The QMUL/HRBDT contribution to the NADI Arabic Dialect Identification Shared Task

Abdulrahman Aloraini
Queen Mary University of London
United Kingdom

Ayman Alhelbawy
University of Essex
United Kingdom

Massimo Poesio
Queen Mary University of London
United Kingdom

{a.aloraini, m.poesio}@qmul.ac.uk
a.alhelbawy@essex.ac.uk

Abstract

We present the Arabic dialect identification system that we used for the country-level subtask of the NADI challenge. Our model consists of three components: BiLSTM-CNN, character-level TF-IDF, and topic modeling features. We represent each tweet using these features and feed them into a deep neural network. We then add an effective heuristic that improves the overall performance. We achieved an F1-Macro score of 20.77% and an accuracy of 34.32% on the test set. The model was also evaluated on the Arabic Online Commentary dataset, achieving results better than the state-of-the-art.

1 Introduction

Arabic is widely spoken in the Middle East and certain parts of Africa–21 countries–and the total number of its speakers is approximately 420 million. It is also one of the six official languages of the United Nations. But Arabic is a general term that can refer to classical Arabic (CA), Modern Standard Arabic (MSA), or several Arabic dialects (ADs). Both classical Arabic and MSA are standardized, while Arabic dialects are not. So identifying different Arabic dialect varieties is quite a challenging task. Thus, many studies have been devoted to Arabic dialect identification, because it benefits automatic speech recognition, remote access, e-health, and other applications (Etman and Beex, 2015).

The main challenge for Arabic dialect identification is the lack of large dataset that can be exploited in computational models. Among the studies addressing this issue, Zaidan and Callison-Burch (2014) created a large dataset for this purpose, called Arabic Online Commentary (AOC). The data consists of texts in MSA and Arabic regional dialects collected from Arabic news sites. Zaghouani and Charfi (2018) collected and annotated Twitter data from 11 regions and 16 countries in the Arab world. Abdul-Mageed et al. (2018) also collected Twitter data and annotated them at the city-level i.e 29 cities of 11 countries. Bouamor et al. (2018) translated the Basic Traveling Expression Corpus (BTEC) (Takezawa et al., 2007) into 25 city dialects of Arab countries; this corpus is referred to as Corpus-25.

Anong the models, some proposals only aim to separating one dialect (e.g., Egyptian) from MSA on the AOC dataset (Elfardy and Diab, 2013; Tillmann et al., 2014). Darwish et al. (2014) also distinguished between Egyptian and MSA, but focusing on twitter data. Zaidan and Callison-Burch (2014) proposed a model to identify MSA and regional dialects including Egyptian, Levantine, and Gulf. Huang (2015) applied a word-level n-gram model to identify MSA and regional dialects, and also considered Facebook posts . Elaraby and Abdul-Mageed (2018) used the AOC dataset to evaluate several machine learning and deep learning models. Salameh et al. (2018) proposed a fine-grained dialect identification module for Corpus-25 by applying a multinomial Naive Bayes classifier with a large set of features.

The Nuanced Arabic Dialect Identification (NADI) shared task aims to incentivise research on identifying different Arabic dialects in every Arab country (Abdul-Mageed et al., 2020). The NADI dataset was collected from Arabic twitter, and each tweet is labeled with two labels, country-label and province-label, for a total of 21 countries and 100 of their provinces. There are two subtasks in NADI: the first is targeting these country-level labels and the second one is targeting the provinces. In the rest of the paper,

Proceedings of the Fifth Arabic Natural Language Processing Workshop, pages 295–301
Barcelona, Spain (Online), December 12, 2020

Country	Train	Dev	Country	Train	Dev	Country	Train	Dev
Egypt	4,473	1,070	UAE	1,070	265	Kuwait	420	70
Iraq	2,556	636	Syria	1,070	265	Qatar	234	104
Saudi Arabia	2,312	579	Yemen	851	206	Mauritania	210	40
Algeria	1,491	359	Tunisia	750	164	Bahrain	210	8
Oman	1,098	249	Lebanon	639	110	Djibouti	210	10
Morocco	1,070	249	Jordan	429	104	Somalia	210	51
Libya	1,070	265	Palestine	420	102	Sudan	210	51

Table 1: Counts of country-level labels in the train and development (Dev) of NADI.

we firt introduce the datasets in Section 2. Our Arabic identification model is discussed in Section 3. We show the results on NADI subtask 1 and AOC in Section 4 and discuss them in Section 5. We conclude with Section 6.

2 Data

The NADI dataset was the main corpus used for training and testing the system. The dataset is partitioned into train, development, and test sets, containing 21,000, 4,957, and 5,000 tweets respectively. The test set was published unlabeled, and the system output was evaluated by the NADI shared task team. The training set was used to train our model while the development set was used to optimize model parameters. In addition, NADI also provided 10 million unlabeled tweets which we used to train a word embedding model for Arabic tweets. Table 1 shows the number of the annotated Arabic tweets for each country in training and development sets. The data samples distribution over the dialectal classes or countries is unbalanced.

Our model was also evaluated against another dataset, the Arabic Online Commentary (AOC). A subset of the data was annotated using crowd-sourcing and has been used in previous work (Zaidan and Callison-Burch, 2014; Cotterell and Callison-Burch, 2014). The AOC dataset classifies Arabic dialects into three dialects in addition to the MSA, so spoken dialects in different countries may be grouped together as one dialect. For example, the "Gulf" dialect label may includes all spoken dialects in Saudi Arabia, United Arab Emirates, Kuwait, Bahrain, Iraq, etc. (Elaraby and Abdul-Mageed, 2018) benchmarked a portion of AOC, and applied various machine learning algorithms to identify MSA and dialects based on three settings:

- Binary: where they classify the data to MSA and or not MSA.
- Three-way: where they classify three dialectal regions (Egyptian, Gulf, and Levantine).
- Four-way: where they classify between the three regional dialects and MSA.

The dataset statistics are shown in Table 2.

	MSA	Egyptian	Gulf	Levantine	Total
Train	50,845	10,022	16,593	9,081	86,541
Dev	6,357	1,253	2,075	1,136	10,821
Test	6,353	1,252	2,073	1,133	10,812

Table 2: AOC portions of MSA and region dialects from (Elaraby and Abdul-Mageed, 2018)

3 System

3.1 Tweet Text Prepossessing

Arabic tweets are very noisy, so we removed URLS, emojis, Latin-characters, numbers, mentions, and any non-Arabic characters. Arabic hashtags were kept as they are because they might contain important

information such as (Lebanon revolts, لبنان ينتفض).

Text normalisation was carried out to normalise different forms of "Alif", "Yaa", removing punctuation, excessive character repetitions, Kashida "tatweel" and diacritics (Althobaiti et al., 2014). The class distribution is highly imbalanced, which could make a model biased towards certain classes. Therefore, random up-sampling for each data class was applied to match the size of the majority class, the Egyptian class.

3.2 Combined Features Model

Our approach to classifying tweets involves three components:

1. BiLSTM-CNN model: to extract word and character representations, we build a BiLSTM-CNN model following the same settings in (Ma and Hovy, 2016). We pre-trained FastText on the 10m unlabeled tweets to represent words. We randomly initialize character embeddings of size 30 and train them into a CNN neural network to learn morphological information, for example, the prefix or suffix of a word (Dos Santos and Zadrozny, 2014). We concatenate each word embedding with its character embedding and feed them into a BiLSTM network to learn the sentence information.

2. Character-level TF-IDF: we applied (1-5) character grams of TF-IDF on the train set. We tried to expand gram range, but that did not improve the performance. The TF-IDF component captures very common patterns of a dialect.

3. Topic modeling: is used to discover a set of topics from large documents where a topic is a distribution over words that are associated with a single subject. We used Latent Dirichlet Allocation (LDA) to learn topic modeling on the train set. We tried different number of topics {1, 10, ..,100 } and we empirically found 50 topics to yield the best results.

For each tweet, we concatenate its BiLSTM-CNN, character TF-IDF, and topic modeling features together and then feed them into a classifier made of two-layer neural network. There is a dropout layer between the two layers of the classifier. The overall model is in Figure 1. Also, we find training Fasttext on the 10 million unlabeled tweets to yield better results than using existing pretrained Arabic word embeddings. We train the model using the train set, and we optimize the hyperparameters based on the evaluation of the development set, the hyperparameter settings in Table 3. We applied the early stopping technique based on the F1-macro score.

In NADI shared task, we used the mentioned model for our first run.

3.3 Combined Features with Heuristic Model

Next, we augmented our model with a heuristic from Samih et al. (2019). The heuristic is based on a list of all Arabic speaking countries and their major cities. If a tweet mentions any country/city in the list, the tweet would be classified to the mentioned country/city. We excluded the cities because they did not boost the overall performance.

Our model combined with the heuristic was our second run to NADI challenge.

Number of units in the first layer	1200
Number of units in the second layer	800
Cell size of BiLSTM	500
Learning rate	1e-4
Dropout rate	0.5
Optimizer	Adam

Table 3: Hyperparameter settings.

We implemented the neural network using Tensorflow (Abadi et al., 2015), and we modelled the topics using Gensim (Řehůřek and Sojka, 2010).

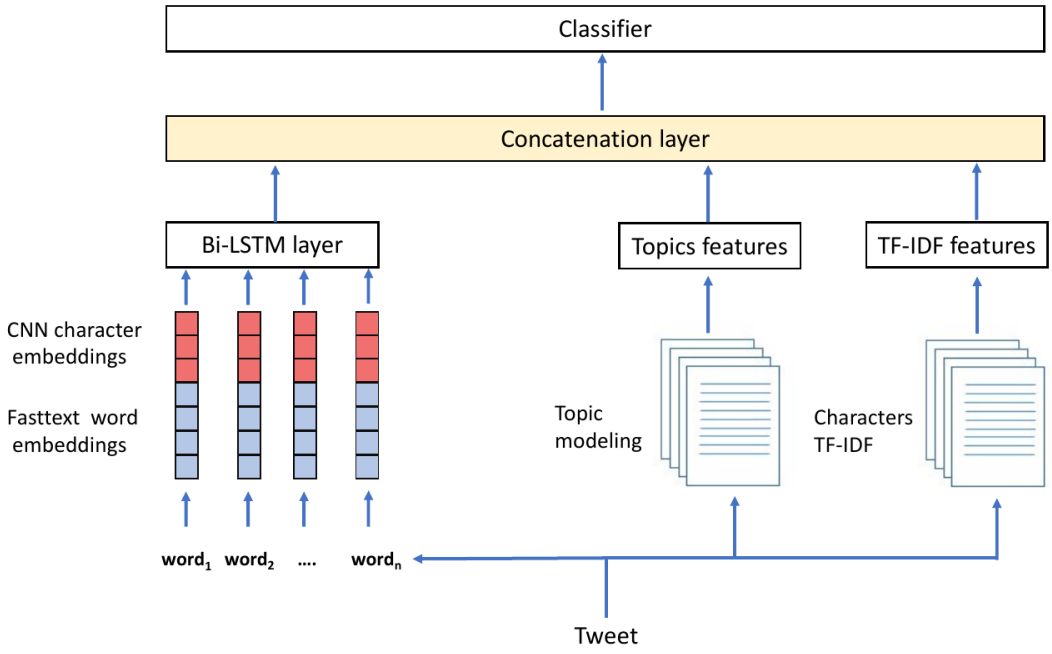

Figure 1: Our model

4 Results

As we can see in Table 4, the basic model achieved a recall, precision, macro-F1 scores of 0.231, 0.231, and 0.227 on the development set and 0.198, 0.203, 0.194 on the test set when we do not apply the heuristic (run 1). When we applied the heuristic (run 2), the scores increased to 0.244 (R), 0.250 (P), and 0.243 (F1) on the development set and .210 (R), 0.216 (P), and 0.207 (F1) on the test set. Investigating the system's outputs, we found the model successfully classified the three largest classes (Egypt, Iraq, and Saudi Arabia), but struggled to classify in particular the classes with fewer samples (Mauritania, Bahrain, Djibouti, Somalia, and Sudan). We also found that the model struggles with tweets containing words which are found in the dialectal varieties of different countries, such as the word *what* (وشو) which is common in Saudi Arabia, Oman, Kuwait and Qatar. In addition, we found that many NADI tweets are in classical Arabic or MSA and are difficult to classify even for native speakers; we discuss this issue more in Section 5.

Run	Dataset	Recall	Precision	F1	Accuracy
run 1 (without heuristic)	Development set	0.231	0.231	0.227	0.387
	Test set	0.198	0.203	0.194	0.337
run 2 (with heuristic)	Development set	0.244	0.250	0.243	0.392
	Test set	0.210	0.216	0.207	0.343

Table 4: Evaluation on NADI country-level dataset on the development and test portions of two settings: run 1 (no heuristic) and run 2 (with heuristic).

The model has also been evaluated on the AOC benchmarks. We trained and optimized the model following the same splits as in (Elaraby and Abdul-Mageed, 2018). As we can see in Table 5, the model achieves the highest accuracy scores compared to previous approaches on all dataset settings, binary, three-way, and four-way.

5 Discussion

Our model was evaluated on two datasets: AOC and NADI. The results on AOC are very high compared to the results on NADI because NADI data contains many difficult cases. In NADI, there are many

Model	Binary		Three-way		Four-way	
	Dev	Test	Dev	Test	Dev	Test
BiGRU (Elaraby and Abdul-Mageed, 2018)	87.65	87.23	87.11	86.18	83.25	82.21
Att. BiLSTM (Elaraby and Abdul-Mageed, 2018)	87.61	87.21	87.81	87.41	83.49	82.45
Our model	88.18	87.64	90.09	89.94	85.56	84.23

Table 5: The results of our experiments on the portion of benchmarked AOC compared with BiGRU and Attention BiLSTM of (Abdul-Mageed et al., 2018). Following prior works, we compare our experimental result in accuracy.

Example	Sentence	Not MSA words	MSA ratio
1	انا عرفت انا هسقط ليه	هسقط	4/5
2	كلهم فتره ويذلفون	ويذلفون	2/3
3	بيخترعو قواعد من عندن	بيخترعو، عندن	2/4

Table 6: The MSA ratio is the number of MSA words of a tweet divided by its total number of words. We consider a word in MSA if it is in AraVec dictionary.

tweets in classical Arabic and MSA, such as religious verses, popular poems, and others. Such cases are very hard to label even for native speakers because they are ubiquitous in all Arabic speaking countries. To gain more insights into the proportion of MSA tweets, we used AraVec dictionary (Soliman et al., 2017). AraVec is a distributed word representation model trained on Arabic Wikipedia articles which are mainly in MSA. Therefore, AraVec dictionary mostly contain MSA words. To know if a tweet is in MSA, we define the MSA ratio which is the number of tweet words in AraVec dictionary divided by the total number of words. We show a few examples in Table 6. The MSA ratio of all tweets in the training and development set is shown in Figure 2: as we can see, many tweets in the corpus have high MSA ratio. We found that 6291 of the 21,000 have an MSA ratio of 1.0 in the train set, and 1508 of the 4957 in the development. These cases can complicate the learning process and associate common MSA words/character-grams to a specific dialect. For example, our model classifies (السلام عليكم/ *Als~lAmu çalykum*[1]) which is used in all dialects, as Somalian because many training instances with السلام عليكم are labeled as Somalian.

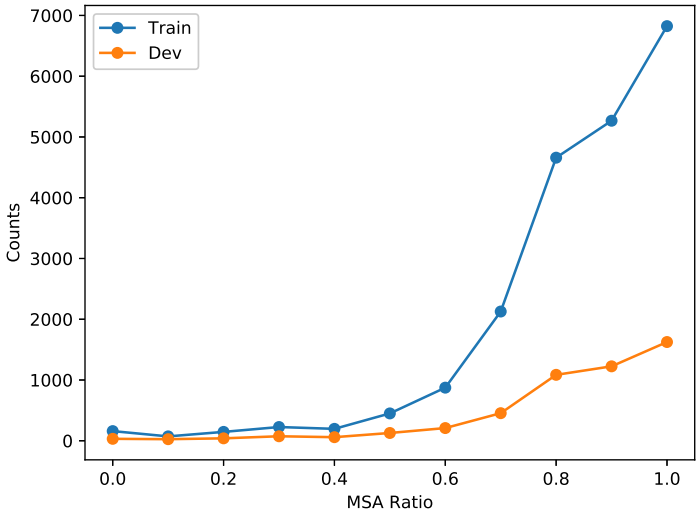

Figure 2: MSA ratios of train and development sets in NADI

[1]Following the Arabic transliteration scheme in (Habash et al., 2007)

6 Conclusion

We presented a model to identify Arabic dialects based on three components: BiLSTM-CNN, character-level TF-IDF, and topic modeling. We evaluated the model on the country-level subtask of NADI, and also on the AOC dataset. We showed their results and discussed the challenges of NADI dataset.

Acknowledgements

The research was in part supported by the UK Economic and Social Research Council (ESRC) through the Big Data Human Rights and Technology project (grant number ES/M010236/1).

References

Martín Abadi, Ashish Agarwal, Paul Barham, et al. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Muhammad Abdul-Mageed, Hassan Alhuzali, and Mohamed Elaraby. 2018. You tweet what you speak: A city-level dataset of arabic dialects. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

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 (WANLP2020)*, Barcelona, Spain.

Maha Althobaiti, Udo Kruschwitz, and Massimo Poesio. 2014. Aranlp: A java-based library for the processing of arabic text.

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)*.

Ryan Cotterell and Chris Callison-Burch. 2014. A multi-dialect, multi-genre corpus of informal written arabic. In *LREC*, pages 241–245.

Kareem Darwish, Hassan Sajjad, and Hamdy Mubarak. 2014. Verifiably effective arabic dialect identification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1465–1468.

Cicero Dos Santos and Bianca Zadrozny. 2014. Learning character-level representations for part-of-speech tagging. In *International Conference on Machine Learning*, pages 1818–1826.

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.

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.

Asmaa Etman and AA Louis Beex. 2015. Language and dialect identification: A survey. In *2015 SAI Intelligent Systems Conference (IntelliSys)*, pages 220–231. IEEE.

Nizar Habash, Abdelhadi Soudi, and Timothy Buckwalter. 2007. On arabic transliteration. In *Arabic computational morphology*, pages 15–22. Springer.

Fei Huang. 2015. Improved arabic dialect classification with social media data. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2118–2126.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA. `http://is.muni.cz/publication/884893/en`.

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.

Younes Samih, Hamdy Mubarak, Ahmed Abdelali, Mohammed Attia, Mohamed Eldesouki, and Kareem Darwish. 2019. Qc-go submission for madar shared task: Arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 290–294.

Abu Bakr Soliman, Kareem Eissa, and Samhaa R El-Beltagy. 2017. Aravec: A set of arabic word embedding models for use in arabic nlp. *Procedia Computer Science*, 117:256–265.

Toshiyuki Takezawa, Genichiro Kikui, Masahide Mizushima, and Eiichiro Sumita. 2007. Multilingual spoken language corpus development for communication research. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 12, Number 3, September 2007: Special Issue on Invited Papers from ISCSLP 2006*, pages 303–324.

Christoph Tillmann, Saab Mansour, and Yaser Al-Onaizan. 2014. Improved sentence-level arabic dialect classification. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 110–119.

Wajdi Zaghouani and Anis Charfi. 2018. Arap-tweet: A large multi-dialect twitter corpus for gender, age and language variety identification. *arXiv preprint arXiv:1808.07674*.

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