Correlation Distance based Information Extraction System at FIRE 2016 Microblog Track

Saptarashmi Bandyopadhyay
Indian Institute of Engineering Science
and Technology, Shibpur
Howrah 711103, India
saptarashmicse@gmail.com

ABSTRACT

The FIRE 2016 Microblog track provided a set of tweets posted during the Nepal earthquake on April 2015, and a set of seven topics. The challenge was to extract all tweets relevant to each topic. In this method, separate word bags are constructed for each topic describing a generic information need during a disaster situation using topic seed words, stemmers, dictionary and other NLP tools. The topic word bags have been populated with scrambled words that generally appear as noise words in tweet texts. The correlation distance between the topic word bag vectors and each tweet text vector is computed. The correlation distance measure is used to compute the relevance score of each tweet to each topic. Special consideration is taken for the topics that are conditioned on the presence of organization names, location names and Geo locations. Organization names and location names are identified on the crawled tweet texts. The presence of geo locations in the crawled tweets is also identified by the tweet parser. The system response is generated by ordering tweet ids in descending order of their relevance score with respect to each topic. The evaluation scores of the submitted system in terms of Precision@20, Recall@1000, Map@1000 and Overall Map have been reported as 0.4357, 0.3420, 0.0869 and 0.1125 respectively. The evaluation scores of the system without the scrambled noisy words in the word bag vectors in terms of Precision@20, Recall@1000, Map@1000 and Overall Map have been reported as 0.4000, 0.3401, 0.0860 and 0.1119 respectively. The results indicate that the precision of the information extraction system depends on considering the presence of scrambled noisy words in the tweet texts.

CCS Concepts

• Information systems → **Information Retrieval → Information Retrieval Query Processing**

Keywords

FIRE 2016; Microblog Track; Twitter Information Extraction; Vector Model; Correlation Distance.

1. INTRODUCTION

User-generated content in microblogging sites like Twitter are important sources of real time information on various events, including disaster events like floods, earthquakes, and terrorist attacks. The aim of the FIRE 2016 Microblog track [14] is to develop Information Retrieval methodologies for extracting important information from microblogs posted during disasters.

A total of 49,774 tweets that were posted during the Nepal earthquake in April 2015, have been provided as the data for the task along with a set of 7 topics in TREC format. Each topic contains an identifier number, a title, a description and a more detailed narrative which describes the types of tweets that would be considered relevant to the topic. Each of the seven topics identifies a broad information need during a disaster, such as – what resources are available (FMT1), what resources are required (FMT2), what medical resources are available (FMT3), what medical resources are required (FMT4), what were the requirements / availability of resources at specified locations (FMT5), what were the activities of various NGOs / Government Organizations (FMT6) and what infrastructure damage / restoration were being reported (FMT7). The corresponding topic ids have been mentioned within brackets.

The task was to develop methodologies for extracting tweets that are relevant to each topic with high precision as well as with high recall. The main challenges involved with the ad-hoc search task are dealing with the noisy nature of the tweets and identification of specific keywords relevant to each topic. Tweet texts contain maximum of 140 characters and are often informally written using abbreviations, colloquial terms, etc. Each individual tweet text might not contain most of the specific keywords even though the tweet is relevant to a topic.In the present system, the tweet parser parses the tweets and extracts the tweet texts. Organization names and locations names are identified on the crawled tweet texts. The presence of geo locations in the crawled tweets is also identified. Separate word bags are constructed for each topic. The topic word bags have been populated with scrambled words that generally appear as noise words in tweet texts. The correlation distance between the topic word bag vectors and each tweet text vector is computed. The correlation distance measure is used to compute the relevance score of each tweet to each topic. Special consideration is taken for the topics that are conditioned on the presence of organization names, location names and geo locations. The system response is generated by ordering tweet ids in descending order of their relevance score with respect to each topic. The evaluation scores of the submitted system in terms of Precision@20, Recall@1000, Map@1000 and Overall Map have been reported as 0.4357, 0.3420, 0.0869 and 0.1125 respectively. The evaluation scores of the system without the scrambled noisy words in the word bag vectors in terms of Precision@20, Recall@1000, Map@1000 and Overall Map have been reported as 0.4000, 0.3401, 0.0860 and 0.1119 respectively. The results indicate that the precision of the information extraction system depends on considering the presence of scrambled noisy words in the tweet texts. In the present work no attempt has been made to identify duplicate tweets. This policy is in line what is mentioned in the problem statement about weeding out the duplicate tweets.

2. PREPROCESSING

The 49,774 tweets have been made available as json files as part of the FIRE 2016 Microblog track. The tweet parser parses the tweets and extracts the tweet text in the format <S> text </S> in which a new line is included after the tweet text. During preprocessing the string as part of the text attribute in a tweet is parsed. Non ASCII characters present in the tweet text are removed by using a python script. It has been observed that some of the tweets contain non ASCII characters in the tweet text which are not necessary during the vector correlation distance computation. The newline character present in the parsed tweet text is also removed.

The StanfordNER (Named Entity Recognizer) Tagger [1] class available as part of the NLTK 3.0 toolkit [3] has been used on the parsed and preprocessed (after removal of non ASCII characters and new line characters) tweet texts to identify the location and organization names in the tweet texts. A big benefit of the Stanford NER (Named Entity Recognizer) tagger is that is provides us with a few different models for pulling out named entities. We can use any of the following:

- 3 class model for recognizing locations, persons, and organizations
- 4 class model for recognizing locations, persons, organizations, and miscellaneous entities
- 7 class model for recognizing locations, persons, organizations, times, money, percents, and dates

The NLTK toolkit provides a wrapper to the StanfordNERTagger so that it can be used in Python. The parameters passed to the StanfordNERTagger class include:

1. Classification model path
2. Stanford tagger jar file path (has been used in the present work)
3. Training data encoding (default of ASCII encoding has been used in the present work)

In the present work, the 3 class model for English has been used as the Classification model, the Stanford-ner-2015-04-20/Stanford-ner.zip file has been used as the Stanford tagger jar file and default ASCII encoding has been used as the training data encoding. The output tags are obtained as UTF-8 encoding for LOCATION, ORGANIZATION and PERSON Named Entities.

It may be observed at this point that identified location names may not belong to the country of Nepal - the place of disaster while the topic with id FMT5 requires that the availability or requirement of resources are referred at specific locations in the place of disaster. Organization names must be present in tweet texts that look for activities of NGOs / Government Organizations (FMT6). It is observed that the identified organization names may not identify NGOs/ Government Organizations who are working in Nepal - the place of disaster. The situation will have an effect on the precision and recall of the information extraction system. A better alternative would have been the development of a list of Locations names in Nepal and a list of the NGOs or Government Organizations in Nepal.

The crawled tweets in the json files are also checked for the presence of geo locations in the tweets. Geo locations present in a tweet identify the location from which the tweet has been submitted. Geo locations are present in the tweets only when the feature is turned ON before sending the tweets. It is observed that geo locations present in a tweet may not belong to Nepal - the place of disaster. It has been observed that location named entities is not always present in the tweet texts. The presence of location named entities or geo locations are considered for relevance of tweets with respect to topic id FMT5 and organization name entities are considered for the relevance of tweets with respect to topic id FMT6.

The following bags of words are initially created considering the information need of seven topics with topic ids FMT1 though FMT7: available, resources, required, medical, working, relief, infrastructure, damage and restoration. The 'working' and 'relief' bags have been considered to take care of topic id FMT6. The word bags have been identified by analyzing the topic descriptions.

These word bags are created in the following manner: first by looking into the narrations in the FMTs, seed words have been identified. For example, the seed words for available word bag as identified from the FMT are 'available' and 'availability'. Then PyDictionary 1.5.2 [2] has been used to find the synonyms of the seed words and these synonyms have been included in the word bag. PyDictionary uses WordNet corpus but not directly. Next stemming has been carried out using the Porter Stemmer module available in NLTK 3.0 [3] toolkit. Other possible stemmers could have been Lancaster stemmer etc. Next the NodeBox toolkit [4] has been used for generating the surface level inflected forms of the words in the word bags. The library bundles WordNet (using Oliver Steele's PyWordNet [5]), NLTK [3], Damian Conway's pluralisation rules [6], Bermi Ferrer's singularization rules [7], Jason Wiener's Brill tagger [8], several algorithms adopted from Michael Granger's Ruby Linguistics module

[9], Charles K. Ogden's list of basic English words [10], and Peter Norvig's spelling corrector [11]. The words in the word bag have been pluralized and the past forms of the verb words have been generated. Finally by looking into the topic narrations appropriate words in each word bag have been identified which fit into the sense with respect to the particular topic.

Scrambled words like ‘avlbl’ for the ‘available’ word bag have also been added by randomly selecting tweets and looking into the narration. Such noisy words are often present in tweets. A separate set of word bags has also been developed for each topic without the inclusion of the scrambled words as above.

Now, separate word bags have been constructed for each of the topics as shown in Table 1.

3. INFORMATION EXTRACTION SYSTEM

The basic task in the Extraction System is to look for co-occurrence of words corresponding to each topic FMT (words in the topic FMT word bag) and in each tweet text. The objective is to assign a relevance score to each tweet text corresponding to each topic FMT. This can be accomplished by converting each topic FMT word bag and each tweet text into separate vectors. The distance between the two vectors will assign a relevance score to each tweet text corresponding to each topic FMT.

Table 1. Topic Word bag

Topic Id	Description	Topic Word Bag
FMT1	availability of resources	available+ resources
FMT2	requirement of resources	required + resources
FMT3	availability of medical resources	available + medical
FMT4	requirement of medical resources	required + medical
FMT5	availability and requirement of general and medical resources at specified locations	available + required + resources + medical + (occurrence of location Named Entities or geo locations in the tweet text is must)
FMT6	activities of various NGOs / Government Organizations	Working + relief + (occurrence of organization Named Entities in the tweet text is must)
FMT7	infrastructure damage and restoration	infrastructure + damage + restoration

Each word bag for each topic FMT is converted to a vector of 200 dimensions by using the Word2Vec package [13] with w2v.twitter.200d.txt as the model file. Each tweet text is also converted to a vector of 200 dimensions in a similar manner.Word2Vec [13] is a two-layer neural net that processes text. Its input is a text corpus and its output is a set of vectors: feature vectors for words in that corpus. While Word2Vec [13] is not a deep neural network, it turns text into a numerical form that deep nets can understand.

The correlation distance between each tweet vector and each topic FMT word bag vector is computed. Such values are stored in an array for each topic FMT. The *distance correlation* of two random variables is obtained by dividing their *distance covariance* by the product of their *distance standard deviations*. The correlation distance (dCor(u,v)) between two one dimensional vectors u and v is defined as

$$\mathrm{dCor}(u,v) = \mathrm{dCov}(u,v) \;/\; \mathrm{SQRT}(\mathrm{dVar}(u)*\mathrm{dVar}(v)) \quad (1)$$

wheredCov(u,v) is the distance covariance of the two vectors u and v and dVar(u) and dVar(v) are the distance standard deviations of the u and v vectors respectively. It may be noted that the distance values are computed as (1-coorrelation distance) in the *scipy* package spatial distance module [12].

The relevance scores of each tweet text for each topic FMT are calculated in the following way. For topic FMTs 1-4 and 7, relevance score of each tweet text corresponding to each topic FMT is computed as

relevance score = 1 - correlation distance as computed by the spatial distance modul (3)

which can be simplified as

relevance score = actual correlation distance (4)

since correlation distance as computed by the spatial distance module = 1 – actual correlation distance (5).

Relevance scores for each tweet as computed by equations 3-5 above will have a lower value if the tweet is relevant to the corresponding topic FMT. Similarly, relevance scores for each tweet as computed will have a higher value if the tweet is less relevant to the corresponding topic FMT. Hence, the relevance scores of each tweet text for each topic FMT are subtracted from 1 and stored as the final relevance score for each tweet text corresponding to that topic FMT. Thus,

final relevance score = 1 – relevance score (6)

This ensures that relevant tweets corresponding to each topic FMT will have a high relevance score.

For topic FMT5, if no location names have been identified in the tweet text or no geo locations have been identified in the tweet, a score of 0.5 is added to the actual correlation distance score already obtained, otherwise a score of 0.05 is added. The above scores of 0.5 or 0.05 have been considered heuristically. It may be noted that tweets with specific location names are considered relevant to topic FMT5 provided other conditions are satisfied.

relevance score = actual correlation distance + 0.5 (no location names in the tweet text or no geo locations in the tweet)

else

relevance score = actual correlation distance + 0.05 (location names in the tweet text or geo locations in the tweet) (7)

The final relevance score for each tweet corresponding to topic FMT 5 will be computed as in equation 6. The objective is to ensure that the relevance scores for tweets with no location names in the tweet text or no geo locations become small with respect to other topic FMT5 relevant tweets. Similarly, the relevance scores for tweets with location names in the tweet text or geo locations become higher with respect to other FMT5 tweets. It may be noted that tweet texts with no location names or geo locations have not been completely rejected since the named entity identification process in tweet texts using the Stanford NER package may have missed such names.

For topic FMT6, if no organization names have been identified in the tweet text, a score of 0.5 is added to the actual correlation distance score already obtained, otherwise a score of 0.05 is added. The above scores of 0.5 or 0.05 have been considered heuristically. It may be noted that tweets with specific NGOs or Government Organization names are considered relevant to topic FMT6 provided other conditions are satisfied.

relevance score = actual correlation distance + 0.5 (no organization names in the tweet text)

else

relevance score = actual correlation distance + 0.05 (organization names in the tweet text) (8)

The final relevance score for each tweet corresponding to topic FMT 6 will be computed as in equation 6. The objective is to ensure that the relevance scores for tweets with no organization names in the tweet text become small with respect to other topic FMT6 relevant tweets. Similarly, the relevance scores for tweets with organization names in the tweet text become higher with respect to other topic FMT6 tweets. It may be noted that tweet texts with no organization names have not been completely rejected since the named entity identification process in tweet texts using the Stanford NER package may have missed such names.

Next, these relevance scores are sorted and in each topic FMT structure, we get the tweet id and relevance score pair in descending order of final relevance score. Highly relevant tweets are placed high in the list.

The final result is submitted in TREC format as <FMTID><Q0><TWEETID><RANK><RELEVANCE SCORE><RUNID>.

4. SYSTEM EVALUATION RESULTS

The submitted system has been evaluated by the FIRE 2016 Microblog Track organizers in terms of Precision@20, Recall@1000, MAP@1000 and Overall Map. The evaluation scores for each topic have been averaged to generate the evaluation scores for the submitted systems. The evaluation scores of the submitted system in terms of Precision@20, Recall@1000, Map@1000 and Overall Map have been reported by the track organizers as 0.4357, 0.3420, 0.0869 and 0.1125 respectively. The evaluation scores of the system without the scrambled noisy words in the topic FMT word bag vectors in terms of Precision@20, Recall@1000, Map@1000 and Overall Map have been reported as 0.4000, 0.3401, 0.0860 and 0.1119 respectively. The results indicate that the precision of the information extraction system improves if scrambled noisy words are included the topic FMT word bag vectors.

5. CONCLUSION

The submitted information extraction system has only considered the correlation distance measure between the vector representations of the topic word bags and the tweet texts. It would be interesting to consider other vector distance measures such as Cosine Similarity or Euclidean distance. In the present system, additional heuristic scores of 0.5 and 0.05 have been considered to identify and rank relevant tweets for topic ids FMT5 and FMT6. It is necessary that multiple experiments are carried out with different values of these additional heuristic scores and the scores that generate the best results are considered. It has been observed identified location names or geo locations or organization names are not checked for their occurrence in the place of disaster. This has an effect on the performance of the system. A better alternative would have been the development of a list of Locations names in Nepal and a list of the NGOs or Government Organizations in Nepal so that identified location or organization names in tweet texts can be checked in these lists. It has been observed that use of general resource or medical resource ontologies as well as infrastructure ontologies in the preparation of the word bags would have produced more relevant results. Scrambled noisy words have been included in the topic word bags in an ad hoc manner. A better alternative will be to collect such scrambled noisy words from large tweet corpus and the development of a methodology for identifying the scrambled noisy words that can be included in a topic word bag.

6. REFERENCES

[1] nlp.stanford.edu/software/Stanford-ner-2015-04-20.zip

[2] https://pypi.python.org/pypi/PyDictionary/1.5.2

[3] https://pypi.python.org/pypi/nltk/3.0.0

[4] https://www.nodebox.net/code/index.php/Linguistics

[5] https://pypi.python.org/pypi/pywordnet

[6] www.csse.monash.edu.au/~damian/papers/extabs/Plurals.htm

[7] *https://github.com/bermi/Python-Inflector/blob/master/rules/english.py*

[8] *pydoc.net/Python/Pattern/1.5/pattern.en.parser/*

[9] *https://github.com/bruce/linguistics. www.nodebox.net/code/index.php/Linguistics*

[10] ogden.basic-english.org/

[11] *norvig.com/spell-correct.html*

[12] *docs.scipy.org/doc/scipy-0.16.1/reference/generated/scipy.spatial.distance.correlation .html*

[13] https://deeplearning4j.org/word2vec

[14] 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.