Capitalizing on a TREC Track to Build a Tweet Summarization Dataset

Alexis Dusart
Karen Pinel-Sauvagnat*
Gilles Hubert*
alexis.dusart@irit.fr
karen.sauvagnat@irit.fr
gilles.hubert@irit.fr
Universite de Toulouse UPS-IRIT

ABSTRACT

Today there is a lack of standard collection for automatic tweet summarization evaluation. The construction of such a large dataset is very tedious. In this paper, we check whether the dataset proposed for the TREC Incident Streams track, which was not created for automatic summary generation, could be used in this way. Indeed, when filtering the TREC Incident Streams (IS) dataset with the assessors' annotations, it appears to respect the citeria identified in the literature related to automatic summarization. For this, we studied the TREC IS dataset and then proposed a subset summarizing each event, based on the assessors' annotations. This subset is evaluated according to the criteria previously mentioned. Several widely used state-of-the-art models for automatic text summarization, adapted to tweet summarization, were finally tested on the proposed dataset. The code, the annotations and the results are provided on our Github.

1 INTRODUCTION

Information Retrieval (IR) is tightly linked with evaluation. Evaluation campaigns were proposed by the community to promote evaluation and allow the researchers to confront their approaches. Since the beginning of TREC[1] in 1992, many evaluation collections were built and widely used to support experiments.

Information sources are simultaneously evolving and the well-formed text from the beginning is now often replaced by less conventional texts such as user-generated contents (UGC). Those contents raise new open issues and require the construction of new evaluation frameworks. Twitter is an example of such UGC that generate a huge and heterogeneous volume of data. Among linked evaluation tracks one can cite the TREC Microblog one between 2001 and 2015 [37] [23], the TREC RTS (Real-Time Summarization) one built between 2016 and 2018 [22], the TREC IS (Incident Streams) one built since 2018 [24], and the INEX Tweet Contextualization task [3].

Those campaigns mainly evaluated adhoc retrieval [37] [23] or filtering [22]. One of the non-evaluated task refers to tweet summarization, although it is a key issue for Twitter users that are often overwhelmed by the continuous stream of information.

Many tweet summarization approaches were proposed by IR researchers [33] [2], [13], but it is difficult to compare them due to the lack of publicly available evaluation collections. Most of the approaches evaluate their performance on self-built collections that are often very small (only 3 or 4 events to summarize) and scarcely accessible. To solve the issue of inadequate evaluation collections, two solutions can be considered: building a new collection, which is very costly, or transform an existing collection of tweets for the purpose of summarization evaluation.

Our research question in this paper is to see if the existing TREC IS collection [24] could potentially be adapted for event summarization evaluation in the context of Twitter.

The rest of the paper is organized as follows. Section 2 presents related works about tweet summarization and evaluation. Section 3 describes the original collection proposed by the organizers of the TREC IS tracks. Section 4 introduces the extension we made to adapt it to summarization and validates our propositions, particularly regarding the proposed Gold Standard. We finally conclude the paper in Section 5.

2 RELATED WORK

The aim of this section is not to present tweet summarization approaches but rather to focus on (tweet) summarization evaluation. Given the Twitter context, we restrict our analysis to multi-document summarization which aims to summarize a set of documents.

2.1 To be or not to be a good summary

Building an evaluation collection for summarization requires on the one hand documents to summarize and on the other hand an "ideal" summary that will be used as gold standard. This summary should be as "ideal" as possible, since it is considered as the summary toward which summarization methods should tend.

What is an "ideal" summary? A first and very simple definition of a summary can be "a text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually, significantly less than that" [31]. To be "ideal" or "perfect", a (multi-document) summary should also meet the following requirements [11, 14]:

- coverage: the main points across documents of the collection are present in the summary;
- anti-redundancy: no redundant information can be found in the summary;

*Both authors contributed equally to this work.

[1]Text REtrieval Conference, http://trec.nist.gov

- cohesion, i.e., the ability to combine text passages in a useful manner for the reader. The authors of [11] mention some ways to guarantee cohesion, such as document ordering, topic cohesion, and time line ordering;
- coherence, i.e., relevance and readability;
- contextualization: the summary should include sufficient context to be understandable to a human being;
- no inconsistencies in the information reported from different sources (e.g., billion vs million, etc.).

These requirements are valid regardless of the type of considered documents (long or short texts, very specific texts such as medical or legal ones, etc.). Tweets are of course no exception, even if some additional requirements may be considered such as source validity [39].

2.2 Existing datasets for tweet summarization

Datasets for the evaluation of automatic summarization systems were mainly built as part of specific evaluation challenges organized during dedicated conferences, such as DUC (Document Understanding Conferences) between 2001 and 2007[2], TAC (Text Analysis Conference) between 2008 and 2015[3], and TREC with the Temporal Summarization track from 2013 to 2015[4]. During all those conferences, many summarization tasks were organized, such as multi-document, query-based, biomedical, temporal, or opinion summarization. In most cases, a Gold Standard summary was built by human assessors, either using participants runs and pooling, or directly on the source documents. Whereas these evaluation collections are known and recognized collections of the domain, similar collections with tweets as source documents are surprisingly missing.

Evaluation collections with tweets however exist: one can cite the TREC Microblog collection built between 2011 and 2015 [37] [23], the TREC RTS (Real-Time Summarization) one built between 2016 and 2018, and the TREC IS one (Incident Streams) built since 2018. The associated research tasks however do not deal with summarization, even for the RTS one, which is in fact rather a filtering task. Indeed, the aim of the RTS task is to filter the Twitter stream according to a given topic in order to push to the user phone only those which are relevant and new. The task focuses on topics that are not limited in time such as 'Women's rights in Saudi Arabia' or 'Sudoku tournaments'. Furthermore, for some topics all the incoming tweets about the topic have to be pushed since there are very few (i.e., sometimes less than one relevant tweet per day). Summarizing these tweets thus would not have any sense.

Among the rare tweet summarization collections that can be found in the literature, one can cite those listed in Table 1, line 1 to 3. Those collections either rely on very few tweets to summarize [35] or very few events [34][28]. They are moreover scarcely accessible to the research community.

2.3 Evaluation metrics

There are two ways to evaluate the performance of text summarization [1, 10]:

- The aim of extrinsic evaluation is to determine to which extent the provided summary help to other tasks such as text classification or question answering.
- The objective of intrinsic evaluation is to measure the quality and informativeness of the provided summary against a human-made summary. The collections dedicated to text summarization aim at doing such type of evaluation [9].

On the one hand, quality refers to properties such as grammatical correctness, structure, or coherence. On the other hand, informativeness is evaluated with metrics such as precision, recall, or ROUGE [20].

ROUGE was widely used for the evaluation of text summarization. It identifies the common units between the summary to evaluate and the gold standard summary. ROUGE includes five measures: ROUGE-N, ROUGE-L, ROUGE-W, ROUGE-S, and ROUGE-SU. ROUGE-1 for instance compares the unigrams of both summaries. BLEU (BiLingual Evaluation Understudy) [30] initially intended for machine translation evaluation was also applied to evaluation of summaries. Other metrics exist in the literature such as METEOR, (Metric for Evaluation of Translation with Explicit ORdering) [6], Pyramid [26], and a ROUGE-based metric integrating word embeddings [27].

2.4 Discussion

As previously introduced, existing tweet summarization collections are rare and small in terms of tweets or topics/events considered, even for the evaluation of approaches published in one of the top conferences of the domain [34].

While participating to the TREC Incident Streams track [7], which was not created for automatic tweet summarization, we noticed that some characteristics of the ground truth could be used to generate a gold standard for tweet summarization. We present this collection in the following section.

3 TREC IS ORIGINAL COLLECTION

3.1 Objectives of the TREC IS track

The TREC Incident Streams track exists since 2018. The purpose of this track is to use social information, in this case information from Twitter, to help emergency services to respond to an appearing incident, as shown in Figure 1 [24]. On the one hand, a set of tweets has to be classified according to predefined categories called Information types. "Information wanted" or "Request for Search and Rescue" are examples of Information types. On the other hand, a priority score has to be associated with each tweet in order to return alerts if necessary. An alert is raised when the priority score exceeds 0.7. Particular attention is paid to information types named "Requests for Goods/ Services", "Requests for Search and Rescue", "Calls to Action for Moving People", "Reports of Emerging Threats", "Reports of Significant Event Changes" and "Reports of Services" that are considered actionable, which means that they should lead to actions from emergency services. Two editions of the track took place, the first in 2018 and the second in 2019 that has run in two parts (one for prototyping and the other for validation).

[2] https://www-nlpir.nist.gov/projects/duc/
[3] https://tac.nist.gov/
[4] http://www.trec-ts.org/

Dataset	# of tweets	# of events	Gold Standard length
trending-topics-2010 [35]	2,500	25	4 tweets
disaster-related-2018 [34]	130,000	3	200 words
events-oct-nov-2016 [28]	60,000	6	N/A
trec-is-news-priority	35,000	26	45 tweets on average

Table 1: Some datasets for tweet summarization

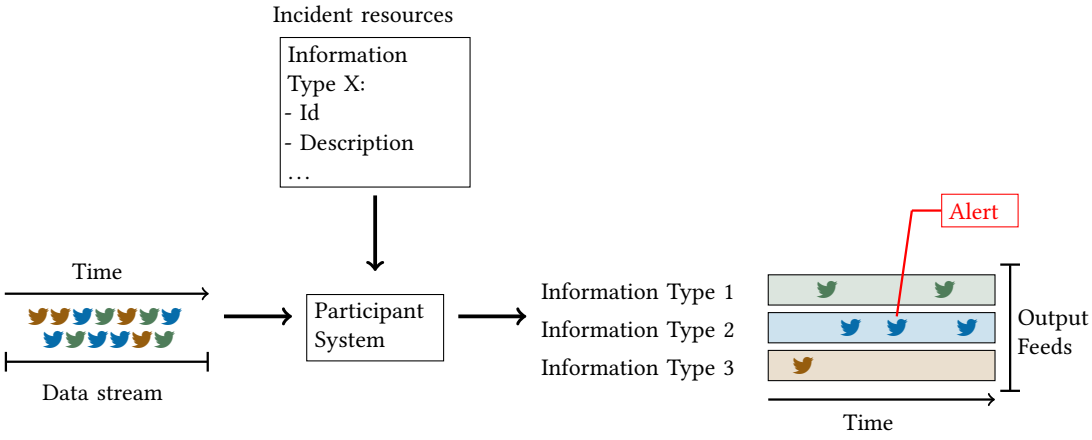

Figure 1: Incident Streams task, as shown on the task's website (http://dcs.gla.ac.uk/~richardm/TREC_IS/)

3.2 Features of the TREC IS collection

This section presents the main features of the original TREC IS collection built since 2018.

3.2.1 Collected tweets. In fact, the collection was incrementally built through three steps. On the one hand, in 2018 the tweets were collected from the CrisisLexT26 source [29]. On the other hand, several sources were used to collect the tweets for the part A and B of the 2019 edition: CrisisLexT26, CrisisNLP Resource #2 [15], CrisisNLP Resource #1 [16], additional tweets crawled by the organizers of the track themselves via Twitter API, other tweets collected via the GNIP service, and finally tweets donated by participant groups. The event related to these collected tweets refers to the following incident types: wildfire, earthquake, flood, typhoon/hurricane, shooting and bombing.

Regarding the CrisisLex and CrisisNLP sources, the organizers used the tweets labeled as relevant to the events. For the remaining events and the corresponding tweets collected by the organizers, a clustering was performed to select one tweet per cluster. Then, an additional keyword filtering led to a reduced, diverse, and likely relevant set. Finally, for all events, the tweets identified by the Twitter's language classifier as written in a language other than English ('en') were removed.

3.2.2 Tweet annotation. For the evaluation process, the organizers created a ground truth according to a set of information type labels for both the training and test sets. To create this ground truth set, a labelling interface was developed by the organizers, see Figure 2. This interface was intended to help the assessors annotating the collected tweets for each event (see section 3.2.1) with:

- information types : "Request–GoodsServices", "Request–SearchAndRescue", "Request-InformationWanted", "CallToAction-Volunteer", "CallToAction-Donations", "CallToAction-MovePeople", "Report-FirstPartyObservation", "Report-ThirdPartyObservation", "Report-Weather", "Report-EmergingThreats", "Report-MultimediaShare", "Report-ServiceAvailable", "Report-Factoid", "Report-Official", 'Report-CleanUp", "Report-Hashtags", 'Report-News", "Report-NewSubEvent", "Report-Location", "Other-Advice", "Other-Sentiment", "Other-Discussion", "Other-ContextualInformation", "Other-Irrelevant", "Other-OriginalEvent".
  Tweets can be multi-labeled. In particular, the information type "Report-News" is the label corresponding to the posts providing/linking to continuous coverage of the event.
- priority levels : "low", "medium", "high", "critical" which correspond respectively to the scores 0.25, 0.5, 0.75, and 1.

The tweets were labeled by several TREC assessors across three assessment periods. Due to the volume of the tweets collected for each event, the assessments were divided among multiple assessors.

3.2.3 Collection statistics. In the end, the merged TREC IS collection totalizes 35,266 assessed tweets according to 33 distinct events. Initially, the organizers mentioned 34 events but only 33 were really assessed. Table 2 reports the considered events and some statistics about the dataset.

Between 96 and 5,863 tweets were labeled for an event with an average of 1068 tweets per event. Regarding tweet size, two distinct

[5]For this event, one tweet was labeled as "Unknown"

Event	# of tweets	# of tweets News	# of tweets Low	# of tweets Medium	# of tweets High	# of tweets Critical	# of tweets in the CGS summary
costaRicaEarthquake2012	247	23	205	22	19	1	1 (0.4 %)
fireColorado2012[5]	263	61	146	82	34	0	0 (0 %)
floodColorado2013	235	41	107	69	56	3	0 (0 %)
typhoonPablo2012	244	39	137	73	34	0	0 (0 %)
laAirportShooting2013	162	60	110	26	24	2	0 (0 %)
westTexasExplosion2013	184	27	110	52	19	3	0 (0 %)
guatemalaEarthquake2012	154	121	46	76	24	8	25 (16.2 %)
bostonBombings2013	535	148	380	86	60	9	58 (10.8 %)
flSchoolShooting2018	1118	106	774	276	62	6	32 (2.9 %)
chileEarthquake2014	311	202	263	35	12	1	12 (3.9 %)
joplinTornado2011	96	58	39	38	12	7	17 (17.7 %)
typhoonYolanda2013	564	401	298	215	48	3	50 (8.9 %)
queenslandFloods2013	713	364	389	178	131	15	136 (19.1 %)
nepalEarthquake2015	5863	836	3950	959	938	16	176 (3.0 %)
australiaBushfire2013	677	435	518	123	34	2	20 (3.0 %)
philipinnesFloods2012	437	84	289	50	96	2	1 (0.2 %)
albertaFloods2013	722	353	651	39	32	0	18 (2.5 %)
typhoonHagupit2014	3941	1277	3086	290	545	20	193 (4.9 %)
manilaFloods2013	411	39	330	38	33	10	6 (1.5 %)
parisAttacks2015	2066	276	1760	182	83	41	80 (3.9 %)
italyEarthquakes2012	103	29	79	6	15	3	17 (16.5 %)
floodChoco2019	389	14	385	4	0	0	0 (0 %)
earthquakeCalifornia2014	127	68	127	0	0	0	0 (0 %)
shootingDallas2017	2000	232	1945	33	18	4	17 (0.9 %)
earthquakeBohol2013	583	125	578	1	4	0	2 (0.3 %)
fireYMM2016	2000	688	1543	218	167	72	134 (6.7 %)
hurricaneFlorence2018	1999	293	1703	216	66	14	36 (1.8 %)
philippinesEarthquake2019	1994	483	1479	312	197	6	94 (4.7 %)
southAfricaFloods2019	1347	408	384	716	241	6	80 (5.9 %)
cycloneKenneth2019	1998	352	1320	386	275	17	86 (4.3 %)
albertaWildfires2019	2000	721	1417	383	152	48	117 (5.9 %)
coloradoStemShooting2019	1147	333	825	38	227	57	246 (21.4 %)
sandiegoSynagogueShooting2019	636	516	220	361	44	11	39 (6.1 %)

Table 2: Some statistics about the TREC IS dataset. For each event, the table reports the number of tweets labeled as "Report-news" as information type (column 3), and the number of tweets per priority level (columns 4 to 7). The last column indicates the number of tweets in the Candidate Gold Standard summaries. The percentage in parenthesis corresponds to the number of tweets kept from the event (column 2).

groups can be observed, the tweets before 2018 and those after 2018. From 2018 tweets have a maximum size around 280 characters while it is around 140 before, which corresponds to the increase of the maximum tweet size decided by Twitter on 7 November 2017. However, the events flschoolshooting2018 and floodchoco2019 have surprisingly a maximum size up to 140 characters although the limit has been already extended to 280 by Twitter for this period.

4 PROPOSED DATASET FOR TWEET SUMMARIZATION EVALUATION

4.1 First intuitions

By analyzing the way the collection was annotated, we had the following intuitions that we discuss in more detail in the following sections:

- tweets labeled as "Report-News" bring novel information and are thus a priori by definition non-redundant between them,
- these tweets, when associated to a "High" or "Critical" priority level are also essential to cover the event.

These two properties meet two of the aforementioned requirements of what is an "ideal" summary, i.e., anti-redundancy and

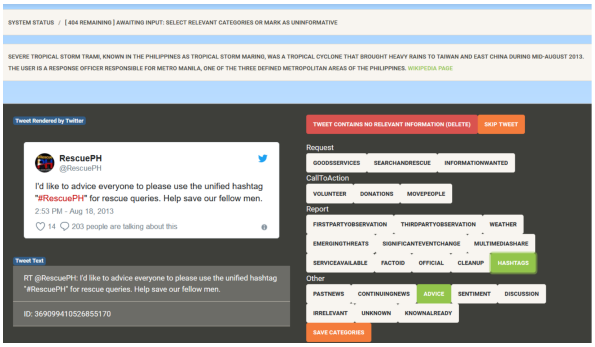

Figure 2: TREC-IS Task Assessment Interface [24]

coverage. For each event, the summary composed of tweets both labeled "Report-News" as information type and "High" or "Critical" as priority can thus be considered as a candidate gold standard, named CGS summary in the remaining of the paper. It is important to notice that all CGSs we consider are chronologically ordered according to the tweet creation dates.

Some statistics about the CGSs are reported in the last column of Table 2. Notice that the events named earthquakeCalifornia2014, fireColorado2012, floodChoco2019, floodColorado2013, laAirportShooting2013, typhoonPablo2012, westTexasExplosion2013 have empty CGSs. They are thus skipped in the remaining of the paper.

On this basis, we decided to further investigate the possibilities of the candidate gold standard to fulfill all the requirements of an ideal summary.

4.2 Analysis of the candidate gold standard summaries

4.2.1 Anti-redundancy. As reported in [24], an event could have been annotated by several assessors on disjointed sets of tweets. Some redundancy could thus remain in some events. To deal with those cases we automatically remove redundant tweets from the candidate gold standard (CGS) as follows. We took all the tweets per event and according to their creation timestamp. For each considered tweet, we evaluated the ROUGE-2 metric between this tweet and all the preceding ones. All tweets getting a ROUGE-2 score greater or equal to 0.3 were removed from the CGS summaries. We set the threshold to 0.3 according to our experiments. In total 349 tweets were removed (see Table 3, second column). One can find below some examples of redundant tweets automatically detected:

- cases of correct redundancy removals
  - exactly the same tweet because of retweet
    Tweet 1: 'RT : Italy earthquake: modern buildings, not ancient ones, pose biggest threat'
    Tweet 2: 'Italy earthquake: modern buildings, not ancient ones, pose biggest threat (video) -'
  - same information but differently expressed
    Tweet 1: 'RT : Just released by the FBI: Suspect 1 and Suspect 2 in the Boston Marathon bombing'
    Tweet 2: 'Photo of Suspects 1 and 2 in the Boston Marathon Bombings. #boston #FBI #help'
- cases of questionable removals
  - case of update of the event information using pattern
    Tweet 1: 'RT : Revised (7.5 -> 7.4): 7.4 earthquake, 24km S of Champerico, Guatemala. Nov 7 10:35 at epicenter (20m ago, depth 42 ...'
    Tweet 2: 'RT : Revised (6.6 -> 6.2): 6.2 earthquake, 24km WSW of Champerico, Guatemala. Nov 11 16:15 at epicenter (14m ago, depth ...'
  - reference to the event oversizing the redundancy (11 tweets only in this case)
    Tweet 1: 'Google exec dies in Mt. Everest avalanche after #NepalEarthquake'
    Tweet 2: 'Mt. Everest avalanche survivor speaks.'

We then manually checked the filtered CGSs using an assessment tool we specifically developed. A screen of this tool interface is depicted on Figure 3. For a given event, tweets are presented in chronological order in the area numbered 1. For each tweet, the contained information is annotated as New (area 2) or Already known (area 3). Tweets labeled as New are added in the area numbered 4 to be used as annotator memory and help him/her to label next tweets. Each labeling action can be undone thanks to the button in area 5. After this phase, 174 tweets were skipped from the summaries (see Table 3, third column).

At the end of the redundancy-check phase, 1,170 tweets were kept to form the "new" CGS summaries, simply referred as CGS in the remainder of the paper. The smallest CGS is now composed of one tweet, while the biggest contains 161 tweets.

4.2.2 Coverage. Regarding coverage, we hypothesized that it is intrinsic to the original TREC IS collection construction. Indeed, important news (i.e., high or critical ones) from the original dataset are kept to build our CGS summaries.

To support this hypothesis we analysed the CGS with respect to the sub-events associated to our events reported in the Wikipedia portal Current_events[6]. This portal daily reports world-wide current events.

For each event, we extracted from the portal the sub-events corresponding to the period covered by the original TREC IS collection. Each event of the TREC IS collection corresponds at least to one sub-event in the portal (see Table 4 column 2). The total number of sub-events is 154 with 6 sub-events per event on average.

Here is an example of the extracted sub-events for the event albertaFloods2013.

- June 20, 2013 (Thursday): Extensive flooding begins throughout southern Alberta, Canada, leading to the evacuation of more than 100,000 people, notably in the City of Calgary and Town of High River. It would become the costliest natural disaster in Canadian history.
- June 21, 2013 (Friday): 75,000 people are evacuated from their homes during flooding in Calgary, Alberta, Canada. (CNN)
- June 22, 2013 (Saturday): 100,000 residents are displaced on the third day of flooding in Alberta. (CBC)

In an additional step, we compared the sub-events represented in the CGS with the sub-events extracted from the Wikipedia portal. To do so, we made another assessment tool (see Figure 3). With

[6]https://en.wikipedia.org/wiki/Portal:Current_events

this tool, for each pair (tweet of the CGS, sub-event) we determined if the tweet is related to either the whole sub-event, a part of it, or not at all. Note that a tweet can thus be associated to multiple sub-events. In total, we annotated 6,560 pairs as 'not at all' (92 %), 543 pairs as 'partly' (about 7 %), and 29 pairs as 'whole' (less than 1 %). 68 % of the CGSs tweets do not refer to any sub-events of the Wiki portal.

Among the 154 sub-events appearing in the Wikipedia portal, 103 are represented in the CGS. The 51 sub-events for which there are no tweet references are spread over 10 of the 26 events, which represents one third of the sub-events and affects 38 % of the events. The distribution of the number of sub-events per event in the Wiki portal and in the CGS is presented in Table 4, third column.

A limitation of our proposed dataset is highlighted here: it does not cover all sub-events reported in the Wikipedia portal for each event. One of the reasons can be that the Wiki events portal covers collateral consequences of the main events, such as cancellation of sporting fixtures. These consequences, although probably present in the original IS collection, are not labelled as High or Critical by assessors and are thus not appearing in our CGSs.

4.2.3 Cohesion. The cohesion criterion is defined by [11] as the way to organize the different passages of text (here the different tweets) in a way that is efficient for the reader. Cohesion can be expressed, among others, in terms of topic-cohesion or time line ordering.

First, we evaluated the time line cohesion of our CGSs by comparing their chronological ordering with the chronological order of the sub-events in the Wikipedia portal. The comparison results showed that the chronological order of the tweets does not strictly follow the chronological appearance of the sub-events. Indeed, some tweets referring to a previous sub-event can continue to be written after other tweets referring to a new sub-event.

Second, the manual checking for coverage presented in section 4.2.2 allowed to shed some light on topic-cohesion. As previously indicated, 68 % of the CGS tweets do not refer to a sub-event reported in the Wikipedia portal. Consequently, topic-cohesion is guaranteed only for the subset comprising the 32 % of the CGS tweets that refer to Wikipedia portal sub-events.

To sum up we provide a subset of tweets of the CGS that respect the cohesion criteria.

4.2.4 Coherence. The coherence criterion is defined by [11] as the ability for a summary to be relevant and readable to the reader. On the one hand, the relevance is induced by the annotations of the TREC IS task. On the other hand, we evaluated readability with state-of-the-art metrics. These metrics are Flesch-Kincaid [17], Gunning-Fog [12], Coleman-Liau [4], Dale-Chall [5], Automatic readability index (Ari) [36], Linsear write [18] and Spache [38]. The features used by these metrics focus on number of words, number of characters, number of syllables, number of difficult words. We used the U.S. grade score obtained by these metrics with the library py-readability-metrics[7] in our readability evaluation. We performed the evaluation of the entire dataset and the CGS for each event whose CGS contains more than 100 words, which affect the events earthquakeBohol2013, costaRicaEarthquake2012, philipinnesFloods2012, manilaFloods2013 that do not contain 100 words. Table 4 presents results for the Flesch-Kincaid, Coleman-Liau, and Dale-Chall metrics. The results for the Gunning-Fog, Ari, Linsear write, and Spache metrics were highly correlated to Flesch-Kincaid (Pearson correlations, $r > 0.9$, $p - value < 0.01$).

We can observe that the readability scores are, except for the event italyEarthquake2012, between the 7th grade and the college graduate grade (college grade is greater than 12 and college graduate grade is greater than college grade). We can also notice that readability scores are weaker (i.e., higher values) for CGS than for all tweets per event in most cases (80 %). However, except for the Coleman-Liau metric, the difference is not significant (Student t-test, without the outlier italyEarthquake, $p - value < 0.05$). The Coleman-Liau measure takes into account the number of words in sentences and the number of characters in words. Indeed, the average length of words in terms of characters is 4.3 for the entire TREC IS dataset and 4.8 for the CGS, and the average length of sentences in terms of words is 11.6 for the entire TREC IS dataset and 13.8 for the CGS. We think that these reserved results are due to the technical aspect of the tweets labeled News and High Priority.

4.2.5 Context and inconsistencies. We assume that the title of an event is sufficient to understand the context. We also assume that the inconsistencies were verified by the TREC IS assessors.

4.3 Experiments on state-of-the-art approaches

We tested several state-of-the-art methods for summarizing in order to verify the following hypothesis. If the CGS represents good gold standard summaries then the summaries generated by the state-of-the-art methods should get better evaluation scores than random summaries. To do so, we generated 50 random summaries per event from the entire TREC IS dataset. All the random summaries comprised the same number of tweets as the CGS for each event. Then, we used the average number of terms of these random summaries and of the CGS to generate summaries by applying the state-of-the-art methods from the entire TREC IS dataset. The tested methods are Centroid-based [32], ClusterCMRW [40], Coverage, LEAD, LexPageRank [8], Submodular 1 [19] and 2 [21], as well as Textrank [25]. We used the available PKUSUMSUM implementation[8].

We evaluated the summaries (i.e., the random summaries and the summaries generated by the state-of-the-art methods) according to ROUGE metrics. The ROUGE metrics compute the overlap between the gold standard and the evaluated summary in terms of n-grams. We used the ROUGE-1 score based on the unigram overlap, the ROUGE-2 score based on the bigram overlap, and the ROUGE-SU score relying on the overlap of both unigrams and skip-bigrams. With regard to the results presented in Table 5 there is no significant ROUGE score differences between the state-of-the-art methods and the random summaries. This can be due to the fact that the state-of-the-art methods are not particularly dedicated to tweet summarization. Further experiments with tweet summarization methods are needed, but the preceding experiments can be seen as a good preliminary step.

[7] https://pypi.org/project/py-readability-metrics/

[8] https://github.com/PKULCWM/PKUSUMSUM

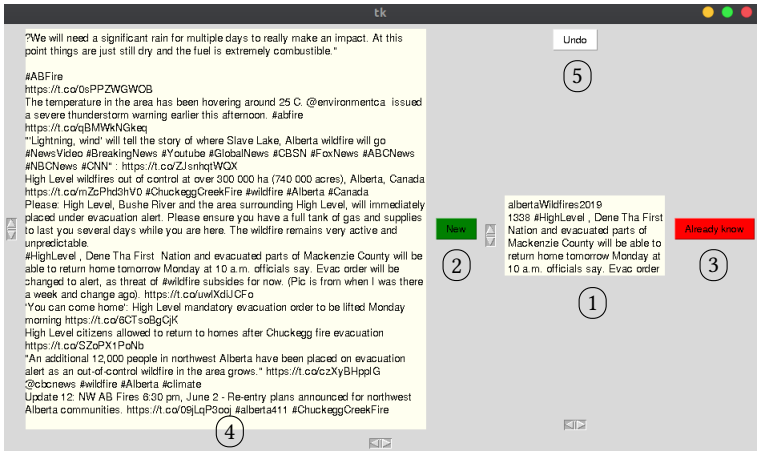

Figure 3: Redundancy Tool Interface

Event	# of tweets after automatic redundancy removal	# of tweets after manually redundancy removal	min. # of characters*	max. # of characters*	min. # of words**	max. # of words**
joplinTornado2011	17(100 %)	17(100 %)	46	136	8	30
costaRicaEarthquake2012	1(100 %)	1(100 %)	111	111	19	19
guatemalaEarthquake2012	19(76%)	15(78.9%)	44	130	7	32
philipinnesFloods2012	1(100 %)	1(100 %)	88	88	17	17
italyEarthquakes2012	16(94.1 %)	13(81.3 %)	43	115	8	22
bostonBombings2013	54(93.1 %)	51(94.4 %)	36	136	9	30
typhoonYolanda2013	47(94 %)	43(91.5 %)	32	133	6	25
australiaBushfire2013	20(100 %)	20(100 %)	65	133	11	28
queenslandFloods2013	127(93.4 %)	117(92.1 %)	22	139	5	29
earthquakeBohol2013	2(100 %)	2(100 %)	111	115	20	22
albertaFloods2013	18(100 %)	18(100 %)	51	132	11	28
manilaFloods2013	6(100 %)	6(100 %)	57	127	12	23
chileEarthquake2014	12(100 %)	7(58.3 %)	37	114	5	21
typhoonHagupit2014	168(87 %)	159(94.6 %)	15	139	4	31
parisAttacks2015	75(93.8 %)	52(69.3 %)	42	129	10	27
nepalEarthquake2015	169(97.7 %)	161(95.3 %)	5	140	1	33
fireYMM2016	129(96.3 %)	111(86.0 %)	36	140	8	33
shootingDallas2017	17(100 %)	17(100 %)	47	140	8	30
flSchoolShooting2018	29(90.7 %)	22(75.9 %)	63	135	7	26
hurricaneFlorence2018	36(100 %)	36(100 %)	53	274	9	56
philippinesEarthquake2019	87(92.6 %)	72(58.6 %)	43	277	6	51
southAfricaFloods2019	60(75 %)	50(83.3 %)	43	241	7	47
cycloneKenneth2019	81(94.2 %)	62(74.1 %)	48	277	7	60
albertaWildfires2019	104(88.9 %)	90(86.5 %)	36	280	6	59
coloradoStemShooting2019	29(11.8 %)	15(51.7 %)	40	252	7	51
sandiegoSynagogueShooting2019	20(51.3 %)	12(60.0 %)	49	157	8	31
Mean	51.69	45.0	48.58	161.15	9.69	33.12
Std	49.95	46.11	23.84	59.67	4.27	12.35

Table 3: Some statistics about the CGS. The percentage in parentheses in the second (resp. third) column corresponds to the number of tweets kept from the CGS (see Table 2), (resp. from the summary of the second column). The columns 4 to 7 show some additional statistics about the tweets in the final CGS (third column). These statistics are evaluated after URLs, useless spaces, HTML characters removal.

Event	Coverage		Readability					
	# of sub-events in Wiki::Current_events	# of sub-events in the CGS	Flesch-Kincaid		Coleman-Liau		Dale-Chall	
			A	GS	A	GS	A	GS
albertaFloods2013	3	3	8.6	8.6	10.7	10.7	11.0	10.9
albertaWildfires2019	2	2	8.8	11.1	10.6	11.6	9.7	10.9
australiaBushfire2013	4	4	10.5	11.8	11.6	14.0	11.1	12.2
bostonBombings2013	24	9	11.5	12.6	11.3	12.3	11.6	12.9
chileEarthquake2014	3	2	-	-	-	-	-	-
coloradoStemShooting2019	1	1	10.8	14.6	11.2	13.4	10.0	10.0
costaRicaEarthquake2012	1	1	-	-	-	-	-	-
cycloneKenneth2019	1	1	11.8	12.7	12.9	13.7	10.8	10.7
earthquakeBohol2013	2	0	-	-	-	-	-	-
fireYMM2016	2	2	6.9	8.1	10.4	11.4	9.9	10.7
flSchoolShooting2018	4	1	15.8	11.4	10.5	12.2	12.0	10.9
guatemalaEarthquake2012	1	1	12.9	14.1	12.0	12.7	14.0	13.7
hurricaneFlorence2018	12	8	8.1	8.9	10.2	12.8	9.6	10.2
italyEarthquakes2012	3	3	15.1	33.4	11.7	13.6	13.9	16.3
joplinTornado2011	20	15	9.5	8.1	9.9	10.3	11.9	11.6
manilaFloods2013	2	2	-	-	-	-	-	-
nepalEarthquake2015	7	7	9.7	13.6	11.1	14.3	11.1	12.4
parisAttacks2015	10	7	15.8	19.1	12.5	13.9	12.2	13.1
philipinnesFloods2012	8	0	-	-	-	-	-	-
philippinesEarthquake2019	5	5	8.4	13.1	10.7	14.3	10.9	12.0
queenslandFloods2013	4	4	10.8	11.5	11.7	12.8	11.2	12.1
sandiegoSynagogueShooting2019	2	2	15.8	16.1	13.0	13.4	11.7	10.7
shootingDallas2017	2	1	7.5	10.0	9.9	14.1	9.2	11.3
southAfricaFloods2019	1	1	11.9	14.7	11.4	14.3	10.3	11.1
typhoonHagupit2014	4	4	11.5	12.5	13.2	12.7	12.2	13.1
typhoonYolanda2013	26	17	11.9	16.5	12.2	13.4	12.3	13.4
Mean	5.92	3.96	11.1	13.5	11.4	12.9	11.3	11.9
Std	6.91	4.24	2.7	5.3	1.0	1.1	1.3	1.4

Table 4: On the left side, the coverage statistics for each event with the number of sub-events in the Wiki portal and the number of these sub-events present in the CGS. On the right side, the readability scores for the different measures regarding the entire dataset (noted A) and CGS (noted GS) for each event.

Summary	ROUGE-2	ROUGE-1	ROUGE-SU
centroid	**0.20926**	0.4626	**0.22939**
clustercmrw	**0.20901**	0.45821	**0.23681**
coverage	0.18985	**0.4639**	0.22641
lead	0.19352	**0.46848**	**0.22988**
lexpagerank	**0.21217**	**0.46685**	**0.22988**
submodular1	**0.20046**	**0.46585**	**0.24169**
submodular2	0.17876	0.43346	0.21173
textrank	**0.22771**	**0.46512**	**0.23522**
Mean random summaries	0.1937	0.4632	0.2283
Std random summaries	0.009	0.009	0.007

Table 5: ROUGE scores for each summary considering CGSs as gold standards, using the t-test there is no significant difference between methods and random summaries.

5 CONCLUSION

In this paper, we presented the benefits of the TREC IS dataset for the task of tweet summarization. We considered a particular subset of this dataset for which we produced additional annotations. These annotations, as well as the code and the results, are available in our Github[9]. We showed that a subset of this dataset can be used as gold standard for this task. We studied this subset with respect to the properties a good summary has to reach. We showed that all these properties are satisfied for tweet summarization, although some limits of this dataset can been found, specially regarding cohesion or readability. The well-known document summarization methods showed limits to outperform random summaries. These limits might be explained by the fact that the TREC IS dataset is oriented towards event reporting. Future work includes a study of more specific summarization methods.

[9]https://github.com/AlexisDusart/SetSummTweet

REFERENCES

[1] Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., and Kochut, K. (2017). Text summarization techniques: a brief survey. In *arXiv preprint arXiv:1707.02268*.

[2] Belkaroui, R. and Faiz, R. (2017). Conversational based method for tweet contextualization. *Vietnam J. Computer Science*, 4(4):223–232.

[3] Bellot, P., Moriceau, V., Mothe, J., SanJuan, E., and Tannier, X. (2016). Inex tweet contextualization task: Evaluation, results and lesson learned. *Information Processing & Management*, 52(5):801 – 819.

[4] Coleman, M. and Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60:283–284.

[5] Dale, E. and Chall, J. S. (1948). A formula for predicting readability. *Educational Research Bulletin*, 27(1):11–28.

[6] Denkowski, M. and Lavie, A. (2011). Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 85–91, Edinburgh, Scotland. Association for Computational Linguistics.

[7] Dusart, A., Hubert, G., and Pinel-Sauvagnat, K. (2019). Irit at trec 2019: Incident streams and complex answer retrieval tracks. In *Proceedings of the TREC 2019 Conference*.

[8] Erkan, G. and Radev, D. R. (2004). Lexpagerank: Prestige in multi-document text summarization. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing , EMNLP 2004, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004, 25-26 July 2004, Barcelona, Spain*, pages 365–371. ACL.

[9] Ermakova, L., Cossu, J., and Mothe, J. (2019). A survey on evaluation of summarization methods. *Inf. Process. Manage.*, 56(5):1794–1814.

[10] Gambhir, M., G. V. (2017). Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, 47:1–66.

[11] Goldstein, J., Mittal, V., Carbonell, J., and Kantrowitz, M. (2000). Multi-document summarization by sentence extraction. In *Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic Summarization - Volume 4*, NAACL-ANLP-AutoSum '00, page 40–48, USA. Association for Computational Linguistics.

[12] Gunning, R. (1968). *The technique of clear writing*. McGraw-Hill.

[13] He, R., Liu, Y., Yu, G., Tang, J., Hu, Q., and Dang, J. (2017). Twitter summarization with social-temporal context. *World Wide Web*, 20(2):267–290.

[14] Huang, L., He, Y., Wei, F., and Li, W. (2010). Modeling document summarization as multi-objective optimization. In *Third International Symposium on Intelligent Information Technology and Security Informatics, IITSI 2010, Jinggangshan, China, April 2-4, 2010*, pages 382–386. IEEE Computer Society.

[15] Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., and Meier, P. (2013). Extracting information nuggets from disaster- related messages in social media. In Comes, T., Fiedrich, F., Fortier, S., Geldermann, J., and Müller, T., editors, *10th Proceedings of the International Conference on Information Systems for Crisis Response and Management, Baden-Baden, Germany, May 12-15, 2013*. ISCRAM Association.

[16] Imran, M., Mitra, P., and Castillo, C. (2016). Twitter as a lifeline: Human-annotated twitter corpora for NLP of crisis-related messages. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).

[17] Kincaid, J. (1975). *Derivation of New Readability Formulas: (automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*. Research Branch report. Chief of Naval Technical Training, Naval Air Station Memphis.

[18] Klare, G. R. (1974). Assessing readability. *Reading Research Quarterly*, 10(1):62–102.

[19] Li, J., Li, L., and Li, T. (2012). Multi-document summarization via submodularity. *Appl. Intell.*, 37(3):420–430.

[20] Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

[21] Lin, H. and Bilmes, J. A. (2010). Multi-document summarization via budgeted maximization of submodular functions. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 2-4, 2010, Los Angeles, California, USA*, pages 912–920. The Association for Computational Linguistics.

[22] Lin, J., Mohammed, S., Sequiera, R., Tan, L., Ghelani, N., Abualsaud, M., McCreadie, R., Milajevs, D., and Voorhees, E. M. (2017). Overview of the TREC 2017 real-time summarization track. In Voorhees, E. M. and Ellis, A., editors, *Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA, November 15-17, 2017*, volume Special Publication 500-324. National Institute of Standards and Technology (NIST).

[23] Lin, J. J. and Efron, M. (2013). Overview of the TREC-2013 microblog track. In Voorhees, E. M., editor, *Proceedings of The Twenty-Second Text REtrieval Conference, TREC 2013, Gaithersburg, Maryland, USA, November 19-22, 2013*, volume Special Publication 500-302. National Institute of Standards and Technology (NIST).

[24] Mccreadie, R., Buntain, C., and Soboroff, I. (2019). Trec incident streams: Finding actionable information on social media.

[25] Mihalcea, R. and Tarau, P. (2004). Textrank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing , EMNLP 2004, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004, 25-26 July 2004, Barcelona, Spain*, pages 404–411. ACL.

[26] Nenkova, A., Passonneau, R., and McKeown, K. (2007). The pyramid method: Incorporating human content selection variation in summarization evaluation. *TSLP*, 4.

[27] Ng, J.-P. and Abrecht, V. (2015). Better summarization evaluation with word embeddings for ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1925–1930, Lisbon, Portugal. Association for Computational Linguistics.

[28] Nguyen, M., Viet, L. D., Nguyen, H., and Nguyen, M. (2018). Tsix: A human-involved-creation dataset for tweet summarization. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

[29] Olteanu, A., Vieweg, S., and Castillo, C. (2015). What to expect when the unexpected happens: Social media communications across crises. In Cosley, D., Forte, A., Ciolfi, L., and McDonald, D., editors, *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, CSCW 2015, Vancouver, BC, Canada, March 14 - 18, 2015*, pages 994–1009. ACM.

[30] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

[31] Radev, D. R., Hovy, E., and McKeown, K. (2002). Introduction to the special issue on summarization. *Computational linguistics*, 28(4):399–408.

[32] Radev, D. R., Jing, H., Sty, M., and Tam, D. (2004). Centroid-based summarization of multiple documents. *Inf. Process. Manag.*, 40(6):919–938.

[33] Rudra, K., Goyal, P., Ganguly, N., Imran, M., and Mitra, P. (2019). Summarizing situational tweets in crisis scenarios: An extractive-abstractive approach. *IEEE Trans. Comput. Social Systems*, 6(5):981–993.

[34] Rudra, K., Goyal, P., Ganguly, N., Mitra, P., and Imran, M. (2018). Identifying sub-events and summarizing disaster-related information from microblogs. In Collins-Thompson, K., Mei, Q., Davison, B. D., Liu, Y., and Yilmaz, E., editors, *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pages 265–274. ACM.

[35] Sharifi, B., Inouye, D. I., and Kalita, J. K. (2014). Summarization of twitter microblogs. *Comput. J.*, 57(3):378–402.

[36] Smith, E., Senter, R., and (U.S.), A. F. A. M. R. L. (1967). *Automated Readability Index*. AMRL-TR. Aerospace Medical Research Laboratories.

[37] Soboroff, I., Ounis, I., Macdonald, C., and Lin, J. J. (2012). Overview of the TREC-2012 microblog track. In Voorhees, E. M. and Buckland, L. P., editors, *Proceedings of The Twenty-First Text REtrieval Conference, TREC 2012, Gaithersburg, Maryland, USA, November 6-9, 2012*, volume Special Publication 500-298. National Institute of Standards and Technology (NIST).

[38] Spache, G. (1953). A new readability formula for primary-grade reading materials. *The Elementary School Journal*, 53(7):410–413.

[39] Vosoughi, S., Mohsenvand, M. N., and Roy, D. (2017). Rumor gauge: Predicting the veracity of rumors on twitter. *ACM transactions on knowledge discovery from data (TKDD)*, 11(4):1–36.

[40] Wan, X. and Yang, J. (2008). Multi-document summarization using cluster-based link analysis. In Myaeng, S., Oard, D. W., Sebastiani, F., Chua, T., and Leong, M., editors, *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, Singapore, July 20-24, 2008*, pages 299–306. ACM.