MED: Multi-domain and Multi-modality Event Detection Dataset

Zhenguo Yang, Qing Li, Wenyin Liu, and Jianming Lv

1 Overview

Real-world events can be defined as something happened or happening, which attract people's attention and impact our lives. In reality, whenever an event happens, it can be witnessed by different groups of people (e.g., professionals and amateurs) and get the attention of the public. Therefore, tremendous data related to the same events may be distributed in different data domains (e.g., online news domain and social media domain), due to the convenience of sharing data in the era of Web 2.0. In this work, we construct and release a multi-domain and multi-modality event detection dataset (MED), containing 9,722 textual news articles collected from hundreds of news media sites (e.g., Yahoo News, Google News, CNN News, etc.) and 2,3500 image posts shared on Flickr social media, which are annotated according to 100 real-world events. The dataset is collected to explore the problem of organizing heterogeneous data contributed by professionals and amateurs in different data domains, and the problem of transferring event knowledge obtained from one data domain to a heterogeneous data domain, thus summarizing the data with different contributors. We hope that the release of the MED dataset can stimulate innovate research on this challenging problem.

2 Collection and Annotation

The event labels in MED cover a wide range of event categories (or event types) like emergency, natural disaster, sport, ceremony, election, protest, military intervention, economic crisis, etc. We collect the dataset considering the aspects of high relevance in supporting the application needs, wide range of event types, non-ambiguity of the event labels, imbalance of the event clusters, and difficulty of discriminating the event labels, etc.

To collect textual news articles, there are two challenging problems. The first problem is the dispersibility of the data. For instance, news articles reporting the same events may be reported by any news media sites in the world, thus it is laborious and even impossible to collect the data by accessing each news media sites for annotation. In addition, most online news media sites do not provide effective and convenient interface for retrieval. The second problem is the quality of the data, such as authority, credibility, etc. We have to ensure the news articles are publicly-accepted, and avoid rumors, fake news, or inaccurate news stories as much as possible. Therefore, instead of collecting the data directly from the individual news media sites, we resort to Wikipedia and collect the data in a top-down manner. More specifically, we manually look up the Wikipedia entries about events recently, and check the crowd-sourced articles describing the events that are available online, which have cited quite a few news articles in the references as shown in Fig. 1. These articles are contributed by different news media sites, and usually focus on different aspects or hold different views on the events, but are quite high in terms of quality and authority. Furthermore, we crawl these news articles further by accessing the links of the articles in different news media sites. As a result, the name of the Wikipedia entry about the event (i.e., event label) are well-accepted by the public as they are edited in a crowdsourcing manner, and there is no ambiguity in terms of event labels. In addition, Wikipedia provides quite a few portals to access similar events in terms of event types, regions, etc., which helps to collect similar events in certain aspects to ensure the difficulty of discriminating the event labels.

For the image posts on social media, we take Flickr as an example for collecting data. Given the event names collected from Wikipedia, we retrieve the related data by using different queries (i.e., keywords) and some strategies like filtering by time to obtain a number of returned results and remove the replicated ones. On one hand, we annotate the image posts manually to verify whether the samples are related to the query events, and crawl the responding data samples. On the other hand, we further access the Flickr albums (as shown in Fig. 2) of these image posts. Whenever the topics of the albums are in accord with the events, we will craw them further. The event labels of the Flickr image posts are consistent with the ones obtained from Wikipedia to ensure non-ambiguity and difficulty of differentiating.

3 The MED Dataset

3.1 Statistics of the MED Dataset

The statistics of the data samples in the dataset are summarized in Table 1. Specifically, 9,722 textual news articles are collected from hundreds of data sources in the online news media domain, including Yahoo News, Google News, Huffington Post, CNN News, New York Times, NBC News, Fox News, Washington Post, The Guardian, etc. A number of 23,500 Flickr image posts are included, which are shared by 269 Flickr social media users.

References [edit]

1. ^ http://www.hkilang.org/NEW_WEB/page/dictionary Association for Conversation of Hong Kong Indigenous Languages Online Dictionary for Hong Kong Hakka and Hong Kong Punti (Weitou dialect)
2. ^ Tang, Baiqiao (26 November 2014). "香港雨傘運動及未來中國民主革命之展望" (in Chinese). New Tang Dynasty Television. Retrieved 29 November 2014.
3. ^ Phillips, Keri (28 October 2014). "Tracing the history of Hong Kong's umbrella movement". *ABC Radio National*. Australian Broadcasting Corporation. Retrieved 29 November 2014.
4. ^ "Beijing rejects full Hong Kong democracy". Deutsche Welle. 31 August 2014. Retrieved 31 August 2014.
5. ^ a b c d e "Hong Kong's 'Occupy' leaders now face quiet but persistent harassment". *The Christian Science Monitor*.
6. ^ "Hong Kong: #umbrellarevolution, anatomie d'un hashtag". *Slate*. Retrieved 3 October 2014.
7. ^ "HK police surprise protesters with tear gas". *The New Paper*. Retrieved 3 October 2014.
8. ^ "Hong Kong protests in pictures: The 'Umbrella Revolution' ". *The Independent*. Retrieved 3 October 2014.
9. ^ " 'Umbrella Revolution' Protests Spread in Hong Kong". *The Huffington Post*. 29 September 2014. Retrieved 3 October 2014.
10. ^ Images of Hong Kong' s 'Umbrella Revolution' Tell a Story. *The New York Times*, 29 September 2014.
87. ^ Wendy Tang. "Texting apps required gear for Hong Kong protests". *The Seattle Times*. Retrieved 6 October 2014.
88. ^ Timmons, Heather (19 October 2014). "The US is no role model in Hong Kong's democracy fight". *Quartz*. Retrieved 30 November 2014.
89. ^ a b Schumacher, Mary Louise (6 November 2014). "The enchanting art of Hong Kong's Umbrella Revolution". *TAP*. Journal Sentinel. Retrieved 8 November 2014.
90. ^ Noack, Rick (7 October 2014). "Photos: The colorful world of Hong Kong's protest art 7 October 2014". *The Washington Post*. Retrieved 8 October 2014.
91. ^ "Art bursts from Hong Kong protests". live5news. 8 October 2014. Retrieved 30 November 2014.
92. ^ "The powerful art behind Hong Kong's protests". Channel 4. 6 October 2014. Retrieved 8 October 2014.
93. ^ Blair, David (7 October 2014). "The public artwork of the Hong Kong protests". *The Daily Telegraph*. Retrieved 8 October 2014.
94. ^ Bradsher, Keith (5 October 2014). "Sinosphere – New Image of the Hong Kong Protests: 'Umbrella Man' ". *The New York Times*. Retrieved 30 November 2014.
95. ^ "Interview With Hong Kong's 'Umbrella Man' Statue Artist". Malaysia: Yahoo News. 5 October 2014. Retrieved 8 October 2014.
96. ^ Watson, Ivan; Boykoff, Pamela & Kam, Vivian (8 October 2014). "Street becomes canvas for 'silent protest' in Hong Kong". CNN. Retrieved 25 October 2014.
97. ^ "Pro-Democracy Banner Occupies Hong Kong's Iconic Lion Rock, Spawns Memes". *The Wall Street Journal*. Retrieved 25 October 2014.

Fig. 1. Snapshot of References of "Umbrella Movement" in the Wikipedia Entry.

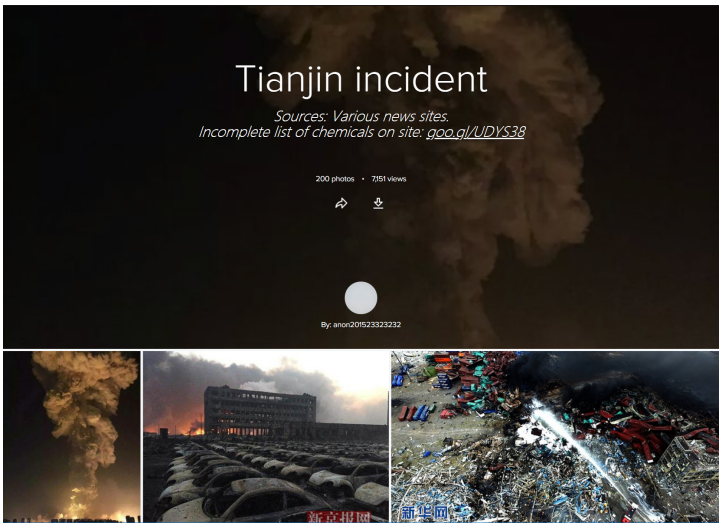

Fig. 2. Snapshot of an Album containing 200 photos about "Tianjin Explosions" on Flickr.

TABLE 1
Statistics of the MED dataset (X: online news media, Y: social media). In particular, new articles do not possess GPS-tags, and we extract them from their content. For the Flickr images, we use the prior knowledge that if two images are shared by the same users in a short time period, the GPS-tags can be shared with the one does not posses GPS-tags.

Domain	#Events	#Items	Time (%)	Location (%)	Title (%)	Data source / Flickr user (%)	Tags (%)	Textual content / Description (%)
X	100	9,722	100	0 → 93.74	99.54	100	n.a.	100
Y	100	23,500	100	13.54 → 49.17	100	100	81.08	99.97

3.2 Dataset Partition

For fair comparisons, we partition the dataset into fixed training set (i.e., seed set) and test set. Specifically, we randomly select 10 samples from the data samples of each event in each domain to generate two seed sets (i.e., training set), consisting of 1,000 news articles and 1,000 image posts. The rest of the data samples constitute two test sets for the two domains, respectively. For transfer learning scenarios such as $X \rightarrow Y$, we can use the seed set in domain X for training, and the test set in domain Y for testing, which is similarly for $Y \rightarrow X$. For unsupervised approaches, the two test sets can be used alone.

3.3 Event Labels of the MED Dataset

The 100 real-world events in the dataset can be roughly summarized by several categories, i.e., public security (e.g., shooting, attack, killing, conflict, explosion, epidemic, crash, etc.), natural disaster (e.g. earthquake, flood, fire, etc.), protest, sport, election, festival, etc. The following Tables summarize the names of event labels and the number of data samples in each domain for the six categories. In particular, there is a Wikipedia entry for the name of each event label, and more details about the events can be obtained by directly accessing these entries in Wikipedia.

TABLE 2
Events of public security (32 events) in the dataset.

Event Label	X	Y
2014 Isla Vista killings	81	143
2015 Copenhagen shootings	131	100
2015 San Bernardino attack	269	272
Central African Republic Civil War (2012-present)	45	397
Charleston church shooting	45	154
Charlie Hebdo shooting	73	80
Death of Eric Garner	218	399
Death of Freddie Gray	210	397
Death of Nelson Mandela	65	76
Death of Sandra Bland	90	60
Kurdish-Turkish conflict (2015-present)	54	42
Mamasapano clash	80	207
Marysville Pilchuck High School shooting	48	73
Military intervention against ISIL	98	207
November 2015 Paris attacks	48	90
Post-coup unrest in Egypt (2013-2014)	94	400
Shooting of Michael Brown	43	400
Taliban insurgency	50	100
Umpqua Community College shooting	47	1145
War in Afghanistan (2015-present)	46	146
Bill Cosby sexual assault allegations	43	170
2013 Savar building collapse	90	96
2014 Kaohsiung gas explosions	158	214
2015 Ankara bombings	55	369
2015 Bangkok bombing	103	68
2015 Tianjin explosions	58	125
Boston Marathon bombing	155	286
2015-16 Zika virus epidemic	146	390
Greek government-debt crisis	213	80
Asiana Airlines Flight 214	148	167
Indonesia AirAsia Flight 8501	71	91
Refugio oil spill	47	299

TABLE 3
Events of protest (13 events) in the dataset.

Event Label	X	Y
2013 Shahbag protests	106	83
2014 Hong Kong protests	46	64
Belfast City Hall flag protests	63	73
Boko Haram insurgency	63	80
Euromaidan	105	100
Extreme Rules (2012)	153	412
Ferguson unrest	69	62
Gezi Park protests	104	398
June 2013 Egyptian protests	111	124
Nuit debout	49	186
Sunflower Student Movement	51	131
Umbrella Movement	60	49
Azadi March	128	106

TABLE 4
Events of festival (9 events) in the dataset.

Event Label	X	Y
2014 Sundance Film Festival	75	52
2015 Tour of California	89	399
85th Academy Awards	342	1154
86th Academy Awards	159	67
Miss USA 2015	214	184
Miss Universe 2013	86	176
Miss Universe 2015	56	102
Prince George of Cambridge	52	400
APEC Philippines 2015	77	81

TABLE 5
Events of natural disaster (17 events) in the dataset.

Event Label	X	Y
2013 Bohol earthquake	64	189
2015 Nepal blockade	51	58
2015 Sabah earthquake	52	399
April 2015 Nepal earthquake	91	850
Cyclone Oswald	75	254
Cyclone Pam	94	261
Cyclone Winston	225	58
December 2013 North American storm complex	157	69
Hurricane Joaquin	46	400
Hurricane Patricia	54	195
Tropical Storm Erika	56	116
Typhoon Haiyan	42	398
2014 Southeast Europe floods	105	123
2014-15 Malaysia floods	61	179
2013 New South Wales bushfires	215	398
2016 Fort McMurray wildfire	71	150
Colectiv nightclub fire	58	92

TABLE 6
Events of election (14 events) in the dataset.

Event Label	X	Y
Australian federal election, 2013	69	117
Canadian federal election, 2015	66	369
Egyptian presidential election, 2014	72	112
European Parliament election, 2014	112	122
Indian general election, 2014	190	125
Israeli legislative election, 2015	382	112
Nigerian general election, 2015	150	745
Romanian presidential election, 2014	61	421
Scottish independence referendum, 2014	68	397
Singaporean general election, 2015	51	105
Tunisian presidential election, 2014	44	58
Turkish general election, June 2015	48	273
United Kingdom general election, 2015	109	46
United States Senate election in Kentucky, 2014	46	399

TABLE 7
Events of sport (15 events) in the dataset.

Event Label	X	Y
2013 All-Ireland Senior Hurling Championship Final	64	200
2013 International V8 Supercars Championship	49	109
2013 Stanley Cup playoffs	62	130
2013 World Series	94	200
2014 24 Hours of Le Mans	164	113
2014 Commonwealth Games	84	1080
2014 FIFA World Cup	67	416
2014 Formula One season	187	792
2014 GP2 Series	106	375
2014 NASCAR Sprint Cup Series	124	122
2014 Winter Olympics	44	71
2015 24 Hours of Le Mans	48	107
2015 MotoGP season	58	97
2015 National Rugby Championship	135	170
2015 Rugby World Cup	71	202