

# Classification of Tweet Posters for Diseases by Combined Rule-Based and Machine Learning Method in NTCIR-13: MedWeb Twitter Task (Japanese Subtask)

Masahito Sakishita

Faculty of Informatics, Shizuoka University  
msakishita@kanolab.net

Yoshinobu Kano

Faculty of Informatics, Shizuoka University  
kano@inf.shizuoka.ac.jp

## ABSTRACT

We propose methods which automatically classify Japanese tweet posters who may have diseases (positive,  $p$ ) or not (negative,  $n$ ). Our methods combine a rule-based method and a machine learning method. Our rule-based method is derived from our observation on the training data. Our machine learning method uses SVM (Support Vector Machine) with word level features. Our system achieved 0.802 of the  $p/n$  exact match rate and 0.871 of F1 score, better than the baseline system of NTCIR-13 MedWeb. In the F1 score, our best configuration is ranked third of all participating teams of this task. We found that individual methods and their combinations have different advantages. Other combinations which we have not tried could raise the accuracy rate in future.

## Team name

KIS

## SubTask

MedWeb Twitter task (Japanese subtask)

## Keywords

Classification, Natural language Processing, Twitter, Rule-base, SVM

## 1. INTRODUCTION

We propose methods which automatically classify Japanese tweet posters who have diseases (positive,  $p$ ) or not (negative,  $n$ ) on Twitter<sup>1</sup>. The diseases and the criteria of their recognitions are defined by NTCIR-13 MedWeb organizers [1]. In this task, target diseases are *influenza*, *hay fever*, *diarrhea*, *cough*, *headache*, *fever*, *runny nose* and *cold*. We give some examples of  $p/n$  classification here: a tweet “The *cold* makes my whole body weak. (風邪で体がだるいのだよね。)” is *cold*  $p$  and  $n$  for other diseases; another tweet, “Dogs also have stuffy nose. (犬も鼻づまりになるんだ。)” is  $n$  for all of the diseases.

In previous studies, we classified patients for diseases using electronic health records (EHRs) in MedNLPDoc of NTCIR-12 [2], where we participated in [3]. We assigned ICD-10 codes given an electronic health record written in Japanese. ICD is an abbreviation for International Codes for Diseases, which is defined by World Health Organization (WHO)<sup>2</sup>. It is required to

see whether the patient is diagnosed by the doctor or not based on the given her in this task. Our system tries to find diagnosis names and then assigns ICD codes. The MedWeb task has three differences from the MedNLPDoc task. The first difference is that possible types of the diseases are limited to eight kinds in the MedWeb task. Next difference is that the target corpus is changed to tweets in Twitter from EHRs. The third difference is that target people are changed to posters of tweets and the people around the posters.

Our methods are combinations of a rule-based method and a machine learning method. Our machine learning method uses SVM (Support Vector Machine) based on word level features. We propose two kinds of combination configurations, showing our experimental results.

Section 2 explains details of our methods, Section 3 shows experimental configurations, which result is shown in Section 4. We discuss our methods and results in Section 5, our future work in Section 6, and we conclude this paper in Section 7.

## 2. METHOD

We propose a rule-based method and a machine learning method to solve this task. The  $p/n$  classification can be done by the individual method alone.

### 2.1 Rule-Based Method

This method is grounded on our observation of all training data which is distributed by the MedWeb organizers. The training data includes 1920 Japanese tweets.

First of all, in order to find out the diseases mentioned by the tweets, we use the morphological analyzer Kuromoji<sup>3</sup> implemented in Java. If the morpheme of a specific disease is included in the morphological analysis’s result, our system could notice the corresponding diseases mentioned in this tweet. In the morphological analysis, we use our own user dictionary which includes Wikipedia entry words and the disease name entries of the standard disease name master [4].

Next, we focus on six types of expressions or words for the  $p/n$  classifications of diseases mentioned by the above morphological analysis.

- 1) Words expressing close relatives  
These expressions contains 43 kinds of family nicknames such as “dad (パパ)” and “older brother (兄)”.
- 2) Recovery expressions  
We use eight expressions such as “healed (治った)”.
- 3) Suspicious expressions  
We use nine expressions such as “it seems that... (っぽい)”.

<sup>1</sup> <https://twitter.com>

<sup>2</sup> <http://www.who.int/classifications/icd/en/>

<sup>3</sup> <https://www.atilika.com/>

- 4) Hearsay expressions  
We use four expressions such as “hear that... (ようだ)”, “news (ニュース)” and so on. Regarding the “news”, tweets with the word “news” are mostly intended to inform the news came from others to his/her twitter followers. Therefore, the “news” is added to the hearsay expressions.
- 5) Prevention expressions  
We choose seven expressions such as “hand wash (手洗い)”, “vaccine (ワクチン)”.
- 6) Words representing creatures  
“dog (犬)”, “bird (鳥)” and “bee (蜂)” are manually registered from the MedWeb training data.

Finally, we assumed dependencies between disease labels: if *influenza* is *p*, then *fever* is also *p*; if *hay fever* is *p*, then *runny nose* is also *p*. We assign *p/n* classifications using these rules.

We show the steps of our *p/n* classifications in Figure 1 and Figure 2. Figure 1 shows that the steps of our *p/n* classifications for tweets that mention *influenza* only. Figure 2 shows the steps of our *p/n* classifications for tweets that mentions a disease except for *influenza* (*hay fever, diarrhea, cough, headache, fever, runny nose and cold*)

For a tweet that mentions *influenza*, our system first checks expressions of 1) explained above (Step 1 in Figure 1). If there are expressions of 1) found in the tweet, our system checks the expressions of 2) (Step 2 in Figure 1). If there are the expressions of 2) in the tweet, we classify the tweet as *influenza n*. If they are not included, we classify the tweet as *influenza p*. If expressions of 1) are not in the tweet (Step 1 is false), our system checks other expressions of 2), 3), 4), 5) or 6) in the tweet (Step 3 in Figure 1). If these expressions are found in the tweet, we classify the tweet as *influenza n*. If these expressions are not found, we classify the tweet as *influenza p*.

In a tweet that mentions a disease except for *influenza* (*hay fever, diarrhea, cough, headache, fever, runny nose and cold*), our system first checks the expressions of 1) and 3) (Step 1 in Figure 2). If these expressions are found in the tweet, we classify the tweet as *p* for the corresponding disease. If they are not found in the tweet, we classify the tweet as *n* for the corresponding disease. Next, our system checks the expressions of 4), 5) and 6) in the tweet (Step 2 in Figure 2). If we found these expressions in the tweet, we classify the tweet as *n* for the target disease. If they are not found in the tweet, we classify the tweet as *p* for the target diseases.

Figure 1. The steps of *p/n* classifications for tweets that mention *influenza* only

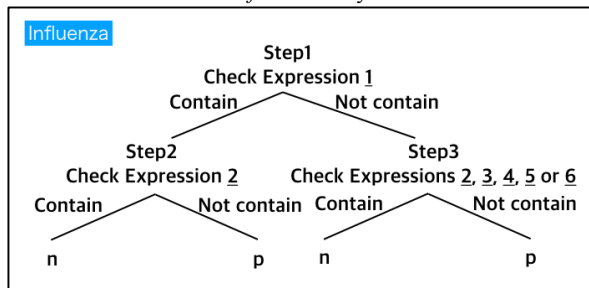
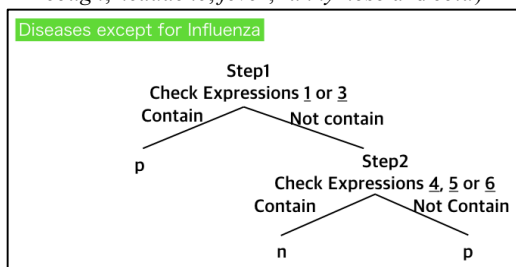


Figure 2. The steps of our *p/n* classifications for tweets that mention a disease except for *influenza* (*hay fever, diarrhea, cough, headache, fever, runny nose and cold*)



## 2.2 Machine Learning Method

This section describes our machine learning method. We complement words necessary for discriminations that could not be picked up by our rule-based method explained in the previous sections.

First, we perform a morphological analysis same as the rule-based method. At this time, a morpheme in the entry of the user dictionary is converted into a one-hot vector. From this feature vector, we create a classifier for each target disease by SVM using libSVM<sup>4</sup>. The training data includes 1536 tweets, which is 80% of the total training data distributed from the MedWeb organizers. The development data for evaluation at the training time includes 384 tweets, which is 20% of the total training data.

## 3. EXPERIMENT

We describe combinations of our methods introduced in Section 2. In the formal run, we use the distributed test data (640 tweets) for our experiment.

Configuration 1) The first configuration uses the rule-based method and the SVM method individually for each disease. For *influenza, fever and cold*, we use output of our rule-based method. For other diseases, we use our SVM method. We created this configuration by selecting better methods for each disease based on the development evaluation.

Configuration 2) The second configuration focuses on the confidence value by libSVM. We adopt the SVM classification result, if the confidence value is greater than 80 points. This threshold of 80 points was determined because the best score was obtained in the development evaluation by 80 points among 90, 80, and 70 points. When the confidence value is less than 80 points, we adopt the output of our rule-based method.

## 4. RESULT

We describe results of the formal run in this section. Section 4.1 introduces comparison between our configurations. Section 4.2 introduces comparison with other teams. For their comparisons, we refer to the evaluation scores of the MedWeb organizer’s overview paper [1].

### 4.1 Comparison between Configurations

We prepared our own baseline method which only uses our rule-based classification result. Table 1 compares our formal run results with our baseline. Configuration 1 is superior to Configuration 2 except for micro and macro Precisions. Regarding Configuration 1, exact match accuracy, micro F1 and macro F1 scores are 0.802, 0.871 and 0.856, respectively. Regarding Configuration 2, exact match accuracy, micro F1 and

<sup>4</sup> <https://www.csie.ntu.edu.tw/~cjlin/libsvm/>

macro F1 scores are 0.784, 0.855 and 0.831, respectively. Both configurations performed with higher precision than our baseline, which classifies by our rule-based method only. As for the way of combination, Configuration 1 (selecting better method for each disease based on development performance) was better than Configuration 2 (combination by SVM confidence value). We found that the performance of classifiers in Section 2.2 was not good for *influenza*, *fever*, and *cold*.

## 4.2 Comparison with Other Teams

We show comparison of our results with organizers’ baseline results in Table 2. The organizers’ baseline method employs SVM with unigram and bigram features. The scores of the unigram version of the SVM baseline for exact match accuracy, micro F1 and macro F1 are 0.761, 0.849 and 0.835, respectively. The scores of the bigram version of the SVM baseline for exact match accuracy, micro F1 and macro F1 are 0.752, 0.843 and 0.830, respectively.

In the exact match accuracy, the score of our rule-based baseline method was between two SVM baselines by the organizers. The best result in our configurations, which incorporates the SVM method for specific diseases was the fourth best among the participating teams. In the micro F1 score, our best configuration is ranked third. In the micro and macro recall, our best configuration is ranked second.

## 5. DISCUSSION

We show some concrete examples that were successfully/not successfully classified by our rule-based method and our machine learning method.

The first example was correctly answered by our rule-based method for *influenza*. This tweet “I wish I could get the *influenza* vaccine cheaply. (インフルのワクチンが安く手に入れたらなあ。)” contains the prevention word “vaccine”. The prevention word is mentioned in section 2.1.

Next example of successful classification is “What I say *diarrhea* in English? (下痢を英語で言うと何だろう?)” Because the word “English (英語)” is in the training data, it can be considered that classification could be done by SVM method. “*Headache*” which is customarily used in Japanese is effective for finding tweets about true *headache*. For example, “I am painful because I am doing the work of translation, and I have a *headache*. (翻訳の仕事をしていたら苦痛すぎて頭痛がしてきた。)", our system can classify this tweet correctly. On the contrary, the word “shrimp (海老)” of the tweet “Do shrimps come down with the *flu*? (インフルエンザって海老もなる

の?)” does not appear in the training data. Therefore, the system judges that *influenza* is *p*, while this output is an incorrect.

## 6. FUTURE WORK

As mentioned in Section 4.1, Configuration 1 performed the best classification. Since Configuration 2 can also classify with higher classification accuracy than rule-based baseline method, we could achieve better score by Configuration 2 with SVM confidence value for diseases except for *influenza*, *fever*, and *cold*. The result could be raised with such an appropriate combination of methods, so additional verification would be necessary in the future.

## 7. CONCLUSION

We created two methods and proposed an effective combination of these methods to classify the posters of tweets whether they have disease (positive) or not (negative). Our methods were rule-based and machine learning (SVM). The most effective combination was to use the rule-based method and the SVM method individually for specific diseases. In addition, both of our configurations were able to give higher exact match rate and F1 score than the baseline system of MedWeb organizers. Although the best configuration was to use the rule-based method and the SVM method individually for specific diseases in this task, other appropriate combinations of methods could raise the classification accuracy in future.

## 8. ACKNOWLEDGEMENT

This research was partially supported by JST CREST and MEXT Kakenhi.

## REFERENCES

- [1] E. Aramaki, S. Wakamiya, M. Morita, Y. Kano, and T. Ohkuma, “Overview of the NTCIR-13 : MedWeb Task,” 2017.
- [2] E. Aramaki, M. Morita, Y. Kano, and T. Ohkuma, “Overview of the NTCIR-12 MedNLPDoc Task,” *NTCIR-12*, pp. 167–179, 2016.
- [3] M. Sakishita and Y. Kano, “Inference of ICD Codes by Rule-Based Method from Medical Record in NTCIR-12 MedNLPDoc,” *NTCIR-12*, 2016.
- [4] K. Hatano and Kazuhiko Ohe, “Information Retrieval System for Japanese Standard Disease-Code Master Using XML Web Service,” 2003.

Table 1. Comparison between two combination configurations and rule-base baseline method

Configuration	Exact match	F1		Precision		Recall	
		micro	macro	micro	macro	micro	macro
Configuration 1	0.802	0.871	0.856	0.831	0.815	0.915	0.904
Configuration 2	0.784	0.855	0.831	0.840	0.816	0.871	0.850
Only Rule-based	0.758	0.849	0.833	0.798	0.782	0.906	0.899

Table 2. Comparison between our combination configurations and baseline system provided by the MedWeb organizer

Configuration	Exact match	F1		Precision		Recall	
		micro	macro	micro	macro	micro	macro
Configuration 1	0.802	0.871	0.856	0.831	0.815	0.915	0.904
Configuration 2	0.784	0.855	0.831	0.840	0.816	0.871	0.850
Baseline: SVM (unigram)	0.761	0.849	0.835	0.843	0.828	0.854	0.842
Only Rule-based	0.758	0.849	0.833	0.798	0.782	0.906	0.899
Baseline: SVM (bigram)	0.752	0.843	0.830	0.838	0.82	0.848	0.845