Keyword-based Challenges at the NTCIR-13 MedWeb

How to share mutual information?

Mamoru Sakai & Hiroki Tanioka
Tokushima University
tanioka.hiroki@tokushima-u.ac.jp

Faculty of Engineering
Tokushima University

abstract

The AITOK team participated in the MedWeb Japanese subtask of the NTCIR-13. This report describes our approaches to challenging the multi-label classification problem of disease/symptom-related texts and reports some improvements after the formal-run. There are three approaches. The first one is a Keyword-based approach, the second one is a Logistic Regression approach, and the third one is a Support Vector Machine(SVM) approach.

Introduction

- Those approaches are based on keyword and Logistic Regression for formal-run challenges, and machine learnings for extra challenges.
- Basically, our approaches are based on keyword, because the whole tweets are supposed to have some keyword related to disease or symptom.
- To predict if a tweet should be labeled, keyword-based approach is a Boolean method if a keyword contained in the tweet. To label predictively, Logistic Regression approach uses a model with explanatory variables. To label more properly, Support Vector Machine approach is used for non-linear method.
- Through the challenges by three types of approaches, the effects of rule-based approach, statistics based approach, and machine learning based approach are measured and compared on the formal-run results of the MedWeb Japanese subtask.

Approach

Keyword-based Approach

Keyword-based approach assesses if the tweet should be labeled by disease/symptom-related labels with keywords, which is called *feature keyword*. Feature keywords shown in Table1 are extracted from tweets in training data. For the formal-run, every tweet is assessed if the tweet contains the feature keywords by label.

Table 1: Feature keywords for each label

Label	Keyword(s)
Influenza	インフル, いんふる
Diarrhea	下痢, ゲリ, お腹を下, おなかをくだ
Hayfever	花粉症, かふんしょう
Cough	咳, せきが, せきだ, せき、, せき。
Headache	頭痛, 頭が痛い, あたまが痛い, あたまがいたい, 頭がいたい
Fever	熱があ, 高熱, 熱が出, 熱がで, ねつがで, ねつがあ
Runnynose	鼻水が, 鼻水, はなみずが, ハナミズが, 鼻が出,
Cold	風邪

Logistic Regression Approach

The tweets of training data which contain keywords for each label shown in Table2, are the target for analyzing by Logistic Regression.

Table 2: Targeting keywords for each label

Label	Keyword(s)
Influenza	インフル
Diarrhea	下痢
Hayfever	花粉
Cough	咳, 痰
Headache	頭, 痛
Fever	熱
Runnynose	鼻
Cold	風邪

Every tweet of the target training data is parsed to separated terms by MeCab with the ipadic. A keyword which is the separated term is accepted as a candidate feature keyword in case of the POS (Part Of Speech) tag of the keyword is verb, adjective, adverb, or auxiliary verb. The candidate features are explanatory variable to each label of Logistic Regression using RStudio.

	tweet	おいしい	おもい	こわい	だるい	つらい	ない	ありがとうございます	やばい	よい	悪い	安い	寒い	苦しい	激しい	高い	辛い	多い	痛い	怖い	良い	楽しい	っぽい	ほしい	やすい	よい	う	じゃ
1	インフルエンザのワクチン打ちに行ってきた。	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	今年二回目のインフルになったんだけど、これって原発事故による放…	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
3	まさかインフルにかかると思わなかったぜ。ワクチン打ったのにな～	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
4	インフル対策に外出時は、マスクをしてるよ。	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	いつの間にかインフルエンザの季節になったわ。	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	今日インフルの手術じゃないただの注射なのにビビる	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
7	担任がインフルという危機的な状況。	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
8	インフルで部活休む～	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	鳥インフルの季節だってのに国内での検査の体制が整ってなくて全く…	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
10	寝すぎて背中痛い。インフルまじ勘弁	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0

Figure 1: dataframe was made by Rstudio

The equation of Logistic Regression model is given as (1). Logistic Regression thinks in likelihoods of the tweet is positive. If probability was 0.5(50%) and over,it was predicted that the tweet is positive. On the other hand, If probability was under 0.5(50%),it was predicted that the tweet is negative.

$$p(x) = \frac{1}{1 + \exp -(b_0 + b_1 x_1 ... b_k x_k)} \tag{1}$$

Support Vector Machine Approach

Support Vector Machine (SVM) approach is applied as an extra challenge, which is also another keyword-based method. Particularly, linear kernel is essentially the same as the keyword-based approach and the Logistic Regression approach, because linear kernel never solves non-linear problems. On the other hand, non-linear kernel solves non-linear problems, and has a potential to solve the false-positive problem. The whole tweets of train data per label are learned by libsvm. Here, the classification type was C-SVC (-s 0), the kernel method was RBF (Radial Basis Function) kernel (-t 2), and the constant value was $3,000$ (-c $3,000$).

Results

Table shows our results including formal-run and extra results. Keyword-based approach Res2 and Logistic Regression approach Res3 are not good result of rank in the formal-run. The Support Vector Machine approach Res4 is based on keyword unigram, and Res5 is based on keyword bigram. Unofficially, these results were ranked relatively good in the formal-run.

Table 3: The official baseline result and our approach results

Group ID	Exact_match	F1-micro	Precision-micro	Recall-micro
Vanilla-SVM-unigram	0.761	0.849	0.843	0.854
AITOK_medweb_result-ja-5	0.814	0.894	0.854	0.938
AITOK_medweb_result-ja-4	0.780	0.867	0.830	0.908
AITOK_medweb_result-ja-3	0.633	0.728	0.761	0.698
AITOK_medweb_result-ja-2	0.503	0.706	0.726	0.687
AITOK_medweb_result-ja-1	0.092	0.368	0.243	0.757

Table 4: The official baseline result and our approach results

Group ID	F1-macro	Precision-macro	Recall-macro
Vanilla-SVM-unigram	0.835	0.828	0.842
AITOK_medweb_result-ja-5	0.877	0.830	0.933
AITOK_medweb_result-ja-4	0.851	0.808	0.904
AITOK_medweb_result-ja-3	0.715	0.741	0.706
AITOK_medweb_result-ja-2	0.696	0.738	0.767
AITOK_medweb_result-ja-1	0.355	0.238	0.765

Conclusions

- Keyword-based approach and Logistic Regression approach are submitted to the formal-run, both approaches not enough accuracy compared to the official baseline system.
- Hence, SVM approach is added as another machine learning approach.
- The SVM approach is non-linear machine learning, based on keyword unigram and bigram. These challenges realize that the machine learning approach with SVMs is really good compared with rule-based approach and statistic based approach.