A Hybrid Model For Information Retrieval From Microblogs During Disaster

Du Xin

School of Computer Science and
Technology, Heilongjiang Institute of
Technology, Harbin, China

duxin111@outlook.com

Wang Xiaoyu

School of Computer Science and
Technology, Heilongjiang Institute of
Technology, Harbin, China

wongxiaoyu1946@gmail.com

Zhuang Ziyao

Faculty of Science, Agriculture and
Engineering, University of Newcastle upon
Tyne, UK

zhuangziyao1@outlook.com

Qi Limin

School of Physical Sciences, Harbin
Normal University, Harbin, China

qilimin111@outlook.com

Abstract

When the disaster occurs, the social network site such as Twitter is increasingly being used for helping direct rescue operations. This article describes the methods we used in the Fire2017. We regarded the distinction of need-tweets and availability-tweets as classification tasks, and the logistic regression and Support Vector Machine are used to decide the type of the tweets. In the need and availability matching, we regard it as an information retrieval task, using the retrieval model to complete the task.

KEYWORDS

information retrieval, microblog, identify need-tweets and availability-tweets, matching need-tweets and availability-tweets

1. Introduction

Now users on the social networking site to publish real-time content, become the most extensive information sharing, one of the fastest ways to spread information. The published material can contain specific locations, specific needs, especially when the disaster comes, can provide accurate rescue information to guide the rescue efficiently.

For Task 1, we need to judge the need-tweets, availability-tweets and others. We deem it as a classification problem. We use the classifiers to distinguish the useful need-tweets and the availability-tweets from other tweets such as useless tweets, and repeat forwarding tweets.

For Task 2, Using the results identified by Task 1, we implement the matching task. We should match need-tweets with availability-tweets. Also, a need-tweet and a matching availability-tweets can be in different language. we deem this task as an information retrieval problem. For each of the need-tweets in the query collection, retrieve the most matching availability-tweets. In the match between need-tweets and availability-tweets.

2. Task 1

2.1 Problem Description

In Task 1, we judge that the text is need-tweets, availability-tweets, or other using the correlation between text and category. Based on the description of classification problem, the classification of text content includes the expression of the text, the selection and training of the classifier, text preprocessing, feature extraction and so on. We use the training data (x_1, y_1), (x_2, y_2), … , (x_N, y_N) to learn a classifier Y=f(x), so that the classification system can use the classifier for new data to mark. x_i represents the characteristic vector of data. y_i represents the classification system output. y_i labels as $\{0,1\}$.

2.2 Classifier Models

In this working note, two groups adopt SVM, which is a binary model, which is defined as the most significant liner classifier in the feature space. The learning strategy is the interval maximization. Solving the problem into Convex Quadratic Programming Problem. We chose the RBF kernel, which can map the samples non-linearly to a higher dimension. The RBF core has fewer numerical complexity characteristics. Ultimately two groups completed Task 1 with LibSVM.

The third group uses LR model. Each data point of the LR model has an impact on the classification plane. Its influence is far from its distance to the classification plane, and if the data dimension is high, the LR model will match the parameters regularization method. In classification, the computation is minimal, and the speed is breakneck, so the storage resources are low, so it is convenient to observe the probability scores of samples.

2.3 Feature Selection

For feature selection, we conduct experiments with words and n-gram (n = 2, 3, 4, 5) respectively. The results show that we have the best effect of using words as features. Table 1 shows the comparison of the test results of each feature. For the preprocessing of tweet text, each team adopts different methods to preprocess the data. Table 2 specifically describes the different teams in the pre-treatment differences exist. (√ indicates that this method is used, and × is not used)

Note:

HLJIT2017-IRMIDIS_1_task1_1 and HLJIT2017-IRMIDIS_1_task1_2 is different in stop words list.

Table 1: The Result of various characteristic

Feather	Need		Availability	
	Pre	Re	Pre	Re
2-gram	0.281	0.245	0.477	0.474
3-gram	0.482	0.254	0.671	0.597
4-gram	0.428	0.245	0.641	0.606
5-gram	0.385	0.2	0.604	0.556
word	0.517	0.281	0.709	0.644

Table 2: Preprocess tweet text

ID	1_1	1_2	1_3
Remove stop word	√	√	×
Remove @username	√	√	√
Remove punctuation	√	√	√
Remove URL	√	√	√
ID	2_1	2_2	2_3
Remove stop word	√	√	×
Remove @user name	√	√	√
Remove punctuation	√	√	√
Remove URL	√	√	√

Note:

(1) HLJIT2017-IRMIDIS_1_task1_1 = 1_1

(2) HLJIT2017-IRMIDIS_1_task1_2 = 1_2

(3) HLJIT2017-IRMIDIS_1_task1_3 = 1_3

(4) HLJIT2017-IRMIDIS_1_task2_1 = 2_1

(5) HLJIT2017-IRMIDIS_1_task2_2 = 2_2

(6) HLJIT2017-IRMIDIS_1_task2_3 = 2_3

3. Task 2

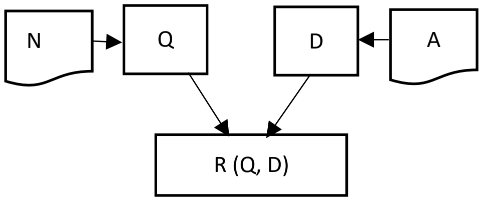
3.1 Problem Description

Task 2 requires that the need-tweets match in Task 1 be searched by Task 1. Need-tweets is used as a query set Q. Availability-tweets can be used as a collection of documents D. We use statistical language models to solve the problem of Task 2. Language models are used to assess what kind of word sequences are more typical according to language usages, and inject the right bias accordingly into the system to prefer an output sequence of words with high probability according to the language model. If a document language model gives the query a high probability, the query words must have top opportunities according to the document language model, which further means that the query words frequently occur in the document.

3.2 Relation

The correlation calculation can be expressed briefly as shown in Figure 1, using the Need-Tweets (N) as the query set Q, the Availability-tweets (A) as the document set D, and then the correlation calculation to obtain the correlation R (Q, D).

Figure 1: Relevance calculation



3.3 Language Model

We use the indri open-source retrieval tool and use the Dirichlet smoothing language model. The formula is as follows:

$$P_s(w|d) = \frac{c(w;d) + \mu p(w|C)}{\sum_w c(w;d) + \mu}$$

$$a_d = \frac{\mu}{\sum_w c(w;d) + \mu}$$

We give a discount to the probability of the word appearing in the document and provide extra value to the likelihood of the word that does not appear in the document.

$$score(q,d) = log\, p(q|d)$$

$$= \sum_{i:c(i:c(q_i:d)>0)} log \frac{p_s(p_i|d)}{a_d p(q_i|c)}$$

$$+ \sum_i logp(q_i|c) + nloga_d$$

$$p(w|d) = \begin{cases} p_{seen}(w|d) & w \in d \\ a_d P(w|C) & oterwise \end{cases}$$

4. Experimental Result

4.1 Data Set

At the start of the track, about 20,000 tweets released (training set), along with a sample of need-tweets and availability-tweets in these 20 K tweets. Later, a new set of 50,000 tweets released (test set).

4.2 Evaluation index

Precision = number of correct messages extracted / number of extracted messages

Recall = the number of correct messages extracted / the number of messages in the sample

$$F - score = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

β is the parameter, P is the precision, and R is the recall.

Map：

$$\hat{\theta}_{ML} = argmax_\theta f(x|\theta)$$

$$\hat{\theta}_{MAP} = argmax_\theta \frac{f(x|\theta)g(\theta)}{\int \ominus f(x|\theta')g(\theta')d\theta'}$$

4.3 Parameter setting

Table 3 describes the parameters of the two tools selected in this article.

Table 3: Parameter setting

SVM	svm_type=c_svc,kernel_type=rbf,
	Gamma=0.1，nr_class=2，
	total_sv=8660，Rho=0.313597

LR	studyRate=0.01，theta = 0.05
	iterNum=1000

4.4 Experimental results

Table 4 shows the experimental results of Task 1, the average Map of LR algorithm is higher than that of two groups of using LibSVM. The Availability-Tweet and Need-Tweets Map values of the two groups using LibSVM showed a contrast. And they all were much higher than the other one, which led to the low Average Map value. The preliminary analysis may be due to improper selection of stop words caused by the occurrence of the above phenomenon.

Table 5 shows the experimental results of Task 2. Task 2 of the input is based on task 1 is output. The precision of the three sets of values was almost the same, but the first group of Recall was much lower than the other two teams. Probably the result of the submission in Task 1 is too much, and there is a mistake on the threshold.

Table 4: Task 1 Experimental Result

	ID	1_1	1_2	1_3
Availability	Precision @100	0.550	0.100	0.760
	Recall @1000	0.100	0.017	0.159
	Map	0.760	0.001	0.112
Need	Precision @100	0.060	0.630	0.610
	Recall @1000	0.021	0.217	0.173
	MAP	0.001	0.140	0.128
Average MAP		0.031	0.071	0.120

Table 5: Task 2 Experimental Result

	Precision@5	Recall	F-score
2_1	0.088	0.021	0.034
2_2	0.088	0.217	0.125
2_3	0.082	0.147	0.105

5. Conclusion

Through the above experimental results, we found that the classification and sorting of text content for some informal occasions are very regarding words. In the future experiments, we will deepen the study of machine learning and try to select more different features, to filter and choose text content of informal occasions.

Acknowledgments

This work is supported by Philosophy and Social Science Planning Project of Heilongjiang Province, China (No. 16EDD05)

References

[1] M. Basu, S. Ghosh, K. Ghosh and M. Choudhury. Overview of the FIRE 2017 track: Information Retrieval from Microblogs during Disasters (IRMiDis). In Working notes of FIRE 2017 - Forum for Information Retrieval Evaluation, Bangalore, India, December 8-10, 2017, CEUR

[2] T. Joachims, Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.

[3] QI Haoliang1,CHENG Xiaolong1, YANG Muyun2, HE Xiaoning3, LI Sheng2, LEI Guohua1. High Performance Chinese Spam Filter.2010, 24(2): 76-84