Identification of Answer-Seeking Questions in Arabic Microblogs

Maram Hasanain[1]
maram.hasanain@qu.edu.qa

Tamer Elsayed[1]
telsayed@qu.edu.qa

Walid Magdy[2]
wmagdy@qf.org.qa

[1]Department of Computer Science and Engineering, Qatar University
[2]Qatar Computing Research Institute, Qatar Foundation
Doha, Qatar

ABSTRACT

Over the past years, Twitter has earned a growing reputation as a hub for communication, and events advertisement and tracking. However, several recent research studies have shown that Twitter users (and microblogging platforms' users in general) are increasingly posting microblogs containing questions seeking answers from their readers. To help those users answer or route their questions, the problem of question identification in tweets has been studied over English tweets; up to our knowledge, no study has attempted it over Arabic (not to mention dialectal Arabic) tweets.

In this paper, we tackle the problem of identifying answer-seeking questions in different dialects over a large collection of Arabic tweets. Our approach is 2-stage. We first used a rule-based filter to extract tweets with interrogative questions. We then leverage a binary classifier (trained using a carefully-developed set of features) to detect tweets with answer-seeking questions. In evaluating the classifier, we used a set of randomly-sampled dialectal Arabic tweets that were labeled using crowdsourcing. Our approach achieved a relatively-good performance as a first study of that problem on the Arabic domain, exhibiting 64% recall with 80% precision in identifying tweets with answer-seeking questions.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Text Mining

Keywords

Question Identification; Arabic; Twitter; Crowdsourcing

1. INTRODUCTION

With the increasing popularity and the wide spread of microblogging platforms such as Twitter, more patterns of usage tend to emerge. Among those patterns is posing questions, where users post questions to their followers or even to other users who might have common interests [11, 14, 12]. In an earlier study, Efron and Winget [8] reported that about 13% of a random sample of 2-million tweets were questions. This constitutes a large portion of the tweets and thus indicates a strong need for studying such behavior. Other studies suggested that about 50% of those questions seek answers [16]. Identifying this type of questions would help at several fronts such as understanding the information needs of such questions as well as building systems that either automatically answer them by finding existing answers or even route them to users who might be able to answer.

While the problem of automatic identification of questions in Twitter is not novel [10, 16], the focus of earlier studies was only on English tweets. In this paper, we present a first study that tackles the problem in the domain of dialectal Arabic tweets. Besides having different linguistic structure than English, the Arabic language imposes more challenges as the tweets are posted in several dialects [6].

We define the problem as follows. We first aim to automatically identify tweets that contain questions, i.e., *interrogative* tweets, denoted by *i*tweets. There are many different types of *i*tweets, such as tweets with rhetorical questions, quoted questions, or questions that are followed by answers in the same tweet [8]. Among those types of *i*tweets, we are interested in identifying those tweets with questions that are seeking answers, denoted by *qweets* [10]. Qweets are tweets whose authors expect answers from other Twitter users, or more formally, tweets that convey real information needs. In this study, our research question is simply: can we automatically identify qweets from Arabic tweets?

To tackle the problem, we formulated it as a two-stage classification problem. We first identify Arabic *i*tweets using a rule-based classifier enriched with a large collection of question words and phrases in different Arabic dialects. We then identify qweets from *i*tweets using a binary classifier that leverages a large set of features including lexical, structural, question-specific, tweet-specific, and (in)formality aspects of the tweets. We trained our classifier using manually-annotated tweets collected through crowdsourcing.

We summarize our contributions in this work as follows:

- A ***first*** study on question identification in *Arabic* microblogs is presented. A large dataset of about **865 millions** Arabic tweets spanning **9 months** was used in the study.
- A comprehensive list of question phrases in different Arabic dialects, with mapping to corresponding Mod-

CIKM'14, November 3–7, 2014, Shanghai, China.

ACM 978-1-4503-2598-1/14/11 ...$15.00.
http://dx.doi.org/10.1145/2661829.2661959.

ern Standard Arabic (MSA) question phrases, was constructed. List is made available online[1].
- Two labeled sets of Arabic tweets were developed and made available online[1]: one includes 5000 tweets labeled for *i*tweet identification, and the other contains 3954 tweets labeled for qweet identification.
- Three new categories of features for question identification in Twitter were proposed and evaluated.

The remainder of the paper is organized as follows. We first introduce related work in Section 2. A detailed description of our approach is presented in Section 3. Experimental setup and results are discussed in Section 4, followed by the conclusion and some guidelines for future work in Section 5.

2. RELATED WORK

Question identification in text has been explored in different domains including community question answering platforms [15] and online forums [3]. In Twitter, understanding question-asking behavior of Twitter users has grabbed much attention in the past few years [11, 13, 14, 16]. Some studies on question-asking in Twitter focused on analyzing types and topics of questions asked by users [11, 14]. Others focused on establishing a taxonomy of questions in tweets [8].

Identifying tweets with questions is another problem investigated in literature. One of the approaches used to detect questions (not necessarily answer-seeking ones) is based on applying a set of rules to tweets [8, 14, 2]. This approach showed good recall, yet it introduced many false positives (i.e. tweets that did not have questions) [2]. Dent and Paul [7] applied natural language processing techniques adapted to handle challenges in language used in Twitter to identify questions in tweets. This approach managed to successfully identify tweets matching the syntactic form of a question, but it introduced noise since many filtered tweets did not have answer-seeking questions [7].

Other recent studies focused on using automatic classification to identify qweets specifically [10, 16]. Both of these studies started with a rule-based approach to filter candidate *i*tweets. A set of features was used in a learning approach for qweet identification. Li et al. [10] have utilized question-specific, context-specific and metadata features in classification achieving 77.5% accuracy. Zhao and Mei [16] focused more on lexical features including unigrams, bigrams and trigrams in tweets. They have also attempted to add more semantics to tweets by using WordNet synonyms and part of speech tagging (POS). Their approach achieved a classification accuracy of 86.6%.

Almost all of the previously mentioned studies have focused on English tweets. Up to our knowledge, no studies on *i*tweets/qweet identification in Arabic tweets exist.

3. QWEET IDENTIFICATION

Tweets are very short in length (maximum of 140 characters), usually informal, and naturally conversational. This implies that automatically-detecting qweets is not a trivial task due to the lack of context in tweets. The problem is indeed more challenging with dialectal Arabic. In our study, we focused on dialects of Arab countries with the highest tweeting rate over the past two years, according to a recently-conducted study [1]. We "grouped" those dialects into three groups: Levantine, Egyptian, and Gulf, which was similarly adopted by Cotterell and Callison-Burch [5]. The Gulf group also covered the dialect of Iraq as it shares multiple question phrases with Gulf dialects. We also added MSA to the groups we cover.

In this section, we discuss our 2-stage approach of qweet identification. We first describe *i*tweet identification as a pre-filtering step which provides a list of potential *i*tweets. That list is then classified by a binary classifier to detect qweets. The process of manual annotatation of tweets needed for training the classifier is outlined next. Finally, we present the features developed for qweet classification[2].

3.1 Pre-Filtering

One of the approaches that showed reasonable effectiveness in detecting interrogative tweets uses a set of rules designed to capture questions in tweets [8, 10]. We follow a similar approach to pre-filter tweets in order to get candidate *i*tweets. A tweet is considered an *i*tweet if it contains a question mark (considering both ? or ؟) or a *question phrase*. A *question phrase* in Arabic (such as: إلى أين، ماذا) is a consecutive sequence of (one or more) words that is anlaogous to one of the 5W1H question keywords in English.

Since we are handling dialectal tweets, we could not find a comprehensive list of *dialectal question phrases* covering all dialect groups of interest in this work. Moreover, we wanted to obtain a rich set of dialectal question phrases to maximize the recall of detecting *i*tweets. To overcome this problem, we developed such list using an online survey. We asked participants speaking Arabic in different dialects to provide a list of dialectal question phrases they use in their native dialect. The survey was answered by 105 participants resulting in a list of 348 unique phrases covering 6 dialect groups: Levantine, Gulf, Iraqi, Egyptian, Sudanese, and Maghrebi.

As pointed out earlier, we focus on 3 dialect groups: Egyptian, Levantine, and Gulf. We excluded phrases in other dialects from our initial list to get 264 unique phrases. We further extended this list by (a) augmenting it with question phrases manually-collected by searching online forum posts and Wikipedia pages listing dialectal question phrases, and (b) MSA equivalents of the dialectal phrases, where a dialectal phrase was manually-translated to one or more MSA phrases. Eventually, the list used in pre-filtering had 488 unique phrases, including both MSA and dialectal phrases. We consider any tweet with either a question mark or any of the collected question phrases as an *i*tweet.

3.2 Human Annotations

The pre-filtering step produces a list of identified *i*tweets that are next classified into qweets and non-qweets. To build such classifier, we need a set of manually-labeled *i*tweets for training. Since the *i*tweet identification was automatic, we also need to judge the accuracy of the pre-filtering step by manually-labeling them as true *i*tweets or not.

To do both labeling tasks, we recruited annotators from CrowdFlower[3]. In the first task, workers were asked to label whether an Arabic tweet contains *at least one question* (i.e., is the tweet an *i*tweet or not?). All tweets labeled as tweets containing a question were passed to the second la-

[1]http://faculty.qu.edu.qa/telsayed/datasets.aspx

[2]We thank Linah Lotfi and Nada Aboueata for their valuable help in earlier versions of the question phrases and feature set.

[3]http://www.crowdflower.com/

beling task, in which workers were asked whether a tweet contains an *answer-seeking question* (i.e., is the tweet is a qweet or not?). In each task, workers were provided with labeling guidelines in addition to example tweets on the labeling classes.

To ensure that our annotators understood the labeling tasks and the language of tweets (especially that a tweet can be in dialectal Arabic), only annotators residing in Arab countries were allowed to label tweets. Moreover, annotators were required to pass a qualification test on *golden tweets* (tweets we manually labeled based on our labeling guidelines) to be allowed to annotate tweets. Within each of the tasks, *golden tweets* were also employed to ensure the quality of labeling during the task. If an annotator failed to maintain a labeling accuracy above 70% over the golden tweets, her judgments were not considered. In both tasks, each tweet was labeled by 3 annotators. A final label was chosen for a tweet based on a labeling *confidence* level computed by the CrowdFlower platform. The confidence score is a measure of the annotators agreement weighted by their accuracy over the golden set.

3.3 Feature Extraction

To develop features, we reviewed the literature and analyzed a sample of Arabic tweets we manually labeled as *i*tweets or not and qweets or not. We identified a set of 29 features that can help filtering qweets, in addition to two sets of standard word features. We grouped the features into 6 groups: tweet-specific, structural, formality, question-specific, lexical, and question phrases.

Tweet-specific: In addition to text, tweets have other content elements that have been widely used in different learning problems on tweets. We used 6 features specific to these elements: a feature to indicate whether the tweet has a URL [10, 16], count of *question* hashtags (hashtags Twitter users sometimes use to tag tweets with questions such as: #سؤال equivalent to #question), mentions count, hashtags count, and similarity between the tweet and the title of a webpage posted in tweet [10]. Similarity is calculated using Jaccard coefficient.

Structural: Based on observing tweets, we realized that the structure of the tweet can be a good indicator for qweet classification. For example, we noticed that a tweet having a question and a quoted string is usually not seeking an answer. Features under this category include: length of the tweet in characters and words, and a feature to indicate if the tweet has a quoted string or not.

Formality: Features in this category are used to measure the level of formality of a tweet. Based on our analysis of tweets, we noticed that many tweets with answer-seeking questions are written in a more formal way than expected in tweets. The 5 features under this category are: count of emojis and emoticons (such as: :-)), a feature to indicate if signs of jokes appear in text, e.g., "هههههههههه" (equivalent to "hhhhh" in English), count of diacritics on Arabic letters, and count of special characters such as: {}*(), etc.

Question-specific: We noticed that earlier studies [10, 16] on this problem on English tweets have not developed many features that are specific enough to describe questions in Tweets. Thus, we worked extensively on developing such features given the collected question phrases. Examples on such features include: count of single question marks, count of blobs of question and exclamation marks indicating strong feeling e.g., "!!؟", number of characters and words following the last question mark in tweet, and a feature to indicate if question phrase(s) in a tweet is in MSA. We also developed some features related to the *question sentence* in tweet. We consider a sentence in a tweet as a question sentence if it has a question mark or one of the question phrases in our list. Examples on features related to question sentences include: length of the sentence in characters and words, length of text before and after the question sentence, and count of question sentences found in the tweet.

Lexical: A recent study [16] has showed that using words in tweets as lexical features has good effectiveness in classifying qweets. We used the set of unigrams and bigrams in the tweets as the lexical feature set.

Question Phrases: In addition to the above lexical features, we also used the question phrases extracted from the tweets as a separate feature set.

4. EXPERIMENTAL EVALUATION

4.1 Experimental Setup

In this study, we used a large set of Arabic tweets, collected through Twitter search API over a period of more than 9 months (from May 9^{th}, 2012 to February 21^{st}, 2013), resulting in about 865 million tweets. A tweet in the dataset can be in MSA, dialectal Arabic or a mix of both. We applied pre-filtering to this dataset resulting in about 69 million tweets (8%) marked as candidate *i*tweets. 5% of tweets in the dataset contained question mark(s) (both ? and ؟ are considered), which is about half of the ratio reported in [16] on a billion English tweets collection.

5000 tweets were randomly sampled out of the 69M tweets to be used in our experiments. In the first labeling task, annotators labeled the 5000 tweets as *i*tweets or not; only 3954 tweets of them were labeled as *i*tweets with a confidence level ≥ 0.5, showing a precision of about 79% in detecting *i*tweets. In the second labeling task, 1001 tweets out of the 3954 *i*tweets (25.3%) were labeled as qweets with a confidence level ≥ 0.5.

Our main focus is on extracting qweets from a collection of candidate *i*tweets. We used Support Vector Machines (SVM) [4] classifier, and specifically SVM-light [9], to train and test our binary qweet classifier. The classifier is evaluated using Leave-one-out cross-validation over the labeled tweets produced during the second labeling task in addition to all non-*i*tweets found during the first labeling. We chose to only use the tweets labeled with labeling confidence level ≥ 0.7 out of the original sets, resulting in 3342 tweets.

4.2 Results and Discussion

We followed the classification approach described above using each of the feature groups individually. In evaluation, we mainly focus on the F1-measure. Results in Table 1 show that the structural group had the best performance compared to each of the other individual groups. We then attempted to find the best combination of feature groups (maximizing F1) by gradually adding a group at a time to the structural group. Adding all groups together except for the question phrases group had the best performance. Table 1 summarizes the performance of different combinations of feature groups in addition to the full set. Note that with some feature groups, all tweets were classified as non-qweets

Groups	Precision	Recall	F1
Tweet-specific(TS)	–	0	–
Structural(S)	0.7857	0.4310	**0.5566**
Formality(F)	–	0	–
Question-specific(QS)	–	0	–
Question phrases(QP)	0.4565	0.0392	0.0722
Lexical(L)	0.6944	0.0466	0.08734
S+TS	0.8114	0.5299	0.6411
S+TS+L	0.796	0.5896	0.6774
S+TS+L+QS	0.8119	0.6362	0.7134
S+TS+L+QS+F	0.8061	0.6437	**0.7158**
All	0.7968	0.6437	0.7121

Table 1: Results of classification using each of the feature groups in addition to the best performing combinations.

(getting a zero recall) and thus we marked the precision and F1 in these cases by dashes.

The structural features (the best performing group) mainly focused on length of tweet (including URLs, mentions and hashtags) and length of text in tweet, in addition to detecting existence of quoted strings. Further analysis is needed on a feature-level to determine which individual feature is the best contributor to these results. The performance improvement resulting by adding simple tweet-specific features to structural features increased F1 by 15%. We believe that tweet-specific features added more context to the tweet allowing for more distinctive representation which improved classification.

Enhancement resulting from adding the lexical features was 5.7% which might indicate that they are not as strong as expected in characterizing qweets. It is interesting to observe that adding lexical features resulted in a slight drop in precision (implying that it introduced noise), yet it enhanced recall by 11%. This enhancement in recall is probably due to the fact that lexical features were able to cover common question phrases used in asking questions.

Adding the Question-specific features enhanced performance by 5.3% over the combination S+TS+L indicating that this group might have captured aspects of qweets that were not fully covered yet. We emphasize here the fact that many features within this group were related to the question structure relevant to the tweet, indicating the importance of the structural aspects. Formality features had minimal improvement on F1 when furtherly added. Adding question phrases features did not enhance performance; a possible explanation is that many of them were already covered by the lexical features and thus were redundant.

5. CONCLUSION AND FUTURE WORK

In this work, we presented a first study on the problem of identifying answer-seeking questions in Arabic tweets. The reported preliminary results were encouraging as our approach achieved about 80% precision with 64% recall, which constitutes a strong reference point for future work.

Further result analysis is required especially on a feature-level. Since this is a work in progress, we will be experimenting using feature selection methods to reduce the feature space. Furthermore, as the results reported here are based on using one classifier (SVM), we will be exploring other types of classifiers as well. Moreover, more analysis of the identified qweets is needed to better-understand the information needs of Arabic users of Twitter.

6. ACKNOWLEDGMENTS

This work was made possible by NPRP grant# NPRP 6-1377-1-257 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

7. REFERENCES

[1] Arab Social Media Report. Technical report, Dubai School of Government, June 2013.

[2] A crowd-powered socially embedded search engine. In *ICWSM*, 2013.

[3] G. Cong, L. Wang, C.-Y. Lin, Y.-I. Song, and Y. Sun. Finding question-answer pairs from online forums. SIGIR '08, 2008.

[4] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[5] R. Cotterell and C. Callison-Burch. A multi-dialect, multi-genre corpus of informal written arabic. In *LREC '14*, 2014.

[6] K. Darwish and W. Magdy. Arabic information retrieval. *Foundations and Trends in Information Retrieval*, 7(4):239–342, 2014.

[7] K. D. Dent and S. A. Paul. Through the twitter glass: Detecting questions in micro-text. In *Analyzing Microtext*, 2011.

[8] M. Efron and M. Winget. Questions are content: A taxonomy of questions in a microblogging environment. ASIS&T '10, 2010.

[9] T. Joachims. Making large-scale svm learning practical. advances in kernel methods-support vector learning. schölkopf b. and burges c. and smola a, 1999.

[10] B. Li, X. Si, M. R. Lyu, I. King, and E. Y. Chang. Question identification on twitter. CIKM '11, 2011.

[11] M. R. Morris, J. Teevan, and K. Panovich. What do people ask their social networks, and why?: A survey study of status message q&a behavior. CHI '10, 2010.

[12] A. Oeldorf-Hirsch, B. Hecht, M. R. Morris, J. Teevan, and D. Gergle. To search or to ask: The routing of information needs between traditional search engines and social networks. CSCW'14, 2014.

[13] S. A. Paul, L. Hong, and E. Chi. What is a question? crowdsourcing tweet categorization. CHI'11, 2011.

[14] L. H. S.A. Paul and E. Chi. Is twitter a good place for asking questions? a characterization study. ICWSM'11, 2011.

[15] K. Wang and T.-S. Chua. Exploiting salient patterns for question detection and question retrieval in community-based question answering. COLING '10, 2010.

[16] Z. Zhao and Q. Mei. Questions about questions: An empirical analysis of information needs on twitter. WWW '13, 2013.