HLJIT2017-IRMIDIS@IRMiDis-FIRE2017:Information Retrieval from Microblogs during Disasters

Zhao Zicheng
School of Computer Science and Technology, Harbin Engineering University, Harbin, China
zichengzhao888@gmail.com

Ning Hui
School of Computer Science and Technology, Harbin Engineering University, Harbin, China
ninghui@hrbeu.edu.cn

Zhuang Ziyao
Faculty of Science, Agriculture and Engineering, University of Newcastle upon Tyne, UK
zhuangziyao1@outlook.com

Zhao Jinmei
School of Continuing Education, Harbin University of Commerce, Harbin, China
zhaojinmei1@outlook.com

Li Jun
School of Computer Science and Technology, Heilongjiang Institute of Technology, Harbin, China
lijun34667@outlook.com

ABSTRACT

This paper describes the work of HLJIT-IRMIDIS for the Information Retrieval from Microblogs during Disasters. This track is divided into two sub-tasks. Task 1 is to solve the identification problem of need-tweets and availability-tweets during the disaster. Task 2 is to solve the matching problem between need-tweets and availability-tweets. For Task 1, the identification of need-tweets and availability-tweets is formalized into a classification problem. This paper presents a classification method for distinguishing the need-tweets and availability-tweets. For Task 2, the match of need-tweets and availability-tweets is formalized into a retrieve problem. This paper proposes a matching method based on language model. The evaluation shows the performance of our approach, which achieved 0.0687 on MAP in Task 1 and 0.1671 on F-Score in Task 2.

KEYWORDS

Information Retrieval, Microblogs during Disasters, tweets, classification

1 Introduction

Microblogging sites such as Twitter have become important sources of situational information during disaster events [2, 6]. However, dealing with identifying specific tweets and matching relevant tweets are challenging due to micro-blog content is short, contains different language and interference information and so on. The FIRE 2017 Microblog task [1] is motivated by this scenario and aims to promote development of information retrieval (IR) methods to Identifying specific tweets from microblogs posted during disasters. This track is divided into two sub-tasks. Task 1 is called recognition need-tweets and availability-tweets. Need-tweets which inform about the need or requirement of some specific resource. Availability-tweets which inform about the availability of some specific resources. Task 2 is called Matching need-tweets and availability-tweets. Participants' goals are to match need-tweet and availability-tweet. The goal of the participants is to push multiple availability-tweets for a need-tweet.

For Task 1 is considered as a classification problem in this paper. We selected three classifiers, AdaBoost [3], SVM [4] of linear kernel and SVM of nonlinear kernel to resolve this problem, denoted as AdaBoost (task1_2), SVM-L(task1_1) and SVM-NL (task1_3). For the feature of the classifier, this paper presents a feature selection method based on the logistic regression. For Task 2, this paper deems it as a retrieval problem. The need-tweets is used as a query and the retrieval model is used to retrieve the most matching documents with need-tweets in the document collection composed of availability-tweets. The evaluation scores of our best submitted in terms of Overall Map and F-score have been reported as 0.0687 and 0.1671 respectively on IRMiDis Fire2017 dataset.

2 Method of Task 1

Intuitively, Task 1 can be viewed as a two-category classification. If we formalize Task 1 of recognition tweet as a classification problem, our objectives focus on answering the following two questions: (1) Which classification-based methods can effectively be applied to the recognition tweet, and (2) which features should be used in the classifier.

2.1 Method Selection

For classification tasks $\mathrm{D} = \{(\mathrm{x}_1, \mathrm{y}_1), (\mathrm{x}_2, \mathrm{y}_2), \cdots, (\mathrm{x}_\mathrm{m}, \mathrm{y}_\mathrm{m})\}$, $\mathrm{y}_\mathrm{i} \epsilon \{0,1\}$, where x_i is a feature vector and y_i is a feature label. IRMiDis Fire2017 submitted three groups of run. We use AdaBoost, SVM-L and SVM-NL classifiers to predict need-tweets and availability-tweets, respectively.

For task1_1, we use the SVM-L classification model. The principle of the model is to classify the data using the hyperplane. The distance from the positive sample point to the hyperplane as the sorting result.

For task1_2, we use Adaboost, which is a family of algorithms that can enhance weak learners to strong learners. The working mechanism of the classifier is to start from the initial training set training at a base learner, according to the performance of the base learner to the training sample distribution of new adjustments. In the previous course, the training samples of the wrong learners received more attention in the follow-up, and then the next-based learner was trained based on the adjusted sample distribution. A probability value with a positive probability greater than 0.5 is used as the sorting result.

For task1_3, we use SVM-NL. The classification principle is to use the inner product kernel function instead of the high-dimensional space to the non-linear mapping of positive and negative examples of separation. During the test, the classifier generates a prediction probability for the positive case. We use the probability value as the sorting result.

2.2 Feature Selection

Content-based microblogging filtering method, affecting a microblogging is need-tweets or availability-tweets factors are the features of the microblogging. For content-based filtering methods, words are natural features. For the Fire2017 task, we applied the logistic regression model to select 1116 disaster-related words as microblogging features. Feature words can filter out the noise word, but also improve the classification efficiency of the classifier. In this paper, the weight of the feature in the feature library is updated by the method of gradient descending. Using the gradient descent method, select the appropriate feature learning rate to ensure the appropriate learning rate. Table 1 shows the top 20 features.

Table 1: Feature of top20

No	Term	No	Term
1	राहत	11	relief
2	Anyone	12	planes
3	Ambulance	13	meals
4	NEA	14	Doctors
5	supplying	15	Hospital
6	medical	16	electricity
7	send	17	packets
8	Food	18	blood
9	pitched	19	चीन
10	emergency	20	भेजे

By analyzing the selected keywords, we found that medical, doctors, blood, hospital, ambulance and so on for medical information. Relief, electricity, food and meals are people's living security items. The extracted words can represent information about the microblogging in the disaster.

3 Method of Task 2

According to the description of Matching need-tweets and availability-tweets, we formalize the problem as follows. Denote a retrieval problem as IR = $\left(Q, D, F, R(q_i, d_i)\right)$, where Q is need-tweet and D is availability-tweet, F is the rule that satisfies the relevance sorting model, $R(q_i, d_i)$ for query q_i and document d_i relevance. Where$q_i$and d_i are predicted need-tweet and availability-tweet in Task 1. The open source retrieval tool indri[1] is used in Task 2. We use the language model based on the Dirichlet [5] smoothing and select the KL distance as the sorting model. The language model based on Dirichlet smoothing and the KL distance sorting model are defined as follows:

$$KL(Q|D) = \sum_{w} P(w|Q) log \frac{P(w|Q)}{P(w|D)} \qquad (1)$$

where Q is query model, D is document model, we would compute an estimate of the corresponding Q and D, and w is the set of all the words in vocabulary.

$$P(w|D) = \frac{c(w, D) + \mu P_{ml}(w)}{|D| + \mu} \qquad (2)$$

where $P_{ml}(w)$ is language model and μ is a smoothing parameter.

4 Result

We begin this section by summarising details of the dataset, performance measures, experimental settings, and then describe our experiments result.

4.1 Data

This section describes the dataset provided to the shared task participants. 20000 training data with answer and 50000 testing data was provided by the organizers during the Nepal earth-quake in April 2015.

4.2 Performance Measures

For Task 1, evaluation is Mean Average Precision (MAP) considering the retrieved ranked list. For Task 2, evaluation is F-Score. F-Score = 2 * Precision@5 * Recall / (Precision@5 + Recall). Precision@5, i.e., for each need-tweet that is correctly identified. Recall, i.e., what fraction of overall need-tweets could be correctly matched by at least one availability-tweet.

4.3 Experimental Settings

Pre-processing: remove punctuation, URL and mention. Parameter selection of feature selection: learning rate =

[1] http://www.lemurproject.org/indri/

0.004. Parameter settings for the classifier: the parameters of each classifier are shown in Tables 2.

Tables 2: Parameter Settings

Method	Parameter
SVM-L	kernel=linear, loss=squared_hinge, multi_class=ovr, penalty=l2, tol=0.0001
Adaboost	n_estimators=100, algorithm=SAMME.R, LearningRate=1.0
SVM-NL	kernel=rbf, gamma=auto, probability=true, classweight=12

4.4 Result of Task 1

Table 3 shows the experimental results of Task 1.

Tables 3: Results of Task 1

Submission Detail		Availability-Tweets Evaluation			Need-Tweets Evaluation			Average map
No	Run ID	Precision @100	Recall @1000	Map	Precision @100	Recall @1000	Map	MAP
1	HLJIT-IRMIDIS_task1_3	0.5400	0.1878	0.0905	0.3500	0.1405	0.0468	0.0687
2	HLJIT-IRMIDIS_task1_2	0.7100	0.1276	0.0798	0.3900	0.0913	0.0468	0.0633
3	HLJIT-IRMIDIS_task1_1	0.2300	0.1633	0.0493	0.0200	0.1194	0.0079	0.0286

From the experimental results, we can see that the Run2 achieves higher Precision@100 than others. For Run2, we submitted 73 Need-Tweets and 216 Need-Tweets, so Recall@1000 is lowest. However, too many negative examples may lead to Recall@1000 of three groups result is too low in the training model.

4.5 Result of Task 2

Table 4 shows the experimental results of Task 2.

Table 4: Results of Task 2

Run ID	Precision @5	Recall	F-Score
HLJIT-IRMIDIS_task2_1	0.1819	0.1546	0.1671
HLJIT-IRMIDIS_task2_3	0.2033	0.1405	0.1662
HLJIT-IRMIDIS_task2_2	0.2051	0.0913	0.1264

From the experimental results, we can see that the results of Run1 and Run3 are similar on the F-score.

5 Conclusion and Further Work

We have described our approach to all of the tasks in the context of IRMiDis fire2017 competition. The evaluation shows the performance of our approach, which achieved Map (0.0687) in Task 1 and F-Score (0.1671) in Task 2. As a future work, we work like to explore deep learning to text matching and information retrieval of the tweets. Meanwhile, also includes finding new filtering techniques and parameters to tackle such informally written documents like tweets.

Acknowledgments

This work is supported by the Social Science Fund of Heilongjiang Province of China (No. 16XWB02)

Reference

[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 Workshop Proceedings. CEUR-WS.org, 2017.

[2] Imran M, Castillo C, Diaz F, et al. Processing social media messages in mass emergency: A survey [J]. ACM Computing Surveys (CSUR), 2015, 47(4): 67.

[3] Rätsch G, Onoda T, Müller K R. Soft margins for AdaBoost [J]. Machine learning, 2001, 42(3): 287-320.

[4] Cortes C, Vapnik V. Support vector machine [J]. Machine learning, 1995, 20(3): 273-297.

[5] MacKay D J C, Peto L C B. A hierarchical Dirichlet language model [J]. Natural language engineering, 1995, 1(3): 289-308.

[6] Vieweg S, Hughes A L, Starbird K, et al. Microblogging during two natural hazards events: what twitter may contribute to situational awareness [C]//Proceedings of the SIGCHI conference on human factors in computing systems. ACM, 2010: 1079-1088.