HERMEVENT: A News Collection for Emerging-Event Detection

Cristiano Di Crescenzo[a] Giulia Gavazzi[a] Giacomo Legnaro[a]
Elena Troccoli[a] Ilaria Bordino[b] Francesco Gullo[b]

[a]Sapienza University of Rome, Italy

[b]UniCredit, R&D Dept., Italy

7th ACM International Conference on Web Intelligence, Mining and Semantics - June 19-22, 2017 - Amantea, Italy

Test collections for event detection

- news portals and microblogging platforms
  - for breaking news and unexpected events
- scarcity of publicly-available test collections
- most of the work on event detection exploits Twitter data

Our Contribution

A test collection typically consists of

- A set of documents
- A list of topics or events
- A set of relevance annotations

Our main contributions:

- HERMEVENT: A new test collection for event detection (tweets and news articles, 3 months in 2016 / 2017)
- A set of knowledge graphs with different semantic and temporal granularity
- Evaluation of two state-of-the-art graph-based event-detection methods

The HERMEVENT Collection

Includes news from a list of major italian newspapers

it.euronews.com	www.ilsole24ore.com
it.reuters.com	www.ingv.it
tg24.sky.it	www.interno.gov.it
www.agi.it	www.ladige.it
www.ansa.it	www.lagazzettadelmezzogiorno.it
www.corriere.it	www.lastampa.it
www.esteri.it	www.milanofinanza.it
www.gazzettadiparma.it	www.protezionecivile.gov.it
www.ilfattoquotidiano.it	www.rai.it
www.ilgiornale.it	www.repubblica.it
www.ilmattino.it	www.tgcom24.mediaset.it
www.ilmessaggero.it	www.viaggiaresicuri.it

The HERMEVENT Collection

- Includes news and tweets in Italian
- Useful for language-independent event detection methods, such as graph-based approaches
- Words and entities can be easily translated in other languages by using multi-language resources (e.g., Wikipedia inter-language links).

- Time Horizon: 3 months from December 12th, 2016 to March 7th, 2017
- News are collected by exploiting the news-crawling, RSS-feed-processing, and data-cleaning functionalities embedded in the Hermes [1] tool
- Overall number of news is 88092

Two different semantic granularities: words and entities

1. *Word-based representation*:
   - Word vocabulary $\mathcal{V}_w$: union of all words in the news.
   - Cleaning: stopword-removal, stemming, words with less than 10 occurrences
2. *Entity-based representation*:
   - Entity vocabulary $\mathcal{V}_e$: the entities extracted solving ERD
   - ERD: TagMe algorithm (Ferragina et al., CIKM'10), implemented in Hermes.
   - Discard entities matching stopwords or over-popular (frequency > 3600).

1. Split the period in intervals of 3h, 6h, 12h and 1D
2. Define an undirected temporal graph $\mathcal{G}^{\mathcal{T}} = (V, \{E_t, w_t\}_{t \in \mathcal{T}})$ for each interval $[t_i, t_{i+1})$, semantic and temporal granularity
    - $\mathcal{T}$: time horizon
    - $E_t \subseteq V \times V$: edge set
    - $w_t : E_t \to \mathbb{R}^+$: weights to edges $w_t(u, v) = c_t(u, v) \geq \eta$

Word-Based Graphs

Average statistics of temporal graphs for the word granularity

	3h	*6h*	*12h*	*1d*
#non-singleton vertices	2 007	3 203	5 205	7 820
#edges	189 108	404 081	823 336	1 595 255
min degree	1.83	1.25	1.01	1
avg degree	157.59	216.57	304.21	398.42
median degree	89.48	106.75	126.02	144.63
max degree	1 617.61	2 602.8	4 256.53	6 428.55

Entity-Based Graphs

Average statistics of temporal graphs for the entity granularity

	3h	*6h*	*12h*	*1d*
#non-singleton vertices	231	471	935	1 822
#edges	1 688	3 653	7 697	16 570
min degree	1.51	1.15	1	1
avg degree	11.7	12.59	13.78	15.57
median degree	10.66	10.52	10.66	11.27
max degree	40.61	65.05	108.56	193.24

Comparison of the two state-of-the-art graph-based event-detection methods:

- *BUZZ [3]*: extracts events with a two-step methodology:
  1. Quantify how abnormal the association between two terms is at any time with respect to its history
  2. Identify cohesive subsets of terms
- *Raw-Graph Event Detection (RG-ED)*: running the BUZZ method on the original graph:
  - Edges are weighted with raw term co-occurrence counts
  - Target time window the (unique) time instant

BUZZ Algorithm: Anomaly Score

- Calculate how anomaly is every data point in a temporal sequence
- Anomaly score is the *e*'s percentile weight at time t_i
- Comparison to the median of the corresponding percentiles at three *reference* past instants

BUZZ Algorithm: Dense Substructure

Consider:

- A time window
- Maximum number of terms N
- K subgraphs optimizing a min-degree-based cohesiveness measure

Testbed

Evaluation parameters:

- 10 starting instants:
  - 5 in $\mathcal{T} = 1d$
  - 5 in $\mathcal{T} = 6h$
- Number of words/entities $N = 10$
- Window size:
  - BUZZ: $W \in \{1, 2, 3, 4, 5\}$
  - RG-ED: $W = 1$
- Output subgraphs
  - Entities: $K = 10$
  - Words: $K = 3$

Entities: 600 stories   Words: 180 stories

Evaluation

- Detect if stories (terms and dates) match real-world events
- Eight judges
- Parameters and algorithm used are unknown
- Classified as story if chosen by at least two editors

Graph	Method	$\lvert W \rvert$	# Events	YES Events #	YES Events %	NO Events #	NO Events %
$\mathcal{G}_e^{(1d)}$	RG-ED	1	50	45	90.00	5	10.00
	BUZZ	1	50	40	80.00	10	20.00
		2	50	34	68.00	16	32.00
		3	50	35	70.00	15	30.00
		4	50	41	82.00	9	18.00
		5	50	40	80.00	10	20.00
$\mathcal{G}_e^{(6h)}$	RG-ED	1	51	40	78.43	11	21.57
	BUZZ	1	50	38	76.00	12	24.00
		2	49	36	73.47	13	26.53
		3	50	30	60.00	20	40.00
		4	50	36	72.00	14	28.00
		5	50	38	76.00	12	24.00

Graph	Method	$\|W\|$	# Events	YES Events		NO Events	
				#	%	#	%
$\mathcal{G}_w^{(1d)}$	RG-ED	1	15	14	93.33	1	6.67
	BUZZ	1	15	14	93.33	1	6.67
		2	15	9	60.00	6	40.00
		3	15	8	53.33	7	46.67
		4	15	9	60.00	6	40.00
		5	15	9	60.00	6	40.00
$\mathcal{G}_w^{(6h)}$	RG-ED	1	15	7	46.67	8	53.33
	BUZZ	1	15	14	93.33	1	6.67
		2	15	14	93.33	1	6.67
		3	15	11	73.33	4	26.67
		4	15	13	86.67	2	13.33
		5	15	12	80.00	3	20.00

Editors' Agreement

Krippendorff's Alpha coefficient:

- Every judge evaluated a subset of all extracted stories
- Word graphs: 0.411
- Entity graphs: 0.486

Anecdotal evidence

- BUZZ and RG-ED are able to extract events
- Topics: politics, showbiz, crime news, natural disasters or catastrophic events
- Italian events
- Facts and events with worldwide relevance and echo

Graph: $\mathcal{G}_e^{(1d)}$	**Date** : 2017-01-25	**W**: 3	**N** : 10	**K** : 20

Story

ryan gosling, damien chazelle, manchester, natalie portman, emma stone, meryl streep, hacksaw ridge, mel gibson, casey affleck, la la land

Corresponding News Article

`http://www.ilpost.it/2017/01/24/oscar-2017-nomination/`

Graph: $\mathcal{G}_e^{(1d)}$	**Date** : 2017-03-03	**W**: 5	**N** : 10	**K** : 30

Story

apollo, orbita terrestre bassa, la nasa, phil larson, stazione spaziale internazionale, fra spacex, programma apollo, esplorazione spaziale, space launch system, space launch system e di orion

Corresponding News Article

`http://www.repubblica.it/scienze/2017/02/27/news/spacex_nel_2018_due_turisti_intorno_alla_luna-159397130/`

Graph: $\mathcal{G}_w^{(6h)}$	**Date** : 2017-02-22 18	**W**: 1	**N** : 20	**K** : 10

Story

nana, pianeti, eso, solare, ospitare, astronomi, distante, liegi, telescopio, gillon, temperatura, european, trappist, planetario, sosia, nasa, abitabile, nature, ultrafredda

Corresponding News Article

`http://www.ansa.it/canale_scienza_tecnica/notizie/spazio_astronomia/2017/02/22/scoperto-qualcosa-oltre-il-nostro-sistema-solare_a8647f10-e3ee-42ae-8f98-2d395aae841f.html`

Graph: $\mathcal{G}_w^{(12h)}$	**Date** : 2016-12-23 12	**W**: 2	**N** : 30	**K** : 10

Story

amri, fermato, killer, terrorista, strage, spalla, somatici, deceduto, identificato, stazione, scat, attentato, colpendolo, sparando, sparato, sparatoria, anis, poliziotti, poliziotto, pistola, zaino, tunisino, agente, agenti, berlino, ferito, ucciso, movio, fermata

Corresponding News Article

`http://www.ansa.it/lombardia/notizie/2016/12/23/milano-spara-ad-agenti-durante-un-controllo-ucciso_7dbfa79d-ca32-4d74-ac88-30038a841756.html`

- HERMEVENT is a *structured* test collection for event detection
- The text dump, the graphs and the editorial judgements are made freely available.

References

- I. Bordino and A. Ferretti and M. Firrincieli and F. Gullo and M. Paris and S. Pascolutti and G. Sabena
  Advancing NLP via a distributed-messaging approach
  Proc. of IEEE Int. Conf. on Big Data, 2(1):1561–1568, 2016.

- Paolo Ferragina and Ugo Scaiella
  TAGME: on-the-fly annotation of short text fragments (by wikipedia entities)
  Proc. of ACM Int. Conf. on Information and Knowledge Management (CIKM), 2(1):1625–1628, 2010

- F. Bonchi and I. Bordino and F. Gullo and G. Stilo
  Identifying Buzzing Stories via Anomalous Temporal Subgraph Discovery.
  Proc. of IEEE/WIC/ACM Int. Conf. on Web Intelligence