Semi-automatic keyword based approach for FIRE 2016 Microblog Track

Ganchimeg Lkhagvasuren
Évora University
ganchimeg@seas.num.edu.mn

Teresa Gonçalves
Évora University
tcg@uevora.pt

José Saias
Évora University
jsaias@uevora.pt

ABSTRACT

This paper describes our semi-automatic keyword based approach for the four topics of Information Extraction from Microblogs Posted during Disasters task at Forum for Information Retrieval Evaluation (FIRE) 2016. The approach consists three phases; Keywords extraction, Retrieval, and Classification.

CCS Concepts

• **Computing methodologies** □**Support Vector Machine**
• **Information systems**□**Information Extraction.**

Keywords

Supervised classification; Information extraction; Terrier; Twitter.

1 INTRODUCTION

It is undeniable that microblogging sites have become key resources of significant information during disaster event [1]. One of these microblogging site, Twitter, is a social networking website which enables users to generate 140-character messages named "tweets" everyday. A giant number of tweets is posted including informative and non-informative messages, which makes opportunities for information extraction [3].

However, dealing with tweets and identifying specific keywords are challenging work due to the nature of Twitter. The small, noisy and fragmented *tweets mean they have very simple discourse and pragmatic structure, issues which still challenge state-of-the-art NLP systems* [2].

Task description: The aim is to retrieve a number tweets relevant to each topic provided with high precision as well as high recall. The titles of topics are provided in TREC format as the following:

1. What resources were available
2. What resource were required
3. What medical resource were available
4. What medical resource were required
5. What were the requirements or availability of resources at specific location
6. What were the activities of various NGOs or government organizations
7. What infrastructure damage or restoration were reported

Dataset: Approximately 50,000 tweets that posted during Nepal earthquake disaster were given in JSON format. A main feature of the task is that a gold standard dataset was not provided.

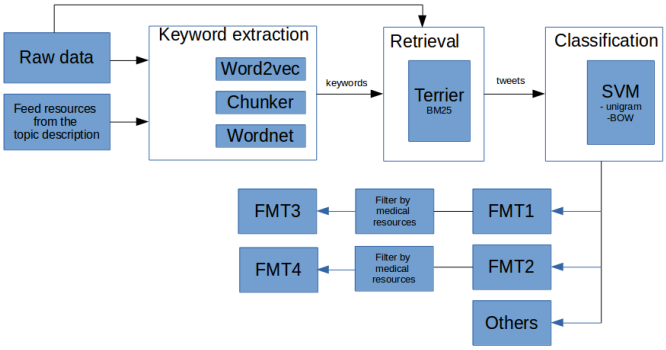

Figure 1. Processing pipeline for the task

In terms of our approach, we propose to achieve the first four topics using keywords extraction with manual work and classification methods.

This paper is organized as follows. First, the components of approach are described separately. Then, the result analysis and conclusion are presented. Our work is submitted in FIRE 2016 Microblog track [7].

2 KEYWORD BASED APPROACH

Our approach for the Microblog track comprises three phases, Keyword extraction, Retrieval, and Classification (see Figure 1). In the first phase, we extracted all relief resources (keywords) that were available or required. Using those keywords and Terrier[1] search engine, we retrieved a number of tweets that each tweet includes a keyword at least in the middle phase. In the last phase, the retrieved tweets are classified into the first and second topics using Support Vector Machine (SVM).

2.1 Extracting keywords

In order to extract keywords, we used separately the following two methods with manual work. The keywords that we first extracted is provided in the topic descriptions, such as food, water, volunteer, money, medicine and transportation. The quantitative results explored in this phase is presented in Table 1.

Since tweets are usually written in an informal style, the most of NLP tools show poor performance on Twitter datasets. So we tried to exploit specific NLP tools which are [4] and [5].

Word embedding: Based on these relief resources we mentioned before, we attempted to obtain more keywords from the given dataset. In order to do that, First of all, we tagged all tweets by GATE twitter Part-of-speech tagger [4]. After distinguished all

1 http://terrier.org

nouns, each noun is represented by a Word2Vec model [5] that was trained particularly on Twitter datasets to deal with noisy tweets. Then 50 nearest neighbor nouns of each of the keywords extracted from the descriptions are found as candidates. From these candidates which are more likely be relief resources during the earthquake, we labeled 86 nouns as keywords manually. However, it was clear that there are more keywords we could not extract, such as Nepali words.

Chunking and Wordnet: One of the basic technique for information extraction, chunking, is used to identify keywords in our approach. We defined some chunk grammar, for example, "*CHUNK: {<NNP.?>*<VB.?>+<DT>?<JJ>*<NN|NNS>+}*" based on tagging by POS in the previous step. Next, the nouns were filtered by Wordnet [6] and specific verbs such as *distribute, give, provide, support* and *hand. Then we enriched the keyword list from filtered nouns manually*.

Table 2. Some numbers of Results in Keywords Extraction phase

Extracted nouns using POS	12236
Extracted keywords from the descriptions	16
Manually extracted keywords using Word2Vec	86
Number of verbs used with Chunking	18
Manually extracted keywords using Chunking	38
Total number of keywords	124

2.2 Retrieval

Once we had a bunch of keywords extracted in the previous phase, we retrieved all tweets (around 8620 tweets) that include at least one keyword using those keywords on the Terrier. There are few open search engines however, we chose Terrier taking some of its advantages into consideration. In term of scoring model, we employed BM25 which is based on probabilistic retrieval framework. The rank and scores are used to compute the relevance of a tweet to a topic in further.

2.3 Classifying into topics

The most of tweets that retrieved in previous phase can be significantly related to the first two topics while some of them cannot. For instance, Even though the following two tweets both includes *water* (a keyword), the former one is related to the first topic what resources were available, whereas the latter tweet is not related to any topic.

Anyone in need of drinking water contact me. Have some can donate #earthquake #Nepal #bhaktapur

#ShameOnYou #nepalgov Rs 20 water cost Rs 40 #earthquakenepal #earthquake #Nepal #fuckoff don't #donate #unknown #website

Therefore we classified the tweets into three classes – available, required and other. In order to do that, first, we annotated 1000 tweets manually. In preprocessing of classification, all URL, usertag and some symbols were removed. Then we employed three classifiers with basic features such as unigram, bag-of-words and some twitter specific features on WEKA[2] open source machine learning software. The best result was executed by SVM (see Table 2).

In term of the third and fourth topic, "What medical resources were available" and "What medical resources were required", we retrieved the relevant tweets from the tweets of the first topic and second topic using medical relief resources respectively.

Table 3. Accuracy Results of Cross-Validation on Training Data

Method	Accuracy
SVM	81.5
MaxEnt	78.9
Naive Bayes	77.2

3 Result

It is impossible to compare our results to other participants results because we submitted the attempts of only three topics to the organizers. However, the results estimated by the organizers was reasonable, which brought us encourage to complete our work. The result is presented in Table 3.

Table 3. Results estimated by the organizers

Run_ID	Precision @ 20	Recall @ 1000	Map @ 1000	Overall MAP
Ganji_1	0,8500	0,4988	0,2204	0,2420

4 Conclusion

In this paper, we have presented our keyword based approach for the four topics of FIRE2016 Microblog Track. Our system is semi-automatic, which includes manual work in the keyword extraction phase. Moreover, the phases are not integrated with each other.

Next, we plan to improve our system to become automatic and to use advanced methods.

5 REFERENCES

[1] M.Imran, S. Elbassuoni, C. Castillo, F. Diaz, P. Meier. Extracting Information Nuggets from Disaster Related Messages in Social Media. *In: Proceeding of the 10th International ISCRAM Conference, 2013*

[2] *A.Ritter, Mausem and O.Etzioni, Open domain event extraction from Twitter. KDD'12 2012.*

[3] J.Piskorski, R.Yangerber Information Extraction: Past, Present and Future. *In: Multi source, Multilingual Information Extraction and Summarisation. 2013*

[4] L. Derczynski, A. Ritter, S. Clarke, and K. Bontcheva. 2013. "Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data". In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, ACL.

[5] Godin, F., Vandersmissen, B., De Neve, W., & Van de Walle, R. (2015). Multimedia Lab @ ACL W-NUT NER shared task: Named entity recognition for Twitter microposts

2 http://www.cs.waikato.ac.nz/ml/weka/

using distributed word representations. *In: Workshop on Noisy User-generated Text, ACL 2015.*

[6] George A. Miller (1995). WordNet: A Lexical Database for English. *Communications of the ACM Vol. 38, No. 11: 39-41.*

[7] 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.