Automatic Dialect Recognition in Arabic Broadcast Media

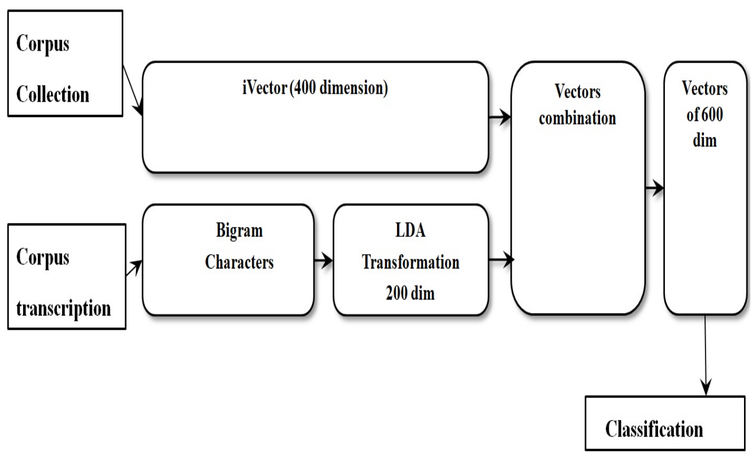
Billal Belainine and Fatiha Sadat
belainine.billal@courrier.uqam.ca, sadat.fatiha @uqam.ca
UQÀM | Université du Québec à Montréal

Introduction

- This work deals with DID using supervised learning methods that emphasize only on labelled data, to discriminate between four major Arabic dialects: Egyptian, Levantine, North African, Gulf and Modern Standard Arabic (MSA) [1].
- Dialect varieties of Arabic: The dialectal Arabic is gregated into five regional language:
  - Egyptian (EGY): dialects of nile valley and Sudan.
  - North African or Maghrebi (NOR): dialects of Morocco, Algeria, Tunisia and Mauritania and Libya.
  - Gulf or Arabian Peninsula (GLF): dialects of Kuwait, Saudi Arabia, Bahrain, Qatar, United Arab Emirates and Oman
  - Levantine (LAV): dialects of Lebanon, Syria, Jordan, Palestine and Israel.
  - Iraqi (IRQ): dialects of Iraq.
- Our team at UQAM is interested in research on Arabic dialect systems in the context of NLP and NLU. [2] [3]

Methodology

- Our proposed approach for Arabic DID focuses on :
  1. Concatenation of i-vector audio representation vectors with bi-gram character-based vectors.
  2. A transformation of the bi-gram character model by LDA.



- The corpus contains two forms of data: [6]
  1. Transcription Buckwalter.
  2. I-Vector audio represantation.

Labeled Corpus(Dialect)	Number of instances
GLF	2744
LAV	2979
NOR	2954
EGY	3115
MSA	2207
Total	13999
Unlabeled Corpus(Test)	Number of instances
All dialects	1492

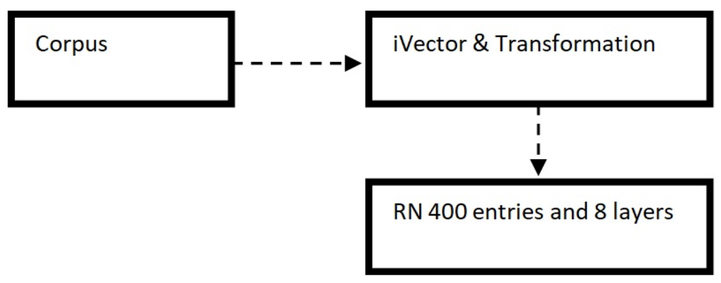
- I-vector is the matrix model of the total variability of a statistics set for each audio track [5].
- Latent Dirichlet Allocation (LDA) for dimensionality reduction[4].

$$p(\theta, z, bigram|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(bigram_n|z_n, \beta) \quad (1)$$

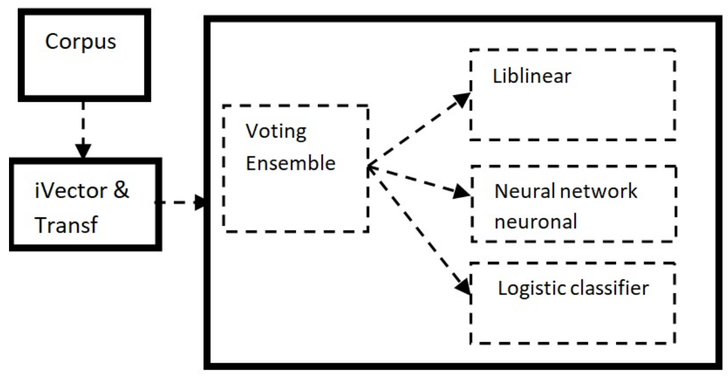
Classification

- Logistic Classifier for Arabic DID:

- Multi-Layer Perceptron for Arabic DID:



- Voting ensemble for Arabic DID:



Experimentations

- We proposed a set of features which improved the NER system performance
  1. Results using cross-validation and the multi-layer perceptron classifiers.

Dialect	Precision	Recall	F-Measure
GLF	0,841	0,840	0,841
LAV	0,771	0,772	0,772
NOR	0,895	0,900	0,897
EGY	0,840	0,857	0,848
MSA	0,849	0,821	0,835
Avg.	0,840	0,840	0,840

  2. Results using cross-validation and the voting ensemble classifiers.

Dialect	Precision	Recall	F-Measure
GLF	0,819	0,863	0,840
LAV	0,768	0,741	0,754
NOR	0,893	0,894	0,893
EGY	0,843	0,856	0,849
MSA	0,849	0,808	0,828
Avg.	0,834	0,834	0,834

  3. Results using cross-validation and the logistic classifiers.

Dialect	Precision	Recall	F-Measure
GLF	0,820	0,818	0,819
LAV	0,737	0,716	0,726
NOR	0,878	0,885	0,881
EGY	0,812	0,821	0,816
MSA	0,790	0,803	0,797
Avg.	0,809	0,810	0,810

- Results on ASRU/MGB test file

Classification algorithms	Accuracy(%)
Voting Ensemble	56.10
Logistic	54.56
Multi-layer perceptron	53.6

- Based on these results, we could notice the following conclusions:
  1. Combining different classification algorithms (Voting ensemble), gives the best results in term of accuracy.
  2. The F-measures for the three algorithms and for each Arabic dialect are located between 72 % and 89%. The average F-measure on the five dialects is located between 81% and 84%, which is very promising.

References

[1] Sadat, F., Kazemi, F. et Farzindar, A. (2014) Automatic identification of arabic dialects in social media. In ACM (pp. 35-40). .
[2] Sadat, F. (2013). Arabic social media analysis for the construction and the enrichment of nlp tools. In Corpus Linguistics.p22-26.
[3] Sadat, F., Kazemi, F. et Farzindar, A. (2014). Automatic identification of arabic language varieties and dialects in social media. In Proceedings of SocialNLP.
[4] BLEI, David M., NG, Andrew Y., et JORDAN, Michael I. (2003). Latent dirichlet allocation. In Journal of machine Learning research,p.993-1022.
[5] Dehak, N., Dehak, R., Kenny, P., Brümmer, N., Ouellet, P., Dumouchel, P. (2009) Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. In Tenth Annual conference of the international speech communication association,p217-250.
[6] Bontcheva, K., Derczynski, L., Funk, A., Greenwood, M. A., Maynard, D. et Aswani,N. (2013). Twitie : An open-source information extraction pipeline for microblog text. In RANLP,p83-90.

Conclusion and perspectives

- Proposed a set of features which improved the DID system performance
- The best results were obtained during this shared task on closed submission using the Voting Ensemble with an overall accuracy of 56.10, followed by the simple logistic and Multi-Layer perceptron with an overall accuracy of 54.56 and 53.55, respectively.
- Future research include:
  - Adding more domain-specific features
  - Exploring semi-supervised learning algorithms using more unlabelled data.
  - Studying a hybrid model for dialect identification with the involvment of character-based and word-based models.