Gold Standard Creation for Microblog Retrieval: Challenges of Completeness in IRMiDis 2017

Ribhav Soni*
Indian Institute of Technology (Banaras Hindu University), Varanasi
India
ribhav.soni.cse13@iitbhu.ac.in

Sukomal Pal
Indian Institute of Technology (Banaras Hindu University), Varanasi
India
spal.cse@iitbhu.ac.in

ABSTRACT

Microblogging sites like Twitter, Facebook, etc., are important sources of first-hand accounts during disaster situations, and have the potential to significantly aid disaster relief efforts. The IRMiDis track at FIRE 2017 focused on developing and comparing IR approaches to automatically identify and match tweets that indicate the need or availability of a resource, leading to the creation of a benchmark dataset for future improvements in this task. However, based on our experiments, we argue that the gold standard data obtained in the track is substantially incomplete. We also discuss some reasons why it may have been so, and provide some suggestions for making more robust ground truth data in such tasks.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; Information systems applications; World Wide Web;

KEYWORDS

Crisis Informatics, Disaster, Microblog Retrieval, Social Media, Gold Standard, Language Models, Word Embeddings, word2vec, GloVe, WordNet, Query Expansion, Relevance Feedback

ACM Reference Format:
Ribhav Soni and Sukomal Pal. 2018. Gold Standard Creation for Microblog Retrieval: Challenges of Completeness in IRMiDis 2017. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3184558.3191622

1 INTRODUCTION

During times of disasters, a huge volume of disaster-related information is posted by users on microblogging sites (like Twitter, Facebook, etc.), which includes first-hand accounts of the situation that can help immensely in knowing about the on-ground situation, and thus in aiding disaster relief efforts.

The Information Retrieval from Microblogs during Disasters (IRMiDis) track [1] in FIRE 2017 [1] in particular focused on developing and comparing automated IR approaches to identify and match "need tweets" and "availability tweets" among a collection of about 46,000 tweets posted during the Nepal earthquake in 2015,[2] where they were defined as:

- Need-tweet: A tweet that indicates the scarcity or requirement of some specific resource, such as food, water, medical aid, shelter, etc.
- Availability-tweet: A tweet that indicates the future or actual availability of some specific resource.

The track consisted of two sub-tasks, with sub-task 1 focused on identifying need tweets and availability tweets separately, from the given collection. Sub-task 2 was to match need-tweets with corresponding availability tweets that could satisfy the need of at least one resource mentioned in the need-tweet.

The track led to the creation of a benchmark dataset for research in IR approaches to identify and match need and availability tweets posted during a disaster situation. However, based on our experiments, we argue that the gold standard creation for sub-task 2 (i.e., the list of correct pairs of need-availability tweets such that at least one resource mentioned in the need tweet is satisfied by the availability tweet, as identified by the human annotators) is substantially incomplete.

The remainder of this paper is organised as follows. We describe the dataset used in the IRMiDis track in Section 2, the gold standard creation method used in Section 3, our experiments and observations in Section 4, some discussion in Section 5, and conclusion, along with some future directions of work, in Section 6.

2 TWEETS DATASET

The organizers had collected a set of about 66k tweets posted during the Nepal earthquake in 2015, which included tweets in English, Nepali, Hindi, etc., as well as code-mixed tweets (i.e., a single tweet containing two or more languages or scripts). A set of 20k tweets was made available as training data, while the remaining 46k tweets were used as test data for the track.

Among the training tweets, the gold standard data (i.e., the list of all need-tweets and availability-tweets for sub-task 1, and the list of all matching need-availability tweet pairs for sub-task 2) were also provided to the participants.

3 GOLD STANDARD CREATION PROCESS ADOPTED IN IRMIDIS 2017

The track organizers employed three annotators for creating the gold standard.

*This is the corresponding author
[1] http://fire.irsi.res.in/fire/2017/home

[2] https://en.wikipedia.org/wiki/April_2015_Nepal_earthquake

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3191622

The gold standard creation for sub-task 1 involved the following three stages.

(1) First, each annotator independently searched for relevant need or availability tweets using manual runs, after the tweets were indexed.
(2) Then, the annotators mutually discussed and finalized the relevance of the tweets that at least one of them had found in stage 1.
(3) Finally, the top-100 results from each of the submitted runs were pooled, and judged by the annotators.

For sub-task 2 (i.e., matching the need and availability tweets that were separately listed in the gold standard for sub-task 1), the annotators were asked to manually match each need-tweet with the availability-tweets that could satisfy the need of at least one resource that was mentioned as lacking in the need-tweet. Additionally, pooling was applied over the submitted runs to judge the matching pairs that may have been missed by the annotators. As per the task instructions, only the top 5 matching availability-tweets were to be output for each need-tweet by a run submission, all of which by each run were taken in the pool for manual assessment by the annotators.

There were only 10 runs submitted for sub-task 2, and they included only 4 different kinds of models (others differing only in the parameters of the model).

4 EXPERIMENTS AND OBSERVATIONS

As the sub-task 1 gold standard, 211 need-tweets and 718 availability-tweets were identified by the annotators in the training data of 20k tweets. In the test data of about 46k tweets, 427 need-tweets and 980 availability-tweets were identified.

For sub-task 2, the annotators identified a total of 3091 need-availability tweet pairs from the training tweets, for 200 need tweets, i.e., an average of 15.46 availability tweets identified for each need tweet. For the test data, they found 4117 need-availability tweet pairs, with 427 need tweets, i.e., an average of only 9.64 availability tweets for each need tweet.

While there was no need-tweet in the training data for sub-task 2 for which no matching availability-tweet could be identified by the annotators, 126 out of 427 need-tweets in the test data were such that no matching availability-tweet was identified by the annotators. On manual inspection, we were easily able to identify about 10 matching availability-tweets for each of at least 10 of those 126 tweets. In addition, even for many of the remaining need-tweets for which some availability-tweets were identified by the annotators, we could easily find at least 10 more matching availability-tweets that were missed by the annotators.

Some examples of need-tweets for which the annotators failed to find a single availability-tweet, along with some availability-tweets that should have been found, are shown in Table 1.

4.1 Discovering relevant pairs from i runs

We randomly selected 20 need-tweets from the 427 need-tweets identified in the Gold Standard for the test data in sub-task 1.

We also applied a total of 8 IR methods: (1) Lucene [3] default model (which uses a variant of Tf-idf for scoring), (2) Word2vec [5] vectors pre-trained using a Google News dataset, (3) GloVe [7] vectors trained on a Twitter dataset, (4) GloVe vectors trained on a Wikipedia dataset, (5) a unigram language model, (6) a bigram language model, (7) searching using Lucene after Query Expansion using WordNet [6], (8) searching using Lucene after manual relevance feedback.

For each of the 8 models, for each of the randomly selected 20 need-tweets, we output the top five matching availability tweets among the availability tweets in the test data identified in the gold standard for sub-task 1. We manually checked the relevance of each of the resulting 8*20*5=800 pairs of tweets, and found 327 need-availability tweet pairs to be relevant (i.e., the availability-tweet mentions the availability of at least one resource mentioned as lacking in the corresponding need-tweet, according to our judgment), and only 49 of them (i.e., only about 15%) were present in the gold standard for sub-task 2 (as identified by the task's annotators). The number of relevant pairs found by each of these 8 methods, along with how many of them were present in the gold standard, are shown in Table 2.

To analyze how the number of relevant need-availability tweet pairs discovered increases with the number of participating systems, we performed experiments by taking all combinations of the 8 methods, first 1 method at a time, then 2 at a time, and so on. The average, minimum, and maximum number of distinct pairs discovered by the methods when taking i, $1 \leq i \leq 8$, methods at a time, are shown in Table 3, and plotted in Figure 1.

Since the graph has not saturated for 8 runs, extrapolating for 10 run submissions (which was the actual number of run submissions in sub-task 2), we can expect a significant increase in the number of relevant pairs discovered if more runs are added to the pool.

5 DISCUSSION

The final gold standard for sub-task 2 obtained in the track does not intuitively seem to be complete enough for a reliable ranking of systems other than those which contributed to the pool.

The pooling method failed to find many of the relevant need-availability tweet pairs. Some reasons why this happened may be:

- There were only 10 runs submitted for sub-task 2, with only 4 different kinds of models applied (others only differed from these in terms of the parameters of the model). Clearly, a pool made from such a collection is neither diverse nor large enough to include most relevant need-availability tweet pairs. This problem also occurred in the corresponding track last year [2], as reported in [9].
- The organizers had set a limit of outputting a maximum of 5 matching availability-tweets for each need-tweet by a run submission. In the Microblog Track at TREC,[4] there used to be more than 100 run submissions, and generally the pooling depth was set as 100 [3, 8]. Setting a pooling depth of only 5 here, with only 10 run submissions, caused the pool to not be reliable. In fact, there were more than 100

[3] https://lucene.apache.org/
[4] http://trec.nist.gov/data/microblog.html

Table 1: Examples of need-tweets for which no matching availability-tweets were found by the annotators, along with some availability-tweets that should have been found (the text has been translated to English wherever needed)

Need-tweets	Matching Availability-tweets
There is need for 4 lakh tents, food for 3.5 million people in Nepal.http://hindi.news-roompost.com/45248/nepal-needs-more-help/ âĂę (id:594732773539237888)	Gud job by—>@RaviNepal, has put together a list of places to getfood, water and shelter in Kathmandu. http://www.bit.ly/nepalrelief15#Earthquake (id:592751043164938240)
Nepal AFTER QUAKE: Today the need is tent, food and medicine. (id:593199294527643648)	Today our volunteers distributed food to more than 1500 earthquake affected people at #ArtofLiving center at Raxaul #NepalEarthquakeRelief (id:592757889535737860)
In the coming days, people of Nepal need basic essentials such as food, clean water and shelter. #earthquake (id:596366795931406336)	at bharatpur hospital ...really all the earthquake victims are well treated with food, clothes, mats, medicines, etc #Earthquake #Nepalquake (id:593783416333717504)

Table 2: Number of relevant tweet pairs found by each method

S. No.	Method	No. of relevant pairs found (out of 100)	No. of pairs also in GS	Percentage
1	Lucene default model	49	10	20.41 %
2	Word2vec on Google News data	41	7	17.07 %
3	GloVe on Twitter data	39	3	7.69 %
4	GloVe on Wikipedia data	44	4	9.09 %
5	Unigram LM	29	6	20.69 %
6	Bigram LM	16	1	6.25 %
7	Lucene, with QE using WordNet	36	5	13.89 %
8	Lucene, with QE using relevance feedback	73	13	17.81 %

Table 3: Number of relevant tweet pairs discovered by i systems, $1 \leq i \leq 8$

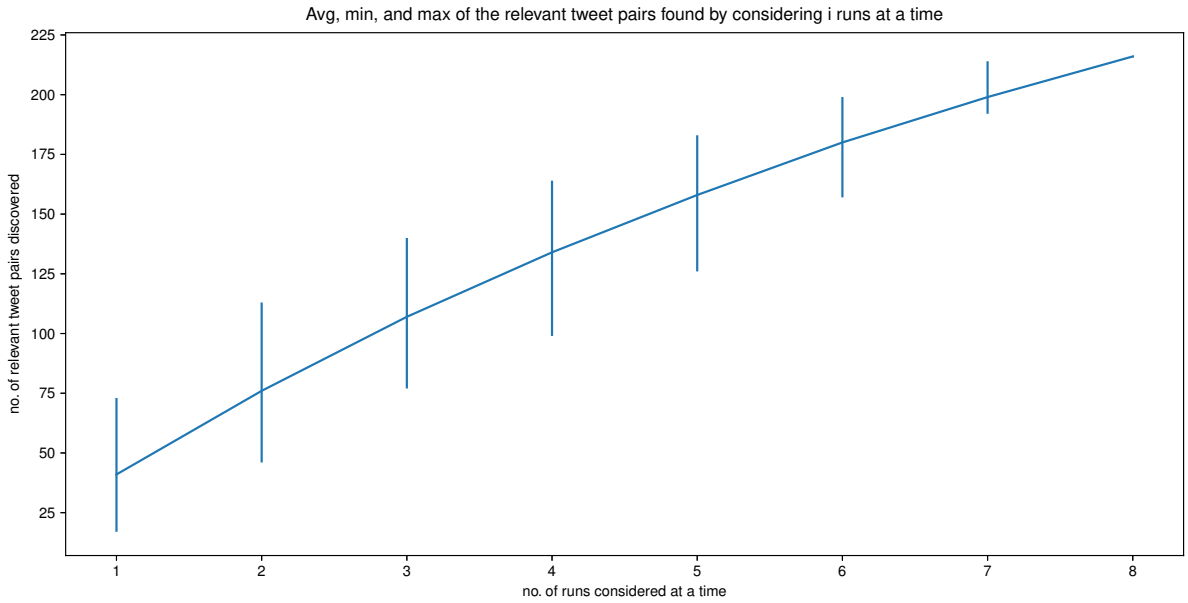
Number of methods taken at a time	Number of combinations	Number of relevant tweet pairs found		
		Average	Minimum	Maximum
1	8	41	17	73
2	28	76	46	113
3	56	107	77	140
4	70	134	99	164
5	56	158	126	183
6	28	180	157	199
7	8	199	192	214
8	1	216	216	216

tweets in the gold standard for availability-tweets in subtask 1 that included the keyword "food", indicating that there may have been about 100 or more availability-tweets that would match any need-tweet that mentions the need of food. Furthermore, the P@5 value for even the best submissions was close to 0.2, which means that, on average, it could retrieve only 1 matching availability-tweet for each need-tweet. With only 10 submissions, we can only expect 10 matching availability-tweets for each need-tweet from this pooling approach, which falls drastically short of the number of potential matches. So, the allowance for the number of matching availability-tweets for each need-tweet in a run should have been much more liberal, perhaps up to 100 tweets.

Such a gold standard cannot reliably judge the performance of new systems, which may use a model different from the participating systems that made up the pool, as evident from the results for our methods, for which only about 15% (49 out of 327) of the correct need-availability tweet pairs were present in the gold standard.

One way to make pools more robust even with a low number of participating systems is for the organizers themselves to implement about 10-15 different approaches, and add their results to the pool. Another option is to employ continuous evaluations [10], rather than working with a static collection of relevance judgments. This way, a new system that is able to output many relevant tweet pairs that are not in the gold standard would not be unfairly ranked low.

Figure 1: Variation in the no. of relevant tweet pairs found by considering different number of systems



6 CONCLUSION AND FUTURE WORK

The IRMiDis track at FIRE 2017 involved the development and comparison of many IR approaches to the important task of identifying and matching need and availability tweets during disaster situations. However, based on our experiments, we found that the gold standard data obtained in the task was substantially incomplete. We discussed some reasons for it, as well as some ways to make a more robust gold standard for similar tasks in the future.

Future directions of work include exploring the feasibility of different approaches for pooling (like [4]) to make more robust gold standards for future tasks in this domain.

REFERENCES

[1] Basu, M., Ghosh, S., Ghosh, K., and Choudhury, M. (2017). Overview of the fire 2017 track: Information retrieval from microblogs during disasters (irmidis). *Working notes of FIRE*, pages 8–10.

[2] Ghosh, S. and Ghosh, K. (2016). Overview of the fire 2016 microblog track: Information extraction from microblogs posted during disasters. In *FIRE (Working Notes)*, pages 56–61.

[3] Lin, J., Efron, M., Wang, Y., and Sherman, G. (2014). Overview of the trec-2014 microblog track. Technical report, MARYLAND UNIV COLLEGE PARK.

[4] Losada, D. E., Parapar, J., and Barreiro, A. (2017). Multi-armed bandits for adjudicating documents in pooling-based evaluation of information retrieval systems. *Information Processing & Management*, 53(5):1005–1025.

[5] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[6] Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

[7] Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

[8] Soboroff, I., Ounis, I., Macdonald, C., and Lin, J. J. (2012). Overview of the trec-2012 microblog track. In *TREC*, volume 2012, page 20.

[9] Soni, R. and Pal, S. (2017). Microblog retrieval for disaster relief: How to create ground truths? In *SMERP@ ECIR*, pages 42–51.

[10] Tonon, A., Demartini, G., and Cudré-Mauroux, P. (2015). Pooling-based continuous evaluation of information retrieval systems. *Information Retrieval Journal*, 18(5):445–472.