
Proceedings Paper:
Alshutayri, A and Atwell, E (2018) Arabic dialects annotation using an online game. In: ICNLSP 2018: 2nd International Conference on Natural Language and Speech Processing. 2nd International Conference on Natural Language and Speech Processing (ICNLSP 2018), 25-26 Apr 2018, Algiers, Algeria. IEEE. ISBN: 978-1-5386-4543-7.

https://doi.org/10.1109/ICNLSP.2018.8374371

Arabic Dialects Annotation using an Online Game

1st Areej Alshutayri
Faculty of Computing and Information Technology
King Abdul Aziz University
Jeddah, Saudi Arabia
aalshetary@kau.edu.sa
School of Computing
University of Leeds
Leeds, United Kingdom
ml14aooa@leeds.ac.uk

2nd Eric Atwell
School of Computing
University of Leeds
Leeds, United Kingdom
E.S.Atwell@leeds.ac.uk

***Abstract*—Modern Standard Arabic is the written standard across the Arab world; but there is an increasing use of Arabic dialects in social media, so this is appropriate as a source of a corpus for research on classifying Arabic dialect texts using machine learning algorithms. An important first step is annotation of the text corpus with correct dialect tags. We collected tweets from Twitter and comments from Facebook and online newspapers, aiming for representative samples of five groups of Arabic dialects: Gulf, Iraqi, Egyptian, Levantine, and North African. Then, we explored an approach to crowdsourcing corpus annotation. The task of annotation was developed as an online game, where players can test their dialect classification skills and get a score of their knowledge. This approach has so far achieved 24K annotated documents containing 587K tokens; 16,179 tagged as a dialect and 7,821 as Modern Standard Arabic.**

***Index Terms*—Arabic, Dialects, Corpus, Annotation, Crowdsourcing**

I. Introduction

Modern Standard Arabic (MSA) is the formal written standard across the Arab world; but there is an increasing use of Arabic dialect in a range of informal text sources. The classification of dialects becomes an important pre-process for other tasks, such as machine translation, dialect-to-dialect lexicons, and information retrieval [1]. To improve the classification of Arabic dialect, we developed a new approach to annotate Arabic dialect texts. We used two sources of social media: tweets from Twitter [2], and comments from Facebook, in addition to readers' comments from online Newspaper as a web source. The corpus contains dialectal Arabic texts collected from Arab's countries to cover the main Arabic dialects which are: The Gulf Dialect (GLF), the Iraqi Dialect (IRQ), the Levantine Dialect (LEV), the Egyptian Dialect (EGY), and the North African (Maghrebi) Dialect (NOR).

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, NOR includes the dialects of Morocco, Algeria, Tunisia and Libya [3]–[5], as in fig. 1.

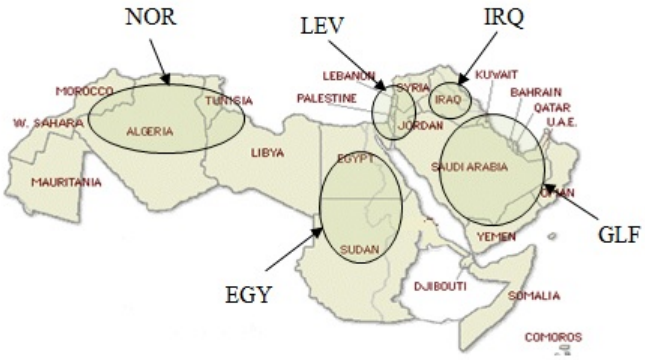

Fig. 1. Arab World Map.

Some tweets were collected based on location points and some tweets based on seed terms which are distinguished words that are very common in one dialect and not used in any other dialects [2], the total number of tweets are 280K, beside 2M comments from Facebook. In addition to 10K comments by crawling the newspaper websites for a period of two months. Table 1 shows the total number of words for each source of text.

TABLE I
THE TOTAL NUMBER OF WORDS FROM EACH SOURCE OF TEXT

Source	Number of Words
Twitter	6,827,733
Facebook	7,056,812
Newspaper	3,318,717

In [6] the method of the annotation was used through the workers on Amazons Mechanical Turk. They showed 10 sentences per screen. The worker was asked to label each sentence with two labels: the amount of dialect in the sentence, and the type of the dialect. They collected 330K labelled

documents in about 4.5 months. But, compared to our method they pay to the workers a reward of $0.10 per screen. The total cost of annotation process was $2,773.20 in addition to $277.32 for Amazons commission.

In this paper, the second section presents why annotation process is important. The third section describes the method used to annotate the collected dataset to build a corpus of Arabic dialect texts. The fourth section shows how we evaluate the annotated results. The fifth section presents the result and the number of annotated documents. Finally, the last section presents the conclusion and future work.

II. Importance of the Annotation Process

We participated in the VarDial2016 workshop at COLING 2016 Discriminating Similar Languages (DSL) shared task [7]. The shared task offered two tasks: first task worked on identification of very similar languages in newswire texts. The second task focused on Arabic dialect identification in speech transcripts [8]. The Arabic dialect text used for training and testing were developed using the QCRI Automatic Speech Recognition (ASR) QATS system [9] to label each document with a dialect [10]. Some evidently mislabeled documents were found which affected the accuracy of classification; so, to avoid this problem a new text corpus and labelling method were created.

In the first step of labelling the corpus, we initially assumed each tweet could be labelled based on the location that appears in the user's profile and the location points which could be used to collect the tweets from Twitter. As for the comments were collected from online newspapers, each comment labelled based on the country where the newspaper is published. Finally, for the comments collected from Facebook posts, each comment labelled based on the country of the Facebook page depended on the nationality of the owner of the Facebook page if it is a famous public group or person. However, through the inspection of the corpus, we noticed some mislabeled documents, due to disagreement between the locations of the users and their dialects. So, must be verified that each document is labelled with the correct dialect. Fig. 2 and 3 give an example of the confusion between the user location and their dialect.

The user location in fig. 2 is England while the tweets are written using Arabic language not English language. Similarly, for Facebook comments, the Facebook page's country based on the nationality of the page owner is Saudi Arabia, but the comments were not written in GLF dialect, such as the shaded comment in the fig. 3.

III. Method

To annotate each document with the correct dialect, 100K documents were randomly selected from the corpus (tweets and comments), then created an annotation tool and hosted this tool in a website.

In the developed annotation tool, the player annotates 15 documents (tweets and comments) per screen. Each of these

Fig. 2. Example of user location and his tweets.

Fig. 3. An example of the Facebook page's country and the user's comment.

documents is labelled with four labels, so the player must read the document and make four judgments about this document. The first judgment is the level of dialectal content in the document. The second judgment is the type of dialect if the document not MSA. The third judgment is the reason which makes the player to select this dialect. Finally, the forth judgment if the reason selected in the third judgment is dialectal terms; then in the fourth judgment the player needs to write the dialectal words found in the document.

The following list shows the options under each judgment to let the player choose one of them.

- The level of dialectal content
  - MSA (for document written in MSA)
  - Little bit of dialect (for document written in MSA but it contains some words of dialect)
  - Mix of MSA and dialect (for document written in MSA and dialect (code switch))
  - Dialect (for document written in dialect)
- The type of dialect if the document written in dialect
  - Egyptian
  - Gulf
  - Iraqi
  - Levantine
  - North Africa

  - Not Sure
- The reason that make this document dialectal
  - Sentence Structure
  - Dialectal Terms
- The words which identify the dialect (we need to use these word as a dictionary for each dialect)

To annotate the collected data, an interface was built as a web page to display a group of Arabic documents randomly selected from the dataset. Fig. 4 shows the interface of the Annotation Tool in the website http://www.alshutayri.com/index.jsp.

Each page displays 15 documents randomly selected from the dataset. As shown in fig. 5, the first label indicates the amount of dialectal content in the document to decide whether the document is MSA or contains dialectal content. If the document is MSA the other labels will be inactive, and the player needs to move to the next document. But, if the document is not MSA, then all labels are required. The second label specifies the document dialect if it is one of the five dialects (EGY, GLF, LEV, IRQ, and NOR), or Not Sure if the document written using dialect but difficult to decide which dialect. The third and fourth labels to explain the causes to choose the selected dialect: for example, the sentence structure if the words in the document are all MSA words, but the structure of the sentence is not based on the MSA grammar rules, and/or the dialectal terms which are famous words help to identify the dialect. In fact, there is no agreed standard for writing Arabic dialects because MSA is the formal standard form of written Arabic [11]; therefore, some documents apparently contain only MSA vocabulary but are annotated as dialect based on non-standard sentence structure.

Before submitting the annotated documents, the mother dialect must be chosen. This may help to decide which annotated document must be accepted if one document has different annotations. Fig. 5 shows an example of one annotated document. Finally, by submitting the annotated documents the score will be shown in the screen by comparing the labelled documents with our pre-labelled sample as shown in fig. 6.

As a control to be sure that the player reads the document before selecting the options, three MSA documents collected from a newspaper articles [12], were mixed with 12 documents selected from the dataset; so these three MSA documents used as a control because they must be labelled as MSA, so, if the player labels all the three MSA documents as a dialect then the player's submitted documents are not counted in the annotated corpus. Furthermore, to verify the annotation process, each document is redundantly being annotated three times.

Fig. 4. The annotation interface.

Fig. 5. Example of the annotated document.

Fig. 6. Example of the player's score.

IV. Evaluation

To ensure that each document got a correct label, every document was annotated by three players besides the gold standard, which is an initial label that have been used to label each document based on the source of comments and tweets as mentioned in section 2. In addition to the mother dialect for each player which help to decide which label must be counted as a correct label if the players gave different labels for one document. The result of annotated documents was evaluated in two cases:

- Agreement between annotators: All the players label one document with same label as in fig. 7 and 8. The agreed label considered as a correct label even if the agreed label is different from the original label because as mentioned in section 2 the initial label sometimes is not correct.
- Disagreement between annotators: Some of the players label the document with different label of the other players as in fig. 9. In this case the mother dialect could

help to decide which label must be accepted as a correct label for this document.

Text	Original Dialect	Dialect level	Dialect	Mother Dialect
احذر بدء يومك بتناول الموز	NOR	MSA	MSA	GLF
احذر بدء يومك بتناول الموز	NOR	MAS	MSA	GLF
احذر بدء يومك بتناول الموز	NOR	MSA	MSA	LEV

Fig. 7. Example 1 of the agreement between annotator.

Text	Original Dialect	Dialect level	Dialect	Mother Dialect
احنا مبنضيعش وقت	EGY	Dialect	EGY	GLF
احنا مبنضيعش وقت	EGY	Dialect	EGY	GLF
احنا مبنضيعش وقت	EGY	Dialect	EGY	LEV

Fig. 8. Example 2 of the agreement between annotator.

Text	Original Dialect	Dialect level	Dialect	Mother Dialect
افاااا عليك معكم ان شاء الله بالدعاء والفعل بس ها إن جاكم شئ لاتنسوا أهل الطائف تحياتي للجميع	GLF	MSA	MSA	NOR
افاااا عليك معكم ان شاء الله بالدعاء والفعل بس ها إن جاكم شئ لاتنسوا أهل الطائف تحياتي للجميع	GLF	Dialect	GLF	GLF
افاااا عليك معكم ان شاء الله بالدعاء والفعل بس ها إن جاكم شئ لاتنسوا أهل الطائف تحياتي للجميع	GLF	MSA	MSA	EGY

Fig. 9. Example of the disagreement between annotator.

To evaluate the quality of the annotation, the inter-annotator agreement was calculated using Fleiss Kappa [13] to calculate the annotator agreement for more than two annotators. The result equal to 0.787 around 79% which is substantial agreement according to [14].

V. Result

The result of the annotation tool is a set of documents which are labelled with four labels: the first label is the dialect level, which is an option from three choices: little_of_dialect, Mix_of_MSA_and_dialect, or Dialect. The second label is the specific dialect which is one of the five dialects: GLF, EGY, LEV, IRQ, or NOR. The third label shows the reasons that help to identify the document's dialect. The last label shows the dialectal words which help to identify the document's dialect. Fig. 10 shows the result of one annotated document in the corpus.

We launched the website via Twitter and WhatsApp at the beginning of August 2017. At the time of paper submission, we have been running the annotation website for around four months, and we have accumulated 24,000 annotated documents with total numbers of words equal to 586,952. The distribution of dialects of the annotated corpus shown in fig. 11, where GLF dialect consist of 5K documents, EGY dialect 4K documents, NOR dialect 2K documents, LEV dialect 3K, and IRQ dialect 2K documents. The number of users (players) equal to 1,575 from different countries around the world, fig. 12 shows the distributions of users on the days. For our immediate research on Arabic dialects classification the annotated documents which we have already collected could be sufficient, but we decided to continue with this experiment to collect a large annotated Arabic dialect text corpus and let the corpus be available for other research by the end of 2018.

Fig. 10. Result of the annotated document.

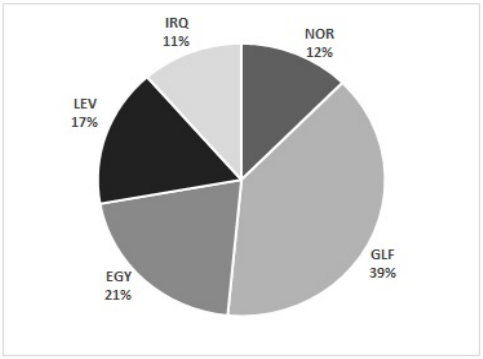

Fig. 11. The distribution of labels (dialects) of the annotated corpus.

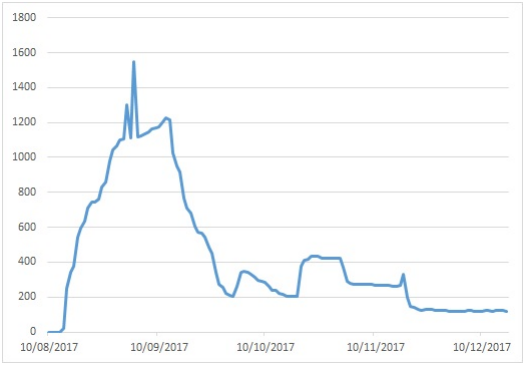

Fig. 12. The distribution of the number of users during days.

VI. Conclusion and Future Work

In this paper, we presented a new approach to annotate the dataset were collected from Twitter, Facebook, and Online Newspaper for the five main Arabic dialects: Gulf, Iraqi, Egyptian, Levantine and North African. The annotation website was created as an online game to gather more users who talk different Arabic dialects and free to pay in comparing with other crowdsourcing websites. This experiment is a new approach help to annotate the sufficient dataset for text researches in Arabic dialect classification. The number of users

has decreased now in comparison with the beginning because we need to redistribute the website widely. In future work, we could modify the interface to be more attractive and easy to explore. In addition, we could make this annotation game as an application can be downloaded in the smart phones and tablets.

REFERENCES

[1] S. Malmasi, E. Refaee, M. Dras, "Arabic dialect identification using a parallel multidialectal corpus," Pacific Association for Computational Linguistics, 2015, pp. 203–211.

[2] A. Alshutayri, E. Atwell, "Exploring Twitter as a source of an Arabic dialect corpus," International Journal of Computational Linguistics (IJCL), vol. 8, 2017, pp. 37–44.

[3] F. S. Alorifi, "Automatic identification of Arabic dialects using hidden markov models," Doctor of Philosophy thesis, University of Pittsburgh, 2008.

[4] F. Biadsy, J. Hirschberg, N. Habash, "Spoken Arabic dialect identification using phonotactic modeling," In proceedings of the EACL Workshop on Computational Approaches to Semitic Languages, Association for Computational Linguistics, Athens 2009, pp. 53–61.

[5] N. Habash, Introduction to Arabic Natural Language Processing. Morgan and Claypool. 2010.

[6] F.O. Zaidan, C. Callison-Burch, "Arabic dialect identification," Computational Linguistics. vol. 40, 2014, pp. 171–202.

[7] A. Alshutayri, E. Atwell, A. AlOsaimy, J. Dickins, M. Ingleby, J. Watson, "Arabic language WEKA-Based dialect classifier for Arabic automatic speech recognition transcripts," Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects, 2016, pp. 204–211.

[8] S. Malmasi, M. Zampieri, N. Ljubešić, P. Nakov, A. Ali, J. Tiedemann, "Discriminating between similar languages and Arabic dialect identification: a report on the third DSL shared task," Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3), Osaka, Japan 2016, pp. 1–14.

[9] S. Khurana, A. Ali, "QCRI advanced transcription system (QATS) for the Arabic multi-dialect broadcast media recognition: MGB-2 challenge," IEEE Spoken Language Technology Workshop (SLT), 2016, pp. 292–298.

[10] A. Ali, N. Dehak, P. Cardinal, S. Khurana, S. Yella, J. Glass, P. Bell, S. Renals, "Automatic dialect detection in Arabic broadcast speech," Interspeech2016, 2016, pp. 2934–2938.

[11] H. Elfardy, M. Diab, "Token level identification of linguistic code switching," In Proceedings of COLING, 2016, pp. 287–296.

[12] L. Al-Sulaiti, E. Atwell, "Designing and developing a corpus of contemporary Arabic," Proceedings of TALC 2004: the sixth Teaching and Language Corpora conference, Granada 2004, pp. 92-93.

[13] J. L. Fleiss,"Measuring nominal scale agreement among many raters," Psychological Bulletin, vol. (76)5, 1971, pp. 378–382

[14] J. Richard Landis and Gary G. Koch, "The Measurement of Observer Agreement for Categorical Data," Biometrics, Wiley, International Biometric Society, vol. 33(1), 1977, pp. 159–174.