Medweb Task: Identify Multi-Symptoms from Tweets Based on Active Learning and Semantic Information

Chao Li
Faculty of Engineering
Tokushima University
2-1 Minami Josanjima,
Tokushima, 770-8506, Japan
c501447002@tokushima-u.ac.jp

Xin Kang
Faculty of Engineering
Tokushima University
2-1 Minami Josanjima,
Tokushima, 770-8506, Japan
kang-xin@is.tokushima-u.ac.jp

Fuji Ren
Faculty of Engineering
Tokushima University
2-1 Minami Josanjima,
Tokushima, 770-8506, Japan
ren@is.tokushima-u.ac.jp

ABSTRACT

Recently, the web have an ever increasing number of medical or clinical related information. Among all the data sources, social media is the most valuable. Ntcir-13 Medweb (medical Natural Language Processing for Web documents) releases the task which identifying the patient symptoms from text. This task exploits pseudo twitter messages as a cross language corpus covering Japanese, English, and Chinese. This paper focuses on the task in Chinese. To finish this task, Active Learning and Semantic Information are exploited in the experiments. Active Learning method is used to find out the new massage which is difficult to be discriminated. With the message from Weibo, the new message are labeled and added to the training data gradually. Word embedding is used as Semantic Information and used to complement the features for each massage. According to the experimental results, the proposal method outperforms other methods in terms of recall. And the overall performance approximates to the baseline. The results also show that the additional training data used in the experiments can only increase recall for this task and the semantic information based on word embedding can increase the overall performance.

Team Name

TUA1

Subtasks

MedWeb (Chinese)

Keywords

Classification, Natural language processing, Twitter, symptom, Weibo, Active learning, Word embedding

1. INTRODUCTION

Paper media are now replacing by electronic media, which are used in the medical reports. Therefore, it become more and more important that using the information techniques to process the electronic data in medical fields. To assist precise and timely treatments, the knowledge obtained from the medical reports is necessary by processing large amounts of them. To promote developing practical tools to support medical decisions, Ntcir13 organizes the MedWeb task. The NTCIR-13 MedWeb task aims to mine the symptom information from the social media data [1].

In this paper, our goal is to find out practical methods to predict the symptoms with the social media data for the Medweb task. In the experiments, Active Learning and Semantic Information are investigated to finish the task. Active Learning method is used to to find out the new data which are difficult to be classified. Gradually, the new message are labeled and added to the training data with the message from Weibo. Word embedding is used as Semantic Information and used to complement the features for each massage. The difference between the proposal method and the baseline method is adding the new data for training classifier and adding the semantic feature for representing the data.

For example, there are the labeled massages and the unlabeled massages at the beginning. By using the active learning method, a specific amount of massages is selected from the unlabeled massages, which are the most difficult to be classified. And then, the selected massages are labeled manually. By looping the above steps, the key messages are found and labeled early to improve the classifier fast. After representing the messages with bag-of-word method, the features based on the word embedding are added to the feature vectors for each massage. Finally, the machine learning method is used to training the classifier and predict the symptoms for the new massages.

According to the experimental results, the proposal method outperforms other methods in terms of recall. And the overall performance approximates to the baseline. The results also show that the additional training data used in the experiments can only increase recall for this task and the semantic information based on word embedding can increase the overall performance.

The remainder of this paper is organized as follows. Section 2 introduces reviews some related work. Section 3 introduces the proposed method. Section 4 presents the details of the experiments. Finally, Section 5 concludes our work and outlines the future work.

2. RELATED WORK

Cardoso et al. (2017) proposed the ranked batch-mode active learning method [2]. In their study, the proposal method used the model to gradually generate the list of unlabeled samples that ranked by discrimination in ascend. And then, these selected samples were labeled by the expert to update the training dataset. Based on this idea, updating the dataset with unlabeled data is exploited in this paper.

Sohrab et al. (2015) proposed an approach to Infusing Word Embeddings into Features for Text Classification [3]. Combining with the TF-IDF method, the new features were used to represent the texts, which were extracted based on the word embedding model. In this paper, this method is used to complement the semantic information for the posts from Twitter and Weibo.

3. METHODS

3.1 Overview

Figure 1 shows the work flow of the proposal method. There are two groups of the steps in the work flow. The upper part is the steps with the active learning method. The under part is the steps with the semantic information for extracting the features. The red blocks are the main methods used in this study.

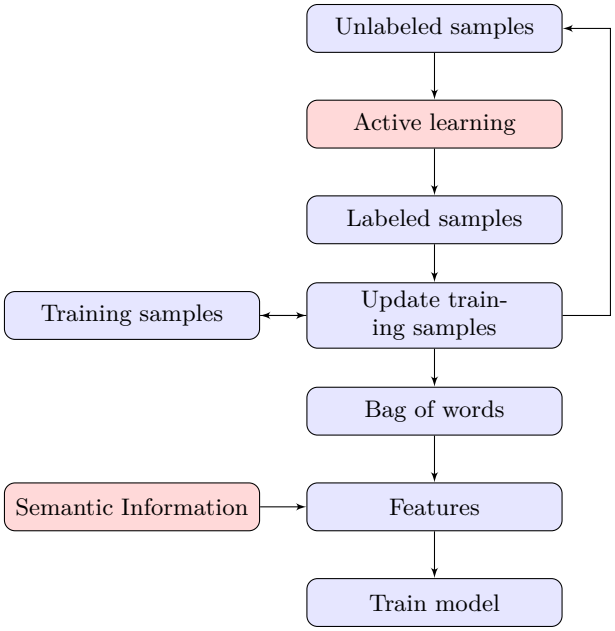

Figure 1: Work flow

The active learning method is used to select the key samples from the unlabeled samples, label the unlabeled samples and add the new samples for training model. Specially, these steps are looped to label the unlabeled samples, and the training samples are updated with the new samples in each loop. And then, the bag of words (BOW) method is used to represent the features. The semantic information is used to extract the semantic features for each samples. Finally, the model is trained by combining the BOW features and the semantic features.

3.2 Active Learning

Figure 2 shows the steps of the active learning based on Cardoso et al.'s method [2].

In these steps, the key step is using the model to generate the ranked list of the samples for labeling. The ranking is based on the prediction for the unlabeled samples by the models. Because there are eight symptoms in the task, some number of the models are trained with the training

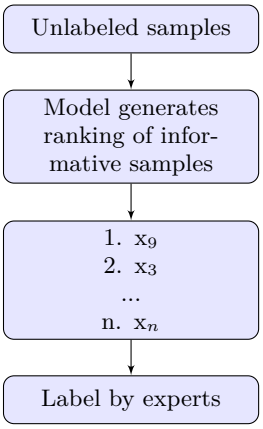

Figure 2: Active learning steps.

samples to selected the unlabeled samples for each classes (symptoms). And then, these models are used to predict the probabilities that the samples belongs to each classes. With the probabilities, the confidence entropy is computed for each samples by equation 1 and 2.

$$confid\_entropy(x_i) = -\sum probs_{norm} * \log probs_{norm} \tag{1}$$

$$probs_{norm}(x_i) = probs \times (\frac{1}{\sum probs}) \tag{2}$$

With the features, the representation score is computed for each samples by equation 3 and 4, where $|C|$ is the number of classes and *Euclidean* is the Euclidean distance.

$$rep\_score(x_i) = \frac{rep\_distances(x_i)}{|C|} \tag{3}$$

$$rep\_distances(x_i) = Euclidean(x_i, x_{j \neq i}) \tag{4}$$

With the confidence entropy and the representation score, the confidence-representation-score is computed by equation 5 and 6, where w is the uncertainty weight and $Mean(\cdot)$ compute the mean value with all the samples.

$$confid\_rep\_score(x_i) = w \times confid\_entropy + rep\_score \tag{5}$$

$$w = \frac{Mean(rep\_score(X))}{Mean(confid\_entropy(X))} \tag{6}$$

After computing the confidence representation score, it computes the euclidean distance between the unlabeled samples and the training samples, and select the minimum value as the variety score for each unlabeled samples. With the variety score, the minimum value is selected from the concatenation of the variety score and the *rep_distances* in equation 4, and then this minimum value add the confidence-representation-score as the confidence representation variety score. Finally, the samples are ranked by the confidence representation variety score in descent and selected with specific number.

3.3 Semantic Information

Figure 3 shows steps of extracting the semantic information from the texts based on word embedding model according to Sohrab et al.'s work[3].

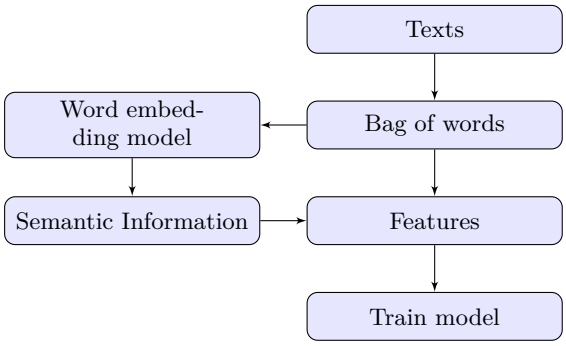

Figure 3: Extract semantic information based on word embedding model

Commonly, the texts are represented by a feature vectors based on bag-of-word and TF-IDF method. Therefore, with all the feature vectors, they form a matrix as **T**, the rows represent the texts, and the columns represent the features. The size of rows equals the number of the texts, and the size of columns equals to the number of the words.

By using the word embedding, the algorithm generates a matrix, as **E**, in which the rows represent the words and the columns represent the features. The size of rows is the number of the words from the training texts, and the size of the features is specified by the algorithm setting.

To extract the semantic information **S** from the texts with the word embedding model, the matrix **T** produces the matrix **E** as the equation 7.

$$\mathbf{S} = \mathbf{T} \times \mathbf{E} \tag{7}$$

Finally, the matrix **T** and the matrix **S** are combined together to represent the texts and train the classification model.

For example, the size of the matrix **T** is 10x100, and the size of the matrix **E** is 100x50, which means 10 texts, 100 words, and 50 embedding size. The **S** will be the matrix as 10x50, and the final result will be the matrix as 10x150. And then, the final matrix is used to train classifier and predict new samples.

4. EXPERIMENTS

4.1 Data and Tools

In the experiments, there are three groups of dataset, the Twitter dataset, the Weibo dataset, and the Wikipedia dataset in Chinese. The Twitter dataset is provided by the Medweb task with 1920 training texts and 640 test texts, which is used for training classifier and predict the symptoms [1]. In the corpus, each massage is labeled with 8 labels (True/False), and each label corresponds one symptom. The Weibo dataset is downloaded by searching posts with the symptom names in Chinese[1], which is used as the unlabeled data to complement the training data by the active learning method. The additional corpus in Chinese is used to training the word embedding model because its size is large.

[1]https://weibo.cn

Table 1: The symptom names in Chinese

English	Chinese
Influenza	流感
Diarrhea	腹泻, 痢疾, 闹肚子, 拉肚子
Hayfever	花粉症
Cough	咳嗽, 痰, 止痰
Headache	头疼, 头痛
Fever	退烧, 低烧, 发烧, 高烧
Runnynose	鼻塞, 鼻涕
Cold	伤风, 感冒
Keywords in training texts	登革热, 腮腺炎

In the experiments, 3 toolkits are employed, including Scikit-learn[2], Gensim[3], NLPIR[4]. The Scikit-learn package is used to train the models including the active learning model and classifiers for predicting the symptoms. The Gensim package is used to train the word embedding model. The NLPIR is used to segment the texts. The Anaconda[5] is used as the experimental platform.

4.2 Details

To prepare the unlabeled data, the related posts are download with the symptom names in Chinese as showed in table 1. Because the average length of the training texts less than 80, the texts larger than 80 are removed in the processing. Other steps include a lot of regex, removing non-Chinese, deduplication, segmenting and checking manually. After the preprocessing, 12200 massages are prepared as the unlabeled data.

To generate the ranked list of unlabeled texts for labeling, 8 logistic regression classifiers are trained with the training data corresponding to each symptom. In each loop of the active learning steps, 1000 unlabeled texts are selected for labeling manually, and Annotation Guideline is provided by the Medweb task[6]. Totally, 7000 unlabeled texts are labeled steps by steps in the experiments. Finally, 8920 texts are fed to training the classifiers.

To predict the symptom for the new massages, three groups of the experiments are performed:

- TUA1-zh-1: bag-of-word features + Logistic Regression model
- TUA1-zh-2: bag-of-word features + SVM (support vector machine)
- TUA1-zh-3: bag-of-word features + Semantic information + Logistic Regression model

4.3 Results

Figure 4.2 shows performances of the proposed methods and the baseline methods according to the overview [1]. At

[2]http://scikit-learn.org/
[3]https://radimrehurek.com/gensim/
[4]https://github.com/NLPIR-team/NLPIR
[5]https://www.anaconda.com/
[6]http://mednlp.jp/medweb/NTCIR-13/#task

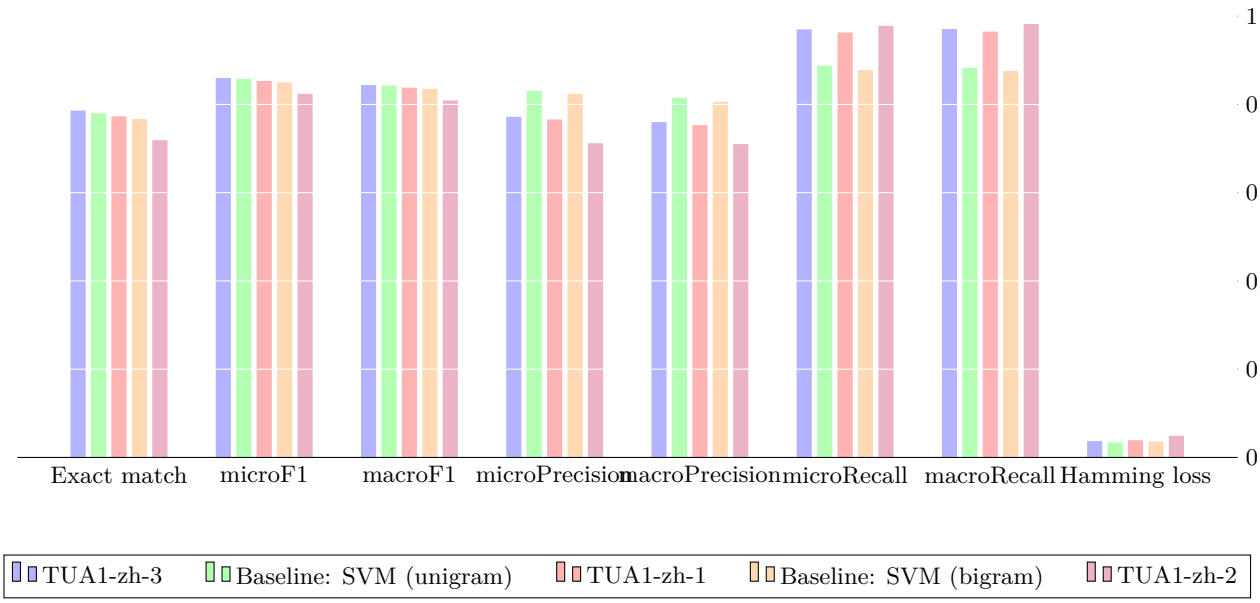

Figure 4: Performances

the Exact match score and the F1 score, the best results are from the third group of the experiments which employ the word embedding model to complement the semantic information. This indicates that the semantic information from the additional corpus could improve the performance for the task. At the precision, the baseline methods outperform the proposed methods, and the results from the third group of the experiments also outperform other groups. At the recall, all the experimental groups outperform the baseline methods. According to the recall and precision, it indicates that the proposed method could improve the recall, but reduce the precision. And the semantic information could improve performance for all the experiments. Therefor, the new training texts reduce the performance, which are selected by active learning method and labeled manually.

4.4 Analysis

In order to analyze the effect of the different size of the additional training samples, figure 5 - 10 are drawn corresponding to MicroF1, Micro precision, Micro recall, MacroF1, Macro Precision and Macro recall. In each figure, the x-axis represents how much additional data is used in the experiments which are selected by active learning and labeled manually, and the y-axis is the value of each metric.

These figures show that the jitter is obvious, especially at the beginning of the curves. Meanwhile, the jitter also becomes smaller gradually, along with increasing the selected massages. And the micro and macro recall have the obvious growth trend.

To analyze the jitter, the average length of the massages in the corpora is computed as showed in table 2. It shows the length of the training and test data is similar. But the length of the selected data and the task data is very different. Also, the jitter also becomes smaller gradually, along with the average length of the the selected data reducing. Note that, the 1st and 2nd average length of words are 3 times the length of the task data. The length is possible to affect the experimental results. The 2nd possible reason for these results is that the method did not select the suitable massages for labeling manually, because the downloaded Weibo massages are not suitable for the task, for example many of them are too long. The 3rd possible reasons is incorrectly labeling the unlabeled massages, because the lack of experience and misunderstanding the labeling guideline for the massages which are selected by active learning method, especially at the beginning.

5. CONCLUSIONS

In this paper, our goal is to find out practical methods to predict the symptoms with the social media data for the Medweb task. In order to complement the amount of the samples, the active learning method is used to rank the unlabeled data with uncertainty, and then the selected data is labeled manually to update the training data. To rich the information of the training data, the semantic information based on the word embedding model is used to complement the features for the data.

The results show that the semantic information could improve the performance for the task. With the active learning method, adding the labeled data could improve the recall for the task, however, this also reduce the precision.

In the future, the reason of reducing the precision would be investigated, for example using the samples with similar length to the task samples. The new ML methods will be also need to be investigated in the task. And the specific properties about the kind of the data in the task will be studied.

6. ACKNOWLEDGMENTS

This research has been partially supported by JSPS KAKENHI Grant Number 15H01712.

Table 2: The average length of the massages in each corpora

Corpora	Avg. chars	Avg. words
Training(original)	15.6	10.53
Test	14.5	9.6
1st 1k massages	44.8	32.2
2nd 1k massages	37.1	36.3
3rd 1k massages	36.9	26.1
4th 1k massages	31.1	21.2
5th 1k massages	28.5	20.1
6th 1k massages	24.9	17.7
7th 1k massages	24.1	16.9

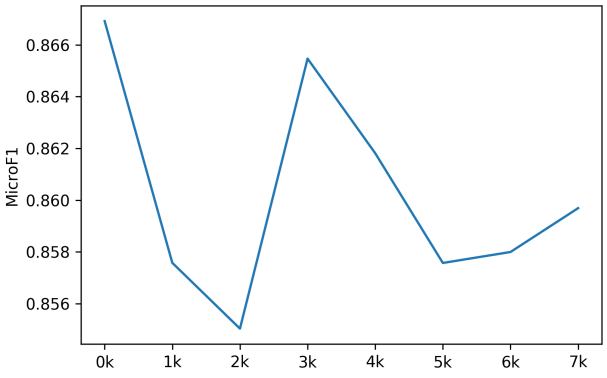

Figure 5: The MicroF1 curve with different size additional training samples

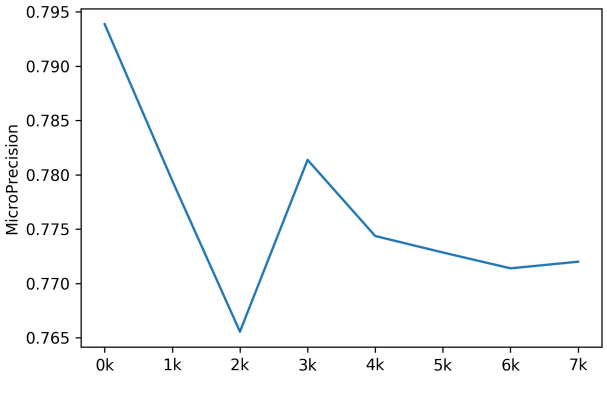

Figure 6: The MicroPrecision curve with different size additional training samples

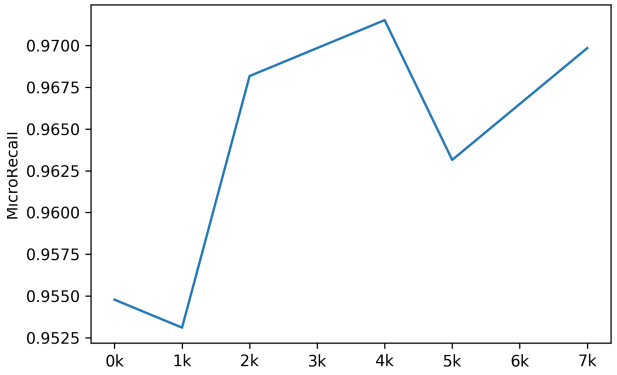

Figure 7: The MicroRecall curve with different size additional training samples

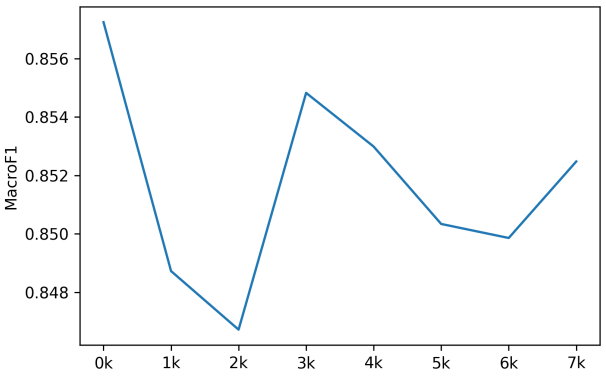

Figure 8: The MacroF1 curve with different size additional training samples

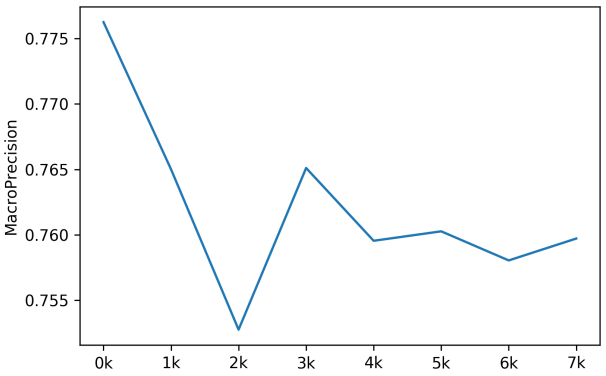

Figure 9: The MacroPrecision curve with different size additional training samples

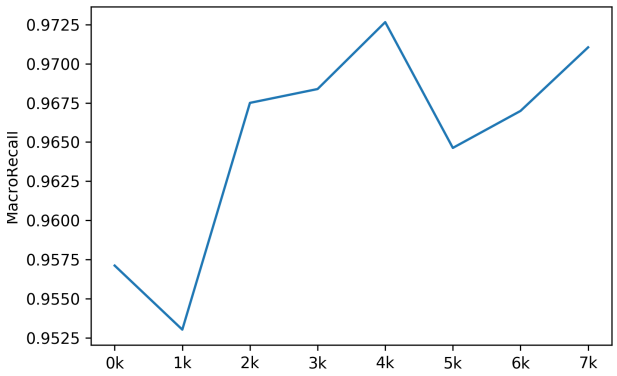

Figure 10: The MacroRecall curve with different size additional training samples

7. REFERENCES

[1] E. Aramaki, S. Wakamiya, M. Morita, Y. Kano, and T. Ohkuma. Overview of the ntcir-13: Medweb task. In *In Proceedings of The NTCIR-13 Conference*, 2017.

[2] T. N. Cardoso, R. M. Silva, S. Canuto, M. M. Moro, and M. A. Gonçalves. Ranked batch-mode active learning. *Information Sciences*, 379(Supplement C):313 – 337, 2017.

[3] M. G. Sohrab, M. Miwa, and Y. Sasaki. *Centroid-Means-Embedding: An Approach to Infusing Word Embeddings into Features for Text Classification*, pages 289–300. Springer International Publishing, Cham, 2015.