ST MADAR 2019 Shared Task:
Arabic Fine-Grained Dialect Identification

Mourad Abbas, Mohamed Lichouri
Computational Linguistics Department-CRSTDLA
Algeria
{m.abbas, m.lichouri}@crstdla.dz

Abed Alhakim Freihat
Trento University
Italy
abed.freihat@unitn.it

Abstract

This paper describes the solution that we propose on MADAR 2019 Arabic Fine-Grained Dialect Identification task. The proposed solution utilized a set of classifiers that we trained on character and word features. These classifiers are: Support Vector Machines (SVM), Bernoulli Naive Bayes (BNB), Multinomial Naive Bayes (MNB), Logistic Regression (LR), Stochastic Gradient Descent (SGD), Passive Aggressive(PA) and Perceptron (PC). The system achieved competitive results, with a performance of 62.87% and 62.12% for both development and test sets.

1 Introduction

Dialect identification (Zaidan and Callison-Burch, 2014) is a sub field of language identification which can be coarse-grained or fine-grained. Coarse-grained dialect identification or simply dialect identification (Meftouh et al., 2015) refers to the process of dividing a language into the main dialects that belong to that language. On the other hand, fine-grained dialect identification (Salameh et al., 2018) considers the differences between the sub dialects inside a dialect of some language.

In this paper, we describe a fine grained dialect identification systems that participated in MADAR 2019 Arabic Fine-Grained Dialect Identification task (Bouamor et al., 2019) In this task, our system was trained on a data-set of short sentences in the travel domain. A sentence in this data set belongs to one or more Arabic fine-grained dialects. These dialects are -Aleppo (ALE), Algiers (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), Tunis (TUN) and Modern Standards Arabic (MSA) (Bouamor et al., 2018). The task of our system is to identify the dialect of a given sentence that belong to these 26 dialects.

The multi-way classification system that we propose uses word n-grams and char n-grams as features, and MNB, BNB and SVM as classifiers.

The rest of the paper is organized as follows. In Section 2, we describe the data-set. In Section 3.1, we address the task as a multiway text-classification task; where we describe the proposed system in 3. We report our experiments and results in 4 and conclude with suggestions for future research and conclusion in 5 and 6.

2 Dataset

In this work, we used the MADAR Travel Domain dataset built by translating the Basic Traveling Expression Corpus (BTEC) (Takezawa et al., 2007). The whole sentences have been translated manually from English and French to the different Arabic dialects by speakers of 25 dialects (Salameh et al., 2018; Bouamor et al., 2019). The training data is composed of 1600 sentences for each of the 25 dialects in addition to MSA. The size of the development and test sets is 200 sentences per dialect. The sentences are short, ranging from 4 to 15 words each. Each sentence is annotated with the speaker dialect. In table 1, we provide some statistics on the used corpora.

Arabic dialects can be considered as variants of Modern Standard Arabic. However, the absence of a standard orthography (Habash et al., 2018) (Habash et al., 2012) for dialects generates many different shapes of the same word. Despite this, there are still similarities between these dialects which make their identification difficult under textual format. In figure 3, we present respectively the number of words and sentences, shared be-

Proceedings of the Fourth Arabic Natural Language Processing Workshop, pages 269–273
Florence, Italy, August 1, 2019. ©2019 Association for Computational Linguistics

	Train	Dev	Test	Total
# sentences	41,600	5,200	5,200	52,000
# distinct sentences	38,506	4,873	4,870	48,249
# words	294,718	37,383	36,810	368,911
# distinct Words	27,501	6,136	6,062	39,699

Table 1: Madar Task 1 Dataset statistics

tween n dialects where n varies from 2 to 26.

3 System

The presentation of our proposed approach is shown in figure 2.

3.1 Feature extraction

We applied a light preprocessing step where a simple blank tokenization and punctuation filtering have been achieved. It is worthy to say, that we deployed in our preliminary experiments Low level NLP processing such as POS-tagging (Freihat et al., 2018b) features and lemmatization (Freihat et al., 2018a) but without a significant enhancement of the achieved results. Besides the word and character n-grams features used in previous work such as (Salameh et al., 2018; Lichouri et al., 2018), we added the character-word_boundary (char_wb). In the following, we present a description of the three adopted features.

- **Word n-grams**: We extract word n-grams, with n ranging from 1 to 3.
- **Char n-grams**: The character first to third grams are used as features.
- **Char_wb n-grams**: This feature creates character n-grams only from text inside word boundaries; n-grams at the edges of words are padded with space.

The count matrix obtained using these features are transformed to a tfidf representation.

3.2 Classification Models

Our model is based on a set of classifiers using the scikit-learn library (Pedregosa et al., 2011), namely: Support Vector Machines (SVM), Bernoulli Naive Bayes (BNB), Multinomial Naive Bayes (MNB), Logistic Regression (LR), Stochastic Gradient Descent (SGD), Passive Aggressive (PA) and Perceptron (PC). In the following, we present the selected parameters for each classifier.

- **SVM_r:** C:1.0, kernel:”rbf”, degree:3, decision-function-shape:”One-vs-Rest”.
- **SVM_l:** C:10, kernel:”linear”, degree:3, decision-function-shape:”One-vs-Rest”.
- **BNB:** alpha:1.0, fit-prior:True.
- **MNB:** alpha:1.0, fit-prior:True.
- **LR:** penalty:”l2”, C:1.0, solver:”sag”, max-iter:100.
- **SGD:** loss:”hinge”, penalty:”l2”, alpha:0.0001, l1-ratio:0.15, max-iter:1000, shuffle:True, epsilon:0.1, learning-rate:”optimal”.
- **PA:** C:1.0, max-iter:1000, shuffle:True, loss:”epsilon-insensitive”, epsilon:0.1.
- **PC:** alpha:0.0001, max-iter:1000, shuffle:True, eta0:1.0.

4 Results

Using the aforementioned classifiers, the best achieved performance (F1-Macro) for coarse-grained and fine-grained dialect identification was 90.55% (table 4) and 62.87% (table 3) respectively. The best results are obtained using the three classifiers: **SVM_l**, BNB and MNB with F1-Macro of 61.94%, 62.72% and 62.87% respectively (table 3). Based on these findings, we adopted the three models for test phase. The results are presented in table 2.

Model	Precision	Recall	F1	Accuracy
MNB	63.13	62.17	62.12	62.17
BNB	62.85	62.13	62.07	62.13
SVM_l	60.41	60.48	60.26	60.48

Table 2: Three first best results achieved by **MNB**, **BNB** and **SVM_l** (Test Phase). The F1, Precision and Recall Metrics are in Macro Mode.

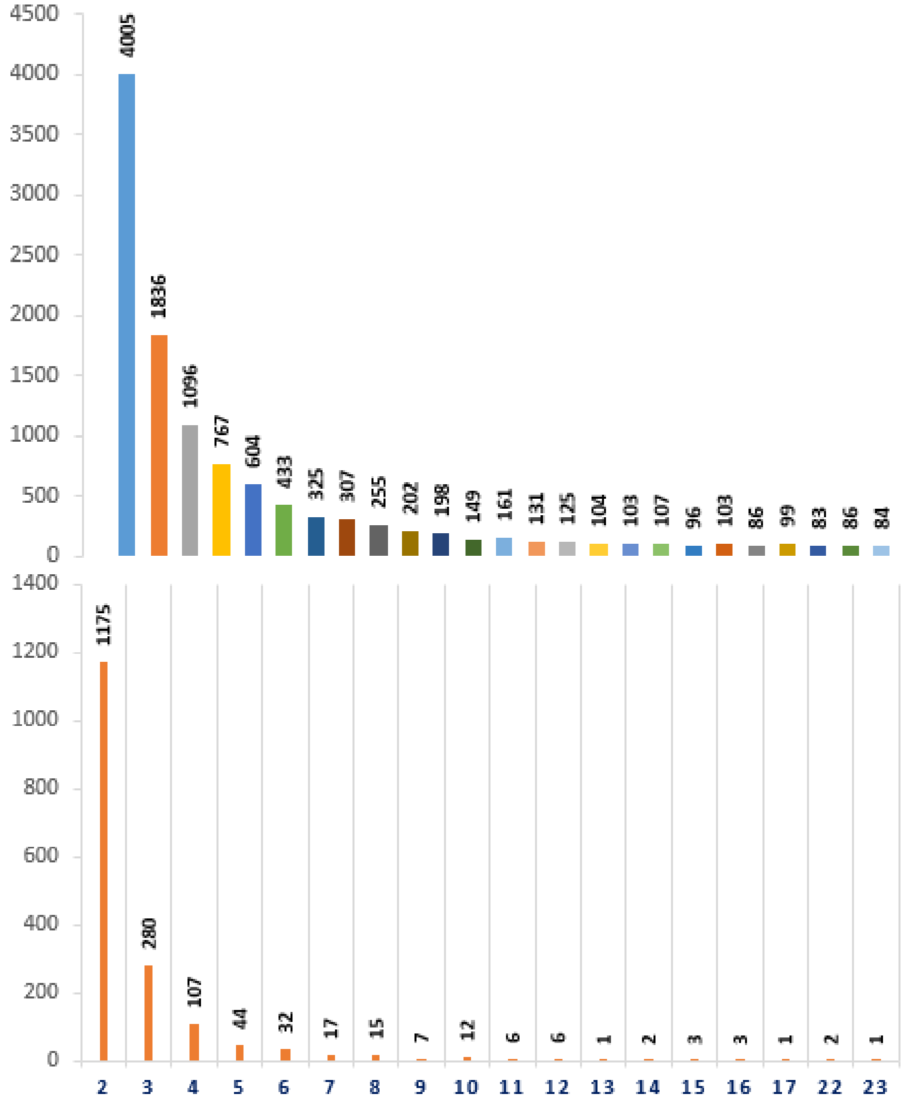

Figure 1: Number of tokens (**above**) and sentences (**below**) shared between the different dialects.

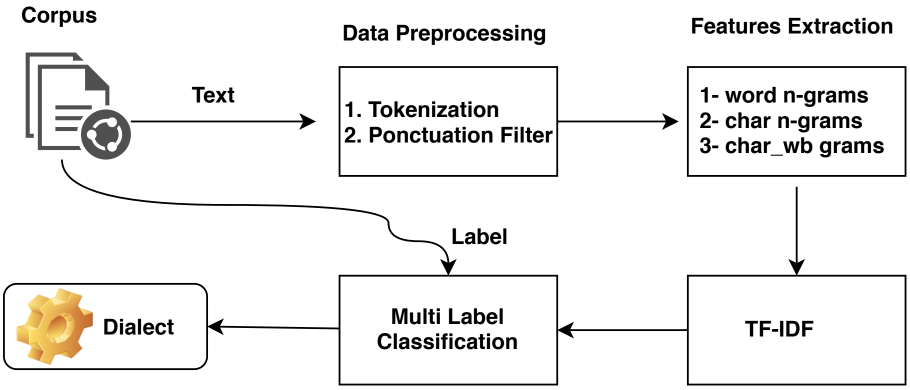

Figure 2: Dialect identification system

	SVM_r	SVM_l	BNB	MNB	LR	PA	SGD	PC
word n-grams	n=2	-	n=1	n=1	n=1	n=2	n=2	n=2
char_wb n-grams	-	n=3	-	-	-	-	-	-
Precision-Macro	60.09	62.29	64.09	64.28	59.67	60.55	58.40	56.97
Recall-Macro	59.19	62.19	62.73	62.87	59.33	60.10	57.88	55.90
F1-Macro	59.17	**61.94**	**62.72**	**62.87**	59.08	60.06	57.30	55.89
Accuracy	59.19	**62.19**	**62.73**	**62.87**	59.33	60.10	57.88	55.90

Table 3: Best results on the development dataset (**Corpus-26**) using the word n-grams and char_wb n-grams.

	SVM_r	SVM_r	BNB	MNB	LR	PA	SGD	PC
word n-grams	n=3	n=3	n=1 n=2	n=1 n=2	n=1 n=2	n=3	n=1 n=2	n=3
Precision-Macro	88.78	89.81	90.47	90.63	88.41	89.48	87.68	87.37
Recall-Macro	88.53	89.65	90.2	90.53	88.28	89.33	87.5	87.22
F1-Macro	88.59	**89.68**	**90.26**	**90.55**	88.32	89.36	87.53	87.24
Accuracy	88.53	**89.65**	**90.2**	**90.53**	88.28	89.33	87.5	87.22

Table 4: Best results on the development dataset (**Corpus-6**) using the word n-grams.

5 Discussion

We experimented different classifiers and a set of features to solve fine-grained dialect identification, i.e. a 26-way classification problem. The results show that fine grained dialect identification is more difficult given the similarity between dialects on one side, and on the other side, the non-standardization of writing dialectal texts that generates unpredictable texts. In addition, we noted the presence of MSA texts in several dialectal tweets which distorts the results. By using the test data-set, we calculated the accuracy achieved by our best model and presented in table 2. In addition, we dress in table 5 our best results compared to the baseline.

	Precision	Recall	F1	Accuracy
Baseline	69.00	68.00	69.00	67.90
ST Team	63.13	62.17	62.12	62.17

Table 5: Speech Translation team results compared to the baseline system -evaluated on test dataset-

In table 6, we note that the best results using both dev and test datasets were obtained for the MOS dialect with an accuracy of 80% and 78%. Whereas the (ALG and TRI) dialects have achieved, for both datasets, an F1-score of more than 70%. For Tunisian dialects (SFX, TUN), more than 69%. For Morrocan ones (FES, RAB), the best result was around 64%. The last results for both (AMM and MUS) showed an accuracy below 49%.

Dialect	**Precision**		**Recall**		**F1**	
	Test	**Dev**	**Test**	**Dev**	**Test**	**Dev**
ALE	55	62	62	57	58	60
ALG	71	73	76	80	73	76
ALX	72	70	76	78	74	74
AMM	49	43	54	54	51	48
ASW	53	47	66	60	58	53
BAG	65	74	61	58	63	65
BAS	70	68	62	64	66	66
BEI	75	77	56	56	64	65
BEN	62	65	68	70	65	68
CAI	64	65	41	41	50	50
DAM	56	65	54	49	55	56
DOH	64	57	61	61	63	59
FES	65	63	62	69	64	66
JED	53	63	56	61	55	62
JER	50	45	60	58	55	51
KHA	55	49	72	68	62	57
MOS	78	82	78	78	78	80
MSA	62	60	71	82	66	69
MUS	60	60	44	41	51	49
RAB	68	74	59	56	63	64
RIY	54	52	57	61	56	56
SAL	51	55	50	47	51	51
SAN	66	82	67	69	66	75
SFX	63	68	72	77	67	72
TRI	74	73	70	73	72	73
TUN	78	79	61	63	69	70
macro avg	63	64	62	63	62	63

Table 6: Best Results for the Test and Dev datasets, in terms of Precision, Recall and F1.

In figure 3, we show the average accuracy of the 5-regions and MSA, as described in (Salameh et al., 2018), for both development and test set. We notice that the best results were achieved for Yemen region with an accuracy of 75%, and an average accuracy of over 67% for the Maghreb Region.

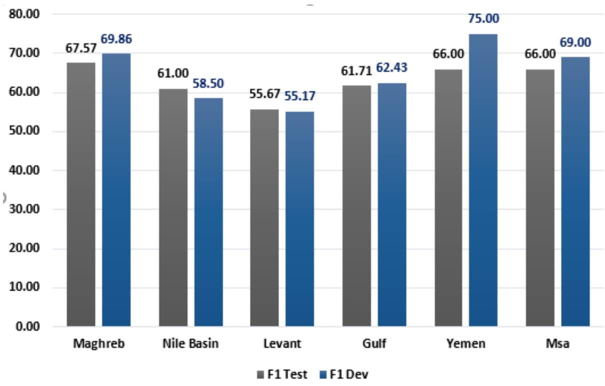

Figure 3: Average accuracy per region

6 Conclusion

In this paper, we proposed an Arabic fine-grained dialect identification system. Our best run on the test data yielded an F1-Macro score of 62% using Naive Bayes classifier and word n-gram features. Despite the simplicity of these features, the results were promising. In order to improve performance, we intend to investigate alternative methods as deep learning architectures and rule-based techniques in future work.

References

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR Arabic Dialect Corpus and Lexicon. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

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.

Abed Alhakim Freihat, Mourad Abbas, Gábor Bella, and Fausto Giunchiglia. 2018a. Towards an optimal solution to lemmatization in arabic. *Procedia computer science*, 142:132–140.

Abed Alhakim Freihat, Gabor Bella, Hamdy Mubarak, and Fausto Giunchiglia. 2018b. A single-model approach for arabic segmentation, pos tagging, and named entity recognition. In *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, pages 1–8. IEEE.

Nizar Habash, Mona Diab, and Owen Rambow. 2012. Conventional orthography for dialectal Arabic. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 711–718, Istanbul, Turkey. European Language Resources Association (ELRA).

Nizar Habash, Fadhl Eryani, Salam Khalifa, Owen Rambow, Dana Abdulrahim, Alexander Erdmann, Reem Faraj, Wajdi Zaghouani, Houda Bouamor, Nasser Zalmout, Sara Hassan, Faisal Al-Shargi, Sakhar Alkhereyf, Basma Abdulkareem, Ramy Eskander, Mohammad Salameh, and Hind Saddiki. 2018. Unified guidelines and resources for Arabic dialect orthography. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan. European Language Resource Association.

Mohamed Lichouri, Mourad Abbas, Abed Alhakim Freihat, and Dhiya El Hak Megtouf. 2018. Word-level vs sentence-level language identification: Application to algerian and arabic dialects. In *Proceedins of the 4th International Conference on Arabic Computational Linguistics (ACLing 2018)*.

Karima Meftouh, Salima Harrat, Salma Jamoussi, Mourad Abbas, and Kamel Smaïli. 2015. Machine translation experiments on PADIC: A parallel arabic dialect corpus. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation, PACLIC 29, Shanghai, China, October 30 - November 1, 2015*.

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.

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.

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

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