LIUM-MIRACL Participation in the MADAR Arabic Dialect Identification Shared Task

Saméh Kchaou
University of Sfax, Tunisia
samehkchaou4@gmail.com

Fethi Bougares
University of Le Mans, France
fethi.bougares@univ-lemans.fr

Lamia Hadrich Belguith
University of Sfax, Tunisia
lamia.belguith@gmail.com

Abstract

This paper describes the joint participation of the LIUM and MIRACL Laboratories at the Arabic dialect identification challenge of the MADAR Shared Task (Bouamor et al., 2019) conducted during the Fourth Arabic Natural Language Processing Workshop (WANLP 2019). We participated to the Travel Domain Dialect Identification subtask. We built several systems and explored different techniques including conventional machine learning methods and deep learning algorithms. Deep learning approaches did not perform well on this task. We experimented several classification systems and we were able to identify the dialect of an input sentence with an F1-score of **65.41%** on the official test set using only the training data supplied by the shared task organizers.

1 Introduction

Dialect can be defined as the language characteristics of a specific community (Etman and Beex, 2015). For all their daily communications, Arabic speakers use their local dialect. Dialects are commonly known as spoken or colloquial Arabic, acquired naturally as their mother tongue.

Being able to identify the dialect of a given sentence is a fundamental step for various applications such as machine translation, speech recognition and multiple Natural Language Processing (NLP) related services. Therefore, the dialect identification task has been the subject of several earlier research and exploration activities. For instance, Arabic dialect identification in speech transcripts was introduced as a subtask of the Discriminating between Similar Languages (DSL) Shared Task of the Third, Fourth and the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial) (Malmasi et al., 2016; Zampieri et al., 2017, 2018).

In practice, the number of existing dialects are as many as there are cities in the Arab world. Going to the city-level of granularity for dialect identification is a complex and expensive task. It is for this reason that earlier work in dialect identification generally study the problem at a region or country level (Zaidan and Callison-Burch, 2014). In this respect, dialects are generally classified into five main groups: *Maghrebi*, *Egyptian*, *Levantine*, *Gulf*, and *Iraqi* (El Haj et al., 2018; Elaraby and Abdul-Mageed, 2018).

Quite recently, and in contrast with the overall previous work, (Bouamor et al., 2018) presented the MADAR corpus which is now the existing resource with the greatest dialectal coverage. Indeed, the MADAR corpus includes 25 Arabic different dialects from east to west. The MADAR shared task is organized to make the most efficient use of this corpus. Dialect Identification (DID) is already a hard task, even when taking into account only 5 groups. This task became more perplexing when taking into consideration 25 groups of MADAR shared task. Indeed, taking into consideration additional dialects will reduce the overall classes dissimilarity and thus make the discrimination process harder. In the following sections of this paper, we will describe our participation to the MADAR Shared task. We investigate different classification techniques based on conventional machine learning algorithms with different kinds of features and various deep learning sequence2sequence architectures.

The paper is structured as follows: Section 2 describes the MADAR shared task and presents brief description of the training data. Section 3 presents a detailed overview of our systems and a discussion of our results. Finally, section 4 will draw a brief conclusion.

Proceedings of the Fourth Arabic Natural Language Processing Workshop, pages 219–223
Florence, Italy, August 1, 2019. ©2019 Association for Computational Linguistics

2 MADAR Shared Task

Arabic Dialect processing is a challenging task since dialects are mainly *spoken* and do not have an explicit written set of grammar rules. In this context, the MADAR corpus (Bouamor et al., 2018) is a valuable resource to push forward the field Arabic Dialect processing. The MADAR Dialect Identification shared task is partially based on this corpus. MADAR Shared task is the first DID shared task to target a large set of dialects. The challenge offered two subtasks: subtask 1 focuses on Travel Domain Dialect Identification, whereas subtask 2 is centred around Twitter User Dialect Identification. We will describe only the subtask 1 in which we have participated.

Subtask 1 Dataset: The provided data-sets are presented in table 1. The organizers provided a training and development sets from two sources created by translating the Basic Traveling Expression Corpus (BTEC) (Takezawa, 2006): *(i)* corpus-6, a large-scale additional sentences of the BTEC corpus of 5 regional representative dialects and MSA, *(ii)* corpus-26, a smaller-scale parallel corpus of 25 dialects in addition to MSA.

Corpus	#lines	#token
Corpus-26-train	41.6k	343.7k
Corpus-26-Dev	5.2k	43.7k
Corpus-6-train	54k	452.3k
Corpus-6-Dev	6k	49.6k
Corpus-26-Test	5.2k	43.1k

Table 1: Statistics of MADAR Subtask 1 Data Sets.

The shared task allowed participants to exploit all the data presented in Table 1 for the development of their DID systems[1]. Submission must be constrained in the sense that external manually annotated data sets are prohibited. Submissions are evaluated automatically on the test set *Corpus-26-Test*, using F1 score. Both Corpus-26-dev and Corpus-26-Test consist of 5.2k dialectal sentences, uniformly distributed over the 26 addressed dialects (200 sentences for each dialect). We note that we only use *corpus-26* train and dev to develop our DID systems.

3 LIU-MIR Submission

3.1 Data pre-processing

All Arabic dialects came from the same source, use the same character set, and share a large number of common words seen throughout their substantial vocabulary overlap. None of the existing Arabic dialects, at least until now, has an official status and none is regulated and taught in schools. As there is at present no dialect-specific computationally motivated pre-processing methods, our pre-processing is limited to a few steps of cleaning up applied to all the data without distinction. This includes the normalization of few arabic characters (أ, آ, إ, ي, ئ, ؤ[2]), the deletion of short vowels and tatweel[3] and the deletion of punctuation numbers and non arabic words.

3.2 Dialect Identification systems

In this section we present our experiments for MADAR Travel Domain Dialect Identification task. All our systems are constrained as we only used the supplied data from table 1.

3.3 Baseline systems

As a baseline for our DID system, we tried to reproduce the results presented in (Salameh et al., 2018). Just like them, we trained a Multinomial Naive Bayes (MNB) classifier using Word and character n-gram features. We also used Term Frequency-Inverse Document Frequency (Tf-Idf) scores learned on extracted character n-grams ranging from 1-grams to 5-grams.

	N-Gram Features		**F1 score**	
	Word	Char	Dev-26	Test-26
1.	1	-	59.64	57.42
2.	-	1	10.96	9.99
3.	-	1→5	56.44	54.34
4.	1	1	59.14	57.15
5.	1	1→3	60.07	58.51
6.	1	1→5	**60.97**	**59.21**

Table 2: MNB system using pre-processed training and evaluation data..

Table 2 reports the results of our baseline systems accuracy on the development set of CORPUS-26 (Dev-26). We performed several experiments using TF-IDF features of word and char

[1] Training data from MADAR-Shared-Task-Subtask-2 are also allowed but not used for our submission.

[2] This corresponds to >, |, <, y, and & with Buckwalter transliteration.

[3] A type of justification using characters elongation.

level n-grams. The best identification accuracy is obtained using uni-gram word level 1→5-gram character level.

We have tested also to evaluate the above systems without any pre-processing or normalization of the training data. The results are presented in the following table.

N-Gram Features		**F1 score**	
Word	Char	Dev-26	Test-26
1	-	64.42	63.33
-	1	16.84	15.51
-	1→5	62.16	60.63
1	1	64.45	63.52
1	1→3	65.29	64.52
1	1→5	**66.41**	**65.41**

Table 3: MNB system using raw training and evaluation data.

As shown in the table above, the dialect identification results are better when the systems are trained using the raw data. Knowing that data were created by translating sentences from English and French, we suspect that the translation was performed by one single person per dialect. If this is true, This could explain this result since the system learns as a side effect, to distinguish between the style of the translators (*i.e* spelling, punctuation, lexical choices, with or without vowels ...).

3.4 LM-based Systems

In addition to the baseline system presented in the previous section, we evaluated the identification performance using only n-gram word and character level Language Models(LM) trained using only corpus 26 training data. This has been done by directly comparing LM perplexity for each input sentence: Given an input sentence S to classify into one of k dialects $d_1, d_2, ..., d_k$ we select the dialect d^* of the model that gives the lowest perplexity on this sentence (*i.e* equation 1).

$$d^* = \arg\min_k PP_k(S) \quad (1)$$

For this experiment, we considered forward and backward word (LMWF, LMWB) and character-level (LMCF, LMCB) LMs, trained using sequences of words and characters in the reverse order. All LMs are 5-gram order, trained using KenLM toolkit with default parameters and Kneser–Ney smoothing (Heafield, 2011).

Table 4 presents the results of the DID systems with only LMs. While the character level LMs is lightly better than word-level LMs, both shows a lower accuracy compared to the MNB with word n-gram features (line 1 in table 3). The best LM based DID system is obtained using Backward character level LM with F1-score of **60.63** on the test-26 set.

	Word LMs		Char LMs	
	LMWF	LMWB	LMCF	LMCB
dev-26	60.34	60.34	61.07	**61.36**
test-26	60.07	60.15	60.21	**60.63**

Table 4: F1-scores of DID system using LM Scores

3.5 LM Scores as Features

In this section we present our attempt to integrate LMs scores as extra feature to the MNB classifier. Each sentence is evaluated using the 26 trained LMs presented in section 3.4 and their scores are used as input features to the MNB.

N-Gram Feat.		**LM Feat.**		**F1 score**
Word	Char	Char	Word	Dev-26
1	1→5	-	-	**66.41** (**65.41**)
1	1→5	F	-	65.97 (64.60)
1	1→5	B	-	66.03 (64.58)
1	1→5	FB	-	65.74 (63.73)
1	1→5	-	F	45.19 (61.51)
1	1→5	-	B	45.19 (61.78)
1	1→5	-	FB	44.76 (61.49)
1	1→5	F	F	46.09 (61.51)

Table 5: MNB system with N-Gram and LMs features.

As shown in table 5, adding LMs scores as features to MNB results in a decrease of F1-score on both dev-26 and test-26 sets. We tried both word and character level LMs trained in either Forward or Backward fashion. However, none of them and not even their combination had a positive effect on the system's accuracy.

3.6 Analysis and Discussion

In this section, we present an analysis of the classification results of our DID system. Table 6 presents the details of the F1-score per dialect. Overall, we can see that the system has a good prediction (high F1 score) for several dialects wheres the identification of other dialects are more challenging. This result is in accordance with the rate of token dissimilarity presented in Figure 2 of (Salameh et al., 2018). For example, ALG, SAN and MOS dialects have high pairwise token dissimilarity rate and they are also the easier to identify compared to CAI and RIY for instance.

Dialect	F1-score	Dialect	F1-score
ALE	0.64	JED	0.62
ALG	**0.78**	JER	0.57
ALX	0.74	KHA	0.69
AMM	0.56	MOS	**0.84**
ASW	0.60	MSA	0.69
BAG	0.65	MUS	0.54
BAS	0.68	RAB	0.70
BEI	0.67	RIY	0.56
BEN	0.67	SAL	0.56
CAI	0.53	SAN	0.71
DAM	0.56	SFX	0.72
DOH	0.63	TRI	0.75
FES	0.68	TUN	0.70

Table 6: F1-score per dialect of the best system on the test-26 set

After analyzing the full confusion matrix, we figured out that identification confusion tends to be bigger for geographical close cities. This is expected since sentences from close cities has a big vocabulary overlap and thus harder to discriminate. In order to investigate further this, we conducted a deeper analysis of two dialects belonging to the same geographical area. The primary objective of this study is to measure the upper bound of the classification accuracy for these dialects (*i.e.* the best possible prediction accuracy).

We conducted this analysis for the Top-2 most confused dialects: TUN (TUNIS) and SFX (SFAX). These dialects belongs to the same country, Tunisia, and present a high level of lexical similarity. In addition, there are the only ones for which we have native speakers. Table 7, presents the TUN *vs.* SFX dialects confusion matrix.

		Predicted	
		SFX	TUN
Actual	SFX	153	18
	TUN	41	123

Table 7: TUN - SFX dialects Confusion matrix.

As shown in Table 7, TUN and SFX dialects are source of confusion for the system. For instance, from among all the 200 SFX test sentences, 153 were well predicted and 18 predicted TUN. Similarly, 123 from the 200 TUN test sentences are well identified whereas 41 are predicted SFX.

In order to understand this substantial SFX-TUN confusion, we conducted a manual evaluation of the "*18 + 41*" sentences. This evaluation was performed as following: each of the incorrectly predicted 18 SFX sentences was presented to a TUN native speaker who decide whether the sentence seems to be a natural in his view, or whether he will formulate it differently. Table 8 presents examples of sentences from the test-26 and their transliteration.

Sentences	Label
فما حانوت قريب من هوني؟ fmA .hAnwt qryb mn hwny?	TUN
نحب مايو قياس اس، يعيشك n.hb mAyw qyAs As, y‘ y^sk	TUN
نحب تفاح يعيشك n.hb tfA.h y‘y^sk	TUN
وقتاش تسكر البوسته؟ wqtAy^s tskr Albwsth?	SFX
باهي، وتيل الشيراتون قريب من هوني؟ bAhy, wtyl Aly^syrAtwn qryb mn hwny?	SFX
دجاج، يعيشك djAj, y‘yy^sk	SFX

Table 8: Examples of TUN and SFX mis-classified sentences from test-26 with their ground truth label.

The evaluation has shown that, almost all the "*18 + 41*" studied examples may belongs to both dialect and hardly distinguishable even for native speakers. This exemplifies the increasing complexity the dialect identification task when we consider close dialects of a large lexical overlap.

4 Conclusion

In this paper we described our participation to the MADAR dialect identification task. We participated to the the Travel Domain Dialect Identification subtask where the goal is to design a system able to predict the correct dialect among 26 considered classes. We performed several experiments showing that the DID is a very challenging task. We were able to reach a F1-score of **65.41** on the official corpus-26-test set. We also conducted a manual assessment of the Top-2 most confused classes (SFX and TUN). We have found that almost all the confused SFX-TUN sentences are cases for which even a native speaker cannot decide. This shows that dialect identification is reaching its effective limit when considered dialects have many commonalities.

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

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.

Mahmoud El Haj, Paul Edward Rayson, and Mariam Aboelezz. 2018. Arabic dialect identification in the context of bivalency and code-switching. In *LREC 2018, Eleventh International Conference on Language Resources and Evaluation*, pages 3622–3627.

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, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

A. Etman and A. A. L. Beex. 2015. Language and dialect identification: A survey. pages 220–231.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, WMT '11, pages 187–197, Stroudsburg, PA, USA. Association for Computational Linguistics.

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 Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 1–14, Osaka, Japan.

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. 2006. Multilingual spoken language corpus development for communication research. In *Chinese Spoken Language Processing*, pages 781–791, Berlin, Heidelberg. Springer Berlin Heidelberg.

Omar F. Zaidan and Chris Callison-Burch. 2014. Arabic dialect identification. *Comput. Linguist.*, 40(1).

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 Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 1–15, Valencia, Spain. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Ahmed Ali, Suwon Shon, James Glass, Yves Scherrer, Tanja Samardžić, Nikola Ljubešić, Jörg Tiedemann, Chris van der Lee, Stefan Grondelaers, Nelleke Oostdijk, Antal van den Bosch, Ritesh Kumar, Bornini Lahiri, and Mayank Jain. 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 (VarDial)*, Santa Fe, USA.