Comparative Evaluation of Query Expansion Methods for Enhanced Search on Microblog Data: DCU ADAPT @ SMERP 2017 Workshop Data Challenge

Wei Li and Gareth J.F. Jones

ADAPT Centre,
School of Computing
Dublin City University, Dublin 9, Ireland
wli@computing.dcu.ie, gjones@computing.dcu.ie

Abstract. The rapid growth in the availability of social media content posted during emergency situations is creating significant interest in research into how this information can be exploited to assist emergency relief operations and to help with emergency preparedness and in early warning systems. We describe the DCU ADAPT Centre participation in the microblog search data challenge at the SMERP 2017 workshop. This task aimed to promote development of information retrieval (IR) methods for practical challenges that need to be addressed during an emergency event, along with comparative evaluation of the methodologies developed for this task. The task is based on a large dataset of microblogs posted during the earthquake in Italy in August 2016, together with a set of query topics provided by the task organisers. For our participation in this task we explored use of three different IR techniques: standard IR query expansion based on an external resource, query expansion based on WordNet and use of query expansion involving manual intervention in order to enhance the search results.

Keywords: exploitation of social media in emergency situations, microblog search, query expansion, external collections, WordNet

1 Introduction

Timely and reliable access to and exploitation of all available information is a key factor in effective relief operations in emergency situations. An emerging and potentially very valuable source of real-time information is the rapidly increasing amount of social media content posted during emergency situations on websites such as *Twitter*. Rapid identification and exploitation of relevant information from within the huge volumes of information posted on these social media channels requires effective real-time information retrieval (IR) methods, and potentially integration of social media content with other information sources. This raises the challenge of developing suitable methods to identify the true relevant

information from among the vast volume of content posted to mainstream social media channels in these situations.

The earlier FIRE 2016 Microblog task was motivated by this scenario with the aim of developing IR methods to extract important information from microblogs posted during disasters [1]. Our analysis of the data provided for this task showed that a significant problem in addressing this task is the difficulty in matching the words present in each search topic and those used in the very short microblog documents. Such differences relate both to different choice of vocabulary in the topics and documents, and also to differing levels of specificity in the words used. For the SMERP 2017 Data Challenge we investigated two approaches to addressing this vocabulary mismatch problem. The first applies synonym-based query expansion methods using WordNet for each search topic that we successfully used for the FIRE 2016 Microblog task [2]. The motivation for this approach is to expand the search topic statement to increase the potential of matching it with relevant tweets in the target collection. The risk in adopting this strategy using a general resource such as WordNet without taking context within the topic into account, is that as well as encouraging matching with relevant items we might also match large numbers of non-relevant items. The second method which we explore for the first time for the SMERP 2017 Data Challenge is to use query expansion using an unstructured external information resource [3]. This approach was previously used successfully within our group for a short-document text metadata-based image retrieval task. This method has been shown to be able to identify relevant expansion terms for each query. However, in order to meaningfully apply this technique, it is important to be able identify a suitable external information source. Because this task has the specific topic of emergency response to an earthquake situation, we used the FIRE 2016 Microblog task dataset which is also focused on response to an earthquake, as the external resource for the source of query expansion terms.

In this task our objective to increase the *recall* of relevant tweets, is tempered the potential for *low precision* which can arise from retrieval of non-relevant tweets. However, the methods that we applied for the task were overall found to be effective, and our results for the data challenge track ranked well in both the level 1 and level 2 tasks.

In the following sections, we first describe the SMERP 2017 Data Challenge task in overview, we then introduce our methods and the experiments that we carried out using them, including details of the dataset and external resources that we used, finally we present the results that we obtained and draw conclusions.

2 Track Description and Dataset

The motivation of the data challenge track is to promote development of IR methods that can be used to extract important information from social media during emergency events, and to arrange for comparative evaluation of the

methods used to address this challenge. The challenge had two sub-tracks / challenges:

- Text Retrieval
- Text Summarization

We participated only in the text retrieval challenge this year.

2.1 Text Retrieval sub-track

In this sub-track, participants were required to develop methods to extract tweets that are relevant to each topic with high precision (i.e., ideally, only the relevant tweets should be identified) as well as high recall (i.e., ideally, all relevant tweets should be identified). This sub-track had two levels:

Level 1 In this level, the tweets collected during the first day (24 hours) after the earthquake were provided, and task participants were asked to extract tweets relevant to each of a number of specified topics.

Level 2 In this level, the tweets collected during the second day (24 hours) after the earthquake were provided. Participants were again asked to retrieve tweets relevant to each topic from among the tweets posted during the second day.

2.2 Dataset

Tweet collection: A large dataset of microblogs (tweets) posted on Twitter during the earthquake in Italy in August 2016 was provided to participants. Since the Twitter terms of usage do not allow public sharing of tweets, only the tweetids of the tweets were provided, along with a Python script that can be used to download the tweets using the Twitter API.

For Level 1 The task organizers provided:

- a text file of 52,469 tweetids;
- a Python script along with the libraries that are required by this script.

For Level 2 The task organizers provided:

- a text file of 19,751 tweetids;
- same as level 1, a Python script along with the libraries that are required by this script.

For both levels, the same topics were used: A set of 4 topics (information needs) were provided. Each topic identifies a broad information need encountered during a disaster, such as "what resources are needed in the disaster affected area?", "what resources are available?", "what damages are being reported?", etc. Specifically, each topic is presented in TREC format, and as such contains: a title, a brief description, and a more detailed narrative on what type of tweets can be considered relevant to the topic. An example of TREC format topic is shown in Figure 1.

```
<top>

<num> Number: SMERP-T1
<title> WHAT RESOURCES ARE AVAILABLE

<desc> Description:
Identify the messages which describe the availability of some resources.

<narr> Narrative:
A relevant message must mention the availability of some resource like
food, drinking water, shelter, clothes, blankets, blood, human resources
like volunteers, resources to build or support infrastructure, like tents,
water filter, power supply,etc. Messages informing the availability of
transport vehicles for assisting the resource distribution process would
also be relevant. Also, messages indicating any services like free wifi,
sms, calling facility etc. will also be relevant. In addition,any message
or announcement about donation of money will also be relevant.However,
generalized statements without reference to any resource would not be
relevant.

</top>
```

Fig. 1. Example of TREC Format Topic

We used the task organizers' instructions to download the listed tweets arising from the Italian earthquake in August 2016. A total of 52,331 tweets were downloaded and written into a Json file for level 1 and 19,406 tweets downloaded for level 2. We then prepared a Json parser to decode and extract the information that we needed which consisted of only the tweet id and the content of each tweet. An example decoder is shown in Figure 2.

2.3 External Resource

As mentioned above, we used the FIRE 2016 task dataset as the external resource for query expansion based on external information. The FIRE 2016 task data focuses on the same emergency setting of an earthquake, in this case consisting

```
import json
if __name__ == "__main__":
     aDict = {}
     source = open('*.jsonl', 'r')
     for line in source:
         data = json.loads(line)
          tid = '<id>' + str(data['id']) + '</id>'
          ttext = '<text>' + data['text'].encode('utf8') + '</text>'
          aDict[tid] = ttext
      source.close()

      target = open('*.txt', 'w')

      for i in aDict:
        line = str(i) +' '+ aDict[i]
        target.write(line)
        target.write('\n')
        print i
      target.close()
```

Fig. 2. Example of Jason Decoder in Python

of tweets from the Nepal earthquake which happened in April 2015. A total of 49,894 tweets for this dataset based on the provided list of tweetids were downloaded. We used these tweets as the external resource to carry out query expansion method to extract the expansion terms for each of the SMERP topics.

3 Experimental Methods and Procedures

The same methods were applied for both the level 1 and 2 activities. For the SMERP 2017 task, we submitted three runs based on three different query expansion methods.

We first give the motivation for our exploration of WordNet based query expansion methods for this task, and then give details of the exact method used in our experiments. In seeking to address this microblog task, we first considered the requirements of the task and the nature of the data. The task seeks to retrieve relevant items at the highest possible rank in the retrieved list for each topic, but also to identify as many relevant items as possible from within the collection. Both of these make sense in the context of the task, in the sense that the first requirement points towards efficiency of investigation by the searcher, and the latter one towards exhaustivity in coverage of the answers to the information need available within the collection. An important overall consideration being that this can be seen as a recall-focused task, for which quality of ranking of relevant items should be maximised.

Based on our analysis of the task and the likelihood of query-document mismatch problems arising from the short length of the tweets and the differing use

of vocabulary in the topics and the tweets, we hypothesised that query expansion to encourage greater levels of matching between topics and tweets is likely to be an important component of a successful retrieval strategy. Our earlier work on the similar FIRE 2016 Microblog task looked at a simple application of WordNet for query expansion which was shown to be successful for this task [2]. In this paper we extend this earlier work to explore three approaches which seek to improve the reliability of query-document matching for enhanced retrieval effectiveness:

1. Use WordNet as an external resource to generate the synonyms for topic terms to perform query expansion;
2. Use a query expansion method based on an external resource (using FIRE 2016 task dataset as the external resource) to carry out query expansion for each topic;
3. Use a semi-automatic method which involves a manual selection of relevant tweets and use the selected tweets to identify expansion terms. Then apply WordNet on the expanded query to expand them again.

3.1 Information Retrieval

For our runs we used the Lucene[1] IR framework to index the tweet data collection and to perform the subsequent IR runs. The indexing process used the following steps:

1. Entries from a list of 500 stop words were removed, stopwords used in all these experiments were generated from the collection itself. We computed the term frequency across the whole dataset and ranked the terms. The 500 most frequently used terms were then selected as stopwords;
2. The standard BM25 IR model was used for retrieval with k_1=1.2, b=0.75 [5], these parameter values were set empirically based on the experiments we carried out for the FIRE 2016 task.

The initial search query used for all experiments in this paper is formed from the combination of title and narrative fields of the topic, since the title field alone is too short and also too similar for the individual queries to be distinct from each other. This problem can be seen clearly from the Table 1, which shows the title part of the 4 topics provided. It can be seen that topic 1 and 2 are too similar to make a distinction between them.

3.2 Query Expansion using WordNet (Run_id: dcu_ADAPT_run2)

WordNet[2] is an electronic lexical database and is regarded as one of the most important resources available to researchers in computational linguistics, text analysis, and many related areas. Its design is inspired by psycholinguistic and

[1] `https://lucene.apache.org`
[2] `https://wordnet.princeton.edu/`

Table 1. Title fields of provided topics.

Title of Topics	
Topic 1	WHAT RESOURCES ARE AVAILABLE
Topic 2	WHAT RESOURCES ARE REQUIRED
Topic 3	WHAT INFRASTRUCTURE DAMAGE, RESTORATION AND CASUALTIES ARE REPORTED
Topic 4	WHAT ARE THE RESCUE ACTIVITIES OF VARIOUS NGOs / GOVERNMENT ORGANIZATIONS

computational theories of human lexical memory. English nouns, verbs, adjectives, and adverbs are organized into synonym sets, each representing one underlying lexicalized concept. Different relations link the synonym sets [6].

WordNet has long been regarded as a potentially useful resource for query expansion in IR. However, it has met with limited success due to its tendency to include contextually unrelated synonyms for query words which are unrelated to document relevance. One of the successful applications of WordNEt in IR is found in [8] which uses the comprehensive WordNet thesaurus and its semantic relatedness measure modules to perform query expansion in a document retrieval task. The authors obtained a 7% improvement on retrieval effectiveness compare to the performance of using original query for search. [7] combined terms obtained from three different resources, including WordNet for use as expansion terms, Their method was tested on a TREC ad hoc test collection with impressive results.

In this experiment, we also used WordNet to carry out query expansion. Two types of terms were extracted from WordNet for each topic term, its synonyms and its hyponyms. This run was one of our automatic runs submitted for the task, and was carried out as follows:

1. Remove stop words from each topic.
2. Use WordNet to generate the synsets for each of the remaining non-stopword topic terms.
3. Use WordNet to generate the hyponyms for each term of the remaining non-stopword topic terms.
4. Both obtained synonyms and hyponyms are ranked based on their distance to the original topic term. We selected the top 10 synonyms and the top 10 hyponyms for use as expansion terms and add them to the original topic statement. The value of 10 expansion terms for both synonyms and hyponyms are selected based on our development runs carried out using the FIRE 2016 data set, for which using 10 for each obtained the best results.
5. Use the expanded topic as the new topic to search the document collection using the BM25 retrieval model. Parameters are where k_1=1.2, b=0.75, training on the FIRE 2016 dataset.

3.3 Query Expansion using External Resources (Run_id: dcu_ADAPT_run1)

Query expansion using external resources (QEE) is another method which seeks to address the mismatch problem between the query and the available documents in the collection. QEE seeks to include additional topically relevant terms in the query [3]. QEE was used to improve retrieval effectiveness for an image retrieval task using expansion of both query and the source documents. In this paper, we only use QEE to expand the query. During the expansion process, queries are searched on an external collection of documents. A number of terms from this external collection are then selected from the top ranked documents returned by the search process and then added to the original query. As described above, the FIRE 2016 Microblog collection was used as the external resource to perform document expansion for queries. This formed the other of our automatic runs for the task. The procedure of the expansion process was as follows:

1. A list of 500 stop words were created for the FIRE 2016 document collection using the same procedure described for the SMERP document collection. These stop words were then removed from the FIRE document collection.
2. The FIRE 2016 document collection was then indexed using Lucene.
3. The BM25 retrieval model used to retrieve a document list for each topic where again k_1=1.2, b=0.75.
4. The top 30 returned documents were assumed to be relevant, and 20 terms from these documents were then selected as the query expansion based on computer term offer weight. The offer weight for each term t for selection of expansion terms was computed as follows [3]:

$$OW(t) = r(t) * idf(t)$$

   where $r(t)$ is the number of documents which contain term t and $idf(t)$ indicates the inverse document frequency of the term t, computed as follows:

$$idf(t) = log\frac{N - n(t) + 0.5}{n(t) + 0.5}$$

   where t is the term, N is the total number of documents in this collection and $n(t)$ is the number of documents which contain the term t.
5. The selected terms were then added to the original topic. Note that the selected expansion terms exclude the original query terms.
6. The expanded topics were then used to search on the SMERP 2017 earthquake tweet collection using BM25 retrieval model on Lucene with same parameters above.

For this experiment, we inspected the top 50 retrieved FIRE collection tweets for each of the original topic, and observed that retrieved relevant tweets are more likely occur in top 30 rather than appear on the rank from 30 to 50. So in order to reduce the noise, we chose the top 30 as the considered document

list number in this experiments. However, in future work, a different number of assumed relevant document can be investigated. Both the value of k_1, b and number of 20 expansion terms are selected empirically based on experiments carried out using the FIRE 2016 task.

3.4 Semi-Auto Run Involvimg Manual Selection (Run_id: dcu_ADAPT_run3)

Our third run was a semi-automatic run which involved manual selection of relevant documents. This run was carried out using the following steps:

1. Use the original topic to search and obtain a ranked list from the indexed SMERP document collection.
2. Go through top 30 tanked tweets from the ranked list and select 1-2 relevant tweets and then perform query expansion using the SMERP target collection. The umber 30 was to help to ensure that we should be able to identify at least one relevant tweet for each topic.
3. Remove the stopwords and duplicate terms from the selected tweets, add the remaining terms to the original topic.
4. Use WordNet to expand the terms of the query, the same as for the run conducted in section 3.2, the top 10 synonyms and the top 10 hyponyms were selected as expansion terms.
5. Add the chosen expansion terms to each manually expanded topic to generate new topics and use them as query to search again to obtain the final search results.

4 Experiment Results

Since the aim of this task is to identify a set of tweets that are relevant to each topic, set-based evaluation metrics of precision, recall, bpref and MAP were used for evaluation.

4.1 Results for Task Level 1

Table 2 shows the evaluation results for the text retrieval task level 1. The table is ranked according to the bpref measure, because the relevance judgements are known to be far from complete, so bpref is a suitable metric to compute for evaluation. From the table we can see that we are the only group who submitted a fully-automatic runs for this task. Our two fully-automatic runs (dcu_ADAPT_run2 and dcu_ADAPT_run1) are ranked in first and second place based on bpref values which are 0.3652 and 0.3549 respectively, which are much higher than other groups' bpref value. Our semi-automatic run (dcu_ADAPT_run3) ranked on the third place with a bpref value 0.2558 which still outperforms the run in fourth place by 34.7%.

These numbers show that using QEE and WordNet are positive ways to carry out query expansion for this Microblog task. Using WordNet to generate

Table 2. Level 1 Results

Text Retrieval Task Level 1 Evalution Results						
Rank	Run-id	Run type	bpref	P@20	Recall@1000	MAP
1	dcu_ADAPT_run2	Full-automatic	0.3654	0.3875	0.1065	0.0314
2	dcu_ADAPT_run1	Full-automatic	0.3549	0.2375	0.1078	0.0344
3	dcu_ADAPT_run3	Semi-automatic	0.2558	0.1500	0.0720	0.0188
4	USI_1	Semi-automatic	0.1899	0.5000	0.1825	0.0789
5	daiict_irlab_2	Semi-automatic	0.1777	0.2000	0.1777	0.0204
6	daiict_irlab_1	Semi-automatic	0.1736	0.2000	0.1701	0.0198
7	rel_ru_nl_lang_analy	Semi-automatic	0.1542	0.3750	0.1119	0.0314
8	rel_ru_nl_ml	Semi-automatic	0.1353	0.4250	0.0528	0.0294
9	charusat_smerp17_1	Semi-automatic	0.1130	0.1500	0.1010	0.0100
10	USI_2	Semi-automatic	0.1063	0.6250	0.1063	0.0553

Table 3. Level 2 Results

Text Retrieval Task Level 2 Evalution Results						
Rank	Run-id	Run type	bpref	P@20	Recall@1000	MAP
1	dcu_ADAPT_run2	Full-automatic	0.7767	0.2125	0.2378	0.0600
2	dcu_ADAPT_run1	Full-automatic	0.6861	0.2375	0.2190	0.0627
3	ru_nl_ml0	Semi-automatic	0.4724	0.4125	0.3367	0.1295
4	rel_ru_nl_lang_analy1	Semi-automatic	0.3846	0.4625	0.2771	0.1323
5	rel_ru_nl_lang_analy0	Semi-automatic	0.3846	0.4125	0.2210	0.0853
6	dcu_ADAPT_run3	Semi-automatic	0.3821	0.2500	0.2572	0.0399
7	ru_nl_ml1	Semi-automatic	0.3097	0.4125	0.2143	0.1093
8	USI_2.1	Semi-automatic	0.3029	0.7000	0.3029	0.1549
9	daiict_irlab_l2_2	Semi-automatic	0.2869	0.3750	0.2869	0.0635
10	daiict_irlab_l2_1	Semi-automatic	0.2869	0.2875	0.2869	0.0571
11	USI_2.2	Semi-automatic	0.2425	0.7250	0.2425	0.1462
12	USI_2.3	Semi-automatic	0.1828	0.6500	0.1828	0.1266
13	daiict_irlab_l2_3	Semi-automatic	0.1204	0.3000	0.1204	0.0433
14	charusat_smerp17_2	Semi-automatic	0.0218	0.0875	0.0218	0.0072

synonyms and hyponyms for topic term methods performs better than using document expansion approach where WordNet bpref value is 0.0105 better than the QEE one.

4.2 Results for Task Level 2

Table 3 shows the evaluation results for Level 2. Our fully automatic runs are again ranked in first and second places with bpref values of 0.7767 and 0.6861 respectively, which are much higher than the third place run whose bpref value is 0.4724. However, our semi-automatic run, dropped from third place in level 1 to sixth place, however the bpref is only slightly lower than the runs in fourth and fifth places. This still demonstrates our conclusion from level 1 that the WordNet and QEE methods work effectively on the Microblog retrieval task.

5 Conclusions and Further Work

For our submissions to the SMERP 2017 data challenge, we examined two methods for query expansion: the first is exploit query expansion based on external resource which used FIRE 2016 dataset as external resources to extract expansion terms to carry out query expansion for topics. The other employed WordNet as an external resource to carry out query expansion by retrieve the synonyms of each topic term and use them as the additional query terms to reformulate each topic. We conducted two runs using this method, an automatic run and a semi-automatic run. The semi-automatic involved manual selection of relevant tweets from a first run, using this information to expand the original topic, and then applying WordNet to expand the topic further. Our automatic runs received the first two places among all submissions for both levels. Our semi-automatic run ranked in third and sixth place for levels 1 and 2 respectively, but still obtained a relatively good bpref value. These positive results show that when a topic is too general and does not contain the necessary terms to match with relevant documents, exploiting external resources, such as WordNet and similar tweets, is a good way to carry out query expansion. Potential further work could be to examine alternative parameter settings for these methods or to merge them. Also applying document expansion for each document in the target collection is another method to potentially improve results which could be investigated [3].

6 Acknowledgement

This research is supported by Science Foundation Ireland in the ADAPT Centre (Grant 13/RC/2106) (www.adaptcentre.ie) at Dublin City University.

References

1. S. Ghosh and K. Ghosh. Overview of the FIRE 2016 microblog track: Information extraction from microblogs posted during disasters. In Working notes of FIRE 2016 - Forum for Information Retrieval Evaluation, pages 5661, Kolkata, India, (2016)
2. W. Li, D. Ganguly, and G. J. F. Jones: Using wordnet for query expansion: ADAPT @ FIRE 2016 microblog track. In Working notes of FIRE 2016 - Forum for Information Retrieval Evaluation, pages 6265,Kolkata, India, (2016)
3. J.M. Min, J. Leveling, D. Zhou and G.J.F. Jones: Document expansion for image retrieval. In proceeding of RIAO '10 Adaptivity, Personalization and Fusion of Heterogeneous Information Pages 65-71, (2010)
4. A. Singhal and F. Pereira: Document Expansion for Speech Retrieval. In Proceedings of the 22nd annual international ACM SIGIR conference on research nad development in information retrieval, page 34-41, Berkeley, California, USA. (1999)
5. S. Robertson and H. Zaragoza. The probabilistic relevance framework: BM25 and beyond.Foundations and Trends in Information Retrieval, 3(4) (2009)
6. G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, andK. Miller. Wordnet: An on-line lexical database.International Journal of Lexicography, 3:235244, (1990)

7. D. Pal, M. Mitra, and K. Datta. Improving query expansion using wordnet. CoRR, abs/1309.4938, (2013)
8. J. Zhang, B. Deng, and X. Li. Concept based query expansion using wordnet. In Proceedings of the 2009International e-Conference on Advanced Science and Technology, AST 09, pages 5255, Washington, DC,USA, IEEE Computer Society. (2009)