Crowdsource a little to label a lot: Labeling a Speech Corpus of Dialectal Arabic

Samantha Wray, Ahmed Ali

Qatar Computing Research Institute
Hamad Bin Khalifa University
Doha, Qatar
{swray,amali}@qf.org.qa

10.21437/Interspeech.2015-594

Abstract

Arabic is a language with great dialectal variety, with Modern Standard Arabic (MSA) being the only standardized dialect. Spoken Arabic is characterized by frequent code-switching between MSA and Dialectal Arabic (DA). DA varieties are typically differentiated by region, but despite their wide-spread usage, they are under-resourced and lack viable corpora and tools necessary for speech recognition and natural language processing. Existing DA speech corpora are limited in scope, consisting of mainly telephone conversations and scripted speech.

In this paper we describe our efforts for using crowdsourcing to create a labeled multi-dialectal speech corpus. We obtained utterance-level dialect labels for 57 hours of high-quality audio from Al Jazeera consisting of four major varieties of DA: Egyptian, Levantine, Gulf, and North African. Using speaker linking to identify utterances spoken by the same speaker, and measures of label accuracy likelihood based on annotator behavior, we automatically labeled an additional 94 hours. The complete corpus contains 850 hours with approximately 18% DA speech.

Index Terms: crowdsourcing, human computation, dialect classification, Arabic, corpora creation, speech corpora

1. Introduction

Arabic as a language consists of numerous varieties. Modern Standard Arabic (MSA) is the standardized dialect of news media and schooling, and the varieties of Dialectal Arabic (DA) that characterize day-to-day usage can be very roughly categorized into four broad categories based on region of usage: Egyptian, Levantine (spoken in Syria, Lebanon, Jordan, and Palestine), Gulf (spoken in Saudi Arabia, Qatar, the United Arab Emirates, Bahrain, Oman, Kuwait, Yemen, and Iraq[1]), and North African (spoken in Morocco, Libya, Tunisia, Algeria, and Mauritania.)

Existing Arabic speech corpora are dominated by MSA, and the few colloquial resources (with notable exceptions: 20 hours of Egyptian [1], 45 hours of Levantine [2], 32 hours of Gulf, Levantine, and Egyptian [3], 15 hours of Gulf [4]) consist of narrow bandwidth telephone conversations.

Crowdsourcing has become a standard method for accessing large numbers of participants who are demographically diverse and harbor a number of skillsets that can be utilized for collection and annotation of data in various speech and language processing studies, such as text corpus construction [5] and acquisition of translations [6]. Within the specific domain of speech data, crowdsourcing has been used effectively for transcription of speech [7], and collection of speech via prompts [8], [9], [10], among other tasks. Numerous studies have investigated the development of quality control mechanisms which can be used to obtain expert-level quality data at a much lower cost [11], [12], making crowdsourcing a viable method for efforts of speech corpus building and labeling.

In this paper, we present a multi-dialectal speech corpus of DA created from high-quality broadcast, debate and discussion programs from Al Jazeera, and as such contains a combination of spontaneous and scripted speech. We utilize human computation by means of crowdsourcing, and develop methods for selecting representative utterances for each speaker to minimize the necessity of complete human annotation for the whole corpus. The paper is organized as follows: in Section 2, we describe the process of collecting speech data from Al Jazeera and selecting representative samples from each speaker for manual classification. Section 3 presents the role of human annotation and development of best practices for obtaining reliable classifications. Then, annotator behavior and implications for perception are described in Section 4. Section 5 shows the results of generalizing dialect information for all speech data based on annotations for representative samples. Finally, Section 6 summarizes results and outlines ongoing and future research.

2. Speech Data

The Qatar Computing Research Institute (QCRI) has worked closely with Al Jazeera to develop a transcription queue which allows journalists and editors at Al Jazeera to choose episodes to be automatically transcribed by the QCRI Advanced Transcription System (QATS) [13]. All videos processed by QATS appear on Al Jazeera's Arabic site aljazeera.net. The transcriptions have been formatted into SRT and DFXP subtitles and have been uploaded to the Brightcove video platform. The audio which makes up the corpus in the current study was pulled from videos in the transcription queue in the time period between June 2014 and January 2015, with an average of 33 videos per day. In total there were more than 8500 video files which contain approximately 850 hours of speech. The audio is a mix of programs, reports, and conversational debates. The data is 44Khz with the highest quality which has been uploaded directly from Al Jazeera to Brightcove. After downloading the video files, we ran ffmpeg to downsample to 16Khz, and then ran each audio file through a pre-processing pipeline before submitting it to annotators.

The pre-processing stage consisted of the following steps:

[1] It can be argued that Yemeni and Iraqi Arabic are distinct enough to be excluded from the Gulf dialect group. The authors have plans to visit a more fine-grained classification in the future.

First, for each episode, we ran Voice Activation Detection (VAD) to remove as many non-speech segments (such as music or white noise) as possible. Then, speaker diarization was performed to determine who speaks when, and to assign each segment a speaker ID. All the aforementioned data pre-processing was carried out using LIUM SpkDiarization [14]. The output from LIUM segmentation is typically small chunks of audio files containing information about speaker ID, speaker gender and duration of utterance.

2.1. Segmentation and Speaker Linking

As a result of processing the data using LIUM, the audio was split into 167,000 segments. Then, we ran a second step in which we concatenated consecutive segments from the same speaker if a one-second or less period of silence or non-speech separated them. The aim of this step was to reduce the number of segments to submit for manual labeling. At this stage, we also discarded any segment less than three seconds as we felt dialect assessment would be too difficult for the annotator in such a short span of time. After concatenation, 121,000 segments remained. These 121,000 represent the "Expanded" data set which contains every utterance.

LIUM also provided speaker linking information in which different speech segments produced by the same speaker were assigned to the same ID within the same file. From the 121,000 segments, two segments per speaker per video were selected, typically first and last segments, resulting in a total of 47,696 segments of unknown dialects to be labeled by human annotators. This subset represents the "Sample" data set. The assumption was that labels for the Sample set can be generalized to segments from the same speaker in the Expanded data set. The crowdsourced labeling of the Sample set is described in Section 3 and the process of expansion of the Sample set to the Expanded set is evaluated in Section 5.

3. Crowdsource Task

Crowdsourced classifications were obtained via CrowdFlower (henceforth CF) [15], a service that utilizes various worker channels including other microworking and rewards sites. Workers can also be targeted by country of user origin. The service also employs optional verification stages in which gold standard data can be used to verify contributor answers as they are submitted. Additionally, it also makes use of a dynamic judgment system in which more annotators are recruited for items which have low inter-annotator agreement.

The output of CF tasks takes various forms. First, the service outputs the full collection of contributor answers. CF also aggregates these answers together so each task item is assigned a single answer based on inter-annotator agreement scores and contributor trust calculated from their performance in previous tasks. The combined trust and inter-annotator agreement calculation is quantified in a value called "confidence" which is also released with the aggregate data. Thus, each item in the aggregate data set is assigned an answer and a confidence value for that answer. CF also outputs a list of contributors which worked on the task with information about which worker channel they were recruited from and location data.

3.1. Task anatomy

The task was restricted to users in the Arab world. All directions for the task were in written Modern Standard Arabic. Contributors were directed to listen to the short speech segments described in Section 2.1 and determine which dialect they thought the speaker was speaking. Contributors were asked to listen only as long as necessary to determine the dialect being spoken. Compensation for this task was USD 0.03 per page of 10 items.

Dialect judgment was answered by a seven-way forced-choice between Modern Standard Arabic (MSA), Levantine Arabic (LEV), Egyptian Arabic (EGY), North African Arabic (NOR), Gulf Arabic (GLF), non-Arabic speech, and non-speech. The non-Arabic speech in the data included foreign speakers who were not dubbed over. The non-speech included white noise, music, and other non-speech sounds such as traffic and gunfire, which were mislabeled by LIUM as speech data. For each regional variety of DA, contributors were explicitly instructed which countries belonged to which dialect groups.

3.2. Development of quality measures

Existing CF quality control options were utilized to reduce the amount of noisy data and post-crowdsource cleanup necessary. Twenty-five audio files were manually annotated to create a gold standard data set in order to use CF automatic quality control. These files were selected to be unambiguous and clear, and the answers distributed across categories with little potential for dispute, such as non-speech, non-Arabic, MSA, in addition to clear examples of DA. Pilot testing confirmed that the gold standard items were appropriately unambiguous.

Live quality control was accomplished in two ways. First, CF optional Quiz Mode was engaged, which required contributors to answer five gold standard items before entering the main portion of the task. Second, for every five items, contributors were presented with a gold standard item that was not discernible from the task items. Contributors had to maintain an accuracy of at least 65% on these hidden gold standard items or else their participation in the task was ended. Although this cutoff point may appear too forgiving, pilot work showed that spammy annotators had an average accuracy of 31% on test questions whereas the remainder of annotators had an average of 94% accuracy. In addition to utilizing live quality control, efforts were also made to reduce the amount of data with low inter-annotator agreement. Recall that CF features a built-in mechanism for fetching additional contributors to provide judgments for items with low inter-annotator agreement. Recall also that each item is assigned a confidence value based on inter-annotator agreement and contributor trust. To determine the most effective way to utilize this feature, an experiment was performed on a random sample of 500 segments. This sample was submitted for contributor judgments three times on CF with different manipulations of both confidence thresholds and maximum number of contributors per item. Suggested settings for the dynamic judgments feature which automatically submits low-agreement items are to resubmit an item with lower than 70% to one additional contributor. This feature was tested on the 500 set, as well as a higher threshold of 75%, and two maximum contributor-per-item numbers: 7 and 9. This experiment demonstrated a gain in total percentage of high-confidence items, as the threshold was made stricter and the number of annotators higher. These results are summarized in Table 1. After determining best practices for dynamic judgments and quality control, the 47,696 sample files representing 404 hours of speech were classified over a period of three weeks, costing a total of USD 971.

Threshold	Minimum contributors	Maximum contributors	% items above 70% confidence
70%	3	4	79%
75%	3	7	89%
75%	3	9	92%

Table 1: Percentage of high-confidence answers for 500 segments annotated with three dynamic judgment options

3.3. Contributor Demographics

A total of 2,053 users contributed to the labeling task, with 39% of contributors hailing from Egypt, the single highest country by contributor count. Complete contributor counts by country[2] are shown in Table 2. In comparing the numbers of contributors

Egypt	795	Saudi Arabia	80	Oman	10
Algeria	422	Palestine	51	Bahrain	4
Tunisia	303	Yemen	45	Qatar	3
Jordan	177	UAE	33		
Morocco	117	Kuwait	13	Total	2053

Table 2: Contributor count by country

based on their dialect group, North African speakers contributed the highest total percentage to the task. Lowest participation by number of contributors was from countries in the Gulf. Percentages of total contributors per dialect group are shown in the map in Figure 1.

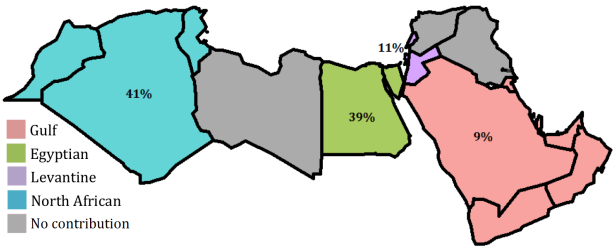

Figure 1: Map of contributor origin by dialect group

Note in Figure 1 that although the Gulf region is a large multi-national group, it contributed a minority of the participants. Potential implications for this and other contributor origin-related phenomena are discussed in the following section.

4. Dialect perception

Although the aim of this paper is primarily concerned with resource improvement and data collection through crowdsourcing, insights on human perception were also investigated based on contributor behavior. We considered the possibilities of annotator bias during the process of labeling, and explored implications of labels which regularly co-occurred.

4.1. Contributor bias

Overall, of the four major DA varieties, labels assigned to Egyptian had the overall highest average confidence value and labels for Gulf exhibited the lowest average confidence value. Percentages for confidence values for items are shown by label in Figure 2. Items were binned according to three confidence thresholds: less than 50% confidence, between 50% and 75% confidence, and finally above 75%.

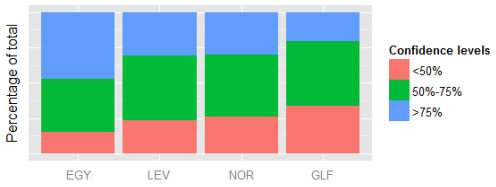

Figure 2: Distribution of confidence by dialect

As for the relation between annotator origin and label assigned, Zaidan and Callison-Burch [16] present evidence of annotator bias in a task identifying dialectal content in text mined from comments on on-line news articles. They found that annotators were biased towards selecting their own native dialect when asked to provide dialect judgments. Thus, Egyptian speakers often mistakenly annotated non-Egyptian comments as being Egyptian, Levantine speakers over-annotated sentences as Levantine, and so forth. This raises the question of the current study: Is there any evidence that contributors were biased towards selecting their own dialect when presenting with speech of unknown origin? To determine this, we also presented annotators with twenty-five manually-annotated items per DA category to compare behavior across origins of annotator.

A chi-square test of independence was performed to determine whether an annotator's dialect of origin affected their selection and therefore whether DA selections were equally distributed. Annotators chose their own dialect 22 $\pm$ 3.7% of the time, which was not significantly different from their probability of choosing other dialects 22.2 $\pm$ 2.7% of the time; ($\chi^2(12)=12.7$, p=0.4)). Thus, contributors did not exhibit a bias to their own native dialect group in the process of making dialect judgments.

Although it may be difficult in certain contexts to determine the dialect of a written comment if it contains graphemic cognates common across multiple dialects of colloquial Arabic and even Modern Standard, this ambiguity is absent in spoken utterances. The specific cues that lead to differences in dialect confusability across written and spoken modalities is beyond the scope of this paper, but merit further investigation.

4.2. Interdialectal confusability

Recall that a label is assigned to an item based on the judgments of several annotators and in the event an item exhibited low inter-annotator agreement, more annotators would automatically be obtained to provide additional judgments. Each label then is the product of judgments from 3-9 different annotators. However, what was the cause of low agreement in the first place, and was there a pattern to contributor disagreement? To investigate the rates of confusability between dialects and the amount of ambiguity which led to high competition between multiple dialect judgments for one item, we counted each judgment provided to each label. Percentages are shown in Table 3.

Results suggest that Egyptian is easily distinguished from other varieties of DA, likely due to its wide-spread representation in media consumed throughout the Arabic-speaking world.

[2] At the time of this study, CF was not available for residents of Iraq, Syria, Libya, and Lebanon. Although the task was available for users in South Sudan, no contributors participated from this country.

	Percentage of total judgments				
Label	EGY	GLF	LEV	NOR	MSA
EGY	79.6%	1.3%	2.6%	1.1%	15.1%
GLF	1.4%	61.3%	11.9%	5.0%	20.2%
LEV	1.7%	6.8%	73.8%	3.3%	14.1%
NOR	0.6%	5.1%	5.3%	70.5%	18.3%

Table 3: Percentages of judgments by label

This interpretation is consistent with the high confidence values for EGY labels as shown in 2. Although the GLF label exhibits the highest percentage of competition between GLF judgments and MSA when compared to other DA varieties (20.2% of GLF labels contained MSA judgments, whereas $15.8 \pm 2.2\%$ of EGY, LEV and NOR labels contained MSA judgments), a chi-square test of independence shows this difference was not significant (($\chi^2(1)$=0.91, p=0.3)).

5. Expansion Results

Recall that the annotated audio set was a subset of the larger audio set. In the process of linking annotated Sample files to the Expanded set in order to generalize contributor judgments, we explored three possible confidence threshold levels for expansion. First, we started with no threshold. All Sample items were eligible for expansion, and whatever answer was selected based on highest inter-annotator agreement and contributor trust was linked to the other files in the Expanded set. The second threshold was set at 50% confidence. At this threshold, any item with at least 50% confidence contributed dialect labels to the files it was linked to in the Expanded set. Items with less than 50% confidence were discarded. Finally, the strictest threshold was the 75% confidence level.

5.1. Validating the expansion process

In order to compare the three possible thresholds of expansion, a sample of randomly-selected previously-unseen 200 items per confidence threshold per dialect from the expanded sets were submitted to CF for manual annotation. The purpose of this was to determine if propagating labels from the Sample set to the Expanded set resulted in accurate labels. Table 4 shows the results of manual annotation of the selected sample of items from each confidence threshold. Common sense would predict that discarding items which were labeled with low confidence values even after multiple additional annotators improves the total percentage of dialect data during the expansion process, and the manually annotated results confirm this: the total percentage of predicted dialect increases as the confidence threshold becomes more restrictive.

However, as shown in Table 4, given that even a strict threshold of 75% doesn't produce full coverage of the predicted dialect, a question presents itself: what other speech is contained in the files and what makes it so easily confused with the predicted dialect?

5.2. Codeswitching

In looking at the results of the highest confidence threshold and the manually annotated dialect labels versus the expected dialect labels, it is clear that using an sample-expansion system doesn't result in completely generalizable labels. However, a closer look reveals that this could be due to the nature of codeswitching. Arabic as a language is characterized by frequent bi-dialectal codeswitching, meaning a speaker alternates between their native dialect and MSA [17]. Because of this fact, much of the remaining percentage of expected dialect data is in fact MSA, as shown in Table 5. (Remaining percentages belonged to Non-Arabic and Non-Speech categories.)

Confidence threshold	Expected Dialect	Hours linked	Confirmed % of sample
None	EGY	32h 59m	17%
	GLF	27h 11m	25%
	LAV	55h 42m	19%
	NOR	27h 02m	16%
50%	EGY	31h 31m	36%
	GLF	22h 17m	39%
	LAV	50h 30m	31%
	NOR	24h 32m	36%
75%	EGY	26h 37m	65%
	GLF	12h 30m	41%
	LAV	38h 49m	53%
	NOR	18h 24m	69%

Table 4: Results of manually-annotated expansion sets

Expected Dialect	EGY	GLF	LAV	NOR	MSA
EGY	65%				32%
GLF		41%	4%		53%
LAV	1%	1%	53%		39%
NOR	1%			69%	28%

Table 5: Percentages of expected dialect (from expansion) of segment by actual dialect (from manual annotation).

For speakers whose samples were labeled as a particular DA variety, the majority of their speech was indeed in that variety, with a minority being in MSA. The exception to this is the Gulf variety. It is therefore possible that Gulf speakers in the corpus used more MSA in their speech than their native dialect, but a comprehensive account of the differences in codeswitching for different DA varieties is warranted.

6. Conclusions and Future Research

This paper presents our efforts to create a multi-dialectal corpus of Arabic speech[3] using audio from Al Jazeera. We showed that using CrowdFlower to label samples from each speaker at the beginning and end of an audio segment results in labels for all of that speaker's speech and that results are suggestive of a regular practice of code-switching between one's native dialect and MSA. The corpus has been automatically transcribed, and utterances determined as DA have also begun to be manually transcribed using crowdsourcing.

7. Acknowledgments

We gratefully acknowledge Al Jazeera for their contributions to this study. We would like to thank them for their ongoing commitment to our research. Thanks are also due to Hamdy Mubarak, Stephan Vogel, Rolando Coto-Solano, and Dane Bell for their valuable input.

[3]The corpus can be accessed at http://alt.qcri.org/resources/aljazeeraSpeechCorpus/

8. References

[1] *A-SpeechDB v. 1.0*. European Language Resources Association, 2014.

[2] BBNTechnologies, (with American University of Beirut a subcontractor), J. Makhoul, B. Zawaydeh, F. Choi, and D. Stallard, *BBN/AUB DARPA Babylon Levantine Arabic Speech and Transcripts*. Linguistics Data Consortium, 2005.

[3] K. Almeman, M. Lee, and A. A. Almiman, "Multi dialect arabic speech parallel corpora," in *Communications, Signal Processing, and their Applications (ICCSPA), 2013 1st International Conference on*. IEEE, 2013, pp. 1–6.

[4] M. Elmahdy, M. Hasegawa-Johnson, and E. Mustafawi, "Development of a tv broadcasts speech recognition system for qatari arabic," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, N. C. C. Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, Eds. Reykjavik, Iceland: European Language Resources Association (ELRA), may 2014.

[5] W. B. Dolan and C. Brockett, "Automatically constructing a corpus of sentential paraphrases," in *Third International Workshop on Paraphrasing (IWP2005)*. Asia Federation of Natural Language Processing, 2005. [Online]. Available: http://research.microsoft.com/apps/pubs/default.aspx?id=101076

[6] O. F. Zaidan and C. Callison-Burch, "Crowdsourcing translation: Professional quality from non-professionals," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011, pp. 1220–1229.

[7] M. Marge, S. Banerjee, and A. I. Rudnicky, "Using the amazon mechanical turk for transcription of spoken language," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 5270–5273.

[8] I. Lane, A. Waibel, M. Eck, and K. Rottmann, "Tools for collecting speech corpora via mechanical-turk," in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics, 2010, pp. 184–187.

[9] M. H. Davel, C. J. v. Heerden, and E. Barnard, "Validating smartphone-collected speech corpora," in *Spoken Language Technologies for Under-Resourced Languages*, 2012.

[10] S. Novotney and C. Callison-Burch, "Shared task: crowdsourced accessibility elicitation of wikipedia articles," in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics, 2010, pp. 41–44.

[11] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, "Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '08. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008, pp. 254–263. [Online]. Available: http://dl.acm.org/citation.cfm?id=1613715.1613751

[12] S. Novotney and C. Callison-Burch, "Cheap, fast and good enough: Automatic speech recognition with non-expert transcription," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, ser. HLT '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 207–215. [Online]. Available: http://dl.acm.org/citation.cfm?id=1857999.1858023

[13] A. Ali, Y. Zhang, and S. Vogel, "Qcri advanced transcription system (qats)," in *Spoken Language Technology Workshop (SLT), IEEE*, 2014.

[14] S. Meignier and T. Merlin, "Lium spkdiarization: an open source toolkit for diarization," in *CMU SPUD Workshop*, 2010.

[15] Crowdflower. [Online]. Available: http://www.crowdflower.com

[16] O. Zaidan and C. Callison-Burch, "Arabic dialect identification," *Computational Linguistics*, vol. 40.1, pp. 171–202, 2014.

[17] C. A. Ferguson, "Diglossia," *Word-Journal of the International Linguistic Association*, vol. 15, no. 2, pp. 325–340, 1959.