Enhancing Event Descriptions through Twitter Mining

Hristo Tanev and **Maud Ehrmann**
J.R.C., European Commission
Via Fermi 2749
Ispra, Italy

Jakub Piskorski
Frontex
Rondo ONZ 1
Warsaw, Poland

Vanni Zavarella
J.R.C., European Commission
Via Fermi 2749
Ispra, Italy

Abstract

We describe a simple IR approach for linking news about events, detected by an event extraction system, to messages from Twitter (tweets).[1] In particular, we explore several methods for creating event-specific queries for Twitter and provide a quantitative and qualitative evaluation of the relevance and usefulness of the information obtained from the tweets. We showed that methods based on utilization of word co-occurrence clustering, domain-specific keywords and named entity recognition improve the performance with respect to a basic approach .

Introduction

Classical online news represent useful sources for collecting information on security-related events. However, microblogging services, in particular those which are accessible through mobile clients, might provide important near real-time updates on ongoing crises. A specific feature of such services in the context of reporting on crisis situations is the fact that they can provide complementary information and offer less popular and/or alternative views that differ from conventional news media reporting.

We describe a simple IR approach which links news about events, detected by our event extraction system, to messages from Twitter (tweets), a micro-blogging service that proved to be an useful source of information for situational awareness. While the previous research has focused on detecting unknown events from tweets, we think that event extraction can be carried out in a more accurate and successful way from news articles and that tweets, short and usually rather noisy, can be used to retrieve additional real-time content related to those events. In particular, we explore several methods for creating event-specific queries for Twitter and prove that these methods, specifically those based on utilization of word co-occurrence clustering, domain-specific keywords and named entity recognition improve the performance w.r.t. the basic configuration. We present a quantitative and qualitative evaluation of the relevance and usefulness of the information obtained from the tweets and from the online sources to which they point.

Related Work

There are several research groups, who work on classical event extraction from online news and other suchlike publicly accessible information sources, e.g., (Yangarber et al. 2005) reports on a system for extraction of outbreaks of infectious diseases from medical alerts, whereas (Tanev, Piskorski, and Atkinson 2008) present systems that extract structured information on violent events and man-made/natural disasters from online news.

In the recent years, utilization of Twitter for gathering information for situational awareness during crisis situation has attracted a lot of attention, e.g., in the context of the earthquakes in Japan (Sakaki, Okazaki, and Matsuo 2010), Red River floods and Oklahoma Grassfires (Vieweg et al. 2010). However, most of the research centered around methods for detection and classification of unknown events in Twitter (e.g., (Becker, Naaman, and Gravano 2010; Weng and Lee 2011)). Relatively little work focused on extracting content about events which are known *a priori*, e.g., (Becker et al. 2011) present a system that relies on strategies for automatically constructing queries for retrieving Twitter content about events, based on structured information on planned events available on publicly available web sites. The aforementioned work is to a certain extent similar in nature to our work.

We utilise some NLP techniques in the process of query creation and tweet relevance ranking. Other authors report on techniques for classification of 'event' and 'non-event' tweets that deploy shallow linguistic analysis, e.g., (Verma et al. 2011) reports on an approach of using linguistic features (e.g., objectivity, impersonality, formality, etc.) for detecting tweets with content relevant to situational awareness during mass emergencies. Some work elaborate on ML-based techniques for Named Entity Recognition (NER) from tweets, which deploy linguistic features (e.g., part-of-speech, shallow parsing) (Ritter et al. 2011)). We benefit from NER in the process of tweets relevance ranking, although we use for this purpose a NER component tuned to process classical news.

[1] The research work presented in this paper did not involve and does not concern the processing of personal data of any kind.

Event Extraction Framework

We deploy a real-time, multilingual engine for security-related event extraction (crisis and border events, medical hazards) from online news (Tanev, Piskorski, and Atkinson 2008; Atkinson et al. 2011). The system is capable of processing news in 8 languages and is based on a blend of manually created linguistic rules, semi-automatically acquired linear patterns and domain knowledge, represented as domain-specific heuristics and taxonomies. The system receives news articles from a news aggregation platform, which processes ca. 100,000 news articles per day in 50 languages. It automatically detects event reporting from them and fills an event description template representing key information about the event, including the event type, time, location, actors and victims.

Retrieving Twitter Messages

We aim at retrieving tweets related to specific security-related events previously extracted by our event extraction system from news articles. The event extraction system processes the title and the lead sentences of the news article. Consequently, our approach considers only this text. It is based on the following schema:

1. The title and the first sentence of the article are represented as a weighted vector of keywords and key phrases.
2. Several Boolean queries are formulated from these keywords and phrases.
3. The queries are submitted to the Twitter API and each of the retrieved tweets is represented as a term vector in the same way as it is done in step 1.
4. The tweets, obtained in step 3, are ranked according to the similarity with the vector representation of the news article.

Basic approach

The main resource in our method is an index of word unigrams and bigrams, obtained from a corpus of 1 million English language news articles. In the index, each word unigram and bigram is accompanied with its frequency and the frequency of the co-occurrences with the other uni/bigrams. Singletons are not included in the index. This index is used to: (a) detect known bigrams in the input news article, (b) calculate IDF for each term, and (c) suggest classes of terms which are used to formulate the queries to Twitter based on the co-occurrence information.

The basic approach for creating the term vector and the queries is as follows. First, all the words unigrams and bigrams from the input news article, which also appear in the index are extracted. The term vector is formed from these uni- and bigrams - the term weights are defined by their inverse document frequency (IDF). Next, terms are sorted in an descending order of their IDF scores. Subsequently, the following four Boolean queries are formed from the terms t_1, t_2 and t_3 with the highest IDF scores: t_1 AND t_2; t_1 AND t_3; t_2 AND t_3; t_1. The first three queries are AND clauses, while the last one is a single term. These queries are submitted to the Twitter Search API in this order, until a certain number of tweets is retrieved, i.e., sometimes only a part of these queries is used. The idea is to submit first more specific queries and only if they do not return enough tweets, more generic terms are used.

We experimented with various ways of creating the term vector and formulating the Boolean queries, including utilisation of: (a) word clustering, based on word co-occurrences, (b) named entity recognition, and (c) a dictionary of keywords, specific for the domain of violence and crises. To better illustrate the three approaches we will use the following text from a news article about an armed conflict, detected by our event extraction system (the text comprises the title and the lead sentence):

> *Two Yemeni soldiers, 13 Islamists killed in clashes*
> *Fresh fighting between suspected Al-Qaeda militants and army troops in Yemen's restive Abyan province have killed 2 soldiers and 13 Islamists, military officials said on Tuesday*

First, a basic term vector is created, in which the top scored terms and their weights are: (*abyan province* 0.17, *fresh fighting* 0.03, *abyan* 0.008, *qaeda militants* 0.009, *army troops* 0.004, ...) Following the basic query formulation approach, the following queries are created: 'abyan province' AND 'fresh fighting', 'abyan province' AND 'qaeda militants', 'fresh fighting' AND 'qaeda militants', 'abyan province'. These queries are then submitted in this order until certain number of tweets are acquired. Now, let's see how our vector and query formulation approaches change this query and the term vector.

Using word co-occurrences

Using word co-occurrences for query expansion is a known IR technique, however its usefulness has been disputed (Peat and P. Willett 1991). A potential problem in using co-occurrence data is the ambiguity of the words, e.g., 'bank' co-occurs with words like 'river' and 'boat', which correspond to its meaning 'river bank' or 'money', 'loan' and others, which correspond to the meaning 'financial institution'. To avoid this issue, we performed query expansion using co-occurring terms which: (a) appear in the input news article, (b) are in the first half of the ordered term list.

More precisely, the algorithm works as follows. We build a graph of co-occurrences from the extracted article terms, where each vertex is a term and each edge is labeled with a number which shows the co-occurrence weight between the two terms (taken from the co-occurrence index). Then, we apply the Newman-Girvan graph clustering algorithm (Girvan. and Newman 2002) to detect term clusters. Finally, terms in the basic query are expanded with all the terms from the same cluster, which are in the first half of the term list. Terms from the same cluster are connected in an OR clause.

In our example, the term 'abiyan province' is expanded with the term 'restive' and the term 'qaeda militants' is expanded with 'islamists', 'yemen', 'clashes' and 'suspected'. The term vector for our example remains the same as the basic one, while the queries become: ('abyan province' OR 'restive') AND ('fresh fighting'), ('abyan province' OR 'restive') AND ('qaeda militants' OR 'islamists' OR 'yemen' OR 'clashes' OR 'suspected'), ('fresh fighting')

AND (‘qaeda militants’ OR ‘islamists’ OR ‘yemen’ OR ‘clashes’ OR ‘suspected’), (‘abyan province’ OR ‘restive’). We experimented with this expansion and evaluated it before making the final experiments. Since the results from this preliminary evaluation seemed encouraging, we used the expanded queries as a base for the experiments with named entities and keywords, rather than the basic queries.

Named entities

Named entities identify important features of the event, e.g., participating actors, organizations and the location. Therefore, we expect to increase the precision through introduction of an additional restriction that named entities of certain types which appear in the input text, appear also in the retrieved tweets. We experimented with considering: (a) person and organization names, (b) locations, and (c) combination of the former two types. When considering a certain type of named entities, we augment the weight of this type of named entities in the term vector.

In our example ‘Al-Qaeda’ is detected as a name of organization and ‘Yemen’ as a name of location (our system for location detection failed to recognize ‘Abyan province’). Consequently, in the run in which we consider persons and organizations, the suffix *AND ‘al-qaeda’* is added to the queries, e.g., the first query becomes (‘abyan province’ OR ‘restive’) AND (‘fresh fighting’) AND ‘al-qaeda’. When considering locations, the query becomes (‘abyan province’ OR ‘restive’) AND (‘fresh fighting’) AND ‘yemen’; when considering both persons/organizations and locations, it becomes (‘abyan province’ OR ‘restive’) AND (‘fresh fighting’) AND ‘al-qaeda’ AND ‘yemen’.

Domain-specific keywords

Our event extraction system utilises a set of keywords from the domain of violence and crisis to detect the event type. We used this list to boost the terms which contain such keywords. In our example, ‘militants’, ‘troops’, ‘soldiers’, and ‘army’ and the bigrams which contain them are boosted by multiplying their weight by 4 - an empirically-set coefficient. The list of terms is reordered accordingly, which might yield a different list of n-grams used to created the query.

In our example, the bigrams ‘qaeda militants’ and ‘army troops’ go on the top of the list and enter the queries, together with their co-occurrence clusters. Consequently, the queries will become (‘qaeda militants’ OR ‘islamists’ OR ‘clashes’ OR ‘soldiers’) AND (‘army troops’) , (‘qaeda militants’ OR ‘islamists’ OR ‘clashes’ OR ‘soldiers’) AND (‘abyan province’ OR ‘restive’), (‘army troops’) AND (‘abyan province’ OR ‘restive’), (‘qaeda militants’ OR ‘islamists’ OR ‘clashes’ OR ‘soldiers’).

Experiments

Experiments were carried out with 7 different query expansion methods, that use: (a) just the terms with highest IDF (BS) - the baseline, (b) word co-occurrence clustering (CO), (c) domain-specific keywords (KW), (d) locations (LOC), (e) person and organizations (PE), (f) all named entities (PE-LOC), and (g) keywords and named entities (KW-PE-LOC). Methods (c)-(g) all use word co-occurence clustering.

Table 1: Relevance evaluation

	yield ($\geq$ 0.49)	yield	precision ($\geq$ 0.49)	precision
BS	11.02	16.23	73	60
CO	11.15	17.64	74	53
KW	14.54	**20.75**	57	46
LOC	12.51	15.62	**75**	**67**
PE	10.83	15.27	69	54
PE-LOC	11.12	13.59	68	61
KW-PE-LOC	**14.63**	17.41	68	62

For carrying out the experiments we selected 40 English news articles published in the second half of December 2011, from which our event extraction system has detected an important security-related event. 28 of the events were related to violence (e.g., armed conflicts, terrorist attacks), 9 were related to disasters (i.e., aircraft crashes, maritime accidents, floodings and storms), 2 were related to biological threats, and 1 was related to a political meeting.

For each query expansion configuration Twitter API was called using the generated queries until at least 50 tweets were returned or the list of generated queries was exhausted. Tweets published more than 2 days before the news article were excluded. Still, this time window was too long for some events and resulted in capturing similar events from the previous days; however, it allowed for retrieving earlier information preceding the publication of the news article. This was especially relevant for disasters and armed conflicts for which the article reflected some intermediate stage of their development, while earlier information was still relevant.

We evaluated two aspects of the retrieved tweets: (a) their relevance to the event described in the news article, and (b) the complementarity of the information contained in them or in the information sources to which they point to, with respect to the information in the news article.

Relevance The relevance evaluation was performed on all 40 events by three evaluators, in such a way that the events were distributed evenly and the tweets for each event were evaluated by two people. Between each pair of evaluators, there were 13-14 shared events. Each tweet was marked as relevant to the input news article, if it referred to the main event described therein. It could be an opinion, question, update or just stating that the event took place.

We measured the kappa between the evaluators on the evaluation of the BS retrieval method. The kappa between the various evaluator pairs was as follows: 1st and 2nd - 0.98, 1st and 3rd - 0.79 and 2nd and 3rd - 0.69. These figures show that the task for measuring the relevance was well grounded.

We used two measures to evaluate the relevance: (a) *yield* - the number of relevant tweets in the response per event, and (b) *precision* - the percentage of the relevant tweets with respect to all retrieved tweets. Since each tweet was evaluated by two different evaluators, we calculated the average yield and precision for all events and all evaluators.

For each retrieval method we report two results - yield and precision measured on tweets with similarity score ≥ 0.49 and on all the retrieved tweets with similarity score > 0. The few 0-scored tweets were discarded from the evalua-

Table 2: Complementarity evaluation

	C	CP	CP_T	CP_{LT}	CP_{LI}	CP_{LV}
BS	0.714	0.233	0.012	0.171	0.157	0.089
CO	0.714	0.213	0.006	0.169	0.148	0.086
KW	0.857	0.219	0.012	0.145	0.112	0.041
LOC	0.714	0.262	0	0.187	0.2	0.114
PE	0.714	0.217	0.017	0.140	0.141	0.034
PE-LOC	0.714	0.275	0.011	0.16	0.187	0.049
KW-LOC-PE	0.714	0.272	0.011	0.182	0.133	0.046

tion, since these were mostly badly formatted tweets our system was not able to process. The '0.49' threshold was set based on experiments with a smaller development corpus. Table 1 shows the relevance figures. It can be observed that co-occurrence expansion improves the yield, however it deteriorates the total precision. Using keywords gives the best yield, while using locations to restrict the queries results in the best precision. Using person and organization names does not seem to improve the results. The combination of all methods (KW-PE-LOC) gives reasonable performance both in terms of yield and precision.

Complementarity The complementarity evaluation was performed only on 7 events by one evaluator due to the laborious nature of this task. In particular, we calculated: (a) *general complementarity* (C), which is set to 1 in case there is at least one tweet in the response, which provides new information or a link to a source that provides new information not available in the original news article (e.g., new facts), or 0 otherwise, (b) *percentaged complementarity* (CP), i.e., the fraction of tweets in the response, which provide new information or a link to a source that provides new information, (c) CP_T - fraction of tweets in the response, which provide new information in the body of the tweet itself (disregarding the links), (d) fraction of tweets with a link to a source that provides new information in form of: a text (CP_{LT}), an image (CP_{LI}), or a video (CP_{LV}).

Table 2 provides average complementarity scores. For computing the complementarity figures the entire response was used (i.e., the 0.49 threshold was not used). The complementarity evaluation on such a small dataset does not constitute a conclusive evidence, but we can hypothesize that the fraction of tweets which provide new information in the body of the tweet is low (CP_T 1%) and the query expansion methods which utilise keywords and named entities yield higher percentaged complementarity. Most of the complementary information is available through sources to which tweets provides links, however a vast majority of such links point to other news sources, whereas only a tiny fraction point to non-news portals, e.g., other social media, etc. Finally, we could hypothesize that there is no correlation between the relevance rank and complementarity of the tweets.

Discussion

Our experiments show that news articles on security-related events can be linked with Twitter messages (possibly containing complementary information) precisely enough for practical purposes. We explored several methods for query expansion, which improved the performance. We intend to further improve the precision through considering: (a) hashtags, (b) ranking authoritativeness of tweet sources, (c) deployment of sentiment analysis to filter out subjective tweets, which usually do not provide any new content, and (d) automatically generating queries in different languages to broaden the coverage. Finally, in order to have a deeper insight into the real value of Twitter for obtaining complementary information, more domain-tailored evaluations over a longer period of time are envisaged, e.g., in the domain of border security, which will focus on security-related events in Africa and the middle East.

References

Atkinson, M.; Piskorski, J.; Yangarber, R.; and van der Goot, E. 2011. Multilingual Real-Time Event Extraction for Border Security Intelligence Gathering. In *Open Source Intelligence and Counter-terrorism*. Springer, LNCS, Vol. 2.

Becker, H.; Chen, F.; Iter, D.; Naaman, M.; and Gravano, L. 2011. Automatic identification and presentation of twitter content for planned events. In *Proceedings of ICWSM 2011*.

Becker, H.; Naaman, M.; and Gravano, L. 2010. Learning similarity metrics for event identification in social media. In *Proceeedings of WSDM 2010*, 291–300.

Girvan., M., and Newman, M. 2002. Community structure in social and biological networks. In *Proceedings of Nationall Academy of Sciences of USA*.

Peat, H. J., and P. Willett, P. 1991. The limitations of term co-occurrence data for query expansion in document retrieval systems. *Journal of the American Society for Information Science* 42(5):378–383.

Ritter, A.; Clark, S.; Mausam; and Etzioni, O. 2011. Named Entity Recognition in Tweets: An Experimental Study. In *Proceedings of EMNLP 2011*, 1524–1534.

Sakaki, T.; Okazaki, M.; and Matsuo, Y. 2010. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. In *Proceedings of WWW 2010*, 851–860. ACM.

Tanev, H.; Piskorski, J.; and Atkinson, M. 2008. Real-Time News Event Extraction for Global Crisis Monitoring. In *Proceedings of NLDB 2008.*, 207–218.

Verma, S.; Vieweg, S.; Corvey, W.; Palen, L.; Martin, J.; Palmer, M.; Schram, A.; and Anderson, K. 2011. Natural Language Processing to the Rescue? Extracting "Situational Awareness"' Tweets During Mass Emergency. In *Proceedings of ICWSM 2011*, 385–392. AAAI.

Vieweg, S.; Hughes, A.; Starbird, K.; and Palen, L. 2010. Microblogging During Two Natural Hazards Events: What Twitter May Contribute to Situational Awareness. In *Proceedings of CHI 2010*, 1079–1088.

Weng, J., and Lee, B.-S. 2011. Event Detection in Twitter. In *Proceedings of ICWSM 2011*, 401–408. AAAI.

Yangarber, R.; Jokipii, L.; Rauramo, A.; and Huttunen, S. 2005. Extracting Information about Outbreaks of Infectious Epidemics. In *Proceedings of the HLT-EMNLP 2005*.