USI Participation at SMERP 2017 Text Retrieval Task

Anastasia Giachanou, Ida Mele, and Fabio Crestani

Faculty of Informatics, Università della Svizzera italiana (USI), Switzerland
{anastasia.giachanou,ida.mele,fabio.crestani}@usi.ch

Abstract. This report describes the participation of the Università della Svizzera italiana (USI) at the SMERP Workshop Data Challenge Track for the text retrieval task for both Level 1 and Level 2. For this task, we propose a methodology based on query expansion and boolean expressions. For Level 1, we submitted two different methods based on query expansion, where queries were expanded using terms mined from an earthquake-related collection of tweets. In this way, we managed to extract useful expansion terms for each query. In addition to the query expansion, we tried to improve the quality of the retrieved results by incorporating Part-Of-Speech tags. For Level 2, we additionally used information from the partial ground truth that was provided by the organizers in relation to our submitted runs on Level 1. The results showed that our query expansion method had the highest performance in terms of MAP and precision on both levels. In addition, we managed to achieve the second best performance on Level 1 among the submitted semi-supervised approaches in terms of bpref metric.

Keywords: Twitter, emergency situations, text retrieval, query expansion

1 Introduction

The advent of social media has changed the way in which people communicate and exchange information during emergency situations. A large number of user generated data is posted online during emergencies (e.g., earthquake, hurricane) with the aim to share information or assist relief operations [13]. For example, in case of an earthquake people post information about resource-distribution centers (i.e., where people can find shelters or pick up food), or emergency call numbers, and money-donation campaigns. However, the amount of posted data is very large and therefore effective methodologies are needed to help people extract content relevant to their information needs.

One of the most well known microblogs used to share information on emergencies is Twitter[1]. A large number of researchers have used Twitter to address different problems that range from microblog retrieval [2, 9] and tweet recommendation [1] to sentiment analysis [6, 7] and from irony detection [4, 12] to sentiment

[1] http://twitter.com/

dynamics [3, 5]. Extracting useful and relevant information from Twitter is very challenging since tweets are very short and contain a lot of abbreviations and slang language. Expanding the query with more relevant terms is an effective way to address the vocabulary mismatch problem that is mainly caused by their short length [8, 11].

In this report, we describe our participation for the text retrieval task at Exploitation of Social Media for Emergency Relief and Preparedness (SMERP) Data Challenge Track. The evaluation campaign proposed two different tasks, text retrieval and text summarization. In this report we present our methodologies on the text retrieval task for both Level 1 (tweets posted the first day of earthquake) and Level 2 (tweets posted the second day of earthquake). To address the text retrieval task, we propose to expand the initial query with relevant terms and form boolean expressions for each of the provided queries.

For Level 1, we submitted two different methods based on query expansion, where queries were expanded using terms mined from an earthquake-related collection of tweets. In this way, we managed to extract useful expansion terms for each query. The terms were then manually selected in order to create a subset of terms to use in the query expansion. In addition to the query expansion, we tried to improve the quality of the retrieved results by incorporating Part-Of-Speech (POS) tags. For Level 2, organizers provided us with information about which tweets submitted in our runs for Level 1 were actually relevant. In other words, we were provided with a ground truth for the tweets retrieved with our submitted runs. To this end, for Level 2 we expanded each query using information about the relevant tweets from Level 1. We also tried to further improve the performance using a classifier and information from POS tags.

The results showed that plain query expansion is more effective than incorporating information from POS tags. We also noticed that the query expansion method managed to obtain the highest performance in terms of MAP and precision on both levels. These measures are two of the most well known performance measures for evaluation of information-retrieval methods. In addition, we managed to achieve the second best performance for the text retrieval task (Level 1) among the submitted semi-supervised approaches in terms of bpref metric, the official performance measure used by the organizers for the final ranking of the participants.

This report is organized as follows. In Section 2 we present in detail our methodology for the task of text retrieval. In Section 3 we present and discuss the results of our experiments, whereas Section 4 concludes our participation in text retrieval task.

2 Methodology

In this section first we briefly present the task of text retrieval and the provided queries/topics. Then, we present our methodology for the text retrieval task for both Level 1 and Level 2.

2.1 The Text Retrieval Task

For the text retrieval task the organizers released a large collection of tweets that were posted on Twitter during the earthquake in Italy in August 2016. The text retrieval task was divided in two different phases/levels. For Level 1 the organizers released a collection of 52,469 tweets that were posted on the first day of the earthquake. For Level 2 the organizers released 19,751 tweets posted on the second and third day. In addition, the organizers provided information on which tweets among the ones we submitted in our runs for Level 1 were actually relevant. This information could be used for the submissions of Level 2.

Besides data, the organizers gave us four different topics representing different information needs. The aim was to retrieve the relevant tweets for each provided topic. A brief description of the topics is the following:

1. SMERP-T1: Identify the messages which describe the availability of some resources.
2. SMERP-T2: Identify the messages which describe the requirement or need of some resources.
3. SMERP-T3: Identify the messages which contain information related to infrastructure damage, restoration, and casualties.
4. SMERP-T4: Identify the messages which describe on-ground rescue activities of different NGOs and Government organizations.

2.2 Text Retrieval on the First Day (Level 1)

The task of text retrieval consists in retrieving the relevant tweets for four different queries (topics). For Level 1, we expanded each query with terms that were selected from another collection containing the tweets posted during Nepal earthquake that occurred on the 25th of April 2015. To be more specific, the collection contained 90,000 tweets posted from the 1st to the 5th of May 2015.

One issue when using a collection related to a different but similar event (both events are earthquakes, but one occurred in Nepal and the other one in Italy) is that there could be terms specific to the country (e.g., names of locations, people). Hence, we aimed at creating a general collection about earthquakes by using the tweets posted during the earthquake in Nepal and removing all terms related to Nepal. To do so, we first removed URLs and some specific characters (e.g., @, #), then we extracted the entities from the Nepal collection (e.g., Kathmandu, Mahadevstan, Rahul Gandhi) and filter them out. The last step consisted in removing the retweets. At the end of this cleaning process, we had 22,017 tweets, 198,280 tokens, and 12,379 unique tokens.

After cleaning the Nepal collection, we got a collection that is made of general tweets about earthquake and could be used to learn the representative terminology used when an earthquake occurs. We will refer to this collection as $\mathcal{C}_e$. Since we did not have any training data for Level 1, we decided to follow a semi-automatic method where useful terms for expanding the queries were extracted as follows:

1. For each query, we retrieved tweets from $\mathcal{C}_e$. These tweets were retrieved using the terms that appear on the query's description. For the first two queries we also included some terms related to means of transportation such as *helicopter, airplane, train, car, truck, bus,* and *plane.*
2. Given the tweets retrieved for each query, we calculated the $TF - IDF$ of their single terms, bigrams, and trigrams.
3. We manually selected some verb phrases and noun phrases for each query which were either synonyms or additional terms that complemented the description of the query.

At this stage, we had a list of expansion terms and phrases for each query. We submitted two runs, and for both of them our methodology was based on the combination of query expansion (QE) and boolean conjunctions of two different phrases (Ph_1 AND Ph_2). Regarding the two first queries (SMERP-T1 & SMERP-T2), Ph_1 consisted of two lists of candidate phrases that described the availability (for SMERP-T1) or the requirement (for SMERP-T2) of the resources, whereas Ph_2 was the same for both queries and referred to the different resources available/requested. For SMERP-T3, Ph_1 included phrases that described damage or restoration, whereas Ph_2 referred to keywords that described the infrastructure. Finally, for SMERP-T4 we combined keywords that showed rescue and relief activities (Ph_1) with phrases that referred to Non-Governmental Organizations (NGOs) (Ph_2). To learn the NGOs we used a method based on Kullback-Leibler divergence that is described in Section 2.4.

For our first run (USI_1_1) we used boolean queries and we did not consider the POS of the different phrases. This method was expected to retrieve a lot of the relevant tweets but with low precision.

For our second run (USI_1_2) we used the POS tags and forced Ph_1 to be a verb phrase and Ph_2 to be a noun phrase. The NLTK toolkit[2] was used for the POS tagging. However, SMERP-T1 & SMERP-T2 were very similar and required additional information to differentiate keywords that might be relevant for both of them. For example, consider the following two tweets: "People donated quite a bit of money to help the victims," "Consider to help by donating money" they have a significant overlap of keywords, however, the first tweet is more relevant to SMERP-T1 and the second one to SMERP-T2. Therefore, for the first two queries in USI_1_2 we additionally differentiated the queries based on specific POS tags. We considered that only the following POS tags were useful to show announcement or availability of resources or of donations: the past tense (VBD), present participle (be + VBG), future tense (will + VB), present tense (PRP + VB), or past participle (VBN). The verbs that appeared in any of these forms were useful for SMERP-T1. For SMERP-T2 we considered that the verbs *raise, donate, give* had to be in the base form (VB), whereas for the rest of the verbs we did not make any differentiation (they can be in any verb form). Finally, as explained earlier, for SMERP-T3 & SMERP-T4, we considered that the keywords of Ph_1 are only verbs.

[2] http://www.nltk.org/

In Table 1 we report the summary of the two submitted runs for the task of text retrieval (Level 1).

Table 1. Summary of runs submitted for Level 1

Run_id	Task	Description of the run
USI_1_1	Retrieval	QE
USI_1_2	Retrieval	QE + POS

2.3 Text Retrieval on the Second Day (Level 2)

For Level 2 we applied a similar methodology adopted for Level 1 with the difference that instead of using the external collection (the one about Nepal earthquake filtered by Nepal's entities), we expanded the queries with terms extracted from the relevant tweets of the first day of the SMERP collection. Such tweets were annotated as relevant from SMERP organizers and released after Level 1 was completed.

We expected that this would improve the results of our runs because the Nepal collection, despite our filtering based on entities specific of Nepal, may contain contry-specific terms that can be noisy. Similar to the methodology, adopted for Level 1, we decided to manually select the expansion terms. Hence, our methods are characterized as semi-automatic.

We submitted three different runs. Similar to Level 1, the first run (USI_2_1) was based on the combination of query expansion and boolean conjunctions of two different phrases (Ph_1 AND Ph_2). Regarding SMERP-T1 & SMERP-T2, the first phase (Ph_1) consisted of two lists of candidate phrases related to the availability (for SMERP-T1) or the requirement (for SMERP-T2) of the resources, while the second phase (Ph_2) was the same for both queries and refers to the different resources available/requested. For SMERP-T3, Ph_1 included phrases that describe damage or restoration, whereas Ph_2 referred to keywords that describe infrastructure. Concerning SMERP-T4, we used keywords related to rescue and relief activities (Ph_1) together with phrases that refer to NGOs (Ph_2).

In the first run (USI_2_1) we did not consider POS tags. For example, we did not differentiate between the terms *donation* and *donate*. This approach is similar to methodologies based on term stemming. We expected that this method would retrieve a lot of relevant tweets, but its precision would be low.

As already mentioned, one of the main challenges for the text retrieval task was that SMERP-T1 & SMERP-T2 were very similar and additional information was required to differentiate keywords that might be relevant for both of them. We submitted two additional runs in the attempt to address this problem. For the second run (USI_2_2) we built a binary classifier for each of the four topics that were trained to differentiate between relevant and non-relevant tweets. We used a Naïve Bayes classifier that was trained on unigrams, bigrams, and POS tags. Also, we used the same number of training data for the two classes in the training phase.

For the third run (USI_2_3), we leveraged POS tags at query time. For SMERP-T1, we assumed that only the following POS tags were useful to show announcement or availability of a resource or a donation: (1) the present tense for the verbs *provide, send, offer*, (2) the present participle for the verbs *send, offer, gather, collect, raise*, and (3) the past participle for the verbs *donate, raise, collect.* For SMERP-T2, we considered that the verbs *raise, donate* had to be in the base form, the verbs *appeal, ask* in present participle whereas the verbs *require, need* in past participle form. Finally, for the topic SMERP-T4 a list of relevant NGOs was required. For our runs on Level 1, we had created an initial list of NGOs using the Nepal collection. For Level 2, we used this initial list but we kept only the NGOs that also appeared in the relevant tweets for SMERP-T4 (annotated as relevant from SMERP organizers).

For the text retrieval task of Level 2, we submitted three runs. Table 2 shows the summary of the submitted runs.

Table 2. Summary of runs submitted for Level 2

Run_id	Task	Description of the run
USI_2_1	Retrieval	QE
USI_2_2	Retrieval	QE + classifier
USI_2_3	Retrieval	QE + POS-on-query-terms

2.4 Learning the Non-Governmental Organization

As already mentioned, regarding SMERP-T4, we additionally learned the Non-Governmental Organizations (NGOs). To this end, we considered an initial query that should be able to retrieve the tweets mentioning different NGOs. Such query is a single-term query containing the term {donate}. Then, we made the assumption that users refer to the NGOs using their usernames (e.g., @*crocerossa*), so we built a language model for the query and the collection using as tokens the usernames (@*username*). We calculated Kullback-Leibler divergence (KLD) [10] between the query language model Q and the collection language model C. We expect that the usernames with high divergence are good indicators of an NGO. Formally, let w be a word that refers to a username that appears in the collection, and Q be the model of the query q (e.g., the query {donate}), then we can estimate the KLD of the username w as:

$$KLD(w) = P(w|Q) \log \frac{P(w|Q)}{P(w|C)}$$

where $P(w|C)$ is the probability of the username w in the collection and is estimated as follows:

$$P(w|C) = \frac{tf(w,C)}{\sum_{D \in C} |D|}$$

while $P(w|Q)$ is the probability of a word in the query model Q and is estimated as follows:

$$P(w|Q) = \frac{tf(w,Q)}{\sum_{D \in Q} |D|}$$

where $D \in Q$ are the documents relevant to the query q.

We used smoothing to address the problem of zero-frequencies. We managed to have a list of candidates where the higher is the KLD value and more likely the candidate is one NGO's name. From this list, after we normalized the KLD values, we kept the candidates with a value over 0.1. With this approach, we could learn some NGOs (e.g., crocerossa, globalgiving). The final query is a boolean query in the form of Ph_1 AND Ph_2, where Ph_1 shows a rescue activity and Ph_2 can be any of the extracted NGOs.

3 Results and Discussion

Table 3 shows the performance of the submitted runs for the text retrieval task for Level 1 ranked according to MAP, that is widely used to compare the performance of different information retrieval methods. We managed to get the best performance in terms of MAP metric. Moreover, our methods performed very well in terms of precision and recall acquiring some of the highest ranks. This result is very important because shows that simple query expansion methods can be very effective in ranking relevant tweets as the most relevant in the result list.

SMERP organizers used bpref metric as the official evaluation metric for ranking the methodologies proposed in the text retrieval task. The bpref measure is used when there are partial relevance judgments (i.e., just a subset of the documents is annotated). It is defined as the number of documents that are labeled as not relevant and are ranked before those documents that are labeled as relevant. The measure is called bpref because the preference relations are *binary.* It is computed using the preference relation of whether judged relevant documents are retrieved ahead of judged irrelevant documents. In terms of bpref, USI_1_1 was ranked as the second best run among the semi-supervised approaches. In general, we observe that USI_1_1 was better than USI_1_2, showing that POS tags were not very effective. However, at the time this report is written, we do not have access to per topic performance and we can not do further analysis.

Table 4 shows the performance of the submitted runs for the text retrieval task for Level 2 ranked according to MAP. Similar to the performance results of Level 1, we had the highest scores in terms of MAP, precision and recall whereas in terms of bpref we obtained lower performance. From the results we

Table 3. Performance results of the submitted runs on Level 1 text retrieval task

Run_id	Description of the run	MAP	bpref	Precision@20	Recall@1000
USI_1_1	QE	0.0789 (1st)	0.1899	0.5000	0.1825
USI_1_2	QE + POS-on-query-terms	0.0553 (2nd)	0.1063	0.6250	0.1063

can observe that combining query expansion with boolean expressions allows to get the best scores among our submitted runs. In other words, the classifier and the use of POS tags did not manage to improve the performance. For the classification we had limited training data and we believe that this could be one of the reasons of the poor performance. However, further analysis is required to better understand the reasons why the classifier or the information from POS tags did not allow to improve the performance of the query-expansion method.

Table 4. Performance results of the submitted runs on Level 2 text retrieval task

Run_id	Description of the run	MAP	bpref	Precision@20	Recall@1000
USI_2_1	QE	0.1549 (1st)	0.3029	0.7000	0.3029
USI_2_2	QE + classifier	0.1462 (2nd)	0.2425	0.7250	0.2425
USI_2_3	QE + POS-on-query-terms	0.1266 (5th)	0.1828	0.6500	0.1828

4 Conclusions

In this report we presented the participation of the Università della Svizzera italiana (USI) at the SMERP Workshop Data Challenge Track for the task of text retrieval and the two levels (Level 1 and Level 2). Our methodology was based on query expansion and boolean expressions. For Level 1, we submitted two different methods based on query expansion, where queries were expanded using terms mined from an earthquake-related collection of tweets. In addition to the query expansion, we tried to improve the quality of retrieved results by incorporating POS tags. In addition, we submitted three different runs for Level 2 that were also based on query expansion and boolean expressions. For Level 2, we used information from the partial ground truth that was provided by the organizers in relation to our submitted runs on Level 1.

The results showed that our runs had the highest performance in terms of MAP and precision, two metrics that are usually applied to evaluate the performance of information retrieval systems. In addition, we managed to achieve the second best performance in terms of bpref measure for the text retrieval task among the submitted semi-supervised approaches of Level 1.

Acknowledgments. This research was partially funded by the Swiss National Science Foundation (SNSF) under the project OpiTrack.

References

1. Alawad, N.A., Anagnostopoulos, A., Leonardi, S., Mele, I., Silvestri, F.: Network-aware recommendations of novel tweets. In: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016. pp. 913–916 (2016)
2. Bandyopadhyay, A., Ghosh, K., Majumder, P., Mitra, M.: Query expansion for microblog retrieval. International Journal of Web Science 1(4), 368–380 (2012)
3. Bollen, J., Mao, H., Pepe, A.: Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. vol. 11, pp. 450–453 (2011)
4. Farías, D.I.H., Patti, V., Rosso, P.: Irony detection in twitter: The role of affective content. ACM Trans. Internet Technol. 16(3), 19:1–19:24 (2016)
5. Giachanou, A., Crestani, F.: Tracking sentiment by time series analysis. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '16, ACM, New York, NY, USA (2016)
6. Go, A., Huang, L., Bhayani, R.: Twitter sentiment analysis. Entropy 17, 252 (2009)
7. Kouloumpis, E., Wilson, T., Moore, J.D.: Twitter sentiment analysis: The good the bad and the omg! vol. 11, p. 164 (2011)
8. Lau, C.H., Li, Y., Tjondronegoro, D.: Microblog retrieval using topical features and query expansion. In: TREC (2011)
9. Liang, S., de Rijke, M.: Burst-aware data fusion for microblog search. Information Processing & Management 51(2), 89–113 (2015)
10. Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA, USA (1999)
11. Massoudi, K., Tsagkias, M., de Rijke, M., Weerkamp, W.: Incorporating query expansion and quality indicators in searching microblog posts. In: Proceedings of the 33rd European Conference on Advances in Information Retrieval. pp. 362–367. ECIR'11, Springer-Verlag, Berlin, Heidelberg (2011)
12. Reyes, A., Rosso, P., Veale, T.: A multidimensional approach for detecting irony in twitter. Language Resources and Evaluation 47(1), 239–268 (2013)
13. Yin, J., Lampert, A., Cameron, M., Robinson, B., Power, R.: Using social media to enhance emergency situation awareness. IEEE Intelligent Systems 27(6), 52–59 (2012)