The Arabic Online Commentary Dataset:
an Annotated Dataset of Informal Arabic with High Dialectal Content

Omar F. Zaidan and **Chris Callison-Burch**
Dept. of Computer Science, Johns Hopkins University
Baltimore, MD 21218, USA
{ozaidan,ccb}@cs.jhu.edu

Abstract

The written form of Arabic, *Modern Standard Arabic* (MSA), differs quite a bit from the spoken dialects of Arabic, which are the true "native" languages of Arabic speakers used in daily life. However, due to MSA's prevalence in written form, almost all Arabic datasets have predominantly MSA content. We present the *Arabic Online Commentary Dataset*, a 52M-word monolingual dataset rich in dialectal content, and we describe our long-term annotation effort to identify the dialect level (and dialect itself) in each sentence of the dataset. So far, we have labeled 108K sentences, 41% of which as having dialectal content. We also present experimental results on the task of automatic dialect identification, using the collected labels for training and evaluation.

1 Introduction

The Arabic language is characterized by an interesting linguistic dichotomy, whereby the written form of the language, *Modern Standard Arabic* (MSA), differs in a non-trivial fashion from the various *spoken varieties* of Arabic. As the variant of choice for written and official communication, MSA content significantly dominates dialectal content, and in turn MSA dominates in datasets available for linguistic research, especially in textual form.

The abundance of MSA data has greatly aided research on computational methods applied to Arabic, but only the MSA variant of it. A state-of-the-art Arabic-to-English machine translation system performs quite well when translating MSA source sentences, but often produces incomprehensible output when the input is dialectal. For example, most words

Figure 1: Two roughly equivalent Arabic sentences, one in MSA and one in Levantine Arabic, translated by the same MT system into English. An acceptable translation would be *When will we see this group of criminals undergo trial* (or *tried*)?. The MSA variant is handled well, while the dialectal variant is mostly transliterated.

of the dialectal sentence of Figure 1 are transliterated.[1] Granted, it is conceivable that processing dialectal content is more difficult than MSA, but the main problem is the lack of dialectal training data.[2]

In this paper, we present our efforts to create a dataset of **dialectal** Arabic, the **Arabic Online Commentary Dataset**, by extracting reader commentary from the online versions of three Arabic newspapers, which have a high degree (about half) of dialectal content (Levantine, Gulf, and Egyptian). Furthermore, we describe a long-term crowdsourced effort to have the sentences labeled by Arabic speakers for the level of dialect in each sentence and the dialect itself. Finally, we present experimental results on the task of automatic dialect classification with systems trained on the collected dialect labels.

---

[1] The high transliteration rate is somewhat alarming, as the first two words of the sentence are relatively frequent: *AymtY* means 'when' and *rH* corresponds to the modal 'will'.

[2] It can in fact be argued that MSA is the variant with the more complex sentence structure and richer morphology.
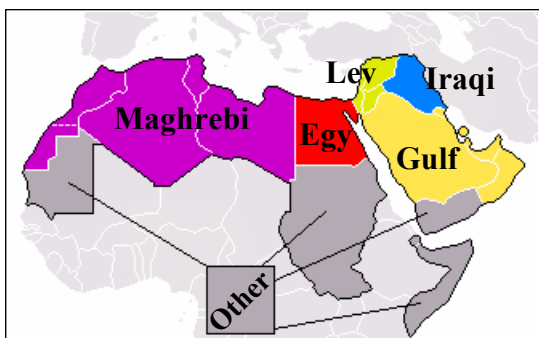
Figure 2: One possible breakdown of spoken Arabic into dialect groups: Maghrebi, Egyptian, Levantine, Gulf, and Iraqi. Habash (2010) also gives a very similar breakdown.

2 The AOC Dataset

Arabic is the official language in over 20 countries, spoken by more than 250 million people. The official status only refers to a written form of Arabic known as *Modern Standard Arabic* (MSA). The *spoken dialects* of Arabic (Figure 2) differ quite a bit from MSA and from each other. The dominance of MSA in available Arabic text makes dialectal Arabic datasets hard to come by.[3]

We set out to create a dataset of dialectal Arabic to address this need. 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 weblogs, forums, and chat transcripts. However, weblogs usually contain relatively little data, and a writer might use dialect in their writing only occasionaly, forums usually have content that is of little interest or relevance to actual applications, and chat transcripts are difficult to obtain and extract.

We instead diverted our attention to **online commentary** by readers of online content. This source of data has several advantages:

- A large amount of data, with more data becoming available on a daily basis.
- The data is publicly accessible, exists in a structured, consistent format, and is easy to extract.
- A high level of topic relevance.

News Source	Al-Ghad	Al-Riyadh	Al-Youm Al-Sabe'
# articles	6.30K	34.2K	45.7K
# comments	26.6K	805K	565K
# sentences	63.3K	1,686K	1,384K
# words	1.24M	18.8M	32.1M
comments/article	4.23	23.56	12.37
sentences/comment	2.38	2.09	2.45
words/sentence	19.51	11.14	23.22

Table 1: A summary of the different components of the AOC dataset. Overall, 1.4M comments were harvested from 86.1K articles, corresponding to 52.1M words.

- The prevalence of dialectal Arabic.

The *Arabic Online Commentary* dataset that we created was based on reader commentary from the online versions of three Arabic newspapers: **Al-Ghad** from Jordan, **Al-Riyadh** from Saudi Arabia, and **Al-Youm Al-Sabe'** from Egypt.[4] The common dialects in those countries are Levantine, Gulf, and Egyptian, respectively.

We crawled webpages corresponding to articles published during a roughly-6-month period, covering early April 2010 to early October 2010. This resulted in crawling about 150K URL's, 86.1K of which included reader commentary (Table 1). The data consists of 1.4M comments, corresponding to 52.1M words.

We also extract the following information for each comment, whenever available:

- The URL of the relevant newspaper article.
- The date and time of the comment.
- The author ID associated with the comment.[5]
- The subtitle header.[5]
- The author's e-mail address.[5]
- The author's geographical location.[5]

The AOC dataset (and the dialect labels of Section 3) is fully documented and publicly available.[6]

---

[3]The problem is somewhat mitigated in the speech domain, since dialectal data exists in the form of phone conversations and television program recordings.

[4]URL's: www.alghad.com, www.alriyadh.com, and www.youm7.com.

[5]These fields are provided by the author.

[6]Data URL: http://cs.jhu.edu/~ozaidan/AOC/. The release also includes all sentences from *articles* in the 150K crawled webpages.

3 Augmenting the AOC with Dialect Labels

We have started an ongoing effort to have each sentence in the AOC dataset labeled with dialect labels. For each sentence, we would like to know whether or not it has dialectal content, how much dialect there is, and which variant of Arabic it is. Having those labels would greatly aid researchers interested in dialect by helping them focus on the sentences identified as having dialectal content.

3.1 Amazon's Mechanical Turk

The dialect labeling task requires knowledge of Arabic at a native level. To gain access to native Arabic speakers, and a large number of them, we crowd-sourced the annotation task to Amazon's Mechanical Turk (MTurk), an online marketplace that allows "Requesters" to create simple tasks requiring human knowledge, and have them completed by "Workers" from all over the world.

3.2 The Annotation Task

Of the 3.1M available sentences, we selected a 'small' subset of 142,530 sentences to be labeled by MTurk Workers.[7] We kept the annotation instructions relatively simple, augmenting them with the map from Figure 2 (with the Arabic names of the dialects) to illustrate the different dialect classes.

The sentences were randomly grouped into 14,253 sets of 10 sentences each. When a Worker chooses to perform our task, they are shown the 10 sentences of some random set, on a single HTML page. For each sentence, they indicate the level of dialectal Arabic, and which dialect it is (if any). We offer a reward of $0.05 per screen, and request each one be completed by three distinct Workers.

3.3 Quality Control

To ensure high annotation quality, we insert two additional *control* sentences into each screen, taken from the *article bodies*. Such sentences are almost always in MSA Arabic. Hence, a careless Worker can be easily identified if they label many control sentences as having dialect in them.

---

[7]There are far fewer sentences available from *Al-Ghad* than the other two sources (fourth line of Table 1). We have taken this imbalance into accout and heavily oversampled *Al-Ghad* sentences when choosing sentences to be labeled.

News Source	# MSA sentences	# words	# dialectal sentences	# words
Al-Ghad	18,947	409K	11,350	240K
Al-Riyadh	31,096	378K	20,741	288K
Al-Youm Al-Sabe'	13,512	334K	12,527	327K
ALL	63,555	1,121K	44,618	855K

Table 2: A breakdown of sentences for which ≥ 2 annotators agreed on whether dialectal content exists or not.

Another effective method to judge a Worker's quality of work is to examine their label distribution within each news source. For instance, within the sentences from *Al-Youm Al-Sabe'*, most sentences judged as having dialectal content should be classified as Egyptian. A similar strong prior exists for Levantine within *Al-Ghad* sentences, and for Gulf within *Al-Riyadh* sentences.

Using those two criteria, there is a very clear distinction between Workers who are faithful and those who are not (mostly spammers), and 13.8% of assignments are rejected on these grounds and re-posted to MTurk.

3.4 Dataset Statistics

We have been collecting labels from MTurk for a period of about four and a half months. In that period, 11,031 HITs were performed to completion (corresponding to 110,310 sentences, each labeled by three distinct annotators). Overall, 455 annotators took part, 63 of whom judged at least 50 screens. Our most prolific annotator completed over 6,000 screens, with the top 25 annotators supplying about 80% of the labels, and the top 50 annotators supplying about 90% of the labels.

We consider a sentence to be dialectal if it is labeled as such by at least two annotators. Similarly, a sentence is considered to be MSA if it has at least two MSA labels. For a small set of sentences (2%), no such agreement existed, and those sentences were discarded (they are mostly sentences identified as being non-Arabic). Table 2 shows a breakdown of the rest of the sentences.[8]

---

[8]Data URL: `http://cs.jhu.edu/~ozaidan/RCLMT/`.

Classification Task	Accuracy (%)	Precision (%)	Recall (%)
Al-Ghad MSA vs. LEV	79.6	70.6	78.2
Al-Riyadh MSA vs. GLF	75.1	66.9	74.6
Al-Youm Al-Sabe' MSA vs. EGY	80.9	77.7	84.4
MSA vs. dialect	77.8	71.2	77.6
LEV vs. GLF vs. EGY	83.5	N/A	N/A
MSA vs. LEV vs. GLF vs. EGY	69.4	N/A	N/A

Table 3: Accuracy, dialect precision, and dialect recall (10-fold cross validation) for various classification tasks.

4   Automatic Dialect Classification

One can think of dialect classification as a language identification task, and techniques for language identification can be applied to dialect classification. We use the collected labels to investigate how well a machine learner can distinguish dialectal Arabic from MSA, and how well it can distinguish between the different Arabic dialects.

We experiment with a language modeling approach. In a classification task with c classes, we build c language models, one per class. At test time, we score a test sentence with all c models, and choose the class label of the model assigning the highest score (i.e. lowest perplexity). We use the SRILM toolkit to build word trigram models, with modified Kneser-Ney as a smoothing method, and report the results of 10-fold cross validation.

Table 3 illustrates the performance of this method under various two-, three-, and four-way scenarios. We find that it is quite good at distinguishing each dialect from the corresponding MSA content, and distinguishing the dialects from each other.

We should note that, in practice, accuracy is probably not as important of a measure as (dialect) precision, since we are mainly interested in identifying dialectal data, and much less so MSA data. To that end, one can significantly increase the precision rate (at the expense of recall, naturally) by biasing classification towards MSA, and choosing the dialectal label only if the ratio of the two LM scores exceeds a certain threshold. Figure 3 illustrates this tradeoff for the classification task over *Al-Ghad* sentences.
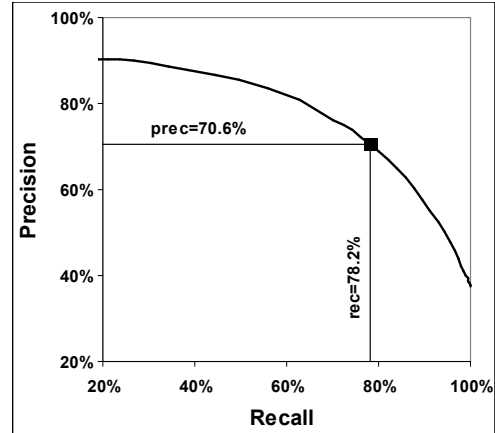

Figure 3: Dialect precision vs. recall for the classification task over *Al-Ghad* sentences (MSA vs. Levantine). The square point corresponds to the first line in Table 3.

5   Related Work

The COLABA project (Diab et al., 2010) is another large effort to create dialectal Arabic resources (and tools). They too focus on online sources such as blogs and forums, and use information retrieval tasks for measuring their ability to properly process dialectal Arabic content.

The work of Irvine and Klementiev (2010) is similar to ours in spirit, as they too use MTurk to find annotators with relatively uncommon linguistic skills, to create translation lexicons between English and 42 rare languages. In the same vein, Zaidan and Callison-Burch (2011) solicit English translations of Urdu sentences from non-professional translators, and show that translation quality can rival that of professionals, for a fraction of the cost.

Lei and Hansen (2011) build Gaussian mixture models to identify the same three dialects we consider, and are able to achieve an accuracy rate of 71.7% using about 10 hours of speech data for training. Biadsy et al. (2009) utilize a much larger dataset (170 hours of speech data) and take a phone recognition and language modeling approach (Zissman, 1996). In a four-way classification task (with Iraqi as a fourth dialect), they achieve a 78.5% accuracy rate. It must be noted that both works use *speech* data, and that dialect identification is done on the *speaker* level, not the sentence level as we do.

6 Current and Future Work

We have already utilized the dialect labels to identify dialectal sentences to be *translated* into English, in an effort to create a Dialectal Arabic-to-English parallel dataset (also taking a crowdsourcing approach) to aid machine translation of dialectal Arabic.

Given the recent political unrest in the Middle East (early 2011), another rich source of dialectal Arabic are Twitter posts (e.g. with the `#Egypt` tag) and discussions on various political Facebook groups. Here again, given the topic at hand and the individualistic nature of the posts, they are very likely to contain a high degree of dialectal data.

References

Fadi Biadsy, Julia Hirschberg, and Nizar Habash. 2009. Spoken arabic dialect identification using phonotactic modeling. In *Proceedings of the EACL Workshop on Computational Approaches to Semitic Languages*, pages 53–61.

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.

Nizar Y. Habash. 2010. *Introduction to Arabic Natural Language Processing*. Morgan & Claypool.

Ann Irvine and Alexandre Klementiev. 2010. Using Mechanical Turk to annotate lexicons for less commonly used languages. In *Proceedings of the NAACL HLT Workshop on Creating Speech and Language Data With Amazon's Mechanical Turk*, pages 108–113.

Yun Lei and John H. L. Hansen. 2011. Dialect classification via text-independent training and testing for arabic, spanish, and chinese. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(1):85–96.

Omar F. Zaidan and Chris Callison-Burch. 2011. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of ACL (this volume)*.

Marc A. Zissman. 1996. Comparison of four approaches to automatic language identification of telephone speech. *IEEE Transactions on Speech and Audio Processing*, 4(1):31–44.