Using Twitter to Collect a Multi-Dialectal Corpus of Arabic

Hamdy Mubarak, Kareem Darwish
Qatar Computing Research Institute
Qatar Foundation
{hmubarak,kdarwish}@qf.org.qa

Abstract

Learning from Rational* Behavior This paper describes the collection and classification of a multi-dialectal corpus of Arabic based on the geographical information of tweets. We mapped information of user locations to one of the Arab countries, and extracted tweets that have dialectal word(s). Manual evaluation of the extracted corpus shows that the accuracy of assignment of tweets to some countries (like Saudi Arabia and Egypt) is above 93% while the accuracy for other countries, such Algeria and Syria is below 70%.

1 Introduction

Arabic is a morphologically complex language (Holes, 2004). With more than 380 million people whose mother tongue is Arabic, it is the fifth most widely spoken language. Modern Standard Arabic (MSA) is the lingua franca amongst Arabic native speakers, and is used in formal communications, such as newspaper, official speeches, and news broadcasts. However, MSA is rarely used in day to day communication. Nearly all the Arabic speakers use dialectal Arabic (DA) in everyday communication (Cotterell et al., 2014). DA may differ from MSA in morphology and phonology (Habash et al., 2012). These dialects may differ also in vocabulary and spelling from MSA and most do not have standard spellings. There is often large lexical overlap between dialects and MSA. Performing proper Arabic dialect identification may positively impact many Natural Language Processing (NLP) application. For example, transcribing dialectal speech or automatically translating into a particular dialect would be aided by the use of targeted language models that are trained on texts in that dialect. This has led to recent interest in the automatic collection large dialectal corpora and the identification of different Arabic dialects (Al-Mannai et al., 2014; Elfardy et al., 2013; Cotterell et al., 2014; Zaidan et al., 2014).

There are many varieties of dialectal Arabic distributed over the 22 countries in the Arabic world. There are often several variants of a dialect within the same country. There is also the difference between Bedouin and Sedentary speech, which runs across all Arabic countries. However, in natural language processing, researchers have merged dialectal Arabic into five regional language groups, namely: Egyptian, Maghrebi, Gulf (Arabian Peninsula), Iraqi, and Levantine (Cotterell et al., 2014; Al-Sabbagh and Girju, 2012).

In this paper, we use geographical information in user Twitter profiles to collect a dialectal corpus for different Arab countries. The contributions of this paper are:

1. We show that we can use Twitter as a source for collecting dialectal corpra for specific Arab countries with reasonable accuracy.

2. We show that most Arabic dialectal words are used in more than one country, and cannot be used separately to collect a dialectal corpus per country.

The paper is organized as follows: Section 2 surveys pervious work on dialect classification; Section 3 describes dialectal Arabic and some of the possible ways to breakdown Arabic dialects; section 4 describes how tweets are collected and classified; section 4 shows how to extract dialectal words and shows that many of them are used in more than one country; Section 5 describes our evaluation approach and reports on evaluation results; and Section 6 contains conclusion and future work.

2 Previous Work

Previous work on Arabic dialect identification uses n-gram based features at both word-level and character-level to identify dialectal sentences (Elfardy et al., 2013; Cotterell et al., 2014; Zaidan et al., 2011; Zaidan et al., 2014). Zaidan et al. (2011) created a dataset of dialectal Arabic. They performed cross-validation experiments for dialect identification using word n-gram based features. Elfardy et al. (2013) built a system to distinguish between Egyptian and MSA. They used word n-gram features combined with core (token-based and perplexity-based features) and meta features for training. Their system showed a 5% improvement over the system of Zaidan and Callison-Burch (2011). Later, Zaidan et al. (2014) used several word n-gram based and character n-gram based features for dialect identification. The system trained on word unigram-based feature performed the best with character five-gram-based feature being second best. A similar result is shown by Cotterell et al. (2014) where word unigram model performs the best. Recent work by Darwish et al. (2014) indicates that using a dialectal word list to identify dialectal Egyptian tweets is better than training on one of the existing dialect corpora.

All of the previous work except Cotterell et al. (2014)[1] evaluated their systems using cross-validation. These models heavily rely on the coverage of training data to achieve better identification. This limits the robustness of identification to genres inline with the training data. In this paper, we exploit geographic information supplied by users to properly identify the dialect of tweets.

There is also increasing interest in the literature to geotag tweets due to its importance for some applications such as event detection, local search, news recommendation and targeted advertising. For example, Mahmud et el. (2012) (Mahmud et al., 2012) presented a new algoritm for inferring home locations of Twitter users by collecting tweets from the top 100 US cities using the geo-tag filter option of Twitter and latitude and longitude for each city using Googles geo-coding API. Bo Han et al. (2014) (Han et al., 2014) presented an integrated geolocation prediction framework and investigated what factors impact on prediction accuracy. They exploited the tweets and profile information of a given user to infer their primary city-level location.

3 Dialectal Arabic (DA)

DA refers to the spoken language used for daily communication in Arab countries. There are considerable geographical distinctions between DAs within countries, across country borders, and even between cities and villages as shown in Figure 1. According to Ethnologue (http://www.ethnologue.com/browse/names), there are 34 variations of spoken Arabic or dialects in Arabic countries in addition to the Modern Standard Arabic (MSA).

Some recent works (Zbib et al., 2012; Cotterell et al., 2014) are based on a coarser classification of Arabic dialects into five groups namely: Egyptian (EGY), Gulf (GLF), Maghrebi (MGR), Levantine (LEV), and Iraqi (IRQ). Other dialects are classified as OTHER.

Zaidan and Callison-Burch (2014) mentioned that this is one possible breakdown but it is relatively coarse and can be further divided into more dialect groups, especially in large regions such as Maghreb. The goal of this paper is to collect a large, clean corpus for each country and study empirically if some of these dialects can be merged together.

We found that there are very few dialectal words that are used in a country and not used in any other country. For example, we took the most frequent Egyptian dialectal words in the Arabic Online Commentary Dataset (AOCD) described in Zaidan and Callison-Burch (2014) according to what they call the dialectness factor, which is akin to mutual information. The AOCD contains comments from newspapers from dialect groups and these comments were classified into different dialects using crowd souring. We examined whether they appear in different dialects or not. As shown in Table 1, most Egyptian dialectal words are being used in different dialects.

With this finding, we realized that unique dialectal words for each country are not common in the sense that they are few and in the sense that relying on them to filter tweets would likely yield a small number of tweets. Thus, we opted not to use such

[1] Zaidan et al. (2014) applied their classifier to a different genre but did not evaluate it's performance.

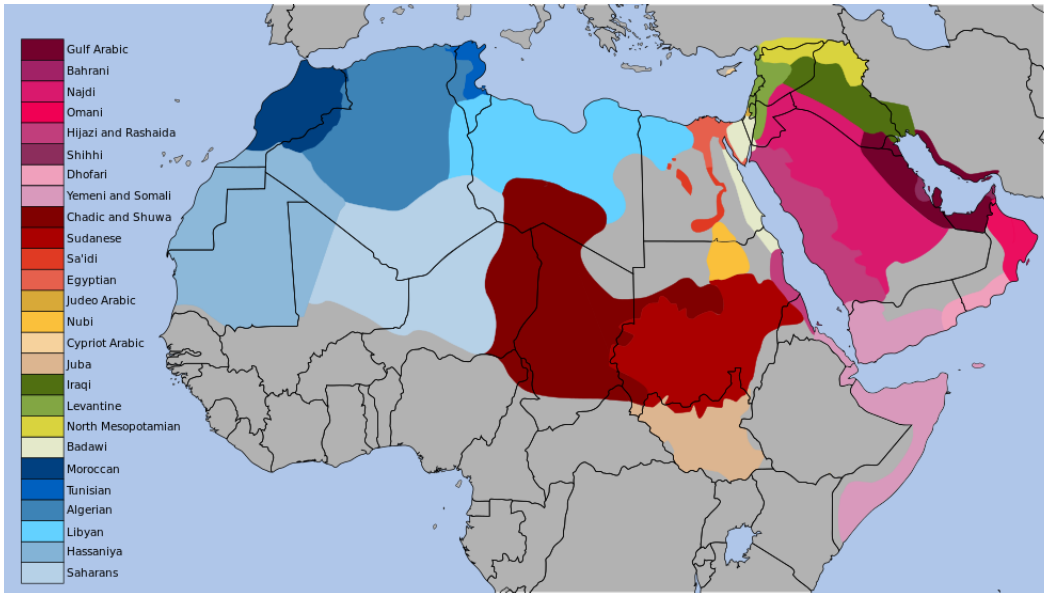

Figure 1: Different Arabic Dialects in the Arab World (`http://en.wikipedia.org/wiki/Arabic_dialects`)

Word	Word in Tweet	Dialect
دي (dy)	الشمس الايامات دي شغالة اوفر تايم عديل كده	Sudan
ده (dh)	كلام كتير داير تقوله لكن بيقيف في طرف لسانك! ده الطبيعي بتاعي	Sudan
عشان (E$An)	انخلقنا عشان نبني لنا مكان في الجنة هذي هي الخلاصة	Gulf
تاني (tAny)	وبعدين ماحكا شي تاني هيك اظن	Levantine

Table 1: Egyptian dialectal Words in other Dialects. We use Buckwalter transliteration in this paper

words to extract tweets for each dialect. From the AOCD, we extracted all unique uni-grams, bigrams, and trigrams, and counted the occurrence of these n-grams from the comments that were marked to belong to different dialects and also in a large MSA corpus composed of 10 years worth of Aljazeera articles, containing 114M tokens [2]. We retained the n-grams that appeared at least 3 or more times in either the dialectal comments. In all, we extracted roughly 45,000 n-grams. The n-grams were manually judged as dialectal or not, and also to which dialect they are most commonly used in. The judgments were performed by one person who is a native Arabic speaker with good exposure to different dialects.

Table 2 lists some words along with their frequencies and to which dialect (or MSA) they belong. Since MSA words compose more than 50% of the words in dialectal text, it is not surprising that words that appear frequently in the corpora of different dialects are indeed MSA. Further, we found that Aljazeera articles contain many dialectal words. Upon further investigation, we found the articles contain transcripts of interviews, where often times the interviewees used dialects, and quotes within articles, where the quoted persons used dialectal utterances. We also found that this was not unique to Aljazeera articles.

When we examined the Arabic GigaWord corpus [3], which is a commonly used MSA corpus, we found that it contains many dialectal words as well. For example, the word كده (kdh) is mentioned 2,574 times and the word علشان (El$An) is mentioned 974 times). This was the main motivating factor for manually judging n-grams as dialectal or not. Of the n-grams we manually tagged, approximately 2,500

[2] `http://aljazeera.net`

[3] `https://catalog.ldc.upenn.edu/LDC2011T11`

Word	EGY	LEV	GLF	IRQ	MGR	MSA	Classification
دي (dy)	541	1	3	0	7	98	EGY
ليه (lyh)	380	23	73	0	22	3734	EGY
ليش (ly$)	28	218	193	18	12	6118	LEV
هيك (hyk)	20	348	9	0	2	4891	LEV
ايش (Ay$)	10	53	87	2	2	87	GLF
يبي (yby)	1	3	99	1	2	21	GLF
شنو ($nw)	0	1	5	5	1	850	IRQ
اكو (Akw)	1	0	1	4	0	0	IRQ
واش (wA$)	2	8	32	5	477	0	MGR
كيما (kymA)	4	3	3	0	246	0	MGR
حاجة (HAjp)	317	8	10	0	120	24468	MSA
صار (SAr)	24	153	79	3	16	12348	MSA

Table 2: Dialectal Words Frequencies in AOCD and MSA (Aljazeera)

were dialectal. We assumed that if a sentence contained one of these n-grams, then the sentence is dialectal. This assumption is consistent with recent published work by Darwish et al. (2014). The distribution of these dialectal n-grams was: 54% unigrams like مش (m$), 39% bigrams like هم دول (hm dwl), and 7% trigrams such as ما أنا عارف (mA >nA EArf). We plan to make the list of dialectal n-grams available to the research community.

Based on interaction with people at Twitter, the estimated number of Arabic microblogs on Twitter is in excess of 15 million per day. The ubiquity of Arabic tweets has been one of the strongest motivations for us to investigate the building of an Arabic dialectal corpus from tweets. Also, tweets are more similar to verbal utterances than formal text, which may be helpful in building language models that are better suited for dialectal Arabic speech recognition.

4 Collecting and Classifying Tweets

4.1 Tweets Collection

We collected 175 M Arabic tweets in March 2014 (5.6M tweets per day) by issuing the query lang:ar against Twitter API [4]. Each tweet has a user ID, and following this ID we can extract the following information from users profile: user name, user time zone, and **user location**. The user location has the user declared geographical location. This could be in the form of a city name, country name, landmark name, country nickname, etc. Such information is available for roughly 70% of tweets. Precise geotagging of tweets, namely latitude and longitude, was available for a very small percentage of tweets. Further, due to the fact that some countries, particularly in the Gulf region, have large expat communities, geo-tagging of tweets only indicate where the tweet is authored but cannot reliably indicate the dialect. By retaining tweets where the user declared a location, we were left with 123M tweets, i.e. 70% of the tweets.

4.2 Tweet Normalization

Tweets and user locations were normalized and cleaned in the manner described in Darwish et al. (2012) by mapping frequent non-Arabic characters and decoration to their mappings, handling repeated characters, etc. Below in an example that shows a tweet before and after normalization:

[4] `http://dev.twitter.com`

Before: مبروووووووك يا باشا mbrwwwwwwwk yA bA$A.
After: مبروك يا باشا mbrwk yA bA$A.
Translation: Congratulations sir.

4.3 User Locations

By looking at user locations, we found that the top unique 10K user locations cover 92M tweets. This is approximately 75% of tweets that have user locations. We used the GeoNames [5] geographical database, which contains eight million place names for each country, to initially assign a user location to one of the Arab countries.

GeoNames has many places without Arabic transliteration, and also users write their locations in Arabic or English, in full or abbreviated forms, and using formal or informal writings. Thus, we manually revised mapping that matched in GeoNames and attempted to map non-matching ones to countries. Examples of such mappings are shown in Table 3.

There were some cases where we could not map a user location to a single Arab country because it is not unique to a particular Arab country or it is not indicative of any country. Such examples include: الوطن العربي الكبير “Great Arab Homeland,” الجزيرة العربية “Arabian Peninsula,” or الشرقية “the Eastern.” In all, approximately 3,500 user locations were mapped to specific countries and the remaining were not. By excluding tweets with non-deterministic user locations, we were left with 62M tweets that have deterministic mappings between user locations and Arab countries. We plan to make the manually reviewed list of user locations publicly available.

4.4 Filtering on Dialectal Words

We used the aforementioned list of dialectal n-grams that we manually extracted to filter the tweets, by retaining those that contain at least one of the n-grams. By doing so, we extracted 6.5M tweets (i.e. 3.7% of the original tweets). Their by-country breakdown is as follows: 3.99M (61%) from Saudi Arabia (SA), 880K (13%) from Egypt (EG), 707K (11%) from Kuwait (KW), 302K (5%) from United Arab Emirates (AE), 65k (2%) from Qatar (QA), and the remaining (8%) from other countries such as Morocco and Sudan. The distribution of tweets per-country is shown in Figure 2.

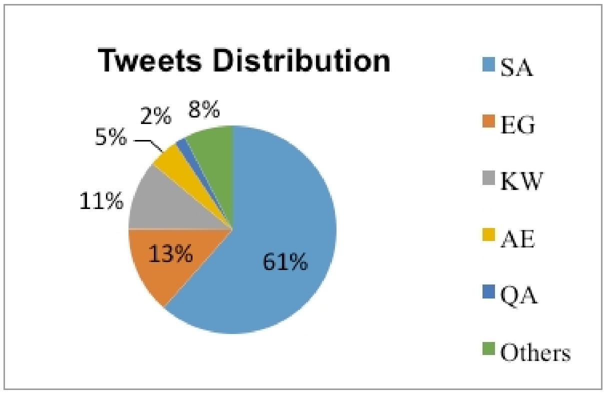

Figure 2: Dialectal Tweets Distribution

5 Evaluation of Dialectal Tweets

To evaluate the accuracy of tweets belonging to the dialect commonly spoken in the different countries that they were assigned to, we randomly extracted 100 tweets per dialect to be manually tagged for dialect.

We used CrowdFlower crowd-sourcing website [6] to evaluate the dialects of tweets. We asked people from the countries associated with each of the associated tweet to judge whether the tweets indeed match the dialect in their country or not. We asked for 3 judgments per tweet. We utilized 20 challenge questions to verify that the judges were doing a good job. We were able to get a sufficient number of judges to finish task for some countries but not all. For example, we were not able to find judges from Qatar and Bahrain. Table 4 lists the accuracy of classification using dialectal words filter and user location.

Errors occurred because some words are mostly used in dialects but less frequently used in MSA

[5] http://www.geonames.org/

[6] http://www.crowdflower.com/

User Location in Profile	Country
الرياض (AlryAD), Riyadh, Saudi Arabia, KSA, الحجاز (AlHjAz)	Saudi Arabia
الكويت (Alkwyt), Q8, kwt, الجهراء AljhrA, كويت العز kwyt AlEz	Kuwait
Egypt, مصر (mSr), Cairo, Alex, أم الدنيا >m AldnyA, جيزة jyzp	Egypt

Table 3: Mapping User Location to Arab Countries

(like تـطـلـع (tTlE)), and the second reason is sometimes a user profile has user location that was mapped to an Arab country, but the user writes tweets using another dialect that is different than one for the stated country.

Examples of tweets that were tagged as Egyptian correctly and incorrectly are shown in table 5.

Dialect	Accuracy
Saudi	95%
Egyptian	94%
Iraqi	82%
Lebanese	75%
Syrian	66%
Algerian	60%

Table 4: Per country classification accuracy

6 Conclusion

Twitter can be used to collect dialectal tweets for each Arab country with high accuracy for some countries and average accuracy for other countries using the geographical information associated with Twitter user profiles. We were able to find dialectal tweets belonging to different dialects with good accuracy by identifying tweets where users used dialectal word n-grams and declared their user locations to belong to particular countries. We tabulated a list of roughly 2,500 dialectal n-grams and 3,500 countries/user locations pairs that we used for identification. We plan to release them publicly. Also, we showed that cross-dialect dialectal words overlap is common, which adds to the complexity of identifying tweets that belong to specific dialects. Thus, using geographical information can greatly enhance dialect identification.

For future work, we plan to analyze the correctness of users' claims on their locations by different methods like tweet geographical information (latitude and longitude), collecting dialectal words for each country, etc. Also, we plan to empirically reexamine the dialect conflation schemes that are commonly used in the literature. Existing schemes for example tend to conflate dialects of all Gulf countries, include Saudi Arabia, Kuwait, Bahrain, Qatar, United Arab Emirates, and Oman. We believe that the dialect spoken in the Western part of Saudi Arabia is sufficiently different from that in Kuwait for example. We would like to study the overlap between dialects spoken in different countries to ascertain dialects of which countries can be safely conflated.

References

Kamla Al-Mannai, Hassan Sajjad, Alaa Khader, Fahad Al Obaidli, Preslav Nakov and Stephan Vogel. 2014. Unsupervised Word Segmentation Improves Dialectal Arabic to English Machine Translation. Arabic Natural Language Processing Workshop, EMNLP-2014.

R. Al-Sabbagh and R. Girju. 2012. YADAC: Yet another Dialectal Arabic Corpus. In LREC. pp. 28822889.

Leo Breiman. 2001. Random Forests. Machine Learning. 45(1):5-32.

Ryan Cotterell and Chris Callison-Burch. 2014. A Multi-Dialect, Multi-Genre Corpus of Informal Written Arabic. LREC-2014, pages 241–245.

Kareem Darwish, Walid Magdy, Ahmed Mourad. 2012. Language Processing for Arabic Microblog Retrieval. CIKM-2012, pages 2427–2430.

Kareem Darwish, Hassan Sajjad, Hamdy Mubarak. 2014. Verifiably Effective Arabic Dialect Identification. EMNLP-2014.

Heba Elfardy, Mona Diab. 2013. Sentence Level Dialect Identification in Arabic. ACL-2013, pages 456–461.

Habash, Nizar, Ramy Eskander, and Abdelati Hawwari. 2012. A morphological analyzer for Egyptian Arabic. Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology, Association for Computational Linguistics, 2012.

Tweet	User Location	Is EGY?
الّي ماعاش حياته ايام المدرسه ده فاته نص عمره	Cairo Egypt	Yes
):احساس صعب اوي لما يكون معاك دقايق مجانيه وموش لاقي حد تكلمه	Masr	Yes
من ادرك ركعة من الصبح قبل ان تطلع الشمس	Alex	No (MSA)
انا عارف فيه ناس كتير عايزة تعرف المعاد بس مكسوفة	Egyptian	Yes
محرك الستة اسطوانات مايقدرش ايدير قوة حصانية زي محرك الثمانية	cairo	No (MGR)

Table 5: Examples of Collected Egyptian Tweets

Han, Bo, Paul Cook, and Timothy Baldwin. 2014. Text-Based Twitter User Geolocation Prediction. Journal Artificial Intelligence Res.(JAIR) 49 (2014): 451-500.

Clive Holes. 2004 Modern Arabic: Structures, functions, and varieties. Georgetown University Press, 2004.

Mahmud, Jalal, Jeffrey Nichols, and Clemens Drews. 2012. Where Is This Tweet From? Inferring Home Locations of Twitter Users. ICWSM. 2012.

Omar F. Zaidan, Chris Callison-Burch. 2011. The Arabic Online Commentary Dataset: An Annotated Dataset of Informal Arabic with High Dialectal Content. ACL-11, pages 37–41.

Omar F. Zaidan, Chris Callison-Burch. 2014. Arabic Dialect Identification. CL-11, 52(1).

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan, Chris Callison-Burch. 2012. Machine translation of Arabic dialects. NAACL-2012, pages 49–59.