A Data Collection for Evaluating the Retrieval of Related Tweets to News Articles

Axel Suarez[1,2], Dyaa Albakour[2], David Corney[2], Miguel Martinez[2], and José Esquivel[2]

[1] School of Computer Science and Electronic Engineering, University of Essex, UK
[2] Signal Media Ltd., 32-38 Leman Street, London, E1 8EW, UK
research@signalmedia.co

Abstract. Nowadays, social media users react in real-time to local and global events. Therefore, social media can be used to measure the impact of particular topics or events and to analyze public opinion. To this end, identifying and ranking social media posts, such as tweets, associated with a news article is an important information retrieval task. In this paper, we devise a new data collection to evaluate approaches for the task of related-tweet retrieval for news articles. Using two sets of a) mainstream news articles and b) tweets from curated newsworthy sources from the same period, we use a TREC-like pooling approach to associate news articles with relevant tweets. We also provide a benchmark for the related-tweet retrieval task by evaluating a number of retrieval approaches on this new data collection.

1 Introduction

In recent years, the way people produce and consume news has radically changed [1]. Where people used to read printed newspapers, many now read news websites and blogs. Along with these sites, social media platforms, such as Twitter, include content from many mainstream news sources, such as the BBC and The New York Times. However, mainstream news editors are not the only source of news, as individual social media users can report local or global events in real-time or comment on them afterwards. Recent studies have shown that social media posts can help understanding public opinion and sensing the state of the world. Indeed, Twitter has been used to replace opinion polls [2], to predict stock market movements [3], and to discover local events in a city [4]. Therefore, finding tweets related to news articles can be useful to analyze an event's context. For example, Figure 1 shows two tweets that are related to a news article on the topic of Obamacare. In the first tweet, the user is sharing a link to the article to inform their followers about the event that Trump criticized Obamacare. In the second tweet, the user is commenting on the event and informing followers on their views about it. Finding, ranking and aggregating tweets about a particular news article may be helpful to understand the popularity and virality of the article, and also to analyze what the public thinks about the topic.

In this paper, we consider the task of *related-tweet retrieval*, where the aim is to identify and rank tweets that are related to a published news article. Examples of previous related work include the tasks of classifying tweets associated with news articles

Fig. 1. Examples of tweets related to a news article published in the New York Times and titled: *"Let Obamacare Fail", Trump says as G.O.P Health Bill Collapses*.

according to their subjectivity [5] and associating news articles with relevant twitter hashtags [6]. However, the data collections used for these tasks are not suitable for evaluating the retrieval of related tweets to news articles. Therefore, we create and present a new TREC-like data collection of relevance judgments for evaluating the task of related-tweet retrieval. In particular, we use the Signal "One-Million News Articles Dataset" [7], which contains articles from multiple sources, and a collection of tweets, created by Brigadir *et al.* [8], from a curated list of newsworthy sources. We follow a pooling approach, where we select a sample of 100 news articles, and propose a number of retrieval methods to create a pool of tweets for each news article in the sample. The pool is then annotated with relevance judgments. Furthermore, we use our created data collection to evaluate the retrieval methods used in the pooling process to provide a benchmark for the related-tweet retrieval task. We make the resulting data collection publicly available for research purposes[3] and together with the results of this paper, we aim at encouraging further research on this task.

The rest of the paper is structured as follows. Section 2 describes our methodology to collect the relevance judgments using a pooling approach. Section 3 describes the data collection and gives insights from the annotation process and the results of retrieval evaluation. Finally, Section 4 summarizes our conclusions.

2 Methodology

2.1 The Pooling Approach

There are 3 components in a test data collection for a document retrieval task [9]: (i) *the corpus*; (ii) the set of information need statements, i.e. *information needs*; and (iii) *the relevance judgments* that indicate which documents should be presented for a particular information need. Generally, the most expensive component to produce, in terms of time and effort, are the judgments. Therefore, we follow a pooling approach to reduce the number of annotations, as is common in the TREC evaluation framework [10]. In our case, the corpus is a set of tweets, while the information needs are a random subset of articles selected from a larger set of news articles. To collect relevance judgments, we propose a variety of retrieval methods to retrieve related tweets. We then merge the top k tweets ranked by each retrieval method to create a pool of diverse tweets (see Figure 2).

[3] The data collection can be downloaded through this link:
`http://research.signalmedia.co/datasets/signal1m-tweetir.html`

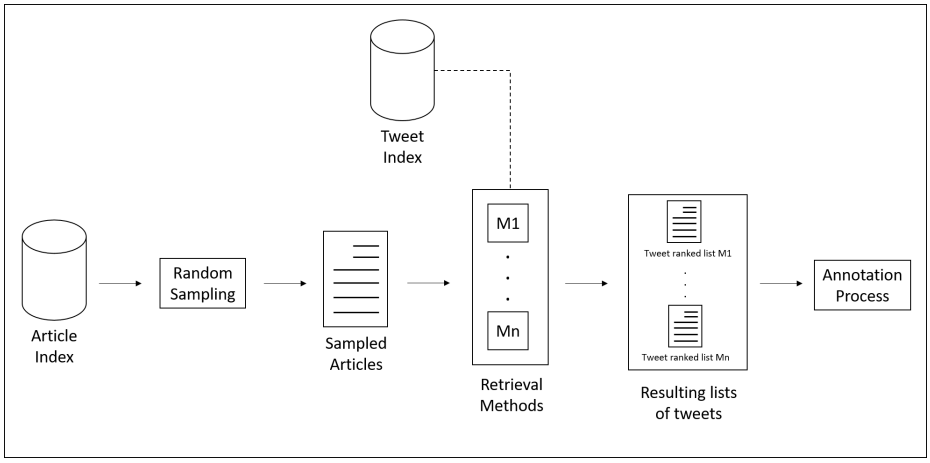

Fig. 2. Illustration of the pooling process

Each proposed related-tweet retrieval method generates a query Q. Using this query, the tweets are ranked with Lucene's Practical Scoring Function[4], which is the sum of the boosted and normalized tf-idf score for all terms in the query Q. We propose and compare eight retrieval methods:

M1. **Title search**: The title of the article is used as a query. The intuition is that the title represents a condensed version of the important subjects of the article.
M2. **Summary search**: We generate a summary for each article and use it as a query. We use the first two sentences of the article as a summary. This is an intuitive, yet effective, summarization approach that reflects the way journalists typically structure news articles.
M3. **Content search**: We use the full content of the article as a query.
M4. **Summary + date search**: Same as M2, but the results are filtered so that only tweets posted on the day that the article is published are retrieved.
M5. **Bi-gram phrase search**: We train a bi-gram phrase recognition model, developed in [11], on a large collection of news articles. Using the trained model, we extract all the phrases in the summary of an article (as in M2) and use them as a query.
M6. **Named entity search**: With this approach, the intuition is to retrieve tweets about people, organizations and places mentioned in the article. We extract named entities using the Stanford NER Tagger configured with the default English 3-class model trained without part-of-speech tagging. We generate a query that consists of all terms representing the people, organizations and places mentioned in the summary.
M7. **Semantic summary search**: This method uses the 10 most 'semantically significant' terms in the news article as a query. To find these, we train a word2vec model [12] on a large collection of news articles to get the vector representation of each term in the article's summary, and compute the summary's centroid as follows:

$$\mu(c) = \frac{1}{|W_{summary}|} \sum_{w \in W_{summary}} v(w) \tag{1}$$

[4] `https://lucene.apache.org/core/4_6_0/core/org/apache/lucene/search/similarities/TFIDFSimilarity`

where $W_{summary}$ is the set of all the terms in the summary, and $v(w)$ is the vector representation of the term w using the word2vec model. After computing the centroid, we rank the terms by their cosine similarity to the centroid and select the top 10 terms as a query. The intuition here is that the closer a term is to the centroid, the more likely that it is relevant to the topic of the article.

M8. **Query expansion search**: This method uses the same 10 terms generated by M7 and expands the query with 10 different terms. In particular, for each of the 10 original terms t_i, we expand the query with the term in the word2vec space that has the highest cosine similarity to t_i. The intuition here is to fill the vocabulary gap where the related tweets do not mention the exact terms in the article.

2.2 Annotation for Relevance Judgments

In order to determine the performance of each of the aforementioned eight methods, we ask human annotators to provide relevance judgments. In particular, each tweet in the generated pool is annotated according to its relatedness to the corresponding article by labelling it with one of three different labels (grades of relevance):

1. **Non-relevant**: The tweet is about a completely different topic, or covers a similar topic but focusses on aspects not covered by the news article.
2. **Somewhat-relevant**: When the tweet refers to an event closely related to the main event of the article, or when it talks about a secondary theme mentioned at least once in the article.
3. **Completely-relevant**: When the tweet talks about the main topic of the article or directly mentions the article itself.

3 Data Collection and Experiments

3.1 Datasets

The Signal One Million News Articles (Signal-1M) dataset[5] contains meta-data about each article, such as title, content, and publication date. The Twitter dataset, by Brigadir *et al.* [8], consists of over 3.2 million tweets from a curated list of major news sources and journalists, which cover all their posts from the same range of dates as Signal-1M[6]. For pooling, we randomly selected 100 articles from Signal-1M, which were then passed into the eight retrieval methods (Section 2.1). Each retrieval method generated a ranked list of tweets for each article. We used a cut-off point k=10 on each ranked list, and merge the tweets, whilst removing duplicates, resulting in 62.3 distinct tweets per article on average. We asked 10 undergraduate students to annotate the pool of tweets, as per Section 2.2. Finally, we train the phrase model used in M5, and the word2vec model used in M7, with a collection of 17 million news articles collected from the same sources of Signal-1M.

[5] http://research.signalmedia.co/newsir16/signal-dataset.html

[6] https://github.com/igorbrigadir/newsir16-data/tree/master/twitter/curated

3.2 Annotation Agreement

We performed an annotator agreement experiment, where we gave the same tweets associated with 10 different news articles to 3 different annotators. Table 1 reports the agreement results. In the first row, we report the pair-wise agreement between annotators when considering all three labels of relevance. We see that the task is not trivial, as annotators only agree on the exact label 70.48% of the time on average. We also consider binary labels of relevance (rows 2 and 3), by merging the 'somewhat relevant' label with 'completely relevant' and 'non-relevant' respectively. Even with binary labels, the agreement is not perfect, but it is line with various retrieval tasks reported in TREC [10]. Next, we use the binary labels obtained for all 100 articles to evaluate the proposed retrieval methods.

Table 1. Agreement rates between the three annotators, A, B and C.

Agreement between annotators	**A and B**	**A and C**	**B and C**	**Average**
3 Labels (Exact match)	68.05%	75.87%	67.53%	70.48%
2 Labels (Relevant = Somewhat relevant)	73.09%	81.94%	74.82%	76.62%
2 Labels (Somewhat relevant = Non-relevant)	**87.15%**	**88.02%**	**84.54%**	**86.57%**

3.3 Retrieval Results

Table 2 summarizes the retrieval results using binary relevance judgments obtained when merging the 'completely relevant' label with 'somewhat relevant'. Retrieval methods (M1-M4) perform well despite their simplicity. In particular, using the summary as a query (M2) outperforms all other methods in terms of MAP, P@5 and P@10. Perhaps surprisingly, M4 shows the worst performance among these four (MAP=0.41), although it is only a slight variation of the best performing retrieval method (M2). In M4, date filtering is used to select tweets posted just before or after the article is published. However, our results suggest that sometimes social media posts may discuss events and topics before they make it to mainstream media, or longer after they are published in mainstream media. The retrieval performance of the more complex methods (M5-M8) is worse than simpler methods (M1-M4), as their MAP scores are markedly lower. It is noteworthy however, that the Query Expansion method (M8) produces a slightly higher MAP score than the non-expanded version (Semantic Summary, M7).

4 Conclusion

We have created a new data collection that combines two existing datasets (news articles and tweets) and adds value to both. Annotating tweets with relevance judgments, on their relatedness to a news article, yields interesting insights, as we have observed that the inter-annotator agreement is not perfect. Furthermore, we have used our collection to evaluate a number of retrieval methods for the task of related-tweet retrieval. Our results show that simple approaches, e.g. using the terms in the title or the summary of the article as a query, can be very effective for this task. More complex approaches,

Table 2. Retrieval scores for the propsed retrieval methods: Mean Average Precision (MAP); Precision at 5 (P@5); Precision at 10 (P@10)

Method		**MAP**	**P@5**	**P@10**
M1	Title search	0.62	0.52	0.48
M2	Summary search	**0.67**	**0.59**	**0.55**
M3	Content search	0.63	0.53	0.51
M4	Summary + sate search	0.48	0.40	0.36
M5	Bi-gram phrase search	0.41	0.32	0.31
M6	Named entity search	0.44	0.35	0.31
M7	Semantic summary search	0.37	0.28	0.29
M8	Query expansion search	0.40	0.28	0.27

such as phrase and entity search, failed to perform well on this task. This opens opportunities to consider more elaborate approaches to effectively bridge the gap between the vocabulary used in mainstream media and social media. To this end, the created data collection and the results presented in this paper will foster developing such approaches. For example, as future work, we aim to use the relevance judgments in our data collection to develop a learning-to-rank model for related-tweet retrieval.

References

1. Martinez-Alvarez, M., Kruschwitz, U., Kazai, G., Hopfgartner, F., Corney, D., Campos, R., Albakour, D.: First international workshop on recent trends in news information retrieval (NewsIR16). In: Proceedings of ECIR. (2016) 878–882
2. O'Connor, B., Balasubramanyan, R., Routledge, B.R., Smith, N.A.: From tweets to polls: Linking text sentiment to public opinion time series. In: Proceedings of ICWSM. (2010)
3. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. Journal of Computational Science **2**(1) (2011) 1–8
4. Albakour, M., Macdonald, C., Ounis, I., et al.: Identifying local events by using microblogs as social sensors. In: Proceedings of OAIR. (2013) 173–180
5. Kothari, A., Magdy, W., Darwish, K., Mourad, A., Taei, A.: Detecting comments on news articles in microblogs. In: Proceedings of ICWSM. (2013)
6. Shi, B., Ifrim, G., Hurley, N.: Be in the know: Connecting news articles to relevant twitter conversations. arXiv:1405.3117 (2014)
7. Corney, D., Albakour, D., Martinez-Alvarez, M., Moussa, S.: What do a million news articles look like? In: Proceedings of NewsIR'16 Workshop at ECIR. (2016) 42–47
8. Brigadir, I., Greene, D., Cunningham, P.: Detecting attention dominating moments across media types. In: Proceedings of NewsIR'16 Workshop at ECIR. (2016)
9. Buckley, C., Dimmick, D., Soboroff, I., Voorhees, E.: Bias and the limits of pooling for large collections. Information Retrieval **10**(6) (2007) 491–508
10. Voorhees, E.M., Harman, D.K., et al.: TREC: Experiment and evaluation in information retrieval. Volume 1. MIT press Cambridge (2005)
11. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of NIPS. (2013) 3111–3119
12. Mikolov, T., Yih, W.t., Zweig, G.: Linguistic regularities in continuous space word representations. In: Proceedings of NAACL. (2013) 746–751