Processing Language ICALP2019 International Arabic Conference

Automatic identification methods on a corpus of twenty five fine-grained Arabic dialects

Salima Harrat, Karima Meftouh
Karima Abidi, Kamel Smaïli

October 16, 2019

Arabic and its dialects

- Standard Arabic is the official language of Arab countries.
- But...all over the Arab world people use Arabic dialects in everyday conversations.
- These dialects are a variant of Arabic language.
- They differ widely between and within Arab countries.
- They share a lot of features with standard Arabic.

Arabic dialects

Arabic dialects are generally classified according to:

- East-West dichotomy: Maghrebi / Middle-East dialects.
- The ethnic and social diversity of Arab speakers: Rural / Bedouin dialects.

Arabic dialects

- Until recently, these dialects were mostly spoken.
- Social media and mobile telephony have promoted their written form.
- Arab people express their opinions by using their dialects instead of Arabic or other standard languages

Arabic dialects identification challenges

In their oral form, Arabic dialects are relatively easy to distinguish. prosody and tone bring important information about them.

- All Arabic dialects share a lot of lexical units with modern standard Arabic.
- In their conversations, Arab people tend to switch to standard Arabic especially when discussing matters relating to religion.
- Some words are shared among Arabic dialects but with different meanings.
- The lack of dialectal resources such as monolingual and multilingual corpora makes the identification task a challenging issue.

Dialect identification approaches

- A Long Short-Term Memory (LSTM) neural network .
- Word embedding.
- Symmetric Kullback-Leibler measure.
- Multinomial Naïve Bayes (MNB) approach.

Long Short-Term Memory neural network approach

- Long Short Term Memory Networks (LSTM) are a special class of neural networks able to learn long-term dependencies.
- LSTM networks have been efficient for many NLP tasks: language modeling, sentiment analysis, etc...
- And other area like automatic speech recognition and image captioning

Processing Language ICALP2019 International Arabic Conference

Long Short-Term Memory neural network approach

We considere dialect identification task as a multi-class classification problem by using a LSTM network that includes:

- An input layer that takes as input a vector of characters/words n-grams (for characters n varies from 1 to 5 and for words it varies between 1 and 2).
- a LSTM layer,
- A drop out layer to prevent over-fitting.
- And an output layer using a softmax function that gives a probability distribution over the different dialect labels.

Word embedding based approach

- The idea: Using semantic information encoded by word embedding for dialect identification.
- We used the CBOW method of Word2Vec model to represent words vectors.
- Each sentence of the training corpus is extended with infra-lexiacl information (2-5 grams of characters).
- We constitued for each dialect, a set of infra-lexical units which do not occur in other dialect: These units are characteristic to each dialect
- We constitued a set of vectors representing the words that are typical for each dialaect, these words include the characteristic infra lexical units of the dialect.

Word embedding based approach

- To assign a label l to a given sentence s, the similarity between the words of s and the list of typical words of each dialect is calculated with the Euclidean distance:

$$d_k = \frac{1}{|s|} \sum_{i=1}^{|s|} \min_{1 \leq j \leq |L_k|} E(s_i, w_j^k) \quad (1)$$

$|s|$: # of words of$s$, L_k: the list of typical words of the dialect k
E: the Euclidean distance, $|D|$ # of dialects/language
w_j^k: the word j belonging to the list of typical words of the dialect k

- Then sentence s is labeled with the label l corresponding to the dialect that gives the smallest distance.

$$i_l = \underset{1 \leq k \leq |D|}{argmin}(d_k) \quad (2)$$

Symmetric Kullback-Leibler for classification

- The General Vocabulary (GV) related to all the corpora is constitued.
- GV includes: all the words, all the 2grams of words and all the 2-5 characters ngrams.
- The distribution of each dialect d_i according to GV is calculated
- Each dialect d_i is assigned a vector where each dimension is given by $P(u_k|d_i)$ (u_k is a unit of GV and d_i corresponds to the dialect i).
- For test purpose, each sentence is assigned a vector in the same way.
- Then SKL measure is calculated between the distributions of the test sentence and each dialect.

$$D(P||Q) = \sum_x ((P(x) - Q(x))Log\frac{P(x)}{Q(x)} \quad (3)$$

- The sentence is then labeled with the dialect that gives the smallest value of SKL.

Multinomial Naïve Bayes (MNB) approach

- We used 1-gram and 2-gram word features,
- 1 to 5 character ngrams features,
- likelihoods estimated by the unigram language models related to each dialect,
- We utilize TermFrequency-Inverse Document Frequency (Tf-Idf) scores instead of count weights.

Data description

For training and testing our classifiers, we used the MADAR shared task data.[1] It consists of two parallel multi-dialect corpora:

- **MADAR-Corpus26**: is composed of parallel sentences translated to 25 dialects of several cities from the Arab countries , in addition to modern standard Arabic. Each dialect/language includes 1600 sentences for training and 200 sentences for test purpose.
- **MADAR-Corpus6**: is a collection of 10K additional sentences translated to the dialects of five selected cities: Beirut, Cairo, Doha, Tunis, and Rabat.

---

1. Bouamor, Houda, Hassan, Sabit, et Habash, Nizar. The MADAR shared task on Arabic fine-grained dialect identification. In : Proceedings of the Fourth Arabic Natural Language Processing Workshop. 2019. p. 199-207.

Dialect identification results

Table 1: Dialect identification results using different approaches.

Training Corpus	MADAR-Corpus 26			MADAR-Corpus 6		
Approach	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Word Embedding	50.11	49.90	49.74	83.96	83.90	83.83
Symmetric Kullback-Leibler	53.21	68.27	53.79	89.05	89.48	89.03
Multinomial Naïve Bayes	69.80	69.15	**69.09**	92.54	92.50	**92.50**
LSTM networks	58.04	61.54	58.33	89.23	89.17	89.18

Table 2: MNB-MADAR-Corpus26 Dialect identification results by dialect /language.

Dial./Lang.	Precision	Recall	F1-score
MOS	83.41	85.50	84.44
ALG	78.08	85.50	81.62
SAN	87.79	75.50	81.18
MSA	71.49	89.00	79.29
ALX	76.17	81.50	78.74
TRI	69.26	80.00	74.25
RAB	78.98	69.50	73.94
FES	72.25	75.50	73.84
SFX	67.52	79.00	72.81
BEI	78.70	66.50	72.09
BEN	70.87	73.00	71.92
TUN	75.14	65.00	69.71
BAG	76.97	63.50	69.59

Dial./Lang.	Precision	Recall	F1-score
ALE	78.12	62.50	69.44
DOH	72.04	67.00	69.43
KHA	63.29	75.00	68.65
BAS	67.15	69.50	68.30
JED	68.45	64.00	66.15
CAI	73.97	54.00	62.43
ASW	59.55	65.50	62.38
RIY	57.14	64.00	60.38
SAL	61.90	58.50	60.15
DAM	56.02	60.50	58.17
JER	54.63	62.00	58.08
MUS	65.52	47.50	55.07
AMM	50.43	59.00	54.38

Dialect identification results/MNB-MADAR-Corpus6

For 6-way classification, the scores are better. The most confused dialects are RAB & TUN followed by CAI & DOH, then BEI & DOH and BEI & CAI (with the same confusion rate), while the most confused dialects with MSA are DOH and CAI.

Table 3: MNB-MADAR-Corpus6 Dialect identification results by dialect/language.

Dialect/Language	Precision	Recall	F1-score
MSA	95.09	96.80	95.94
RAB	94.04	93.10	93.57
TUN	94.25	91.80	93.01
BEI	93.03	90.70	91.85
DOH	88.21	92.80	90.45
CAI	90.72	89.90	90.31

Processing Language ICALP2019 International Arabic Conference

- Using n-grams features helps to increase accuracy
- Character n-grams order varies from 1 to 5, while for word n-grams lower order (1 and 2) achieve the best results

Table 4: The dialect features used in the different approaches.

Approach	Word n-grams features	Character n-grams features
Word Embedding		2-gram to 5-gram
Symmetric Kullback-Leibler	1-gram and 2-gram	1-gram to 5-gram
Multinomial Naïve Bayes	1-gram to 2-gram	1-gram to 5-gram +LMs Prob
LSTM networks	1-gram	4-gram

Processing Language ICALP2019 International Arabic Conference

- Several approaches were explored to tackle the issue of dialect identification using a multi-dialect corpus from 25 arab cities in addition MSA.
- Neural netwoks approaches were considered by using words embedding and LSTM networks.
- The symmetric Kullback-Leibler distance was experimented in addition to Multinomial Naïve Bayes classifier.
- Character and word ngrams features were used.
- The Multinomial Naïve Bayes classifier achieved the best results
- Neural approach results did not achieve high accuracy (as we expected), the training data are not sufficient to learn such classifiers.

Processing Language ICALP2019 International Arabic Conference