The MADAR Shared Task on Arabic Fine-Grained Dialect Identification

Houda Bouamor, Sabit Hassan, Nizar Habash[†]
Carnegie Mellon University in Qatar, Qatar
[†]New York University Abu Dhabi, UAE
{hbouamor,sabith}@qatar.cmu.edu
nizar.habash@nyu.edu

Abstract

In this paper, we present the results and findings of the MADAR Shared Task on Arabic Fine-Grained Dialect Identification. This shared task was organized as part of The Fourth Arabic Natural Language Processing Workshop, collocated with ACL 2019. The shared task includes two subtasks: the MADAR Travel Domain Dialect Identification subtask (Subtask 1) and the MADAR Twitter User Dialect Identification subtask (Subtask 2). This shared task is the first to target a large set of dialect labels at the city and country levels. The data for the shared task was created or collected under the Multi-Arabic Dialect Applications and Resources (MADAR) project. A total of 21 teams from 15 countries participated in the shared task.

1 Introduction

Arabic has a number of diverse dialects from across different regions of the Arab World. Although primarily spoken, written dialectal Arabic has been increasingly used on social media. Automatic dialect identification is helpful for tasks such as sentiment analysis (Al-Twairesh et al., 2016), author profiling (Sadat et al., 2014), and machine translation (Salloum et al., 2014). Most previous work, shared tasks, and evaluation campaigns on Arabic dialect identification were limited in terms of dialectal variety targeting coarse-grained regional dialect classes (around five) plus Modern Standard Arabic (MSA) (Zaidan and Callison-Burch, 2013; Elfardy and Diab, 2013; Darwish et al., 2014; Malmasi et al., 2016; Zampieri et al., 2017; El-Haj et al., 2018). There are of course some recent noteworthy exceptions (Bouamor et al., 2018; Zaghouani and Charfi, 2018; Abdul-Mageed et al., 2018).

In this paper, we present the results and findings of the MADAR Shared Task on Arabic Fine-Grained Dialect Identification. The shared task was organized as part of the Fourth Arabic Natural Language Processing Workshop (WANLP), collocated with ACL 2019.[1] This shared task is the first to target a large set of dialect labels at the city and country levels. The data for the shared task was created under the Multi-Arabic Dialect Applications and Resources (MADAR) project.[2]

The shared task featured two subtasks. First is the MADAR Travel Domain Dialect Identification subtask (Subtask 1), which targeted 25 specific cities in the Arab World. And second is the MADAR Twitter User Dialect Identification (Subtask 2), which targeted 21 Arab countries. All of the datasets created for this shared task will be made publicly available to support further research on Arabic dialect modeling.[3]

A total of 21 teams from 15 countries in four continents submitted runs across the two subtasks and contributed 17 system description papers. All system description papers are included in the WANLP workshop proceedings and cited in this report. The large number of teams and submitted systems suggests that such shared tasks on Arabic NLP can indeed generate significant interest in the research community within and outside of the Arab World.

Next, Section 2 describes the shared task subtasks. Section 3 provides a description of the datasets used in the shared task, including the newly created MADAR Twitter Corpus. Section 4 presents the teams that participated in each subtask with a high-level description of the approaches they adopted. Section 5 discusses the results of the competition. Finally, Section 6 concludes this report and discusses some future directions.

[1] http://wanlp2019.arabic-nlp.net
[2] https://camel.abudhabi.nyu.edu/madar/
[3] http://resources.camel-lab.com.

2 Task Description

The MADAR Shared Task included two subtasks: the MADAR Travel Domain Dialect Identification subtask, and the MADAR Twitter User Dialect Identification subtask.

2.1 Subtask 1: MADAR Travel Domain Dialect Identification

The goal of this subtask is to classify written Arabic sentences into one of 26 labels representing the specific city dialect of the sentences, or MSA. The participants were provided with a dataset from the MADAR corpus (Bouamor et al., 2018), a large-scale collection of parallel sentences in the travel domain covering the dialects of 25 cities from the Arab World in addition to MSA (Table 1 shows the list of cities). This fine-grained dialect identification task was first explored in Salameh et al. (2018), where the authors introduced a system that can identify the exact city with an averaged macro F1 score of 67.9%. The participants in this subtask received the same training, development and test sets used in (Salameh et al., 2018). More details about this dataset are given in Section 3.

2.2 Subtask-2: MADAR Twitter User Dialect Identification

The goal of this subtask is to classify Twitter user profiles into one of 21 labels representing 21 Arab countries, using only the Twitter user tweets. The Twitter user profiles as well as the tweets are part of the MADAR Twitter Corpus, which was created specifically for this shared task. More details about this dataset are given in Section 3.

2.3 Restrictions and Evaluation Metrics

We provided the participants with a set of restrictions for building their systems to ensure a common experimental setup.

Subtask 1 Restrictions Participants were asked not to use any external manually labeled datasets. However, the use of publicly available unlabelled data was allowed. Participants were not allowed to use the development set for training.

Subtask 2 Restrictions First, participants were asked to only use the text of the tweets and the specific information about the tweets provided in the shared task (see Section 3.2). Additional tweets, external manually labelled data sets, or any meta information about the Twitter user or the tweets (e.g., geo-location data) were not allowed. Second, participants were instructed not to include the MADAR Twitter Corpus development set in training. However, any publicly available unlabelled data could be used.

Region	Country	City
Gulf	Yemen	Sana'a
of Aden	Djibouti	
	Somalia	
Gulf	Oman	Muscat
	UAE	
	Qatar	Doha
	Bahrain	
	Kuwait	
	KSA	Riyadh, Jeddah
	Iraq	Baghdad,
		Mosul, Basra
Levant	Syria	Damascus, Aleppo
	Lebanon	Beirut
	Jordan	Amman, Salt
	Palestine	Jerusalem
Nile Basin	Egypt	Cairo, Alexandria,
		Aswan
	Sudan	Khartoum
Maghreb	Libya	Tripoli, Benghazi
	Tunisia	Tunis, Sfax
	Algeria	Algiers
	Morocco	Rabat, Fes
	Mauritania	
		MSA

Table 1: The list of the regions, countries, and cities covered in Subtask 1 (City column) and Subtask 2 (Country column).

Evaluation Metrics Participating systems are ranked based on the macro-averaged F1 scores obtained on blind test sets (official metric). We also report performance in terms of macro-averaged precision, macro-averaged recall and accuracy at different levels: region (Acc_{region}), country ($\text{Acc}_{country}$) and city (Acc_{city}). Accuracy at coarser levels (i.e., country and region in Subtask 1; and region in Subtask 2) is computed by comparing the reference and prediction labels after mapping them to the coarser level. We follow the mapping shown in Table 1. Each participating team was allowed to submit up to three runs for each subtask. Only the highest scoring run was selected to represent the team.

3 Shared Task Data

Next, we discuss the corpora used for the subtasks.

3.1 The MADAR Travel Domain Corpus

In Subtask 1, we use a large-scale collection of parallel sentences covering the dialects of 25 Arab cities (Table 1), in addition to English, French and MSA (Bouamor et al., 2018). This resource was a commissioned translation of the Basic Traveling Expression Corpus (BTEC) (Takezawa et al., 2007) sentences from English and French to the different dialects. It includes two corpora. The first consists of 2,000 sentences translated into 25 Arab city dialects in parallel. We refer to it as Corpus 26 (25 cities plus MSA). The second corpus has 10,000 additional sentences (non-overlapping with the 2,000 sentences) from the BTEC corpus translated to the dialects of only five selected cities: Beirut, Cairo, Doha, Tunis and Rabat. We refer to it as Corpus 6 (5 cities plus MSA). An example of a 27-way parallel sentence (25 cities plus MSA and English) extracted from Corpus 26 is given in Table 2. The train-dev-test splits of the corpora are shown in Table 3. Corpus 6 test set was not included in the shared task.[4]

3.2 The MADAR Twitter Corpus

For Subtask 2, we created a new dataset, the MADAR Twitter Corpus, containing 2,980 Twitter user profiles from 21 different countries.

Corpus collection Inspired by the work of Mubarak and Darwish (2014) we collected a set of Twitter user profiles that reflects the way users from different regions in the Arab World tweet. Unlike previous work (Zaghouani and Charfi, 2018), we do not search Twitter based on specific dialectal keywords. Rather, we search for tweets that contain a set of 25 seed hashtags corresponding to the 22 states of the Arab League (e.g., #Algeria, #Egypt, #Kuwait, etc.), in addition to the hashtags: "#ArabWorld", "#ArabLeague" and "#Arab". We collected an equal number of profiles (175 * 25 = 4,375) from the search results of each of the hashtags. The profiles were all manually labeled by a team of three annotators. For each labeled user profile, only the first 100 available tweets at collection time are kept.

[4] In (Salameh et al., 2018), the Corpus 6 test set corresponds to the 2,000 sentences from Corpus 26 corresponding to the Corpus 6's five cities and MSA.

Dialect	Sentence
Aleppo	بدي كنزة اطفال.
Alexandria	عاوز بلوفر أطفال.
Algiers	راني حاب تريكو تع أطفال.
Amman	بدي بلوفر أطفال.
Aswan	كنت عايز بلوفر اطفال.
Baghdad	اريد سترة مال أطفال.
Basra	اريد سترة جهال.
Beirut	بدي كنزة للولاد.
Benghazi	نبي مالية بتاع أطفال.
Cairo	عايز بلوفر اطفال.
Damascus	بدي كنزة أطفال.
Doha	بغيت فانلة يهال.
Fes	بغيت لبيسة ديال الدراري الصغار.
Jeddah	أبا سترة اطفال.
Jerusalem	بدي جرزة أطفال.
Khartoum	داير بلوفر أطفال.
Mosul	اغيد سترة اطفال.
Muscat	أبغا سترة للأطفال.
Rabat	بغيت تريكو ديال الدراري الصغار.
Riyadh	ابغى سترة للأطفال.
Salt	بدي بلوفر للأطفال .
Sana'a	أشتي سترة أطفالي.
Sfax	نحب مريول للأولاد.
Tripoli	نبي ماليه متاع صغار.
Tunis	نحب مريول متاع صغار.
MSA	أريد جاكيت للأطفال.

Table 2: An example from Corpus 26 for the English sentence 'I'd like a children's sweater.'

	Sentences	* Variant	Total
Corpus 6 train	9,000	* 6	54,000
Corpus 6 dev	1,000	* 6	6,000
Corpus 26 train	1,600	* 26	41,600
Corpus 26 dev	200	* 26	5,200
Corpus 26 test	200	* 26	5,200

Table 3: Distribution of the train, dev and test sets provided for Subtask 1.

Corpus annotation Three annotators, all native speakers of Arabic were hired to complete this task. They were provided with a list of Twitter user profiles and their corresponding URLs. They were asked to inspect each profile by checking if the user indicated his/her location, checking his/her tweets, and label it with its corresponding country when possible. In the context of dialect identification, the country label here refers to the Twitter

DrBehbehaniAM	Kuwait	وتقولون ليش الأطباء داشين تويتر
DrBehbehaniAM	Kuwait	لا القافلة وقفت ، ولا الكلاب تعبت .
DrBehbehaniAM	Kuwait	مسالنور على قوم التويتر مساء ما يليق الا بالطيبين مساء الابتسامة من القلب
HederAshraf	Egypt	الى عايز يجيب جون فى الأهلى يجيب زى ما هو عايز بس الهدف يتحسب أو لا دى بتاعتنا
HederAshraf	Egypt	يعنى هى جت عالجون دة بس الدورى مليان بلاوى
HederAshraf	Egypt	هههههههههه مش عارف انت مكبر الموضوع كدة ليه انا حاسيت اننا هنحرر فلسطين
samykhalildz	Algeria	اعطيلهم فكرة علاش راك غير تنتقد ..
samykhalildz	Algeria	صلاة المغرب رسميا سيصبح اسمها صلاة الجزائر في السعودية
samykhalildz	Algeria	تسريبات : نائب افلاني من بليدة قد يكون متهما في قضية كوكايين العاصمة.

Table 4: Three examples from the MADAR Twitter Corpus.

user geopolitical identity with the assumption that such identity could be expressed either explicitly through the location indicated in the Twitter bio section, or implicitly through dialectal and MSA usage in the tweets. Annotators were instructed not to rely only on the location provided by the user, and were invited to use all the extra-linguistic information available in the profile such as images, proclamations of loyalty and pride, etc. They were also allowed to check other sources such as corresponding Facebook profiles, if available, to confirm the user's country. Profiles may be marked as *Non Person*, *Non Arab* or *too hard to guess*. To measure inter-annotator agreement, a common set of 150 profiles were labeled by all annotators. They obtained an average Cohen Kappa score of 80.16%, which shows substantial agreement.

We discarded all profiles that became unavailable after the collection step, as well as profiles marked as *Non Person*, *Non Arab* or *too hard to guess*. Our final data set contained 2,980 country-labeled profiles. Three examples from the MADAR Twitter Corpus are shown in Table 4.

The distribution of the Twitter profiles by country is given in Table 5. The majority of the users obtained were from Saudi Arabia, representing 35.91% of the total profiles. Since there were zero Twitter user profiles from the Comoros in our dataset, we exclude it from the shared subtask.

Dataset splits and additional features We split the Twitter corpus into train, dev and test sets. The split distribution is given in Table 6. Participants were provided with the pointers to the tweets together with automatically detected language by Twitter, as well as the 26 confidence scores of the Salameh et al. (2018) system for the 26-way classification task applied per tweet.

Country	**Count**	**Percentage**
Saudi Arabia	1,070	35.91
Kuwait	213	7.15
Egypt	173	5.81
UAE	152	5.10
Oman	138	4.63
Yemen	136	4.56
Qatar	126	4.22
Bahrain	113	3.79
Jordan	107	3.59
Sudan	100	3.36
Iraq	99	3.32
Algeria	92	3.09
Libya	78	2.62
Palestine	74	2.48
Lebanon	66	2.21
Somalia	60	2.01
Tunisia	51	1.71
Syria	48	1.61
Morocco	45	1.51
Mauritania	37	1.24
Djibouti	2	0.07
Comoros	0	0
Total Annotated	**2,980**	**100**

Table 5: Distribution of the tweet Profiles by country label in the MADAR Twitter Corpus.

	Users	**Tweets**
Twitter Corpus train	2,180	217,592
Twitter Corpus dev	300	29,869
Twitter Corpus test	500	49,962

Table 6: Distribution of the train, dev and test sets provided for Subtask 2.

Team	Affiliation	Tasks
A3-108 (Mishra and Mujadia, 2019)	International Institute of Information Technology (IIIT), Hyderabad, India	1,2
ADAPT-Epita (De Francony et al., 2019)	Cork Institute of Technology, Ireland; and EPITA, France	1
ArbDialectID (Qwaider and Saad, 2019)	Göteborg Universitet, Sweden; and The Islamic University of Gaza, Palestine	1
CURAISA (Elaraby and Zahran, 2019)	Raisa Energy; and Cairo University, Egypt	2
DNLP	Dalhousie University, Canada	1
JHU (Lippincott et al., 2019)	Johns Hopkins University, USA	1,2
JUST (Talafha et al., 2019a)	Jordan University of Science and Technology, Jordan	1
khalifaaa	Cairo University, Egypt	1
LIU_MIR (Kchaou et al., 2019)	Laboratoire d'Informatique de l'Universitè du Mans (LIUM), France; and Multimedia, InfoRmation Systems, and Advanced Computing Laboratory (MIRACL), Tunisia	1
Mawdoo3 _AI_Team (Ragab et al., 2019; Talafha et al., 2019b)	Mawdoo3, Jordan, Egypt and Italy	1,2
MICHAEL (Ghoul and Lejeune, 2019)	Sorbonne University, France	1
Eldesouki	Qatar Computing Research Institute (QCRI), Qatar	1
OscarGaribo	Universitat Politècnica de València and Autoritas Consulting, Spain	1
QC-GO (Samih et al., 2019)	Qatar Computing Research Institute (QCRI), Qatar; and Google Inc, USA	1,2
QUT (Eltanbouly et al., 2019)	Qatar University, Qatar	1
Safina	Cairo University, Egypt	1
SMarT (Meftouh et al., 2019)	Badji Mokhtar University, Algeria; Lorraine University, France; and École Normale Supérieure de Bouzaréah, Algeria	1
Speech Translation (Abbas et al., 2019)	Le Centre de Recherche Scientifique et Technique pour le Développement de la Langue Arabe (CRSTDLA), Algeria and University of Trento, Italy	1,2
Trends (Fares et al., 2019)	Alexandria University, Egypt	1,2
UBC-NLP (Zhang and Abdul-Mageed, 2019)	The University of British Columbia, Canada	2
ZCU-NLP (Přibáň and Taylor, 2019)	Západočeská Univerzita v Plzni, Czech Republic	1,2

Table 7: List of the 21 teams that participated in Subtasks 1 and 2 of the MADAR Shared Task.

4 Participants and Systems

A total of 21 teams from 15 countries in four continents participated in the shared task. Table 7 presents the names of participating teams and their affiliations. 19 teams participated in Subtask 1; and 9 in Subtask 2. The submitted systems included a diverse set of approaches that incorporated machine learning, ensemble learning and deep learning frameworks, and exploited a various range of features. Table 8 summarizes the approaches adopted by each team for the two subtasks. In the table, ML refers to any non-neural machine learning technique such as multinomial naive Bayes (MNB) and support vector machines (SVM). Neural refers to any neural network based model such as bidirectional long short-term memory (BiLSTM), or convolutional neural network (CNN). In terms of features, word and character ngram features (in Table 8 as WC), sometimes weighted with TFIDF, were among the most commonly used features. Language-model based features (in Table 8 as LM) were also used a lot. A few participants used pre-trained embeddings. All details about the different systems submitted could be found in the papers cited in Table 7.

5 Results and Discussion

5.1 Subtask 1 Results

Table 9 presents the results for Subtask 1. The last two rows are for the state-of-the-art system by Salameh et al. (2018), and the character 5-gram LM based baseline system from Zaidan and Callison-Burch (2013). The best result in terms of macro-averaged F1-score is achieved by the winning team ArbDialectID (67.32%), very closely followed by SMART and Mawdoo3_AI_Team with F1 scores of 67.31% and 67.20%, respectively. The top five systems all used non-neural ML models and word and character features. Two of the top three systems used ensemble methods (See Table 8). Generally, the neural methods did not do well. This is consistent with what Salameh et al. (2018) reported, and is likely the result of limited training data. It is noteworthy that none

		Techniques			Features		
Team	**F1**	**ML**	**Neural**	**Ensemble**	**WC**	**LM**	**Embeddings**
Subtask 1							
ArbDialectID	67.32	X		X	X		
SMarT	67.31	X			X		
Mawdoo3 LTD	67.20	X		X	X	X	
Safina	66.31	X			X	X	
A3-108	66.28	X			X	X	
ZCU-NLP	65.82	X		X	X	X	
Trends	65.66	X	X		X		X
QUT	64.45	X			X		
DNLP	64.20	X					
ADAPT-Epita	63.02	X					X
Eldesouki	63.02	X	X		X		
Speech Translation	62.12						
JHU	61.83			X		X	X
QC-GO	58.72		X				X
OscarGaribo	58.44		X				
LIU_MIR	56.66	X				X	
khalifaaa	53.21	X	X				
MICHAEL	52.96	X	X				
JUST*	66.33	X			X	X	
Subtask 2							
UBC-NLP	71.70		X				X
Mawdoo3 LTD	69.86	X			X		
QC-GO	66.68		X				X
CURAISA	61.54		X		X	X	
A3-108	57.90	X			X	X	
JHU	50.43		X	X			X
ZCU-NLP	47.51	X				X	
Speech Translation	3.82	X			X		
Trends	3.32	X	X		X		X

Table 8: Approaches (techniques and features) adopted by the participating teams in Subtasks 1 and 2. ML refers to any non-neural machine learning technique such as MNB, SVM, etc. Neural refers to any neural network based model such as BILSTM, CNN, GRUs, etc. LM refers to language-model based features. WC corresponds to word and character features.

of the competing systems overcame the previously published Salameh et al. (2018) result.

5.2 Subtask 2 Results

Table 10 presents the results for Subtask 2. The last three rows are for three baselines. First is a maximum likelihood estimate (MLE) baseline, which was to always select Saudi Arabia (the majority class). Second is the state-of-the-art system setup of Salameh et al. (2018) trained on the MADAR Twitter Corpus data. And third is the baseline system from Zaidan and Callison-Burch (2013) using character 5-gram LM models. The winning system is UBC-NLP beating the next system by almost 2% points. The best performer in this subtask used a neural model (See Table 8).

Unavailable Tweets One of the concerns with any Twitter-based evaluation is that some of the tweets and Twitter users included in the manually annotated training, development and test data sets become unavailable at the time of the shared task. In our shared task, the percentage of missing tweets from train and development immediately after the conclusion of the shared task was 12.7%, which is basically the upper limit on unavailability. The corresponding number for unavailable Twitter users was 7.6%. The range of percentages of unavailable tweets as reported by some of the participating teams is between 6.0% and 11.3%. However, there seems to be no significant effect on the systems performance, as the correlation between the percentage of unavailable tweets and performance rank is -62%. The range in percentages of unavailable tweets for the test set is much smaller (11.5% to 12.1%) since all the teams received the test set at the same time and much later after the training and development data release.

6 Conclusion and Outlook

In this paper, we described the framework and the results of the MADAR Shared Task on Arabic Fine-Grained Dialect Identification. In addition to making a previously collected city-level dataset publicly available, we also introduced a new country-level dataset built specifically for this shared task. The unexpected large number of participants is an indication that there is a lot of interest in working on Arabic and Arabic dialects. We plan to run similar shared tasks in the near future, possibly with more naturally occurring (as opposed to commissioned) datasets. We also plan to coordinate with the VarDial Arabic Dialect Identification organizers to explore ways of leveraging the resources created in both competitions.

Acknowledgments

We would like to thank our dedicated annotators who contributed to the building the MADAR Twitter Corpus: Anissa Jrad, Sameh Lakhal, and Syrine Guediche. This publication was made possible by grant NPRP 7-290-1-047 from the Qatar National Research Fund (a member of the Qatar Foundation). The statements made herein are solely the responsibility of the authors.

Team	F1	Precision	Recall	Acc_{city}	$\text{Acc}_{country}$	Acc_{region}
ArbDialectID	**67.32 (1)**	67.60 (2)	67.29 (2)	67.29 (2)	75.23 (2)	84.42 (5)
SMarT	67.31 (2)	**67.73 (1)**	**67.33 (1)**	**67.33 (1)**	**75.69 (1)**	**85.13 (1)**
Mawdoo3 LTD	67.20 (3)	67.53 (3)	67.08 (3)	67.08 (3)	75.19 (3)	84.75 (2)
Safina	66.31 (4)	66.68 (4)	66.48 (4)	66.48 (4)	75.02 (5)	84.48 (4)
A3-108	66.28 (5)	66.56 (5)	66.31 (5)	66.31 (5)	75.15 (4)	84.62 (3)
ZCU-NLP	65.82 (6)	66.45 (6)	65.85 (6)	65.85 (6)	74.27 (6)	84.10 (6)
Trends	65.66 (7)	65.79 (7)	65.75 (7)	65.75 (7)	74.08 (7)	83.46 (7)
QUT	64.45 (8)	64.99 (8)	64.58 (8)	64.58 (8)	73.29 (8)	83.02 (8)
DNLP	64.20 (9)	64.72 (9)	63.98 (9)	63.98 (9)	72.27 (9)	82.52 (10)
ADAPT-Epita	63.02 (10)	63.43 (11)	63.08 (10)	63.08 (10)	72.15 (10)	82.56 (9)
Eldesouki	63.02 (11)	63.53 (10)	63.06 (11)	63.06 (11)	71.96 (11)	82.23 (11)
Speech Translation	62.12 (12)	63.13 (13)	62.17 (12)	62.17 (12)	71.23 (12)	81.71 (13)
JHU	61.83 (13)	62.06 (14)	61.90 (13)	61.90 (13)	71.06 (13)	81.88 (12)
QC-GO	58.72 (14)	59.77 (15)	59.12 (14)	59.12 (14)	69.29 (14)	81.29 (14)
OscarGaribo	58.44 (15)	58.58 (16)	58.52 (15)	58.52 (15)	67.67 (15)	79.31 (15)
LIU_MIR	56.66 (16)	57.06 (17)	56.52 (16)	56.52 (16)	67.62 (16)	78.77 (16)
khalifaaa	53.21 (17)	63.14 (12)	53.37 (17)	53.37 (17)	64.71 (17)	78.19 (17)
MICHAEL	52.96 (18)	53.38 (18)	53.25 (18)	53.25 (18)	62.29 (18)	73.90 (18)
JUST*	66.33 (19)	66.56 (19)	66.42 (19)	66.42 (19)	74.71 (19)	84.54 (19)
Salameh et al (2018)	67.89	68.41	67.75	67.75	76.44	85.96
Character 5-gram LM	64.74	65.01	64.75	64.75	73.65	83.40

Table 9: Results for Subtask 1. Numbers in parentheses are the ranks. The table is sorted on the macro F1 score, the official metric,. The JUST system result was updated after the shared task as their official submission was corrupted. The last two rows are for baselines ((Salameh et al., 2018) and (Zaidan and Callison-Burch, 2013)).

Team	F1	Precision	Recall	$\text{Acc}_{country}$	Acc_{region}
UBC-NLP	**71.70 (1)**	82.59 (3)	**65.63 (1)**	**77.40 (1)**	**88.40 (1)**
Mawdoo3 LTD	69.86 (2)	78.51 (4)	65.20 (2)	76.20 (2)	87.60 (2)
QC-GO	66.68 (3)	82.91 (2)	59.36 (4)	70.60 (4)	80.60 (5)
CURAISA	61.54 (4)	67.27 (7)	60.32 (3)	72.60 (3)	83.40 (3)
A3-108	57.90 (5)	**83.37 (1)**	47.73 (5)	67.20 (5)	81.60 (4)
JHU	50.43 (6)	70.45 (6)	43.18 (6)	62.20 (6)	77.80 (6)
ZCU-NLP	47.51 (7)	74.16 (5)	38.88 (7)	59.00 (7)	72.80 (7)
Speech Translation	3.82 (8)	5.22 (9)	5.37 (8)	5.00 (9)	31.80 (9)
Trends	3.32 (9)	6.82 (8)	4.97 (9)	33.00 (8)	61.40 (8)
MLE - KSA	2.64	1.79	5.00	35.80	64.20
Salameh et al (2018)	13.08	41.91	11.15	42.20	66.80
Character 5gram LM model	50.31	66.15	43.90	65.80	79.20

Table 10: Results for Subtask 2. Numbers in parentheses are the ranks. The table is sorted on the macro F1 score, the official metric. The last three rows are for baselines.

References

Mourad Abbas, Mohamed Lichouri, and Abded Alhakim Freihat. 2019. ST MADAR 2019 Shared Task: Arabic Fine-Grained Dialect Identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop (WANLP 2019)*, Florence, Italy.

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 11th Language Resources and Evaluation Conference*, Miyazaki, Japan. European Language Resource Association.

Nora Al-Twairesh, Hend Al-Khalifa, and Abdulmalik AlSalman. 2016. AraSenTi: Large-Scale Twitter-Specific Arabic Sentiment Lexicons. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, Berlin, Germany.

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.

Kareem Darwish, Hassan Sajjad, and Hamdy Mubarak. 2014. Verifiably Effective Arabic Dialect Identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar.

Gaël De Francony, Victor Guichard, Praveen Joshi, Haithem Afli, and Abdessalam Bouchekif. 2019. Hierarchical Deep Learning for Arabic Dialect Identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop (WANLP 2019)*, Florence, Italy.

Mahmoud El-Haj, Paul Rayson, and Mariam Aboelezz. 2018. Arabic Dialect Identification in the Context of Bivalency and Code-Switching. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

Mohamed Elaraby and Ahmed Zahran. 2019. A Character Level Convolutional Bilstm for Arabic Dialect Identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop (WANLP 2019)*, Florence, Italy.

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.

Sohaila Eltanbouly, May Bashendy, and Tamer Elsayed. 2019. Simple but not Na{"ive: Fine-Grained Arabic Dialect Identification using only N-Grams. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop (WANLP 2019)*, Florence, Italy.

Youssef Fares, Zeyad El-Zanaty, Kareem Abdel-Salam, Muhammed Ezzeldin, Aliaa Mohamed, Karim El-Awaad, and Marwan Torki. 2019. Arabic Dialect Identification with Deep learning and Hybrid Frequency Based Features. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop (WANLP 2019)*, Florence, Italy.

Dhaou Ghoul and Gaël Lejeune. 2019. MICHAEL: Mining Character-level Patterns for Arabic Dialect Identification (MADAR Challenge). In *Proceedings of the Fourth Arabic Natural Language Processing Workshop (WANLP 2019)*, Florence, Italy.

Saméh Kchaou, Fethi Bougares, and Lamia Hadrich-Belguith. 2019. LIUM-MIRACL Participation in the MADAR Arabic Dialect Identification Shared Task. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop (WANLP 2019)*, Florence, Italy.

Tom Lippincott, Pamela Shapiro, Kevin Duh, and Paul McNamee. 2019. JHU System Description for the MADAR Arabic Dialect Identification Shared Task. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop (WANLP 2019)*, Florence, Italy.

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 Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, Osaka, Japan.

Karima Meftouh, Karima Abidi, Salima Harrat, and Kamel Smaili. 2019. The SMarT Classifier for Arabic Fine-Grained Dialect Identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop (WANLP 2019)*, Florence, Italy.

Pruthwik Mishra and Vandan Mujadia. 2019. Arabic Dialect Identification for Travel and Twitter Text. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop (WANLP 2019)*, Florence, Italy.

Hamdy Mubarak and Kareem Darwish. 2014. Using Twitter to collect a multi-dialectal corpus of Arabic. In *Proceedings of the Workshop for Arabic Natural Language Processing (WANLP)*, Doha, Qatar.

Pavel Přibáň and Stephen Taylor. 2019. ZCU-NLP at MADAR 2019: Recognizing Arabic Dialects. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop (WANLP 2019)*, Florence, Italy.

Chatrine Qwaider and Motaz Saad. 2019. ArbDialectID at MADAR Shared Task 1: Language Modelling and Ensemble Learning for Fine Grained Arabic Dialect Identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop (WANLP 2019)*, Florence, Italy.

Ahmad Ragab, Haitham Seelawi, Mostafa Samir, Abdelrahman Mattar, Hesham Al-Bataineh, Mohammad Zaghloul, Ahmad Mustafa, Bashar Talafha, Abed Alhakim Freihat, and Hussein T. Al-Natsheh. 2019. Mawdoo3 AI at MADAR Shared Task: Arabic Fine-Grained Dialect Identification with Ensemble Learning. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop (WANLP 2019)*, Florence, Italy.

Fatiha Sadat, Farnazeh Kazemi, and Atefeh Farzindar. 2014. Automatic Identification of Arabic Dialects in Social Media. In *Proceedings of the Workshop on Natural Language Processing for Social Media (SocialNLP)*, pages 22–27, Dublin, Ireland.

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.

Wael Salloum, Heba Elfardy, Linda Alamir-Salloum, Nizar Habash, and Mona Diab. 2014. Sentence Level Dialect Identification for Machine Translation System Selection. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, Baltimore, Maryland, USA.

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 (WANLP 2019)*, Florence, Italy.

Toshiyuki Takezawa, Genichiro Kikui, Masahide Mizushima, and Eiichiro Sumita. 2007. Multilingual Spoken Language Corpus Development for Communication Research. *Computational Linguistics and Chinese Language Processing*, 12(3):303–324.

Bashar Talafha, Ali Fadel, Mahmoud Al-Ayyoub, Yaser Jaraweh, Mohammad Al-Smadi, and Patrick Juola. 2019a. Team JUST at the MADAR Shared Task on Arabic Fine-Grained Dialect Identification . In *Proceedings of the Fourth Arabic Natural Language Processing Workshop (WANLP 2019)*, Florence, Italy.

Bashar Talafha, Wael Farhan, Ahmed Altakrouri, and Al-Natsheh Hussein. 2019b. Mawdoo3 AI at MADAR Shared Task: Arabic Tweet Dialect Identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop (WANLP 2019)*, Florence, Italy.

Wajdi Zaghouani and Anis Charfi. 2018. ArapTweet: A Large Multi-Dialect Twitter Corpus for Gender, Age and Language Variety Identification. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

Omar Zaidan and Chris Callison-Burch. 2013. Arabic dialect identification. *Computational Linguistics*.

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 Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, Valencia, Spain.

Chiyu Zhang and Muhammad Abdul-Mageed. 2019. No Army, No Navy: BERT Semi-Supervised Learning of Arabic Dialects. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop (WANLP 2019)*, Florence, Italy.