# NIL: Using scoring to analyse the ambiguous messages on the NTCIR-13 MedWeb task

Masao Ito
NIL Software Corp.
2-17-7, Kinuta, Setagaya
Tokyo, 157-0073, JAPAN

## ABSTRACT

To decide the symptom and disease from the short message, we first analyse the text by the lexical and syntactical analysis. Then, we give it the point according to the criteria that the guideline provides. Finally, we decide the positive or negative value for each symptom/disease by using the weighted sum.

## Keywords

Natural Language Processing; Twitter

Team Name: NIL, Subtask: Japanese subtask

## 1. TASK

The given ""task is to classify them whether or not the message contains patient symptom" form the pseudo twitter sentences."[1]

## 2. METHOD

As preprocessing, we make the lexical and syntactical analysis by using tools; MeCab and CaboCha. Next, we analyse the text by the rules that the guideline shows and provide points to each message. The guideline has the various rules: there are eleven ones for the symptom and disease, six for the temporal category and five for the spatial one. We create the mechanism to give the points corresponding to each rule. Finally, we decide the positive or negative value after calculating the weighted average.

### 2.1 Procedure

We explain the detailed procedure we do in the analysis and the way of scoring to indicate the probability of each symptom and disease. We give points in the middle of analysis, but we describe the scoring approach on the last block.

(a) Finding the corresponding messages

Using keywords and its combination to other special words, we find the sentences to be checked. For example, we use keywords for a headache: the combination of the keyword, "頭(head)" and the word:"いたい(pain)", ("頭が痛い(headache)" is a candidate for the symptom. On the other hand, if it combines the word "contract", we exclude this sentence from the candidate list because it is not related to the symptom, headache (cf. training data 103).

(b) Checking the predicate part

Next, we test the predicate of the candidate sentence got from the previous work (a). If in the predicate we can find keywords that are relating to the onset of disease or symptom, we improve the score that is relating to disease/symptom. If there is a word that shows recovering from the disease or symptom, the point of it will decrease.

To analyze the short message, we think that the modality of it is important. For example, we improve the score of the sentence that has next epistemic modality: The sentence, "From colds, maybe?" (Guideline 2.3 f), indicates the probability of his catching a cold.

(c) Searching the time and location of utterance

According to the guideline, the twenty-four-hour time span is a condition to make a disease or symptom shown in the message positive. If the sentence meets this rule, we give the big positive number. Moreover, if the sentence has the word that is relating to the future or past, we give it the negative number. If the sentence does not have the temporal sentence, we give the small positive number.

As for the location of people who has a disease or symptom, we search and use the information of human relation (e.g. family, position in the office) and the location nearby (e.g. "on the train/bus", "in the office/school/convenience store"). If a sentence has this word, we give the positive number to the sentence.

(d) Scoring

We grade a disease or symptom according to the formula showing below:

$$S_{ds} = \sum_{a \in A} p(a)w(a)$$

The set A includes features; predicate, modality, time and location. The *p(a)* is the score given to each disease of symptom (ds). The *w(a)* is the weight value.

Finally we decide the positive/negative of each disease of symptom by this score $S_{ds}$.

## 3. CONCLUSION

This was a hard task for us (, that means this is a good task ?). For example, to the message "my doctor mistook my symptoms for allergies", the guideline insists that we have to make it "hay fever is positive, cold is positive and runny nose is positive". If the doctor was wrong, the hay fever must be negative. If we remember that only the doctor can make the diagnosis, the cold and runny nose might be negative.

We use the scoring system to decide the disease and symptom. This task is symbolised by the figure in the guideline (section 2.4). There is no distinct boundary. So, we develop the scoring system. We would like to do the further research about the weighting function. We also think that we have to re-examine whether linear scoring is good or not.

## 4. REFERENCES

[1] 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.