Using WordNet for Query Expansion: ADAPT @ FIRE 2016 Microblog Track

Wei li Debasis Ganguly Gareth J. F. Jones

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

ABSTRACT

User-generated content on social websites such as *Twitter* is known to be an important source of real-time information on significant events as they occur, for example natural disasters. Our participation in the FIRE 2016 Microblog track, seeks to exploit WordNet as an external resource for synonym-based query expansion to support improved matching between search topics and the target Tweet collection. The results of our participation in this task show that this is an effective method for use with a standard BM25 based information retrieval system for this task.

CCS Concepts

•Information systems → Query reformulation;

Keywords

Microblog Search; WordNet; Query Expansion

1. INTRODUCTION

User-generated content on social media websites such as *Twitter* is known to be an important real-time source of information on various events as they occur, including disaster events like floods, earthquakes and terrorist attacks. If information relevant to these events can be reliably identified automatically, there is huge potential to exploit it in the management of the response to these events by disaster and relief agencies. This raises the challenge of developing methods to identify the true relevant information from among the vast volume of content posted to mainstream social media channels. The FIRE 2016 Microblog task [1] is motivated by this scenario, and aims to promote development of information retrieval (IR) methods to extract important information from microblogs posted during disasters.

Our analysis of the task data 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. In order to address this challenge, we investigate the use of synonym-based query expansion using WordNet for each search topic. The motivation for this approach is to expand the topic to enhance the chance of matching it with relevant tweets in the target collection. The risk of adopting this strategy using resources such as WordNet without taking into account the context within the topic, is that apart from matching with relevant items, we will also match large numbers of non-relevant items. In this case our objective of increasing recall of relevant tweets, will be tempered by low precision arising from retrieval of non-relevant tweets.

In following section, we first describe the task in overview, we then introduce our method and the experiments that we carried out using it, including details of the dataset and the external resources that we used, finally we present the results that we obtained and draw conclusions.

2. TRACK DESCRIPTION

The FIRE 2016 Microblog track [1] requires the identification of relevant items from within a large set of microblogs (tweets) posted during a recent disaster event, for a set of given topics (in TREC format). Each topic identifies a broad information need during a disaster, such as, what resources are needed by the population in the disaster affected area, what resources are available, what resources are required / available in which geographical region, and so on. Specifically, each topic contains a title, a brief description, and a more detailed narrative describing in summary what types of tweets will be considered relevant to the topic. Task participants are required to develop methodologies for extracting tweets that are relevant to each topic with high precision as well as high recall.

The main challenges for this ad-hoc search task are:

- Identifying specific keywords relevant to each broad topic within each tweet which only contains a few words (140 characters at most), most of which do not contain specific keywords relating to the disaster even the tweet itself is relevant to the search topic.

- Dealing with noise in the content of the short tweet documents which are often written in an informal style using abbreviations, colloquial terms, etc;

3. EXPERIMENTAL METHODS AND PROCEDURES

We begin this section by summarising details of the dataset, and then describe our experiments and the results obtained.

3.1 Data

Listing 1: Json Parser Code

```
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()
```

In order to obtain the dataset of tweets for the task, we followed the instruction provided by the task organizers. They provided:

- a text file of 50,068 tweetids;
- a Python script, along with the libraries that are required by this script, to crawl the tweets.

We used their instructions to download the listed tweets arising from the Nepal earthquake in April 2015. A total of 49,894 tweets were downloaded and written into a Json file. We then prepared a Json parser to decode and extract the information that we needed which consisted of only the tweet id and content of each tweet. Listing 1 shows the code.

The provided query set contained 7 topics in TREC format, each of which contains three parts: title, brief description, and a more detailed narrative on what type of tweets will be considered relevant to the topic. Listing 2 presents an example of the TREC format topic:

Listing 2: TREC Topic Example

```
<top>
<num> Number: FMT1
<title> WHAT RESOURCES WERE 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, human resources like volunteers
resources to build or support infrastructure
like tents, water filter, power supply and
so on. Messages informing the availability
of transport vehicles for assisting the
resource distribution process would also
be relevant. However, generalized statements
without reference to any resource or
messages asking for donation of money would
not be relevant.
</top>
```

3.2 Experiments and Results

Based on our observation of the probable query-document mismatch problems arising from the short length of the tweets and the differing use of vocabulary in the topics and the tweet, we explore the use of WordNet[1] to improve the reliability of query-document marching. We used WordNet to generate synonyms for the terms in each topic. Two experiments were conducted based on the WordNet. In these experiments, Lucene was used to index the tweet set and to carry out the IR. The indexing process followed the following steps:

1. entries from a list of 655 stop words were removed;
2. Porter stemmer was used for stemming the words;
3. BM25 was used for indexing with k_1=1.2, b=0.75.

3.2.1 *Query Expansion using WordNet*

WordNet 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 current psycholinguistic and 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 [2].

WordNet has long been regarded as a potentially useful resource for query expansion in IR. However, it has met with

[1] https://wordnet.princeton.edu/

limited success due to its tendency to include contextually unrelated synonyms for query words which are unrelated. One of the successful applications of WordNet in IR is found in [4] which uses the comprehensive WordNet thesaurus and its semantic relatedness measure modules to perform query expansion on a document retrieval task. The authors obtained a 7% improvement on retrieval effectiveness compare to the performance of using original query for search. [3] 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 use WordNet to carry out query expansion. WordNet is used as external resource to generate the synonyms for each topic. We limited the number of synonyms for each topic term to 20 maximum, some terms received less synonyms.

3.2.2 Experiment One

Our first run (named dcu_fmt16_1) uses our first automatic method using WordNet query expansion. In this run, the following 4 steps were applied:

1. remove stop words from each topic;
2. use WordNet to generate the synonyms for each item in every topic;
3. these synonyms are used as expand terms and add them back to each topic;
4. use the expanded topics as new topic to search again (BM25 retrieval model is used for retrieval).

We use the combination of title and narrative fields of the topic in combination as the original topic. An example of an original topic and its extend version are shown in the Appendix.

3.2.3 Experiment Two

Our second run (named dcu_fmt16_2) is an semi-automatic run which means that the manual selection is involved. This run was carried out using the following steps:

1. use the original topic to search and obtain a rank list;
2. go through top 30 tweets from rank list to select 1-2 relevant tweets and to do query expansion. Number 30 is selected to promise we could find at least one relevant tweet for some topics.
3. remove the stopwords and duplicate terms from the select tweets, add the rest term to original topic;
4. then, applied WordNet again on the expanded topics and find synonyms for these terms;
5. finally, add the synonyms to each expanded topic to generate new topics and use them as query to search again to obtain the final search results.

3.2.4 Experimental Results

Since the aim of this track is to identify a set of tweets that are relevant to each topic, set-based evaluation metrics of precision, recall, and MAP are used for evaluation. The gold standard, against which the set of tweets identified by the participants are matched, is generated using a "manual run" where human assessors were given the same set of tweets and topics, and asked to identify all possible relevant tweets using a search engine (Indri). While judging the participants' runs, the track organizers arranged for a second round of assessments to judge the relevance of tweets that are identified by the participants but were not identified during the first round of human assessment.

Results of our two runs are shown in Table 1. The table shows results for 4 runs, two of them are automatic and the other two are semi_automatic. In the automatic runs listed, our submissions were placed third. The Precision@20 of the best automatic run result is 0.4357 where ours is 0.3786. However our automatic run achieved the best MAP@1000 value of 0.1103, which is an increase of 27.93% relative to the best run. Our overall MAP is lower because we only submitted the top 1000 tweets for each topic while other participants submitted more. We received first place for the semi-automatic method where our Precision@20 is 33.35% higher than the second place run. These numbers show that using WordNet to generate synonyms for topic terms is a positive way to carry out query expansion for this Microblog task.

4. CONCLUSIONS AND FURTHER WORK

For our submissions to the FIRE 2016 Microblog Track, we employed WordNet as an external resource to carry out query expansion by retrieving the synonyms of each topic term and using 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 and application of WordNet in a subsequent retrieval stage. Our automatic run received the third place among submission, however with the best MAP value. Our semi-automatic run obtained the overall first place. These positive results show that when a topic is too general and does not contain the necessary terms to match with relevant documents, using WordNet as an external resource to generate synonyms is a good way to make them more effective. Potentially, using WordNet to retrieve hypernym or hyponym for each topic term maybe another method worth attempt for this task.

5. ACKNOWLEDGEMENT

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

6. 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, Kolkata, India, December 7-10, 2016*, CEUR Workshop Proceedings. CEUR-WS.org, 2016.

[2] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. Wordnet: An on-line lexical database. *International Journal of Lexicography*, 3:235–244, 1990.

[3] D. Pal, M. Mitra, and K. Datta. Improving Query Expansion Using Wordnet. *CoRR*, abs/1309.4938, 2013.

Table 1: Our Results and Comparison with Others

Run Type	Run Name	Rank	Precision@20	Recall@1000	MAP@1000	MAP
Automatic run	iiest_saptarashmi_bandyopadhyay_1	1	0.4357	0.3420	0.0869	0.1125
Automatic run	dcu_fmt16_1	3	0.3786	0.3578	0.1103	0.1103
Semi-auto run	dcu_fmt16_2	1	0.4286	0.3445	0.0815	0.0815
Semi-auto run	iitbhu_fmt16_1	2	0.3214	0.2581	0.0670	0.0827

[4] J. Zhang, B. Deng, and X. Li. Concept Based Query Expansion Using Wordnet. In *Proceedings of the 2009 International e-Conference on Advanced Science and Technology*, AST '09, pages 52–55, Washington, DC, USA, 2009. IEEE Computer Society.

Appendix

Original topic:

<num> Number: FMT6

<title> WHAT WERE THE ACTIVITIES OF VARIOUS NGOs / GOVERNMENT ORGANIZATIONS

<narr> Narrative: A relevant message must contain information about relief-related activities of different NGOs and Government organizations in rescue and relief operation. Messages that contain information about the volunteers visiting different geographical locations would also be relevant. However, messages that do not contain the name of any NGO / Government organization would not be relevant.

Expanded topic:

<num> Number: FMT1

<narr>were activities assorted respective several diverse versatile various NGOs government organization organisation arrangement system administration governance governing body establishment brass constitution formation organizations a relevant message mustiness moldiness must incorporate comprise hold bear carry control hold in check curb moderate take turn back arrest stop hold back contain information about relief related activities different organization organisation arrangement system administration governance governing body establishment brass constitution formation organizations indium atomic number four9 indiana hoosier state inwards inward in deliverance delivery saving deliver rescue relief operation. message content subject matter substance messages that incorporate comprise hold bear carry control hold in check curb moderate take turn back arrest stop hold back contain information about volunteers visit see travel call in call inspect inflict bring down impose chew fat shoot breeze chat confabulate confab chitchat chit chat chatter chaffer natter gossip jaw claver visiting different geographical location placement locating position positioning emplacement localization localisation fix locations would besides too likewise well also relevant. message content subject matter substance messages that do not incorporate comprise hold bear carry control hold in check curb moderate take turn back arrest stop hold back contain name whatever whatsoever any NGO government organization would not relevant