A Multi-Dialect, Multi-Genre Corpus of Informal Written Arabic

Ryan Cotterell,[1] **Chris Callison-Burch**[2]

[1] Center for Language and Speech Processing, Johns Hopkins University
[2] Computer and Information Science Department, University of Pennsylvania

Abstract

This paper presents a multi-dialect, multi-genre, human annotated corpus of dialectal Arabic with data obtained from both online newspaper commentary and Twitter. Most Arabic corpora are small and focus on Modern Standard Arabic (MSA). There has been recent interest, however, in the construction of dialectal Arabic corpora (Zaidan and Callison-Burch, 2011a; Al-Sabbagh and Girju, 2012). This work differs from previously constructed corpora in two ways. First, we include coverage of five dialects of Arabic: Egyptian, Gulf, Levantine, Maghrebi and Iraqi. This is the most complete coverage of any dialectal corpus known to the authors. In addition to data, we provide results for the Arabic dialect identification task that outperform those reported in Zaidan and Callison-Burch (2011a).

Keywords: arabic, dialectal arabic, dialect identification, crowd-sourcing annotation

1 Introduction

This paper presents a multi-dialect, multi-genre, human annotated corpus of dialectal Arabic with data obtained from both online newspaper commentary and Twitter. Most Arabic corpora are small and focus on Modern Standard Arabic (MSA). There has been recent interest, however, in the construction of dialectal Arabic corpora (Zaidan and Callison-Burch, 2011a; Al-Sabbagh and Girju, 2012) . This work differs from these in two ways. First, we include coverage of five dialects of Arabic: Egyptian, Gulf, Levantine, Maghrebi and Iraqi. This is the most complete coverage of any dialectal corpus known to the authors. Second, every sentence in the corpus was human annotated on Amazon's Mechanical Turk; this stands in contrast to Al-Sabbagh and Girju (2012) where only a small subset was human annotated in order to train a classifier. In addition to data, we provide results for the Arabic dialect identification task that outperform those reported in Zaidan and Callison-Burch (2011a). The paper is structured as follows: in section 2 we provide a brief overview of the relevant socio-linguistic details of the Arabic language, in section 3 we discuss related work pertaining to dialect corpus creation and dialect identification, in sections 4, 5 and 6 we discuss our methodology and annotation techniques and in section 8 we discuss experiments.

2 Arabic

Arabic exhibits a linguistic phenomenon known as diglossia, in which the written standard differs substantially from the spoken vernacular. The written standard, known as Modern Standard Arabic (MSA), is largely based on the Qur'anic literary register. It is used across the Arabic-speaking world in written news and in broadcast media. The local dialects however, differ substantially from MSA at all linguistic levels. The recent emergence of informal, user-generated text has led to a proliferation of large quantities of written dialectal Arabic on the internet. The Arabic dialects differ for historical reasons and have been individually influenced by the pre-Arabization language spoken by the population, as is the case with Aramaic in the Levant, as well as the European languages from the time of colonization. Such distinctions are important as North African dialects are unique in the quantity of French loanwords, whereas Iraqi Arabic has been historically more influenced by Turkish.

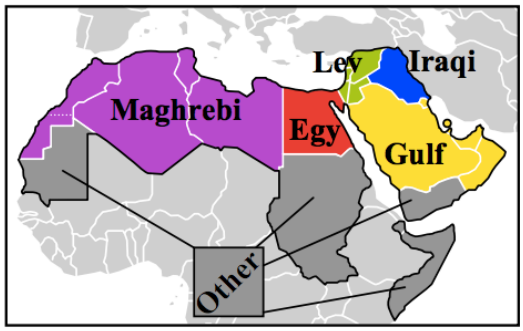

Figure 1: Arabic Map

The stark difference between MSA and the dialects creates a problem for NLP software trained largely on MSA text. As most previous efforts in Arabic NLP have focused on MSA, there is a dearth of resources available to adequately tackle the problem. A natural starting place for Arabic-dialect NLP lies in dialect identification and classification. Since each dialect group should be treated as a separate language from the point of view of downstream processing tasks, classifying them will be key for any pipeline. At the highest level, the Arabic dialects can be divided into two groups: the *maghreb* and the *mashreq* dialects. The *maghreb* dialects consist of the North African dialects spoken west of Egypt and the *mashreq* dialects include everything east of Egypt. Within these broader categories, subdivisions are made whose differences often reflect pre-Arab culture and colonial history. We opted to classify dialect Arabic into five groups

Maghrebi (spoken in all of North Africa)

Egyptian (spoken in Egypt, but understood universally)

Levantine (spoken primarily in the Levant, Syria and Palestine)

Iraqi (spoken in Iraq)

Gulf (spoken primarily in Saudi Araubi, UAE, Kuwait and Qatar)

Automatically, identifying dialects is a more complex task than language identification. The relation between Latin and Romance languages is often brought forth as a European equivalent of the distinction between the MSA and the Arabic dialects. However, Arabic dialect identification is further complicated by the orthography. Traditionally, vowels are omitted from Arabic text leaving obvious phonological clues absent. A similar task would be stripping the vowels from French and Italian text and trying to identify the correct language. A further complication arises in the shared technical vocabulary from MSA.

3 Related Work

Our work is a direct extension of (Zaidan and Callison-Burch, 2011a) in that we use a similar methodology for the collection of the data and the classification task. There are several other dialectal Arabic corpora of note. Al-Sabbagh and Girju (2012) created an Egyptian Arabic corpus through human annotation and classifiers. The COLABA project has similarly constructed dialectal resources from web logs (Diab et al., 2010). Elfardy and Diab (2012c) provide guidelines for the construction of large corpora of mixed Arabic resources. Elfardy and Diab (2012b) introduced AIDA, a system for dialect identification, classification and glossing on both the token and sentence level. Elfardy and Diab (2013) presented a supervised approach for sentence level dialect identification and studied the effects of preprocessing techniques on classifier accuracy. Tratz et al. (2013) made efforts to improve the annotation of Arabic corpora through the creation of a tool specifically designed to facilitate the annotation of social media data. A different line of work that is relevant is the analysis of code-switched data. Owning to the informal nature of conversations in Arabic dialects, the language is often mixed with MSA. Therefore the line between what is dialect and what is MSA is blurred. It thus may be more appropriate to consider the task of token level dialect identification as in Elfardy and Diab (2012a).

4 Newspapers

We set out to create a dataset of dialectal Arabic to address the general lack of resources. The most viable resource of dialectal Arabic text is online data, which is more individual-driven and less institutionalized, and therefore more likely to contain dialectal content. Possible sources of dialectal text include web logs, fora, and chat transcripts. We collected a substantial amount of dialect data from user comments from online newspapers, following Zaidan and Callison-Burch (2011a) We chose 5 Arabic language newspapers for this commentary set: an Egyptian newspaper **Al-Youm Al-Sabe'**, a Saudi-Arabian newspaper **Al-Riyadh**, a Jordanian newspaper **Al-Ghad**, an Algerian newspaper **Ech Chorouk El Youmi** and an Iraqi newspaper **Al-Wefaq**.

5 Collection of Twitter Data

Twitter (`twitter.com`) has provided researchers with an enormous quantity of natural language data. In particular, Twitter is an excellent source for colloquial data in many languages. Users typically tweet short, informal messages that reveal many properties of spoken languages (Eisenstein, 2013). Twitter data varies substantially from newspaper commentary, making it a natural complement to the previously scraped data since the service imposes a 140 character limit on all messages, which encourages non-standard orthography.

Twitter provides a well-documented streaming API that allows easy access to their social media content. Access to the Twitter API is rate limited, however, and no individual stream may contain more than 1% of the total global stream. Since the number of Arabic-language tweets as a percentage of the global stream is much less than 1%, this generally is not a problem.

6 Annotation

Amazon's Mechanical Turk (MTurk) provides the primary annotation platform. MTurk has recently been exploited for large-scale linguistic annotation (Callison-Burch and Dredze, 2010; Zaidan and Callison-Burch, 2011b). MTurk provides an environment in which "requesters" can set up Human Intelligent Tasks (HITs) to be performed by "workers". The tasks typically require human knowledge. To interact with the workers, the requester creates an interactive website that allows the "worker" to perform the task. We randomly divided the sentences into groups of 10 and additionally provided 2 controls, which trusted annotators had previously labeled. The controls were selected from the Arabic Online Commentary Dataset (Zaidan and Callison-Burch, 2011a). The controls were only marked for dialect versus MSA, a much easier task than dialect identification. Each HIT required a worker to give a judgment to specify which dialect (including MSA) the message was written in as well as the "dialectness" of the tweet. We gave each worker 60 minutes to complete the task from start to finish although the average time was substantially less than this. Additionally, we required the workers to fill out a brief survey about their native language, place of birth and current place of residence.

To ensure the quality of the annotation, we monitored performance on the controls. For full payment, we required that the worker achieve 85% accuracy on the control. Most workers, however, achieved well above that mark. Workers who performed better than chance, but below the 85% threshold were compensated proportionally for the number of correct annotations

Two main groups of HITs were put up: one for the newspaper commentary and one for the twitter data. The design was the same and is pictured in Figure 2. Both of these hits required approval to access, which was granted on the basis of performance on an identical "test" HIT with the exception that no restriction was applied. Workers who performed well on the "test" were then granted the qualification on the main HIT. Each group consisted of 100,000 messages from the commentary data set and from Twitter respectively.

	Comments	Words
Al-Ghad	4811	100K
Al-Riyadh	6307	105K
El-Youm El-Sabe3	8927	220K
Al-Wifaq	254	8K
Echourouk	6940	150K

Figure 3: Newspaper commentary annotated on MTurk as having high dialectal content

	Tweets	Words
Levantine	1594	22K
Gulf	36330	611K
Egyptian	2052	27K
Iraqi	154	4K
Maghrebi	99	2K

Figure 4: Tweets annotated on MTurk as having high dialectal content

230 workers attempted the "test" HIT. Of those 230, 23 passed the initial qualification. All of these workers self-reported speaking Arabic natively, although not all resided in Arabic speaking countries. The 23 workers completed all the HITs in both groups over a 4 month period. Figures 5 and 6 show the performance of the individual workers on newspaper commentary annotation task and the tweet annotation task.

7 Automatic Dialect Classification

In addition to the corpus, this work also budges the state of the art in dialect identification. Arabic dialect identification is effectively the task of language identification, with the added difficulty that dialects are far more similar. In contrast to (Zaidan and Callison-Burch, 2011a) , which made use of a language model for the task, we consider Naive Bayes and Support Vector Machines. We further extend the task to include 5 dialects in total, as opposed to three.

Country	# Hits	Acc.	Prec.	Rec.	F1
Algeria	198	0.96	0.97	0.98	0.98
Algeria	2	1.0	1.0	1.0	1.0
Algeria	4	1.0	1.0	1.0	1.0
Egypt	11	0.91	0.89	1.0	0.94
Egypt	1264	0.95	0.96	0.98	0.97
Egypt	743	0.91	0.9	0.99	0.94
Georgia	43	0.93	0.92	1.0	0.96
Greece	193	0.95	0.98	0.95	0.97
Jordan	1228	0.97	0.96	0.99	0.98
Jordan	3160	0.96	0.96	0.99	0.97
Lebanon	62	0.95	0.98	0.96	0.97
Morocco	673	0.95	0.95	0.99	0.97
Morocco	732	0.95	0.99	0.95	0.97
Tunisia	2104	0.92	0.97	0.93	0.95
Tunisia	28	0.82	0.98	0.82	0.89
Tunisia	3766	0.96	0.97	0.97	0.97
U.K.	1193	0.96	0.97	0.98	0.97
U.S.	1668	0.95	0.94	0.99	0.96
U.S.	20	0.97	0.97	1.0	0.99
U.S.	2407	0.97	0.96	0.99	0.98
U.S.	5	0.9	1.0	0.88	0.93

Figure 5: Workers' Performance on Twitter HIT

Country	# Hits	Acc.	Prec.	Rec.	F1
Algeria	1048	.96	.96	.93	.94
Egypt	137	.94	.86	.97	.91
Egypt	139	.9	.74	1.0	.85
Egypt	141	.96	.95	.94	.95
Egypt	4	.75	.5	1.0	.67
Egypt	461	.96	.95	0.92	.94
Egypt	722	.94	.92	.92	.92
Greece	416	.9	.97	.79	.87
Jordan	272	.82	.78	.7	0.74
Jordan	417	.97	.96	.95	.96
Jordan	50	.91	.97	.8	.88
Jordan	93	.96	.94	.96	.95
Lebanon	136	.96	.96	.92	.94
Morocco	473	.97	.94	.95	.94
Tunisia	1404	.9	.96	.8	.87
Tunisia	260	.91	.91	.84	.88
Tunisia	727	.84	.89	.74	.81
U.K.	3	1.0	1.0	1.0	1.0
U.K.	828	.95	.93	.91	.92
U.S.	1337	.95	.89	.98	.93
U.S.	30	.95	.96	.92	.94
U.S.	899	.97	.97	.94	.95

Figure 6: Workers' Performance on Arabic Commentary HIT

كلمة الى قادة القائمة العراقية ..لا تهتموا لما جرى ونحن واثقون من انتصاركم ايا كانت النتائج ..من المهم ان نبني تحالفات برلمانية متينة تبنى على المصلحة الوطنية اولا بلا تنازلات عن الثوابت والمقومات ارمو الماضي وراء ظهوركم وانطلقوا لبناء عراق آمن مزدهر عربي ولا تلتفتوا الى المشككين فقد نلتم اصوات شعبكم وفي اعناقكم امانة هذا الشعب ممن صوت لكم ومن صوت لغيركم فهم كلهم ابناء العراق....شكرا لكم وفقكم الله لخدمة بلدكم ووسام اخر من اوسمة العز والفخر....

Figure 7: Sample Commentary

8 Experiments

We made use of the Python-based Machine Learning library Scikit-learn to train a classifier for the Arabic social media data (Pedregosa et al., 2011). Scikit-learn provides a suite of supervised learning algorithms that can be readily substituted in a generic framework. We used unigram, bigram, and trigram features for the model in combination with the two learning algorithms: SVM with a linear kernel and Naive Bayes. We were primarily interested in the binary classification problem dialect versus MSA with all five dialects. We performed 10-fold cross-validation with each algorithm. Accuracy is reported in figures 8 and 8.
We observe that the unigram models typically outperform the higher order models; the additional features hurt the model performance. This is surprising; it is possible that

Figure 2: Screenshot of HIT

	Egy.	Lev.	Mag.	Gulf	Iraqi
NB Uni	**.89**	**.79**	**.92**	**.88**	**.87**
NB Bi	.88	.78	.89	.84	.66
NB Tri	.88	.77	.88	.84	.65
SVM Uni	.88	.78	.89	.85	85
SVM Bi	.87	.75	.87	.82	.79
SVM Tri	.87	.74	.87	.82	.79

Figure 8: Experiments on Extended AOC (accuracy reported)

	Egy.	Lev.	Mag.	Gulf	Iraqi
NB Uni	**.84**	**.84**	.70	**.87**	**.79**
NB Bi	**.84**	**.84**	**.71**	.86 .	**.79**
NB Tri	.83	.74	.70	.86	.65
SVM Uni	.80	.81	.65	.86	75
SVM Bi	.79	.76	.57	.86	.62
SVM Tri	.77	.76	.57	.86	.62

Figure 9: Experiments on Twitter (accuracy reported)

dialectal words do not typically occur in sequence and it is therefore bigrams and trigrams do not help performance. Another explanation could stem from the informal nature of the text. The varying orthorgrahical conventions typically increase sparsity and make it more difficult to estimate the true n-gram probabilities. We also compared our model to Zaidan and Callison-Burch (2011a). We trained our best performing model, Naive Bayes with unigram features, on the data released with Zaidan and Callison-Burch (2011a) and showed significant improvements. The numbers are reported in table 8.

9 Future Work

The problem of Arabic dialect identification is still very much an open problem despite the introduction of an additional annotated corpus in this work. While language identification is often considered a solved problem, McNamee (2005) points out that problem can be made arbitrarily difficult by using informal text, many languages, short text, and unbalanced data. The Arabic dialect identification task inherently embodies many of these attributes. Dialectal Arabic occurs exclusively in informal text, which is often short due to the nature of social media. Additionally, the task of Arabic dialect identification with short messages may often be impossible. Just as an informal message containing only the text *por que?* could either be Spanish or Portuguese, many short Arabic texts are inherently ambiguous. Future work should attempt to define an empirical error in human classification of data, as it is unlikely rates similar to those seen in traditional language identification can be achieved. Additional efforts should also focus on isolating dialectal features from topical features. As the source of each dialect was a single newspaper, it is reasonable to expect that differences in n-gram counts are due only to the topical coverage of each newspaper and not to inherent differences in the dialects. It is important to mitigate the boost in accuracy attributable to these features to get a better sense of the models' performance on new data. One possible solution to this problem is to scrape comments from multiple websites in the same country and compare the corresponding accuracy. Also leveraging the vast of amount of unannotated data is of interest. It is financially unfeasible to annotate all the data scraped from online fora, but it nevertheless may be possible to improve performance through the use of semi-supervised techniques. It is also necessary to extend the coverage of corpus to additional dialects. No data was collected from newspapers from Sudan, which has its own distinct dialect of Arabic (Comrie, 2013). Additionally, there are stark contrasts between various North African dialects, such as between Tunisian and Moroccan (Comrie, 2013).

	NB	Zaidan et al. (2011a)
MSA vs. Lev.	86.6	79.6
MSA vs. Gulf	82.7	75.1
MSA vs. Egy.	86.6	80.9

Figure 10: Experiments on Extended AOC

Acknowledgements

This material is based on research sponsored by a DARPA Computer Science Study Panel phase 3 award entitled "Crowdsourcing Translation" (contract D12PC00368). The views and conclusions contained in this publication are those of the authors and should not be interpreted as representing official policies or endorsements by DARPA or the U.S. Government. This research was supported by the

Johns Hopkins University Human Language Technology Center of Excellence.

10 References

Rania Al-Sabbagh and Roxana Girju. 2010. Mining the web for the induction of a dialectical Arabic lexicon. In *LREC*.

Rania Al-Sabbagh and Roxana Girju. 2012. YADAC: Yet another dialectal Arabic corpus. In *LREC*, pages 2882–2889.

Chris Callison-Burch and Mark Dredze. 2010. Creating speech and language data with amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 1–12. Association for Computational Linguistics.

Bernard Comrie. 2013. *The Major Languages of South Asia, the Middle East and Africa*. Routledge.

Mona Diab, Nizar Habash, Owen Rambow, Mohamed Altantawy, and Yassine Benajiba. 2010. COLABA: Arabic dialect annotation and processing. In *LREC Workshop on Semitic Language Processing*, pages 66–74.

Jacob Eisenstein. 2013. Phonological factors in social media writing.

Heba Elfardy and Mona Diab. 2012a. Token level identication of linguistic code switching.

Heba Elfardy and Mona T Diab. 2012b. AIDA: Automatic identification and glossing of dialectal Arabic.

Heba Elfardy and Mona T Diab. 2012c. Simplified guidelines for the creation of large scale dialectal Arabic annotations. In *LREC*, pages 371–378.

Heba Elfardy and Mona Diab. 2013. Sentence level dialect identification in Arabic.

Heba Elfardy, Mohamed Al-Badrashiny, and Mona Diab. 2013. Code switch point detection in Arabic. In *Natural Language Processing and Information Systems*, pages 412–416. Springer.

Paul McNamee. 2005. Language identification: a solved problem suitable for undergraduate instruction. *Journal of Computing Sciences in Colleges*, 20(3):94–101.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830.

Stephen Tratz, Douglas Briesch, Jamal Laoudi, and Clare Voss. 2013. Tweet conversation annotation tool with a focus on an Arabic dialect, Moroccan Darija. *LAW VII & ID*, page 135.

Yi Yang and Jacob Eisenstein. 2013. A log-linear model for unsupervised text normalization. In *Proc. of EMNLP*.

Omar F Zaidan and Chris Callison-Burch. 2011a. 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: short papers-Volume 2*, pages 37–41. Association for Computational Linguistics.

Omar F Zaidan and Chris Callison-Burch. 2011b. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1220–1229. Association for Computational Linguistics.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F Zaidan, and Chris Callison-Burch. 2012. Machine translation of Arabic dialects. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 49–59. Association for Computational Linguistics.