Verifiably Effective Arabic Dialect Identification

Kareem Darwish, Hassan Sajjad, Hamdy Mubarak
Qatar Computing Research Institute
Qatar Foundation
{kdarwish,hsajjad,hmubarak}@qf.org.qa

Abstract

Several recent papers on Arabic dialect identification have hinted that using a word unigram model is sufficient and effective for the task. However, most previous work was done on a standard fairly homogeneous dataset of dialectal user comments. In this paper, we show that training on the standard dataset does not generalize, because a unigram model may be tuned to topics in the comments and does not capture the distinguishing features of dialects. We show that effective dialect identification requires that we account for the distinguishing lexical, morphological, and phonological phenomena of dialects. We show that accounting for such can improve dialect detection accuracy by nearly 10% absolute.

1 Introduction

Modern Standard Arabic (MSA) is the lingua franca of the so-called Arab world, which includes northern Africa, the Arabian Peninsula, and Mesopotamia. However, Arabic speakers generally use dramatically different languages (or dialects) in daily interactions and in social media. These dialects may differ in vocabulary, morphology, and spelling from MSA and most do not have standard spellings. There is often large lexical overlap between dialects and MSA. Performing proper Arabic dialect identification may positively impact many Natural Language Processing (NLP) application. For example, transcribing dialectal speech or automatically translating into a particular dialect would be aided by the use of targeted language models that are trained on texts in that dialect.

This has led to recent interest in automatic identification of different Arabic dialects (Elfardy et al., 2013; Cotterell et al., 2014; Zaidan et al., 2014). Though previous work (Cotterell et al., 2014) have reported high accuracies for dialect identification using word unigram model, which implies that this is a solved problem, we argue that the problem is far from being solved. The reason for this assertion stems from the fact that the available dialectal data is drawn from singular sources, namely online news sites, for each dialect. This is problematic because comments on singular news site are likely to have some homogeneity in topics and jargon. Such homogeneity has caused fairly simple classification techniques that use word unigrams and character n-grams to yield very high identification accuracies. Perhaps, this can be attributed to topical similarity and not just differences between dialects. To showcase this, we trained a classifier using the best reported methods, and we tested the classifier on a new test set of 700 tweets, with dialectal Egyptian (ARZ) and MSA tweets, which led to a low accuracy of 83.3%. We also sorted words in the ARZ part from our training dataset by how much they discriminate between ARZ and MSA (using mutual information) and indeed many of the top words were in fact MSA words.

There seems to be a necessity to identify lexical and linguistic features that discriminate between MSA and different dialects. In this paper, we highlight some such features that help in separating between MSA and ARZ. We identify common ARZ words that do not overlap with MSA and identify specific linguistic phenomena that exist in ARZ, and not MSA, such as morphological patterns, word concatenations, and verb negation constructs (Section 3). We also devise methods for capturing the linguistic phenomena, and we use the appearance of such phenomena as features (Section 4). Further, we show the positive impact of using the new features in identifying ARZ (Section 5).

2 Previous Work

Previous work on Arabic dialect identification uses n-gram based features at both word-level and character-level to identify dialectal sentences (Elfardy et al., 2013; Cotterell et al., 2014; Zaidan et al., 2011; Zaidan et al., 2014). Zaidan et al. (2011) created a dataset of dialectal Arabic. They performed cross-validation experiments for dialect identification using word n-gram based features. Elfardy et al. (2013) built a system to distinguish between ARZ and MSA. They used word n-gram features combined with core (token-based and perplexity-based features) and meta features for training. Their system showed a 5% improvement over the system of Zaidan et al. (2011). Later, Zaidan et al. (2014) used several word n-gram based and character n-gram based features for dialect identification. The system trained on word unigram-based feature performed the best with character five-gram-based feature being second best. A similar result is shown by Cotterell et al. (2014) where word unigram model performs

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1465–1468,
October 25-29, 2014, Doha, Qatar.

the best.

All of the previous work except Cotterell et al. (2014)[1] evaluate their systems using cross-validation. These models heavily rely on the coverage of training data to achieve better identification. This limits the robustness of identification to genres inline with the training data.

Language identification is a related area to dialect identification. It has raised some of the issues which we discussed in this paper in the context of dialect identification. Lui et al. (2011) showed that in-domain language identification performs better than cross domain language identification. Tiedemann et al. (2012) argued that the linguistic understanding of the differences between languages can lead to a better language identification system. kilgarriff (2001) discussed the differences between datasets as a poor representation of differences between dialects of English.

In this paper, we exploit the linguistic phenomena that are specific to Arabic dialects to show that they produce significant improvements in accuracy. We show that this also helps in achieving high quality cross-domain dialect identification system.

3 Dialectal Egyptian Phenomena

There are several phenomena in ARZ that set it apart from MSA. Some of them are as follows:

Dialectal words: ARZ uses unique words that do not overlap with MSA and may not overlap with other dialects. Some of the common ARZ words are: "zy" (like), "kdh" (like this), and "Azyk" (how are you) [2]. These dialectal terms stem from the following:

• Using proper Arabic words that are rarely used in MSA such as "$nTp" (bag) and "n$wf" (we see).

• Fusing multiple words together by concatenating and dropping letters such as the word "mEl$" (no worry), which is a fusion of "mA Elyh $y' ".

• Using non-standard spelling of words such as "SAbE" (finger) instead of "<sbE" in MSA. Consequently, broken plurals may also be non-standard.

• using non-Arabic words such as "<y$Arb" (scarf), which is transliterated from the French écharpe.

• altering the forms of some pronouns such as the feminine second person pronoun from "k" to "ky", the second person plural pronoun "tm" to "tw", and the object pronoun "km" to "kw".

Morphological differences: ARZ makes use of particular morphological patterns that do not exist in MSA and often alters some morphological constructs. Some examples include:

• Adding the letter "b" in front of verb in present tense. Ex. MSA: "ylEb" (he plays) → EG: "bylEb".

• Using the letters "H" or "h", instead of "s", to indicate future tense. Ex. MSA: "sylEb" (he will play) → EG: "hylEb" or "HylEb".

• Adding the letters "At" to passive past tense verbs. Ex. MSA: "luEiba" (was played) → "AtlaEab".

• Adding the letters "m" or "mA" before the verb and "$" or "$y" after the verb to express negation. Ex. MSA: "lm ylEb" (he did not play) → "mlEb$".

• the merging of verbs and prepositional phrases of the form (to-pronoun) that follow it. Ex. MSA: "ylEb lh" (he plays for/to him) → "bylEblh".

• Replacing a short vowel with a long vowel in imperative verbs that are derived from hollow roots. Ex. MSA: "qul" (say) → "qwl".

Letter substitution: in ARZ the following letter substitutions are common:

• "v" → "t". Ex. MSA: "kvyr" (a lot) → EG: "ktyr".

• "}" → "y". Ex. MSA: "b}r" (well) → "byr".

• Trailing "y" → "Y". Ex. MSA: "Hqy" (my right) → "HqY".

• "*" → "d". Ex. MSA: "xu*" (take) → "xud".

• middle or trailing ">" → "A". Ex. MSA: "f>r" (mouse) → "fAr".

• "D" → "Z". Ex. MSA: "DAbT" (officer) → "ZAbT".

• "Z" → "D". Ex. MSA: "Zhr" (back) → "Dhr".

• Middle "|" → "yA". Ex. MSA: "ml|n" (full) → "mlyAn".

• Removal of trailing " ' ". Ex. MSA: "AlsmA' " (the sky) → "AlsmA".

Syntactic differences: some of the following phenomena are generally observed:

• Common use of masculine plural or singular noun forms instead dual and feminine plural. Ex. MSA "jnyhyn" (two pounds) → EG: "Atnyn jnyh".

• Dropping some articles and preposition in some syntactic constructs. For example, the preposition "<lY" (to) in ">nA rAyH <lY Al$gl" (I am going to work) is typically dropped. Also, the particle ">n" (to) is dropped in the sentence ">nA mHtAj >n >nAm" (I need to sleep).

• Using only one form of noun and verb suffixes such as "yn" instead of "wn" and "wA" instead of "wn" respectively. Also, so-called "five nouns", are used in only one form (ex. ">bw" (father of) instead of ">bA" or ">by").

4 Detecting Dialectal Peculiarities

ARZ is different from MSA lexically, morphologically, phonetically, and syntactically. Here, we present methods to handle such peculiarities. We chose not to handle syntactic differences, because they may be captured using word n-gram models.

To capture lexical variations, we extracted and sorted by frequency all the unigrams from the Egyptian side of the LDC2012T09 corpus (Zbib et al., 2012), which has ≈ 38k Egyptian-English parallel sentences. A linguist was tasked with manually reviewing the words from the top until 1,300 dialectal words were found. Some of the words on the list included dialectal words, commonly used foreign words, words that exhibit morphological variations, and others with letter substitution.

[1] Zaidan et al. (2014) applied their classifier to a different genre but did not evaluate it's performance.

[2] Buckwalter encoding is used throughout the paper.

For morphological phenomenon, we employed three methods, namely:

• **Unsupervised Morphology Induction:** We employed the unsupervised morpheme segmentation tool, Morfessor (Virpioja et al., 2013). It is a data driven tool that automatically learns morphemes from data in an unsupervised fashion. We used the trained model to segment the training and test sets.

• **Morphological Rules:** In contrast to Morfessor, we developed only 15 morphological rules (based on the analysis proposed in Section 3) to segment ARZ text. These rules would separate prefixes and suffixes like a light stemmer. Example rules would segment a leading "b" and segment a combination of a leading "m" and trailing "$".

• **Morphological Generator:** For morphological generation, we enumerated a list of ≈ 200 morphological patterns that derive dialectal verbs from Arabic roots. One such pattern is ytCCC that would generate the dialectal verb-form ytktb (to be written) from the root "ktb". We used the root list that is distributed with Sebawai (Darwish, 2002). We also expanded the list by attaching negation affixes and pronouns. We retained generated word forms that: a) exist in a large corpus of 63 million Arabic tweets from 2012 with more than 1 billion tokens; and b) don't appear in a large MSA corpus of 10 years worth of Aljazeera articles containing 114 million tokens [3]. The resulting list included 94k verb surface forms such as "mbyEmlhA$" (he does not do it).

For phonological differences, we used a morphological generator that makes use of the aforementioned root list and an inventory of ≈ 605 morphological patterns (with diacritization) to generate possible Arabic stems. The generated stems with their diacritics were checked against a large diacritized Arabic corpus containing more than 200 million diacritized words [4]. If generated words contained the letters "v", "}", "*", and "D", we used the aforementioned letter substitutions. We retained words that exist in the large tweet corpus but not in the Aljazeera corpus. The list contained 8k surface forms.

5 Evaluation Setup

Dataset: We performed dialect identification experiment for ARZ and MSA. For ARZ, we used the Egyptian side of the LDC2012T09 corpus (Zbib et al., 2012) [5]. For MSA, we used the Arabic side of the English/Arabic parallel corpus from the International Workshop on Arabic Language Translation[6] which consists of ≈ 150k sentences. For testing, we constructed an evaluation set that is markedly different from the training set. We crawled Arabic tweets from Twitter during March 2014 and selected those where user location was set to Egypt or a geographic location within Egypt, leading to 880k tweets. We randomly selected 2k tweets, and we manually annotated them as ARZ, MSA, or neither until we obtained 350 ARZ and 350 MSA tweets. We used these tweets for testing. We plan to release the tweet ID's and our annotations. We preprocessed the training and test sets using the method described by Darwish et al. (2012), which includes performing letter and word normalizations, and segmented all data using an open-source MSA word segmentor (Darwish et al., 2012). We also removed punctuations, hashtags, and name mentions from the test set. We used a Random Forest (RF) ensemble classifier that generates many decision trees, each of which is trained on a subset of the features.[7] We used the RF implementation in Weka (Breiman, 2001).

5.1 Classification Runs

Baseline *BL*: In our baseline experiments, we used word unigram, bigram, and trigram models and character unigram to 5-gram models as features. We first performed a cross-validation experiment using ARZ and MSA training sets. The classifier achieved fairly high results (+95%) which are much higher than the results mentioned in the literature. This could be due in part to the fact that we were doing ARZ-MSA classification instead of multi-dialect classification and MSA data is fairly different in genre from ARZ data. We did not further discuss these results. This experiment was a sanity check to see how does in-domain dialect identification perform.

Later, we trained the RF classifier on the complete training set using word n-gram features (WRD), character n-gram features (CHAR), or both (BOTH) and tested it on the tweets test set. We referred to this system as *BL* later on.

Dialectal Egyptian Lexicon S_{lex}: As mentioned earlier, we constructed three word lists containing 1,300 manually reviewed ARZ words (MAN), 94k dialectal verbs (VERB), and 8k words with letter substitutions (SUBT). Using the lists, we counted the number of words in a tweet that exist in the word lists and used it as a standalone feature for classifications. LEX refers to concatenation of all three lists.

Morphological Features: For S_{mrph}, we trained Morfessor separately on the MSA and Egyptian training data and applied to the same training data for segmentation. For S_{rule}, we segmented Egyptian part of the training data using the morphological rules mentioned in Section 4. For both, word and character n-gram features were calculated from the segmented data and the

[3] http://aljazeera.net

[4] http://www.sh.rewayat2.com

[5] We did not use the Arabic Online Commentary data (Zaidan et al., 2011) as annotations were often not reliable.

[6] https://wit3.fbk.eu/mt.php?release=2013-01

[7] We tried also the multi-class Bayesian classifier and SVM classifier. SVM classifier had comparable results with Random Forest classifier. However, it was very slow. So, we decided to go with Random Forest classifier for the rest of the experiments.

SYS	WRD	CHR	BOTH	BEST+LEX
BL	53.0	74.0	83.3	84.7
S_{mrph}	72.0	88.0	62.1	89.3
S_{rule}	53.9	85.9	85.9	90.1

Table 1: Dialect identification accuracy using various classification settings: only word-based (WRD), character-based (CHAR), and both features. BEST+LEX is built on the best feature of that system plus a feature built on the concatenation of all lists

SYS	MAN	+VERB	+SUBT
S_{lex}	93.6	94.6	94.4

Table 2: Accuracy of the dialect identification system with the addition of various types of lexicon

classifier was trained on them and tested on the tweet test set.

5.2 Results

Table 1 summarizes the results. Unlike results in the literature, character-based n-gram features outperformed word-based n-gram features, as they seemed to better generalize to the new test set, where lexical overlap between the training and test sets was low. Except for S_{mrph}, adding both character and word n-gram features led to improved results. We observed that Morfessor over-segmented the text, which in turns created small character segments and enabled the character-based language model to learn the phenomenon inherit in a word. The baseline system achieved an accuracy of 84.7% when combined with the S_{lex} feature. Combining S_{mrph} and S_{rule} features with the S_{lex} feature led to further improvement. However, as shown in Table 2, using the S_{lex} feature alone with the MAN and VERB lists led to the best results (94.6%), outperforming using all other features either alone or in combination. This suggests that having a clean list of dialectal words that cover common dialectal phenomena is more effective than using word and character n-grams. It also highlights the shortcomings of using a homogeneous training set where word unigrams could be capturing topical cues along with dialectal ones.

6 Conclusion

In this paper, we identified lexical, morphological, phonological, and syntactic features that help distinguish between dialectal Egyptian and MSA. Given the substantial lexical overlap between dialectal Egyptian and MSA, targeting words that exhibit distinguishing traits is essential to proper dialect identification. We used some of these features for dialect detection leading to nearly 10% (absolute) improvement in classification accuracy. We plan to extend our work to other dialects.

References

Leo Breiman. 2001. Random Forests. Machine Learning. 45(1):5-32.

Ryan Cotterell, Chris Callison-Burch. 2014. A Multi-Dialect, Multi-Genre Corpus of Informal Written Arabic. LREC-2014, pages 241–245.

Kareem Darwish. 2002. Building a shallow morphological analyzer in one day. In Proceedings of the ACL-2002 Workshop on Computational Approaches to Semitic Languages.

Kareem Darwish, Walid Magdy, Ahmed Mourad. 2012. Language Processing for Arabic Microblog Retrieval. CIKM-2012, pages 2427–2430.

Kareem Darwish, Ahmed Abdelali, Hamdy Mubarak. 2014. Using Stem-Templates to improve Arabic POS and Gender/Number Tagging. LREC-2014.

Heba Elfardy, Mona Diab. 2013. Sentence Level Dialect Identification in Arabic. ACL-2013, pages 456–461.

Sami Virpioja, Peter Smit, Stig-Arne Grnroos, and Mikko Kurimo. 2013. Morfessor 2.0: Python Implementation and Extensions for Morfessor Baseline. Aalto University publication series SCIENCE + TECHNOLOGY, 25/2013. Aalto University, Helsinki, 2013.

Omar F. Zaidan, Chris Callison-Burch. 2011. The Arabic Online Commentary Dataset: An Annotated Dataset of Informal Arabic with High Dialectal Content. ACL-11, pages 37–41.

Omar F. Zaidan, Chris Callison-Burch. 2014. Arabic Dialect Identification. CL-11, 52(1).

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan, Chris Callison-Burch. 2012. Machine translation of Arabic dialects. NAACL-2012, pages 49–59.

Marco Lui and Timothy Baldwin. 2011. Cross-domain feature selection for language identification. IJCNLP-2011, page 553–561.

Jörg Tiedemann and Nikola Ljubesic. 2012. Efficient discrimination between closely related languages. COLING-2012, 2619–2634.

Adam Kilgarriff. 2001. Comparing corpora. CL-01, 6(1).