DrG at NTCIR-13: MedWeb Task

Kazui Morita
University of Tokyo
k.morita@bs.s.u-tokyo.ac.jp

Toshihisa Takagi
University of Tokyo
tt@bs.s.u-tokyo.ac.jp

ABSTRACT
NTCIR13 MedWeb Task provides pseudo-Twitter messages (in Japanese, English, and Chinese) and is to classify these messages whether the message contains patient symptom or not. We participated in this Japanese subtask by using Random Forest and some manual rules.

Keywords
Classification, Random Forest, Natural language processing, Twitter

Team Name
DrG

SubTask
MedWeb Japanese subtask

1.INTRODUCTION
Recently, Analysis of social media content is used in various fields such as consumer behavior, sentiment analysis, etc. One of the fields is predicting the outbreak of epidemics by detecting mentions on social media. Previous strategies mainly relied on simple Natural language processing approach such as keyword counting but it can lead to incorrect predictions[1]. In order to encourage research of the strategies of the prediction, MedWeb task requires participants to perform a multi-label classification that labels for 8 diseases/symptoms must be assigned to each tweet collected by a crowdsourcing[2].

We tackled this classification by keyword matching, machine learning, and a manual rule. According to the result, we propose future work taking advantage of a characteristic of social media.

2.MATERIALS & METHODS
The task organizer created 2,560 pseudo tweets that include at least one keyword of target diseases by means of crowdsourcing. Each message is attached “p” (positive) or “n” (negative) label for 8 symptoms from the clinical viewpoint, considering the medical importance of information[3]. The training data corpuses consist of 1,920 messages with labels. The test data corpuses consist of 640 messages without labels. The task requires classified test data into positive and negative tweets based on training data.

We used decision trees and their extension Random Forests for classifier system in this task because decision trees are easy to observe transition according to guideline’s change and to add methods by hand. Our method consists of the following three steps as shown in Figure 1.

2.1.Candidate Classification
Extracting messages for each disease from a given tweet set by searching specific words. Mach of disease tweet has feature words. For example, every message attached “p” labels for “Influenza” have “インフル”. We collect messages which have these words for each disease and classify message whether the message contains patient symptom or not by following steps.

2.2.Model induction
Classifying the message whether the message contains patient symptom or not by using machine learning. We segment a message using a Japanese morphological analyzer MeCab, an open source application[4]. We employed RandomForest for classifier system[5] implemented with scikit-learn[6].

2.3.Manual rule construction
The twitter message contains some features like the inclusion of disease names, paraphrases, etc. We searched such features in training data sets and made some rules. For example, the message which has “鳥インフル” was attached “n” label for “Influenza” even if the message was attached “p” label in Step2 (in the training data, 34 messages have “鳥インフル”, and every message is attached “n” label). another example, the message which has “スペイン風邪” was attached “n” label for “Cold” even if the message was attached “p” label in Step2(in the training data, 14 messages have “スペイン風邪”, and every message is attached “n” label).

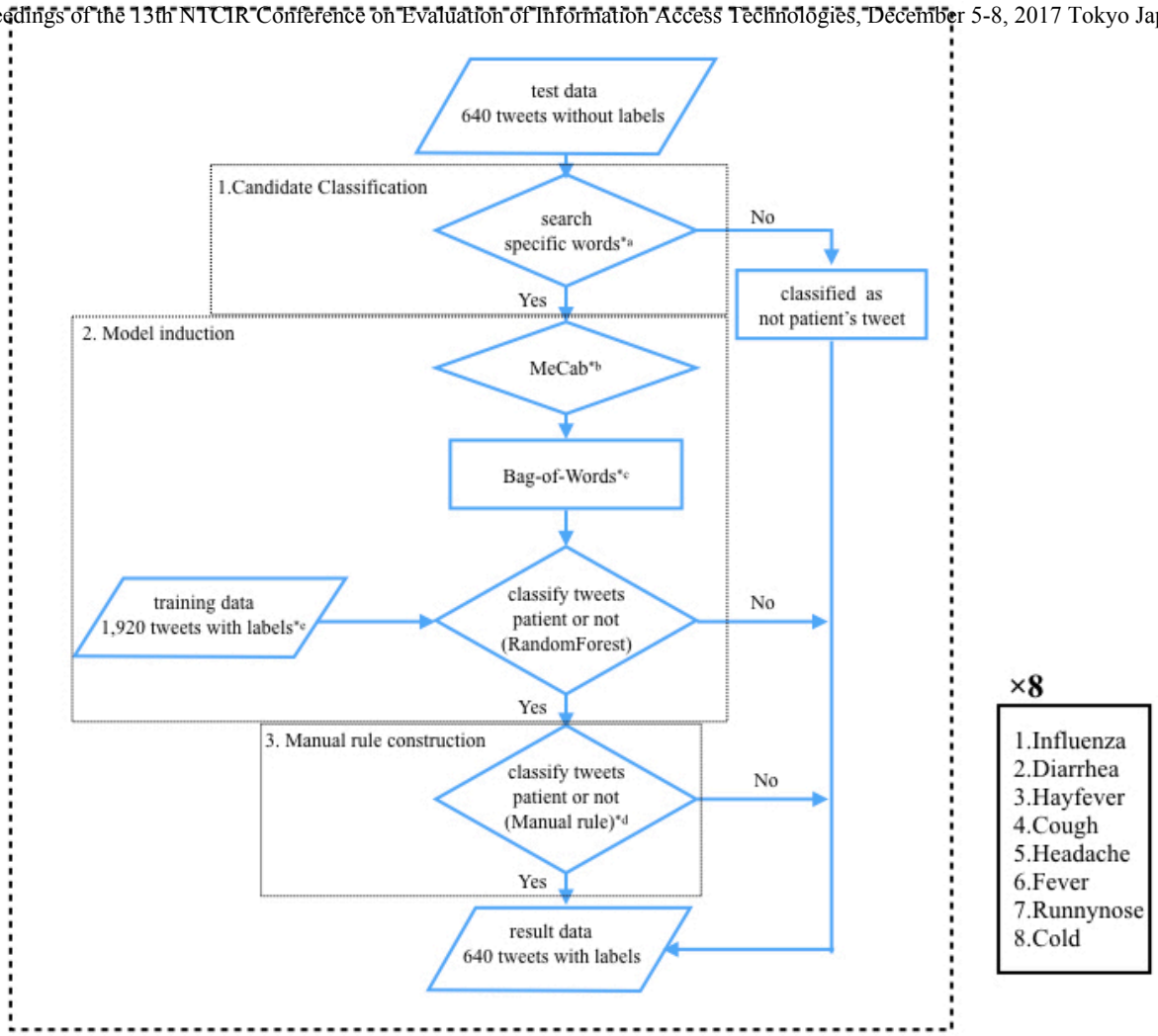

Figure 1: The architecture of our method. a: Searching tweets which have specific words. See 2.1 Candidate Classification for details. b, c: Extracting words from sentence by a Japanese morphological analyzer MeCab. d: Classifying tweets depending on some features. See 2.3 Manual rule construction for details. e: The positive label represents patient's tweet, and the negative label represents not patient's tweet.

3.RESULTS

Table 1 shows the list of specific words for each disease in Candidate Classification. Most positive messages were extracted by searching specific words (97.4% of training data set and 97.8% of test data set).

Table 2 shows the performances of all participating systems and baseline systems. Performance of subtasks was assessed using the exact match accuracy, F-measure (β = 1) based on precision and recall, and hamming loss[2].

4.DISCUSSION

Compared to other work, our system is not the best performance. There are some typical errors in our methods.

In the following test data sentences, our system attached negative labels for "Diarrhea" by Candidate Classification but the organizer attached positive labels.

・弟は熱がありリンパも腫れるしお腹も痛いらしい (positive)

・鼻水ダラダラに熱４０度。お腹は下すは体ボロボロ (positive)

The cause of this mistake is that these sentences don't have specific words. For improving the result, these messages are classified without Candidate Classification. However, we think that this step is necessary for classifying social media. When we actually search patient's messages on Twitter, we cannot afford to classify all message whether patient or not respectively because Twitter messages are too numerous and almost all of the messages are have nothing to do with disease.

In this task, the Basic Guideline set rules for past information, which is meaningless from the viewpoint of surveillance, should be discarded. This rule regards 24 hours as the standard condition. In the case that the timing is ambiguous, a general guide would be that information within 24 hours is regarded as "p"[3]. For example, the following sentence attached the positive label.

Symptoms	Specific words	training data			test data		
		The number of extracted tweets*a	Positive tweets*b	All positive tweets*c	The number of extracted tweets	Positive tweets	All positive tweets
Influenza	インフル	247	106	106	84	24	24
Diarrhea	下痢	234	169	182	86	61	64
Hayfever	花粉症	250	163	163	71	46	46
Cough	咳,痰	237	220	227	86	79	80
Headache	頭痛	279	233	251	86	72	77
Fever	熱,インフル	582	334	345	186	93	93
Runnynose	鼻水,鼻づまり,花粉症,鼻風邪	504	375	375	168	123	123
Cold	風邪	366	265	265	117	86	90

Table 1: List of specific words for each disease. a: Extracted tweets by Candidate Classification. b: The number of 'p' messages in extracted tweets by Candidate Classification. c: The number of 'p' messages in the training data corpus.

ID	Exact_match	F1-micro	Precision-micro	Recall-micro	F1-macro	Precision-macro	Recall-macro	Hamming_loss
NAIST-ja-2	0.880	0.920	0.899	0.941	0.906	0.887	0.925	0.019
NAIST-ja-3	0.878	0.919	0.899	0.940	0.904	0.885	0.924	0.019
NAIST-ja-1	0.877	0.918	0.899	0.938	0.904	0.887	0.921	0.020
AKBLt-ja-3	0.805	0.872	0.896	0.849	0.859	0.883	0.839	0.029
UE-ja-1	0.805	0.865	0.831	0.903	0.855	0.819	0.902	0.033
KIS-ja-2	0.802	0.871	0.831	0.915	0.856	0.815	0.904	0.032
UEt-ja-3	0.800	0.866	0.823	0.913	0.855	0.812	0.911	0.033
AKBL-ja-1	0.800	0.869	0.889	0.849	0.847	0.873	0.825	0.030
AKBL-ja-2	0.795	0.868	0.891	0.846	0.849	0.875	0.827	0.030
KIS-ja-3	0.784	0.855	0.840	0.871	0.831	0.816	0.850	0.034
Baseline-SVM-unigram	0.761	0.849	0.843	0.854	0.835	0.828	0.842	0.036
KIS-ja-1	0.758	0.849	0.798	0.906	0.833	0.782	0.899	0.038
Baseline-SVM-bigram	0.752	0.843	0.838	0.848	0.830	0.820	0.845	0.037
NTTMU-ja-1	0.738	0.835	0.770	0.913	0.829	0.761	0.921	0.042
UE-ja-2	0.706	0.815	0.696	0.983	0.803	0.702	0.984	0.052
NIL-ja-1	0.680	0.749	0.862	0.662	0.742	0.845	0.671	0.052
DrG-ja-1	0.653	0.777	0.825	0.734	0.774	0.808	0.779	0.049
NTTMUt-ja-3	0.614	0.775	0.740	0.814	0.773	0.720	0.840	0.055
NTTMU-ja-2	0.597	0.770	0.741	0.801	0.753	0.706	0.813	0.056
AITOK-ja-2	0.503	0.706	0.726	0.687	0.696	0.738	0.767	0.067
AITOK-ja-1	0.092	0.368	0.243	0.757	0.355	0.238	0.765	0.304

Table 2: Performances of each team in Japanese subtask. The cell highlighted in gray is the performance our team.

・インフルエンザのせいで昨晩大分苦しかった、今日は回復したけどね。(positive)

However, this rule is not suitable for social media because general social media have time information for each message. For improving surveillance, each twitter message should be attached time information respectively. If this message has time information,

attached negative on the day of the tweet and positive on the day before. There is a possibility that time information has the effects of classification that day.

For Influenza, the Basic Guideline don't accept expressions with suspicion such as "feel," "have possibilities," "likely to," "a warning sign," "bad feeling," etc. For example, the following sentence attached the negative label.

・背中がぞくぞくする、インフルエンザかな。(negative)

・インフルエンザかもしれないから部活休もうかな。(negative)

However, these expressions are dealt depending on the epidemic season. When epidemic reaches a peak, these expressions had better be classified as positive. There is a

possibility that tweet season has the effects on classification, and we would like to consider them for our future work.

5.REFERENCES

[1] A. Lamb, M. J. Paul, and M. Dredze. *Separating fact from fear: Tracking flu infections on twitter*, 2013.

[2] E. Aramaki, S. Wakamiya, M. Morita, Yo. Kano, and T. Ohkuma. *Overview of the NTCIR-13: MedWeb task. In Proceeding of the NTCIR-13 Conference*, 2017.

[3] E. Aramaki, S. Wakamiya, M. Morita, Y. Kano, and T. Ohkuma. *NTCIR13 Medical Natural Language Processing for Web Document Annotation Guideline*, 2017.

[4] T. Kudo, K. Yamamoto, and Y. Matsumoto. *Applying conditional random fields to Japanese morphological analysis. In Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2004.

[5] Breiman, L.: *Random Forests*, Machine Learning, 2001.

[6] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E douard Duchesnay. *Scikit-learn: Machine learning in python.* Journal of Machine Learning Research, 2011.