled as ASW and 38 from RAB are labeled as FES. Moreover, among the hardest to classify is CAI, perhaps, due to its high similarity with many dialects. After all, CAI is among the most well-known Arabic and Egyptian dialects due to the cultural influence of Cairo and Egypt on the Arab world, which means that other dialects (especially Egyptian ones) might have been influenced by it. On the other hand, ALG and MOS are among the easiest to classify due to their low similarity with the dialects under consideration.

	F1	Pre	Rec	Acc_{cty}
w/ DA	67.51	69.28	67.29	67.29
w/o DA	66.83	68.69	66.6	66.6

Table 2: Effect of DA

In order to show the effect of DA, we perform an ablation study using the dev set. Table 2 shows the results of this experiment. The results show that DA had a slight effect on improving the performance of the proposed model. Perhaps, this is due to the generative nature of the MNB classifier and its assumption of independence between the features. In the future, we plan on focusing more on DA techniques and their application with neural models, where the intuition is that such models make better use of any additional data in order to learn new things.

5 Conclusion

In this paper, we presented a simple model for the fine-grained ADI subtask. The model's performance was good producing results competitive with the top system for the task. In the future, we plan on exploring approaches based on better DA techniques in addition to the concepts of transfer learning and semi-supervised learning (Talafha and Al-Ayyoub, 2019) in order to obtained better results.

Acknowledgment

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

Mahmoud Al-Ayyoub, Aya Nuseir, Kholoud Alsmearat, Yaser Jararweh, and Brij Gupta. 2018. Deep learning for arabic nlp: A survey. *Journal of computational science*, 26:522–531.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, et al. 2018. The madar arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

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 (WANLP19)*, Florence, Italy.

Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*.

Francois Chollet. 2016. Building powerful image classification models using very little data. The Keras Blog. Https://blog.keras.io/building-powerful-image-classification-models-using-very-little-data.html.

François Chollet. 2017. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Maad Ebrahim, Mohammad Alsmirat, and Mahmoud Al-Ayyoub. 2018. Performance study of augmentation techniques for hep2 cnn classification. In *2018 9th International Conference on Information and Communication Systems (ICICS)*, pages 163–168. IEEE.

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)*, volume 2, pages 456–461.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. *arXiv preprint arXiv:1705.00440*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Kushal Kafle, Mohammed Yousefhussien, and Christopher Kanan. 2017. Data augmentation for visual question answering. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 198–202.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196.

Fatiha Sadat, Farnazeh Kazemi, and Atefeh Farzindar. 2014. Automatic identification of arabic dialects in social media. In *Proceedings of the first international workshop on Social media retrieval and analysis*, pages 35–40. ACM.

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

Bashar Talafha and Mahmoud Al-Ayyoub. 2019. Iohrcnn: Pursue the ingredients of happiness using recurrent convolutional neural networks. In *Proceedings of the 2nd Workshop on Affective Content Analysis (AffCon 2019) co-located with Thirty-Third AAAI Conference on Artificial Intelligence (AAAI 2019), Honolulu, USA, January 27, 2019.*, pages 191–197.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

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

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. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects*. Association for Computational Linguistics.