Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2016"

Moscow, June 1–4, 2016

AUTOMATIC ARABIC DIALECT CLASSIFICATION[1]

Durandin O. V. (durandin@dictum.ru)[1,2],
Strebkov D. Y. (strebkov@dictum.ru)[1],
Hilal N. R. (hilal@dictum.ru)[1,2]

[1]Dictum Ltd., Nizhny Novgorod, Russia
[2]Lobachevsky State University of Nizhni Novgorod,
Nizhni Novgorod

The paper presents work on automatic Arabic dialect classification and proposes machine learning classification method where training dataset consists of two corpora. The first one is a small corpus of manually dialect-annotated instances. The second one contains big amount of instances that were grabbed from the Web automatically using word-marks—most unique and frequent dialectal words used as dialect identifiers. In the paper we considered four dialects that are mostly used by Arabic people: Levantine, Egyptian, Saudi and Iraq. The most important benefit of that approach is the fact that it reduces time expenses on manual annotation of data from social media, because the accent is made on the corpus created automatically. Best results that we got were achieved with Naïve Bayes classifier trained using character-based bigrams, trigrams and word-marks vocabulary: precision of classification reaches 0.92 with F_1-measure equal to 0.91 on the test set of instances taken from manually annotated corpus.

Key words: machine learning, classification, Arabic language, dialects of Arabic language

АВТОМАТИЧЕСКАЯ КЛАССИФИКАЦИЯ АРАБСКИХ ДИАЛЕКТОВ

Дурандин О. В. (durandin@dictum.ru)[1,2],
Стребков Д. Ю. (strebkov@dictum.ru)[1],
Хилал Н. Р. (hilal@dictum.ru)[1,2]

[1]ООО «Диктум», Нижний Новгород, Россия
[2]ННГУ им. Н. И. Лобачевского, Нижний Новгород, Россия

Ключевые слова: машинное обучение, классификация, арабский язык, диалекты арабского языка

[1] The paper is funded by the Russian Federal Foundation for Assistance to Small Innovative Enterprises (FASIE), project number is 0019251. The authors express their gratitude to the Foundation.

1. Introduction

Despite the fact that Arabic is the official language for more than 250 million people, the native Arabic speakers actually speak on different variations of Arabic language—dialects (lahjah, or Dialectal Arabic—DA). At the same time, the standardized variety of Arabic used in writing and in most formal situations is known as Modern Standard Arabic (MSA).

The dialectal variants of Arabic are widely used in everyday life, at live communication or chatting. Dialects are not taught in schools, also there are no special textbooks for the large-scale study of the current dialect. Man absorbs dialect since his birth, inheriting the style of conversation from the society in which he lives.

Taking into account these peculiarities of the Arabic, we could think of such NLP task as identification of Arabic dialect. Practical usage of a system that is capable to detect dialect accurately consists of several aspects. Among them:

- identifying nationality of an author of some social media text;
- revealing common ideas/preferences;
- finding actual topics that are actively discussed in some concrete Arabic country or city.

In fact, many social media provide geolocation functionality that allows identifying location of an author. However, as we will show below, results of such approach are highly ambiguous. Therefore we started our investigation of DA classification problem.

It is essential to highlight that we don't address a problem of classifying text between DA and MSA in this paper. We do have a classifier that is capable to perform distinguishing between these two Arabic varieties. So, from now on, we treat any given text input as an input on DA.

2. Related Work

Currently most of works related to Arabic NLP are based on MSA. As for DA, most of research papers are focused on Egyptian, Iraq and Levantine dialects.

Habash et al. (2008) [6] presented annotation guidelines for the identification of DA content embedded in MSA context. They presented annotation results on a set of around 1,600 Arabic sentences (19k words), with both sentence- and word-level DA annotations.

The next effort was the COLABA project by Diab et al. (2010) [3], it consisted of resources and processing tools for DA blogs. One of such tools was developed to determine the degree to which a text includes DA words. The tool tried to determine how many words are not MSA in some given text input. Results that were achieved stated that 50% of all bigrams and 25% of trigrams contained at least one dialectal word.

Zaidan and Callison-Burch (2011) [10] created the Arabic Online Commentary (AOC) Data set by extracting reader's comments from online Arabic forums. The selected sentences were manually labeled with one of 4 dialect labels with the help of crowdsourcing: Egyptian, Gulf, Iraq and Levantine.

Elfardy et al. (2013) [5] introduced an approach to perform dialect identification between MSA and Egyptian dialect on sentence level. Token level labels from AOC data set were used to generate sentence level features. Generated features were combined with core features to train a classifier that could perform label prediction for each given sentence. The system achieved an accuracy of 85.5% on an AOC data.

Zaidan and Callison-Burch (2014) [11] presented a novel Arabic resource with DA annotations. Using that new resource, they considered the task of DA identification: used the data to train and evaluate classifiers for that task and established that classifiers using dialectal data significantly and dramatically outperform baselines that use MSA-only data, achieving near-human classification accuracy.

Sadat et al. (2014) [9] described the usage of the character n-gram language model and Naïve Bayes classifiers with examination of what models perform best under different conditions in social media context. The classifier that authors trained using character bigram model could identify the 18 different Arabic dialects with a considerable overall accuracy of 98%.

3. Linguistic Background: The MSA/DA Distinction in Arabic

In this paragraph we will provide some essential information regarding Arabic and its dialects, and explain our approach for preparation of DA corpora.

3.1. MSA and dialects

The phrase "Arabic language" includes different variants of one language. It may be:

- Classical Arabic (the language of the Koran);
- MSA;
- DA.

If we work with Arabic literature, books, official letters, it is sufficient to be able to recognize and understand both classical Arabic and MSA. But to recognize the content of blogs, forums and user commentary it is necessary to take into account the features of language in which the author writes posts. In our case, such languages are colloquial regional dialects, with similarities and differences in comparison, both among themselves and with the MSA.

3.2. MSA and dialects: Linguistic differences

MSA is regulated with strict linguistic rules and laws that describe the standards of forming words and sentences established centuries ago. Unlike MSA, DA doesn't have an explicit written set of regulated grammar rules.

The main differences between MSA and DA are ([7] and [11]):

1. MSA has the largest set of negative markers (la, lan, lam, laysa, ma) while the dialects are restricted to three (maa, muš/miš/maš/maši/muu/mub, laa);

2. Lack of spelling standards. For example, the algorithm of verb conjugations in the dialects differs from this algorithm in MSA. Also, in DA the dual form of the verbs and pronouns is almost absent and instead of it plural form is often used;

3. There are new function words and particles, that replacing it's similar versions in MSA, for example:
   - DA pronoun: اللي—ايه—هديك—هذول—هدا—هداك—دول—دي—ده ...
   - DA negative particles: مش—مو ...
   - DA interrogative particles: شو—أشو ...

4. There are additional function words and particles, which have no analogues in the MSA and carrying syntactic or morphological information, for example:
   English: Jordan's King
   MSA: ملك عمان
   DA: الملك تبع عمان (the middle token is similar in meaning to the possessive pronoun, but it is a dialectal word and requires additional syntactic processing).

5. A more complex cliticization system:
   In MSA there are a large number of fused particles (prepositions, pronouns, etc.) and cliticization process occurs in a certain patterns. However, in DA words we observe differences in the clitics themselves and in the logic of cliticization process.
   - "This man":
     هذا الإنسان (2 tokens on MSA)
     هالإنسان (1 token on DA)
   - "And you didn't write this to him":
     و لم تكتبوها له (4 tokens on MSA)
     وماكتبتوهالوش (1 token on DA)

6. The presence of an extensive vocabulary, which has no common roots with its MSA synonyms, which complicates the identification of grammatical information for these words, for example:

Table 1. Differences in word forming between DA and MSA

English word	MSA variant	Dialectal variant
want	أريد	بدي
also	أيضا	كمان

Moreover, each dialect can differ from another [2]. For example, in Tunisian dialect it is observed specific lexico-semantic, morphological and syntactic properties that differ from Saudi and Egyptian dialects [2].

The above linguistic differences and characteristics require a special approach to treat text written in DA.

3.3. Arabic dialects identification

Currently there are more than 30 dialects of Arabic, depending on the region, country and even village where Arabic people live. But according to popularity, the main dialects are: Levantine, Egyptian, Saudi, Algerian and Gulf-Arabic.

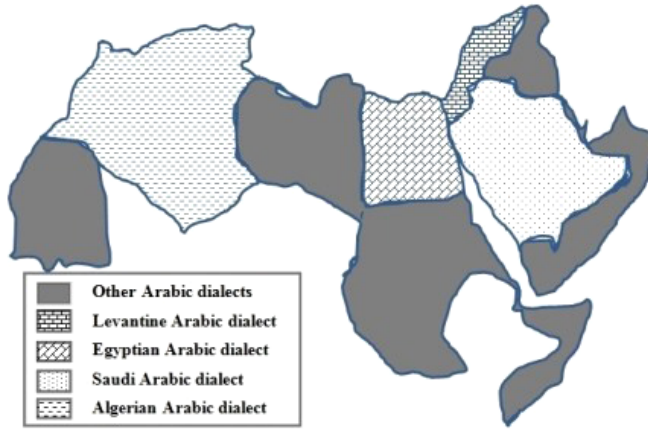

Fig. 1. Arabic dialects map

The ability to identify the kind of a dialect the text was written in makes it possible to recognize (ethnic, civil, etc.) membership of the author. This in its turn will help in monitoring general ideas, preferences, opinions and relevant topics discussed by citizens of a particular region.

In this paper we paid attention to the four most common dialects, which are mostly used by people in Arabic world in social media communications: Levantine, Egyptian, Saudi and Iraq.

It should be noted that Arabic dialects don't fully consist of DA words but are a mix of MSA with DA words, which identify regional affiliation of the dialect. DA words may be common for several regional dialects (example MSA: في الخارج, Levantine, Egyptian and Saudi: برة), and may be unique for a particular dialect (example MSA: أريد, Levantine: بدي, Egyptian: عايزة).

So, we decided to collect the most unique and frequent dialectal words—word-marks—and use them as markers for identifying the type of dialect in which a message/comment is written in the Web. These word-marks include some pronouns, interrogative and negative particles, few nouns and verbs:

Table 2. Examples of word-marks of Arabic dialects

Dialect	Dialect word-marks
Levantine	ليش — منيح — أديش
Egyptian	إزيك — النهارده — دلوقتي
Saudi	الحين — مهوب — أبغي
Iraq	شكو ماكو — بلكت — وينج

3.4. Collecting DA corpus

Above we observed features of DA and stated that several dialects have special words—word-marks that could be met only in one concrete kind of the dialect. So, we could induce an empirical rule: if we find word-mark that belongs to dialect D_i in some given text, then the whole given text is written on that dialect D_i.

Since these word-marks are widely used and met very frequently, it is possible to label messages as dialectal.

We used Twitter to collect all required data. Twitter API doesn't support extracting messages that were published more than 2 weeks ago. To overcome that limitation, we developed utility module TweetMiner that uses Twitter's Web interface to get messages with publishing date until 2 years ago. The module provides several tuning parameters that could customize extraction process:

- time interval
- geolocation
- author of the messages
- search query that has to be contained in found messages.

Finally, the module was executed with word-marks as search queries and grabbed tweets in time interval between July 2015 and January 2016. We collected tweets using word-marks of four Arabic dialects: Egypt, Iraq, Levantine and Saudi. The number of tweets that we got using TweetMiner was 4,387,806. Since dialectal word-marks were used in grabbing process, each tweet among collected was associated with corresponding Arabic dialect. All that information was stored in database forming so-called dialectal database (dialectal DB).

The distribution of tweets from dialectal DB per dialect is presented on figure 2.

In addition to that, we utilized TweetMiner once again and collected tweets that were published during June 2015 (so, these tweets don't intersect with tweets from dialectal DB). Arabic linguists manually annotated a small part of these tweets, so we got 51,589 tweets with correct dialectal labels. It is essential to notice that resulting collection was extended with tweets without word-marks; these tweets were manually found in Twitter and annotated by the linguists.

Figure 3 shows distribution of tweets from that manually annotated corpus per dialect.

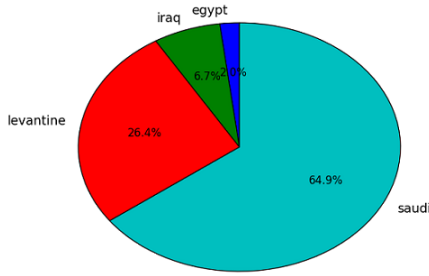

Fig. 2. Dialectal distribution of tweets from dialectal DB

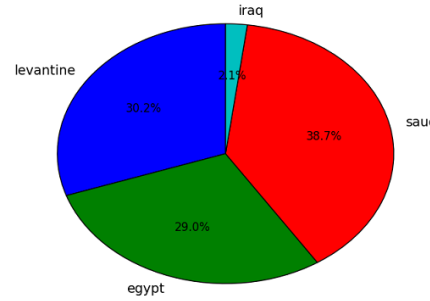

Fig. 3. Dialectal distribution of tweets from the manually annotated corpus

In fact, Twitter API allows executing a Twitter search using geolocation queries. However, that approach doesn't provide relevant results. For example, we can't be sure that some tweet with Iraq geolocation mark does contain text written on Iraq dialect.

A small investigation was made to confirm that fact. Around 30% of all tweets from dialectal DB had associated geolocation marks. So, we took all tweets with geolocation marks that correspond to the some concrete dialect and created a diagram that shows country distribution of these tweets. The process was repeated for all 4 dialects that we were working with.

The figure 4 shows distributions that we finally got.

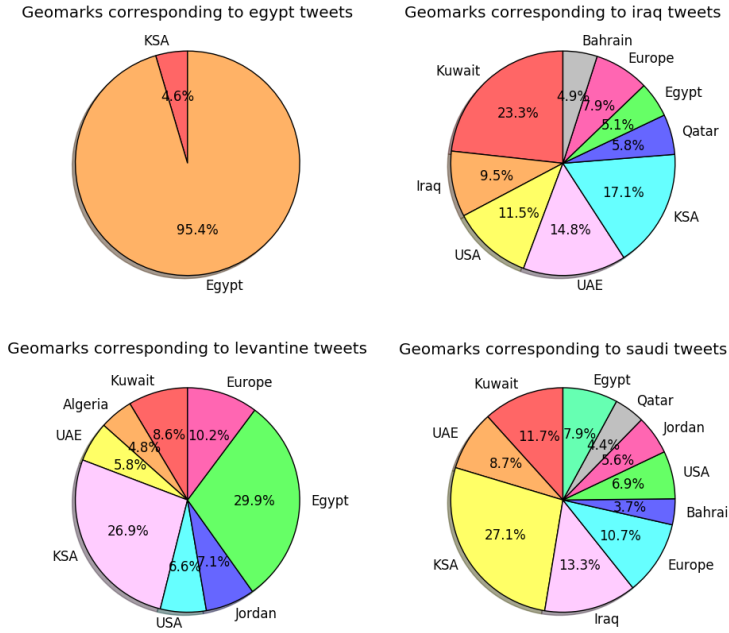

Fig. 4. Country distributions of tweets from dialectal DB that have geolocation marks

As we can see from that figure, geographical location correlates with used dialect very weakly. In might be caused by high migration rate in Arabic countries. Another interesting observation that could be made from the figure is connected with Saudi Arabia geolocation mark: it contains in all diagrams. It means that if we execute Twitter search with Saudi Arabia geolocation mark, we could get Tweets that contain texts not only on Saudi dialect, but on Egypt, Iraq and Levantine dialects as well.

So, to sum up, geolocation marks provide irrelevant results and can't be used in dialect identification task.

4. Model

We treat dialect classification task as a kind of a statistical language identification task [4]. Language identification is the task of identifying the language of a given document [1].

Our approach for Arabic dialect identification is focused on character-based n-grams and Naïve Bayes classifiers.

The reason why we used these n-grams is the fact that in many cases the difference between Arabic dialects is based on used affixes, and these affixes could be easily extracted using character-based n-grams model. In addition to that, since word-marks are unique words that correspond to some particular Arabic dialect, they were added to the set of word-based features.

4.1. Naïve Bayes Classifier

We experimented with several classifiers, but best results for our dialect classification task were achieved using Naïve Bayes classifier [8].

Naïve Bayes classifier is a set of supervised learning algorithms based on applying Bayes' theorem with the "naïve" assumption of independence between every pair of features. Given a class variable y and a dependent feature vector x_1, ..., x_n, Beayes' theorem states the following relationship:

$$P(y|x_1, \dots, x_n) = \frac{P(y)P(x_1, x_2, \dots, x_n|y)}{P(x_1, x_2, \dots, x_n)}$$

Using the naïve independence assumption that: $P(x_i|y, x_1, x_2, x_{i-1}, x_{i+1}, \dots, x_n) = P(x_i|y)$, for all i, this relationship is simplified to:

$$P(y|x_1, \dots, x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1, x_2, \dots, x_n)}$$

Since $P(x_1, x_2, \dots, x_n)$ is constant given the input, we can use the following classification rule:

$$P(y|x_1, \dots, x_n) \propto P(y)\prod_{i=1}^{n} P(x_i|y)$$

So,

$$\hat{y} = \arg\max_{y} P(y) \prod_{i=1}^{n} P(x_i|y)$$

We can use Maximum A Posteriori (MAP) estimation to estimate $P(y)$ and $P(x_i|y)$; the former is then the relative frequency of class y in the training set.

The features x_i in our task represent absence/availability local n-grams, and vocabulary terms.

4.2. Experiment: data and features

Data. As it was mentioned before, we had two dialectal Arabic corpora:

- Dialectal DB (4,387,806 tweets);
- Manually annotated corpus (51,589 tweets).

Taking into account rather small size of the second corpus, we came up with the following train/test data split: we train our model on 80k tweets from dialectal DB and combine them with 33% of the second, manually annotated corpus (~17k tweets). Needed to mention that 80k tweets is just 1.8% of all tweets stored in dialectal DB, but we took that relatively small amount due to PC memory limitations.

As for testing, we tested our model on the rest 66% (~34k tweets) of the manually annotated corpus.

Features. We executed the following preprocessing procedure for each tweet: removed non-Arabic symbols, underlines, smiles, etc. So, finally each tweet contained Arabic text written without anything else. Table 3 below shows all attributes that were tested with our model:

Table 3. Sets of attributes that were tested

Features	Precision	F_1-measure
bigrams	0.877	0.869
bigrams + word-marks vocabulary	0.918	0.907
bigrams + trigrams	0.908	0.899
bigrams + trigrams + word-marks vocabulary	**0.923**	**0.912**
bigrams + trigrams + 4-grams	0.916	0.905
bigrams + trigrams + 4-grams + word-marks vocabulary	0.919	0.909

Best results were achieved using a set of three attributes: bigrams, trigrams and word-marks vocabulary.

4.3. Experiments with the size of train data

We also made experiments to find out how varying of the size of train data affects precision/recall for our model. The same set of attributes was used in all of experiments described further: bigrams, trigrams and word-marks vocabulary.

Firstly, we made an experiment that was connected with varying the size of data taken from manually annotated corpus. So, we fixed training part taken from of dialectal DB to 80k tweets.

We varied the size from 0 to 50% of all manually annotated data. Testing was done on the rest 50%. Figure 5 shows dependencies between precision/recall of the classifier and the size of annotated data.

Dashed line on the plot represents a baseline. In our model we took predicting the most frequent class—Saudi dialect—as a baseline. Therefore, baseline for precision/size dependency is equal to 0.64.

We could also notice that if we will not use data from manually annotated corpus at all, precision has quite low value; however, it beats the baseline. As we increase the size of manually annotated data in train set, both precision and recall increase as well. Maximum value for precision is 0.935; it is observed when 50% of manually annotated data used for training. Such observation could be explained by the fact that manually annotated dataset was cleared from various types of mistakes. In addition to that, it also contains tweets without word-marks.

In our second experiment we tried varying the size of data taken from dialectal DB. The experiment was executed with the following setting: we used 33% of manually annotated data for training, and the 66% for testing.

Figure 6 represents dependencies between precision/recall of the classifier and the size data taken from dialectal DB.

As we could see, increasing the size of data from dialectal DB doesn't lead to significant increase of precision/recall on manually annotated data used for testing. In our opinion, such result is most likely caused by the fact that training dialectal DB data doesn't fully cover all patterns of word formation that exist in DA. Way out here could be, firstly, in grabbing tweets from wider time interval (half-year or more), and secondly, in using other data sources in addition to Twitter (blogs, etc.).

Finally, confusion matrix on Table 4 illustrates situations when classifier tends to make mistakes.

Table 4. Confusion matrix of the classifier

	Egypt	Iraq	Levantine	Saudi
Egypt	**0.969**	0.008	0.012	0.011
Iraq	0.003	**0.937**	0.044	0.016
Levantine	0.002	0.003	**0.957**	0.038
Saudi	0.004	0.010	0.178	**0.808**

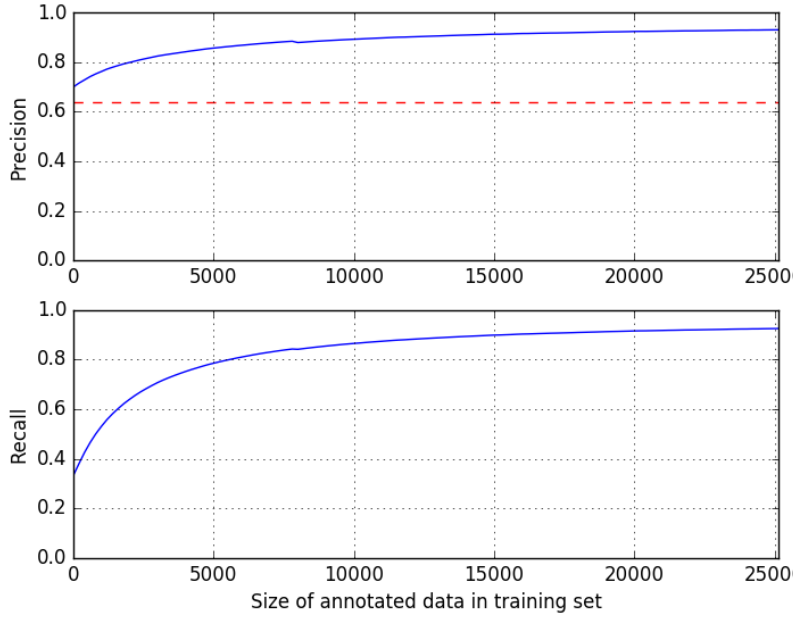

Fig. 5. Dependencies between precision/recall of the classifier and the size of annotated data

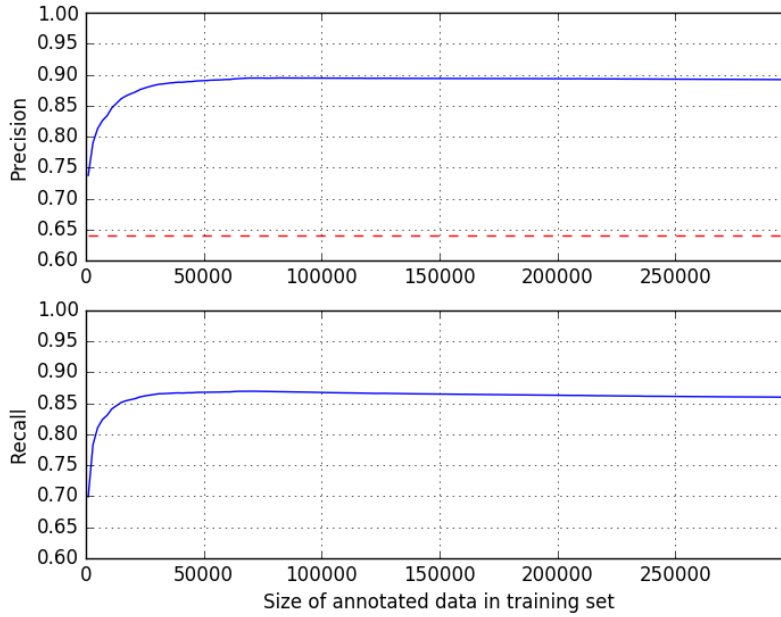

Fig. 6. Dependencies between precision/recall of the classifier and the size of data from dialectal DB

5. Conclusion

The paper presented a study on dialect identification of Arabic language using texts from Twitter and explained insolvency of classification based on Twitter geolocation attributes.

Two datasets were used in the experiment: dataset of tweets grabbed from Twitter using word-marks (the formed dialectal DB) and dataset of tweets that were manually annotated with proper dialectal labels by Arabic linguists (that dataset was also extended with tweets without word-marks).

We showed how different sets of classification attributes (bigrams, trigrams, 4-grams, word-marks) affect quality of Naïve Bayes classifier that was used in experiments.

In the paper we proposed machine learning method where training dataset contains big amount of instances taken from dialectal DB and rather small number of instances from manually annotated corpus. The benefit of that approach is the fact that it reduces time expenses on manual annotation of data from social media. Results that we got look rather promising: precision of classification reaches 0.92 with F_1-measure equal to 0.91.

At the end, we presented results of varying the amount of manually annotated data and data from dialectal DB in training process. The effect that these changes make on precision/recall metrics of the classifier was stated as well.

6. Future Work

In the future, we are planning to increase the number of handled Arabic dialects.

Also it would be interesting to try a semi-supervised approach for our dialect classification task: we expect that such methods will reduce the percentage of errors caused by incorrect dialect labels assigned automatically for instances of our dialectal DB.

References

1. *Baldwin T., Lui M.* (2010), Language Identification: The Long and the Short of the Matter, Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, Los Angeles, California, pp. 229–237.
2. *Blinov A. A.* (2009), Territorial varieties of literary Arabic language and their reflection in press [Territorial'nye varianty arabskogo litaraturnogo jazyka i ih otrazhenie v presse], PhD Thesis, Institute of Oriental Studies of the RAS [Institut vostokovedenija Rossijskoj Akademii Nauk], Moscow.
3. *Diab M., Habash N., Rambow O., Altantawy M., Benajiba Y.* (2010), COLABA: Arabic dialect annotation and processing, In LREC Workshop on Semitic Language Processing, Valleta, Malta, pp. 66–74.
4. *Dunning T.* (1994), Statistical Identification of Language, Technical Report MCCS 94–273, New Mexico State University.
5. *Elfardy H., Diab M.* (2013), Sentence level dialect identification in Arabic, Proceedings of the 51st Annual Meeting of the ACL, Sofia, Bulgaria, pp. 456–461.

6. *Habash N., Rambow O., Diab M., Kanjawi-Faraj R.* (2008), Guidelines for annotation of Arabic dialectness, Proceedings of the LREC Workshop on HLT & NLP within the Arabic World, Marrakech, p.p. 49–53.
7. *Habash N., Roth R., Eskander R., Tomeh N.* (2013), Morphological Analysis and Disambiguation for Dialectal Arabic, Proceedings of HLT-NAACL, Atlanta, Georgia, pp. 426–432.
8. *Manning C., Raghavan P., Schütze H.* (2008), Introduction to Information Retrieval, Cambridge University Press.
9. *Sadat F., Kazemi F., Farzindar A.* (2014), Automatic identification of Arabic language varieties and dialect in social media, Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP) ,Dublin, Ireland, pp. 22–27.
10. *Zaidan O., Callison-Burch C.* (2011), The Arabic Online Commentary Dataset: An annotated dataset of informal Arabic with high dialectal content, Proceeding of the 49th Annual Meeting of the ACL: Human Language Technologies, Portland, OR, pp. 37–41.
11. *Zaidan O., Callison-Burch C.* (2014), Arabic Dialect Identification, Computational Linguistics, March 2014, Vol. 40, No. 1, pp. 171–202.