Arabic Dialect Identification

in the Context of Bivalency and Code-Switching

Mahmoud EL-Haj

SCC, Lancaster University

@DocElhaj

Paul Rayson

SCC, Lancaster University

@perayson

Mariam Aboelezz

British Library

@MariamAboelezz

Overview

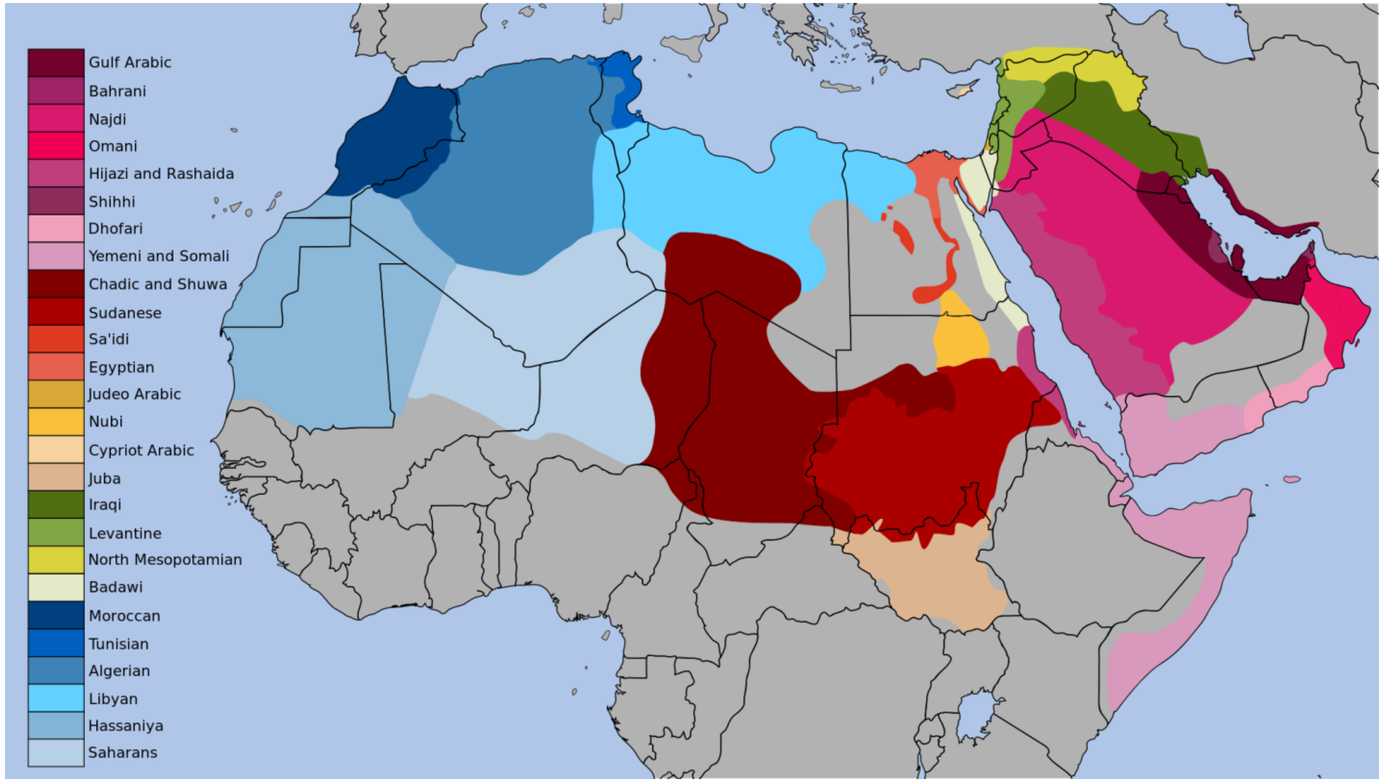
- Automatically Identify Written Arabic Dialects using Machine Learning.
- Incorporate grammatical and stylistic features.
- Enhancing dialect detection by addressing the issue of language bivalency across Arabic dialects.

Arabic Dialects

- (Modern) Standard Arabic – descendant of Classical Arabic
- Standard Arabic vs. Regional dialects
- Diglossic distribution of functions
- Written/Spoken dichotomy
- Code-switching

Arabic Dialects

- Continuum(s) of Regional dialects
- Main dialect groups: Maghrebi, Egyptian, Levantine, Mesopotamian, Gulf



Different Arabic varieties in the Arab world - Wikipedia
https://commons.wikimedia.org/wiki/File%3AArabic_Dialects.svg

Arabic Dialects

English	MSA	Egyptian	Jordanian
Coffee	qahwah	ʾahwah	gahweh
Sugar	sukkar	sukkar	sukkar
Camel	jamal	gamal	jamal
Giraffe	zarāfah	zarāfah	zarāfeh
Chicken	dajāj	firākh	jāj
Man	rajul	rāgil	zalameh
Happy	saʿīd	mabsūt	mabsūt
Car	sayyārah	ʿarabiyyah	sayyārah
Clothes	malābis	hudūm	ʾawāʿī
Mattress	martabah	martabah	farsheh
Grey	ramādī	ramādī	sakanī
Pink	zahrī	bambī	zahrī

What is Bivalency?

- "simultaneous membership of a given linguistic segment in more than one linguistic system in a contact setting" (Woolard 2007: 448)
- Strategic bivalency
- Written bivalency
- Common in spoken Arabic
- Even more common in written Arabic

- Opaqueness of unvoweled Arabic script

- Hegemony of standard Arabic writing system – eg. قلم not ألم

Bivalency in Written Arabic

- Example from Mejdell (2014: 273):

كتابي عن مبارك وعصره ومصره

My Book about Mubarak, his era and his Egypt

Standard Arabic reading:

kitābī ʿan Mubārak wa-ʿaṣri-hi wa-miṣri-hi

Egyptian Arabic reading:

kitābi ʿan Mubārak wi-ʿaṣr-u w-maṣr-u

Bivalency vs. Code-switching

- Code-switching: focus on divergent features.
- Bivalence: focus on convergent features.

E.g. 1 (Egy corpus):

اول مرة اشوف رئيس دولة يحشد جيوشة من اجل كرة قدم

This is the first time I see a head of state mobilising his army for [a game of] football

E.g. 2 (Glf corpus):

إش المعيار الذي يحكم من خلاله

What is the criterion that is used to judge...

Problem

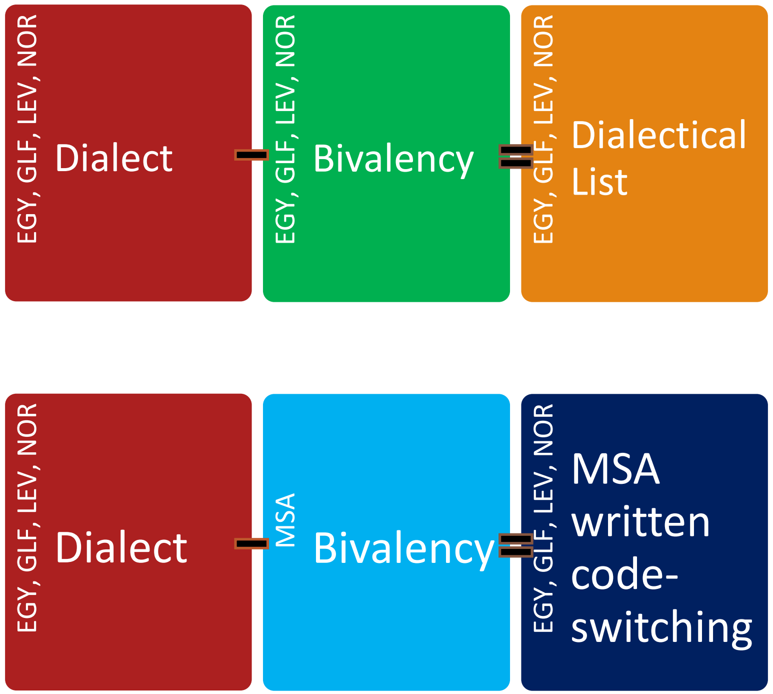
- Identifying written dialects is a hard task even for Arabic native speakers.
- The task of automatically identifying dialects is harder and classifiers trained using only n-grams will perform poorly when tested on new unseen data.
- It requires significant amounts of annotated training data.
- Currently available dialect datasets do not exceed a few hundred thousand sentences.
- Therefore features other than word n-grams are needed.

Methodology

- Use Machine Learning Classifiers
- Apply a novel approach of detecting bivalent words between dialects.
- We call this: Subtractive Bivalency Profiling (SBP).
- In addition to SBP we also incorporate grammatical and stylistic features.

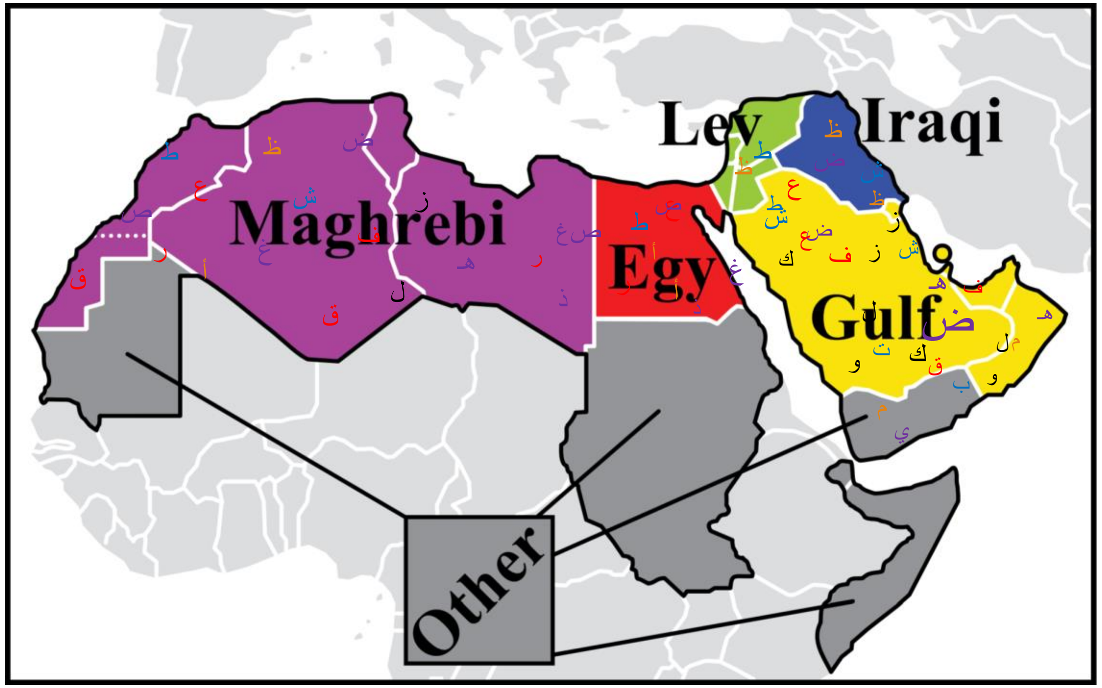
Subtractive Bivalency Profiling (SBP)

- SBP to study closeness and homogeneity between classes.
- Analysing the dataset we found dialect speakers tend to use MSA when writing in their own dialect.
- This is more common in formal conversations (e.g. Political debates)
- We used bivalency and written code-switching to create dialect-specific frequency lists of two types:
  - A) Dialect Bivalency list.
    - Identifying bivalent words between dialects aside from MSA leaving us with more fine grained dialectical lists.
  - B) MSA written code-switching list.
    - Finding bivalent words between dialects and MSA (MSA written code switching)



Dataset

- Four Arabic Dialects: Egyptian (EGY), Levant (LAV), Gulf (GLF), and North (NOR) in addition to Modern Standard Arabic (MSA).
- NOR: http://www.tunisiya.org/
- Filtering Arabic Commentary Dataset (AOC) (Zaidan and Callison-Burch, 2014)*.
- AOC used crowdsourcing (Mechanical Turk).



---

* Zaidan, O. F. and Callison-Burch, C. (2014). Arabic dialect identification. Comput. Linguist., 40(1):171–202.

Machine Learning

- We trained different text classifiers using four algorithms: Naïve Bayes, Support Vector Machine (SVM), k–Nearest Neighbor (KNN) and Decision Trees (J48).

- We divided the data into training and testing

Training Data (~70%)

Dialect Label	Sentences	Words
GLF	2,546	65,752
LAV	2,463	67,976
MSA	3,731	49,985
NOR	3,693	53,204
EGY	4,061	118,152
Total	16,494	355,069

Testing Data (~30%)

Dialect Label	Sentences	Words
GLF	1,741	40,768
LAV	1,092	17,070
MSA	1,056	18,215
NOR	1,600	29,759
EGY	1,584	33,066
Total	7,073	138,878

Baselines

- Baseline_1:

A classifier that always selects the most frequent class (EGY in this case).

Accuracy: 24%

- Baseline_2:

A word-level n-gram features classifier; selecting unigram, bigram and trigram contiguous words using Naïve Bayes classifier.

Accuracy: 52%

Feature Extraction_1

- Grammatical Features
  - POST (Stanford)
    - Tag Frequency: refers to the frequency of each tag found in the POS tagset
    - Uniqueness: refers to the number of tag types introduced in the text.
  - Function words
    - adverbs, adverbials, conjunctions, demonstratives, modals, negations, particles, prepositional, prepositions, pronouns, quantities, question and relatives function words.

Feature Extraction_2

- Stylistic Features
  - Type-Token-Ratio (TTR)
    - The ratio obtained by dividing the total number of different words (types) occurring in a text by the total number of words (tokens).
  - Readability (OSMAN) (http://drelhaj.github.io/OsmanReadability/)
    - Provides readability score between 0 (hard to read) and 100 (easy to read). In addition to syllables, hard words, complex words and Faseeh.

Feature Extraction_3

- Subtractive Bivalency Profiling (SBP)
  - Create two Frequency lists:
    - Dialect bivalency
    - MSA Written code-switching.

Feature Reduction

- Using Information Gain Ration and Feature-Group Filtering
- Reduce large number of features
- Increaser performance and classification speed.

Results / Baselines

- Baseline_1: 24% (most frequent Label: EGY)
- Baseline_2: 52% (Short sentences, High Bivalency (e.g. (رياضة, نعم، تعليم)

Results / Training

- 10-fold cross validation
- Reduced features.
- J48, SVM, Naïve Bayes and KNN
- Best machine learning algorithm c97% (J48)

Algorithm	Accuracy
J48	97.11%
SVM	91.3%
KNN	73.69%
NB	60.89%

Results / Training

- Examining Feature Groups
- Help in better split the dataset, easier for Machine to learn and classify.
- Results show SBP outperformed all other features.
- Combining SBP with Gram and Sty helps increase accuracy.

Feature(s)	J48	SVM	NB	KNN
Sty + SBP	97.11	89.74	74.46	92.98
SBP + Gram	97.08	90.50	61.04	77.75
SBP	97.07	89.10	75.06	96.39
Sty + Gram	51.20	54.35	41.48	46.78
Gram	50.56	52.56	40.47	46.39
Sty	44.87	29.12	32.78	42.62

Results / Testing

- Separate unseen dataset
- Classifiers testing results outperformed the two baselines.
- Using n-gram on new unseen data didn’t work well as expected.
- SBP combined with Sty and Gram features helps the classifier identify dialects even when there are new vocabulary that the classifier has not seen before.

Feature(s)	J48	SVM	NB	KNN
Sty + SBP	**64.31**	59.64	50.82	63.51
SBP + Gram	**64.28**	59.52	50.78	63.40
SBP	63.84	58.56	51.09	**66.32**
All	63.64	62.99	43.29	54.57
Sty + Gram	51.31	53.24	39.92	43.48
Gram	50.38	52.49	38.92	42.17
n-gram	42.78	31.02	32.36	38.86
Sty	41.16	33.09	27.15	31.45

Conclusion

- Built machine learning classifiers to automatically detect Arabic dialects.
- New method SBP helps classifiers split dataset of different and close Arabic dialects.
- SBP outperformed all other individual features.
- Results improve when combining SBP with other Gram and Sty features.
- Code available online:
- https://github.com/drelhaj/ArabicDialects

Questions