Aggregating Hierarchical Dialectal Data for Arabic Dialect Identification

Nurpeiis Baimukan, Nizar Habash, Houda Bouamor[†]

Computational Approaches to Modeling Language Lab, New York University Abu Dhabi, UAE

[†]Carnegie Mellon University in Qatar

{nurpeiis, nizar.habash}@nyu.edu,hbouamor@cmu.edu

Introduction

- Arabic is a collection of dialectal variants that are historically related but significantly different.
- These differences can be seen across regions, countries, and even cities in the same countries.
- Previous work on Arabic Dialect identification has focused mainly on specific dialect levels.
- We define a unified hierarchical schema for dialectal Arabic identification.

Arabic Linguistic Challenges

- Extensive orthographic, phonological, morphological and lexical variations among dialects.
- Arabic speakers tend to code-switch between their dialect and Modern Standard Arabic.

Unified Labeling of Arabic Dialect Data Sets

- **Data Selection**
  - Exclusively Arabic script data sets in Arabic dialects, with some degree of identification.
- **Data Variability**
  - Wide range of genres: speech transcripts, social media texts (tweets, news comments, youtube comments), SMS, forum novels, travel phrases, and song lyrics.
  - Labels vary widely in terms of granularity a nd spread.
  - Different sizes.

Corpus ID	Genre/Domain	Region	Country	Province	City	MSA	Split	# Lines (1000s)	# Words (1000s)
MADAR-ST1	travel domain	(mix)	(mix)	(mix)	25	X	original	112	800
MADAR-EX	travel domain	(mix)	(mix)	(mix)	6	X	original	48	285
NADI	twitter	(mix)	(mix)	100	-	-	original	31	408
MADAR-ST2	twitter	(mix)	22	-	-	-	original	188	2,240
HABIBI	song lyrics	(mix)	18	-	-	-	new	412	2,525
QADI	twitter	(mix)	18	-	-	-	original	499	6,260
ARAP-T	twitter	(mix)	16	-	-	-	new	1,607	18,827
LEV-FISHER	speech transcript	(mix)	6	-	-	-	new	61	326
MDPC	web mixed	(mix)	5	-	-	X	new	6	58
PADIC	speech transcript	(mix)	5	-	-	X	new	45	301
GUMAR	forum novel	(1)	6	-	-	-	original	9,097	85,615
LEV-CTS	speech transcript	(1)	4	-	-	-	new	192	968
LEV-TRANS-1	speech transcript	(1)	4	-	-	-	new	359	1,841
LEV-TRANS-2	speech transcript	(1)	4	-	-	-	original	60	499
SHAMI	web mixed	(1)	4	-	-	-	new	66	1,050
GULF-TRANS	speech transcript	(1)	3	-	-	-	original	58	479
BOLT-SMS	sms	(1)	1	-	-	-	original	67	310
CALLHOME	speech transcript	(1)	1	-	-	-	original	29	147
CALLHOME-EX	speech transcript	(1)	1	-	-	-	new	3	14
CURRAS	web mixed	(1)	1	-	-	-	original	5	57
IRAQ-TRANS	speech transcript	(1)	1	-	-	-	original	27	228
LEV-BABYLON	speech transcript	(1)	1	-	-	-	new	76	336
SUAR	web mixed	(1)	1	-	-	-	new	11	121
MMIC-N	news comments	5	-	-	-	X	new	91	2,999
YOUDACC	youtube comments	5	-	-	-	X	original	510	8,317
MMIC-T	twitter	5	-	-	-	-	new	40	578
AMDTC	web mixed	4	-	-	-	-	new	5,183	50,323
AOC	news comments	3	-	-	-	X	new	108	1,976
ADEPT	web mixed	2	-	-	-	-	new	176	1,689

- **Unified Labeling into a three level hierarchy: Region→Country→City**

Region	Country	City
levant	jo	amman
		aqaba
		salt
		zarqa
	lb	beirut
		halba
		sidon
		tripoli
	ps	gaza
		jerusalem
	sy	al_suwayda
		aleppo
		damascus
		homs
		latakia
nile_basin	eg	alexandria
		aswan
		asyut
		beni_suef
		cairo
		damanhur
		el_arish
		el_tor
		faiyum
		girga
		giza
		hurghada
		ismailia
		kafr_el_sheikh
		luxor
		mansoura
		minya
		port_said
		qena
		shibin_el_kom
		suez
		tanta
		zagazig
	sd	khartoum
gulf	ae	abu_dhabi
		dubai
		fujairah
		ras_al_khaimah
		umm_al_quwain
	bh	manama
	iq	amarah
		baghdad
		basra
		duhok
		erbil
		karbala
		kut
		mosul
		najaf
		ramadi
		samawah
		sulaymaniyah
	kw	hawalli
		jahra
	om	khasab
		muscat
		nizwa
		salalah
		sohar
		sur
	qa	al_rayyan
		doha
	sa	abha
		al_madinah
		buraidah
		dammam
		hail
		jeddah
		jizan
		najran
		riyadh
		tabuk
gulf_aden	dj	djibouti
	so	mogadishu
	ye	aden
		al_hudaydah
		dhamar
		ibb
		sanaa
maghreb	dz	algiers
		annaba
		bechar
		bordj_bou_arreridj
		bouira
		jijel
		khenchela
		oran
		ouargla
	ly	bayda
		benghazi
		misrata
		tobruk
		tripoli
	ma	agadir
		fes
		marrakesh
		meknes
		oujda
		rabat
		tangier
	mr	nouakchott
	tn	ariana
		kairouan
		mahdia
		sfax
		sousse
		tunis
msa	msa	msa

Results

- Aggregated Arabic dialect data from different sources into a unified schema
- Built aggregated character and word language models for dialect identification
- Models and mapping code are publically available

  https://github.com/CAMeL-Lab/HierarchicalArabicDialectID
- Improved results on the city, country, and region levels by extending a SOTA technique for Arabic dialect id (Salameh et al., 2018).

	City		Country		Region	
Classifier Setup	Accuracy	F_1	Accuracy	F_1	Accuracy	F_1
(Salameh et al., 2018)	67.75	67.89	76.44	-	85.96	-
Baseline	67.69	67.83	76.33	74.10	85.75	82.60
+City	67.69	67.87	76.50	74.09	85.94	82.68
+Country	67.90	68.10	**77.12**	**74.83**	86.46	83.52
+Region	**68.13**	**68.27**	76.92	74.65	**87.06**	**83.80**