MICHAEL: Mining Character-level Patterns for Arabic Dialect Identification (MADAR Challenge)

Dhaou Ghoul
STIH Lab, Sorbonne University
dhaou.ghoul@sorbonne-universite.fr

Gaël Lejeune
STIH Lab, Sorbonne University
gael.lejeune@sorbonne-universite.fr

Abstract

We present MICHAEL, a lightweight method developed for the MADAR shared task on travel domain Dialect Identification (DID). It uses character-level features and perform classification without any pre-processing. Character N-grams extracted from the original sentences are used to train a Multinomial Naive Bayes classifier. MICHAEL achieved an official score (accuracy) of 53.25% with $1 \leq N \leq 3$ but showed a much better result with character 4-grams (62.17%).

1 Introduction

The Arabic language is one of the most widely spoken language in the world, currently considered as the fifth language (Chung, 2008) with more than 330 million Arabic speakers. It is the official language of more than 22 countries. In its written form, commonly referred as Literary Arabic, it is divided into two categories: Classical Arabic and Modern Standard Arabic (MSA). However, Arabic speakers mostly use dialects which are a linguistic variant of classical Arabic with their own features, varying with respect to the country or the region. If MSA is used only for written and official communication, dialects are used for oral communication as well as for many device mediated communication forms: email, sms, chat or blogs. Therefore, Arabic dialects identification (DID) has become a very important pre-processing step that attracts many attention from NLP research. Indeed, the knowledge about the dialect of an input text is useful in several NLP tasks such as sentiment analysis (Al-Twairesh et al., 2016).

We propose a simple, light-weight, character-based method to classify Arabic sentences into 26 classes (25 dialects + MSA) based on the MADAR corpus provided for this competition (Bouamor et al., 2019). This paper is organized as follows: in Section 2, we present some related word for DID. In section 3, we describe some aspects of the Arabic dialects and in section 4 we present the MADAR dataset and we introduce MICHAEL, the system we designed to tackle the DID task. Finally, we show our results in Section 5 and give some future directions in section 6

2 Previous Work

Arabic Dialect Identification is a very difficult task because of several factors like the lack of NLP tools that deal with Arabic variants. So far, the researchers have tried to address this task using different methods.

Salameh *et al.* (Salameh et al., 2018), presented a MNB (Multinomial Naive Bayes) classifier trained to identify a tweet among 26 classes (MSA+25 dialects) using a large-scale of parallel sentences (Bouamor et al., 2018). Their models reach 67.9% accuracy for sentences with an average length of 7 word and reached more than 90% with 16 words.
Elfardy and Diab (Elfardy and Diab, 2013) proposed a supervised method for identifying whether a given sentence in prevalently MSA or Egyptian using the Arabic online commentary dataset(AOC) (Zaidan and Callison-Burch, 2011). Their system achieves an accuracy of 85.5% on an Arabic online-commentary dataset.

Najafian *et al.* (Najafian et al., 2018), presented different approaches for Dialect Identification (DID) in Arabic broadcast speech using use Support Vector Machines (SVM), and Convolutional Neural Networks (CNN) as backend classifiers. The final system merges these results and obtains 24.7% and 19.0% relative error rate reduction compared to conventional phonotactic DID, and i-vectors with bottleneck features. Rabee *et al.* (Naser and Hanani, 2018), describes an Automatic Dialect Recognition (ADI) system for the VarDial 2018 challenge, with the goal of distinguishing four major Arabic dialects, as well as Modern Standard Arabic (MSA) using four sys-

Proceedings of the Fourth Arabic Natural Language Processing Workshop, pages 229–233
Florence, Italy, August 1, 2019. ©2019 Association for Computational Linguistics

tems. The first system uses word transcriptions and tries to recognize the speaker's dialect by modeling the word sequence of each dialect. The second one aims to recognize the dialect by modeling the telephonesequences produced by non-Arabic telephone recognition devices. The other two systems use GMM trained in acoustic functions to recognize the dialect. This system reached 68.77% in micro F1. Elaraby *et al.* (Elaraby and Abdul-Mageed, 2018), presented a deep learning models for DID taking advantage of the performance of several conventional machine learning models under different conditions. Their model showed a 87.65% score in accuracy for the binary task (MSA vs. dialects), 87.4% for the 3 class task (Egyptian, Gulf and Levantine).

3 The Dialectal Varieties of Arabic

Arabic language is a rather generic term that refers in fact to many variants and dialects. Nowadays, one can consider that Arabic language is divided into three major categories: classical Arabic, standard Arabic (MSA) and dialectal Arabic. The 2019 MADAR competition focused on the latter.

Dialectal Arabic is a proper form of the Arabic language used in everyday communication, usually called ”darija”. It varies from one country to another and even from one region to another within the same country. All Arab countries have their own dialects that are more or less close to each other. The differences the dialects exhibit depend mainly on the history of each country and its geographical location. For example, the Tunisian dialect (TUN) integrates several borrowings from French language as it has been colonized by France. Words like ”*stylo*” (pen/pencil) or ”*cartable*” (schoolbag) are examples of borrowings completely integrated into TUN. According to Zaidan and Callison-Burch (2014), arabic dialects can be classified into five major classes (these classes can have several subclasses):

- **Egyptian:** The most widely understood dialect, due to a thriving Egyptian television and movie industry (Haeri, 2003).

- **Levantine:** A set of dialects that differ somewhat in pronunciation and intonation, but are largely equivalent in written form. They are closely related to Aramaic (Amara, 2010).

- **Gulf:** Probably the closest to MSA, perhaps because the current form of MSA evolved from an Arabic variety originating in the Gulf region. There are differences between Gulf and MSA but Gulf kept more of MSA's verb conjugation than other dialects (Versteegh, 2001).

- **Iraqi:** Despite its similarity to Gulf dialects it exhibits some very distinctive features in terms of prepositioning, verb conjugation and pronunciation (Mitchell, 1993).

- **Maghrebi:** These dialects were influenced by both French and Berber languages. The Western-most varieties could be unintelligible for speakers from other regions in the Middle East, especially in spoken form. Maghreb is a large region with more variation than regions like the Levant or the Gulf. It makes it probably easier to distinguish its local variants : Tunisia, Algeria, Morocco, Libya. . . (Tilmatine, 1999).

Arabic dialect differ from one another and from MSA on several levels of linguistic representation such as phonology, morphology, lexicon and syntax. Table 1 exhibits examples of differences between some dialects. For instances, the phonem ”qaf” (first column) will not have the same pronounciation in all the dialects. In the second column one can see that the future tense is not marked by the same morpheme in each variant. The syntax of negation (third column) is not the same in Maghrebian dialects and in othe dialects. Regarding lexicon (fourth column) the concept ”car” in ALG and MAR dialects reflects a borowing from the French term ”automobile”.

	Phon.	Morph.	Synt.	Lex.
MSA	qaf	s or swf	mA	sayyaara
ALG	qaf and /g/	ghadi or rH	mA	tomobile
EGY	hamza	h	muw	3arabiyya
GUL	/g/	ba	lA	sayyaara
LEV	hamza	H or rH	muw	sayyaara
MAR	qaf	ghadi	mA	tomobile
TUN	qaf and /g/	bAsh	mA	krhba

Table 1: Examples of differences between MSA and ALG (Algeria), EGY (Egyptian), GUL (Gulf), LEV (levantine), MOR (Moroccan) and TUN (Tunisian) regarding phonetics, morphology, syntax and lexicon.

4 Arabic Dialect Identification: Methods for classification

4.1 Some Difficulties of Arabic DID

Despite the differences between the different dialects, their automatic identification remains a very difficult task, even impossible in some cases. This difficulty is due to several factors:

- Shared lexicon: dialects have a common vocabulary and a dialectal sentence can contain several dialects as well as MSA.
- Grammatical Ambiguity: some identical words are used with different functions. For example, the word ”Tyb” can be an adjective in some dialects and an interjection in others.
- Homonyms: mostly due to the omission of short vowels, a dialectal word can have the same spelling as an MSA word but an entirely different meaning. This includes strongly dialectal words such as *dwl*: it is either the Egyptian (EGY) word for ”these” (pronounced dowl) or the MSA word for ”countries” (pronounced duwal) (Zaidan and Callison-Burch, 2014).

4.2 Data: The MADAR corpus

The purpose of the shared task is to give each short sentence a label among 26 avialable labels. We took advantage of the MADAR corpus supplied for the competition in order to train various classifiers. We did not use anay external resource. The MADAR corpus has been created by translating sentences from the Basic Traveling Expression Corpus (BTEC) from English and French to the different dialects. This corpus has been splitted into Train, Validation and Test sets, they are priefly presented in Table 2.

Datasets	Train	Dev	Test
# sentences	41,600	5,200	5,200
# words	336,342	42,586	36,811
# characters	1,301,599	166,898	162,185

Table 2: Size of the Train, Dev and Test sets

4.3 Method: Character N-grams

MICHAEL has been built on the assumption that the features most prone to discriminate languages are found at character-level. With this idea in mind we tried different classifiers but quickly found that, under the technical constraints we were facing, Naive Bayes algorithms were the most appropriate for such a multi-class problem. The One VS Rest implementation of SVM we tested were unable to reach a result and we did not want to train 26 different classifiers separately. We used the SCI-KIT LEARN implementation of MNB and it proves quickly that among the NB implementations of this library, the Multinomial Naive Bayes (MNB) was the most efficient. We will show in the next section different learning configurations and various size of n-grams for feature engineering.

Trained on Tested on	Train Set Dev Set	Train Set Test Set	Train+Dev Test Set
$N = 1$	19.08	18.46	18.48
$1 \leq N \leq 2$	40.04	37.29	37.44
$N = 2$	42.62	39.90	40.38
$1 \leq N \leq 3$	55.00	**53.25**	53.54
$2 \leq N \leq 3$	56.17	54.31	54.40
$N = 3$	58.25	57.50	57.92
$1 \leq N \leq 4$	60.73	59.62	59.88
$2 \leq N \leq 4$	61.21	60.04	60.25
$3 \leq N \leq 4$	62.44	60.88	61.42
$N = 4$	62.96	**61.94**	**62.17**
$1 \leq N \leq 5$	62.65	60.98	61.71
$2 \leq N \leq 5$	63.17	61.02	61.77
$3 \leq N \leq 5$	**63.48**	61.65	62.12
$5 \leq N \leq 5$	62.62	61.71	61.88
$N = 5$	60.71	59.77	60.48

Table 3: Results for the Multinomial Naive Bayes Classifier, character N-grams with various range of N from $N_{min} = 1$ to $N_{max} = 5$ with different training and testing configurations (blue score is our official score)

5 Results and Error Analysis

5.1 Results

The results obtained by MICHAEL are shown on Table 3. One can see that character $1 - grams$ ($N_{min} = N_{max} = 1$) alone can achieve more than 18% in accuracy which is an interesting result for a 26-class task. Increasing the maximum size of the N-grams increases the accuracy quickly: +19 percentage points (pp) with $N_{max} = 2$ and another 16 points with $N_{max} = 3$. The gain with $N_{max} = 4$ is lower but it is still a 6 pp gain.

Working on the minimal size of the n-grams is also a good way to improve the score. In our particular learning setting, removing short n-grams helps to improve the results. For instance with $N_{max} = 3$, setting $N_{min} = 3$ instead of $N_{min} = 1$ improves the accuracy by 4 percentage points. Finally, the best results were obtained with 4-grams.

	Maghreb							Egyptian				S. Levant			N. Levant			Iraqi			Gulf					
	ALG	BEN	FES	RAB	SFX	TRI	TUN	ALX	ASW	CAI	KHA	AMM	JER	SAL	ALE	BEI	DAM	BAG	BAS	MOS	DOH	JED	MUS	RIY	SAN	MSA
ALG	**153**	3	5	6	4	3	5	1	0	3	2	0	0	2	0	1	0	0	0	0	0	0	5	1	0	1
BEN	7	**127**	2	3	2	8	0	2	0	0	0	4	6	3	1	3	3	3	3	2	3	5	5	9	4	0
FES	8	1	**135**	**36**	1	0	1	1	2	1	2	2	3	1	2	5	2	1	2	0	2	2	2	1	2	0
RAB	7	2	**34**	**138**	3	2	6	1	2	2	1	1	1	3	1	1	0	0	1	0	1	0	0	0	1	0
SFX	3	5	5	4	**149**	3	**47**	0	0	1	1	1	1	2	0	1	1	2	1	0	2	4	1	1	2	2
TRI	2	**11**	0	4	3	**145**	3	0	3	1	3	1	1	1	1	1	1	1	2	6	5	3	0	2	4	0
TUN	1	1	1	1	**22**	3	**119**	0	1	2	0	0	1	0	1	3	1	0	0	0	1	0	1	0	0	0
ALX	2	0	1	1	0	0	1	**143**	**27**	**20**	3	4	3	2	0	2	2	0	1	2	2	3	2	1	0	2
ASW	0	7	1	0	0	3	0	**14**	**116**	**36**	**11**	4	2	4	1	3	3	3	1	0	1	6	3	0	2	0
CAI	1	1	2	0	0	2	1	**12**	**22**	**88**	2	4	2	3	0	4	2	1	1	0	0	2	2	4	3	1
KHA	3	3	1	0	0	5	0	8	3	**14**	**139**	3	2	2	2	4	2	1	2	1	4	7	**10**	2	5	9
AMM	0	4	0	0	1	2	1	5	3	6	1	**108**	**21**	**10**	8	5	**13**	2	1	0	2	4	1	3	2	0
JER	2	3	0	3	2	3	1	2	4	3	2	**18**	**112**	**15**	8	7	9	0	0	0	4	1	0	3	1	0
SAL	0	0	1	0	1	3	3	0	1	2	1	6	**12**	**106**	4	6	**10**	1	2	2	4	5	3	3	3	2
ALE	0	1	0	1	1	2	0	0	0	7	0	6	7	3	**122**	9	**16**	2	0	2	3	0	2	1	0	2
BEI	1	1	0	0	0	2	0	1	0	2	1	5	7	4	6	**113**	**15**	2	1	0	1	2	1	1	0	0
DAM	1	1	0	0	1	0	1	0	2	0	3	9	5	6	**25**	**18**	**100**	1	1	1	3	5	3	0	2	2
BAG	0	1	1	0	2	1	1	1	0	1	0	1	0	2	3	1	7	**123**	**26**	1	3	1	5	3	5	4
BAS	2	1	0	0	0	0	3	0	1	0	2	0	3	3	2	2	2	**31**	**128**	8	3	0	3	3	2	1
MOS	1	0	1	0	2	0	1	1	0	1	2	1	0	3	3	1	0	7	**12**	**165**	4	2	1	6	3	0
DOH	0	3	2	1	1	4	2	0	3	1	4	6	3	4	0	2	2	2	3	0	**119**	9	**12**	5	5	1
JED	2	7	0	1	0	2	3	4	5	4	3	5	3	4	5	1	4	1	2	1	**13**	**115**	4	**21**	6	3
MUS	1	3	3	0	1	1	1	0	1	0	6	4	2	5	2	3	0	0	2	3	9	0	**94**	**13**	2	**23**
RIY	2	**10**	2	0	2	2	0	1	3	1	1	5	2	6	0	3	1	7	3	3	7	**13**	**12**	**102**	7	5
SAN	0	4	3	1	0	4	0	1	1	3	1	2	1	4	1	1	2	5	4	3	3	8	5	**10**	**130**	2
MSA	4	1	3	0	0	2	0	4	1	0	8	2	1	2	2	0	0	2	1	1	0	2	**12**	3	0	**137**

Table 4: Confusion matrix for our best system (MNB with character 4-grams) with dialects grouped with respect to regions, with true positives in blue, and in blod dialect pairs with more than ten false positives.

It appears that the results obtained on the Test Set were worse than those obtained on the Dev Set (third column of Table 3), with an average loss of 1.6 percentage points. Merging the Train and the Dev Set resulted in a gain that in most cases was marginal (+0.26 pp). With $N_{max} > 4$ we did not find much improvement in results, except on the dev set but this can be a bias. This threshold may be related to the fact that character N-grams with $N > 4$ tend to represent the lexicon more than general properties of the dialect itself.

5.2 Error Analysis

Table 4 shows the confusion matrix of our best configuration. The 25 dialects are grouped by regions and MSA appears as the last class. We can see that MUS and SAN are the closest dialects to MSA with respectively 35 and 17 errors involving the MUS-MSA and the SAN-MSA pairs. CAI, MUS and DAM dialects were the most difficult to detect with respectively 112, 106 and 100 False Negatives (FN). Regarding False Positives (FP), the most problematic cases were ASW (106) , RIY (105) and JED (103). Interestingly, the most difficult dialect pairs to discriminate were from Maghreb: FES–RAB (36 and 34 FP) and SFX–TUN (47 and 22). Most of FPs occured between dialects of the same regions with two exceptions : (I) a minor one because North Levant dialects are hard to distinguish from South Levant dialects and (II) a more strange situation with BEN-RIY and KHA-MUS being rather difficult pairs to distinguish despite their apparent distance.

6 Conclusion and Future Work

In this paper, we explored the problem of Arabic dialect classification into 26 classes (covering 25 cities from the Arab World in addition to Modern Standard Arabic(MSA)). We presented MICHAEL a simple, pre-processing free, system design for this DID task. MICHAEL uses character N-Grams features to train a Multinomial Naive Bayes classifier. Beside its simplicity, MICHAEL does not need a huge amount of training data to achieve good results. This system achieved an official score (accuracy) of 53.25% with $1 \leq N \leq 3$ but showed a much better result with only character 4-grams (62.17% accuracy). Using N-grams with $N > 4$ did not seem to improve the results. However, an accurate feature selection technique, like mutual information, may help to get advantage of these longer n-grams that capture more lexical information than shorter N-grams.

Using other types of character features like closed motifs (Buscaldi et al., 2018) would be a first way to assess the influence of the classifier and the features. We plan to explore if adding pre-processing steps like tokenization into words or normalization may improve the results. Another interesting perspective would be to test a Bilstm RNN architecture since this has proven to be adapted to sequential data and Bilstm can exploit both character-level and word-level features. In another perspective it would be very interesting to perform a deeper analysis of classification errors.

References

Nora Al-Twairesh, Hend Al-Khalifa, and Abdulmalik AlSalman. 2016. AraSenTi: Large-scale twitter-specific Arabic sentiment lexicons. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 697–705, Berlin, Germany. Association for Computational Linguistics.

Muhammad Amara. 2010. Reem bassiouney: Arabic sociolinguistics. *Language Policy*, 9:379–381.

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 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.

Davide Buscaldi, Joseph Le Roux, and Gaël Lejeune. 2018. Character-level models for polarity detection in tweets. In *Atelier DEFT 2018*, Rennes, France.

Wingyan Chung. 2008. Web searching in a multilingual world. *Commun. ACM*, 51(5):32–40.

Mohamed Elaraby and Muhammad Abdul-Mageed. 2018. Deep models for arabic dialect identification on benchmarked data. In *VarDial@COLING 2018*.

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, Sofia, Bulgaria. Association for Computational Linguistics.

Niloofar Haeri. 2003. *Sacred Language, Ordinary People: Dilemmas of Culture and Politics in Egypt*. Sacred Language, Ordinary People: Dilemmas of Culture and Politics in Egypt. Palgrave Macmillan.

Terence Frederic Mitchell. 1993. *Pronouncing Arabic*. vol. 2. Clarendon Press.

Maryam Najafian, Sameer Khurana, Suwon Shon, Ahmed Ali, and James Glass. 2018. Exploiting convolutional neural networks for phonotactic based dialect identification. In *ICASSP*, pages 5174–5178. IEEE.

Rabee Naser and Abualsoud Hanani. 2018. Birzeit arabic dialect identification system for the 2018 vardial challenge. In *VarDial@COLING 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.

Mohand Tilmatine. 1999. Substrat et convergences: le berbre et l'arabe nord-africain. *Estudios de dialectolog'ia norteafricana y andalus'i, EDNA*, pages 99–120.

Cornelis Henricus Maria Versteegh. 2001. *The Arabic Language*. Edinburgh University Press Series. Edinburgh University Press.

Omar Zaidan and Chris Callison-Burch. 2011. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1220–1229, Portland, Oregon, USA. Association for Computational Linguistics.

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