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

dhaou.ghoul@sorbonne-universite.fr / dhaou.ghoul@gmail.com
http://cereli.fr/membres/dhaou-ghoul/

26 mars 2020

Paln

- Cadre du travail.
- Travaux antérieurs.
- Les variétés dialectales de l'arabe.
- Identification du dialecte arabe.
- Résultats et analyse des erreurs.
- Conclusions et perspectives.

Cadre du travail

- WANLP 2019 : The Fourth Workshop for Arabic Natural Language Processing : MADAR (Multi-Arabic Dialect Applications and Resources)SHARED TASk.
- Task 1 : MADAR Travel Domain Dialect Identification.
- Conférence ACL, Août 2019, Florence, Italie

Résultats de MADAR Challenge Task 1

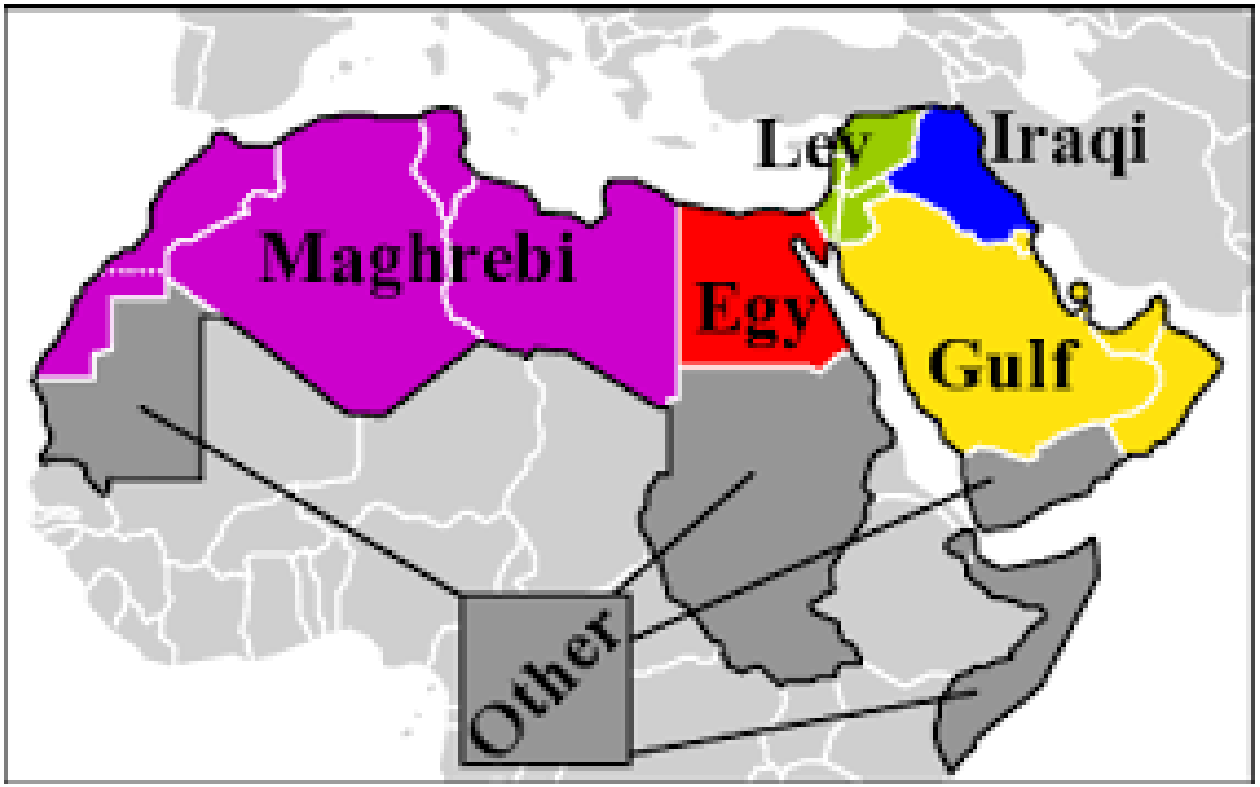
Team	F1	Precision	Recall	Acc_{city}	$\text{Acc}_{country}$	Acc_{region}
ArbDialectID	**67.32 (1)**	67.60 (2)	67.29 (2)	67.29 (2)	75.23 (2)	84.42 (5)
SMarT	67.31 (2)	**67.73 (1)**	**67.33 (1)**	**67.33 (1)**	**75.69 (1)**	**85.13 (1)**
Mawdoo3 LTD	67.20 (3)	67.53 (3)	67.08 (3)	67.08 (3)	75.19 (3)	84.75 (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

Travaux antérieurs

Auteurs	Classificateur	Nb classes	précision	Année
Salameh et al.	MNB	26	68,4%	2018
Elaraby et al.	BILSTM	2/3	87,65%/87.4%	2018
Najafian et al.	CNN/SVM	5	64.5%/62.12%	2018
Elfardy et Diab	Naive Bayes	2	85.5%	2013
Zaidan et al.	MNB	4	65%	2013

Travaux avec Pré-traitement

Carte de dialectes arabes



Exemples

Région	Dialecte	Exemple	Translittération	Traduction
MSA	MSA	كيف أستطيع مساعدتك ؟	kyf ?stTyE msAEdtk	Comment puis-je vous aider ?
Gulf	Doha	شلون اقدر اساعدك ؟	$lwn Aqdr AsAEdk	
	Riyadh	كيف اقدر اساعدك ؟	kyf Aqdr AsAEdk	
	Jeddah	كيف أقدر اساعدك ؟	kyf ?qdr AsAEdk	
	Muscat	كيف اقدر اساعدك ؟	kyf ?qdr AsAEdk	
Iraqi	Mosul	اشون اطيق اساعدك ؟	A$wn ATyq AsAEdk	
	Basra	شلون اكدر اساعدك ؟	$lwn Akdr AsAEdk	
	Baghdad	شلون اكدر اساعدك ؟	$lwn Akdr AsAEdk	
Égypte	Cairo	ازاي أقدر اساعدك ؟	AzAy ?qdr AsAEdk	
	Alexandria	إزاى أقدر أساعدك ؟	AzAy ?qdr AsAEdk	
	Aswan	إزاي أقدر اساعدك ؟	AzAy ?qdr AsAEdk	
Levantine	Aleppo	شلون بقدر أساعدك ؟	$lwn bAkdr AsAEdk	
	Damascus	كيف فيني ساعدك ؟	kyf fyny sAEdk	
	Beirut	كيف فيني ساعدك ؟	kyf fyny sAEdk	
	Amman	كيف بقدر أساعدك ؟	Kyf bAkdr AsAEdk	
	Salt	كيف بقدر اساعدك ؟	Kyf bAkdr AsAEdk	
	Jerusalem	كيف بقدر أساعدك ؟	Kyf bAkdr AsAEdk	
Maghreb	Tunis	كيفاش انجم نعاونك ؟	kyfA$ Angm nEAwnk	
	Sfax	كيفاش النجم نعاونك ؟	kyfA$ Alngm nEAwnk	
	Tripoli	كيف نقدر نساعدك ؟	kyf nqdr nsAEdk	
	Benghazi	كيف نقدر نساعدك ؟	kyf nqdr nsAEdk	
	Alger	كيفاش نقدر نعاونك ؟	kyfA$ nqdr nEAwnk	
	Rabat	كيفاش نقدر نعاونك ؟	kyfA$ nqdr nEAwnk	
	Fes	كيفاش نقدر نساعدك ؟	kyfA$ nqdr nsAEdk	
Other	Khartoum	كيف ممكن اساعدك ؟	kyf mmkn AsAEdk	
	Sana'a	كيف اقدر اساعدك ؟	kyf Aqdr AsAEdk	

Exemples de différences entre MSA et l'arabe dialectale concernant la phonétique, la morphologie, la syntaxe et le lexique

	Phonétique	Morphologie	Syntaxe	Lexique
MSA	qaf	s or swf	mA	sayyaara
ALG	qaf and /g/	ghadi or rH	mA	tomobile
EGY	hamza	h	muw	3arabiyya
GULF	/g/	ba	lA	sayyaara
LEV	hamza	H or rH	muw	sayyaara
MOR	qaf	ghadi	mA	tomobile
TUN	qaf and /g/	bAsh	mA	krhba

Difficultés pour l'identification de l'arabe dialectale

- Lexique partagé : les dialectes ont un vocabulaire commun et une phrase dialectale peut contenir plusieurs dialectes ainsi que MSA.

  لو سمحت خذ وقتك : AMM, ALX, ASW, RIY, SAN...

  S'il te plaît, prends ton temps

- Ambiguïté grammaticale : certains mots identiques sont utilisés avec des fonctions différentes. Par exemple, le mot «Tyb» peut être un adjectif dans certains dialectes et une interjection dans d'autres.
- Homonymes : principalement en raison de l'omission des voyelles courtes, un mot dialectal peut avoir la même orthographe qu'un mot MSA mais une signification entièrement différente. Cela comprend des mots fortement dialectaux tels que dwl : il s'agit soit du mot égyptien (EGY) pour «ces» (hibou prononcé), soit du mot MSA pour «pays» (prononcé duwal) (Zaidan et Callison-Burch, 2014).

Data : The MADAR Corpus

Le corpus MADAR a été créé en traduisant des phrases du corpus "Basic Travel Expression Corpus" (BTEC) de l'anglais et du français vers les différents dialectes (26 classes).

Données	Train	Dev	Test
# phrases	41,600/1600 par classe	5,200/200 par classe	5,200
# mots	336,342	42,586	36,811
# caractères	1,301,599	166,898	162,185

Méthode de classification : Multinomial Naive Bayes Classifier, caractère N-grammes

- Modèle de langage n-grammes basé sur des caractères : Ensembles d'entraînements réduits, divers domaines entre le train et le test.
- Tester plusieurs Classifiers inclus dans la bibliothèque Scikit-Learn (MNB,SVM,OneVsRestClassifier,DecisionTreeClassifier,MLPClassifier).
- Le classifier retenu est MNB.
- Le Classifier SVC (OneVsRestClassifier)n'a pas pu atteindre un résultat.

Résultats pour le classificateur multinomial Naive Bayes

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

Matrice de confusion(MNB avec caractère 4 grammes)

	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

Analyse des erreurs

- MUS et KHA sont les dialectes les plus proches de MSA avec respectivement 35 et 17 erreurs impliquant les paires MUS-MSA et KHA-MSA.
- Les dialectes CAI, DAM,AMM,SAL,MUS et RIY ont été les plus difficiles à détecter avec respectivement 111, 98,94,94,94 et 93 faux négatifs (FN).
- En ce qui concerne les faux positifs (PF), les cas les plus problématiques étaient ASW (105), RIY (98) et JED (104).
- Les paires de dialectes les plus difficiles à distinguer provenaient du Maghreb : FES – RAB (36 et 34 FP) et SFX– TUN (47 et 22) et Egyptian : ALX-ASW (27 et 14) et ASW-CAI (36 et 22) .
- La plupart des PF se sont produits entre des dialectes des mêmes régions à deux exceptions près : (I) mineur parce que les dialectes du Levant Nord sont difficiles à distinguer des dialectes du Levant Sud et (II) une situation plus étrange avec BEN-RIY et KHA-MUS étant plutôt paires difficiles à distinguer malgré leur distance apparente.

Conclusions

- Tâche de classification des dialectes arabes en 26 classes (couvrant 25 villes du monde arabe en plus de l'arabe standard moderne (MSA)).
- MICHAEL une conception de système simple et sans pré-traitement pour cette tâche.
- MICHAEL utilise un modèle de langage au niveau des caractères N-grammes pour entraîner un classifier multinomial Naive Bayes.
- MICHAEL n'a pas besoin d'une énorme quantité de données d'entraînement pour obtenir de bons résultats.
- Ce système a obtenu un score officiel (Précision) de 53,25% avec 1 N 3 mais a montré un bien meilleur résultat avec seulement 4 grammes de caractère (62,17% de précision).

Perspectives

- Vérifier si l'ajout d'étapes de prétraitement comme la tokenisation en mots ou la normalisation peut améliorer les résultats.
- Tester un modèle de réseau de neurone comme CNN, LSTM.
- Analyse plus approfondie des erreurs de classifications.


شكرا لإنتباهكم
MERCI POUR VOTRE
ATTENTION