Category-Level Transfer Learning from Knowledge Base to Microblog Stream for Accurate Event Detection

Weijing Huang, Tengjiao Wang, Wei Chen, Yazhou Wang

School of Electronics Engineering and Computer Science, Peking University

@DASFAA 2017, Suzhou, China

Motivation

Many Web applications need the **accurate event detection** technique on microblog stream, including:

1. public opinion analysis [Chen, SIGIR 2013]
2. public security [Li, ICDE 2012], [Imran, WWW 2014]
3. disaster response [Sakaki,WWW2010]
4. breaking news report[1]

But detecting events on twitter stream accurately is still challenging.

[1] `http://www.theverge.com/2016/12/1/13804542/reuters-algorithm-breaking-news-twitter`

Challenges (1/2)

According to [Huang, WWW 2016], the challenges include,

1. fast changing
2. high noise
3. short length

And, we found another key factor,

1. Small events with fewer tweets ➔ Hard to trade off between precision and recall.

Challenges (2/2)

Exploratory study on the *Edinburgh twitter corpus*: 11/29 events contain less than 50 tweets.

Table: Statistics of labeled events.

Event	Date	Event Size
S&P downgrade US credit rating	05/08/2011	656
Atlantis shuttle lands	21/07/2011	595
US increases debt ceiling	25/07/2011	485
Plane with Russian hocky team Lokomotiv crashes	07/09/2011	286
Amy Winehouse dies	23/07/2011	283
Gunman opens fire in youth camp in Norway	23/07/2011	260
Earthquake in Virginia	24/08/2011	246
First victim of London riots dies	09/08/2011	174
Explosion in French nuclear plant in Marcoule	12/09/2011	135
Google announces plans to bury Motorola Mobility	15/08/2011	127
NASA announces there might be water on Mars	04/08/2011	124
Car bomb explodes in Oslo, Norway	22/07/2011	114
...	...	...
Indian and Bangladesh sign a border pact	06/09/2011	25
Flight 4896 crash	13/07/2011	21
First aritficial organ transplant	12/07/2011	18
three men die in riots in england	10/08/2011	16
rebels capture interational tripoli airport	21/08/2011	13

11 Small events with fewer tweets

How about existing methods? (1/2)

Event detection methods *without extra information*, such as

1. clustering articles
   - LSH[Petrovic, NAACL 2010]
   - need to set threshold to determine whether new article represents a new event.
2. analyzing word frequencies
   - EDCoW[Weng, ICWSM 2011]
   - treat the word as the basic unit in analysis, without regarding polysemy words (words have different meanings, e.g. "apple")
3. finding bursty topics via topic modeling
   - TimeUserLDA[Diao, ACL 2012], BurstyBTM[Yan, AAAI 2015]
   - detects the "large" events but may ignore the "small" ones.

How about existing methods? (2/2)

Event detection methods *by leveraging extra information*

1. typical one: Twevent[Li, CIKM 2012]
   - divides the tweet into segments according to the Microsoft Web N-Gram service and Wikpedia
   - detects the bursty segments and cluster these segments into candidate events
   - still has to trade off between precision and recall

An Example

Much easier to detect the event on the time series of word "hood" related to *Military*, **without adjusting the threshold**.

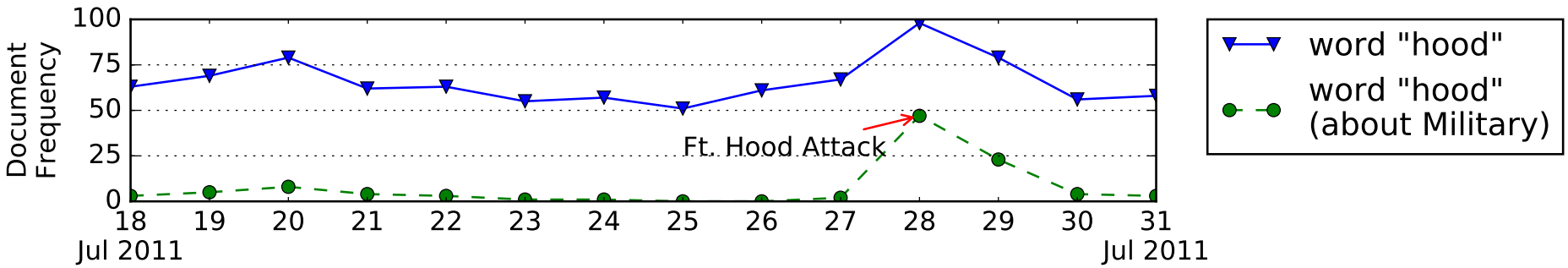

Figure: The comparison of the time series between the raw word *hood* and the *Military* related word *hood*, computed on the *Edinburgh twitter corpus*. Refer the event to `https://en.wikipedia.org/wiki/Fort_Hood#2011_attack_plot`.

Fort Hood

From Wikipedia, the free encyclopedia

Fort Hood is a U.S. military post located in Killeen, Texas. The post is named after Confederate General John Bell Hood. It is

The insights on the example

Knowledge Base

- Well organized
- Constructed elaborately
- full of rich information

Microblog stream

- Short length
- Fast changing
- High noise

The benefit of **enriching the semantics** and **filtering out noise** by Knowledge Base for microblogs is attractive.

But it's expensive to retrieve every word of tweets in the Knowledge Base.

Overview of our solution

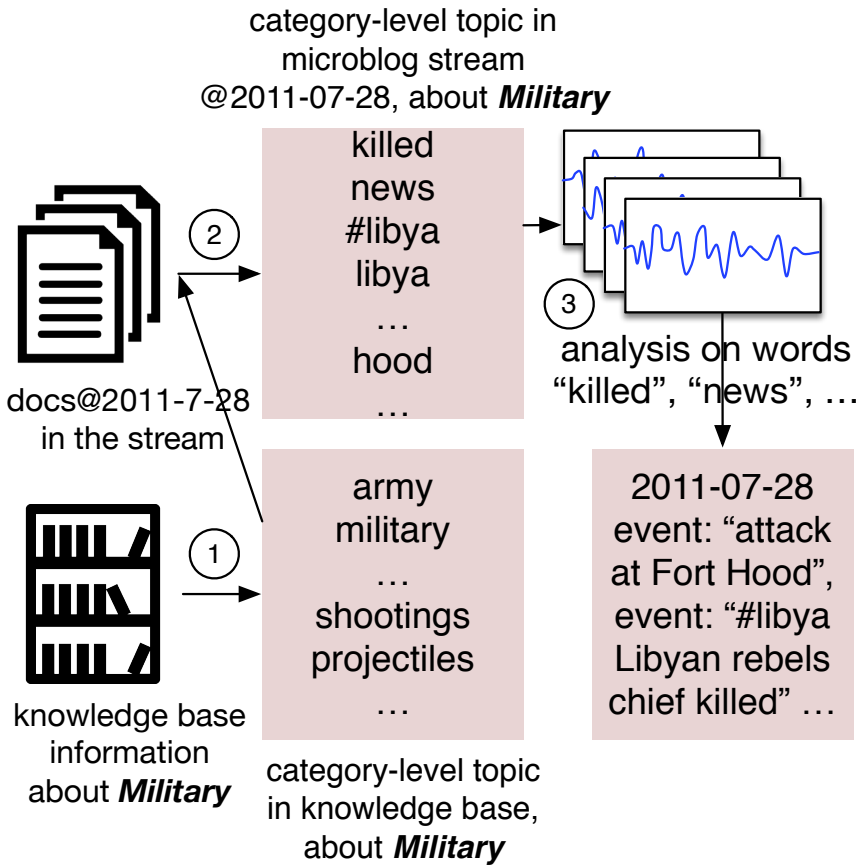

Figure: TransDetector's processing flow, taking *Military* related events in microblogs as an example.