EveTAR: A New Test Collection for Event Detection in Arabic Tweets

Hind Almerekhi*, Maram Hasanain, Tamer Elsayed
Computer Science and Engineering Department, Qatar University, Doha, Qatar
halmerekhi@qf.org.qa, {maram.hasanain, telsayed}@qu.edu.qa

ABSTRACT

Research on event detection in Twitter is often obstructed by the lack of publicly-available evaluation mechanisms such as test collections; this problem is more severe when considering the scarcity of them in languages other than English. In this paper, we present *EveTAR*, the *first* publicly-available test collection for event detection in *Arabic* tweets. The collection includes a crawl of 590M Arabic tweets posted in a month period and covers 66 *significant* events (in 8 different categories) for which more than 134k relevance judgments were gathered using crowdsourcing with high average inter-annotator agreement (Kappa value of 0.6).

We demonstrate the usability of the collection by evaluating 3 state-of-the-art event detection algorithms. The collection is also designed to support other retrieval tasks, as we show in our experiments with ad-hoc search systems.

Keywords

Evaluation; Crowdsourcing; Twitter; Ad-hoc Search

1. INTRODUCTION

The overwhelming popularity of Twitter led several parts of the world (e.g., the Arab region) to use it as a medium for continuous exchange of short messages (i.e., tweets). While it became essential for everyday Arab users, popular Arabic news agencies (e.g., AlJazeera and AlArabiya) are also using it extensively to continuously update their followers on critical events and news as they happen. The Arab social media report[1] shows that, as of March 2014, an average of 17M Arabic tweets are posted every day. Such tweets are extremely noisy, full of typos and redundancy, making it difficult to manually identify events [8]. With the growing events in the Arab region, the need for automatic tools that can reliably track the huge stream of tweets and detect events before being publicly announced is increasing.

To design and develop high quality event detection systems, evaluation mechanisms, such as *test collections*, are evidently required. A test collection typically consists of a set of documents (tweets), topics (events), and relevance judgments (labels) that specify which documents are relevant to the topics [14]. The problem of detecting events in Twitter has been extensively studied [2, 13, 15], however most of event detection systems were tested on English tweets. To the best of our knowledge, Alsaedi and Burnap [1] presented the only study that tackles the problem of event detection in Arabic tweets, however, no test collection was publicly-available for further research.

This paper presents *EveTAR*, the *first* publicly-available test collection for event detection in *Arabic* tweets. More specifically, the collection was designed for the problem of detecting *significant* events. A significant event is defined as an occurrence that happens at a particular time in a specific location and is discussed by the media (e.g., covered by an online news article). This definition is similar to definitions introduced in [8]. However, we emphasize event significance, which is often neglected in most event definitions [1, 2, 12]. *EveTAR* covers 66 manually-identified significant events occurred in the month of January 2015. More than 134k potentially-relevant tweets to the events were judged through crowdsourcing. The design of *EveTAR* allows for the evaluation of systems that are built for other tasks, such as ad-hoc search and filtering. It can also be extended to support other tasks such as summarization.

In building *EveTAR*, we address three research questions:

- **RQ1:** How can we design a test collection that is reusable and supports multiple tasks?
- **RQ2:** How can we use crowdsourcing in building a reliable test collection?
- **RQ3:** How well do existing state-of-the-art event detection techniques perform on Arabic tweets?

Our contribution in this study is 3-fold:

1. *EveTAR* is the *first* test collection for the task of event detection in Arabic tweets.
2. We make the full test collection publicly-available for research[2], including ids of 590M Arabic tweets, 66 events and 134K relevance judgments, inter-annotator agreements, queries used to identify potentially-relevant tweets for the events, and documented design of the crowdsourcing tasks.

*Also affiliated with Qatar Foundation, Doha, Qatar.
[1]http://www.arabsocialmediareport.com

[2]http://qufaculty.qu.edu.qa/telsayed/evetar

SIGIR '16, July 17–21, 2016, Pisa, Italy.

DOI: http://dx.doi.org/10.1145/2911451.2914681

3. We demonstrate that our test collection can be used to evaluate existing state-of-the-art event detection systems in addition to ad-hoc search systems.

2. RELATED WORK

In this section, we discuss some of the existing test collections for event detection and compare them with *EveTAR*.

The study of Alsaedi and Burnap [1] is the first on event detection in Arabic tweets, focusing on detecting events in Abu Dhabi. Around 1M Arabic tweets were collected and labelled by 3 annotators, however, the dataset was not made publicly-available. Additionally, it was restricted to events in Abu Dhabi, which introduces a bias towards types of events that happen in that location.

Petrović et al. [13] built a test collection consisting of 50M English tweets for the task of First Story Detection. Given a set of manually-crafted events, authors recruited expert annotators to collect on-topic tweets for those events. The need for expert annotators makes the creation of the test collection expensive and limits its scale. Furthermore, the authors identified 27 events with 3K relevant tweets only, making it difficult to use the collection for conducting large-scale evaluation of event detection over the Twitter stream.

On a larger scale, McMinn et al. [8] built a publicly-available test collection for evaluating event detection. The authors crawled around 120M English tweets, covering more than 500 events identified using automatic and manual ways, and collected labels for over 150K tweets. We consider their approach as a basis for our work, however, we followed a slightly different approach to construct *EveTAR* with manual identification of events. On average, *EveTAR* has more tweets per event when compared to their collection. Additionally, we designed our test collection to be general enough to support additional tasks like ad-hoc search.

Contrasting *EveTAR* to other event detection test collections, we find it the largest (with 590M tweets) compared to collections described in [3], [6], [9] and [10]. Moreover, events in *EveTAR* were not limited to a specific location as opposed to test collections in [2] or [3] for example.

3. BUILDING EveTAR

We built *EveTAR* based on the following pipeline: collecting the tweet dataset, identifying events in this dataset, extracting potentially-relevant tweets for those events, and finally obtaining their relevance judgments.

3.1 Collecting a Tweet Dataset

We used Twitter's streaming API to collect a dataset of 590M Arabic tweets over the month of January 2015. Tweets were collected by tracking 400 most frequently-used Arabic words extracted from a previously-crawled Twitter stream.

3.2 Identifying Events

Following the approach of McMinn et al. [8], we manually collected a set of 357 events in the month of January 2015 listed over both the English[3] and Arabic[4] Wikipedia's Current Events Portal (WCEP). We then applied our significance criteria over two phases. In the first, we only kept events for which we found at least one online *Arabic* news article discussing the event; only 71 events satisfied that condition. For each of those events, we used Twitter's online search service to manually try to find at least 20 non-redundant tweets related to the event. We restricted the search to 5 days starting 2 days prior to the event date. Eventually, 66 events satisfied the condition, comprising our final list of events. For each of those, we prepared a complete representation as shown in the translated example below:

ID	E12
Title	Discovery of tomb of Egyptian queen Khentakawess III
Date	January 04, 2015
Location	Abusir, Egypt
Category	Arts and Culture
Reference	http://cnn.it/1O6grQK
Keywords	Khentakawess, Egyptian queen, archaeologist
Description	A Czech archeological team discovered the tomb of an Egyptian queen named Khentakawess III who lived during the 5^{th} dynasty.

The final events fall into 8 WCEP categories; Table 1 shows their distribution. Interestingly, the majority belong to Armed Conflicts and Attacks.

Table 1: Distribution of events

Category (Events)	Category (Events)
Armed Conflicts & Attacks (45)	Sports (5)
Business & Economy (1)	Arts & Culture (2)
International Relations (3)	Law & Crime (2)
Disasters & Accidents (3)	Politics & Elections (5)

The above event representation is sufficient to evaluate event detection systems that represent events by any combination of date/time, location, and set of keywords. However, in designing *EveTAR*, we elected to enrich the above representation by adding a list of tweets related (or relevant) to each event. That serves two purposes; first, it helps evaluate several event detection systems that represent an event by a list of tweets, and second, it enables the evaluation of other types of retrieval systems such as ad-hoc search or filtering systems that rely on producing lists of tweets per topic.

We obtained those tweets over 2 steps. We first extracted a list of potentially-relevant tweets for each event from our dataset, then used crowdsourcing to obtain relevance judgments on them; both are described in the following sections.

3.3 Extracting Tweets of Events

For each event, we manually crafted a list of keyword and phrase queries, ran them against a Lucene-constructed index of the dataset, and collected the resulted tweets (up to 10k per event). After removing exact tweet-text duplicates, we finally got a set of 134,069 potentially-relevant tweets.

3.4 Collecting Relevance Judgments

For efficiently labeling that large set of potential event-related tweets, we used CrowdFlower crowdsourcing platform[5]. We ran a set of pilot studies before launching the final labeling tasks, one per event. In each task, annotators were given an introduction to the task, the title, description and date of the event, and the content of a corresponding Arabic news article. Annotators were then asked to label each of the candidate tweets as relevant to the event or not.

[3]https://en.wikipedia.org/wiki/Portal:Current_events

[4]https://ar.wikipedia.org/wiki/بوابة:أحداث جارية

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

Before annotators can start labeling, they were required to pass a qualification test by correctly labeling a minimum of 8 out of 10 gold tweets. Gold tweets were randomly sampled from the collection of potentially-relevant tweets per event. Once they start labeling, annotators had to maintain a minimum accuracy of 80% over gold tweets within the task to continue labeling. An average of 8 annotators were blocked while labeling for an event, and only labels from trusted annotators where included in *EveTAR*. We chose to have each tweet annotated by 3 annotators to ensure a majority label. We also restrict the annotators to be Arabic-speaking with an intermediate level per CrowdFlower's ranking of annotators. An average of 59 hours were spent per event.

To decide the final label of a tweet out of the 3 given labels, we adopted a trust-based voting scheme that utilizes the trust scores provided by CrowdFlower per annotator (describing her accuracy in the current task). We chose the label that has the highest sum of trust scores over corresponding annotators. For all events, 51,424 tweets were labeled as relevant and 82,645 were non-relevant, based on the above voting scheme.

We also computed an overall trust score per tweet using the following Equation[6]:

$$\text{Tweet Trust Score} = \frac{\max(\sum_{i=1}^{r} trust_i, \sum_{i=1}^{n} trust_i)}{\sum_{i=1}^{n+r} trust_i} \quad (1)$$

where r and n are the number of annotators labeling the tweet as relevant or non-relevant respectively, and $trust_i$ is the trust score for an annotator. Averaging this overall trust score over all tweets and all events results in an average quality score of 0.94 out of 1.

We measure the quality of the obtained labels by computing the inter-annotator agreement using Fleiss' Kappa [16] per event. We chose this measure over the widely-used Cohen's Kappa as it allows measuring annotators' agreement when having more than 2 annotators labeling a single data item. In literature, the value of Kappa has been mapped into 6 categories based on how strong the agreement is [16]. Figure 1 illustrates the distribution of events over Fleiss' Kappa categories. Over all events, we got an average Kappa of 0.60, which is considered a moderate agreement. Moreover, more than half of the events got a substantial to an almost perfect agreement. Additionally, eliminating the 3 events with slight annotator agreement results in an average agreement of 0.62, which is considered *substantial*. We observed that events with slight agreement tend to be less popular in the media than those with almost perfect agreement.The figure also shows a slight drop in the values of the average trust score per event with decreasing Kappa values.

4. USING EveTAR

In this section, we demonstrate the usability of *EveTAR* for event detection in addition to other retrieval tasks, taking ad-hoc search as an example.

4.1 Evaluating Event Detection

We experimented with *EveTAR* to show how it can be used to evaluate state-of-the-art event detection algorithms

[6]The equation is stemming from the one used by CrowdFlower to report confidence in the aggregated label given for a data item, see: http://bit.ly/20NmFkU

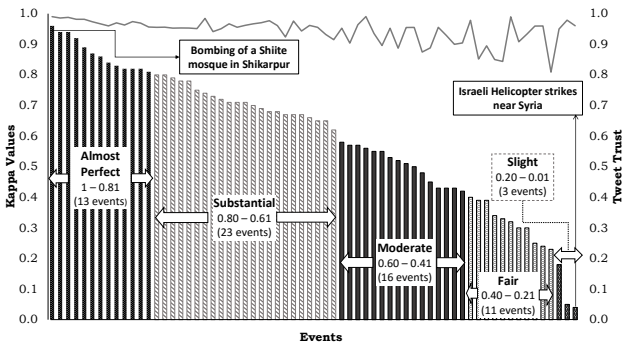

Figure 1: Distribution of events over Fleiss' Kappa categories and average trust scores.

over Arabic tweets. We used an open-source implementation of 3 event detection algorithms provided by SONDY platform[7], namely MABED [4], EDCoW [17], and Peaky Topics [15]. We ran the algorithms using SONDY's default parameter settings.

In a real setting, we should run those algorithms on the full dataset (590M tweets), but due to time limitation and to be able to compute both recall and precision of event detection, we only ran the algorithms on the labeled subset (134K tweets). We evaluated the output of each algorithm by *manually* comparing the output events with the labeled events (based on their descriptions). Moreover, we also *automatically* evaluated the output using the approach of Petrović [11]. Performance was evaluated using the standard precision (P), recall (R) and F_1 measures.

To compare event detection over *EveTAR* with another available test collection, we used McMinn et al. English test collection [8]. We crawled a subset of the collection, that was created similar to *EveTAR*, of 23K judged tweets covering 361 events. We ran and evaluated the same algorithms using the setup explained above. Results over both collections are summarized in Table 2. The label (A) next to some algorithms indicates the automatic evaluation.

Table 2: Event detection performance with *EveTAR* and an English collection

	EveTAR			***McMinn et al.***		
Algorithm	P	R	F_1	P	R	F_1
EDCoW	0.09	0.15	0.11	0.17	0.05	0.07
Peaky Topics	0.11	0.80	0.19	0.17	**0.58**	0.26
Peaky Topics(A)	0.10	0.71	0.16	0.08	0.29	0.13
MABED	0.42	0.64	0.51	0.89	0.24	0.38
MABED(A)	**0.61**	**0.92**	**0.73**	**1.0**	0.28	**0.43**

Over both collections and using both manual and automatic evaluation, MABED consistently exhibited the best performance. Moreover, automatic evaluation of MABED showed that the results are somewhat close to those using the manual method. The same observation stands with Peaky Topics when using *EveTAR*. We believe using automatic evaluation more accurately reflects detection performance as it compares identified events with the ground truth on a tweet-level, not to mention its efficiency. These

[7]mediamining.univ-lyon2.fr/people/guille/sondy.php

results demonstrate that *EveTAR* can be used to automatically evaluate event detection systems.

While F_1 numbers over *EveTAR* (especially for MABED) are relatively high, this is probably an artifact of using only the judged subset of tweets instead of the full dataset for testing. Another factor we noticed is that the default value of a MABED parameter that indicates the number of events to be detected was 100, which is coincidentally close enough to the number of events in *EveTAR*.

4.2 Evaluating Ad-hoc Search

In addition to event detection, *EveTAR* is designed to support evaluation of other tasks like ad-hoc search, as it provides a short query and a list of relevant tweets per topic (i.e., event). Using *EveTAR*, we experimented with two ad-hoc search systems, denoted by query likelihood (QL) and query expansion (QE), which were adopted by one of the top teams in the ad-hoc search task in TREC-2013 microblog track [5, 7]. We ran the search systems (using their reported parameter values) over the Lucene index of the 590M tweets. Both MAP and P@30 measures were used to evaluate the performance of the systems as summarized in Table 3.

Table 3: Ad-hoc search performance with *EveTAR*

Model	*MAP*	*P*@30
QL	0.1283	0.3783
QE	0.1207	0.3384

We notice that P@30 values for both models are in range of those reported in [7], however, MAP values are much lower than expected. The difference in ad-hoc search performance between *EveTAR* and the English collection in [7] might be due to the very big difference in dataset size, as the English dataset is much smaller than *EveTAR*. Surprisingly, the QE model is performing worse than QL; we believe that the lack of proper Arabic text normalization on tweets used to select expansion terms might cause a poor selection of terms for expansion. Further experiments are needed to understand the performance of ad-hoc search systems over *EveTAR*.

5. CONCLUSION & FUTURE WORK

In this paper, we present our work on constructing *EveTAR*, the first publicly-available Arabic test collection for event detection. The collection includes 66 events manually-identified during the month of January 2015. We used crowdsourcing to collect a total of 134K labeled tweets. We demonstrate the usage of our collection in manually and automatically evaluating state-of-the-art event detection algorithms. We also show how *EveTAR* can be used to automatically evaluate other tasks such as ad-hoc search.

There are several possible directions for future work. First, the experimental results obtained in demonstrating the usage of *EveTAR* are just preliminary; therefore we still need to further study the performance of event detection and ad-hoc search systems over Arabic tweets. Second, we plan to run more extensive studies to evaluate systems built for other retrieval tasks (e.g., tweet filtering) over our collection. Finally, we would like to explore the possibility of extending *EveTAR* with more events using the automatically-identified potential events by event detection systems.

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.[8]

6. REFERENCES

[1] N. Alsaedi and P. Burnap. Arabic event detection in social media. In *Computational Linguistics and Intelligent Text Processing*, pages 384–401. 2015.

[2] H. Becker, M. Naaman, and L. Gravano. Beyond trending topics: Real-world event identification on Twitter. In *ICWSM'11*, pages 438–441, 2011.

[3] T. Gottron, O. Radcke, and R. Pickhardt. On the temporal dynamics of influence on the social semantic web. In *Semantic Web and Web Science*, pages 75–87. 2013.

[4] A. Guille and C. Favre. Mention-anomaly-based event detection and tracking in Twitter. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 375–382, 2014.

[5] M. Hasanain, L. Al-Marri, and T. Elsayed. QU at TREC-2013: Expansion Experiments for Microblog Ad hoc Search. In *TREC-2013*, 2014.

[6] F. Kunneman and A. van den Bosch. Event detection in Twitter: A machine-learning approach based on term pivoting. In *BNAIC'14*, 2014.

[7] J. Lin and M. Efron. Overview of the TREC-2013 Microblog Track. In *TREC-2013*, 2013.

[8] A. J. McMinn, Y. Moshfeghi, and J. M. Jose. Building a large-scale corpus for evaluating event detection on Twitter. In *CIKM'13*, pages 409–418, 2013.

[9] T. Mitra and E. Gilbert. Credbank: A large-scale social media corpus with associated credibility annotations. In *ICWSM'15*, 2015.

[10] R. Parikh and K. Karlapalem. ET: Events from Tweets. In *WWW'13 Companion*, pages 613–620, 2013.

[11] S. Petrović. *Real-time event detection in massive streams*. PhD thesis, School of Informatics, University of Edinburgh, 2013.

[12] S. Petrović, M. Osborne, and V. Lavrenko. Streaming first story detection with application to Twitter. In *NAACL HLT'10*, pages 181–189, 2010.

[13] S. Petrović, M. Osborne, and V. Lavrenko. Using paraphrases for improving first story detection in news and Twitter. In *NAACL HLT'12*, pages 338–346, 2012.

[14] M. Sanderson. *Test collection based evaluation of information retrieval systems*. Now Publishers Inc, 2010.

[15] D. A. Shamma, L. Kennedy, and E. F. Churchill. Peaks and persistence: modeling the shape of microblog conversations. In *CSCW'11*, pages 355–358, 2011.

[16] J. Sim and C. C. Wright. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical therapy*, 85(3):257–268, 2005.

[17] J. Weng and B.-S. Lee. Event detection in Twitter. In *ICWSM'11*, pages 401–408, 2011.

[8]The first author was not funded by the grant.