YouDACC: the Youtube Dialectal Arabic Commentary Corpus

Ahmed Salama, Houda Bouamor, Behrang Mohit and Kemal Oflazer

Carnegie Mellon University in Qatar
ahabdelh@microsoft.com, {hbouamor, behrang}@cmu.edu, ko@cs.cmu.edu

Abstract

In the Arab world, while Modern Standard Arabic is commonly used in formal written context, on sites like Youtube, people are increasingly using Dialectal Arabic, the language for everyday use to comment on a video and interact with the community. These user-contributed comments along with the video and user attributes, offer a rich source of multi-dialectal Arabic sentences and expressions from different countries in the Arab world. This paper presents YOUDACC, an automatically annotated large-scale multi-dialectal Arabic corpus collected from user comments on Youtube videos. Our corpus covers different groups of dialects: Egyptian (EG), Gulf (GU), Iraqi (IQ), Maghrebi (MG) and Levantine (LV). We perform an empirical analysis on the crawled corpus and demonstrate that our location-based proposed method is effective for the task of dialect labeling.

Keywords: Dialectal Arabic, multi-dialectal corpus, Youtube

1. Introduction

With the growth of Web2.0, people increasingly express and hare their opinion through social media. Youtube, for example, is one of the most known and widely used participatory sites which provides a new generation of short video sharing service. An analysis of this platform reveals a large amount of community feedback through comments for published videos as well as through meta ratings for these comments (Siersdorfer et al., 2010).

In the Arab world, while MSA (Modern Standard Arabic) is commonly used in formal written context (e.g., newspapers, academic writing, etc.), on sites like Youtube, people are increasingly using Dialectal Arabic (DA), the language for everyday use (e.g., Egyptian, Iraqi, Gulf, etc.) to comment on a video and interact with the community. These user-contributed comments along with the video and user attributes, offer a rich source of multi-dialectal Arabic sentences and expressions from different countries in the Arab world.

In Youtube, a video comment has a title, a content and an author. Each author has a profile containing several pieces of information about her, e.g., hobbies, occupation, location, etc. For Arabic written comments, the user's location could be exploited to assign to each comment label indicating its dialectal class corresponding generally to the geographical location from which the comment was entered.

In this paper we present YOUDACC, an automatically annotated large-scale multi-dialectal Arabic corpus collected from user comments on Youtube videos. Our corpus covers the different groups of dialects defined by Habash (2010): Egyptian (EG), Gulf (GU), Iraqi (IQ), Maghrebi (MG) and Levantine (LV). We perform an empirical analysis on the crawled corpus and demonstrate that our location-based proposed method is effective for the task of dialect labeling.

Unfortunately due to YouTube's policy of data distribution, we are not able to release the dataset publicly. Here we present our framework and analysis of the data in great details. We believe this can be useful for other researchers to replicate and use in future research.

The remainder of this paper is organized as follows. Section 2 outlines the challenges of Arabic Dialects processing and reviews previous efforts in building dailectal Arabic resources. Section 3 explains our motivation behind the use of Youtube as a source of knowledge. Our approach for building YouDACC and annotate it is explained in Section 4. In section 5, we report our initial experiments on dialect classification. Finally, we conclude and describe our future work in Section 6.

2. Dialectal Arabic processing

Dialectal Arabic (DA) or the Arabic languages used for everyday speaking is the result of a linguistic and lexical interference between the Modern Standard Arabic (MSA) and the local and neighboring languages, or different cultural influences caused mainly by colonization and the media. DA is nowadays emerging as the language of informal communication online; in emails, blogs, discussion forums, chats,

The first author of this paper is currently working for Microsoft Corporation.

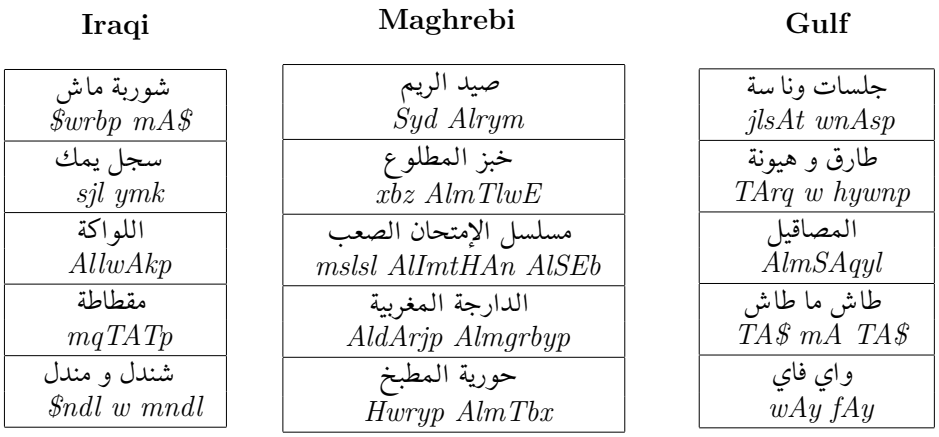

Figure 1: Examples of region specific keywords used for looking up some videos dialects

SMS, etc., as the media which is closer to the spoken form of language. There are several varieties of spoken Arabic language with substantial differences. Every dialect is unique in structure and vocabulary so that it can be a real challenge for speakers of different nationalities to understand each other. A possible breakdown of these varieties into dialect groups was proposed by Habash (2010) where five major Arabic dialects were considered: Egyptian, Gulf, Maghrebi, Levantine and Iraqi. More fine-grained classes can be found within these major classes. As well, there are other minor dialects each can be considered a different class on its own, such as: Sudanese or Yemeni.

Recently, automatic Arabic dialect processing has attracted a considerable amount of research in NLP. Most of these focus on (i) developing DA to English machine translation systems (Zbib et al., 2012; Salloum and Habash, 2013; Sajjad et al., 2013) (ii) creating processing tools (Habash et al., 2013) (iii) or creating different resources.

In the COLABA project, for example, Diab et al. (2010) used online sources such as blogs and forums, and applied these to information retrieval tasks for measuring their ability to properly process dialectal Arabic content. Unfortunately, this corpus is not publicly available. More recently, Zaidan and Callison-Burch (2011b) crawled the websites of three Arabic newspapers and extracted reader commentary on their articles. The resulting Arabic Online Commentary dataset much of which is in dialectal Arabic. The dialectal annotations were collected from the crowd using Mechanical Turk.

Arabic dialects are constantly changing and new words and figures of speech, mainly drawn from Western languages (e.g., English, French) are being added. The continuing and increasing stream of comments in Youtube, makes it an interesting source of dialectal sentences and expressions that could be collected on a regular basis and used to update the lexicon of each of these dialects with the newly appearing words. Contrarily to news websites, in which contributions are mostly done by educated people, Youtube videos have a wider audience and recieve comments from a variety of users with different education backgrounds.

3. Youtube: a rich source of dialectal information

Youtube has become one of the largest and most popular participatory media in the online environment, providing a new generation of video sharing service. Recent studies showed that Youtube alone comprises approximately 20% of all HTTP traffic, or nearly 10% of the whole traffic on the Internet (Cheng et al., 2007).
The growth and success of this sharing environment is driven by the large-scale user involvement in the content annotation. Logged-in users can provide category tags or add comments in a response to a video. This creates threaded discussions containing generally a large number of comments written in different languages, usually corresponding to the language of the video. DA are largely covered in Youtube videos and comments.
Examining video sharing practices on Youtube shows that the Arabic speaking community is quite active in creating channels, sharing videos, and commenting them. Figure 2 illustrates the distribution of Youtube Arabic comments by country.
By mining the users' comment activity and interactions about videos, and exploiting different Youtube features describing videos and comments, we built a Multi-dialectal corpus covering different dialectal groups in which every sentence is labeled with its corresponding dialect.

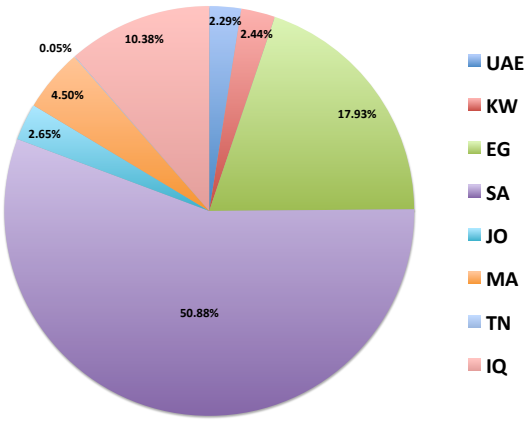

Figure 2: Distribution of Youtube Arabic comments by country

Youtube video and comments have been already exploited as a source of information in different applications. Mukherjee and Bhattacharyya (2012) propose a weakly supervised system for video categorization. (Zhang et al., 2011) and (Wang et al., 2010) use Youtube video comments to analyze the quality of user-contributed comments and mitigate the negative impact of spam and low-quality comments on the sustainability of the social web.

4. Corpus collection and annotation

The Youtube Data API[1] allows users to search for videos that match specific criteria, such as videos published by a particular author or in a specific category. It also lets users retrieve video comments and user profiles. In this study, we simply used a selected set of Arabic keywords[2] to search in all Youtube videos for those commented by users from the Arab world and retrieve the user comments specific to each of our five Arabic dialects.

For each dialect, the list of keywords is provided by native speakers describing the videos they usually watch on Youtube. This narrows our search to the region in which a given dialect is spoken. Figure 1 provides some sample keywords used for harvesting comments for each dialect. For each keyword, the returned video by Youtube, along with their video IDs and web page URLs, were stored.[3]

Given a video url, the video title and the first 999 comments[4] (when available) on the video are retrieved along with their authors, timestamps and comment ratings.

Only user comments written in Arabic letters are kept. They usually contain different kinds of orthographic and typographical errors such as the use of special and decorative characters, letter duplication caused by a speech effect used generally for emphasis, word duplication, missing or added spaces, extra punctuation etc. Following (2010), we designed a procedure to clean up these spelling errors and normalize the comments.

Table 1 provides various statistics for the sentences collected for each dialect. It is interesting to note that the MG dialect has the longest sentences in terms of words. This could be explained by the fact that people in this region tend to write sentences in a Franco-Arabic way. These sentences are discarded later in a processing step. The only sentences we keep are the dialectal ones written in Arabic script. Most of those correspond to quotations from MSA resources. Some examples of comments obtained for each dialect are given in Table 3 .

Table 2 gives the number of unique comments collected for each dialect after normalization (**# of sents**). This table reports the number of sentences which are exclusive for each of the dialects (**# of sents per dialect**).[5] For example, $323,925$ comments out of $661,994$ are found only in the Gulf part of our corpus.

	# of sents	# of sents per dialect
GU	655,578	322,765
EG	201,528	165,579
IQ	118,675	68,869
MG	48,162	32,215
LV	61,651	41,389

Table 2: Number of sentences collected for each dialect group.

In addition to building a multi-dialectal corpus covering several Arabic dialects, our goal is to annotate each sentence/comment extracted, with its corresponding dialect class by exploiting some of the user profile features provided in Youtube and take advantage of the list of the keywords defined by the native

[1]Available at: `https://developers.google.com/Youtube`

[2]corresponding generally to TV series and programs, video clips, etc. covering various categories: entertainment, romance, sports, etc.

[3]We observed that some keywords yield more relevant video than others.

[4]We are constrained by the number of comments provided by the Youtube API.

[5]Some expressions can be shared among two more dialects. Such cases are excluded in our count.

	#tokens	#Sentences	Avg sent. length$_{words}$	Avg sent. length$_{characters}$
GU	2,416,105	322,765	7.48	37.78
EG	2,287,892	165,579	13.81	71.07
IQ	852,438	68,869	12.37	65.01
MG	553,900	32,215	17.19	90.73
LV	411,203	41,389	9.93	50.37

Table 1: Statistics on different parts of our dialectal Arabic corpus

speaker.

In Youtube, a user's profile contains information about each user (video owners or comment authors), such as the user's name, age, occupation, hometown and location. This personal information has been entered by the user for publication on Youtube. We first retrieve different user profiles with their features. Then, we automatically label each comment based on the geographic location of its author. After examining several samples for each dialect, we realized that the location attribute of the user could sometimes be misleading; especially when the video is related to a region different from the location of that user (e.g., expats speaking a dialect different from the one spoken in the reported location). In order to filter out the problematic comments, we keep only those for which the user's location matches the region of the keywords used to retrieve the video.

In order to assess the quality of sentences extracted for each dialect, we randomly selected 1,000 sentences from each dialectal corpus and asked two native speakers to read the sentences and give their judgment by answering these questions:*"Does this sentence look familiar to you?" "Is it in your dialect? In MSA or other dialect/language?"*.

The results for each dialect illustrated in Figure 3 show that the approach we followed to assign a dialect class to each comment is efficient for most of the dialect groups we are studying in this work. The low percentage of the MG dialectal sentences in this sample is due to the presence of sentences written mainly in MSA (40%).

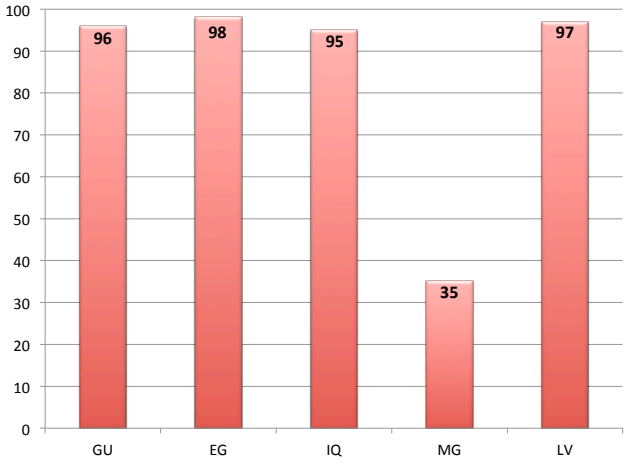

Figure 3: Percentage of correct labels in a sample of 1,000 sentences of each dialect group.

5. Experiments on dialect classification

The most straightforward application in which YouDACC can be used is Dialect Identification. We ran different experiments on dialect classification in order to show the effectiveness and usefulness of our data for such task.

Following (2011a), we formulate the dialect identification problem as a multi-class classification task with five classes, using a language modelling approach. We build several language models, one per class. To test our classifier, we score a given test sentence with all the models, and we assign it to the class of the model assigning the highest score.

Table 4. compares the different models used and their performance for identifying the dialect of a given sentence. We measure the percentage of sentences which are correctly identified using a given model. For example, if a model predicts four correct EG sentences out of 10, its recall is of 40%.

The best performance is achieved when we use word-bigram model. This result is not surprising, since we take into account the local context of each word. It is also important to note that the letter 4-graph is an effective model to distinguish between dialects: 93.10% of EG sentences were correctly identified using this model compared to 62.50% while using only the first letter. In the case of Maghrebi dialects, using the 4-letter based model outperforms the rest of models. Our results confirm that using letter-based models are effective enough for dialect identification, regardless the domain and the nature of the sentences.

Dialect	Examples
GU	يعطيك العافيه يارب ممكن طلب اذا مافي مانع او ازعاج ابغي قصيده او زفه باسم ام حاتم. yETyk AlEAfyh yArb mmkn Tlb A*A mAfy mAnE Aw AzEAj Abgy qSydh Aw zfh bAsm Am Hatm. God bless you, I want to request -if you don't mind- a peom or a song under the name of Um Hatem.
EG	انا بنفسي شفت النهارده في رابعه واحد امين شرطه كان عاوز يخش الاعتصام. AnA bnfsy $ft AlnhArdh fy rAbEh wAHd Amyn $rTh kAn EAwz yx$ AlAEtSAm. I -myself- saw some policemen in Rabaa who wanted to enter the Sit-in
IQ	حبي ماكو فرق ليش ماتعوفون هل اشياء حته شويه يصير براسنه خير. Hby mAkw frq ly$ mAtEwfwn hl A$yA' Hth $wyh ySyr brAsnh xyr My dear, there's no difference, how don't you know these things at least till it gets better.
MG	لاباس بيها بصح كون جات طويله. lAbAs byhA bSH kwn jAt Twylh. It's fine, but I hope it was long.
LV	بتذكر لما كنت تقلي الدنيا كلا حلوه. bt*kr lmA knt tqly AldnyA klA Hlwh. I remember when you used to tell me that life is all good.

Table 3: Examples of YouDACC sentences (along with translation)

	GU	EG	IQ	MG
Word Unigram	96.00	95.60	92.20	87.20
Word Bigram	**97.00**	**96.40**	**93.30**	83.60
Letters 1-graph	65.70	67.10	68.50	63.70
Letters 2-graph	79.50	81.60	81.10	75.40
Letters 3-graph	89.00	88.10	88.30	83.70
Letters 4-graph	**93.10**	91.80	91.60	**85.20**
Word Length	67.70	70.50	68.40	55.40
Initial Letter 1-gram	62.50	63.20	70.20	58.70
Initial Letters 2-gram	75.20	75.50	78.90	73.30
Initial Letters 3-gram	87.50	86.30	87.00	82.00
Initial Letters 4-gram	92.30	90.80	90.90	84.80
Final Letter 1-gram	57.30	61.50	60.60	67.10
Final Letters 2-gram	75.80	79.20	80.50	74.10
Final Letters 3-gram	88.00	88.90	87.90	84.00
Final Letters 4-graam	94.20	93.30	90.50	85.10

Table 4: Dialect recall of different dialects for the different models.

6. Conclusion

We presented an automatically annotated large-scale multi-dialectal Arabic corpus collected from user comments on Youtube videos. We exploit different video and comment features to automatically annotate each comment with its dialectal class. Our corpus covers the different groups of dialects defined by Habash (2010): Egyptian (EG), Gulf (GU), Iraqi (IQ), Maghrebi (MG) and Levantine (LV). We perform an empirical analysis on the crawled corpus and demonstrate that our location-based proposed method is effective for the task of dialect labeling. This corpus represents a valuable and rich resource for NLP applications treating Arabic dialects.

Acknowledgements

This publication was made possible by grants NPRP-09-1140-1-177 and YSREP-1-018-1-004 from the Qatar National Research Fund (a member of the Qatar Foundation). The statements made herein are solely the responsibility of the authors.

7. References

Cheng, Xu, Dale, Cameron, and Liu, Jiangchuan. (2007). Understanding The Characteristics Of Internet Short Video Sharing: YouTube As A Case Study.

Diab, Mona, Habash, Nizar, Rambow, Owen, Altantawy, Mohamed, and Benajiba, Yassine. (2010). COLABA: Arabic Dialect Annotation And Processing. In *LREC Workshop on Semitic Language Processing*, pages 66–74.

Habash, Nizar, Roth, Ryan, Rambow, Owen, Eskander, Ramy, and Tomeh, Nadi. (2013). Morphological Analysis and Disambiguation for Dialectal Arabic. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 426–432, Atlanta, Georgia, June.

Habash, Nizar Y. (2010). *Introduction to Arabic Natural Language Processing*, volume 3. Morgan & Claypool Publishers.

Mukherjee, Subhabrata and Bhattacharyya, Pushpak. (2012). YouCat: Weakly Supervised Youtube Video Categorization System from Meta Data & User Comments using WordNet & Wikipedia. In *Proceedings of COLING 2012*, pages 1865–1882, Mumbai, India, December. The COLING 2012 Organizing Committee.

Sajjad, Hassan, Darwish, Kareem, and Belinkov, Yonatan. (2013). Translating Dialectal Arabic to English. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–6, Sofia, Bulgaria.

Salloum, Wael and Habash, Nizar. (2013). Dialectal Arabic to English Machine Translation: Pivoting through Modern Standard Arabic. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 348–358, Atlanta, Georgia.

Siersdorfer, Stefan, Chelaru, Sergiu, Nejdl, Wolfgang, and San Pedro, Jose. (2010). How Useful are Your Comments?: Analyzing and Predicting Youtube Comments and Comment Ratings. In *Proceedings of WWW*, pages 891–900, New York, NY, USA.

Wang, Zheshen, Zhao, Ming, Song, Yang, Kumar, Sanjiv, and Li, Baoxin. (2010). YoutubeCat: Learning To Categorize Wild Web Videos. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 879–886. IEEE.

Zaidan, Omar F. and Callison-Burch, Chris. (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*, pages 37–41, Portland, Oregon, USA, June. Association for Computational Linguistics.

Zaidan, Omar F. and Callison-Burch, Chris. (2011b). Crowdsourcing Translation: Professional Quality from Non-Professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1220–1229, Portland, Oregon, USA, June.

Zbib, Rabih, Malchiodi, Erika, Devlin, Jacob, Stallard, David, Matsoukas, Spyros, Schwartz, Richard, Makhoul, John, Zaidan, Omar, and Callison-Burch, Chris. (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, Montréal, Canada, June.

Zhang, John R, Song, Yang, and Leung, Thomas. (2011). Improving Video Classification Via Youtube Video Co-Watch Data. In *Proceedings of the 2011 ACM workshop on Social and behavioural networked media access*, pages 21–26. ACM.