Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4586–4596
Marseille, 20-25 June 2022

Hierarchical Aggregation of Dialectal Data for Arabic Dialect Identification

Nurpeiis Baimukan, Houda Bouamor†, Nizar Habash
Computational Approaches to Modeling Language (CAMeL) Lab
New York University Abu Dhabi
†Carnegie Mellon University in Qatar
{nurpeiis,nizar.habash}@nyu.edu
hbouamor@cmu.edu

Abstract

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 (region, country, province, or city) using level-specific resources; and different efforts used different schemas and labels. In this paper, we present the first effort aiming at defining a standard unified three-level hierarchical schema (region-country-city) for dialectal Arabic classification. We map 29 different data sets to this unified schema, and use the common mapping to facilitate aggregating these data sets. We test the value of such aggregation by building language models and using them in dialect identification. We make our label mapping code and aggregated language models publicly available.

Keywords: Arabic Dialects, Dialect Identification, Language Models

1. Introduction

Dialect identification (DID) is a natural language processing (NLP) task that aims at automatically determining the dialect of a given speech fragment or text (Etman and Beex, 2015). Since dialectal difference tend to be more subtle in relative terms to language differences, the DID task is harder than language identification. In this paper we focus on Arabic dialect identification, but we believe that our techniques and insights are extensible to other languages and dialect groups.

Arabic is a collection of dialectal variants that are historically related but significantly different. Arabic dialects are often classified in terms of geography at different levels of granularity: at the regional, country, province and city levels (Zaidan and Callison-Burch, 2012; Bouamor et al., 2019; Abdul-Mageed et al., 2020; Abdul-Mageed et al., 2021). Dialects are different from each other and Modern Standard Arabic (MSA) in terms of phonology, orthography, morphology, and the lexicon. Arabic speakers tend to code-switch between their dialect and MSA, creating sentences with different levels/percentages of "dialectness" (Habash et al., 2008; Elfardy et al., 2013; El-Haj et al., 2018; Ali et al., 2021). That said, many of the differences are not observed in the written forms since in Arabic orthography, writing vowel diacritics is optional.

Several efforts have targeted creating different resources at different hierarchical levels going from cities to regions (Zaidan and Callison-Burch, 2011; Smaïli et al., 2014; Jarrar et al., 2016; Al-Twairesh et al., 2018; Bouamor et al., 2018; Abdul-Mageed et al., 2020). Most of these resources were built in independent efforts using different labeling schemas, making the joint use and comparison of these data sets impractical.

In this paper, we address this issue by defining a unified hierarchical schema for dialectal Arabic identification; and demonstrate its use by mapping a number of data sets to it, and building aggregated language models at different hierarchical levels.

The main contributions of our work are as follows:

- We define a unified 3-level hierarchical schema for labeling dialectal text from different sources: region-country-city.
- We map the various labels from 29 different data sets into our unified schema. We make our mapping code, which is tailored to the different data sets, available.[1]
- We create aggregated n-gram language models (LM) at the region, country, and city levels in character and word spaces from all of the data sets we worked with. We make the LMs publicly available.[1]
- We demonstrate the value of our aggregated dialectal LMs on a standard DID test set for city level identification extending on a well-established state-of-the-art approach for Arabic DID.

The paper is organized as follows. We present some basic facts about the challenges of processing Arabic dialects in Section 2. In Section 3, we provide an overview of related work. Section 4 details our approach and its implementation from selecting the data sets, to building the aggregated LMs. We evaluate the use of the aggregated LMs in Section 5. We conclude the work in Section 6.

[1] https://github.com/CAMeL-Lab/HierarchicalArabicDialectID

2. Arabic Linguistic Challenges

Arabic is a collection of dialectal variants that are historically related but significantly different. Arabic dialects are often classified in terms of geography at different levels of granularity. Typical regional groupings cluster the dialects into Levantine Arabic (Lebanon, Syria, Jordan and Palestine), Gulf Arabic (Qatar, Kuwait, Saudi Arabia, United Arab Emirates and Bahrain, with Iraqi and Omani Arabic included sometimes), Egyptian Arabic (which may include Sudan), North African Arabic (vaguely covering Morocco, Algeria, Tunisia, Libya and Mauritania), and Yemeni Arabic (Habash, 2010). However, within each of these regional groups, there is significant variation down to the country, province, and city levels. Although we acknowledge that there are various dimensions of classifying them (i.e., social class and religious), the automatic regional dialect identification, or less granular classification, has shown to achieve strong results (Zaidan and Callison-Burch, 2012; Althobaiti, 2020). But city level identification has been shown to be very challenging (Bouamor et al., 2019; Abdul-Mageed et al., 2020; Abdul-Mageed et al., 2021).

The differences among the dialects, and between the dialects and MSA extend over the phonological, morphological and lexical dimensions. For example, the Levantine word عالبندورة *ʕalbanadurah* 'about the tomatoes' shows morphological and lexical differences from the MSA phrase عن الطماطم *ʕan AlTamATim*. The word for 'tomato' varies widely across Arabic dialects, e.g., مطيشة *maTiyšah* in Moroccan Arabic, and قوطة *quTah* /ʔūtˤa/ in Egyptian Arabic.

Such differences suggest that the task of dialect identification should be easy. But in fact, distinguishing among different Arabic varieties is quite difficult for a number of reasons. Because short vowels are optionally written in Arabic, many dialectal words end up looking similar to MSA cognates or unrelated forms. For example, the written word يكتب *yktb*[2] maybe pronounced /yaktub/ in MSA, /yoktob/ in Levantine Arabic or /yiktib/ in Egyptian Arabic. Additionally, since there are no standard orthographies for the dialects, there are numerous ways to spell the same word, making it hard to train models for this task. Habash et al. (2018) uses an example of a word with over two dozen spellings. One of the common spelling choices dialect writers make to spell consonants is spelling the word as it appears in MSA, rather than how it sounds. For example, the word قلب *qlb* corresponds to /qalb/ in MSA, /galb/ in Gulf Arabic and /ʔalb/ in Levantine Arabic. It should be noted that Arabic dialects are sometimes also written in an ad hoc romanization called Arabizi (Darwish, 2014; Bies et al., 2014). We do not model Arabizi text in this work.

Furthermore, Arabic speakers tend to code-switch between their dialect and MSA, creating sentences with different levels/percentages of "dialectness" (Habash et al., 2008; Elfardy et al., 2013; El-Haj et al., 2018; Ali et al., 2021). As such, a dialectal sentence might consist entirely of words that are used commonly across all Arabic varieties, including MSA. Some words are used across the varieties with different functions and different meanings. As such, it is important to consider the context in which these words appear for DID.

[2]Arabic transliteration is in the HSB scheme (Habash et al., 2007).

3. Related Work

Recently, there has been an active interest in developing automatic Arabic dialect processing systems working at a different levels of representation and in exploring different dialectal data sets (Shoufan and Alameri, 2015; Jauhiainen et al., 2019; Althobaiti, 2020). This has been facilitated by the newly developed monolingual and multilingual dialectal corpora and lexicons. Several mono-dialectal corpora covering different Arabic dialects were built and made available (Gadalla et al., 1997; Diab et al., 2010; Zaidan and Callison-Burch, 2011; Al-Sabbagh and Girju, 2012; Salama et al., 2014; Sadat et al., 2014; Smaïli et al., 2014; Cotterell and Callison-Burch, 2014; Jarrar et al., 2016; Khalifa et al., 2016; Al-Twairesh et al., 2018; Abu Kwaik et al., 2018; El-Haj, 2020).

The expansion into multi-dialectal data sets was initially done at the regional level (Zaidan and Callison-Burch, 2011; McNeil and Faiza, 2011; Elfardy et al., 2014; Bouamor et al., 2014; Salama et al., 2014; Meftouh et al., 2015). Then, several efforts for creating finer grained parallel dialectal corpus and lexicon has been presented. These include labeling country-level data from similar regions (Sawalha et al., 2019; Jarrar et al., 2016; Meftouh et al., 2015; Zaghouani and Charfi, 2018; Al-Twairesh et al., 2018; Abu Kwaik et al., 2018; El-Haj, 2020; Khalifa et al., 2018; Shon et al., 2020; Abdelali et al., 2020; Abdul-Mageed et al., 2018) and introducing larger-scale data sets covering between 5 and 21 countries (Mubarak and Darwish, 2014; Bouamor et al., 2018; Abdul-Mageed et al., 2018; Zaghouani and Charfi, 2018) with a much more varying label sets. Bouamor et al. (2018) introduced MADAR, the first city-level dialectal data set including dialects from 25 cities. Following this effort, Abdul-Mageed et al. (2020) presented NADI, a large scale data set of Arabic varieties annotated with provinces in addition to cities, countries and regions, covering up 21 countries and 100 provinces.

However, most of these efforts focus primarily on a number of varieties corresponding generally to those spoken in major cities (Cairo, Amman, Baghdad, Tunis, Rabat, etc.), or study different dialects independently. All these data sets have different dialect labeling schema at different hierarchical levels, making their use in any research work impractical.

To the best of our knowledge, our work is the first aiming at aggregating several Arabic dialectal data sets

Corpus ID	Corpus Name	Reference
ADEPT	LDC2012T09: Arabic-Dialect/English Parallel Text	(BBN Technologies et al., 2012)
AMDTC	Arabic Multi Dialect Text Corpus	(Almeman and Lee, 2013)
AOC	Arabic Online Commentary Dataset	(Zaidan and Callison-Burch, 2011)
ARAP-T	Arap-Tweet: The Arabic Author Profiling Project Twitter Corpus	(Zaghouani and Charfi, 2018)
BOLT-SMS	LDC2017T07: BOLT Egyptian SMS	(Chen et al., 2017)
CALLHOME	LDC97T19: CALLHOME Transcripts	(Gadalla et al., 1997)
CALLHOME-EX	LDC2002T38: CALLHOME Supplement Transcripts	(Linguistic Data Consortium, 2002)
CURRAS	Curras: A corpus for the Palestinian Arabic dialect	(Jarrar et al., 2016)
GULF-TRANS	LDC2006T15: Gulf Arabic Transcripts	(Appen Pty Ltd, 2006a)
GUMAR	Gumar: A Gulf Arabic Internet Novel Corpus	(Khalifa et al., 2018)
HABIBI	Habibi: A multi Dialect multi National Arabic Song Lyrics Corpus	(El-Haj, 2020)
IRAQ-TRANS	LDC2006T16: Iraqi Arabic Transcripts	(Appen Pty Ltd, 2006b)
LEV-BABYLON	LDC2005S08: Babylon Levantine Arabic Transcripts	(BBN Technologies, 2005)
LEV-CTS	LDC2005S14: CTS Levantine Arabic Transcripts	(Maamouri et al., 2005)
LEV-FISHER	LDC2007T04: Fisher Levantine Arabic Transcripts	(Maamouri et al., 2007)
LEV-TRANS-1	LDC2006T07: Levantine Arabic Transcripts	(Maamouri et al., 2006)
LEV-TRANS-2	LDC2007T01: Levantine Arabic Transcripts	(Appen Pty Ltd, 2007)
MADAR-EX	MADAR Corpus 6 Extra	(Bouamor et al., 2019)
MADAR-ST1	MADAR Shared Task 1	(Bouamor et al., 2019)
MADAR-ST2	MADAR Shared Task 2	(Bouamor et al., 2019)
MDPC	Multi-dialect parallel corpus	(Bouamor et al., 2014)
MMIC-N	Multi-dialect, multi-genre informal corpus news	(Cotterell and Callison-Burch, 2014)
MMIC-T	Multi-dialect, multi-genre informal corpus twitter	(Cotterell and Callison-Burch, 2014)
NADI	NADI: Nuanced Arabic Dialect Identification Shared Task	(Abdul-Mageed et al., 2020)
PADIC	PADIC: Parallel Arabic Dialect Corpus	(Meftouh et al., 2015)
QADI	QCRI Arabic Dialects Identification Corpus	(Abdelali et al., 2020)
SHAMI	Shami: A Corpus of Levantine Arabic Dialects	(Abu Kwaik et al., 2018)
SUAR	SUAR: Saudi Corpus for NLP Applications and Resources	(Al-Twairesh et al., 2018)
YOUDACC	Youtube Dialectal Arabic Commentary Corpus	(Salama et al., 2014)

Table 1: The list of dialectal data sets used in this work.

from different sources and different levels: region, country, province and city levels, defining a unified hierarchical schema for labeling dialectal text from different sources, and building the largest-scale dialectal Arabic resource, mapped to their MSA, English, and French versions, when available.

In terms of dialect identification, a number of Arabic dialect identification shared tasks were organized first as part of the VarDial workshop. These focused on regional varieties such as Egyptian, Gulf, Levantine, and North African based on speech broadcast transcriptions and integrated acoustic and phonetic features extracted from raw audio (Malmasi et al., 2016; Zampieri et al., 2017; Zampieri et al., 2018). Then, shared tasks dedicated to Arabic dialect identification specifically were created: The MADAR shared task(Bouamor et al., 2019) and the NADI shared task in its two editions (Abdul-Mageed et al., 2020; Abdul-Mageed et al., 2021).

A variety of methods have been introduced to classify the dialectal texts in MADAR and NADI. Most of the work have shown that shallow n-gram based approaches are the state-of-the art in terms of performance for this task (Salameh et al., 2018; Bouamor et al., 2019; Abdul-Mageed et al., 2020; Abdul-Mageed et al., 2021), while using deep learning architectures such as RNN, CNN, or BERT do not achieve a comparable accuracy. Recently, Inoue et al. (2021) compared fine-tuning 12 different BERT models for Arabic and none of them improved over Salameh et al. (2018)'s model. Since fine-tuning tends to be more robust to limited training/tuning, we hypothesize that the BERT-like masked LMs are trained towards modeling deeper semantic similarity and less so towards modeling shallow signals of phonology and spelling differences that can help the task of dialect identification (i.e. in comparison to n-gram features and models).

In this work, we improve on the DID results reported in (Salameh et al., 2018) by adopting the same approach and adding a few additional features.

4. Unified Labeling of Arabic Dialect Data Sets

We discuss next the various data sets we worked with, and the process to unify their dialectal id labels.

4.1. Data Selection

There are numerous data sets for Arabic NLP. In this effort, we selected 29 data sets that fit the following criteria. First, the data sets is primarily or exclusively written in Arabic script. We do not consider Arabizi data sets in this effort. Second, The data sets are primarily or exclusively in Arabic dialects. We do not

Corpus ID	Genre/Domain	Region	Country	Province	City	MSA	Split	# Lines (1000s)	# Words (1000s)	# Chars (1000s)
MADAR-ST1	travel domain	(mix)	(mix)	(mix)	25	X	original	112	800	3,551
MADAR-EX	travel domain	(mix)	(mix)	(mix)	6	X	original	48	285	1,473
NADI	twitter	(mix)	(mix)	100	-	-	original	31	408	2,553
MADAR-ST2	twitter	(mix)	22	-	-	-	original	188	2,240	12,235
HABIBI	song lyrics	(mix)	18	-	-	-	new	412	2,525	12,637
QADI	twitter	(mix)	18	-	-	-	original	499	6,260	32,628
ARAP-T	twitter	(mix)	16	-	-	-	new	1,607	18,827	119,548
LEV-FISHER	speech transcript	(mix)	6	-	-	-	new	61	326	1,307
MDPC	web mixed	(mix)	5	-	-	X	new	6	58	275
PADIC	speech transcript	(mix)	5	-	-	X	new	45	301	1,424
GUMAR	forum novel	(1)	6	-	-	-	original	9,097	85,615	452,570
LEV-CTS	speech transcript	(1)	4	-	-	-	new	192	968	3,648
LEV-TRANS-1	speech transcript	(1)	4	-	-	-	new	359	1,841	7,052
LEV-TRANS-2	speech transcript	(1)	4	-	-	-	original	60	499	2,035
SHAMI	web mixed	(1)	4	-	-	-	new	66	1,050	4,706
GULF-TRANS	speech transcript	(1)	3	-	-	-	original	58	479	1,926
BOLT-SMS	sms	(1)	1	-	-	-	original	67	310	1,421
CALLHOME	speech transcript	(1)	1	-	-	-	original	29	147	669
CALLHOME-EX	speech transcript	(1)	1	-	-	-	new	3	14	63
CURRAS	web mixed	(1)	1	-	-	-	original	5	57	267
IRAQ-TRANS	speech transcript	(1)	1	-	-	-	original	27	228	927
LEV-BABYLON	speech transcript	(1)	1	-	-	-	new	76	336	1,725
SUAR	web mixed	(1)	1	-	-	-	new	11	121	565
MMIC-N	news comments	5	-	-	-	X	new	91	2,999	11,307
YOUDACC	youtube comments	5	-	-	-	X	original	510	8,317	44,468
MMIC-T	twitter	5	-	-	-	-	new	40	578	3,114
AMDTC	web mixed	4	-	-	-	-	new	5,183	50,323	273,545
AOC	news comments	3	-	-	-	X	new	108	1,976	10,221
ADEPT	web mixed	2	-	-	-	-	new	176	1,689	7,755

Table 2: Corpora domain, region, country, province and city level details and statistics in terms of lines, words, and characters. We indicate whether original publish train-dev-test splits are used or new ones defined in this work.

consider primarily or exclusively Standard Arabic data sets. Third, the dialects in the data sets are identified to some degree. We do not expect a specific or common level of granularity in the dialect label. Table 1 presents the list of data sets we used. All of the data sets are publicly available, some are freely downloadable, and others require membership licenses.

4.2. Data Variability

Table compares the data sets we worked with. As can be seen, the data sets cover a wide range of genres: speech transcripts, social media texts (tweets, news comments, youtube comments), SMS, forum novels, travel phrases, and song lyrics.

The data set labels vary widely in terms of granularity and spread. In terms of granularity, we found four levels: city, province, country and region. Some data sets include multiple regions and countries, but are only labeled at the city or province levels. Some only specify the region level. It terms of spread, some are specific to a single country (e.g. CALLHOME), others are regional (e.g., GUMAR and SHAMI) or pan-Arab (e.g., MADAR-ST2 and NADI). Some data sets included and marked MSA texts explicitly.

Furthermore, the data sets vary widely in how the dialect label is identified in terms of textual units. The lowest level of identification is at the sentence/line level. The speech transcript data sets were created by targeting a specific speech community. The MDPC, MADAR-ST1, and MADAR-EX data sets were commissioned translations, so they were created in the target dialect at the sentence level. The MADAR-ST2 data set was labeled by identifying the country of the Tweeter and assigning the label to all their tweets. The GUMAR data set was labeled at the document level. The NADI data set was annotated automatically using Tweet location as proxy for dialect.

Finally, these data sets also vary in terms of size, as well as whether standard (i.e. non-random, and replicable) experimental train-dev-test splits already exist for them.

4.3. Data Preprocessing and Splitting

We minimally processed the data sets using simple punctuation based sentence segmentation and white space tokenization using (Obeid et al., 2020).

If a data set does not have a standard train-dev-test split, we manually divided it as follows: 80% for training, 10%, for development, and 10% for testing. Table 3 shows the aggregated counts for the splits. In this pa-

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

Region	Country	City
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

Region	Country	City
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

Figure 1: Hierarchical classification of the Arabic dialect labels in our data set. Country labels follow the ISO 3166 standard.

Level	Split	# of lines (1000s)
City	Train	591
	Dev	53
	Test	55
Country	Train	11,154
	Dev	1,474
	Test	2,023
Region	Train	14,805
	Dev	1,903
	Test	2,454

Table 3: Aggregated Data Splits

per we restrict ourselves to building models using the training portions, to maximize usability by other researchers. All of the code for preprocessing the data sets and identifying the splits is publicly available.[1]

4.4. Unified Hierarchical Labeling

Our next step is to unify the labeling for the various data sets discussed above. Not only did the data sets come at different levels of annotation, but they used different naming conventions for the labels. Since these data sets were created in different research efforts for different original purposes, they did not have a common format or representation. As such, our challenge was to process the data from each data sets into a common unified representation.

In order to create our labeling schema, we first went through all the collected data sets and created a list of all unique terms they used for *region*, *country*, *province*, and *city* labels. We then created a mapping from each label to a unified set of corresponding labels for *region*, *country* and *city*. We intentionally ignored the province level because only one data set had

(a) **Initial Representation**

sentence	شحال كا يدير الفطور؟
city dialect	rabat

(b) **Extended Representation**

sentence	شحال كا يدير الفطور؟
dialect_city_id	rabat
dialect_country_id	ma
dialect_region_id	maghreb
dataset_name	madar_shared_task1
data_annotation_method	manual_translation
data_source	travel_domain
lexicon_corpus	corpus
split_original_manual	original

Figure 2: Example of hierarchical labeling of a sentence extracted from MADAR and originally labeled at the city-level (Rabat).

it (NADI). We collapsed the province and the city level under the label of city where we choose the capital of the province as the city label. This was only done for the NADI data set. For example, the province label *rabat_sale_kenitra* is mapped to the city label *rabat*.

In total, our label space comprises 113 city labels, 22 country labels and 6 region labels, which we organize hierarchically. At each level we include MSA as an additional dialect label, i.e., the city, country and region are all *msa*. The full list of labels for each level in our hierarchy is provided in Figure 1. Obviously this is not a complete list, but it covers the data points in our data sets.

Finally, we assign the missing hierarchical labels to all the data set points when possible. Basically, a specific fine granularity label identifies all of the higher level labels, e.g., the city label 'Abu Dhabi' identifies its country ('The United Arab Emirates') and its Region ('Arabian Gulf'). However, for the country label 'Syria', we cannot identify the specific cities, but can identify the region as 'Levant'. As such, some data set points will be under-specified for lower levels of the hierarchy.

For all the data points from all the sets, we keep a record of their original corpus, data source, splits, etc. An example of our hierarchical labeling of a sentence extracted from the MADAR-ST1 data set and originally labeled with the Rabat city label is shown in Figure 2.

4.5. Aggregated Language Models

Using the hierarchical labels, we create aggregated data sets for each city, country and region, by combining all the data points from the 29 data sets with shared city, country or region labels, accordingly. These aggregations were done separately for training, development, and test subsets.

From each aggregated training set, we created two n-gram language models at the character and word levels using KenLM (Heafield, 2011) with an order of 5 and discount fallback. We make all the models publicly available.[1] We demonstrate the use of these models in the next section.

5. Evaluation

We assess the value of our aggregated language models on a difficult public shared task.

5.1. Experimental Settings

Task and Data We report our results on the MADAR Shared Task 1, which targets labeling 25 city dialects and MSA (26 labels) (Bouamor et al., 2019). The task makes use of data in the travel phrase domain (Bouamor et al., 2018) consisting in commissioned translations from the English and French versions of the Basic Traveling Expression Corpus (BTEC) (Takezawa et al., 2007).

The official shared task allows the use of training data for the 26 labels (Corpus-26), a larger data set (Corpus-6) which has more examples labeled for five cities and MSA, in addition to unlabeled external data. We use the same labeled data for our baseline and all of our training. We make use of our training split aggregated LMs only to provide features to other machine learning classifiers in a pretrained LM manner. As such it should be noted that indeed our use goes beyond the restrictions of the official shared task since technically our aggregated data was labeled, even if not for the same target set of labels of the shared task. That said, we still think there are very useful insights from these experiments.

Metrics We evaluate our model's outputs in terms of accuracy and F_1 score at the city, country and region level. We classify only at the city level, and generate the country and region labels by simple deterministic mapping from the finer grained city labels using our label hierarchy.

5.2. Dialect Identification Approach

Previous Efforts The MADAR Shared Task 1 is quite a hard task. To our knowledge, the best results reported on it is that by (Salameh et al., 2018), which precedes the shared task itself. Salameh et al. (2018) used a Multinomial Naive Bayes classifier with a set of engineered n-gram features. Salameh et al. (2018) reported worse results using a number of models, including neural model without success. They attributed the weaker performance on the limited training data.

Recently, Inoue et al. (2021) compared fine-tuning 12 different BERT models for Arabic and none of them improved over Salameh et al. (2018)'s model.

In the rest of this paper we discuss the Salameh et al. (2018) approach and describe how we extend it using our aggregated LMs.

Baseline Classifier As a baseline, we consider the Camel Tools implementation (Obeid et al., 2020) of

Classifier Setup	City		Country		Region	
	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**
+City +Country	67.13	67.46	76.33	73.75	86.02	82.90
+City +Region	67.48	67.64	76.46	73.97	86.38	83.17
+Country +Region	67.54	67.76	76.77	74.39	86.56	83.40
+City +Country +Region	67.23	67.51	76.50	73.79	86.38	83.27

Table 4: Dialect identification results on MADAR Shared Task 1 data (MADAR 26 Test). The results for (Salameh et al., 2018) are as reported in (Bouamor et al., 2019).

Salameh et al. (2018)'s best model for dialect identification. This is the only available implementation of Salameh et al. (2018). This implementation is slightly below the reported results in the shared task paper: 0.06% at the city-level accuracy and F_1.

The Salameh et al. (2018) *best* model uses two classifiers: **main** and **supporting**. Both classifiers use word unigram and character uni-, bi-, and trigram features with TF-IDF scores in addition to 5-gram LM scores, all trained on the same training MADAR data set. The **supporting** classifier is trained on Corpus-6 and classifies into its corresponding six labels. The **main** classifier is trained on Corpus-26 and classifies into its corresponding 26 labels. Most importantly, the **main** classifier uses the **supporting** classifier probabilities as additional features.

Aggregated Classifiers We build three new classifiers to use as *additional supporting classifiers* to the **main** classifier design mentioned above. The three classifiers use the Corpus-26 training data and n-gram features, but replace the 5-gram LM scores with those from our aggregated models at the city, country and region levels.[3] In all of these classifiers, the number of additional LM features is equal to two times the number of the labels in the hierarchy level. For example, in the city classifier, we use 113 word 5-gram features and 113 character 5-gram features. But, in all cases, we only classify into the 26 labels of the MADAR 26 Shared Task 1; and we only pass the 26 classification probabilities to the **main** classifier.

We experimented with the use of all combinations of the aggregated classifiers. For example, in **Baseline+City+Region**, we use the exact Baseline setup, but add the classifier output probabilities from City and Region.

[3]We experimented with training on the aggregated data, but that produced consistently lower results. We also experimented with classifying into the full hierarchy and passing on the probabilities of these classifications; but that did worse also.

5.3. Results and Discussion

The results of our experiments are presented in Table 4.

The best results at the city level come from using the aggregated region classifier to support the baseline classifier with an increase of 0.44% in accuracy and 0.43% in F_1. However, these results are **not** statistically significant against the baseline using McNemar's test ($p = 0.07$) (McNemar, 1947). The aggregated region classifier setup also has the best region-level performance with an increase of 1.31% in accuracy and 1.20% in F_1. These results are statistically significant ($p < 0.01$). The setup with the aggregated country classifier outperforms on the country-level labels, with an increase of 0.79% in accuracy and 0.73% in F_1. These results are also statistically significant ($p < 0.05$). The setup with the aggregated city classifier did not help the city-level classification; and did not do well, in general. The combinations of aggregated classifiers did worse than the single aggregated classifiers.

The simple interpretation of the results is that the larger aggregated models, for region and country, help more because they have more data in them. The aggregated region set has 32% more lines of training than the aggregated country set; and 25 times the number of lines in the aggregated city training. Furthermore, most of the data in the aggregated city LMs come from the MADAR data, any way. It's unclear why the combinations of aggregated classifiers do not help; perhaps this is due to noisy signals from the different components.

While our use of the aggregated data for DID is similar to what Salameh et al. (2018) did with the Corpus-6 classifier, there is an important difference. The Corpus-6 contained the same in-domain data as Corpus-26 and included labels all of which appear in the basic Corpus-26 set. The regional aggregated data comes from a much wider variety of genres and domains, and it contains a limited number of low granularity labels. But it is much bigger in size; which is the biggest advantage it has.

6. Conclusion and Future Work

In this work, we defined a general hierarchical dialectal labeling schema and mapped 29 different dialectal data sets into it. We created a number of n-gram language models for specific cities, countries and regions and demonstrated the use of such models in city-level dialect identification task. We make our models and code publicly available.

In the future, we would like to use the aggregated language models in other Arabic dialect NLP tasks, such as speech recognition. We would like to use these models as part of systems for downstream applications such as user-aware (dialect-wise) generation, and text normalization. We also look forward to extending the label set to cover more Arab cities.

Acknowledgments

Part of this work was carried out on the High Performance Computing resources at New York University Abu Dhabi. We thank the anonymous reviewers for their insightful suggestions and comments. We also thank Ossama Obeid for the Camel Tools support, and Bashar Alhafni for the helpful conversations.

7. Bibliographical References

Abdul-Mageed, M., Zhang, C., Bouamor, H., and Habash, N. (2020). NADI 2020: The First Nuanced Arabic Dialect Identification Shared Task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop (WANLP 2020)*, Barcelona, Spain.

Abdul-Mageed, M., Zhang, C., Elmadany, A., Bouamor, H., and Habash, N. (2021). NADI 2021: The second nuanced Arabic dialect identification shared task. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 244–259, Kyiv, Ukraine (Virtual), April. Association for Computational Linguistics.

Al-Twairesh, N., Al-Matham, R. N., Madi, N., Almugren, N., Al-Aljmi, A.-H., Alshalan, S., Alshalan, R., Alrumayyan, N., Al-Manea, S., Bawazeer, S., Al-Mutlaq, N., Almanea, N., Huwaymil, W. B., Alqusair, D., Alotaibi, R., Al-Senaydi, S., and Alfutamani, A. (2018). Suar: Towards building a corpus for the saudi dialect. In *ACLING*.

Ali, A., Chowdhury, S., Hussein, A., and Hifny, Y. (2021). Arabic code-switching speech recognition using monolingual data. *arXiv preprint arXiv:2107.01573*.

Althobaiti, M. J. (2020). Automatic Arabic dialect identification systems for written texts: A survey. *arXiv preprint arXiv:2009.12622*.

Bies, A., Song, Z., Maamouri, M., Grimes, S., Lee, H., Wright, J., Strassel, S., Habash, N., Eskander, R., and Rambow, O. (2014). Transliteration of Arabizi into Arabic Orthography: Developing a Parallel Annotated Arabizi-Arabic Script SMS/Chat Corpus. In *Proceedings of the Workshop for Arabic Natural Language Processing (WANLP)*, Doha, Qatar.

Bouamor, H., Habash, N., Salameh, M., Zaghouani, W., Rambow, O., Abdulrahim, D., Obeid, O., Khalifa, S., Eryani, F., Erdmann, A., and Oflazer, K. (2018). The MADAR Arabic Dialect Corpus and Lexicon. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

Bouamor, H., Hassan, S., and Habash, N. (2019). The MADAR shared task on Arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 199–207, Florence, Italy, August. Association for Computational Linguistics.

Darwish, K. (2014). Arabizi Detection and Conversion to Arabic. In *Proceedings of the Workshop for Arabic Natural Language Processing (WANLP)*, pages 217–224, Doha, Qatar.

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

Elfardy, H., Al-Badrashiny, M., and Diab, M. (2013). Code Switch Point Detection in Arabic. In *Proceedings of the Conference on Application of Natural Language to Information Systems (NLDB)*, MediaCity, UK.

Etman, A. and Beex, L. (2015). Language and Dialect Identification: A Survey. In *Proceedings of the Intelligent Systems Conference (IntelliSys)*, London, UK.

Habash, N., Soudi, A., and Buckwalter, T. (2007). On Arabic Transliteration. In A. van den Bosch et al., editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*, pages 15–22. Springer, Netherlands.

Habash, N., Rambow, O., Diab, M., and Kanjawi-Faraj, R. (2008). Guidelines for Annotation of Arabic Dialectness. In *Proceedings of the Workshop on HLT & NLP within the Arabic World*, Marrakech, Morocco.

Habash, N., Eryani, F., Khalifa, S., Rambow, O., Abdulrahim, D., Erdmann, A., Faraj, R., Zaghouani, W., Bouamor, H., Zalmout, N., Hassan, S., shargi, F. A., Alkhereyf, S., Abdulkareem, B., Eskander, R., Salameh, M., and Saddiki, H. (2018). Unified guidelines and resources for Arabic dialect orthography. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

Habash, N. Y. (2010). *Introduction to Arabic natural language processing*. Morgan & Claypool Publishers.

Heafield, K. (2011). KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the Workshop on Statistical Machine Translation (WMT)*, pages 187–197.

Inoue, G., Alhafni, B., Baimukan, N., Bouamor, H., and Habash, N. (2021). The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natu-*

ral Language Processing Workshop, pages 92–104, Kyiv, Ukraine (Virtual), April. Association for Computational Linguistics.

Jarrar, M., Habash, N., Alrimawi, F., Akra, D., and Zalmout, N. (2016). Curras: an annotated corpus for the Palestinian Arabic dialect. *Language Resources and Evaluation*, pages 1–31.

Jauhiainen, T., Lui, M., Zampieri, M., Baldwin, T., and Lindén, K. (2019). Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research*, 65, Aug.

Malmasi, S., Zampieri, M., Ljubešić, N., Nakov, P., Ali, A., and Tiedemann, J. (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)*, pages 1–14, Osaka, Japan.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Obeid, O., Zalmout, N., Khalifa, S., Taji, D., Oudah, M., Alhafni, B., Inoue, G., Eryani, F., Erdmann, A., and Habash, N. (2020). CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France, May. European Language Resources Association.

Salameh, M., Bouamor, H., and Habash, N. (2018). Fine-grained Arabic dialect identification. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 1332–1344, Santa Fe, New Mexico, USA.

Shoufan, A. and Alameri, S. (2015). Natural language processing for dialectical Arabic: A survey. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 36–48, Beijing, China, July. Association for Computational Linguistics.

Smaïli, K., Abbas, M., Meftouh, K., and Harrat, S. (2014). Building resources for Algerian Arabic dialects. In *Proceedings of the Conference of the International Speech Communication Association (Interspeech)*.

Takezawa, T., Kikui, G., Mizushima, M., and Sumita, E. (2007). Multilingual Spoken Language Corpus Development for Communication Research. *Computational Linguistics and Chinese Language Processing*, 12(3):303–324.

Zaidan, O. F. and Callison-Burch, C. (2011). The Arabic Online Commentary Dataset: an Annotated Dataset of Informal Arabic With High Dialectal Content. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, pages 37–41.

Zaidan, O. F. and Callison-Burch, C. (2012). Arabic dialect identification. *Computational Linguistics*, 40(1):171–202.

Zampieri, M., Malmasi, S., Ljubešić, N., Nakov, P., Ali, A., Tiedemann, J., Scherrer, Y., and Aepli, N. (2017). Findings of the VarDial Evaluation Campaign 2017. In *Proceedings of the Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 1–15, Valencia, Spain.

Zampieri, M., Malmasi, S., Nakov, P., Ali, A., Shon, S., Glass, J., Scherrer, Y., Samardžić, T., Ljubešić, N., Tiedemann, J., van der Lee, C., Grondelaers, S., Oostdijk, N., Speelman, D., van den Bosch, A., Kumar, R., Lahiri, B., and Jain, M. (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 2018)*, pages 1–17, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

8. Language Resource References

Abdelali, A., Mubarak, H., Samih, Y., Hassan, S., and Darwish, K. (2020). Arabic dialect identification in the wild.

Abdul-Mageed, M., Alhuzali, H., and Elaraby, M. (2018). You tweet what you speak: A city-level dataset of Arabic dialects. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

Abdul-Mageed, M., Zhang, C., Bouamor, H., and Habash, N. (2020). NADI 2020: The First Nuanced Arabic Dialect Identification Shared Task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop (WANLP 2020)*, Barcelona, Spain.

Abu Kwaik, K., Saad, M., Chatzikyriakidis, S., and Dobnik, S. (2018). Shami: A corpus of Levantine Arabic dialects. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Al-Sabbagh, R. and Girju, R. (2012). YADAC: Yet another Dialectal Arabic Corpus. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 2882–2889, Istanbul, Turkey.

Al-Twairesh, N., Al-Matham, R. N., Madi, N., Almugren, N., Al-Aljmi, A.-H., Alshalan, S., Alshalan, R., Alrumayyan, N., Al-Manea, S., Bawazeer, S., Al-Mutlaq, N., Almanea, N., Huwaymil, W. B., Alqusair, D., Alotaibi, R., Al-Senaydi, S., and Alfutamani, A. (2018). Suar: Towards building a corpus for the saudi dialect. In *ACLING*.

Almeman, K. and Lee, M. (2013). Automatic building of Arabic multi dialect text corpora by bootstrapping dialect words. In *Proceedings of the International Conference on Communications, Signal Processing, and their Applications (ICCSPA)*, pages 1–6.

Appen Pty Ltd. (2006a). Gulf Arabic Conversational Telephone Speech, Transcripts LDC2006T15.

Appen Pty Ltd. (2006b). Iraqi Arabic Conversational Telephone Speech, Transcripts LDC2006T16.

Appen Pty Ltd. (2007). Levantine Arabic Conversational Telephone Speech, Transcripts LDC2007T01.

BBN Technologies, R., Linguistic Data Consortium, and Sakhr Software. (2012). Arabic-Dialect/English Parallel Text LDC2012T09.

BBN Technologies. (2005). BBN/AUB DARPA Babylon Levantine Arabic Speech and Transcripts LDC2005S08.

Bouamor, H., Habash, N., and Oflazer, K. (2014). A multidialectal parallel corpus of Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland.

Bouamor, H., Habash, N., Salameh, M., Zaghouani, W., Rambow, O., Abdulrahim, D., Obeid, O., Khalifa, S., Eryani, F., Erdmann, A., and Oflazer, K. (2018). The MADAR Arabic Dialect Corpus and Lexicon. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

Bouamor, H., Hassan, S., and Habash, N. (2019). The MADAR shared task on Arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 199–207, Florence, Italy, August. Association for Computational Linguistics.

Chen, S., Fore, D., Strassel, S., Lee, H., and Wright, J. (2017). Bolt egyptian arabic sms/chat and transliteration ldc2017t07.

Cotterell, R. and Callison-Burch, C. (2014). A multi-dialect, multi-genre corpus of informal written Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 241–245, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Diab, M., Habash, N., Rambow, O., AlTantawy, M., and Benajiba, Y. (2010). COLABA: Arabic Dialect Annotation and Processing. In *Proceedings of the Workshop on Language Resources and Human Language Technology for Semitic Languages*.

El-Haj, M. (2020). Habibi - a multi dialect multi national Arabic song lyrics corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1318–1326, Marseille, France, May. European Language Resources Association.

Elfardy, H., Al-Badrashiny, M., and Diab, M. (2014). Aida: Identifying code switching in informal Arabic text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 94–101, Doha, Qatar.

Gadalla, H., Kilany, H., Arram, H., Yacoub, A., El-Habashi, A., Shalaby, A., Karins, K., Rowson, E., MacIntyre, R., Kingsbury, P., Graff, D., and McLemore, C. (1997). CALLHOME Egyptian Arabic Transcripts LDC97T19.

Jarrar, M., Habash, N., Alrimawi, F., Akra, D., and Zalmout, N. (2016). Curras: an annotated corpus for the Palestinian Arabic dialect. *Language Resources and Evaluation*, pages 1–31.

Khalifa, S., Habash, N., Abdulrahim, D., and Hassan, S. (2016). A Large Scale Corpus of Gulf Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Portorož, Slovenia.

Khalifa, S., Habash, N., Eryani, F., Obeid, O., Abdulrahim, D., and Al Kaabi, M. (2018). A morphologically annotated corpus of Emirati Arabic. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Linguistic Data Consortium. (2002). CALLHOME Egyptian Arabic Transcripts Supplement LDC2002T38.

Maamouri, M., Buckwalter, T., and Jin, H. (2005). Levantine Arabic QT Training Data Set 4 (Speech + Transcripts) LDC2005S14.

Maamouri, M., Buckwalter, T., Graff, D., and Jin, H. (2006). Levantine Arabic QT Training Data Set 5, Transcripts LDC2006T07.

Maamouri, M., Buckwalter, T., Graff, D., and Jin, H. (2007). Fisher levantine arabic conversational telephone speech, transcripts ldc2007t04.

McNeil, K. and Faiza, M. (2011). Tunisian Arabic Corpus : Creating a Written Corpus of an "Unwritten" Language. In *Proceedings of the Workshop on Arabic Corpus Linguistics (WACL)*.

Meftouh, K., Harrat, S., Jamoussi, S., Abbas, M., and Smaili, K. (2015). Machine translation experiments on PADIC: A parallel Arabic DIalect corpus. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 26–34, Shanghai, China, October.

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

Sadat, F., Kazemi, F., and Farzindar, A. (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.

Salama, A., Bouamor, H., Mohit, B., and Oflazer, K. (2014). YouDACC: the Youtube Dialectal Arabic Comment Corpus. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 1246–1251, Reykjavik, Iceland.

Sawalha, M., Alshargi, F., AlShdaifat, A., Yagi, S., and Qudah, M. A. (2019). Construction and annotation of the jordan comprehensive contemporary Arabic corpus (JCCA). In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 148–157, Florence, Italy, August. Association for Computational Linguistics.

Shon, S., , Ali, A., Samih, Y., Mubarak, H., and Glass, J. (2020). Adi17: A fine-grained arabic dialect

identification dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8244–8248.

Smaïli, K., Abbas, M., Meftouh, K., and Harrat, S. (2014). Building resources for Algerian Arabic dialects. In *Proceedings of the Conference of the International Speech Communication Association (Interspeech)*.

Zaghouani, W. and Charfi, A. (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.

Zaidan, O. F. and Callison-Burch, C. (2011). The Arabic Online Commentary Dataset: an Annotated Dataset of Informal Arabic With High Dialectal Content. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, pages 37–41.