Principle Base Approach for Classifying Tweets with Flu-related Information in NTCIR-13 MedWeb Task

Josan Wei-San Lin
Graduate Institute of Biomedical Informatics
Taipei Medical University
Taipei, Taiwan
g658102003@tmu.edu.tw

Hong-Jie Dai
Department of Computer Science and Information Engineering
National Taitung University
Taitung, Taiwan
hjdai@nttu.edu.tw

Joni Yu-Hsuan Shao
Graduate Institute of Biomedical Informatics
Taipei Medical University
Taipei, Taiwan
jonishao@tmu.edu.tw

ABSTRACT
Disease surveillance system on social media become more important and difficult recently, not only for the variety data type, but also the difficulty for extracting the disease related words and interpreting correctly. In real world, one disease can have many types of symptoms. A symptom is observed by the patient, which is subjective and cannot be measured directly. For example, people who have the flu often feel some of these symptoms: fever, cough, or runny nose. How to identify the disease according to observed symptoms is a challenging problem. In this study, we propose a principle base method to approach the goal of classifying tweets conveying flu-related information in Japanese. By evaluating the proposed method on the corpus of the NTCIR-13MedWeb task, the proposed method achieves micro/macro F-scores of 0.8352/0.8290 on the training set, and an F-score of 0.835 on the test set. We also report the performance of our models, which are based on support vector machines and recurrent neural networks, developed for the dataset in English. In the future, we will include grammatical information to improve the performance of the developed model.

Keywords
Texting mining, principle base approach, text classification

Team Name
NTTMU

Subtasks
Japanese and English Subtasks of MedWeb, The NTCIR-13 Conference

1. INTRODUCTION
As one of the teams participating the NTCIR-13 MedWeb (Medical Natural Language Processing for Web Document) task[6], we report the details of our systems developed for classifying Twitter-like message texts in Japanese and English, which are related to flu and its related symptom information.

For the dataset in Japanese, our method is based on the statistical principle-based approach (PBA). PBA relies on a prebuilt ontology for the target problem. It uses slots defined in the ontology to label the input text, combines them into representative principles, and employs a matching algorithm to accomplish the goal. This approach has been successfully employed in domains such as sentimental analysis, relation extraction and topic detection [1-3].

For the dataset in English, we formulate the task as a classification problem and employ support vector machines (SVMs) and recurrent neural networks (RNNs) to develop our models.

2. METHOD
First of all, we preprocess the tweet in the given datasets by generating tokens. MeCab [1] is used to tokenize tweets in the Japanese corpus. For tweets in English, the part-of-speech tagger developed by Gimpel, et al. [4] is used for tokenization.

Because the models developed for the English corpus was based on well-known machine learning algorithms, in the following sub-sections, we will focus on our PBA-based system developed for the Japanese dataset.

2.1 Knowledge Construction for Flu-related Tweet Classification
In PBA, we use a hierarchical principle-slot combination scheme to express knowledge in Information Map (InfoMap)[5]. Here a *principle* refers to an organized semantic description of a target. For our application, it could be terms related to flus, its symptoms or tweets conveys flu information. Each principle contains a collection of *slots*, and *relations* specified among them. A *slot* serves as the basic component that holds a piece of information in a particular principle. A slot can consist of a set of words, phrases, semantic categories, or other slots.

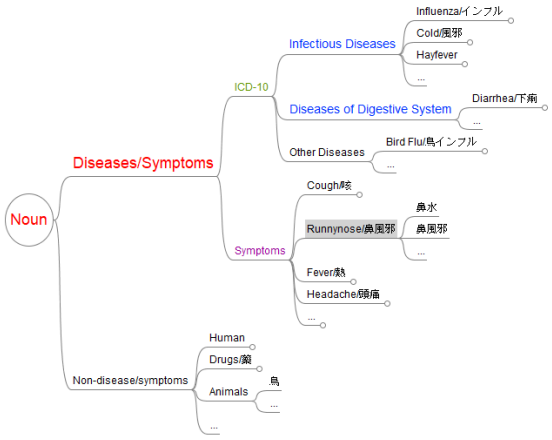

Fig1. Example slots defined for the Japanese subtask.

The knowledge we developed for capturing tweets conveyed flu-related information is manually constructed by considering

[1] An open source morphological analysis engine. Available at http://taku910.github.io/mecab/

common sense and the International Statistical Classification of Diseases and Related Health Problems 10th Revision(ICD-10). Figure 1 illustrates an example. As one can see that a slot in InfoMap can be defined in a hierarchical manner. Each node shown in Figure 1 is a slot defined for the task. For example, the [ICD-10] slot consists of three sub-slots: [Infectious Diseases], [Diseases of Digestive System] and [Other Diseases]. According to ICD-10, we know the term "ウイルスの特効薬" implies effective medicine for Influenza, therefore it is defined under the [Disease/Symptoms] slot. The term "ウイルスの特効薬の研究" means a research related to virus medication, which is defined under the [Non-diseases/symptoms] slot. Figure 2 shows the slots defined for the runny nose concept.

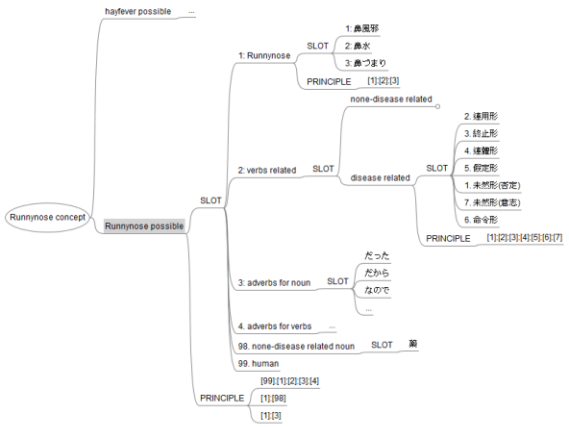

Fig. 2 Slots defined for the runny nose concept.

In addition to the [Noun] slot showed in Figure 1, we define five slots including [Verb], [Adjective], [Adverb], [Phrase], and [Clause]; each of which includes the [Diseases/Symptoms] slot and the [Non-disease/symptoms] slots. The [Verb] slot is the most important slot for the task because only with the rules of verb usages we can express the actions. Take verbs of positive or negative form and the concept of sick or not sick as an example. We divide the knowledge into infection possible concept with possible/negative verbs in sentences and infection negative concept but using possible/negative verbs in sentences, such as [感染名詞(肯定句)]/[感染名詞(否定句)] and[感染句(肯定句)]/[感染句(否定句)].

2.2 Principle Definition

Based on the defined slots, we can define principles by specifying relations such as collocation or positional relation among the defined slots. For this task, we generate principles that can classify tweets conveys flu-related information. A principle in InfoMap is represented as a "PRINCIPLE" node. In Figure 3, the principle node, [1]:[2]:[3]:[4],indicates that if a tweet matching the defined order of slots, it should indicate that the user has a runny nose (鼻風邪). Here [1] refers to the[Drugs/藥] slot defined in the [Non-disease/symptom] slot as shown in Figure 1. With the definition of the combination of different slots in principles, we can form the patterns for matching.

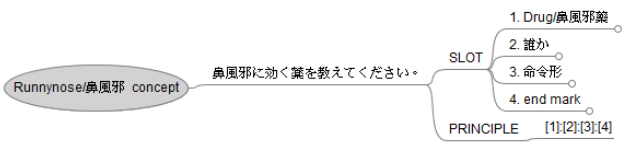

Fig.3 The concept of runny nose represented in PBA.

Furthermore, in InfoMap, a principle along with its related slots can be represented as a concept node. The lead node illustrated in Figure 2 and 3areconcept nodes. In Figure 2 the concept node consists of one principle and 4 slots: [Drugs/藥], [who/誰か], [Verb/命令形] and [end mark]. This is a mechanism provided by InfoMap for reusing previously defined knowledge in another new slot to form new knowledge. The concept node shown in Figure 3 can be reused in Figure 2.

2.3 Principle Definition

After defining principles, a principle matching algorithm [1]is used to match the labeled tweet to determine the categories from eight disease/symptoms including influenza, cold, hay fever, diarrhea, headache, cough, fever and runny nose. Unlike traditional template matching that involves rigid left-right relation of slots in a sentence, a scoring criteria during principle alignment was used in the algorithm. Therefore the more principles corresponding to a disease/symptom we match, the score corresponding to that disease/symptom is larger. We set a threshold to decide whether or not the event of the corresponding disease/symptom described in tweet is fired.

3. RESULTS

3.1 Dataset

The dataset provided by the NTCIR13-MedWeb task contains two kinds of annotations, positive and negative, which indicates whether a tweet related to diseases/symptoms of influenza, diarrhea, hay fever, cough, headache, fever, runny nose and cold. The annotation guide also take the temporal information as consideration in which a tweet written for describing the diseases/symptoms he suffered recently or just a memory long time ago is annotated as positive and negative, respectively.

There are a total of 1,920 tweets in the training set and 640 tweets in test set in both the Japanese and English datasets. Table 1 shows the statistics of the training set.

Table 1: Statistics of the training set.

Disease/Symptoms	Positive	Negative	Total
Influenza	112	1808	1920
Diarrhea	189	1731	1920
Hay fever	201	1719	1920
Cough	237	1683	1920
Headache	254	1666	1920
Fever	355	1565	1920
Runny Nose	417	1503	1920
Cold	284	1636	1920

3.2 Evaluation Results

We submitted three runs for the Japanese dataset. The first run is based on our PBA. The second and third run are based on the SVMs with the linear kernel and *n*-gram features where *n* was set to one to three. The difference between the second and third runs is that the training set of the second run integrates the translated Chinese and English dataset.

Table2: F1-score on the Japanese dataset.

	Run 1	Run 2	Run 3
Training set	0.902	**0.954**	0.895
Test set	**0.835**	0.770	0.775

For the English dataset, we develop two SVM models. The first (Run1)was trained on the original dataset, was trained on the original dataset. As you can see in Table 1, there is a problem between positive data and negative data – imbalanced training data. Considering class imbalance may cause skewed impact during model training, we reviewed some methodologies, for example random oversampling and random under sampling. In order to solve class imbalance issue, the second used the dataset adjusted by the synthetic minority oversampling technique(Chawla, Bowyer, Hall, & Kegelmeyer, 2002) to create new instances for training. Both models used *n*-gram features with TF-IDF as the weighing function. In addition, we applied the bi-directional RNNs to build the model for the third run. Table 3 shows the results on the test set.

Table 3:F1-score on the English test set.

	Test set			Training set		
Config.	P	R	F	P	R	F
Run1	0.734	0.809	0.77	0.865	0.874	0.869
Run2	0.807	0.911	0.856	0.861	0.876	0.869
Run3	0.836	0.854	0.845	not available (data missing)		

4. DISCUSSION

After conducting error analysis of the outcome of our PBA system, we observed that the cause of most of the error cases is due to that the current knowledge implemented in InfoMap ignored the usage of negative verbs in Japanese. In Japanese grammar, we will not know the meaning of a sentence until we get the whole pictures of verb conjugation. For instance, the combination of the tokens [runny nose/鼻風邪]/noun, [negative/ない]/adj, [cure/治]/verb, and [get sick/かかる]/verb can turn into combination of[鼻風邪治さない] or [鼻風邪かからない]. The first combination means that the runny nose cannot be curried, the second combination indicates that there is no runny nose at all. From the above two examples, we can observe that the polarity of a verb can be changed by conjugating the verb with [negative/ない] to form a negative form. If this conjugation happens in disease-related verbs, the meaning will turn into none-disease related. However, sometimes the meaning will remain none disease-related. Most of situations depend on the verbs meaning and make the task become very tricky.

Our current approach ignoring the use of negative verbs result in lots of errors. The following are another two examples.

- [runny nose/ 鼻 風 邪]/noun[に]/adv[effect/ 効 果]/verb [が]/adv [negative/ない]/adj: The cure does not work for runny nose.
- [runny nose/鼻風邪]/noun + [に]/adv + [reccur/再発]/v + [が/adv] + [negative/ない]/adj: Runny nose does not occur again.

In the future, we will include the rules of verb conjugations in our principles construction to overcome the limitation.

5. REFERENCES

[1] N.-W. Chang, H.-J. Dai, Y.-L. Hsieh, and W.-L. Hsu, "Statistical Principle-based Approach for Detecting miRNA-target Gene Interaction Articles," presented at the Proceeding of the IEEE 16th International Conference on BioInformatics and BioEngineering (BIBE), Taichung, Taiwan, 2016.

[2] Y. C. Chang, C. H. Chu, C. C. Chen, and W. L. Hsu, "Linguistic Template Extraction for Recognizing Reader-Emotion," International Journal of Computational Linguistics and Chinese Language Processing, vol. 21, pp. 29-50, September 2016 2016.

[3] Y. C. Chang, Y. L. Hsieh, C. C. Chen, and W. L. Hsu, "A Semantic Frame-based Intelligent Agent for Topic Detection," Soft Computing, vol. 2015, pp. 1-11, July 2015 2015.

[4] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, et al., "Part-of-speech tagging for Twitter: annotation, features, and experiments," presented at the Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2, Portland, Oregon, 2011.

[5] W.-L. Hsu, S.-H. Wu, and Y.-S. Chen, "Event identification based on the information map-INFOMAP," in Proceedings of IEEE International Conference on Systems, Man, and Cybernetics, 2001, pp. 1661-1666.

[6] Eiji Aramaki, Shoko Wakamiya, Mizuki Morita, Yoshinobu Kano, and Tomoko Ohkuma., Overview of the NTCIR-13: MedWeb task. In Proceeding of the NTCIR-13 Conference, 2017.