ArbDialectID at MADAR Shared Task 1: Language Modelling and Ensemble Learning for Fine Grained Arabic Dialect Identification

Kathrein Abu Kwaik
Gothenburg University
Sweden
kathrein.abu.kwaik@gu.se

Motaz Saad
The Islamic University of Gaza
Palestine
motaz.saad@gmail.com

Abstract

In this paper, we present a Dialect Identification system (ArbDialectID) that competed at Task 1 of the MADAR shared task, MADAR Travel Domain Dialect Identification. We build a coarse and a fine grained identification model to predict the label (corresponding to a dialect of Arabic) of a given text. We build two language models by extracting features at two levels (words and characters). We firstly build a coarse identification model to classify each sentence into one out of six dialects, then use this label as a feature for the fine grained model that classifies the sentence among 26 dialects from different Arab cities, after that we apply ensemble voting classifier on both subsystems. Our system ranked 1st that achieving an f-score of 67.32%. Both the models and our feature engineering tools are made available to the research community.

1 Introduction

Arabic Language is one of the most spoken languages in the world. Furthermore, Arabic presents us with a special case of Diglossia (Ferguson, 1959), where the spoken language is different than the formal language. Speakers of Arabic use Modern Standard Arabic (MSA) as the official language in very formal situations like education, religion, media, and politics, while they use an Arabic Dialect (AD) for everyday conversation (Shah, 2008; Versteegh, 2014).

With the emergence of social media, speakers of Arabic use their dialects to tweet, post, socialize and express themselves. The Arabic Dialects (AD) do not have a standardized writing and/or orthography, and they do not have a formal grammar. These characteristics make the task of identifying dialects more challenging.

The task of Arabic Dialect Identification (ADI) has recently attracted research attention, building identification systems able to differentiate among the dialects have been attempted. Even though dialects share similar features in term of lexical, syntax, morphology and semantics, they still have many differences which, of course, complicates the identification task.

Many works addressed the problem of dialect identification. They have reported different dialectal divisions, according to the geo-location, the country or, in some cases, on the level of cities. Most of those works used Machine learning classifiers and language modelling and achieved a good accuracy depending on the level of identification and either they explored the coarse grained identification, where the differences between the individual dialects are clear or a fine grained identification, where the differences become hard to detect in text as the dialects look very similar to each others (Zbib et al., 2012; Cotterell and Callison-Burch, 2014; Zaidan and Callison-Burch, 2014; Qwaider et al., 2018; Elfardy and Diab, 2013).

Other approaches investigated the use of Deep Learning (DL) methods to identify dialects. As such, they tried different DL architectures like LSTMs, CNNs and attention networks, and have employed different word embedding models. Elaraby and Abdul-Mageed (2018) benchmarked the Arabic Online Commentary (AOC) (Zaidan and Callison-Burch, 2011) and tested six different deep learning methods on the ADI task, comparing performance to several classical machine learning models under different conditions (both binary and multi-way classification). Their models reached 87.65% accuracy on the binary task (MSA vs. dialects), 87.4% accuracy on the three-way dialect task (Egyptian, Gulf, Levantine), and 82.45% accuracy on the four-way classification task (MSA, Egyptian, Gulf, Levantine). Similarly, Lulu and Elnagar (2018) explored the DL methods with different networks structure using AOC

Proceedings of the Fourth Arabic Natural Language Processing Workshop, pages 254–258
Florence, Italy, August 1, 2019.

on a three-way classification, with LSTM they achieved 71.4% accuracy

This paper presents our participation in MADAR shared task (Bouamor et al., 2019). We participate in Task 1: MADAR travel domain dialect identification, and we ranked 1st in the task with accuracy of 67.3%. We present our proposed model (ArbDialectID) in details and the code is available at GitHub[1].

The rest of this paper is organized as follow: Section 3 discusses the used data and presents our proposed model, we discusses the results in Section 4 and conclude in Section 5.

2 ArbDialectID: Arabic Dialect Identification System

This section introduces our proposed model which is applied on MADAR corpus for dialect identification shared task. MADAR corpus (Bouamor et al., 2018) is a parallel corpus in travel domain, it contains 25 dialects from different Arab cities in addition to the MSA. This corpus has been used for AID task in (Salameh et al., 2018), where the authors applied language modeling with various combinations of word and character levels and trained the model by MNB classifier. They got 67.9% accuracy for 26 classification task.

Our model consists of two sub models and exploiting two different data set as shown in Figure 1. The first model tries to predict the dialect among six different Arab dialects and known as coarse grained level, followed by the second model which goes much deeper and is known as a fine grained level to classify 26 Arabic dialects.

In both of our sub models we use MADAR data set to build and evaluate the models. Table 1 shows the number of sentences/samples per dialects and the total sentences for each data set. All of the experiments are implemented by Python and with the help of `scikit learn` library (Pedregosa et al., 2011).

MADAR	Split	sentences	Total
Corpus-6	train	9,000	41,600
	dev	1,000	6,000
Corpus-26	train	1,600	41,600
	dev	200	5,200
	test	200	5,200

Table 1: Statistics for MADAR data sets

[1]https://github.com/motazsaad/ArbDialectID

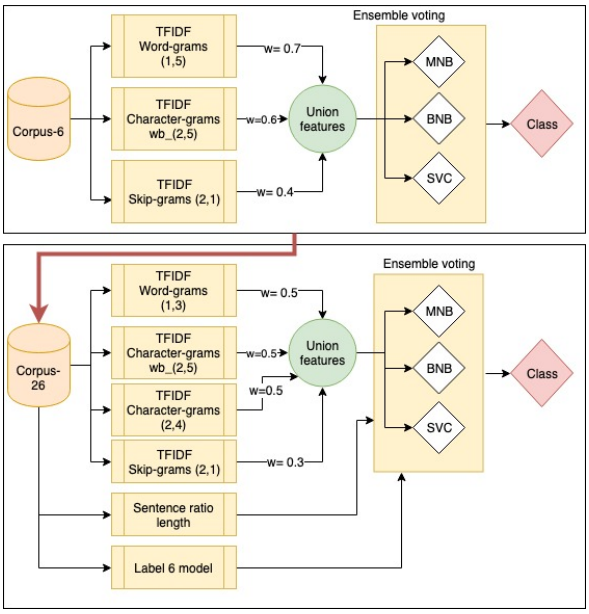

Figure 1: ArbDialectID proposed model

2.1 Coarse Grained Dialect Identification

This is the first model where we classify among five different Arab dialects from five Arabic countries, which are covered by MADAR corpus, they are: Beirut (BEI), Cairo (CAI), Doha (DOH), Rabat (RAB), Tunisia (TUN), In addition to (MSA).

We build a model that depends on the language modelling and exploring different combinations of n-grams in the word level and the character level. We use *FeatureUnion* in *sklearn*, which is an estimator that concatenates results of multiple transformer objects. To build and train the model we extract the following features:

- TF-IDF vectors from the word grams ranged from the unigram to 5-grams. We apply 0.7 weight for vector transformation
- TF-IDF vectors from the character n-grams with word boundary consideration ranged from bigrams to 5-grams and the transformation weight is 0.6
- Apply skip grams , then we extract the unigram words with one word skipping. We give it the lowest transformation weight of 0.4

The transformation weight is a weight used in *FeatureUnion* to give a weight for the feature. We choose these weights empirically after many experiments that investigate various weights with many features combinations.

After features extraction process, we build an ensemble voting classifier with hard voting, where it uses predicted class labels for majority rule voting. The ensemble classifiers consists of the fol-

lowing best standalone Machine Learning algorithms:

- MultinomialNB (MNB) , we set alpha to 0.01
- Linear SVC with l2 penalty and the learning rate sets to 0.0001
- BernoulliNB (BNB), set alpha = 0.01

We trained the model using "MADAR corpus-6" train set, and evaluate it by MADAR corpus-6 development set. We reach an accuracy of 92.7% and macro F-score of 93%. Finally, we combine the train and the dev-set together and rebuild the model again. We call it (MADAR model-6). We will use this model later in the second sub model.

2.2 Fine Grained Dialect Identification

This model is the core of the shared task, where it is going to predict the label for a given sentence and classify it to one of 26 dialects. MADAR corpus covers 25 cities in the Arab countries in addition to the MSA, they are : Aleppo (ALE), Algeria (ALG), Alexandria (ALX), Amman (AMM), Aswan (ASW), Baghdad (BAG), Basra (BAS), Beirut (BEI), Benghazi (BEN), Cairo (CAI), Damascus (DAM), Doha (DOH), Fes (FES), Jeddah (JED), Jerusalem (JER), Khartoum (KHA), Mosul (MOS), Muscat (MUS), Rabat (RAB), Riyadh (RIY), Salt (SAL),Sana'a (SAN), Sfax (SFX), Tripoli (TRI), Tunisia (TUN) and MSA.

In the same manner we build the second model by extracting some features as follow:

- TF-IDF vectors from the word grams with uni-gram, bi-gram and tri-gram words. we apply 0.5 weight for vector transformation
- TF-IDF vectors from the character n-grams with word boundary consideration ranged from bi-grams to 5-grams and the transformation weight is 0.5
- Extract another character n-grams but this time without word boundary consideration from bi-grams to 4 grams and the transformation weight is 0.5
- Again apply skip gram, then we extract the uni-gram words with one work skipping. We assign it 0.3 transformation weight

In addition to theses feature we add another two numerical features, the first is the sentence length ratio for every sentence in the data (train, dev, test) which in turn divides the total number of words appearing in the sentence by the total number of words appearing in the longest sentence. The second features depends on the previous MADAR-model-6. We exploit this model to predict the label for MADAR Corpus-26, so every sentence is combined with a predicted class number with one value from 1 to 6, for example 1 means CAI, 2 is for BEI and so on. So in total we have the TF-IDF vectors features in addition to the two numerical features (the coarse-grained label and the sentence length).

To build the model, we employ ensemble hard voting classifier with the previously mentioned three algorithms (Linear SVC, MNB and BNB). The system is trained on MADAR corpus-26 train set, evaluated by MADAR corpus-26 dev set and finally tested by MADAR corpus-26 test set. Table 1 reports the results for the dev set and test set and Figure 2 shows the classification report which is produced from the test set .

	Accuracy	macro F-score
Dev	68.7	69.00
Test	67.29	67.32

Table 2: Results for 26 dialects Identification system

```
classification report:
              precision    recall  f1-score

         ALE       0.62      0.68      0.65
         ALG       0.77      0.81      0.79
         ALX       0.76      0.76      0.76
         AMM       0.54      0.53      0.54
         ASW       0.57      0.65      0.60
         BAG       0.65      0.68      0.66
         BAS       0.70      0.70      0.70
         BEI       0.73      0.64      0.68
         BEN       0.71      0.69      0.70
         CAI       0.54      0.54      0.54
         DAM       0.54      0.61      0.57
         DOH       0.65      0.67      0.66
         FES       0.77      0.70      0.73
         JED       0.57      0.61      0.59
         JER       0.58      0.60      0.59
         KHA       0.74      0.74      0.74
         MOS       0.89      0.82      0.85
         MSA       0.68      0.79      0.73
         MUS       0.56      0.46      0.50
         RAB       0.76      0.76      0.76
         RIY       0.58      0.60      0.59
         SAL       0.62      0.56      0.59
         SAN       0.75      0.73      0.74
         SFX       0.74      0.73      0.74
         TRI       0.78      0.80      0.79
         TUN       0.78      0.68      0.73

   micro avg       0.67      0.67      0.67
   macro avg       0.68      0.67      0.67
weighted avg       0.68      0.67      0.67
```

Figure 2: Fine Grained Dialect Identification classification report for MADAR corpus-26 test set

3 Discussion

Building a language model for a language or a text is an informative way to describe and represent the language. In this work, we try to extract as

many discriminated features as possible that can be employed efficiently to distinguish among the desired 6 and 26 dialects. In the coarse grained dialect identification with MADAR Corpus-6 the task was more flexible, the dialects have a reasonable differences as they represent a large groups of dialects, for example DOH represents dialects from the Arab Gulf, BEI represents the Levantine dialects and so on. Due to the differences on the lexical level between thees dialects we emphasise the word n-grams by using greater weight transformation, and assign a smaller weight value for the character levels n-grams.

For the task of fine grained dialect identification, the task was more tough and we need more extra features and emphasise some of them more. Hence, we increase the number of n-grams and emphasise the character n-grams and pay attention to the words boundaries. We employ the first model as another feature to enhance the f-score for the second models. Given that, the corpus contains many short sentence that appears in more one dialects, it makes the models to some extent confused, then we add the length of the sentence as an extract helpful feature where some dialects need more words to express an idea, and the other use more suffixes. It is also impossible for Arabic speakers to detect the dialect from a very short sentence with 100% especially if it does not contain any clue words. In some cases the dialects become very similar to each others when they are spoken by neighbourhood, for instance the Jerusalem dialect and the dialect from Amman where they are considered in some researches in Arabic history as the same dialect (Owens, 2015; Bishop, 1998). From the classification report in Figure 2, it is very clear that some dialects were easier to detect than other, for example, the North Africa dialects gain high f-scores compare to others such as the following dialects: TRI (0.79), SFX(0.74), BEN(0.70), ALG(0.79) and TUN(0.73). The confusion matrix in Figure 3 shows the numbers of actual and predicate labels for each dialect. There are some similar pairs of dialects where the system confused like (BAG and BAS), (AMM and JER), (CAI and ASW), (ALE and DAM) and (SFX and TUN).

We investigate the word grams model as well as the character grams model. The best result is obtained when we combine both of these models, given that the differences may occur in terms of

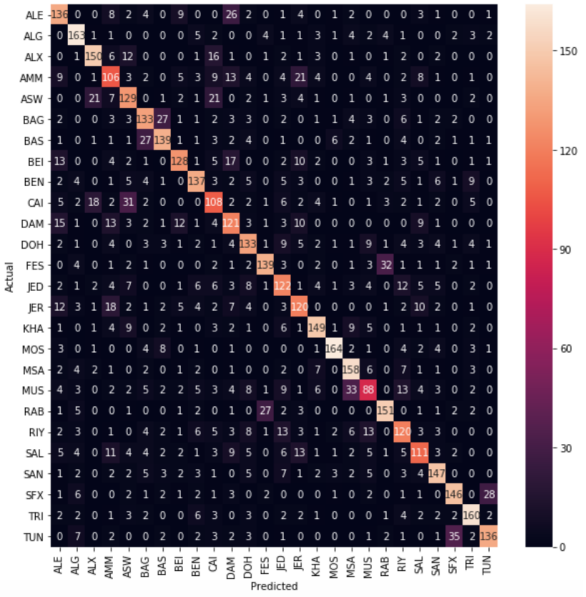

Figure 3: Fine Grained Dialect Identification confusion matrix for MADAR corpus-26 test set

lexical words, however there are many differences that occurred on character levels like different clitics, prefixes and suffixes. We try to exploit the best classifier that has been used for ADI and finally end up by ensemble learning that combines the Linear SVC , MNB and BNB with hard voting where the max probability is chosen as the correct class.

4 Conclusion

We participate in MADAR shared task, Task 1: "MADAR Travel Domain Dialect Identification". We build an ADI system consists of two subsystems. The first is a six dialects classification system, followed by a 26 classification system that classify 26 dialects from 25 cities in the Arab world in addition to MSA. We use different combinations of n-gram models (words, Characters) and skip gram models. In addition to these language modelling features, we compute the ratio length of each input sentence and use the predicted label from the first model. We achieve the best score in the competition with 67.32% f-score and an accuracy of 67.29%.

Acknowledgements

Kathrein Abu Kwaik is supported by grant 2014-39 from the Swedish Research Council, which funds the Centre for Linguistic Theory and Studies in Probability (CLASP) in the Department of Philosophy, Linguistics, and Theory of Science at the University of Gothenburg.

References

Brian Bishop. 1998. A history of the Arabic language. *Department of Linguistics, Brigham Young University*.

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.

Ryan Cotterell and Chris Callison-Burch. 2014. A Multi-Dialect, Multi-Genre Corpus of Informal Written Arabic. In *LREC*, pages 241–245.

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 Conference of the Association for Computational Linguistics (ACL)*, Sofia, Bulgaria.

Charles A. Ferguson. 1959. Diglossia. *word*, 15(2):325–340.

Leena Lulu and Ashraf Elnagar. 2018. Automatic Arabic Dialect Classification Using Deep Learning Models. *Procedia computer science*, 142:262–269.

Jonathan Owens. 2015. Arabic language history and the comparative method. *International Journal of Arabic Linguistics*, 1(1):1–27.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research*, 12(Oct):2825–2830.

Chatrine Qwaider, Motaz Saad, Stergios Chatzikyriakidis, and Simon Dobnik. 2018. Shami: A Corpus of Levantine Arabic Dialects. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

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.

Mustafa Shah. 2008. *The Arabic language*. Routledge.

Kees Versteegh. 2014. *The Arabic language*. Edinburgh University Press.

Omar F. 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: short papers-Volume 2*, pages 37–41. Association for Computational Linguistics.

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

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan, and Chris Callison-Burch. 2012. Machine translation of Arabic dialects. In *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 49–59. Association for Computational Linguistics.