# Arabic Dialect Identification using a Parallel Multidialectal Corpus

Shervin Malmasi
*Centre for Language Technology*
*Macquarie University*
*Sydney, NSW, Australia*
shervin.malmasi@mq.edu.au

Eshrag Refaee
*Interaction Lab*
*Heriot-Watt University*
*EH144AS Edinburgh, UK*
eaar1@hw.ac.uk

Mark Dras
*Centre for Language Technology*
*Macquarie University*
*Sydney, NSW, Australia*
mark.dras@mq.edu.au

***Abstract*—We present a study on sentence-level Arabic Dialect Identification using the newly developed Multidialectal Parallel Corpus of Arabic (MPCA) – the first experiments on such data. Using a set of surface features based on characters and words, we conduct three experiments with a linear Support Vector Machine classifier and a meta-classifier using stacked generalization – a method not previously applied for this task. We first conduct a 6-way multi-dialect classification task in the first experiment, achieving $74\%$ accuracy against a random baseline of $16.7\%$ and demonstrating that meta-classifiers can large performance increases over single classifiers. The second experiment investigates pairwise binary dialect classification within the corpus, yielding results as high as $94\%$, but also highlighting poorer results between closely related dialects such as Palestinian and Jordanian ($76\%$). Our final experiment conducts cross-corpus evaluation on the widely used Arabic Online Commentary (AOC) dataset and demonstrates that despite differing greatly in size and content, models trained with the MPCA generalize to the AOC, and vice versa. Using only $2{,}000$ sentences from the MPCA, we classify over 26k sentences from the radically different AOC dataset with $74\%$ accuracy. We also use this data to classify a new dataset of MSA and Egyptian Arabic tweets with $97\%$ accuracy. We find that character *n*-grams are a very informative feature for this task, in both within- and cross-corpus settings. Contrary to previous results, they outperform word n-grams in several experiments here. Several directions for future work are outlined.**

***Keywords*-Arabic Dialects; Automatic Dialect Identification; Parallel Corpus; Text Classification;**

## I. Introduction

The Arabic language, the official language of more than 20 countries, is comprised of many regional dialects with the Modern Standard Arabic (MSA) variety having the role of a common dialect across the Arabic-speaking population.

Arabic is a morphologically sophisticated language with many morphemes that can appear as prefixes, suffixes or even circumfixes. These mark grammatical information including case, number, gender, and definiteness, amongst others. This leads to a sophisticated morphotactic system. Its orthography is very different to English with right-to-left text that uses connective letters. Moreover, this is further complicated due to the presence of word elongation, common ligatures, zero-width diacritics and allographic variants – resulting in a degree of orthographic ambiguity. All of these properties pose a challenge for NLP [1].

These varieties of Dialectal Arabic (DA) and MSA vary among each other across the major linguistic subsystems, including phonology, morphology, orthography and to a lesser degree, syntax. For written Arabic – the focus of the present work – the greatest differences exist in lexicon, morphology and orthography.[1]

The availability of robust and accurate dialect identification models can be of great benefit to Arabic NLP tasks and this has fuelled the recent drive in investigating Arabic Dialect Identification (ADI). Potential applications of ADI are:

- As a useful preprocessing step for other tasks, such as statistical machine translation. Here they could be used to determine the most suitable dialect-specific models to be used for the input data.
- For building dialect-to-dialect or dialect-to-MSA lexicons, such as the work presented in [3] which uses information mined from the web to induce such lexicons. Another example is [4], which presents an electronic three-way lexicon, Tharwa, comprising Dialectal Arabic, Modern Standard Arabic and English correspondents. This can be helpful in linguistic research and can also aid learners who are studying a specific dialect.
- The generated dialectal mappings can be used in Natural Language Generation (NLG) for selecting the appropriate lexeme or morphological inflection using dialect-based word choice criteria [5]. This is useful for tailoring the output for a particular dialect or region.
- As a tool for Authorship profiling and attribution in the forensic linguistics domain.
- In an Information Retrieval context this method can be used to filter documents according to their dialect. Practical applications include, *inter alia*, filtering of news articles or search engine results according to user preferences.

These potential applications have generated recent interest in the task of automatically identifying the Arabic dialect of given texts.

[1] See [2, §2] for a more detailed discussion.

The rise of microblogs and social media have also spurred researchers to investigate NLP tasks at smaller scales.[2] In this spirit, our work also focuses on dialect identification at the sentence level. This is a more challenging task due to sparsity and the amount of information available per item.

There have been concerns that the word unigram models used in previous research are affected by topic bias, as discussed in section II. We attempt to investigate this by running the first ADI experiments using a parallel corpus that is inherently balanced by topic. We further investigate this issue by using cross-corpus evaluation on previous datasets.

Another limitation with previous work is that almost all studies have distinguished between only two classes. There are likely to be many more classes in practical application and we perform a 6-way dialect identification experiment to evaluate our system.

In sum, the broad aim of the present study is to assess the utility of surface features for multi-class Arabic dialect identification on a parallel corpus that is balanced by topic and size across classes. In addition to the standard single-classifier setup, we also experiment with a meta-classifier approach which to the best of our knowledge, has not hitherto been applied to dialect identification. Finally, we also aim to evaluate the generalizability of models trained on specific datasets through cross-corpus evaluation.

## II. Background

A number of recent works have attempted to perform automatic dialect identification of Arabic texts.[3] In this section we briefly review some of this previous work.

The Arabic Online Commentary (AOC) Dataset, a 52m word monolingual dataset rich in dialectal content was developed in [7]. A total of 108k sentences were labelled for dialect and used for automatic dialect identification. The authors take a Language Model approach and report an accuracy of 69.4% on a 4-way classification task (MSA and three dialects). On a binary classification between Egyptian Arabic and MSA, an accuracy of 80.9% was reported.

Similarly, [8] also take a supervised learning approach to sentence-level binary classification of Egyptian Arabic and MSA data from the AOC dataset. They utilize a Naive Bayes classifier along with word $n$-grams combined with core (token- and perplexity-based features) and meta features. Their system achieves as accuracy of 85.5%, an improvement over the 80.9% reported in [7] for the same task.

In [9] the authors extend their previous work on the AOC dataset to include letter and word features. They report that word unigrams are the best performing feature. They report an accuracy of 81.0% on a 4-way classification task (MSA vs. three dialects). For Egyptian Arabic *vs.* MSA, an accuracy of 87.9% is reported.

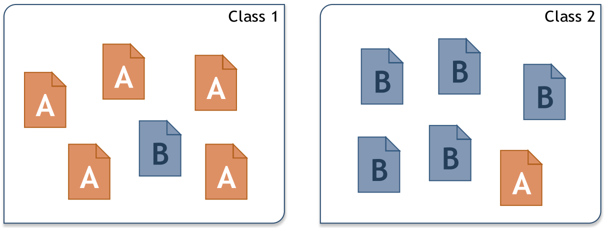

Figure 1. An example of a dataset that is not balanced by topic: class 1 contains mostly documents from topic A while class 2 is dominated by texts from topic B. Here, a learning algorithm may distinguish the classes through other confounding variables related to topic.

Also focusing on the Egyptian-MSA binary classification task, [10] use a range of lexical and morphological features to classify 700 tweets with 95% accuracy against a 50% baseline. This set of 700 tweets was constructed specifically for evaluation and is different to the training data. A total of 880k Arabic tweets were crawled from Twitter in March 2014 and this manually selected subset of 350 Egyptian and MSA tweets were selected to create the test set. We also use this test set in our cross-corpus evaluation.

Much of the previous work in Arabic Dialect Identification has used the Arabic Online Commentary (AOC) dataset. This dataset is not controlled for topic and the the number of sentences across the different dialects are not balanced. The authors of [10] state that since the data are sourced from singular sources, these models may not generalize to other data as they implicitly capture topical cues and are thus susceptible to *topic bias*. This is a claim that we aim to assess in this work.

*Topic bias* can occur as a result of the themes or topics of the texts to be classified not being evenly distributed across the classes, leading to correlations between classes and topics [11] [12]. For example, if in our training data all the texts written by Egyptian Arabic speakers are on topic A, while all the MSA texts refer to topic B, then we have implicitly trained our classifier on the topics as well. In this case the classifier learns to distinguish our target variable through another confounding variable. This concept is illustrated in Figure 1.

## III. Data

For our experiments we use the Multidialectal Parallel Corpus of Arabic (MPCA) which was recently released by [2]. They present the first parallel multidialectal Arabic dataset, comprised of 2,000 sentences in Modern Standard Standard Arabic, five regional dialects, as well as English. This data was transcribed native-speaker translators who translated the source sentences into their dialect. This corpus is a valuable resource as such parallel cross-dialect translations do not occur naturally and are useful for studying dialectal differences while controlling for topic bias. Moreover, as this data has been transcribed, it is not prone to the issues found in noisy social media or web crawled data.

[2] e.g. on short texts such as Tweets, SMS messages and status updates.

[3] Spoken Arabic dialect identification is a another area of research, as discussed in [6].

| Dialect/Language | Example |
|---|---|
| **English** | *Because you are a personality that I can not describe.* |
| **Modern Standard Arabic** | لأنك شخصية لا أستطيع وصفها.<br>*lÂnk šxSyħ lA ÂstTyς wSfhA.* |
| **Egyptian Arabic** | لأنك شخصية **وبجد** مش هعرف أوصفها.<br>*lÂnk šxSyħ wbjd mš hςrf ÂwSfhA.* |
| **Syrian Arabic** | لأنك شخصية **وعنجد** ما رح أعرف أوصفها.<br>*lÂnk šxSyħ wςnjd mA rH Âςrf ÂwSfhA.* |
| **Jordanian Arabic** | انت **جد** شخصية مستحيل اقدر اوصفه<br>*Ant jd šxSyħ mstHyl Aqdr AwSfhA.* |
| **Palestinian Arabic** | **عن جد** ماشاء اللّه عليك شخصيتك ما بتنوصف.<br>*ςn jd mA šA' Allh ςlyk šxSytk mA btnwSf.* |
| **Tunisian Arabic** | على خاطرك شخصية **بلحق** منجمش نوصفها.<br>*ςlý xATrk šxSyħ blHq mnjmš nwSfhA.* |

Figure 2. A comparison of the translations for one sentence in the Multidialectal Parallel Arabic Corpus. We use the six Arabic dialects in our experiments.

The corpus covers seven dialects/languages: Modern Standard Arabic (MSA), English (EN), Egyptian (EG), Tunisian (TN), Syrian (SY), Jordanian (JO) and Palestinian (PA). An example sentence is shown in Figure 2, which highlights the wide ranging differences among the dialects. We use $1{,}000$ sentences[4] from the Arabic data for our experiments, excluding the English translations.

In this work we also explore cross-corpus evaluation and use AOC and Egyptian-MSA tweet datasets, both described in the previous section, to test our system.

## IV. Methodology

We take a supervised classification approach for this task, similar to previous research. Our features, classifier and evaluation method are described in this section.

### A. Features

We employ two lexical surface feature types for this task, as described below. We do not perform any preprocessing steps (e.g. tokenization or orthography normalization) prior to feature extraction.

*Character $n$-grams:* This is a sub-word feature that uses the constituent characters that make up the whole text. When used as $n$-grams, the features are $n$-character slices of the text. From a linguistic point of view, the substrings captured by this feature, depending on the order, can implicitly capture various sub-lexical features including single letters, phonemes, syllables, morphemes and suffixes.

*Word $n$-grams:* The surface forms of words can be used as a feature for classification. Each unique word may be used as a feature (i.e. unigrams), but the use of bigram distributions is also common. In this scenario, the $n$-grams are extracted along with their distributions.

The features frequencies are weighted using the *tf-idf* weighting scheme. This choice is based on our preliminary experiments showing that they outperformed a binary feature representation.

### B. Classifier

We use a linear Support Vector Machine to perform multi-class classification in our experiments. In particular, we use the LIBLINEAR[5] SVM package [13] which has been shown to be efficient for text classification problems with large numbers of features and documents. We use cross-validation to optimize the SVM's $C$ hyperparameter.

Ensemble classifiers have been found to be useful in other multi-class text classification tasks such as Native Language Identification [14] [15]. In this work we also experiment with a stacked generalization model [16]. This is done through creating an ensemble of classifiers by training a single linear SVM classifier for each feature type and using the class probability outputs from each of these classifiers to train a higher level classifier. This meta-classifier, also a linear SVM, may be able to map the outputs from the lower level classifiers to their true labels by learning patterns such as certain classifiers being more likely to misclassify some classes [17, §3.6].

### C. Evaluation

We report our results as classification accuracy under cross-validation. We experiment with two types of cross-validation.

Consistent with most previous studies, we use $k$-fold cross-validation, with $k = 10$. For creating our folds, we employ stratified cross-validation which aims to ensure that the proportion of classes within each partition is equal [18].

The accuracy estimated by $k$-fold cross-validation is a variable value that depends on the randomly chosen splits of the data. To reduce the variability introduced by this random splitting we also experiment with Leave-one-out (LOO) cross-validation where each data point is predicted by a learner trained on every other data point.[6]

No previous baselines are available here as this is the first application of dialect identification to this data. We use a *random baseline* for comparison purposes. This is commonly employed in classification tasks where it is calculated by randomly assigning labels to documents. It is a good measure of overall performance in instances where the training data is evenly distributed across the classes, as is the case here. For example, an 11-class dataset has a random baseline of $\frac{1}{11} = 9.1\%$.

Additionally, we also compare against the oracle baseline used by [19]. Here the oracle correctly classifies a text if any single feature type alone correctly predicts its label. It is useful in defining an upper-bound for classification accuracy.

[4] Given that this is a parallel corpus, this is $1{,}000$ sentences per dialect, $6{,}000$ sentences in total.

[5] http://www.csie.ntu.edu.tw/%7Ecjlin/liblinear/

[6] For a dataset with $n$ items, this is equivalent to $n$-fold cross-validation.

Table I
ARABIC DIALECT IDENTIFICATION CLASSIFICATION ACCURACY FOR ALL SIX DIALECTS, USING OUR FEATURE SET. THE BEST RESULTS FOR EACH COLUMN ARE IN BOLD.

| **Feature** | **Accuracy (%)** | |
|---|---|---|
| | 10-fold CV | LOO CV |
| Random Baseline | 16.67 | 16.67 |
| Oracle Baseline | 81.21 | 81.74 |
| (1) Character unigrams | 46.12 | 46.27 |
| (2) Character bigrams | 62.16 | 62.40 |
| (3) Character trigrams | 65.26 | 65.60 |
| (4) Character 4-grams | 59.62 | 60.12 |
| (5) Word unigrams | 57.53 | 57.76 |
| (6) Word bigrams | 24.10 | 24.27 |
| All Character $n$-grams (1–4) | 65.60 | 66.10 |
| Character $1/2/3$-grams (1–3) | 66.48 | 66.63 |
| All Word $n$-grams (5–6) | 54.40 | 54.44 |
| All features combined (1–6) | 65.25 | 66.07 |
| Meta-classifier (all features) | **74.32** | **74.35** |

## V. EXPERIMENTS AND RESULTS

### A. *Multi-dialect Classification*

Our first experiment evaluates our feature set for distinguishing all of the dialects from each other. This is a 6-way classification task with a random baseline of 16.67%. The oracle baseline – the estimated maximum accuracy possible on this data – is 81%, meaning that not any of our feature types can correctly classify around 19% of this data. The results for all of our features under both cross-validation methods are shown in Table I.

These results show that character $n$-grams are the best feature type, with trigrams yielding the highest accuracy and performance dropping sharply with 4-grams. Word unigrams are also an informative feature, although not as accurate as the other features.

We also experiment with combining different feature types into a single feature vector, with results shown in the third section of Table I. Here we observe that a combination of character $1/2/3$-grams provides the best result for this type of simple combination.

Finally, we also test our stacked generalization model for this task with all 6 feature types, achieving an accuracy of 74%. This is an 8% increase over the best single-classifier model and is only 7% lower than the oracle upper-bound.

We can also assess the degree of confusion between classes; a confusion matrix of the results obtained using the stacked generalization model is presented in Figure 3. Egyptian Arabic has the highest degree of confusion, mostly with MSA and Palestinian Arabic. We also see a significant amount of confusion between Jordanian Arabic and the Syrian and Palestinian varieties. This is not surprising and likely a result of geographical proximity as all three classes are Levantine dialects. MSA and Tunisian are the dialects that are most accurately identified.

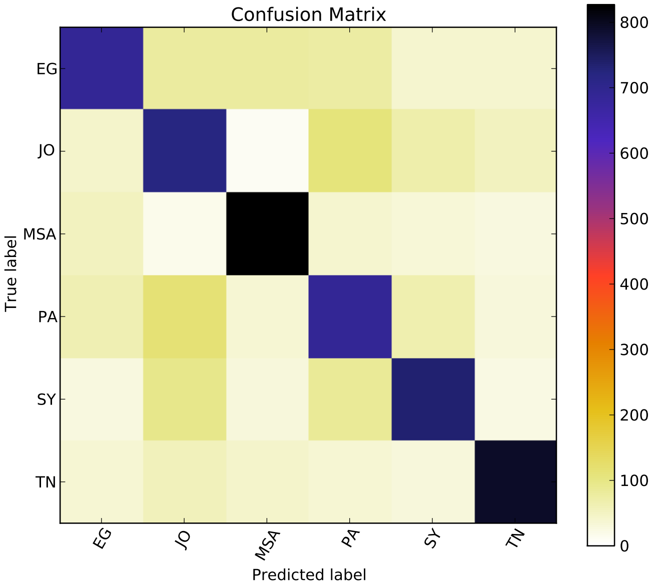

Figure 3. The confusion matrix of our multi-class classification using stacked generalization with all features, visualized as a heatmap.

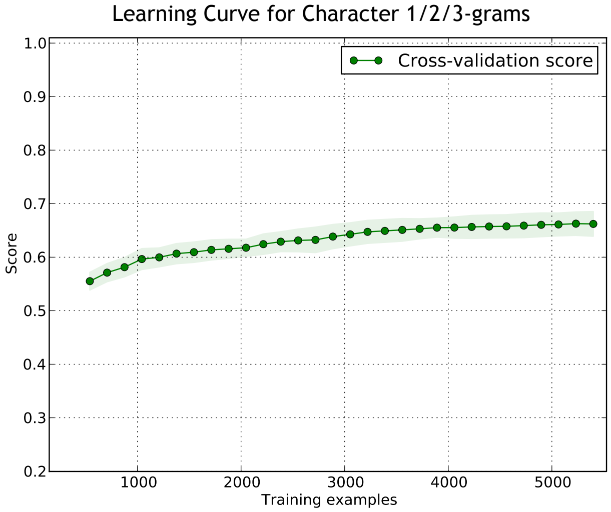

Figure 4. The learning curve for Character $1/2/3$-grams. The curve begins to stabilize after around 4k training sentences.

Finally, we can also assess the learning curve for our best feature, a combination of character $1/2/3$-grams. This is shown in Figure 4. It can be seen that there is rapid increase with the first 1,000 training instances and steady increases until the curve begins to stabilize at around 4,000 training examples. The accuracy does continue to increase after this, albeit at a slower pace. This suggests that the addition of more training data could help increase performance.

Table II
PAIRWISE ARABIC DIALECT IDENTIFICATION CLASSIFICATION ACCURACY USING CHARACTER $1/2/3$-GRAMS.

| | EG | JO | MSA | PA | SY | TN |
|---|---|---|---|---|---|---|
| EG | | 83.0 | 82.8 | 74.9 | 81.6 | 82.3 |
| JO | 83.0 | | 93.8 | 76.3 | 80.8 | 86.0 |
| MSA | 82.8 | 93.8 | | 91.4 | 90.7 | 90.2 |
| PA | 74.9 | 76.3 | 91.4 | | 83.1 | 87.3 |
| SY | 81.6 | 80.8 | 90.7 | 83.1 | | 87.9 |
| TN | 82.3 | 86.0 | 90.2 | 87.3 | 87.9 | |

### B. Pairwise Classification

Given that much of the previous dialect identification work has focused on binary classification of two dialects, we perform pairwise classification between all six varieties in the MPCA dataset. The results are shown in Table II.

MSA and Jordanian Arabic are the most distinguishable pair with an accuracy of 93.8%. Conversely, Palestinian and Egyptian Arabic are the most challenging to discriminate, resulting in the lowest accuracy of 74.9%. Most other pairs are discriminated well. The accuracy for the widely-investigated MSA-Egyptian pair is similar to the previous results reported in Section II.

### C. Cross-Corpus Evaluation

Our final experiment aims to assess the generalizability of the features learned by our system. We do this through a cross-corpus evaluation using the MPCA and the AOC dataset described in Section II. Additionally, we also test our system on the set of tweets constructed by [10].

As the datasets cover different dialects, we use the overlapping MSA and Egyptian dialects for binary classification. We take 2,000 sentences from the MPCA dataset and 26,039 sentences from the MSA-Egyptian portion of the AOC dataset.[7] We also test against the Twitter dataset composed of 700 tweets.

What is interesting about this setup is that the data differ significantly in size and content; one is a parallel corpus while the other contains web-sourced user comments.

Using our Character $1/2/3$-gram and Word unigram features, we train on the MPCA and test on the AOC dataset, and vice versa. We also train a single model using both the AOC and MPCA data and test it against the tweets. The results for all of these evaluations are listed in Table III.

These results again show that the character features perform very well in both cross-corpus scenarios. The accuracy for training on the MPCA is over 20% higher than the AOC baseline. This is particularly impressive considering that we are using only 2k sentences from one corpus to classify over 26k sentences from a radically different corpus with 73.6% accuracy. Word unigrams are also useful and only a few percentage points behind the character $n$-grams.

Table III
MSA VS. EGYPTIAN ARABIC CROSS-CORPUS CLASSIFICATION RESULTS FOR TRAINING ON ONE DATASET AND TESTING ON THE OTHER, AND VICE VERSA. BOLD INDICATES BEST RESULT IN COLUMN.

| | **Cross-Corpus Accuracy (%)** | | |
|---|---|---|---|
| **Train** | MPCA[a] | AOC | AOC+MPCA |
| **Test** | AOC[b] | MPCA | Tweets[c] |
| Baseline | 51.89 | 50.00 | 50.00 |
| Character $1/2/3$-grams | **73.60** | **83.35** | 94.00 |
| Word unigrams | 68.82 | 80.20 | **96.71** |
| All Features | 73.16 | 83.00 | 95.14 |

[a] Includes 2,000 sentences, distributed evenly across the two classes.
[b] Has 26,039 sentences, majority baseline used as not evenly distributed
[c] Includes 700 Tweets distributed equally across both classes.

This pattern is mirrored for training and the larger AOC and testing on the MPCA, but with higher accuracies. This is not surprising given that the training data is 13 times larger. Character $n$-grams provide the best cross-corpus accuracy of 83.85% compared to 80.20% for the word unigrams, both of which are against a 50% random baseline.

A key finding here is that the models trained here do generalize across datasets with a high degree of accuracy, despite their striking differences in size and content. Although this result does not evidence the absence of topic bias, it may indicate that its negative effects are tolerable.

These results also suggest that, at least for small dataset like the MPCA, character $n$-grams generalize the most. However, it may be the case that word unigrams may perform better with a large enough dataset; character $n$-grams may be performing better here as there may not be much lexical overlap between the unrelated datasets.

## VI. ERROR ANALYSIS

In this section we isolate and analyze the misclassified sentences in the MPCA data to gain a better understanding of the challenges for sentence-level dialect identification.

### A. Sentence Length Analysis

Sentence length, measured by the number of tokens, is an important factor to consider in sentence-level classification tasks [20] [21]. There may not be enough distinguishing features if a sentence is too short. Conversely, very long sentences will likely have more features that facilitate correct classification. Here we investigate the length of misclassified items.

The MPCA data has a mean sentence length of 8.9 tokens ($SD$=5.3) while the misclassified subset has a substantially smaller average length of 6.8 tokens per sentence ($SD$=4.07). Histograms for this data are shown in Figure

[7] This contains 13,512 MSA sentences, resulting in a majority class baseline of 51.89%

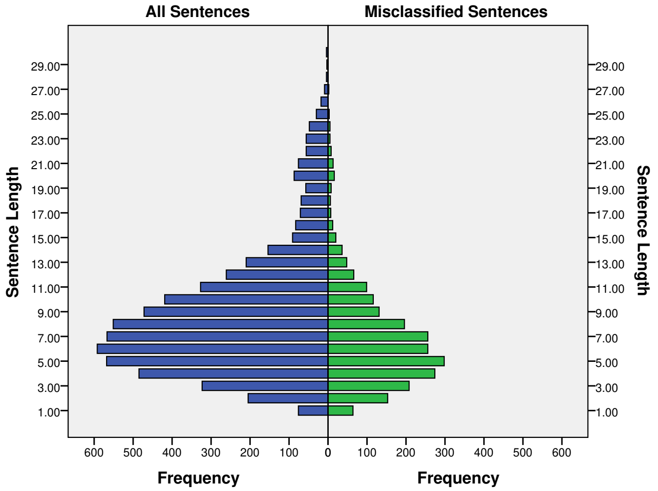

Figure 5. Histograms of the sentence lengths (tokens) in our data for the entire corpus (left) and only for misclassified sentences (right).

| Egyptian Arabic | English |
|---|---|
| ده, دى, دي | Egyptian specific references |
| كد, كدا, كده | Like that |
| ايه, بجد | Denoting questions |
| اوي, اوى | Egyptian specific intensifiers |
| حصل | Verb meaning ‘happened’ |
| بقى, بق | Word denotes inferring/reasoning |
| واد | Noun meaning “boy” |
| ريس | Noun meaning “president” |
| فيش, دش, حدش, مش | Negators |
| بتا | Token denoting possession |

Figure 6. Discriminative features for Egyptian Arabic.

5. We also observe that very few of the longer sentences are misclassified.

An analysis of the cumulative frequency shown that 65% of the misclassified sentences have 7 tokens or less. In sum, evidence from this analysis points to the challenges of distinguishing smaller sentences.

### B. Human Evaluation of Misclassified Sentences

We also perform a human evaluation on the misclassified sentences. Such analyses of misclassified items can help better understand the difficulty of a task [19] and provide further insights about the task.

For this analysis 20 misclassified sentences from each dialect were randomly selected to create a set of 120 sentences. The second author, a native speaker with experience in dialectal Arabic research, was then required to label each sentence with the most probable dialect.

Only 23 sentences (19.17%) were correctly classified, only slightly above the random baseline of 16.67%. Analysis and evaluator feedback from the task provided some relevant insights:

- A large proportion of the sentences are very short and therefore lack contextual and dialect-specific cues that can be effective in determining the dialect class accurately.
- The above issue results in many texts being acceptable into any of dialect classes.
- A number of other instances can be confidently ruled out as being MSA, but it is not clear which non-MSA dialect they belong to.
- A narrower subset of sentences can fit within any of the Levantine dialects.
- Most of the correctly labelled sentences (65%) were MSA or Egyptian Arabic.

These results and highlighted issues comport with our confusion matrix and sentence length analyses. All of these findings could also explain why the MPCA data has an oracle baseline of 81%.

Future work can use an oracle classifier [19] to isolate the subset of sentences that no feature type can predict correctly.

## VII. Feature Analysis

In this section we perform an analysis of the most discriminative features associated with each class in the MPCA data. We do this using the method proposed by [22] to extract lists of features associated with each dialect.

### A. Egyptian Arabic

A large portion of discriminative features here are dialect specific function words and highly dialectal content words. Some discriminative features are shown in Figure 6.

### B. Jordanian Arabic

The discriminative features of Jordanian Arabic tend to be more content words rather than function words. We also note that some of the content words are conversational (*i.e.* بحكي ، حكي ، حكتلو). This might reflect a genuine trend in this dialect or it could merely be an artefact due to the size of the dataset. Example features are listed in Figure 7.

### C. Palestinian Arabic

This dialect is also distinguished by more unique content words rather than function words. Some examples from this dialect are listed in Figure 8.

### D. Syrian Arabic

This dialect has some features that overlap with the other Levantine dialects. Examples are shown in Figure 9.

| Jordanian Arabic | English |
|---|---|
| عال | Common token meaning “good/fine” |
| اشي | Common token meaning “something” |
| حدا | Common token meaning “someone” |
| زلم | Common token meaning “men” |
| هاي | A reference, “this” |
| لسا, مش | Negators |
| رح | Denoting future actions, “will” |
| ليش, شو | Denoting questions |

Figure 7. Discriminative features for Jordanian Arabic.

| Palestinian Arabic | English |
|---|---|
| هسه | “Now” |
| في, عن, على, حتى | Prepositions overused by this dialect |
| مره | A woman |
| مش | Negator |
| شو | Denoting questions |
| هدا, هد | Denoting references |
| بد | A common prefix in Levantine dialects that denotes a desire to do something. “I’d like to”. |

Figure 8. Discriminative features for Palestinian Arabic.

| Syrian Arabic | English |
|---|---|
| عم | Denote actions currently happening |
| رح | Denotes future actions, “will” |
| متل | Syrian spelling variation “like/similar to” |
| هلئ | Word shared among Levantine dialects meaning “now” |
| تعي | Syrian specific verb “come” |
| مو , ليش | Negators |
| هاد, هنن, حنا, عنا | References |

Figure 9. Discriminative features for Syrian Arabic.

| Tunisian Arabic | English |
|---|---|
| نيش, فما, موش, بلا | Negators |
| باش | Function word meaning “In order to” |
| شن, شنو, شكو | Interrogative prefixes |
| ش, وش | Usually appear as affixes of an integrative word |
| نحب | Verb meaning “to like” |
| برشة | A Tunisian-specific intensifier |
| لحق | Adverb means “Truly” |
| هكا | “Like that” |

Figure 10. Discriminative features for Tunisian Arabic.

### E. *Tunisian Arabic*

As shown by the features in Figure 10, this dialect has a set of highly specific negators, prefixes, intensifiers, interrogative prefixes and verbs.

### F. *Modern Standard Arabic*

Function words are the most discriminative features for MSA, some of which are listed in Figure 11.

## VIII. Discussion and Conclusion

We presented a number of Arabic dialect identification experiments using the newly released MPCA dataset. These results inform current research in several ways.

We demonstrated the utility of a parallel corpus for ADI, achieving 74% accuracy on a 6-dialect classification task with a random baseline of 16.7%. Pairwise binary dialect classification within this corpus also yielded results as high as 94%, but also highlighted poorer results between closely related dialects such as Palestinian and Jordanian (76%). This was also evident in our feature analysis where we observed that the Levantine dialects share a lot of common features, making it harder to distinguish them. The results also show that leave-one-out cross-validation leads to very similar results as 10-fold cross-validation.

We demonstrated that a meta-classifier can provide significant increases for multi-class dialect identification. This is a direction that requires further investigation as this is the first application of such a method for this task.

Our cross-corpus experiment demonstrated that models trained with the MPCA generalize to other data. This was also the case for the AOC dataset when tested against the MPCA. Data from both corpora was also used to classify 700 Egyptian Arabic and MSA tweets with 97% accuracy.

Similar to the results of [10], we find that character $n$-grams are in most scenarios the best single feature for this task, in both within- and cross-corpus settings. This is in contrast to the results of [7]–[9] that establish word unigrams as being the best feature type. This discrepancy merits further scrutiny and we plan to investigate it in future research.

| MSA | English |
|---|---|
| سوف, قد | Denoting future actions, “will” |
| ليس, لم | Negators |
| ماذا | “What” |
| له | Denoting possession (masculine) |
| هذه | A reference (feminine) |
| أن | That (connects parts of a sentence) |
| قوم | People |
| ندم | A verb or noun meaning “regret” |
| وجد, تجد | Present and past verbs meaning “find” |

Figure 11. Discriminative features for Modern Standard Arabic (MSA).

One possibility is that previous experiments report higher accuracies due to topic bias *within* their corpora, which is most strongly present in words. This may also be due to the smaller size of our dataset; there may not be a sufficient amount of data and unique tokens to train a learner on words.

The use of the two feature types is not mutually exclusive. In a system that operates on both the token and sentence levels, such as that of [8], the character $n$-grams could be used to classify out of vocabulary (OOV) tokens which are previously unseen.

The key shortcoming of this study, albeit beyond our control, is the limited amount of data available for the experiments. In this regard, we are surprised by relatively high classification accuracy of our system, given the restricted amount of training data available. Future work includes the application of our methods to additional data as it becomes available. Only 1k parallel sentences from the MPCA dataset were available to us at the time of our study, but this is to be expanded in the future.

Another limitation is the absence of data preprocessing. The integration of additional task-specific preprocessing steps, namely tokenization and orthography normalization, could lead to improved performance according to the results reported by [8, p. 459].

There are also a number of other potential directions for future work. The overall accuracy can be increased by focusing on improving the most commonly confused classes (as shown in the confusion matrix in Section V-A) and the worst performing dialect pairs from the pairwise classification analysis.

We also note that conducting an even more comprehensive error analysis could also provide to be a fruitful line of future inquiry. This analysis could provide valuable insights about the most common errors being committed by the current system - such as those related to the above-mentioned class confusion - thus helping guide future efforts in this area.

Another possibility is to experiment with a wider range of features and to assess the diversity of these features, for example, using the method proposed by [23].

Further to increasing the dataset sizes, the number of dialects can also be increased. More datasets could also be used to perform additional cross-corpus experiments. This data can be sourced from everyday natural language productions found in social media and Twitter.

## Acknowledgments

We would like to thank Houda Bouamor for making the MPCA data available. We also thank the three anonymous reviewers for their helpful comments.

## References

[1] N. Y. Habash, “Introduction to Arabic natural language processing,” *Synthesis Lectures on Human Language Technologies*, vol. 3, no. 1, pp. 1–187, 2010.

[2] H. Bouamor, N. Habash, and K. Oflazer, “A Multidialectal Parallel Corpus of Arabic,” in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014.

[3] R. Al-Sabbagh and R. Girju, “Mining the Web for the Induction of a Dialectical Arabic Lexicon,” in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*. Valletta, Malta: European Language Resources Association (ELRA), May 2010.

[4] M. Diab, M. Albadrashiny, M. Aminian, M. Attia, H. Elfardy, N. Habash, A. Hawwari, W. Salloum, P. Dasigi, and R. Eskander, “Tharwa: A Large Scale Dialectal Arabic - Standard Arabic - English Lexicon,” May 2014.

[5] M. Stede, “Lexical choice criteria in language generation,” in *Proceedings of the sixth conference on European chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 1993, pp. 454–459.

[6] F. Biadsy, J. Hirschberg, and N. Habash, “Spoken Arabic dialect identification using phonotactic modeling,” in *Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages*. Association for Computational Linguistics, 2009, pp. 53–61.

[7] O. F. Zaidan and C. Callison-Burch, “The Arabic online commentary dataset: an annotated dataset of informal Arabic with high dialectal content,” in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2011, pp. 37–41.

[8] H. Elfardy and M. T. Diab, “Sentence Level Dialect Identification in Arabic,” in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2013, pp. 456–461.

[9] O. F. Zaidan and C. Callison-Burch, “Arabic Dialect Identification,” *Computational Linguistics*, vol. 40, no. 1, pp. 171–202, 2014.

[10] K. Darwish, H. Sajjad, and H. Mubarak, “Verifiably Effective Arabic Dialect Identification,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, October 2014.

[11] J. Brooke and G. Hirst, “Measuring interlanguage: Native language identification with L1-influence metrics,” in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12)*, Istanbul, Turkey, May 2012, pp. 779–784.

[12] S. Malmasi and M. Dras, “Arabic Native Language Identification,” in *Proceedings of the Arabic Natural Language Processing Workshop (EMNLP 2014)*. Doha, Qatar: Association for Computational Linguistics, October 2014, pp. 180–186. [Online]. Available: http://aclweb.org/anthology/W14-3625

[13] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, “LIBLINEAR: A Library for Large Linear Classification,” *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[14] S. Malmasi, S.-M. J. Wong, and M. Dras, “NLI Shared Task 2013: MQ Submission,” in *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*. Atlanta, Georgia: Association for Computational Linguistics, June 2013, pp. 124–133. [Online]. Available: http://www.aclweb.org/anthology/W13-1716

[15] S. Malmasi and M. Dras, “Large-scale Native Language Identification with Cross-Corpus Evaluation,” in *Proceedings of NAACL-HLT 2015*. Denver, Colorado: Association for Computational Linguistics, June 2015.

[16] D. H. Wolpert, “Stacked Generalization,” *Neural networks*, vol. 5, no. 2, pp. 241–259, 1992.

[17] R. Polikar, “Ensemble based systems in decision making,” *Circuits and systems magazine, IEEE*, vol. 6, no. 3, pp. 21–45, 2006.

[18] R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in *IJCAI*, vol. 14, no. 2, 1995, pp. 1137–1145.

[19] S. Malmasi, J. Tetreault, and M. Dras, “Oracle and Human Baselines for Native Language Identification,” in *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*. Denver, Colorado: Association for Computational Linguistics, June 2015.

[20] S. Malmasi and M. Dras, “Automatic Language Identification for Persian and Dari texts,” in *Proceedings of the 14th Conference of the Pacific Association for Computational Linguistics (PACLING 2015)*, Bali, Indonesia, May 2015.

[21] T. Gottron and N. Lipka, “A Comparison of Language Identification Approaches on Short, Query-Style Texts,” in *Advances in Information Retrieval*, ser. Lecture Notes in Computer Science, C. Gurrin, Y. He, G. Kazai, U. Kruschwitz, S. Little, T. Roelleke, S. Rger, and K. van Rijsbergen, Eds. Springer Berlin Heidelberg, 2010, vol. 5993, pp. 611–614.

[22] S. Malmasi and M. Dras, “Language Transfer Hypotheses with Linear SVM Weights,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, 10 2014, pp. 1385–1390. [Online]. Available: http://aclweb.org/anthology/D14-1144

[23] S. Malmasi and A. Cahill, “Measuring Feature Diversity in Native Language Identification,” in *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*. Denver, Colorado: Association for Computational Linguistics, June 2015.