Keyword-based Challenges at the NTCIR-13 MedWeb

Mamoru Sakai
Department of Information Science
and Intelligent Systems
Tokushima University, Japan
c501406003@tokushima-u.ac.jp

Hiroki Tanioka
Center for Administration of
Information Technology
Tokushima University, Japan
tanioka.hiroki@tokushima-u.ac.jp

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. Keyword-based and Logistic Regression approaches were submitted to the formal-run. SVM-based approach was not submitted to the formal-run. These challenges made some improvements and have realized a machine learning approach is really good compared to other approaches.

Team Name

AITOK

Subtasks

MedWeb (Japanese)

Keywords

Keyword-based,Logistic Regression,Support Vector Machine

1. INTRODUCTION

The AITOK team participated in the MedWeb Japanese subtask of the NTCIR-13 [3]. 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. 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 keywords related to disease or symptom. To predict if a tweet should be labeled, the keyword-based approach is a Boolean method if a keyword contained in the tweet. To label predictively, the Logistic Regression approach [2] uses a model with explanatory variables. To label more properly, the Support Vector Machine [5] approach is used for non-linear method.

Through the challenges by three types of approaches, the effects of rule-based approach with keywords, statistics based approach with Logistic Regression, and machine learning based approach with Support Vector Machines were measured and have been compared on the formal-run results of the MedWeb Japanese subtask.

2. APPROACHES

2.1 Keyword-based Approach

When it is observed that a keyword "インフル" (influ) is contained in some tweets with Influenza label at a high frequency. Other tweets also contain characteristic keywords in common to each disease/symptom-related label. Therefore, the 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 Table. 1 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

2.2 Logistic Regression Approach

Logistic regression [2] approach is also keyword-based approach. The tweets of training data which contain keywords for each label shown in Table 2, 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 [6] with the ipadic [1]. 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 [8] [9] using RStudio [7]. The generalized linear model function *glm* is used for analyzing as binomial model. The model of *glm* on RStudio is passed to a function *predict.* The function obtains predictions from a fitted generalized linear model.

Table 3: Top five results in Formal-run and our approach results. Unofficially, Res1 is replaced with Res3 due to a fault in predicting phase of the logistic regression approach.

Group ID	Exact_ match	F1-micro	Precision-micro	Recall-micro	F1-macro	Precision-macro	Recall-macro
NAIST_medweb_result-ja-2	0.880	0.920	0.899	0.941	0.906	0.887	0.925
NAIST_medweb_result-ja-3	0.878	0.919	0.899	0.940	0.904	0.885	0.924
NAIST_medweb_result-ja-1	0.877	0.918	0.899	0.938	0.904	0.887	0.921
AKBL_medweb_result-ja-3	0.805	0.872	0.896	0.849	0.859	0.883	0.839
UE_medweb_result-ja-1	0.805	0.865	0.831	0.903	0.855	0.819	0.902
Vanilla-SVM-unigram	0.761	0.849	0.843	0.854	0.835	0.828	0.842
AITOK_medweb_result-ja-5 [Res5]	0.814	0.894	0.854	0.938	0.877	0.830	0.933
AITOK_medweb_result-ja-4 [Res4]	0.780	0.867	0.830	0.908	0.851	0.808	0.904
AITOK_medweb_result-ja-3 [Res3]	0.633	0.728	0.761	0.698	0.715	0.741	0.706
AITOK_medweb_result-ja-2 [Res2]	0.503	0.706	0.726	0.687	0.696	0.738	0.767
~~AITOK_medweb_result-ja-1~~ [Res1]	~~0.092~~	~~0.368~~	~~0.243~~	~~0.757~~	~~0.355~~	~~0.238~~	~~0.765~~

2.3 Support Vector Machine Approach

Keyword-based approach and Logistic Regression approach are submitted for the formal-run. Both approaches are not good result of rank in the formal-run. The reason is supposed that both approaches are linear methods, which have some false-positive to impregnable labeling tasks.

For instance, when a tweet contains "インフル" (influ), there are some cases where the tweet should not be labeled as Influenza. Hence, 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 [4].

3. RESULTS

Table 3 shows our results including formal-run and extra results. Both keyword-based approach and Logistic Regression approach are submitted for the formal-run. Besides, the keyword-based approach Res2 is not good at the rank as expected. Therefore, it is considered that another approach based on non-linear approach should be applied for the MedWeb labeling task.

3.1 Result of Keyword-based

The Keyword-based approach Res2 is a truly simple rule-based approach. The result is reasonable. The advantage of this approach is that manual adjustment by adding other keywords and balancing thresholds. However, the scores of Recall, Precision, and F1 were totally under baseline system, *Vanilla-SVM-unigram* and *Vanilla-SVM-bigram*. The main reason is supposed that the result includes false-positive.

3.2 Result of Logistic Regression

Res1 was undesirable, because the whole tweets of training data were used in the predicting phase, even though only the tweets of training data which contain the targeting keywords were used in the training phase. Hence, Res3 was appended to these results instead of Res1 after formal-run. The scores were under the baseline SVM systems, because the result had lots of false-positive. The difference from the keyword-based approach is that the prediction method is based on a probabilistic method with Logistic Regression model.

3.3 Result of Support Vector Machine

The Support Vector Machine approach Res4 is based on keyword unigram, and Res5 is based on keyword bigram. 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 $3,000$), which were tuned on the train data $ja_{t}rain_{2}0170501.xlsx$. Unofficially, these results were ranked relatively good in the formal-run.

4. CONCLUSIONS

There were three approaches in our challenges. The first one was a Keyword-based approach, the second one was a Logistic Regression approach, and the third one was a Support Vector Machine(SVM) approach. Although keyword-based and Logistic Regression approaches were submitted to the formal-run, both approaches were not enough. Then, SVM approach was added as another approach after formal-run. The SVM approach was non-linear machine learning, based on keyword unigram and bigram. These challenges have realized that the machine learning approach with SVMs is really good compared with the rule-based approach and the statistic based approach.

5. REFERENCES

[1] IPA dictionary: mecab-ipadic-2.7.0-20070801, 2007. (Accessed 4 Aug 2017).

[2] A. Agresti. An introduction to categorical data analysis, volume 135. Wiley New York, 1996.

[3] E. Aramaki, S. Wakamiya, M. Morita, Y. Kano, and T. Ohkuma. Overview of the NTCIR-13: MedWeb Task. In Proceedings of NTCIR-13, 2017.

[4] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27, 2011.

[5] C. Cortes and V. Vapnik. Support-vector networks. Mach. Learn., 20(3):273–297, Sept. 1995.

[6] T. KUDO. Mecab : Yet another part-of-speech and morphological analyzer. 2005.

[7] RStudio Team. RStudio: Integrated Development Environment for R. RStudio, Inc., Boston, MA, 2017.

[8] H. Toyota. Kaiki Bunseki Nyumon (Introductory Regression Analysis). R de Manabu Saishin Data Kaiseki (The latest data analysis learned by R). Tokyo Tosho, 2012.

[9] Y. Yamamoto, T. Fujino, and T. Kubota. R niyoru Data Mining Nyumon (Introductory Regression Analysis with R). Ohmsha, 2015.