Classification of Closely Related Sub-dialects of Arabic Using Support-Vector Machines

Samantha Wray
New York University Abu Dhabi
United Arab Emirates
samantha.wray@nyu.edu

Abstract

Colloquial dialects of Arabic can be roughly categorized into five groups based on relatedness and geographic location (Egyptian, North African/Maghrebi, Gulf, Iraqi, and Levantine), but given that all dialects utilize much of the same writing system and share overlapping features and vocabulary, dialect identification and text classification is no trivial task. Furthermore, text classification by dialect is often performed at a coarse-grained level into these five groups or a subset thereof, and there is little work on sub-dialectal classification. The current study utilizes an n-gram based SVM to classify on a fine-grained sub-dialectal level, and compares it to methods used in dialect classification such as vocabulary pruning of shared items across dialects. A test case of the dialect Levantine is presented here, and results of 65% accuracy on a four-way classification experiment to sub-dialects of Levantine (Jordanian, Lebanese, Palestinian and Syrian) are presented and discussed. This paper also examines the possibility of leveraging existing mixed-dialectal resources to determine their sub-dialectal makeup by automatic classification.

Keywords: text classification, validation of language resources, language identification

1. Introduction

Arabic is a language rich in dialectal variety, with Modern Standard Arabic (MSA) used for official purposes alongside a colloquial dialect local to a speaker's given region. Although these dialects are typically written throughout informal contexts such as social media, they are nonstandardized and exhibit spontaneous spelling when compared to the standardized variety.

In the text domain, previous work on dialect identification has been performed for purposes ranging from training automatic speech recognition to bootstrapping corpus building efforts. For some of these studies, the focus is on a binary classification between MSA and a colloquial dialect, such as the Egyptian-MSA classification performed by Tillmann et al. (2014), Mansour et al. (2014) and Elfardy and Diab (2013).

A more fine-grained classification is performed by (Zaidan and Callison-Burch, 2011), as they crowdsource the human classification of 100,000 sentences of dialectal Arabic, and train language models to classify additional dialectal material scraped from online newspaper comment sections. The classification categories for this study were MSA, Levantine, Egyptian, and Gulf. This 4-way classification achieves an accuracy of 69.4% (Zaidan and Callison-Burch, 2011). However, their accuracy when classifying between just the non-standard dialects Levantine, Egyptian, and Gulf jumps to 83.5%.

Typically, dialect classification is performed into these coarse-grained categories. However, work by Malmasi et al. (2015) attempted classification with a finer-grained distinction. They classify text into seven categories: MSA, Tunisian, Egyptian, Jordanian, Syrian, and Palestinian, achieving a 74.3% accuracy for this 6-way classification.

However, the training set used to construct the language models was based on materials created by human translation from one dialect to another. This parallel corpus is composed of the above 5 non-standard dialects in addition to MSA from (Bouamor et al., 2014), but the corpus is not spontaneous and was created under prompts in such a way that it is not entirely parallel. For example, when speakers who contributed to the corpus were prompted to translate a base phrase into their native dialects, some speakers inserted additional material (such as /maʃa al:a/ 'wow!') while others did not, despite these phrases being pan-Arabic and having no likely preponderance to occur more frequently in one dialect over another. It's likely then that these phrases emerge in the corpus as an artifact of a particular contributor's inclination to use them at that time. Finally, Arabic dialect identification has been the subject of shared tasks focusing on discriminating between similar languages (Malmasi et al., 2016), again, at a coarse-grained level between MSA and the colloquial dialects North African, Egyptian, Levantine, and Gulf. The methodology that has emerged from this line of research is the use of *support vector machines* (SVM) utilizing character n-grams. Given the relatively small size of training data utilized in these tasks, SVM has outperformed other methods such as deep neural networks (Malmasi et al., 2016).

2. Materials

To focus on training and testing materials that were annotated at the sub-dialectal level and produced spontaneously, I utilized 10,000 tweets harvested from each of the four countries that make up the speakers of the Levantine dialect (Jordan, Palestine, Syria, and Lebanon). These 10,000 tweets contained a total of approximately 100,000 words.

I drew partially from annotated material from (Mubarak and Darwish, 2014) which contains Twitter data from Jordan, Palestine, Syria, and Lebanon and was collected by exploiting user-provided location and geo-tagging info. I supplemented this with additional tweets from each country, which were harvested in the same manner throughout 2014-2015.

There are approximately 15 million tweets produced in Arabic per day (Mubarak and Darwish, 2014). Each tweet produced has associated metadata that includes the time of publication, the username of the author, the language the tweet is written in, as well as location information of where the tweet was written from. This metadata is searchable and made available for harvesting via the Twitter API[1]. Given that all colloquial dialects of Arabic utilize the same script for encoding language, and because the writing of colloquial language is not a phenomenon unique to any given dialect, the approach used in previous studies pioneering tweet harvesting such as (Ljubešic et al., 2014) will not provide fine-grained dialect information given that the metadata included for Twitter does not include Arabic dialect info. A selection of 'Arabic' would produce all colloquial dialects, as well as material written in MSA.

As for location information, it comes in two forms. The first is a raw latitude longitude geo-tag which can give precise indication to region. However, this is an optional feature most users do not use: only about 2% of tweets are geo-tagged for location (Huck et al., 2012). Another optional feature is a user-provided location, which approximately 70% of tweets include (Mubarak and Darwish, 2014). These are locations which the user has custom written, and may or may not be informative as to actual location. For example, a user can write they are located in 'Amman, Jordan' or simply write 'My house'. I used the Twitter API and engaged with it via the Twython library[2] for Python. Using these tools, I requested the Twitter API search through all tweets in the Arabic language which met either of the following criteria:

- Geo-tag: tweet was tagged within the geographic borders of the regions of interest
- User location: tweet was written by user with location matching major cities in Levantine countries, as well as the country names. Following Mubarak and Darwish (2014), possible user location names associated with each country were drawn from the geographic name database GeoNames[3]. In addition, each time a user from a specific country was identified, manual inspection of user locations of their followers was performed to collect new permutations of possible ways to indicate residence in that country

This approach operated under a couple of assumptions. First, that users tweet from a location where speakers speak the same dialect as the user. Given patterns of immigration in the Levant, users who produce geo-coded tweets in a particular country may have origins, and therefore native dialects, associated with a different region. Furthermore, a user may be temporarily producing tweets with a different geo location other than their default location if they are traveling. To partially circumvent this potential confound, tweets which exhibited mismatch between the user-provided location and the geocode were not retained in the dataset. Furthermore, a potential confound arises from the fact that a particular author's features may be easily identified across testing and training sets (Rangel et al., 2017). However, the data has been completely anonymized and therefore further analysis regarding the influence of authorship is unavailable.

Tweets were collected for the greater part of a year to help reduce the possibility of cyclic effects of frequency, such as the tendency for holiday greetings to spike during Ramadan, for example (Eisenstein, 2013; Refaee and Rieser, 2014). To help maximize the likelihood that the collected tweets contained colloquial material, I discarded tweets which contained vowel diacritics, as this was typically indicative of a user tweeting a verse from the Quran, and not likely to contain colloquial data.

Data was prepared for study by removing additional non-textual information included in the tweets such as emojis, other usernames (mentions), punctuation, and links.

In addition to spontaneously-produced tweets, two additional language resources were utilized: two different spoken Arabic telephone corpora in which the speaker's country of origin is annotated (Appen, 2007; Maamouri et al., 2007). The transcribed speech corpora differed from the tweets as they were professionally transcribed, thus losing any character-level features that may emerge as a result of being written by the speaker. The telephone transcripts were used to generate words unique to each dialect; that is, words which appeared in telephone conversations in one dialect but not in the other three. These vocabularies were thus pruned to remove overlapping dialectal items.

3. Classification Experiments

Following recent work on discriminating between dialects of Arabic for transcriptions of speech (Malmasi et al., 2016), a state-of-art method of Support Vector Machines including features of character n-grams (from unigrams to 5-grams) was explored (Eldesouki et al., 2016). In addition to this, word n-grams (from unigrams to trigrams) were included in the models.

In addition to the inclusion of n-gram based feature models, models which focused on unique words for each dialect were included to prune vocabulary that overlaps across dialects. This method has been demonstrated to be successful in large-scale text identification and classification tasks (Madsen et al., 2004, among others), including discrimination tasks for other closely-related languages with low resources and a large lexical overlap, such as Tagalog, Cebuano, and Bicolano (Dimalen and Roxas, 2007).

3.1. Procedure

A multiclass linear kernel Support Vector Machine (Joachims, 1999) was used to perform a four-way classification into the sub-dialects Jordanian, Palestinian, Lebanese, and Syrian, on the Twitter data described above. The model included features for character unigrams up to 5-grams, and word unigrams up to trigrams. Furthermore, a model incorporating features of the pruned vocabulary lists built from the telephone corpora were explored. Classification was performed on each tweet individually.

To generate n-gram probabilities used as features for the SVM, I used the SRI Language Modeling (SRILM) toolkit

[1] https://dev.twitter.com/

[2] https://github.com/ryanmcgrath/twython

[3] geonames.org

(Stolcke and others, 2002) encompassing the following:

- *Word n-grams:* Modeled the probability of occurrences of contiguous word strings in a text. For the current study, three word n-gram models were built: a unigram model determining the likelihood of a word occurring in a text, a bigram model determining the likelihood of a word occurring given the previous word, and a trigram model determining the likelihood of a word occurring given the previous two words. Unknown words encountered by the model were mapped to <unk>, and sentence start and end markers were considered in the models.
- *Character n-grams:* Modeled the probability of occurrences of contiguous character strings in a text. Five character n-gram models were built: a unigram model determining the likelihood of a character appearing in a text, a bigram model determining the likelihood of a character occurring given a previous character, and so on, until the 5-gram model.[4]

An additional feature type considered was the inclusion of dialect-specific words generated by the unique words list from the telephone corpora. These features were binary: either the word was present, or not.

3.2. Results

I used k-fold cross-validation (k = 10), in which a sampling of 9/10 of the data was used as the training data with the remaining 1/10 being used as testing data, and this process was performed 10 times. The accuracy reported in 1 is calculated as an average across all 10 folds.

Method	Accuracy
N-gram model:	
1-5 character; 1-3 word	**65%**
Vocabulary pruning	54%
N-gram + vocabulary pruning	46%

Table 1: Classification accuracy for training and testing sets. Best results are in bold.

Table 1, shows the performance of three models: (i) The n-gram based SVM trained on the Twitter data which did not incorporate unique words from the telephone corpora, (ii) SVM trained on only the unique words per dialect from the telephone corpora and (iii) a combination SVM using both of the above feature sets. The first model, the n-gram based SVM trained on the Twitter data, had the highest accuracy at 65%.

4. Classification Estimates for Existing Resources

Given the SVM n-gram model presented in the previous section achieves an accuracy of 65% when classifying tweets, it was utilized to perform further classification on existing Levantine resources in order to provide estimates on their sub-dialectal makeup. I used the model to classify the mixed Levantine corpora from Almeman and Lee (2013) and Zaidan and Callison-Burch (2011) and performed classification on each sentence in the corpus. These estimates are shown in Table 2.

Existing mixed-dialect resource	Classification
(Zaidan and Callison-Burch, 2011)	JOR 42%
	LEB 19%
	PAL 9%
	SYR 30%
(Almeman and Lee, 2013)	JOR 70%
	LEB 9%
	PAL 17%
	SYR 4%

Table 2: Classification estimates for existing mixed-dialect resources using the best performing model (n-gram SVM)

As shown in Table 2, a majority of the data was classified as Jordanian. The high estimates for Jordanian are not surprising, especially for classification of (Zaidan and Callison-Burch, 2011) given that the resource is composed of newspaper commentary and the newspaper selected to represent Levantine for that multidialectal dataset was indeed a Jordanian newspaper.

5. Conclusions and Future Work

Colloquial/Spoken Arabic dialect identification and discrimination in text is a nontrivial task due to several factors, including the fact that the dialects are closely related and thus exhibit overlapping lexicons and a shared writing system. For fine-grained distinctions at the sub-dialectal level, this problem is exacerbated by the lack of data annotated by sub-dialect and a lack of previous research into discrimination at the sub-dialectal level. In this paper, the state-of-the-art methodology for discriminating between colloquial/spoken dialects of Arabic in text (n-gram based SVMs) (Malmasi et al., 2016) was applied at a more fine-grained level. I also explored utilizing this methodology to leverage existing coarse-grained mixed dialectal resources with the intention of repurposing them as fine-grained resources with classification at the sub-dialectal level. Further improvements are currently being explored, including the utilization of tf-idf for features. Furthermore, vocabulary pruning from different sources than the telephone corpora should be explored, given that the telephone corpora are quite sparse at only approximately 200,000 words. For example, Darwish et al. (2014) demonstrated success with dialect-specific words in text-based classification, specifically, manually-identified Egyptian words at the high end of the frequency range from (Zbib, Rabih and Malchiodi, Erika and Devlin, Jacob and Stallard, David and Matsoukas, Spyros and Schwartz, Richard and Makhoul, John and Zaidan, Omar F and Callison-Burch, Chris, 2012). The current study focuses on a test case of the Arabic dialect Levantine, which is broadly made up of four geopolitical entities that can serve as a more fine-grained classification (Jordanian, Lebanese, Palestinian, and Syrian.)

[4]Models were also run utilizing a binary system in which presence or absence of an n-gram was indicated. Results did not differ with regard to predictive performance, but did benefit a slight time cost and would be therefore lighter weight to deploy if scaled up.

However, this methodology and improvements upon it can be easily adapted to other closely-related dialects, such as those spoken throughout the Arab Gulf, and to other non-geopolitical dialectal distinctions such as the urban vs. rural varieties.

6. Acknowledgements

Preliminary work on this research appeared as part of a doctoral dissertation (Wray, 2016). Funding for this research is gratefully acknowledged from the National Science Foundation under BCS-1533780, and the University of Arizona Graduate and Professional Student Council. An allocation of computer time is also gratefully acknowledged from the UofA Research Computing HPC and HTC, as well as HPC at New York University Abu Dhabi.

7. Bibliographical References

Almeman, K. and Lee, M. (2013). Automatic building of Arabic Multi Dialect text corpora by bootstrapping dialect words. In *Communications, Signal Processing, and their Applications (ICCSPA), 2013 1st International Conference on*, pages 1–6. IEEE.

Darwish, K., Sajjad, H., and Mubarak, H. (2014). Verifiably Effective Arabic Dialect Identification. *EMNLP-2014*.

Dimalen, D. M. and Roxas, R. E. O. (2007). AutoCor: A Query Based Automatic Acquisition of Corpora of Closely-related Languages. In *PACLIC*.

Eisenstein, J. (2013). What to do about bad language on the internet. In *HLT-NAACL*, pages 359–369.

Eldesouki, M., Dalvi, F., Sajjad, H., and Darwish, K. (2016). QCRI@ DSL 2016: Spoken Arabic Dialect Identification Using Textual. *VarDial 3*, page 221.

Elfardy, H. and Diab, M. T. (2013). Sentence Level Dialect Identification in Arabic. In *ACL (2)*, pages 456–461.

Huck, J., Whyatt, D., and Coulton, P. (2012). Challenges in geocoding socially-generated data.

Joachims, T. (1999). Svmlight: Support vector machine. *SVM-Light Support Vector Machine http://svmlight. joachims. org/, University of Dortmund*, 19(4).

Ljubešic, N., Fišer, D., and Erjavec, T. (2014). TweetCaT: a Tool for Building Twitter Corpora of Smaller Languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland. European Language Resources Association (ELRA)*.

Madsen, R. E., Sigurdsson, S., Hansen, L. K., and Larsen, J. (2004). Pruning the vocabulary for better context recognition. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 2, pages 483–488. IEEE.

Malmasi, S., Refaee, E., and Dras, M. (2015). Arabic dialect identification using a parallel multidialectal corpus. In *Proceedings of the 14th Conference of the Pacific Association for Computational Linguistics (PACLING 2015), Bali, Indonesia*, pages 209–217.

Malmasi, S., Zampieri, M., Ljubešic, N., Nakov, P., Ali, A., and Tiedemann, J. (2016). Discriminating between similar languages and Arabic dialect identification: A report on the third DSL shared task. *VarDial 3*, page 1.

Mansour, S., Al-Onaizan, Y., Blackwood, G., and Tillmann, C. (2014). Automatic Dialect Classification for Statistical Machine Translation. In *Proc. of the Conf. of the Assoc. for Machine Translation in the Americas (AMTA), Vancouver, BC, Canada*.

Mubarak, H. and Darwish, K. (2014). Using Twitter to collect a multi-dialectal corpus of Arabic. *ANLP 2014*.

Rangel, F., Rosso, P., Potthast, M., and Stein, B. (2017). Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. *Working Notes Papers of the CLEF*.

Refaee, E. and Rieser, V. (2014). Subjectivity and sentiment analysis of Arabic twitter feeds with limited resources. In *Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools*.

Stolcke, A. et al. (2002). SRILM - An extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002)*, volume 2002, pages 901–904.

Tillmann, C., Al-Onaizan, Y., and Mansour, S. (2014). Improved sentence-level Arabic dialect classification. In *Proceedings of the VarDial Workshop*, pages 110–119.

Wray, S. (2016). Decomposability and the effects of morpheme frequency in lexical access. *University of Arizona*.

Zaidan, O. F. and Callison-Burch, C. (2011). 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.

8. Language Resource References

Almeman, K. and Lee, M. (2013). Automatic building of Arabic Multi Dialect text corpora by bootstrapping dialect words. In *Communications, Signal Processing, and their Applications (ICCSPA), 2013 1st International Conference on*, pages 1–6. IEEE.

Appen. (2007). *Levantine Arabic Conversational Telephone Speech LDC2007T01*. Linguistics Data Consortium.

Bouamor, H., Habash, N., and Oflazer, K. (2014). A Multidialectal Parallel Corpus of Arabic. In *LREC*, pages 1240–1245.

Maamouri, M., Buckwalter, T., Graff, D., and Jin, H. (2007). *Fisher Levantine Arabic Conversational Telephone Speech LDC2007S02*. LDC.

Mubarak, H. and Darwish, K. (2014). Using Twitter to collect a multi-dialectal corpus of Arabic. *ANLP 2014*.

Zaidan, O. F. and Callison-Burch, C. (2011). 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.

Zbib, Rabih and Malchiodi, Erika and Devlin, Jacob and Stallard, David and Matsoukas, Spyros and Schwartz, Richard and Makhoul, John and Zaidan, Omar F and Callison-Burch, Chris. (2012). *Machine translation of Arabic dialects*.