DART: A Large Dataset of Dialectal Arabic Tweets

Israa Alsarsour, Esraa Mohamed, Reem Suwaileh, Tamer Elsayed
Computer Science and Engineering Department, Qatar University, Doha, Qatar
{ia1205702, em1205267, reem.suwaileh, telsayed}@qu.edu.qa

Abstract

In this paper, we present a new large manually-annotated multi-dialect dataset of Arabic tweets that is publicly available. The Dialectal ARabic Tweets (DART) dataset has about 25K tweets that are annotated via crowdsourcing and it is well-balanced over five main groups of Arabic dialects: Egyptian, Maghrebi, Levantine, Gulf, and Iraqi. The paper outlines the pipeline of constructing the dataset from crawling tweets that match a list of dialect phrases to annotating the tweets by the crowd. We also touch some challenges that we face during the process. We evaluate the quality of the dataset from two perspectives: the inter-annotator agreement and the accuracy of the final labels. Results show that both measures were substantially high for the Egyptian, Gulf, and Levantine dialect groups, but lower for the Iraqi and Maghrebi dialects, which indicates the difficulty of identifying those two dialects manually and hence automatically.

Keywords: Arabic, Multi-Dialect, Twitter, Crowdsourcing, Annotations, Corpus

1. Introduction

The Arabic language is the fifth most widely spoken language in the world; more than 380 million people speak and write in Arabic(Darwish et al., 2014). Additionally, approximately 41.7% of Arabic speakers are using the Internet[1] which necessitates the need for developing language-specific tools for Arabic. The Arabic language has many dialects (varieties), besides the Modern standard Arabic (MSA), that are broadly used in daily life(Huang, 2015). Although dialects have common linguistic uses, they greatly differ making Arab people themselves face difficulty in understanding each other. The variations and similarities of Arabic dialects stem from different factors, e.g., social class, education level, religion, gender, and geographical region (Benajiba and Diab, 2010).

Arab users write in dialects over the Internet and extensively in social media. This introduces many challenges to researchers in areas such as Natural Language Programming, Information Retrieval, and Machine Learning, who deal with the spoken and/or written language one way or the other. The advancement in these areas is remarkably restricted by the shortage of high-quality Arabic language resources.

In this paper, we tackle the problem of building a large dialectal Arabic tweets dataset that somewhat remedies the lack of Arabic resources and opens the door of support to tackle various research problems such as dialect detection, words segmentation, translation, cross-dialect search, and speech recognition. Our contribution in this work is two folds:

1. We introduce ***DART*** dataset; a large well-balanced publicly-available[2] **D**ialectal **AR**abic **T**weets dataset that we believe will enable research on different areas.
2. We provide an analysis on the quality of the dataset in terms of inter-annotator agreement (measured in Kappa), and accuracy of final labels (measured by in-house annotators).

The rest of this paper is organized as follows. Section 2 reviews the literature and describes the publicly-available datasets. We layout the process of collecting and annotating DART dataset in Sections 3 and 4. Section 5 provides a comparison between DART and the other similar datasets. Finally, we conclude and discuss possible future directions in Section 6.

2. Related Work

In this section, we review the available dialectal Arabic datasets and discuss their properties.

The Arabic Online news Commentary (AOC) (Zaidan and Callison-Burch, 2011) is the first available dialectal dataset that contains 3.1M comments gathered from Egyptian, Gulf, and Levantine news websites. The authors initially labeled only around 0.05% of the dataset by Amazon's Mechanical Turk (MTurk) crowdsourcing platform [3]. Many researchers used AOC dataset by either extending the annotations (Cotterell and Callison-Burch, 2014) or directly using it for different purposes such as extracting dialectal n-grams to automatically label tweets by their dialect (Mubarak and Darwish, 2014). (Cotterell and Callison-Burch, 2014) extended the AOC dataset to cover Maghrebi (MG) and Iraqi dialects. They also crawled tweets using Twitter API and labeled them using MTurk. The dataset is not balanced across dialect groups. Moreover, it contains lots of noise such as Arabizi and French tweets.

(Bouamor et al., 2014) used an Egyptian-English corpus (Zbib et al., 2012) as seed corpus and asked four in-house Arabic native speakers from Palestine, Syria, Jordan, and Tunisia to translate 2,000 Egyptian sentences into their dialects. A major issue of this dataset is the approach that generated sentences that do not reflect the natural way of writing and speaking in dialects. Additionally, as annotators were selected from few countries, the dataset provides biased labeled data for only two dialectal groups (Levantine and Maghrebi) besides Egyptian group.

Thus far, we discussed datasets that were manually labeled (either by in-house or crowdsourcing annotators). We fur-

[1] http://www.internetworldstats.com/stats19.htm
[2] http://qufaculty.qu.edu.qa/telsayed/datasets/
[3] www.mturk.com

ther discuss datasets that are collected and labeled automatically. (Mubarak and Darwish, 2014) used the geo-location attribute of tweets to automatically label them by their corresponding dialect. (Eldesouki et al., 2017) selected 350 tweets from this corpus for the five-dialectal groups and labeled them manually. Similarly, (Huang, 2015) also used the geographical location of Facebook posts to label them and create a week classifier to detect the dialect of posts. The classifier was trained to detect the five dialects groups. All these datasets are rather small with respect to the current standards. (Salama et al., 2015) also labeled Youtube comments and videos description using their geographic location. They randomly selected 1,000 sentences from each dialectal corpus and asked two native speakers to judge them. Differently, (Almeman and Lee, 2013) proposed an automatic approach to collect dialectal Arabic web pages. They covered only four of the common dialectal groups (they combine Iraqi and Gulf dialects). Their approach has a pipeline of gathering and filtering steps. The major issue with this dataset is that the pages might also contain MSA sentences which are hard to separate.

3. Collecting Dialectal Data

In this section, we describe the pipeline of collecting *DART* dataset. We started by manually collecting popular dialectal phrases for each Arabic dialect group as listed in table 1. After filtering out inappropriate and common phrases (i.e., those used in more than one dialect group), we tracked the unique phrases over Twitter stream. As the stream is flooded by spam and retweets, we cleaned the collected data annotating it. We elaborate thoroughly on each step in the following subsections.

3.1. Collecting Dialectal Phrases

For each dialect[4], we target distinct phrases that are spoken by only the native speakers of that dialect. We collected a list of dialectal phrases from two sources. We first acquired a list of 1,000 dialectal words collected by (Almeman and Lee, 2013). The list covers only four dialects: EGY, GLF, LEV, MGH. To diversify the sources from where we collected the dialectal phrases, we extended the list with phrases from *Mo3jam* website[5], which allows Arab users to contribute with dialectal phrases spoken in their countries. For each dialect group, we randomly selected phrases from the list of phrases of each country under that dialect from that website.

We performed several filtering steps on both lists of dialectal phrases. We first manually dropped inappropriate phrases. We then issued each phrase against Twitter Live search interface [6] and excluded any phrase that returns inappropriate or no results. We also filtered out phrases that returned tweets in different dialects. We ended up with 232 phrases on average for each dialect: 278 for EGY, 246 for GLF, 244 for LEV, 121 for IRQ, and 273 for MGH. We share all of the final lists of phrases in our released dataset.

[4]We will use "dialect" to denote "dialect group" from now on.

[5]ar.mo3jam.com/

[6]twitter.com/search-home

3.2. Tracking Tweets

To construct a potential dialectal dataset, we tracked the list of phrases using Twitter streaming API for tracking[7]. The tracking period spanned about two months sporadicly (from 25 of February to 5 of May 2017). Table 1 shows (in the third column) examples of tracked tweets and their corresponding tracked phrases (in bold face). We also report the number of tweets crawled for each dialect group in the fourth column of the table.

3.3. Cleaning Tweets

Although we crawled the potential dialectal dataset via tracking dialectal phrases, the dataset might still contain lots of noise such as multilingual tweets (i.e., tweets written in other languages besides Arabic), inappropriate tweets, etc. Therefore, to have a better-quality potential dataset for labeling, we cleaned the dataset as follows:

- *Filtering out non-Arabic tweets*: Many Arab users post tweets written in multiple languages. For example, users from North Africa tend to write in the French language besides Arabic. Moreover, many other Arab users prefer to communicate in Arabizi [8] (Arabic phrases written in English alphabet) on Twitter. Although foreign words and phrases might be good indicator of the dialect of the text, we opted to drop the tweets that are mostly written in non-Arabic language.
- *Filtering out inappropriate tweets*: Having a list of around 300 manually-collected Arabic inappropriate phrases, we cleaned the dataset by dropping tweets that contain any phrase from that list.
- *Filtering out short tweets*: To avoid ambiguity in very short tweets and hence difficulty and confusion in annotations, we also eliminated tweets that have less than three words.
- *Filtering out duplicates*: To save time of annotating duplicate tweets, we also excluded the retweets.

Applying the above cleaning steps, we finally obtained a cleaned version of the potential dataset that contains around 145K tweets, as shown in the right-most column of table 1.

4. Annotations via Crowdsourcing

Our objective is to create a good quality dialectal dataset that is of lasting value to researchers interested in working on Arabic dialects. Having a clean Arabic tweets dataset with pseudo-labels (i.e., tweets labeled by the dialect of the corresponding tracking phrase), the next step is to annotate the tweets in a more reliable way.

To accurately and reliably annotate the tweets with their corresponding dialects, we need multiple Arabic native speakers for each dialect. That is indeed challenging to find in a surrounding community. More importantly, the dataset size we want to annotate is quite large, hence it is too expensive to recruit in-house annotators for that task. Therefore, we used CrowdFlower crowdsourcing platform[9] to acquire

[7]developer.twitter.com/en/docs/tutorials/consuming-streaming-data#track

[8]en.wikipedia.org/wiki/Arabic_chat_alphabet

[9]www.crowdflower.com

Dialect	Countries	Example Tweet	Collected	Clean
EGY	Egypt	انا **عاوزه** اعدي التيرم ده يا عاطف	89,424	37,834
GLF	UAE, KSA, Bahrain, Kuwait, Qatar	ودي اشوفكم **دحين** تعالو بسرعة	89,709	35,137
LEV	Palestine, Jordan, Syria, Lebanon	فوتو احكوا شو **بدكن**	75,549	23,039
IRQ	Iraq	**شكد** حلوين شلون كاعدين	55,464	23,236
MGH	Morocco, Algeria, Tunisia	**واش بيك** واش راك تحكي مافهمتش عليك	36,991	16,350
Total	-	-	347,137	144,596

Table 1: Dialect groups and corresponding collected/cleaned dialectal data.

Dialect	Test Qs	Dial. Tweets (%)	Kappa	Acc.
EGY	338	5,265 (75%)	0.71	97%
GLF	340	5,893 (84%)	0.71	100%
LEV	347	3,939 (55%)	0.62	96%
IRQ	234	5,253 (75%)	0.42	78%
MGH	398	3,930 (55%)	0.28	88%
Total	1,657	24,280 (69%)	-	92%

Table 2: Different statistics about DART.

annotations. In the following subsections, we describe our annotation process.

4.1. Task Design

For each dialectal group, we randomly selected about 7K tweets from the potential dialect dataset to be annotated by contributors on CrowdFlower. In order to increase both accuracy and reliability, we designed one job (with relatively simple instructions) for each dialectal group, aiming at finding native speakers for that group. For each job, we allow only contributors from the corresponding countries of the dialect group to work on the job. For example, for Gulf dialect job, only participants from Bahrain, Oman, Kuwait, Qatar, Saudi Arabia, and United Arab Emirates are eligible to work. We designed the annotation task to show contributors ten tweets per page, and asked them to label each tweet by either the corresponding dialect (indicated by the pseudo-label of the tweet), MSA, or other (in case the tweet is written in other dialect or annotators could not identify its dialect).

4.2. Quality Control

We adopted a common quality control method to ensure high-quality labels. We sought five native speakers, one for each dialectal group, to label around 300 to 400 tweets (1,657 tweets in total) and use those as the source of quiz/test questions in our crowdsourcing jobs. We required the contributors to attain at least level 2 (moderate) according to CrowdFlower rating. We randomly selected 10 tweets from that set as quiz questions to examine the contributors before they start the job. A minimum accuracy of 90% was required to pass the quiz.

We also used the full set of labeled tweets as "gold" questions to ensure a consistent performance of the contributors throughout the job. When contributors accuracy fell under the predefined accuracy level, they were excluded. Finally, we collected 3 annotations for each tweet from different contributors to increase confidence in labeling.

4.3. Pilot Studies

Before we launch the actual jobs, we conducted several pilot studies for each dialectal group separately (using 100 tweets for each). We aimed at estimating the required budget (e.g., cost and time) and improving the instructions, design, and setup of the jobs. We list here the major challenges we encountered while we ran these small-scale studies.

- Inaccessible Tweets: We used Twitter widget to display the tweets on the task interface using the tweet ID. However, when tweets are deleted or the author make his profile private, the annotators are no longer able to label them. This is critical especially if the tweets are used as gold questions. In such case, the annotator would arbitrarily choose a label, which in turn affects their performance. To resolve this issue, we periodically checked the test questions during the period when the tasks were running and removed the inaccessible ones.

- Lack of country-specific contributors: Some Arab countries do not have contributors on CrowdFlower. This was evident specifically for the Iraqi dialect, hence, we opted to disable the geographical constraint on the contributors of the Iraqi job. As this decision has the potential to affect the quality of the labels, we limited the contributors to be only from the Gulf countries for that job, as they are the closest group to the Iraqi dialect.

4.4. Aggregation and Agreement

To aggregate the multiple labels per tweet, we opted to use majority voting which requires at least two annotators to agree on the label. This resulted in 24,280 dialectal tweets. To asses the reliability of the agreement, we measured the inter-annotator agreement using Fleiss Kappa (Fleiss, 1971). Fleiss Kappa is used when more than two annotators labeled a data item (a tweet in our case) to measure the degree of agreement over what would be expected by chance. We found the degree of agreements substantial for three dialects EGY, GLF, and LEV, moderate for IRQ, and fair for MGH. We show the exact kappa values in Table 2.

4.5. Accuracy

To evaluate the accuracy of the crowdsourcing labels, we randomly selected 100 tweets per dialectal group and asked

Dataset	Source	Size	Dial. Groups	Labels	Public?
(Cotterell and Callison-Burch, 2014)	Twitter and AOC	67,468	5G	Manual	✓
(Zaidan and Callison-Burch, 2011)	News Comments	44,618	5G-{IRQ, MGH}	Manual	✓
DART	Twitter	24,280	5G	Manual	✓
(Bouamor et al., 2014)	Egy-Eng Corpus	5,000	5G-{GLF, IRQ}	Manual	✓
(Eldesouki et al., 2017)	Twitter	1,400	5G-{IRQ}	Manual	✓
(Huang, 2015)	FaceBook	66M	5G	Auto	✗
(Mubarak and Darwish, 2014)	Twitter	6.5M	5G	Auto	✗
(Almeman and Lee, 2013)	Web Corpus	2M	5G-{IRQ}	Auto	✓
(Salama et al., 2015)	YouTube	640,817	5G	Auto	✗

Table 3: A comparison between DART and datasets used in literature.

one native speaker from each group to *re-label* the corresponding tweets. The last column in Table 2 shows the accuracy for each group. It indicates that accuracy for GLF, EGY, and LEV is high (ranges between 100% and 96%), a little lower for MGH (88%), and much lower for IRQ (78%). This is somewhat aligned with the inter-annotator agreement values. In fact, both indicate that manually-identifying Iraqi and Maghrebi tweets is very challenging, which in turn hints about the difficulty dialect identification systems would face in identifying them too.

5. DART among Others

Table 3 illustrates a comparison between DART and the Arabic dialectal datasets used in the literature. For each dataset, the table indicates the data source, the size of the dataset (in sentences or tweets), the dialectal groups covered (5G denotes the five groups we covered), the type of annotations (manually or automatically), and whether it is publicly available or not. While the table shows that DART is the third largest dataset among the manually-annotated ones, the largest two have balancing and coverage limitations that make them less usable. The first is not well balanced over the five groups as it has too many GLF (63%) but very few IRQ ($<$1%) and MGH (10%) sentences or tweets, while the second covers only three dialectal groups. On the contrary, DART is well-balanced over the five groups it covers. Moreover, DART is exclusively composed of tweets, which makes it more homogeneous and thus suitable for training Twitter-specific dialect identification systems.

6. Conclusion and Future Work

We introduced DART, a large multi-dialect dataset of Arabic tweets that is publicly-available. The dataset is composed of about 25k labeled tweets and is well balanced over five common dialect groups. DART is constructed over a well-planned pipeline and was annotated via crowdsourcing. Measures of inter-annotator agreement as well as accuracy of final labels showed high quality and hence high potential of utilizing the dataset as a rich resource for the community.

DART opens several possible future research directions. It can be used as a gold-standard for training and evaluating Arabic dialect detection systems. It can also be extended to a more fine-grained level of annotations per country. It can even enable further studies on the differences and commonalities between Arabic dialects.

7. Acknowledgments

This work was made possible by NPRP grant# NPRP 7-1313-1-245 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors. The work was also supported by grant QUST-CENG-SPR-2017-21 from College of Engineering at Qatar University.

8. Bibliographical References

Al-Mannai, K., Sajjad, H., Khader, A., Al Obaidli, F., Nakov, P., and Vogel, S. (2014). Unsupervised word segmentation improves dialectal Arabic to English machine translation. *ANLP 2014*, page 207.

Almeman, K. and Lee, M. (2013). Automatic building of Arabic multi dialect text corpora by bootstrapping dialect words. In *Communications, signal Processing, and their Applications (ICCSPA), 2013 1st international conference on*, pages 1–6. IEEE.

Benajiba, Y. and Diab, M. (2010). A Web application for dialectal arabic text annotation. In *Proceedings of the LREC workshop for language resources (LRS) and human language technologies (HLT) for semitic languages: Status, updates, and prospects*.

Bouamor, H., Habash, N., and Oflazer, K. (2014). A multidialectal parallel corpus of Arabic. In *LREC*, pages 1240–1245.

Cotterell, R. and Callison-Burch, C. (2014). A multi-dialect, multi-genre corpus of informal written Arabic. In *LREC*, pages 241–245.

Darwish, K., Sajjad, H., and Mubarak, H. (2014). Verifiably effective Arabic dialect identification. In *EMNLP*, pages 1465–1468.

Eldesouki, M., Samih, Y., Abdelali, A., Attia, M., Mubarak, H., Darwish, K., and Laura, K. (2017). Arabic multi-dialect segmentation: bi-lstm-crf vs. svm. *arXiv preprint arXiv:1708.05891*.

Elfardy, H. and Diab, M. T. (2013). Sentence level dialect identification in Arabic. In *ACL (2)*, pages 456–461.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Huang, F. (2015). Improved Arabic dialect classification with social media data. In *EMNLP*, pages 2118–2126.

Mubarak, H. and Darwish, K. (2014). Using Twitter to collect a multi-dialectal corpus of arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 1–7.

Salama, A., Bouamor, H., Mohit, B., and Oflazer, K. (2015). Youdacc: the youtube dialectal Arabic commentary corpus.

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 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 37–41. Association for Computational Linguistics.

Zbib, R., Malchiodi, E., Devlin, J., Stallard, D., Matsoukas, S., Schwartz, R., Makhoul, J., Zaidan, O. F., and Callison-Burch, C. (2012). Machine translation of arabic dialects. In *Proceedings of the 2012 conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 49–59. Association for Computational Linguistics.