مؤسسة قطر
Qatar Foundation

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute

Verifiably Effective Arabic Dialect Identification

Kareem Darwish, Hassan Sajjad, Hamdy Mubarak

Background

Arabic Language

Modern Standard Arabic (MSA) is the lingua franca of the Arab world

Arabic dialects are generally used in daily interactions and in social media

Dialects differ from MSA and from each other. Differences are: **lexical**, **morphological**, **phonological** and **syntactic**

Arabic Dialect Identification

Previous work:

- Claims that word unigram models are sufficient and effective for the dialect identification

Problems:

- Homogeneity of training and test sets (in topics and jargon) can be attributed to "good" dialect identification
- Unigram models do not learn dialect specific linguistic phenomena

Focus

- Identify phenomena in Egyptian Arabic (ARZ) that set it apart from MSA, namely:
  - Dialectal words, Morphological differences, Letter substitution, and syntactic differences
- Present methods to handle such peculiarities

Egyptian Arabic

Lexical Differences

ARZ specific words stem from:

- Archaic Arabic words: **"$nTp"** (bag)
- Fusing multiple words together by concatenating and dropping letters: **"mA Elyh $y"** → **"mEl$"** (no worry)
- Non-standard spelling: **"<sbE"** → **"SAbE"** (finger)

Morphological Differences

Some examples include:

- Addition of the letter "b" in front of verb in present tense: **"ylEb"** → **"bylEb"** (he plays)
- Use of letters "H" or "h" instead of "s" for future tense: **"sylEb"** → **"hylEb"** (he will play)
- Replacement of a short vowel with a long vowel in imperative verbs: **"qul"** → **"qwl"** (say)

Phonological/Letter Substitution

Common letter substitutions:

- "v" → "t": **"kvyr"** → **"ktyr"** (a lot)
- "}" → "y": **"b}r"** → **"byr"** (well)
- "*" → "d": **"xu*"** → **"xud"** (take)
- "D" → "Z": **"DAbT"** → **"ZAbT"** (officer)
- Removal of trailing " ' ": **"AlsmA' "** → **"AlsmA"** (the sky)

Detecting Differences

Lexical Coverage

Create a dictionary of dialectal words (**DICT1**)

- Take Egyptian side of the LDC2012T09 corpus
- Sort unigrams by frequency
- Manually identify top 1300 dialectal words

Morphological Phenomenon

Employ three methods:

1. Unsupervised morphology induction using Morfessor
   - Segmented the training and test set
2. Morphological rules
   - Developed 15 morphological rules to segment ARZ
3. Morphological generator
   - Enumerated 200 morphological patterns that derive dialectal verbs
   - Resulting list is of 94k verb forms (**DICT2**)

Phonological variations

Generate possible Arabic stem (diacritized)

- Use root list (Darwish, 2002) and 605 morphological patterns (Darwish et al., 2014)

Check the generated stems against a large diacritized Arabic corpus

- If generated words contained the letters "v", "}", "*", and "D", apply dialectal letter substitution
- Final list is of 8k words (**DICT3**)

Evaluation

Dataset

Training set:

- ARZ -> Egyptian side of the LDC2012T09 corpus
- MSA -> Arabic side of the English/Arabic parallel corpus from IWSLT

Test set:

- Collected Arabic tweets from Twitter during March 2014
- Filtered based on user location set to Egypt (880k)
- Randomly selected 2k tweets and manually annotated them to obtain 350 ARZ and 350 MSA tweets

Classification Model

- Random Forest ensemble classifier
- Baseline features:
  - Word unigrams, bigrams, trigrams (WRD)
  - Character unigram to 5-gram (CHAR)
- Lex features based on DICT{1-3}

Results

Baseline: trained on the full training data (ARZ + MSA)

S-morfessor: training data is segmented using morfessor

S-rule: training data is segmented using morphological rules

LEX: concatenation of three lists (1.3k, 94k, 8k). We count the number of words in a tweet that exist in the word lists and used it as a standalone feature

SYS	WRD	CHR	WRD+CHR	BEST+LEX
Baseline	53.0	74.0	83.3	84.7
S-morfessor	72.0	88.0	62.1	89.3
S-rule	53.9	85.9	85.9	90.1

Classification results using only dictionaries as features in the model

SYS	DICT1	+DICT2	+DICT3
S-lex	93.6	94.6	94.4