Exploring Twitter as a Source of an Arabic Dialect Corpus

Areej Alshutayri *aalshetary@kau.edu.sa*
Faculty of Computing and Information Technology
King Abdul Aziz University
Jeddah, Saudi Arabia
and
School of Computing
University of Leeds
Leeds, LS2 9JT, United Kingdom

Eric Atwell *E.S.Atwell@leeds.ac.uk*
School of Computing
University of Leeds
Leeds, LS2 9JT, United Kingdom

Abstract

Given the lack of Arabic dialect text corpora in comparison with what is available for dialects of English and other languages, there is a need to create dialect text corpora for use in Arabic natural language processing. What is more, there is an increasing use of Arabic dialects in social media, so this text is now considered quite appropriate as a source of a corpus. We collected 210,915K tweets from five groups of Arabic dialects Gulf, Iraqi, Egyptian, Levantine, and North African. This paper explores Twitter as a source and describes the methods that we used to extract tweets and classify them according to the geographic location of the sender. We classified Arabic dialects by using Waikato Environment for Knowledge Analysis (WEKA) data analytic tool which contains many alternative filters and classifiers for machine learning. Our approach in classification tweets achieved an accuracy equal to 79%.

Keywords: Dialectal Arabic, Phonological Variations, Social Media, Multi Dialect, Twitter, Tweet.

1. INTRODUCTION

There are many languages spoken and written of the world's population and each language has different dialects, which are divided mainly by their geographical locations. The Arabic language is one of the world's most widely-spoken languages. It is considered the fifth most-spoken language and one of the oldest languages in the world [1]. Additionally, the Arabic language consists of multiple variants, some formal and some informal [2]. Modern Standard Arabic (MSA) is a formal variant in the Arab world, and it is understood by almost all people in the Arab world. MSA is based on Classical Arabic, which is the language of the Qur'an, the Holy Book of Islam. MSA used in media, newspaper, culture and education; additionally, most of the Automatic Speech Recognition (ASR) and Language Identification (LID) systems are based on MSA. The Dialectal Arabic (DA), in contrast, is an informal variant used in daily life communication, TV shows, songs and movies (ibid). In contrast to MSA, Arabic dialects are less closely related to Classical Arabic. DA is a mix of Classical Arabic and other ancient forms from different neighbouring countries that developed as a result of social interaction between people in Arab countries and people in the neighbouring countries [1].

There are many Arabic dialects that are spoken and written around the Arab world. The main Arabic dialects are: the Gulf Dialect (GLF), the Iraqi Dialect (IRQ), the Levantine Dialect (LEV), Egyptian Dialect (EGY), and the North African (Maghrebi) Dialect (MAG) as shown in Figure 1.

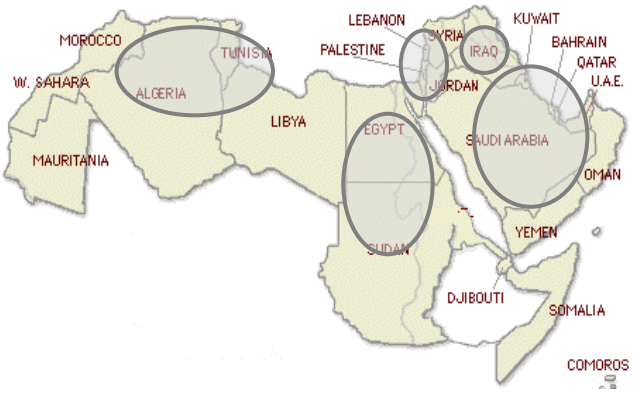

FIGURE 1: Arab World Map.

GLF is spoken in countries around the Arabian Gulf, and includes dialects of Saudi Arabia, Kuwait, Qatar, United Arab Emirates, Bahrain, Oman and Yemen. IRQ is spoken in Iraq, and it is a sub-dialect of GLF. LEV is spoken in countries around the Mediterranean east coast, and covers the dialects of Lebanon, Syria, Jordan and Palestine. EGY includes the dialects of Egypt and Sudan. Finally, MAG includes the dialects of Morocco, Algeria, Tunisia and Libya [1, 2, 3].

Arabic dialects are spread rapidly on the Internet. As a result, there is an essential need to know the dialect used by speakers or writers to communicate with each other; and to identify the dialect before translation takes place, in order to ensure spell checkers work, or to accurately search and retrieve data.

In general, natural language processing for spoken and written English and other languages has been the subject of most studies in the last fifty years [1]. However, Arabic language research has been growing very slowly in comparison to English language research [3]. This slow growth is due to the lack of recent studies on the nature of the acoustic-phonetics of the Arabic language resulting from a lack of a database of Arabic dialects (ibid). In addition, assessing the similarities and differences between dialects of a language is a challenge in natural language processing. Most research in Arabic dialectology focus on phonetic variation based on audio recordings and listening to dialect speakers [3, 1, 5, 4]. Horesh and Cotter (2016) confirmed that past and current research is focussed on phonetic and phonological variation between Arabic dialects: all examples that they presented are of phoneme variation, and they did not mention any work on text, or corpus-based research, or of lexical or morpho-syntactic or grammar variation. Therefore, most Arabic dialectology research collected audio recording to use it in their research [5].

In this paper, we use Twitter to create a dialectal Arabic text corpus by using some seed words. Seed words are distinguished words that are very common in one dialect and not used in any other dialects. In addition to user geographical location information to be sure about the results.

The paper is organized as follows: in section 2 we review related work on using Twitter as a source of Arabic Dialects. In section 3 we describe the major variations between Arabic dialects. In section 4 we present our method on how to extract tweets and dialectal words. In section 5 we show the result of classification process. Finally section 6 contains conclusion.

2. RELATED WORK

Arabic dialect studies have developed rapidly in recent years and most of the previous work has focused on a spoken dialect. Recently people have started using dialect in social media which makes Twitter a source of written Arabic dialect. A related research project created Malay text corpus using Twitter [6]. To collect Malay texts, the researchers define the boundary of the

desired population and select Twitter user IDs for the users who set their location to Malaysia. They did not depend in location only but also checked the language to be sure that they wrote using a non-formal Malay language; therefore, any commercial and political tweets are ignored and they tried to cover different writing style considered the differences in using grammar, lexis, and discourse features. After applying these criteria, researchers found that the sample frame was equal to 321 users who posted their tweets in chat-style Malay language, out of 4,500 users. Then, they used a computer application to extract 3,200 tweets from each user to create a corpus containing one million tweets consists of 14,484,384 words and 646,807 terms.

A multi dialect Arabic speech parallel corpus was built by an Arabic Dialects study [7] which created a speech corpus focused on four main Arabic dialects: MSA, GLF, EGY and LEV; in a domain of travel and tourism. They obtain 67132 speech files, 15492 for MSA, 15492 for GLF, 25820 for EGY and 10328 for LEV by recording the dialectal prompts from 52 speakers with an age range between 16 and 60 years, 49 males and 3 females. They obtained 32 hours of speech with the average length of prompt being 37 minutes. After recording they started to segment prompts into audio files in which each file contained one sentence.

Mubarak and Darwish (2014) used Twitter to collect an Arabic multi-dialect corpus [8]. The researchers classified dialects as Saudi Arabian, Egyptian, Algerian, Iraqi, Lebanese and Syrian. They used a general query which is lang:ar, and issued it against Twitter API to get the tweets which were written in the Arabic language. They collected 175M Arabic tweets, then extracted the user location from each tweet to classify it as a specific dialect according to the location.

Then, Mubarak and Darwish (2014) classified these tweets as dialectal or not dialectal by using the dialectal words from the Arabic Online Commentary Dataset (AOCD) described in [10]. Each dialectal tweet was mapped to a country according to the user location mentioned in the user's profile, with the help of the GeoNames geographical database.

The next step was normalization to delete any non-Arabic characters and also to delete the repetition of characters. Finally, they asked native speakers from the countries identified as tweet locations to confirm whether this tweet used their dialects or not. At the end of this classification, the total tweets number about 6.5M in the following distribution: 3.99M from Saudi Arabia (SA), 880K from Egypt (EG), 707K from Kuwait (KW), 302K from United Arab Emirates (AE), 65k from Qatar (QA), and the remaining 8% from other countries such as Morocco and Sudan. Figure 2 shows the distribution of tweets per-country [8].

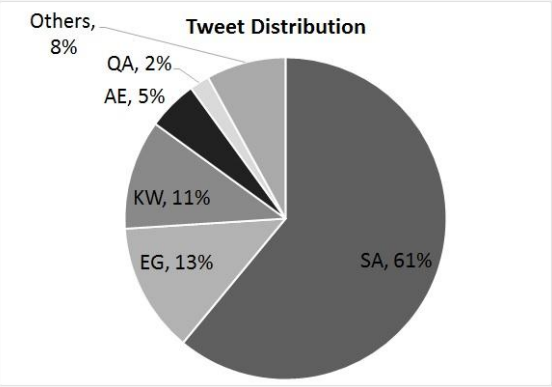

FIGURE 2: Dialectal Tweets Distribution.

Another research team, Ali, Mubarak, and Vogel (2014) used the same corpus that was described in [8] to build a language model for the Egyptian dialect as a basis for a speech recognition system able to distinguish whether the dialect spoken is Egyptian or not and to

recognize the speech accurately [11]. They used 880K tweets written in Egyptian dialect and for speech data they recorded 12.5 hours from Aljazeera Arabic channels (ibid).

In this paper instead of extracting all Arabic tweets like the previous work we tried to extract dialectal tweets by using a filter based on the seed words belonging to each dialect in the Twitter extractor program that we used. In addition, we tried to create a balanced corpus by running the Twitter extractor program for a specific time for each dialect to collect same number of tweets for all dialects.

3. PHONOLOGICAL VARIATIONS BETWEEN ARABIC DIALECTS

Arabic dialects differ phonologically from MSA and each other [2, 12]. They suggested that these variations between Arabic dialects help users distinguish and recognize one dialect from another. The following summary presents some common variations in the pronunciation of some Arabic consonants.

The MSA consonant Qaaf (/q/) is pronounced as a glottal stop (/ʔ/) in EGY and LEV, as /g/ in GLF, and IRQ [2, 3, 12]. For instance, the word “road” in MSA is pronounced as Tariq, in EGY and LEV is pronounced as Tariʔ, and in GLF and IRQ is pronounced as Tarig. Also sometimes Qaaf (/q/) is pronounced as (/k/) in IRQ; for instance, the word “time” in MSA is pronounced as Waqt, while in IRQ it is pronounced as Wkt. Another variation is in consonant Jiim (/dʒ/) which pronounced as (/g/) in EGY and LEV and as /y/ in GLF, for example, the word “beautiful” is pronounced as dʒamil in MSA, IRQ and MAG, while in EGY it is pronounced as gamil and in GLF as yamil, which means tend to (ibid). Moreover, the consonant Thaa (/θ/) in MSA is pronounced as (/t/) or (/s/) in EGY, LEV and MAG. For example, the word “three” is pronounced θalaθa in MSA, GLF, and IRQ whereas in EGY and LEV is pronounced talata. Another example, the word “then” is pronounced as θuma in MSA and GLF; however, in EGY and LEV, it is pronounced as suma.

A final difference is in consonant Dhaa (/ðˤ/), which is pronounced as (/z/) in EGY, LEV, and MAG. The word “appear” is pronounced as ðˤhar in MSA, GLF, and IRQ while in EGY and LEV it is pronounced as zhar, which means flower. Table 1 summarizes the major regional variations in the pronunciation of alphabetic characters in Arabic.

MSA		GLF	EGY	MAG	LEV	IRQ
ق	q	g	ʔ	g	ʔ	k
ج	dʒ	ʤ (or) y	g	ʤ	ʤ	ʤ
ث	θ	θ	s (or) t	t	s (or) t	θ
ذ	ð	ð (or) d	z (or) d	ð	z	ð
ظ	ðˤ	ðˤ	z	ðˤ	z	ðˤ

TABLE 1: Regional Variations in Arabic Phonetics.

4. COLLECTING TWEETS

This section is about how to collect tweets and label them by the name of the dialect that they represent. In our experiment, we tried to collect dialectal tweets for country groups (5 groups) which are GLF, IRQ, LEV, EGY, and MAG. We created an app which connects with the Twitter API1 to access the Twitter data programmatically.

Our plan for collecting tweets depends on identifying seed words for every dialect. Seed words are distinguished words that are used very common and used very frequently in one dialect and not used in any other dialects. One source for a dialectal word is an AOCD, but we do not have access to this dataset; instead, we have chosen some seed words from Zaidan and Callison-

[1] http://apps.twitter.com

Burch's (2014) paper that described this dataset. The authors collected words for all dialects from readers' comments on the online websites of three Arabic newspapers: Al-Ghad from Jordan to cover the Levantine dialect, Al-Riyadh from Saudi Arabia to cover the Gulf dialect, and Al-Youm Al-Sabe from Egypt to cover the Egyptian dialect (ibid). In addition, we used some seed words from Almeman and Lee (2013) paper [13]. The researchers collected 1,500 words and phrases by exploring the web and extracting dialects' words and phrases, which must be found in one dialect of the four main dialects which are GLF, LEV, EGY, and MAG. We did not find a corpus for the Iraqi dialect, but we extracted some IRQ seed words from [14]. All dialect seed words we have chosen seem to be popular and frequently used in this dialect and usually we hear them from native speakers for each dialect or on TV programs or movies.

We tried to use words that could be found in one dialect not in other dialects, such as the word مصاري (Msary), which means "Money" and is used only in LEV dialect; we also used the word دلوقتي (Dlwqty), which means "now" and is used only in EGY dialect, while in GLF speakers used the word الحين (Alhyn). In IRQ, speakers change Qaaf (/q/) to (/k/) so they say وكت (wkt), which means "time". Finally, for MAG, which is the dialect most affected by French colonialism and neighbouring countries, speakers used the words بزاف (Bzaf) and برشا (Brʃā), which mean "much". Table 2 shows examples of the seed words that we used in our experiment.

GLF	IRQ	LEV	EGY	MAG
Lbyh لبيه	ʃtryd شتريد	Mnyh منيح	ʕayz عايز	Dyalk ديالك
ʃlwn شلون	bawʕ باوع	Xtyār ختيار	Bs بص	ʕlāʃ علاش
Amhq أمحق	ʕlmwd علمود	Zlmh زلمه	mfyʃ مفيش	qʕmz قعمز

TABLE 2: Examples of some seed words for each dialect

We collected Arabic dialect tweets by using the query lang:ar which extracts all tweets written in the Arabic language, and we tracked 35 seed words all unigram in each dialect. Each tweet has a user name and user location. In addition to the tracking of seed words, we used the user location to show the geographical location of the tweets, to be sure that tweets belong to this dialect. The user location sometimes was not available in some tweets, and sometimes could be a sport club name, street name or landmark name. However, in general, it might be a country or city name. To verify these tweets and assert that they belong to this dialect we have chosen 500 tweets and asked some native speakers of each dialect to confirm if this tweet belongs to their dialect or not. By running the Twitter extractor for 144 hours, we collected 210,915K tweets with the total number of words equal to 3,627,733 words; these included 44,894K tweets from GLF during 9 hours, 39,582K from EGY during 10 hours, 45,149K from IRQ during 29 hours, 40,248K from LEV during 52 hours, and 41,042K from MAG during 44 hours. Figure 3 shows the distribution of tweets per dialect and Table 3 shows the number of words that were extracted for each dialect.

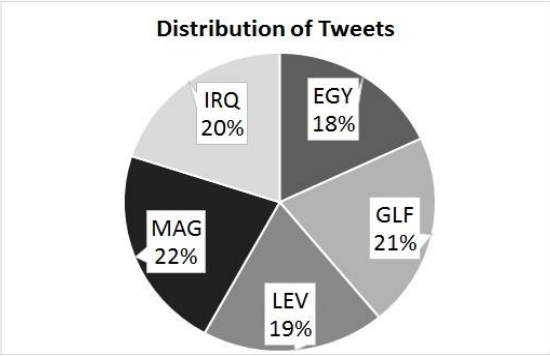

FIGURE 3: Dialect Tweet Distribution

Dialect	Number of words
GLF	658,893
EGY	558,236
IRQ	905,072
LEV	628,184
MAG	877,348

TABLE 3: Number of words extracted for each dialect

5. RESEARCH EXPERIMENTS AND RESULTS

In this section, we describe how we classified the samples of our five major Arabic dialects collected from Twitter using WEKA toolkit [15], a widely used tool for data mining that provides a lot of machine learning algorithms.

To classify dialects the data set divided into two sets: the first set contains 8,090 labelled tweets used for training and divided unequally between the Arabic dialects: 2,152K from GLF, 1,541K from EGY, 1,585K from MAG, 1,533K from LEV, and 1,279K from IRQ. The second set is for testing contains 1,764 labelled tweets: 450 from GLF, 326 from EGY, 377 from MAG, 286 from LEV, and 223 from IRQ. For testing set, we have tried to collect new tweets depending on locations only, without using any seeds words, then we have manually classified these tweets into appropriate dialect.

We achieved 79% accuracy by using Multinomial Naive Bayes (MNB) algorithm with WordTokenizer feature to extract words between spaces or any other delimiters such as (full-stop, comma, semi colon, colon, parenthesis, question, quotation and exclamation mark).

6. CONCLUSION AND FUTURE WORK

Most of Arabic corpora are audio recording, so in this paper we explored Twitter as a source of Arabic dialect texts to create written corpus of Arabic dialects which is more directly useful for natural language processing research. Our dialect text corpus is more useful for building classifier to classify dialects than the corpus produced from [8] because we collected a balanced corpus.

We have achieved a large corpus of written Arabic dialects texts by divided the Arab countries into five groups, one for each of the five main dialects: Gulf, Iraqi, Egyptian, Levantine and North African.

To distinguish between one dialect and another we used some seed words that are spoken in one dialect and not in the other dialects. In addition, we extracted the user location to help us to enhance dialect classification and specify the country and dialect to which each tweet belongs.

In general, Twitter can be used as a reference to collect an Arabic dialect text corpus but to make our corpus balanced we had to run the tweet extractor in one dialect longer than another as we notice that a lot of tweets come from Saudi Arabia whereas we had fewer tweets from North African countries and Iraq.

To classify Arabic dialects we used WEKA and created two set of data: one as a training set and another as a testing set. We achieved the accuracy up to 79%. We think that we might combine WordTokenizer and CharacterNGram in the future to improve the results using an ensemble method. As a future work, we can explore other sources of informal Arabic dialect language such as Facebook and the comments in online newspaper, in addition to using speech recognition on spoken Arabic dialects to extend our text corpus. We can compare Arabic dialect texts against other variants of Arabic, such as Classical Arabic of the Quran [16, 17]. Furthermore, we could try using word embedding classifiers to compare the result with WEKA classifiers as well as other task such as check the similarity of Arabic sentences [9].

7. ACKNOWLEDGMENT

Areej Alshutayri is grateful for funding from King AbdulAziz University, Jeddah, Saudi Arabia. Eric Atwell is grateful for EPSRC grant EP/K015206/1 Natural Language Processing working together with Arabic and Islamic Studies.

8. REFERENCES

[1] F. Biadsy, J. Hirschberg, N. Habash. (2009). “Spoken Arabic dialect identification using phonotactic modeling”. *In: Proceedings of the EACL workshop on computational approaches to Semitic languages*, pp. 53-61, 31 March, Athens, Greece. ACL, Stroudsburg, PA, USA.

[2] N. Habash. (2010). “Introduction to Arabic natural language processing”. Morgan & Claypool Publishers, Synthesis Lectures on Human Language Technology. 10, ebook isbn 978-1-59829-796-6.

[3] F. Alorifi. (2008). “Automatic identification of Arabic dialects using Hidden Markov Models”. PhD thesis, University of Pittsburgh, Department of Electrical Engineering and Computer Science.

[4] F. Sadat, F. Kazemi, and A. Farzindar. (2014). “Automatic identification of arabic language varieties and dialects in social media”. In Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP), pages 22–27.

[5] U. Horesh and W. M. Cotter. (2016). “Current research on linguistic variation in the arabic-speaking world”. Language and Linguistics Compass, 10(8):370–381.

[6] M. Saloot, N. Idris, A. Aw, and D. Thorleuchter. (2016). “Twitter corpus creation: The case of a Malay Chat-style-text Corpus (MCC)”. Digital Scholarship in the Humanities, 31(2), pp.227-243.

[7] K. Almeman, M. Lee, and A. Almiman. (2013). “Multi Dialect Arabic Speech Parallel Corpora”. In: Communications, Signal Processing, and their Applications (ICCSPA), 1st International Conference, Sharjah, UAE. IEEE.

[8] H. Mubarak, K. Darwish. (2014). “Using Twitter to collect a multi-dialectal corpus of Arabic”. In: Proceedings of the EMNLP workshop on natural language processing. Doha, Qatar, 25 October, 2014, pp. 1-7.

[9] E. Nagoudi, and D. Schwab. (2017). “Semantic Similarity of Arabic Sentences with Word Embeddings”. Association for Computational Linguistics. pp.18-24. [workshop publication]. Available from: http://aclweb.org/anthology/W17-1303.

[10] O. Zaidan, C. Callison-Burch. (2014). “Arabic dialect identification”. In: Computational Linguistics. 40(1): pp. 171-202.

[11] A. Ali, H. Mubarak, and S. Vogel. (2014). “Advances in Dialectal Arabic speech recognition”. In: Proceedings of the of the international workshop on spoken language translation (IWSLT) Dec 4-5, Lake Tahoe CA, USA. pp.156-162.

[12] M. Elmahdy, R. Gruhn, W. Minker, S. Abdennadher. (2009). “Cross-lingual acoustic modeling for Dialectal Arabic speech recognition”. In: ACM SIGKDD Explorations Newsletter 11(1):101-118, November 2009.

[13] K. Almeman, M. Lee. (2013). “Automatic building of Arabic multi-dialect text corpora by bootstrapping dialect words”. In: The Proceedings of the 1st International Conference on

Communications, Signal Processing, and their Applications (ICCSPA'13), Sharjah, UAE, 12-14 Feb., IEEE.

[14] M. Khoshaba. (2006). “Iraqi dialect vs. Standard Arabic”, Medium Corporation, San Jose, CA, USA.

[15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, H. Witten. (2009). “The WEKA Data Mining Software: An update”. In ACM SIGKDD Explorations Newsletter, 11(1): pp. 10-18, November 2009.

[16] M. Alrabiah, N. Alhelewh, A. Al-Salman, E. Atwell. (2014). “An Empirical Study On The Holy Quran Based On A Large Classical Arabic Corpus”. International Journal of Computational Linguistics 5(1):pp.1-13.

[17] M. Alrabiah, A. Al-Salman, E. Atwell, N. Alhelewh. (2014). “KSUCCA: A Key To Exploring Arabic Historical Linguistics”. International Journal of Computational Linguistics 5(2):pp.27-36.