Extracting Humanitarian Information from Tweets

Ali Hürriyetoğlu[1] and Nelleke Oostdijk[2]

[1] Statistics Netherlands,
P.O. Box 4481, 6401 CZ Heerlen, the Netherlands
a.hurriyetoglu@cbs.nl
[2] Centre for Language Studies, Radboud University,
P.O. Box 9103, 6500 HD, Nijmegen, the Netherlands
n.oostdijk@let.ru.nl

Abstract. In this paper we describe the application of our methods to humanitarian information extraction from tweets and their performance in the scope of the SMERP 2017 Data Challenge task. Detecting and extracting the (scarce) relevant information from tweet collections as precisely, completely, and rapidly as possible is of the utmost importance during natural disasters and other emergency events. We applied a machine learning and a linguistically motivated approach. Both are designed to satisfy the information needs of an expert by allowing experts to define and find the target information. We found that the performance of this effort highly depends on the task definition and the ability to facilitate the feedback iteratively. The results of the current data challenge task demonstrate that it is realistic to expect a balanced performance across multiple metrics even under poor conditions.

Keywords: information extraction, text mining, social media analysis, machine learning, linguistic analysis, Italy earthquake

1 Introduction

This contribution describes our approach used in the text retrieval sub-track that was organized as part of the Exploitation of Social Media for Emergency Relief and Preparedness (SMERP 2017) Data Challenge Track, Task 1. In this task, participants were required to develop methodologies for extracting from a collection of microblogs (tweets) those tweets that are relevant to one or more of a given set of topics with high precision as well as high recall.[3] The extracted tweets should be ranked based on their relevance. The topics were the following: resources available (T1), resources needed (T2), damage, restoration, and casualties (T3), and rescue activities of various NGOs and government organizations (T4). With each of the topics there was a short (one sentence) description and a more elaborate description in the form of a one-paragraph narrative.

[3] See also http://computing.dcu.ie/~dganguly/smerp2017/index.html

The challenge was organized in two rounds.[4] The task in both rounds was essentially the same, but participants could benefit from the feedback they received after submitting their results for the first round. The data were provided by the organizers of the challenge and consisted of tweets about the earthquake that occurred in Central Italy in August 2016.[5] The data for the first round of the challenge were tweets posted during the first day (24 hours) after the earthquake happened, while for the second round the data set was a collection of tweets posted during the next two days (day two and three) after the earthquake occurred. The data for the second round were released after round one had been completed. All data were made available in the form of tweet IDs (52,469 and 19,751 for rounds 1 and 2 respectively), along with a Python script for downloading them by means of the Twitter API. In our case the downloaded data sets comprised 52,422 (round 1) and 19,443 tweets (round 2) respectively.[6]

For each round, we discarded tweets that (i) were not marked as English by the Twitter API; (ii) did not contain the country name Italy or any region, city, municipality, or earthquake-related place in Italy; (iii) had been posted by users that have a profile location other than Italy; this was determined by manually checking the locations that occurred most frequently; (iv) originated from an excluded user time zone; the time zones were identified manually and covered the ones that appeared to be the most common in the data sets; (v) had a country meta-field other than Italy; and (vi) had fewer than 4 tokens after normalization and basic cleaning. The filtering was applied in the order given above. After filtering the data sets consisted of 40,780 (round 1) and 17,019 (round 2) tweets.

We participated in this challenge with two completely different approaches which we developed and applied independently of each other. The first approach is a machine-learning approach implemented in the Relevancer tool [2], which offers a complete pipeline for analyzing tweet collections. The second approach is a linguistically motivated approach in which a lexicon and a set of handcrafted rules are used in order to generate search queries. As the two approaches each have their particular strengths and weaknesses and we wanted to find out whether the combination would outperform each of the underlying approaches, we also submitted a run in which we combined them.

The structure of the remainder of this paper is as follows. We first give a more elaborate description of our two separate approaches, starting with the machine learning approach in Section 2 and the rule-based approach in Section 3. Then in Section 4 we describe the combination. Next, in Section 5, the results are presented, while in Section 6 the most salient findings are discussed. Section 7 concludes this paper with a summary of the main findings and suggestions for future research.

[4] The organizers consistently referred to these as 'levels'.
[5] `https://en.wikipedia.org/wiki/August_2016_Central_Italy_earthquake`
[6] At the time of download some tweets apparently had been removed.

2 Approach 1: Identifying topics using Relevancer

The main analysis steps supported by Relevancer are: preprocessing, clustering, manual labeling of coherent clusters, and creating a classifier for labeling previously unseen data. Below we note the general characteristics of the approach and provide details of the configuration that we used for the present task.

Pre-processing The aim of the pre-processing steps is to convert the collection to more standard text without loosing any information. An expert may choose to apply all or only some of the following steps. As a result, the expert is in control of any bias that might arise due to the preprocessing.

- **RT Elimination** Tweets that start with 'RT @' or that in their meta-information have some indication that they are retweets are excluded.
- **Normalization** User names and URLs that are in a tweet text are converted to 'usrusr' and 'urlurl' respectively.
- **Text cleaning** Text parts that are auto-generated, meta-informative, and immediate repetitions of the same word(s) are removed.
- **Duplicate elimination** Tweets that after normalization have an exact equivalent in terms of tweet text are excluded from the collection.
- **Near-duplicate elimination** Leave only one of tweet from a group of tweets that has a cosine similarity above a threshold.

RT elimination, duplicate elimination, and near-duplicate elimination steps that are part of the standard Relevancer approach and in which retweets, exact- and near-duplicates are detected and eliminated were not applied in the scope of this task.

Clustering First we split the tweet set in buckets determined by periods, which are extended at each iteration of clustering. The bucket length starts with 1 hour and stops after the iteration in which the length of the period is equal to the whole period covered by the tweet set. We run KMeans clustering on each bucket recursively in search of coherent clusters (i.e. clusters that meet certain distribution thresholds around the cluster center), using character n-grams (tri-, four- and five-grams). Tweets that appear in the automatically identified coherent clusters are kept apart from the subsequent iterations of the clustering. The iterations continue by relaxing the coherency criteria until the requested number of coherent clusters is obtained[7] or the maximum allowed coherency thresholds are reached.

Annotation Coherent clusters that were identified by the algorithm automatically are presented to an expert who is asked to judge whether indeed a cluster is coherent and if so, to provide the appropriate label.[8] For the present task, the topic labels are the ones determined by the task organization team (T1-T4). We introduced two additional labels: the irrelevant label for coherent clusters that are about any irrelevant topic and the incoherent label for

[7] This number is determined manually based on the quantity of the available data.

[8] The experts in this setting were the task participants. The organizers of the task contributed to the expert knowledge in terms of topic definition and providing feedback on the topic assignment of the round 1 data.

clusters that contain tweets about multiple relevant topics or combinations of relevant and irrelevant topics.[9] The expert does not have to label all available clusters. For this task we annotated only the one quarter of the clusters from each hour.

Classifier Generation The labeled clusters are used to create an automatic classifier. For the current task we trained a state-of-the-art Support Vector Machine (SVM) classifier by using standard default parameters. We trained the classifier with 90% of the labeled tweets by cross-validation. The classifier was tested on the remaining labeled tweets.[10]

Ranking For our submission in round 1 we used the labels as they were obtained by the classifier. No ranking was involved. However, as the evaluation metrics used in this shared task expected the results to be ranked, for our submission in round 2 we classified the relevant tweets by means of a classifier based on these tweets and used the classifier confidence score for each tweet for ranking them.

We applied clustering with the value for the requested clusters parameter set to 50 (round 1) and 6 (round 2) per bucket, which yielded a total of 1,341 and 315 coherent clusters respectively. In the annotation step, 611 clusters from the round 1 and 315 clusters from round 2 were labeled with one of the topic labels T1-T4, irrelevant, or incoherent. Since most of the annotated clusters were irrelevant or T3, we enriched this training set with the annotated tweets featuring in the FIRE - Forum for Information Retrieval Evaluation, Information Extraction from Microblogs Posted during Disasters - 2016 task[1].

In preparing our submission for round 2 of the current task, we also included the positive feedback for both of our approaches from the round 1 submissions in the training set.

We used a relatively sophisticated method in preparing the submission for round 2. First, we put the tweets from the clusters that were annotated with one of the task topics (T1-T4) directly in the submission with rank 1. The number of these tweets per topic were as follows: 331 – T1, 33 – T2, 647 – T3, and 134 – T4. Then, the tweets from the incoherent clusters and the tweets that were not included in any cluster were classified by the SVM classifier created for this round. The tweets that were classified as one of the task topics were included in the submission. The second part is ranked lower than the annotated part and ranked based on the confidence of a classifier trained specifically for this subset.

Table 1 gives an overview of the submissions in rounds 1 and 2 that are based on the approach using the Relevancer tool. The submissions are identified as rel_ru_nl_ml and rel_ru_ml0 (see also Section 5).

[9] The label definition affects the coherence judgment. Specificity of the labels determines the required level of the tweet similarity in a cluster.

[10] The performance scores are not considered to be representative due to the high degree of similarity of the training and test data.

	Round 1		Round 2	
Topic(s)	*# tweets*	*% tweets*	*# tweets*	*% tweets*
T1	52	0.0012	855	5
T2	22	0.00054	173	1
T3	5,622	0.13	3,422	20
T4	50	0.0012	507	3
T0	35,034	0.85	12,062	71
Total	40,780	100.00	17,019	100.00

Table 1. Topic assignment using Relevancer

3 Approach 2: Topic assignment by rule-based search query generation

The second approach is based on that which is currently being developed by the second author for extracting information from social media data, more specifically from (but not limited to) forum posts.[11] In essence in this approach a lexicon and a set of hand-crafted rules are used to generate search queries. As such it continues the line of research described in Oostdijk & van Halteren [3, 4] and Oostdijk et al. [5] in which word n-grams are used for search.

For the present task we compiled a dedicated (task-specific) lexicon and rule set from scratch. In the lexicon with each lexical item information is provided about the part of speech (e.g. noun, verb), semantic class (e.g. casualties, building structures, roads, equipment) and topic class (e.g. T1, T2). A typical example of a lexicon entry thus look as follows:

deaths N2B T3

where *deaths* is listed as a relevant term with N2B encoding the information that it is a plural noun of the semantic class [casualties] which is associated with topic 3. In addition to the four topic classes defined for the task, in the lexicon we introduced a fifth topic class, viz. T0, for items that rendered a tweet irrelevant. Thus T0 was used to mark a small set of words each word of which referred to a geographical location (country, city) outside Italy, for example *Nepal*, *Myanmar* and *Thailand*.[12]

The rule set consists of finite state rules that describe how lexical items can (combine to) form search queries made up of (multi-)word n-grams. Moreover, the rules also specify which of the constituent words determines the topic class for the search query. An example of a rule is

NB1B *V10D

Here NB1B refers to items such as *houses* and *flats* while V10D refers to past participle verb forms expressing [damage] (*damaged*, *destroyed*, etc.). The asterisk indicates that in cases covered by this rule it is the verb that is deemed to determine the topic class for the multi-word n-gram that the rule describes.

[11] A paper describing this approach is in preparation.

[12] More generally, T0 was assigned to all irrelevant tweets. See below.

This means that if the lexicon lists *destroyed* as V10D and T3, upon parsing the bigram *houses destroyed* the rule will yield T3 as the result.

The ability to recognize multi-word n-grams is essential in the context of this challenge as most single key words on their own are not specific enough to identify relevant instances: with each topic the task is to identify tweets with specific mentions of resources, damage, etc. Thus the task/topic description for topic 3 explicitly states that tweets should be identified 'which contain information related to infrastructure damage, restoration and casualties', where 'a relevant message must mention the damage or restoration of some specific infrastructure resources such as structures (e.g., dams, houses, mobile towers), communication facilities, ...' and that 'generalized statements without reference to infrastructure resources would not be relevant'. Accordingly, it is only when the words *bridge* and *collapsed* co-occur that a relevant instance is identified.

As for each tweet we seek to match all possible search queries specified by the rules and the lexicon, it is possible that more than one match is found for a given tweet. If this is the case we apply the following heuristics: (a) multiple instances of the same topic class label are reduced to one (e.g. T3-T3-T3 becomes T3); (b) where more than one topic class label is assigned but one of these happens to be T0, then all labels except T0 are discarded (thus T0-T3 becomes T0); (c) where more than one topic label is assigned and these labels are different, we maintain the labels (e.g. T1-T3-T4 is a possible result). Tweets for which no matches were found were assigned the T0 label.

The lexicon used for round 1 comprised around 950 items, while the rule set consisted of some 550 rules. For round 2 we extended both the lexicon and the rule set (to around 1,400 items and 1,750 rules respectively) with the aim to increase the coverage especially with respect to topics 1, 2 and 4. Here we should note that, although upon declaration of the results for round 1 each participant received some feedback, we found that it contributed very little to improving our understanding of what exactly we were targeting with each of the topics. We only got confirmation – and then only for a subset of tweets – that tweets had been assigned the right topic. Thus we were left in the dark about whether tweets we deemed irrelevant were indeed irrelevant, while also for relevant tweets that might have been assigned the right topic but were not included in the evaluation set we were none the wiser.[13]

The topic assignments we obtained for the two data sets are presented in Table 2.

In both data sets T3 (Damage, restoration and casualties reported) is by far the most frequent of the relevant topics. The number of cases where multiple topics were assigned to a tweet is relatively small (151/40,780 and 194/17,019 tweets resp.). Also in both datasets there is a large proportion of tweets that were labelled as irrelevant (T0, 81.22% and 75.03% resp.). We note that in the majority of cases it is the lack of positive evidence for one of the relevant topics

[13] The results from round 1 are discussed in more detail in Section 6.

	Round 1		Round 2	
Topic(s)	*# tweets*	*% tweets*	*# tweets*	*% tweets*
T1	91	0.22	206	1.21
T2	55	0.13	115	0.68
T3	7,002	17.17	3,558	20.91
T4	115	0.28	177	1.04
Mult.	151	0.37	194	1.14
T0	33,366	81.82	12,769	75.03
Total	40,780	100.00	17,019	100.00

Table 2. Topic assignment rule-based approach

that leads to the assignment of the irrelevant label.[14] Thus for the data in round 1 only 2,514/33,366 tweets were assigned the T0 label on the basis of the lexicon (words referring to geographical locations outside Italy, see above). For the data in round 2 the same was true for 1,774/12,769 tweets.[15]

For round 1 we submitted the output of this approach without any ranking (rel_ru_nl_lang_analy in Table 4). For round 2 (cf. Table 5) there were two submissions based on this approach: one (rel_ru_nl_lang_analy0) similar to the one in round 1 and another one for which the results were ranked (rel_ru_nl_lang_analy1). In the latter case ranking was done by means of an SVM classifier trained on the results. The confidence score of the classifier was used as a rank.

4 Combined Approach

While analyzing the feedback on our submissions in round 1, we noted that, although the two approaches were partly in agreement as to what topic should be assigned to a given tweet, there was a tendency for the two approaches to obtain complementary sets of results, especially with the topic classes that had remained underrepresented in both submissions.[16] We speculated that this was due to the fact that each approach has its strengths and weaknesses. This then invited the question as to how we might benefit from combining the two approaches.

Below we first provide a brief overview of how the approaches differ with regard to a number of aspects, before describing our first attempt at combining them.

Role of the expert Each approach requires and utilizes expert knowledge and effort at different stages. In the machine learning approach using Relevancer

[14] In other words, it might be the case that these are not just truly irrelevant tweets, but also tweets that are falsely rejected because the lexicon and/or the rules are incomplete.

[15] Actually, in 209/2,514 tweets (round 1) and 281/1,774 tweets (round 2) one or more of the relevant topics were identified; yet these tweets were discarded on the basis that they presumably were not about Italy.

[16] Thus for T2 there was no overlap at all in the confirmed results for the two submissions.

the expert is expected to (manually) verify the clusters and label them. In the rule-based approach the expert is needed for providing the lexicon and/or the rules.

Granularity The granularity of the information to be used as input and targeted as output is not the same across the approaches. The Relevancer approach can only control clusters. This can be inefficient in case the clusters contain information about multiple topics. By contrast, the linguistic approach has full control on the granularity of the details.

Exploration Unsupervised clustering helps the expert to understand what is in the data. The linguistic approach, on the other hand, relies on the interpretation of the expert. To the extent development data are available, they can be explored by the expert and contribute to insights as regards what linguistic rules are needed.

Cost of start The linguistic, rule-based approach does not require any training data. It can immediately start the analysis and yield results. The machine learning approach requires large quantities of annotated data to be able to make reasonable predictions. These may be data that have already been annotated, or when no such data are available as yet, these may be obtained by annotating the clusters produced in the clustering step of Relevancer. The filtering and preprocessing of the data plays an important role in machine learning.

Control over the output In case of the rule-based approach it is always clear why a given tweet was assigned a particular topic: the output can straightforwardly be traced back to the rules and the lexicon. With the machine learning approach it is sometimes hard to understand why a particular tweet is picked as relevant or not.

Reusability Both approaches can re-use the knowledge they receive from experts in terms of annotations or linguistic definitions. The fine-grained definitions are more transferable than the basic topic label-based annotations.

One can imagine various ways in which to combine the two approaches. However, it is less obvious how to obtain the optimal combination. As a first attempt in round 2 we created a submission based on the intersection of the results of the two approaches (rel_ru_nl_lang_analy0 and ru_nl_ml0). The intersection contains only those tweets that were identified as relevant by both approaches and for which both approaches agreed on the topic class. We left the ranking created in ru_nl_ml0 untouched. The results obtained by the combined approach were submitted under run ID rel_ru_nl_ml1. In Table 3 details for this submission are given.

5 Results

The submissions were evaluated by the organizers.[17] Apart from the mean average precision (MAP) and recall that had originally been announced as evaluation

[17] For more detailed information on the task and organization of the challenge, its participants, and the results achieved see **REF TO ORGANIZERS' PAPER**.

	Round 2	
Topic(s)	*# tweets*	*% tweets*
T1	305	2
T2	120	1
T3	2,844	17
T4	149	1
T0	13,601	79
Total	17,019	100

Table 3. Topic assignment combined approach

metrics, two further metrics were used viz. bpref and precision@20, while recall was evaluated as recall@1000. As the organizers arranged for 'evaluation for some of the top-ranked results of each submission' but eventually did not communicate what data of the submissions was evaluated (especially in the case of the non-ranked submissions), it remains unclear how the performance scores were arrived at. In Tables 4 and 5 the results for our submissions are summarized. The various runs and run IDs are as follows:

Run ID	*Description*
rel_ru_ml	Relevancer without ranking
rel_ru_ml0	Relevancer with ranking
rel_ru_ml1	Combined approach
rel_ru_nl_lang_analy	Rule-based approach, no ranking of results
rel_ru_nl_lang_analy0	Rule-based approach, no ranking of results
rel_ru_nl_lang_analy1	Rule-based approach, results ranked

Run ID	*bpref*	*precision@20*	*recall@1000*	*MAP*
rel_ru_nl_ml	0.1973	0.2625	0.0855	0.0375
rel_ru_nl_lang_analy	0.3153	0.2125	0.1913	0.0678

Table 4. Results obtained in Round 1 as evaluated by the organizers

Run ID	*bpref*	*precision@20*	*recall@1000*	*MAP*
ru_nl_ml0	0.4724	0.4125	0.3367	0.1295
rel_ru_nl_lang_analy0	0.3846	0.4125	0.2210	0.0853
rel_ru_nl_lang_analy1	0.3846	0.4625	0.2771	0.1323
ru_nl_ml1	0.3097	0.4125	0.2143	0.1093

Table 5. Results obtained in Round 2 as evaluated by the organizers

6 Discussion

The task in this challenge proved quite hard. This was due to a number of factors. One of these was the selection and definition of the topics: topics T1 and T2 specifically were quite close, as both were concerned with resources; T1 was to be assigned to tweets in which the availability of some resource was mentioned while in the case of T2 tweets should mention the need of some resource. The definitions of the different topics left some room for interpretation and the absence of annotation guidelines was experienced to be a problem.

Another factor was the data set which in both rounds we perceived to be highly imbalanced as regards to the distribution of the targeted topics. Although we appreciate that this realistically reflects the development of an event – you would indeed expect the tweets posted within the first 24 hours after the earthquake occurred to be about casualties and damage and only later tweets to ask for or report the availabilty of resources – the underrepresentation in the data of all topics except T3 made it quite difficult to achieve a decent performance.

As already mentioned in Section 3, the feedback on the submissions for round 1 was only about the positively evaluated entries of our own submissions. There was no information about the negatively evaluated submission entries. Moreover, not having any insight about the total annotated subset of the tweets made it impossible to infer anything about the subset that was marked as positive. This put the teams in unpredictably different conditions for round 2. Since the feedback was in proportion to the submission, having only two submissions was to our disadvantage.

As can be seen from the results in Tables 4 and 5 the performance achieved in round 2 shows an increase on all metrics when compared to that achieved in round 1. Since our approaches are inherently designed to benefit from experts over multiple interactions with the data, we consider this increase in performance significantly positive.

The overall results also show that from all our submissions the one in round 2 using the Relevancer approach achieves the highest scores in terms of the bpref and recall@1000 metrics, while the ranked results from the rule-based approach has the highest scores for precision@20 and MAP. The Relevancer approach clearly benefited from the increase in the training data (feedback for the round 1 results for both our approaches and additional data from the FIRE 2016 task). For the rule-based approach the extensions to the lexicon and the rules presumably largely explain the increased performance, while the different scores for the two submissions in round 2 (one in which the results were ranked, the other without ranking) show how ranking boosts the scores. For reasons we have not begun to understand the combined approach was not as successful as we expected. Further experimentation is needed to determine how the two approaches are best combined.

7 Conclusion

In this report we have described the approaches we used to prepare our submissions for the SMERP Data Challenge Task. Over the two rounds of the challenge we succeeded in improving our results, based on the experience we gained in round 1.

This task along with the issues that we came across provided us with a realistic setting in which we could measure the performance of our approaches. In a real use case, we would not have had any control on the information need of an expert, her annotation quality, her feedback on the output, and her performance evaluation. Therefore, from this point of view, we consider our participation and the results we achieved a success.

As regards future research, this will be directed at improving each of the approaches individually, while we will also continue exploring their combination. The machine learning approach, Relevancer, missed relatively 'small' topics. The clustering and classifying steps should be improved to yield coherent clusters for the small topics and to utilize this information about small topics in the automatic classification respectively. The rule-based approach currently employs contiguous n-grams. Extending it with skip-grams will help increase the coverage as patterns can be matched more flexibly. As observed before, we expect that eventually the best result can be obtained by combining the two approaches. The combination of the outputs we attempted in the context of the current challenge is but one option, which as it turns out may be too simplistic. We intend to explore the possibilities of having the two approaches interact and produce truly joint output.

Acknowledgments

Relevancer has been developed in a project funded by COMMIT and with support of Statistics Netherlands and Floodtags. The software module used in the rule-based approach to interpret the rules, parse the tweets and structure the output is being developed by Polderlink bv.

References

1. Ghosh, S., Ghosh, K.: Overview of the fire 2016 microblog track: Information extraction from microblogs posted during disasters. Working notes of FIRE pp. 7–10 (2016), http://ceur-ws.org/Vol-1737/T2-1.pdf
2. Hürriyetoğlu, A., Gudehus, C., Oostdijk, N., van den Bosch, A.: Relevancer: Finding and Labeling Relevant Information in Tweet Collections, pp. 210–224. Springer International Publishing, Cham (2016), http://dx.doi.org/10.1007/978-3-319-47874-6_15
3. Oostdijk, N., van Halteren, H.: N-gram-based recognition of threatening tweets. In: Gelbukh, A. (ed.) CICLing 2013, Part II, LNCS7817. pp. 183–196. Springer Verlag, Berlin – Heidelberg (2013)

4. Oostdijk, N., van Halteren, H.: Shallow parsing for recognizing threats in dutch tweets. In: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013, Niagara Falls, Canada, August 25-28, 2013). pp. 1034–1041 (2013)
5. Oostdijk, N., Hürriyetoglu, A., Puts, M., Daas, P., van den Bosch, A.: Information extraction from social media: A linguistically motivated approach. PARIS Inalco du 4 au 8 juillet 2016 10, 21–33 (2016)