Keyword-based Challenges
at the NTCIR-13 MedWeb Japanese subtask

Mamoru Sakai
Tokushima University , Faculty of Engineering

Hiroki Tanioka
Tokushima University, Center for Administration of Information Technology

Method

Formal-run

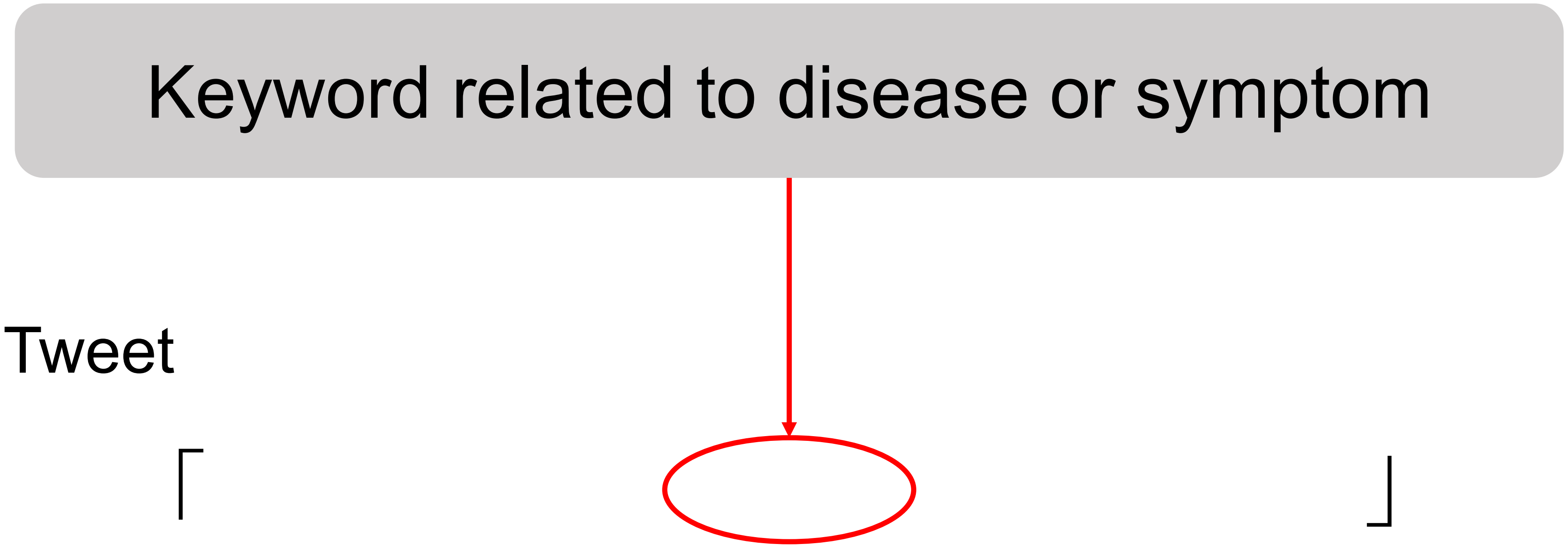
1. Keyword-based Approach
2. Logistic Regression Approach

Additional challenge

3. Support Vector Machine(SVM) Approach

Introduction

Why keyword-based?



Keyword related to disease or symptom

Tweet

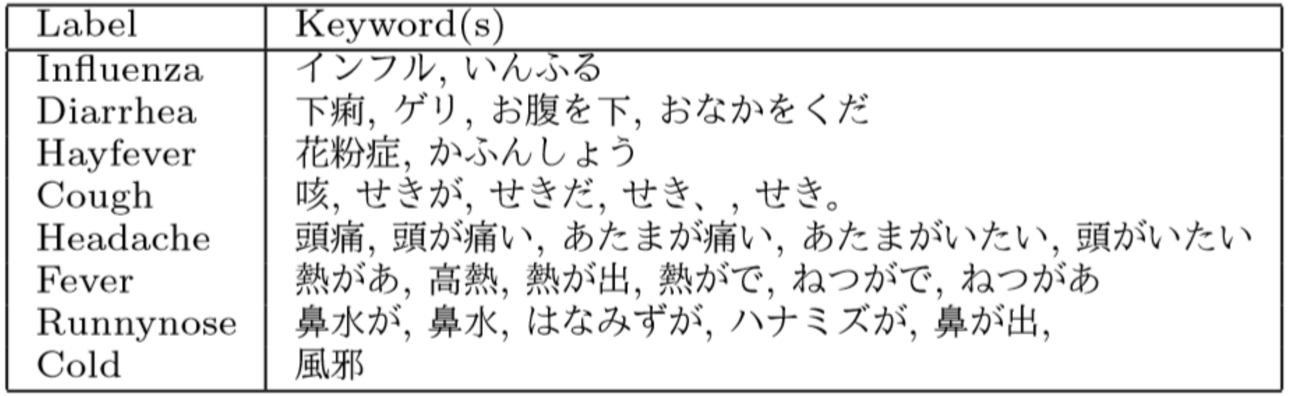
「　　　　　　　　」

Keyword-based Approach

Every tweet is assessed

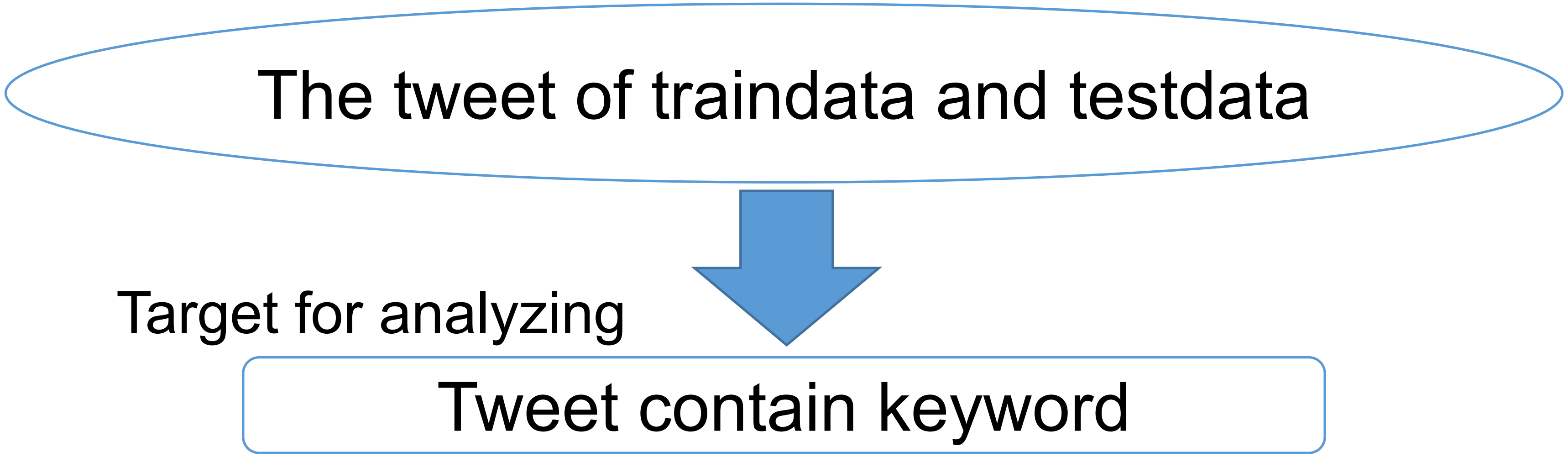
if the tweet contains the feature keywords by label

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

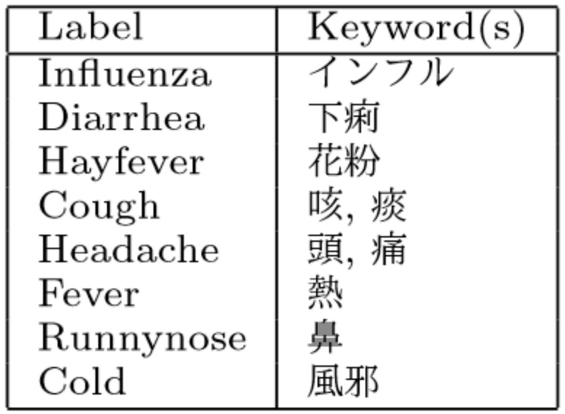


The tweet of traindata and testdata

Target for analyzing

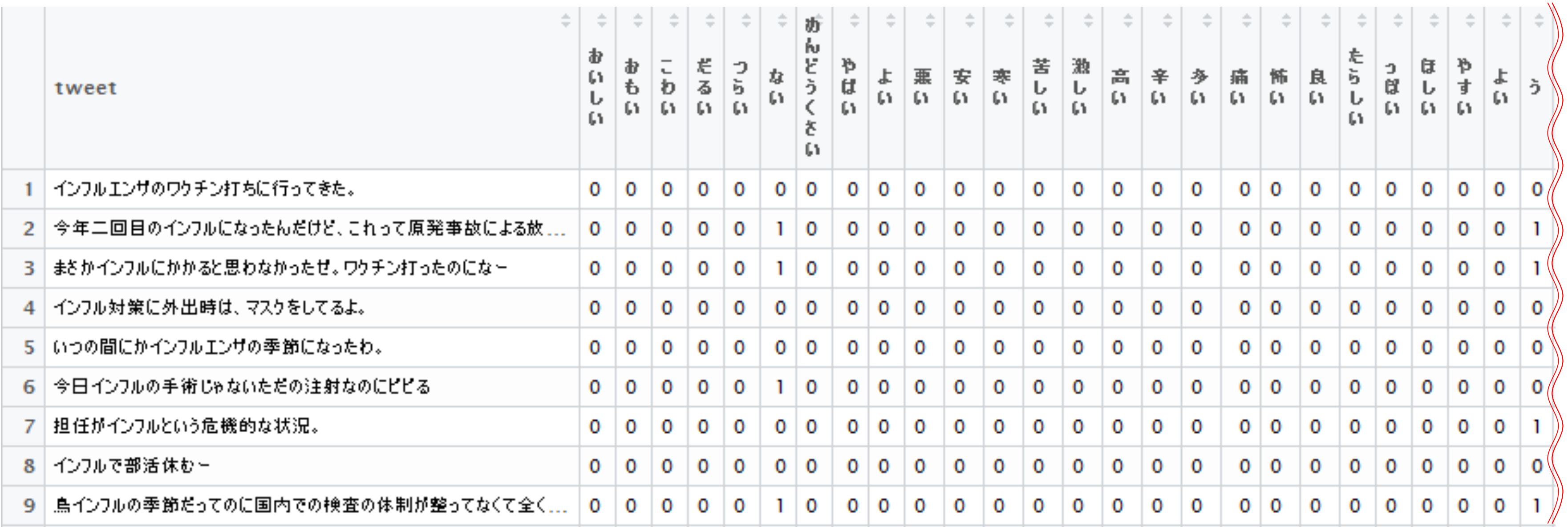
Tweet contain keyword

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

Logistic Regression Approach



	tweet	おいしい	おもい	こわい	だるい	つらい	ない	めんどうくさい	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	たらしい	っぽい	ほしい	やすい	よい	う
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
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
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ー	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
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
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
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
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
8	インフルで部活休むー	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

$$p(x) = \frac{1}{1+e^{-(b_0+b_1x_1+\cdots b_kx_k)}}$$

p >= 0.5 ➡ 1(positive)

p < 0.5 ➡ 0(negative)

Support Vector Machine Approach

Liner method

Keyword-based
Logistic Regression

Some false-positive

Non-Liner method

SVM with RBF kernel → sloves to non-liner problems

Result

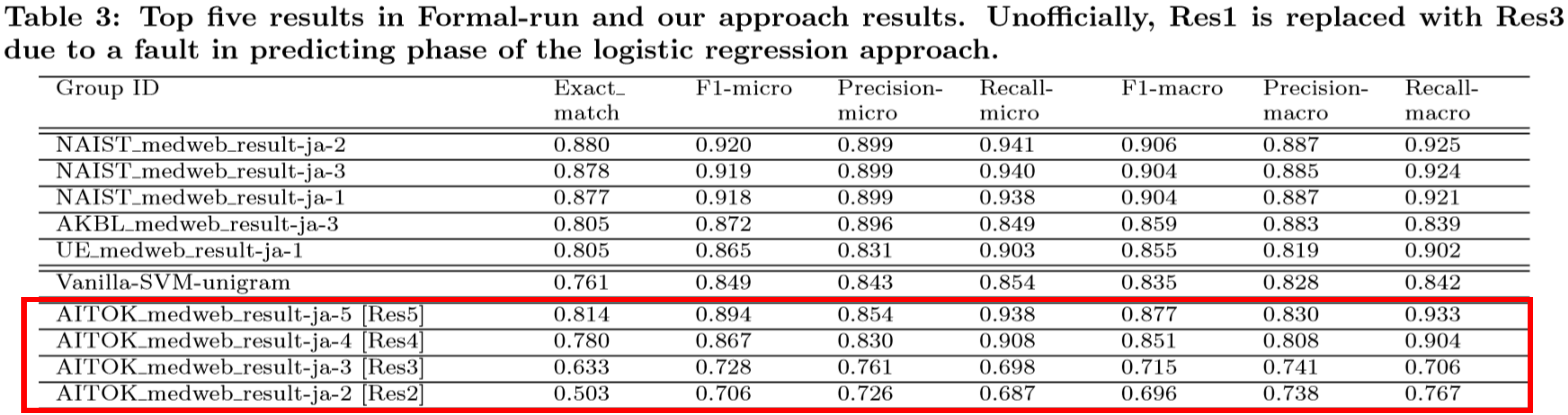


Table 3: Top five results in Formal-run and our approach results. Unofficially, Res1 is replaced with Res3 due to a fault in predicting phase of the logistic regression approach.

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

- Keyword-based(Res2) and Logistic Regression(Res3) were not good result

- SVM approach (Res4,Res5) was better than vanilla-SVM

Conclusion

- Keyword-based and Logistic Regression approach were not enough.

- Both approach were very simple.

- SVM approach is good compared with both approach.