Arabic Dialects Identification for All Arabic countries

Ahmed Hussein Aliwy
Faculty of CS and math
University of Kufa
Iraq

Hawraa Ali Taher
Faculty of education for girls
University of Kufa
Iraq

Zena A. Abutiheen
Dep. of CS
University of Kerbala
Iraq

{ahmed.almajidy,hawraaa.alshimirty}@uokufa.edu.iq, z.aboaltaheen@uokerbala.edu.iq

Abstract

Arabic dialects are among of three main variant of Arabic language (Classical Arabic, modern standard Arabic and dialectal Arabic). It has many variants according to the country, city (provinces) or town. In this paper, several techniques with multiple algorithms are applied for Arabic dialects identification starting from removing noise till classification task using all Arabic countries as 21 classes. Three types of classifiers (Naïve Bayes, Logistic Regression, and Decision Tree) are combined using voting with two different methodologies. Also clustering technique is used for decreasing the noise that result from the existing of MSA tweets in the data set for training phase. The results of f-measure were 27.17, 41.34 and 52.38 for first methodology without clustering, second methodology without clustering, and second methodology with clustering, the used data set is NADI shared task data set.

1 Introduction

Arabic Dialects is one of three variations of Arabic language (Classical Arabic CA, Modern Standard Arabic MSA and dialectal Arabic DA). It is a native language of Arabic people used in the communication among them and in the social media (Itani, 2018). Each country of 21 countries, in the Arab world, has its own dialect, sometimes they are written in same script with different pronunciation. Recently, dialect identification (DI) is interesting field in Natural Language Processing (NLP) and conducted by number of researchers because it is increasing rapidly on the web and social media.

There are nine distinct dialectal categories in Arab world: Egyptian, Gulf, Iraqi, Levantine, Maghrebi (El-Haj et al., 2018), Yemeni, Somali, Sudanese and Mauritania. Each one of them has many varieties according to the city and town.

It is clear that there are four levels of Arabic dialectal identification (ADI): (1) identification of dialectal Arabic from MSA and CA, it is very easy task similar to identification of Arabic language among other languages, (2) identification of the main category of dialectal Arabic of(nine or five categories), it is more difficult than previous point, (3) identification of country level out of 21 countries, it more difficult than the previous two points, (4) identification of city level or town level which is subfield of country level, it is the most difficult among all these levels. Table 1 show dialects of the word "أريكة- Ârykħ"-(sofa) [1] in some Arabic countries*.

Country	Native word	Country	Native word	Country	Native word
Iraq	كرويت- krwyt قنفة- qnfħ	Kuwait	كنبه- knbh قنفه- qnfħ	Saudi Arabia	باطرمه- bATrmh كنبه- knbh
Bahrain	قنفة- qnfħ	Lebanon	كنبايه- knbAyh	Syria	كنبايه- knbAyh
Egypt	كنبه- knbh	Oman	قنفة- qnfħ	Morocco & Tunisia	فوطوي- fwTwy
Algeria	فوطوي- fwTwy	Palestine	كنبايه- knbAyh	Yemen	كنبه- knbh
Jordan	كنبايه- knbAyh	Qatar	كنبه- knbh	United Arab Emirates	انتريه- Antryh, قنفة- qnfħ كنبه- knbh

Table 1: dialects of the word "أريكة- Ârykħ" (sofa) in 16 Arabic countries.

[1] We used Habash-Soudi-Buckwalter transliteration in the form: "Arabic word-its transliteration"- (its translation)

Proceedings of the Fifth Arabic Natural Language Processing Workshop, pages 302–307
Barcelona, Spain (Online), December 12, 2020

There are many challenges in ADI over identification of other types of Arabic language such as:

1. In case of the levels 2,3 and 4, all the countries and cities use MSA for writing in some times, therefore these will be noise in the dialectal identification.
2. Existing of a city in a specific country use dialect that very close to other country more than its country such as Basrah in Iraq use dialect very close to Kuwait dialect.
3. Some countries are much similar in most words and different in little words therefore identification of their dialects are much difficult.

this paper is a part of NADI 2020 shared task (subtask1) where the tweets in Arabic dialect has been classified into the country belong it by voting among three classifiers (Naïve Bayes NB, Logistic Regression LR, and Decision Tree DT) in different methodologies to make final decision. Also a preprocessing phase is done, before implementation of the classifier, such as noisy redundant, removing stopwords, feature extraction and feature selection.

2 Related work

There are many works for identification of Arabic dialects in all the levels. We chose some of the close works to our work.

Belgacem et al. (2010) worked on nine dialects (Tunisia, Algeria, Syria, Lebanon, Yemen, Egypt, Golf's Countries, Morocco and Iraq) using the platform Alize and Gaussian Mixture Models (GMM). They showed the complexity of the automatic identification of Arabic dialects. Elfardy & Diab (2013) used a supervised approach for identification of Arabic dialects. They got an accuracy of 85.5% on an Arabic online-commentary. Cotterell et al. (2014) presented a multi-dialect, multi-genre, human annotated corpus of dialectal Arabic with data obtained from both online commentary on the newspaper and Twitter. They used five Arabic dialects (Egyptian, Levantine, Gulf, Maghrebi and Iraqi). Sadat et al.(2014) presented a set of experiments of letter-based (n-gram) Markov language model and NB classifiers on social media. Experimental results showed that NB classifier using character-level bigram model can identify the 18 different Arabic dialects with a considerable accuracy. Malmasi & Zampieri (2016) described a system to identify of four regional Arabic dialects (Egyptian, Levantine, Gulf, North African) and Modern Standard Arabic (MSA) in a transcribed speech corpus as a DSL shared task. They used ensemble classifier of set of linear models as base classifiers and they achieved a score of 0.51 in the closed training track. El-Haj et al. (2018) presented Subtractive Bivalency Profiling (SBP) for identification of four Arabic dialects(Egyptian, Levant , Gulf , and North African) as well as MSA where the accuracy were 76%. Mishra, & Mujadia (2019) explored the use of different features (char, word n-gram, language model probabilities, etc) on different classifiers for Arabic dialects identification. The work is part of Multi Arabic Dialect Applications and Resources (MADAR) Shared Task (Bouamor, et al.,2019) in WANLP 2019 on Arabic Fine-Grained Dialect Identification. They showed that traditional machine learning classifier tends to perform better when compared to neural network models in a low resource setting. Salameh et al. (2018) presented a fine-grained dialect classification task covering 25 specific cities from across the Arab World, in addition to Standard Arabic. They used several classification systems with large space of features. Their results show that the exact city of a speaker can be identified at an accuracy of 67.9%.

ADI, in our work, is achieved by (i) identifying Arabic language from other similar languages, (ii) identifying dialects from MSA & CA, and (iii) identifying dialects among 21 Arabic dialects. The final step is achieved by voting among three well-known and very different classifiers (NB, LR and DT) in two different methodologies. Also, because the used data is none golden standard, little steps of noisy removal are done.

3 Data set

The used data set is NADI Tweeter data set (Abdul-Mageed et al., 2020). It consists of three parts training part of 21,000 tweets, development part of 4957 Tweets and Test set of 5,000 tweets. It is not golden standard corpus and has different noise levels such as existing of 405 non-Arabic tweets (Kurdish and Persian). Also, this data set is mixed of DA, MSA and CA. therefore a noisy removal should be taken as preprocessing. The data sets are labeled in two levels; first level (country level) of 21 countries and second level (provinces level) of 100 provinces. Table 2 show the statistics of training data set **with /without** non-Arabic tweets.

4 Our system

Our system consist of five phases: (i) preprocessing, (ii) noisy tweets removal, (iii) formal clitics and stop words removal, (iv) features extraction and selection, and (v) the classification. Figure 1 explained the proposed method. They will be explained in the next few sections.

Country	# doc with non-Arabic	# doc without non-Arabic	Country	# doc with non-Arabic	# doc without non-Arabic
Algeria	1491	1486	Oman	1098	1072
Bahrain	210	210	Palestine	420	420
Djibouti	210	210	Qatar	234	234
Egypt	4473	4471	Saudi_Arabia	2312	2306
Iraq	2556	2174	Somalia	210	210
Jordan	426	426	Sudan	210	210
Kuwait	420	420	Syria	1070	1067
Lebanon	639	639	Tunisia	750	748
Libya	1070	1070	United_Arab_Emirates	1070	1063
Mauritania	210	210	Yemen	851	781
Morocco	1070	1069	**Totally**	21,000	20,496

Table 2: statistics of training data set with and without non-Arabic tweets

4.1 Preprocessing

Preprocessing is a step used in almost all NLP applications. In this work, this stage consists of two main steps **noisy preprocessing**, and **non-Arabic letter removal**. The first step is done before **noisy tweets removal**, it is achieved by deleting English letters, special symbols, numbers, tweeter mark-up, Emoticons, repetition letters etc., and unification of letter variants (normalization). The second step of preprocessing is removing non-Arabic letters from Arabic tweets (only), it is applied after noisy tweets removal.

4.2 Noisy tweets removal

As was mentioned previously the data set has levels of errors such as foreign tweets and MSA. Approximately 504 tweets were recorded manually as non-Arabic tweets. These tweets and the others are used for learning the binary classifier in character-level and word-level for identifying the foreign languages such as Persian and Kurdish languages. From many tests, the best was unigram for word-level and bigram for character-level with Naïve Bayes classifier. In the classification step of development and test sets, we should see that all the classified tweets as foreign language will be classified as Iraq class because of the probability of being foreign as Iraq class is the highest (2556-2174)/504 ≈ 0.76, see table 2 for more details.

In case of dialects and non-dialects (CA & MSA), 2,000 tweets are classified manually from training set as DA or non-DA. Then these two types of tweets (2,000 tweets) are used as centers for two clusters. Kmean clustering was used for clustering the remaining tweets into the two clusters using Simi-supervised clustering where the manually classified tweets will not be changed their clusters in the iterations of Kmean clustering but stay always in the specific cluster. The non-DA cluster is checked manually only because it was small and all the dialects tweets were removed. The final clusters, in our case are two classes, are used for leaning a binary classifier for using it in the identification of dialects from non-dialects. In the classification step of development and test sets, we should see that all the classified tweets as non-DA will be classified as Egypt dialects but it is bad selection therefore it will produce extra errors in the evaluation.

4.3 Formal clitics removal

The third phase is formal clitics removal such as “ال- Al”- (the), “وال-wAl”-(and the) and “وللـ-wll”-(and for the) but not the letter “هـ” “h” in word “هألعب- hÂlςb”-(I will play) because it is dialectal clitics in Egyptian dialects but not formal clitics (not in modern standard Arabic). Also, the stop words will be deleted in this stage but a very simple list of stop words is taken and they are tuned using the training set according to threshold. For example, the word “عننه- ɕnnh”-(about us) is dialect therefore it will be not deleted but the word “عنه- ɕnh”-(about us) will be deleted.

4.4 Feature selection

The fourth phase is feature selection which started by selecting the effective prefixes and suffixes of size of (1-4) letters according to their threshold and weights. Also, all the words (without formal clitics) are taken as features where their features are TF-IDF according to the equation below.

$$\mathrm{IDF}_{w} = \log\left(\frac{\text{Number of classes}}{\textit{Number of classes that contain word } w}\right) \ldots(11)$$

$$TFIDF_{w,i} = \mathrm{TF}_{w,i} * \mathrm{IDF}_{w} \ldots.(12)$$

Where *w* represents the word and *i* represents the class.

4.5 Classification

The last phase, the classification process, is achieved by voting among three well-known and very different classifiers (Naïve Bayes, Logistic Regression, and Decision Tree). The classification process is done using voting in two methodologies; the first is normal classification using 21 classes. The second methodology is done using binary classifiers where each one is learned from two classes, the first class represent one class of 21 classes and the second class represent the other 20 classes. For each tweet, there are 21 classification processes done for two classes (“specific country” or “others”). If all classifiers produce “others” class except one give country then this class (country) will be selected as the class for this tweets otherwise other classifier is will be used for classifying among the candidate countries only. For example suppose we try to classify tweet *t*, 18 classifiers gave[2] “others” and 3 classifiers gave Iraq, Kuwait and Qatar respectively then this tweet will be feed to other classifier that learned from training tweets of these 3 classes only and the output will be the final class. But if, in our example, 20 classifiers gave “others” and one gave “Iraq” then the class of the tweet *t* will be Iraq directly without using extra classifier. If all the twenty one classifiers are classified the tweet *t* as “other”, then this tweet is unknown tweet and it will be classified as Egypt class because Egypt class has the highest probability among other classes. We should see that there is not any possibility for getting two appearance of one country from two classifiers because each classifier of 21 classifiers is used for one country.

5 Results

The system was implemented on NADI shared task data set (subtask 1). The results for identification of foreign tweets in development set were 0.985, 1 and 0.992 for precision, recall and f-measure respectively where the total foreign tweets were 132, as part of noisy tweets removal.

The classification results (official[3]) is 27.17 as f-measure for voting without using clustering. For voting with clustering (unofficial), the f-measure is 41.34. For voting (21 binary classifiers) with clustering, the f-measure is 52.38. all these percent’s are for development data set. Table 3 shows the results for these three types of tests. Table 3 shows the summary of all tests results.

We should know that the foreign tweets are classified as Iraq and the MSA tweets are classified as Egypt for evaluation purpose which it in most cases cause dropdown in the scores. Aslo, unknown tweets in the last test are classified as Egypt class.

Test type	Dev data set(F-measure)	Test Data set(F-measure)
voting without clustering	27.17	12.45
voting with clustering	41.34	17.61
voting (21 binary classifiers) with clustering	52.38	20.05

Table 3: All classification tests results

[2] Each classifier is done using voting among three classifiers (NB, LR and DT)

[3] Official means that the results are sent to NADI shared task team before the deadline.

6 Discussion

In this paper, ADI is implemented and applied on NADI shared task dataset of very close 21 classes. The proposed system starts from removing noise till the classification process for 21 classes of all Arabic countries. Three classifiers were combined and used in two methodologies (one classifier for classification of 21 classes or using twenty one binary classifiers). Selecting the classes Iraq, Egypt and Egypt classes for Foreign, MSA and unknown tweets respectively dropped down the macro-average score but really foreign and MSA are noise. The results of classification were very low for many reasons: (i) existing of 504 non-Arabic tweets in training set, (ii) existing of MSA tweets which is used in all the Arabic countries result in high noise in learning phase, (iii) some dialects are close to each other's, (iv) existing of ambiguous tweets where it written as MSA but can be pronounced in many dialects, (v) existing of a city in a country that used dialect close to other country more than its country, (vi) the tweets were classified, in the data set, according to user location but not user dialect which produce errors in the classes of training set and hence the wrong learning.

The result of using twenty two binary classifiers with clustering (removing MSA from training data) gave us the best results because the system will focus on the dialect of each country in the training process without noise tweets.

7 Conclusions and future works

Arabic Dialects are native languages for each country or city in Arab world where sometimes the writing of dialects is same but the pronunciations are different.

We can see four levels of Arabic dialectal identification (ADI) from the easiest task to the most difficult task. The first level is identification of dialectal Arabic from the other two Arabic language varieties. The second level is identification of the main category of dialectal Arabic of (nine or five categories). The third level is identification of country level out of 21 countries. The fourth level is identification of city dialects or town dialects. There are many challenges in ADI according to the level number.

Simply we can conclude that ADI is a hard task for many reasons as was mentioned in discussion section. The task need to golden standard corpus and a dictionary for almost all words used in each dialect. The learning from low noise data set gives a good results but the task is still need to special learning techniques and huge dataset (golden standard).

References

Itani. M. 2018 "*Sentiment analysis and resources for informal Arabic text on social media*." PhD diss., Sheffield Hallam University.

El-Haj M., Ryson P., and Aboelezz M., 2018 "*Arabic Dialect Identification in the Context of Bivalency and Code-Switching,*" Eleventh international conference on language Resources and Evaluation LREC 2018.

Belgacem, M., Antoniadis, G., & Besacier, L. 2010. *Automatic Identification of Arabic Dialects*. In LREC2010.

Darwish, K., Sajjad, H., & Mubarak, H., 2014. *Verifiably effective Arabic dialect identification*. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1465-1468).

Elfardy, H., & Diab, M. (2013, August). *Sentence level dialect identification in Arabic*. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 456-461).

Cotterell, R., & Callison-Burch, C. 2014. *A Multi-Dialect, Multi-Genre Corpus of Informal Written Arabic*. In LREC (pp. 241-245).

Sadat, F., Kazemi, F., & Farzindar, A. 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) (pp. 22-27).

Malmasi, S., & Zampieri, M. 2016. *Arabic dialect identification in speech transcripts*. In Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3) (pp. 106-113).

Mishra, P., & Mujadia, V. 2019, *Arabic Dialect Identification for Travel and Twitter Text*. In Proceedings of the Fourth Arabic Natural Language Processing Workshop (pp. 234-238).

Bouamor, H., Hassan, S., & Habash, N. 2019. "*The MADAR shared task on Arabic fine-grained dialect identification*. In Proceedings of the Fourth Arabic Natural Language Processing Workshop (pp. 199-207).

Salameh, M., Bouamor, H., & Habash, N. 2018. *Fine-grained Arabic dialect identification*. In Proceedings of the 27th International Conference on Computational Linguistics (pp. 1332-1344).

Abdul-Mageed, Muhammad, Zhang, Chiyu, Bouamor, Houda and Habash, Nizar 2020 "*The Shared Task on Nuanced Arabic Dialect Identification (NADI)*", Proceedings of the Fifth Arabic Natural Language Processing Workshop (WANLP2020), Barcelona, Spain