

Overview of the NTCIR-13: MedWeb Task

Shoko Wakamiya
Nara Institute of Science and
Technology, Japan
wakamiya@is.naist.jp

Tomoko Ohkuma
Fuji Xerox Co., Ltd., Japan
ohkuma.tomoko@fujixerox.
co.jp

Mizuki Morita
Okayama University, Japan
morita.mizuki@gmail.com

Yoshinobu Kano
Shizuoka University, Japan
kano@inf.shizuoka.ac.jp

Eiji Aramaki
Nara Institute of Science and
Technology, Japan
aramaki@is.naist.jp

ABSTRACT

The amount of medical and clinical-related information on the Web is increasing. Among the various types of information, Web-based data are particularly valuable, with Twitter-based medical research garnering much attention. The NTCIR-13 MedWeb (Medical Natural Language Processing for Web Document) provides pseudo-Twitter messages in a cross-language and multi-label corpus, covering three languages (Japanese, English, and Chinese), and annotated with eight labels (e.g., cold, fever, flu, and so on). The MedWeb task classifies each tweet into one of two categories: those containing a patient's symptom, and those that do not. Because our task settings can be formalized as the factualization of text, the achievement of this task can be applied directly to practical clinical applications. In all, eight groups (19 systems) participated in the Japanese subtask, four groups (12 systems) participated in the English subtask, and two groups (six systems) participated in the Chinese subtask. This paper presents the results of these systems, along with relevant discussions, to clarify the issues that need to be resolved in medical natural language processing.

Keywords

Medical natural language processing, Twitter, Social media, Shared task, Evaluation

1. INTRODUCTION

Medical reports using electronic media are now replacing those of paper media. As a result, the importance of natural language processing techniques in various medical fields has increased significantly. Our goal is to promote practical tools to assist precise and timely medical decisions. In order to achieve this goal, a series of "shared tasks" (or contests, competitions, challenge evaluations, critical assessments) are being used to encourage research in information retrieval. Several shared tasks have already been organized, such as the Informatics for Integrating Biology and the Bedside (i2b2) task [14], organized by the National Institute of Health (NIH)¹ in the United States, the Text Retrieval Conference (TREC) [19], the ShARe/CLEF eHealth Evaluation Lab² in the European Union, and NTCIR Medical tasks and MedNLP workshops [13, 3, 4] had been held in Japan.

On the other hand, with the widespread use of the Internet, lots of materials concerning medical care or health have been shared on the Web and web mining techniques for utilizing the materials have been developed. One of the most popular medical applications of web mining is flu surveillance, which aims to predict influenza epidemics based on the use of flu-related terms [2, 10, 9, 20, 7, 8, 6, 15, 18, 12, 1, 17]. Most previous studies have relied on shallow textual clues in Twitter messages, such as the number of occurrences of specific keywords (e.g., "flu" or "influenza") on Twitter. However, such simple approaches have difficulty coping with the volume of noisy tweets. Thus, in order to increase their accuracy, recent approaches [2, 12] have employed a binary classifier to filter out noisy tweets. Typical examples of noisy tweets are those that simply express concern or awareness about flu (e.g., "Starting to get worried about swine flu").

Given this situation, the NTCIR-13³ MedWeb (Medical Natural Language Processing for Web Document) task⁴ is designed for obtaining health-related information by web mining, focusing in particular on social media. Specifically we propose a generalized task setting for public health surveillance, referring to the following two characteristics:

- **Multi-label:** this task handles not only a single symptom (*influenza*), but also multiple symptoms such as *cold*, *cough/sore throat*, *diarrhea/stomachache*, *fever*, *hay fever*, *headache*, and *runny nose*. Because a single message could contain multiple symptoms, this is one of the multi-labeling tasks.
- **Cross-language:** in contrast to the previous shared tasks [14, 19, 13, 3, 4], this task covers multiple languages: *Japanese*, *English*, and *Chinese*. To build parallel corpora, we translated the original Japanese messages to English and Chinese.

We distributed each corpus to the participants, of whom nine groups submitted results (37 systems). Specifically, eight groups (19 systems) participated in the Japanese subtask, four groups (12 systems) participated in the English subtask, and two groups (six systems) participated in the Chinese subtask. Table 1 shows the list of participating groups, and Table 2 summarizes the number of participating groups for each subtask. This report presents the results

¹<https://www.nih.gov>

²<https://sites.google.com/site/shareclefhealth/>

³<http://research.nii.ac.jp/ntcir/ntcir-13/index.html>

⁴<http://mednlp.jp/medweb/NTCIR-13/>

Table 1: Organization of groups participating in MedWeb (listed in alphabetical order by Group ID)

Group ID	Organization
AITOK	Tokushima University, Japan
AKBL	Toyohashi University of Technology, Japan
DrG	The University of Tokyo, Japan
KIS	Shizuoka University, Japan
NAIST	Nara Institute of Science and Technology, Japan
NIL	NIL Software Corp., Japan
NTTMU	Taipei Medical University, Taiwan
TUA1	Tokushima University, Japan
UE	University of Evora, Portugal

Table 2: Statistics of result submissions (listed in alphabetical order by Group ID)

Group ID	Japanese	English	Chinese
AITOK	2		
AKBL	3	3	
DrG	1		
KIS	3		
NAIST	3	3	3
NIL	1		
NTTMU	3	3	
TUA1			3
UE	3	3	

of these groups, along with discussions, in order to clarify the issues that need to be resolved in the field of medical natural language processing.

2. CORPUS

The material for the MedWeb task is a collection of tweets that include at least one keyword of target diseases or symptoms (for brevity, we refer to these simply as **symptoms** hereafter). We set eight symptoms, including *cold*, *cough/sore throat* (which we refer to as “cough”), *diarrhea/stomachache* (“diarrhea”), *fever*, *hay fever*, *headache*, *influenza* (“flu”), and *runny nose*.

Owing to the Twitter developer policy on data redistribution⁵, the tweet data crawled using the API are not publicly available. Therefore, our data consist of pseudo-tweets created by a crowdsourcing service.

In order to obtain the pseudo-tweets, we first collected Japanese tweets related to each symptom from Twitter. Then, we classified these tweets as positive or negative, based on the work of [2]. Next, we extracted keyword sets that appeared frequently in positive tweets and negative tweets. We call these keywords **seed words**.

We then had a group of people create pseudo-tweets consisting of 100 to 140 characters that included a symptom and at least one of the seed words of the symptom. Each person created 32 pseudo-tweets (two tweets × two keyword sets (positive and negative) × eight symptoms). As a result, 80 people were able to generate 2,560 Japanese pseudo-tweets.

⁵<https://developer.twitter.com/en/developer-terms/agreement-and-policy>

Table 3: Samples of pseudo-tweets of the eight symptoms. English messages (en) and Chinese messages (zh) were translated from Japanese messages (ja).

Symptom	Pseudo-tweet message
Cold	風邪を引くと全身がだるくなる。 The cold makes my whole body weak.
Cough	一感冒就浑身酸软无力。 あかん。咳込みすぎて頭まで痛くなってきた This is not good. I coughed too much and I got a headache from it. 糟了。咳得太厉害，头都疼起来了。
Diarrhea	下痢ひどすぎて笑うわ I gotta laugh. My diarrhea is so bad. 腹泻过于严重，很搞笑。
Fever	熱が出なくてもリンパが腫れることがよくある。 It's not unusual for lymph nodes to get swollen, even when there's no fever. 很多时候就算不发热淋巴也肿。
Hay fever	花粉症の症状が出てきたのは久し振りだ。 It's been a while since I've had allergy symptoms. 好久没有出现花粉症的症状了。
Headache	頭痛がやばいから帰宅して寝るー My headache is killing me, so I'm going to go home and sleep. 因为头疼得厉害，我回家睡觉了。
Flu	インフルエンザのワクチン打ちに行ってきた。 I went to get vaccinated for the flu. 去打了流感的疫苗。
Runny nose	鼻づまりで今日は休むわー I'm not going today, because my stuffy nose is killing me. 因为鼻塞，今天休息吧！

In the last step, we had the Japanese pseudo-tweets translated into English and Chinese by relevant first-language practitioners. Therefore, we also had 2,560 pseudo-tweets in both English and Chinese. Table 3 shows samples of each set of pseudo-tweets.

3. SYMPTOM LABELING

This section describes the criteria used for symptom labeling. These consist of basic criteria (Section 3.1) and symptom-specific criteria (Section 3.2). The inter-annotator agreement ratio ($n=2$) was 0.9851 (=20174/(2560×8)).

3.1 Basic criteria

The most basic criterion is that the labeling is examined from a clinical viewpoint, considering the medical importance of the information. Thus, non-clinical information should be disregarded.

For example, older information (by several weeks) and non-severe symptoms (headache due to over-drinking) should be labeled “n” (negative). The following three criteria describe the basic principles:

- **Factuality:** The Twitter user (or someone close to the

Table 4: Samples of the training data corpus for the English subtask. ID corresponds to the corpora of other language (e.g., the tweet of “1en” corresponds to the tweets of “1ja” and “1zh”).

ID	Message	Cold	Cough	Diarrhea	Fever	Hay fever	Headache	Flu	Runny nose
1en	The cold makes my whole body weak.	p	n	n	n	n	n	n	n
2en	It's been a while since I've had allergy symptoms.	n	n	n	n	p	n	n	p
3en	I'm so feverish and out of it because of my allergies. I'm so sleepy.	n	n	n	p	p	n	n	p
4en	I took some medicine for my runny nose, but it won't stop.	n	n	n	n	n	n	n	p
5en	I had a bad case of diarrhea when I traveled to Nepal.	n	n	n	n	n	n	n	n
6en	It takes a millennial wimp to call in sick just because they're coughing. It's always important to go to work, no matter what.	n	p	n	n	n	n	n	n
7en	I'm not going today, because my stuffy nose is killing me.	n	n	n	n	n	n	n	p
8en	I never thought I would have allergies.	n	n	n	n	p	n	n	p
9en	I have a fever but I don't think it's the kind of cold that will make it to my stomach.	p	n	n	p	n	n	n	n
10en	My phlegm has blood in it and it's really gross.	n	p	n	n	n	n	n	n

Table 5: Exceptions for symptom labels.

Symptom	Accept expressions with suspicion	Accept just a word of a symptom	“p” (positive)	Exceptions “n” (negative)
Cold	✓	✓	-	-
Cough	✓	✓	Alcohol drinking Pungently flavored food	-
Diarrhea	✓	✓	Overeating Indigestion Alcohol drinking Medication Pungently flavored food	-
Fever	✓	✓	Side-effect of injection	-
Hay fever	✓	✓(only slight fever)	-	-
Headache	✓	✓	-	Due to a sense of sight or smell
Flu	-	-	-	-
Runny nose	✓	-	Hay fever	Change in temperature

user) should be affected by a certain disease or have a symptom of the disease. A tweet that includes only a disease name or a symptom as a topic is removed by labeling it as “n” (negative).

- **Tense (time):** Older information, which is meaningless from the viewpoint of surveillance, should be discarded. Such information should also be labeled “n” (negative). Here, we regard 24 hours as the standard condition. When the precise date time is ambiguous, the general guideline is that information within 24 hours (e.g., information related to today or yesterday) is labeled as “p” (positive).

- **Location:** The location of the disease should be specified as follows. If a Twitter user is affected, the information is labeled as “p” (positive) because the location of the user is the place of onset of the symptom. In

cases where the user is not affected personally, the information is labeled as “p” (positive) if it is within the same vicinity (prefecture) as the user, and “n” (negative) otherwise.

3.2 Symptom-specific criteria

The fundamental annotation principles are described in Section 3.1. However, there are several exceptions to the above principles.

For example, a remark about a “headache” might not relate to a clinical disease (e.g., excessive drinking). When conducting disease surveillance, such statements should be regarded as noise. To deal with disease-specific phenomena, we build a guideline that addresses exceptions for each disease. For example, cases such as “excessive drinking,” “medication,” “pungently flavored food (including irritant),” “spiritual,” “motion sickness,” “morning,” “menstrual pain,” and so on, should be excluded for “headache.” The exceptions

Table 6: Participating systems in Japanese subtask (19 participating systems and two baseline systems). * indicates that the method was tested after the submission of the formal run and, thus, was not included in the results.

System ID	Models/Methods	Language resources
AITOK-ja	Keyword-based, Logistic regression Support vector machine (SVM)*	-
AKBL-ja	Support vector machine (SVM), Fisher's exact test	Patient symptom feature word dict Disease-X feature word dict1 Disease-X feature word dict2
DrG-ja	Random forest	-
KIS-ja	Rule-based, SVM	-
NAIST-ja	Ensembles of hierarchical attention network (HAN) and deep character-level convolutional neural network (CNN) with loss functions (negative loss function, hinge, and hinge squared)	-
NIL-ja	Rule-based	-
NTTMU-ja	Principle-based approach	Manually constructed knowledge for capturing tweets that conveyed flu-related information, using common sense and ICD-10 Custom dictionary consisting of nouns selected from the dry-run data set and heuristics
UE-ja	Rule-based, Random forest	-
Baseline	SVM (unigram, bigram)	-

Table 7: Participating systems in English subtask (12 participating systems and two baseline systems)

System ID	Models/Methods	Language resources
AKBL-en	Support vector machine (SVM), Fisher's exact test	Patient symptom feature word dict Disease-X feature word dict1 Disease-X feature word dict2
NAIST-en	Ensembles of hierarchical attention network (HAN) and deep character-level convolutional neural network (CNN) with loss functions (negative loss function, hinge, and hinge squared)	-
NTTMU-en	SVM, Recurrent neural network (RNNs)	Manually constructed knowledge for capturing tweets that conveyed flu-related information, using common sense and ICD-10 Custom dictionary that consists of nouns selected from the dry-run data set and heuristics
UE-en	Rule-based, Random forests Skip-gram neural network for word2vec	-
Baseline	SVM (unigram, bigram)	-

are summarized in Table 5.

4. METHODS

4.1 Task settings

In the MedWeb task, we organized three subtasks: a Japanese subtask, an English subtask, and a Chinese subtask.

Step 1: Training corpus distribution: The training data corpus and the annotation criteria were sent to the participant groups for development. The training data corpus consists of 1,920 messages (75% of the whole corpus), with labels. Each message is labeled “p” (positive) or “n” (negative) for each of the eight symptoms.

Step 2: Formal run result submission: After about a three-month development period, the test data corpus was sent to each participant group. The test data corpus consists of 640 messages (25% of the whole corpus), without labels. Then, the participant groups submitted their annotated results within two weeks. Multiple results with up to three systems were allowed to be submitted.

Step 3: Evaluation result release: After a one-month evaluation period, the evaluation results and the annotated test data were sent to each participant group.

4.2 Evaluation metrics

The performance in the subtasks was assessed using the exact match accuracy, F-measure ($\beta = 1$) (F1) based on precision and recall, and Hamming loss [21]. The details of the metrics are as follows.

- Exact match accuracy: the most strict metric.
- F1-micro and macro: the harmonic mean of precision and recall.
- Hamming loss: xor loss (lower scores are better).

Note that “micro” is to calculate metrics globally by counting all true positives, false negatives, and false positives. On the other hand, “macro” calculates the metrics for each symptom label, and then determines their unweighted mean. Therefore, label imbalance is not taken into account.

5. RESULTS

Table 8: Participating systems in Chinese subtask (six participating systems and two baseline systems)

System ID	Models/Methods	Language resources
NAIST-zh	Ensembles of hierarchical attention network (HAN) and deep character-level convolutional neural network (CNN) with loss functions (negative loss function, hinge, and hinge squared)	-
TUA1-zh	Logistic regression, Support vector machine (SVM) Logistic Regression with semantic information	Updated training samples using active learning unlabeled posts downloaded with the symptom names in Chinese
Baseline	SVM (unigram, bigram)	-

Table 9: Performance in the Japanese subtask (19 participating systems and two baseline systems). The results are ordered by exact match accuracy.

System ID	Exact match	F1		Precision		Recall		Hamming loss
		micro	macro	micro	macro	micro	macro	
NAIST-ja-2	0.880	0.920	0.906	0.899	0.887	0.941	0.925	0.019
NAIST-ja-3	0.878	0.919	0.904	0.899	0.885	0.940	0.924	0.019
NAIST-ja-1	0.877	0.918	0.904	0.899	0.887	0.938	0.921	0.020
AKBL-ja-3	0.805	0.872	0.859	0.896	0.883	0.849	0.839	0.029
UE-ja-1	0.805	0.865	0.855	0.831	0.819	0.903	0.902	0.033
KIS-ja-2	0.802	0.871	0.856	0.831	0.815	0.915	0.904	0.032
AKBL-ja-1	0.800	0.869	0.847	0.889	0.873	0.849	0.825	0.030
UE-ja-3	0.800	0.866	0.855	0.823	0.812	0.913	0.911	0.033
AKBL-ja-2	0.795	0.868	0.849	0.891	0.875	0.846	0.827	0.030
KIS-ja-3	0.784	0.855	0.831	0.840	0.816	0.871	0.850	0.034
Baseline: SVM (unigram)	0.761	0.849	0.835	0.843	0.828	0.854	0.842	0.036
KIS-ja-1	0.758	0.849	0.833	0.798	0.782	0.906	0.899	0.038
Baseline: SVM (bigram)	0.752	0.843	0.830	0.838	0.820	0.848	0.845	0.037
NTTMU-ja-1	0.738	0.835	0.829	0.770	0.761	0.913	0.921	0.042
UE-ja-2	0.706	0.815	0.803	0.696	0.702	0.983	0.984	0.052
NIL-ja-1	0.680	0.749	0.742	0.862	0.845	0.662	0.671	0.052
DrG-ja-1	0.653	0.777	0.774	0.825	0.808	0.734	0.779	0.049
NTTMU-ja-3	0.614	0.775	0.773	0.740	0.720	0.814	0.840	0.055
NTTMU-ja-2	0.597	0.770	0.753	0.741	0.706	0.801	0.813	0.056
AITOK-ja-2	0.503	0.706	0.696	0.726	0.738	0.687	0.767	0.067
AITOK-ja-1	0.092	0.368	0.355	0.243	0.238	0.757	0.765	0.304

5.1 Baseline systems

5.1.1 Overview

As a baseline, two systems were constructed using a support vector machine (SVM) based on unigram features and bigram features. For feature representation, the bag-of-words (BoW) model is used in each system. A tweet message is segmented using MeCab [11] for Japanese messages, NLTK TweetTokenizer⁶ [5] for English messages, and jieba⁷ for Chinese messages. The two systems have a linear kernel, and the parameter for regularization C is set on 1.0. The baseline systems are implemented using scikit-learn (sklearn)⁸ [16].

5.1.2 Performance

The performance of the baseline measured using all the evaluation metrics is described in Section 4.2. Table 9, Ta-

ble 10, and Table 11 show the results for the Japanese, English, and Chinese subtasks, respectively.

For the Japanese and Chinese subtasks, unigram SVM performed better than bigram SVM did. On the other hand, bigram SVM outperformed unigram SVM in the English subtask. The highest average of exact match accuracy was 0.791 (English subtask) and the lowest was 0.756 (Japanese subtask).

5.2 Participating systems

5.2.1 Overview

In all, 37 systems (of nine groups) participated and had their results submitted in the MedWeb. Of these, 19 systems (of eight groups) submitted results for the Japanese subtask, 12 systems (of four groups) submitted results for the English subtask, and six systems (of two groups) submitted results for the Chinese subtask. The participating systems for the Japanese, English, and Chinese subtasks are summarized in Table 6, Table 7, and Table 8, respectively.

Table 6 shows that most of the groups applied machine

⁶<http://www.nltk.org/api/nltk.tokenize.html>

⁷<https://github.com/fxsjy/jieba>

⁸<http://scikit-learn.org/stable/>

Table 10: Performance in the English subtask (12 participating systems and two baseline systems). The results are ordered by exact match accuracy.

System ID	Exact match	F1		Precision		Recall		Hamming loss
		micro	macro	micro	macro	micro	macro	
NAIST-en-2	0.880	0.920	0.906	0.899	0.887	0.941	0.925	0.019
NAIST-en-3	0.878	0.919	0.904	0.899	0.885	0.940	0.924	0.019
NAIST-en-1	0.877	0.918	0.904	0.899	0.887	0.938	0.921	0.020
Baseline: SVM (bigram)	0.800	0.866	0.856	0.865	0.849	0.868	0.865	0.031
UE-en-1	0.789	0.858	0.848	0.846	0.831	0.871	0.876	0.034
Baseline: SVM (unigram)	0.783	0.858	0.845	0.851	0.830	0.864	0.864	0.033
NTTMU-en-2	0.773	0.856	0.849	0.807	0.796	0.911	0.918	0.036
NTTMU-en-3	0.758	0.845	0.828	0.836	0.818	0.854	0.844	0.037
UE-en-2	0.745	0.821	0.809	0.861	0.838	0.786	0.800	0.040
UE-en-3	0.739	0.820	0.815	0.870	0.851	0.776	0.795	0.040
AKBL-en-2	0.734	0.819	0.799	0.832	0.808	0.806	0.793	0.042
AKBL-en-3	0.716	0.804	0.787	0.853	0.834	0.760	0.747	0.043
NTTMU-en-1	0.619	0.770	0.777	0.734	0.733	0.809	0.835	0.056
AKBL-en-1	0.613	0.772	0.755	0.656	0.649	0.936	0.945	0.065

Table 11: Performance in the Chinese subtask (six participating systems and two baseline systems). The results are ordered by exact match accuracy.

System ID	Exact match	F1		Precision		Recall		Hamming loss
		micro	macro	micro	macro	micro	macro	
NAIST-zh-2	0.880	0.920	0.906	0.899	0.887	0.941	0.925	0.019
NAIST-zh-3	0.878	0.919	0.904	0.899	0.885	0.940	0.924	0.019
NAIST-zh-1	0.877	0.918	0.904	0.899	0.887	0.938	0.921	0.020
TUA1-zh-3	0.786	0.860	0.844	0.772	0.760	0.970	0.971	0.037
Baseline: SVM (unigram)	0.780	0.858	0.843	0.831	0.815	0.888	0.883	0.034
TUA1-zh-1	0.773	0.853	0.838	0.766	0.753	0.963	0.965	0.039
Baseline: SVM (bigram)	0.767	0.850	0.835	0.824	0.806	0.878	0.876	0.036
TUA1-zh-2	0.719	0.824	0.809	0.712	0.710	0.978	0.982	0.049

learning approaches, such as SVM (as in the baseline systems), random forests, and neural networks. Several groups constructed their own resources to enhance the original training corpus.

Similarly, for the English subtask, most of the groups applied machine learning approaches, such as SVM, random forests, and neural networks, as shown in Table 7.

The Chinese subtask had two participating groups. The one applied the same methods as the other subtasks, and the other used a logistic regression and SVM, and updated the training data using active learning.

5.2.2 Performance

The performance of the participating systems was also measured using all the evaluation metrics described in Section 4.2. Table 9, Table 10, and Table 11 show the results for the Japanese, English, and Chinese subtasks, respectively. The results in these tables are ordered by the exact match accuracy of the systems. In addition, Figure 1, Figure 2, and Figure 3 illustrate the results in the respective subtasks, ordered by (a) exact match accuracy, (b) F1 micro, and (c) Hamming loss.

For the Japanese subtask, the best system, NAIST-ja-2, achieved 0.88 in exact match accuracy, 0.92 in F-measure,

and 0.019 in Hamming loss, as shown in Table 9. The averages across the participating groups and the baseline systems were 0.72, 0.82, and 0.051, respectively. The rank order of the top four systems was the same in all measures. Ten of the 17 participating systems outperformed both baseline systems, as shown in Figure 1. The systems of the AKBL group and the KIS group were constructed using an SVM, as in the baseline systems. The AKBL group’s results indicate that their system is effective in terms of using additional language resources. The KIS group switched their methods between an SVM and a rule-based method, depending on the confidence factor.

For the English subtask, the best system, NAIST-en-2, achieved 0.88 in exact match accuracy, 0.92 in F-measure, and 0.019 in Hamming loss, as shown in Table 10. The system is constructed using the same method as that used in the Japanese subtask. The averages across the participating groups and the baseline systems were 0.77, 0.85, and 0.037, respectively. Only the top three of the 12 participating systems showed better performance than both baseline systems, as shown in Figure 2.

For the Chinese subtask, the best system, NAIST-zh-2, achieved 0.88 in exact match accuracy, 0.92 in F-measure, and 0.019 in Hamming loss, as shown in Table 11. The

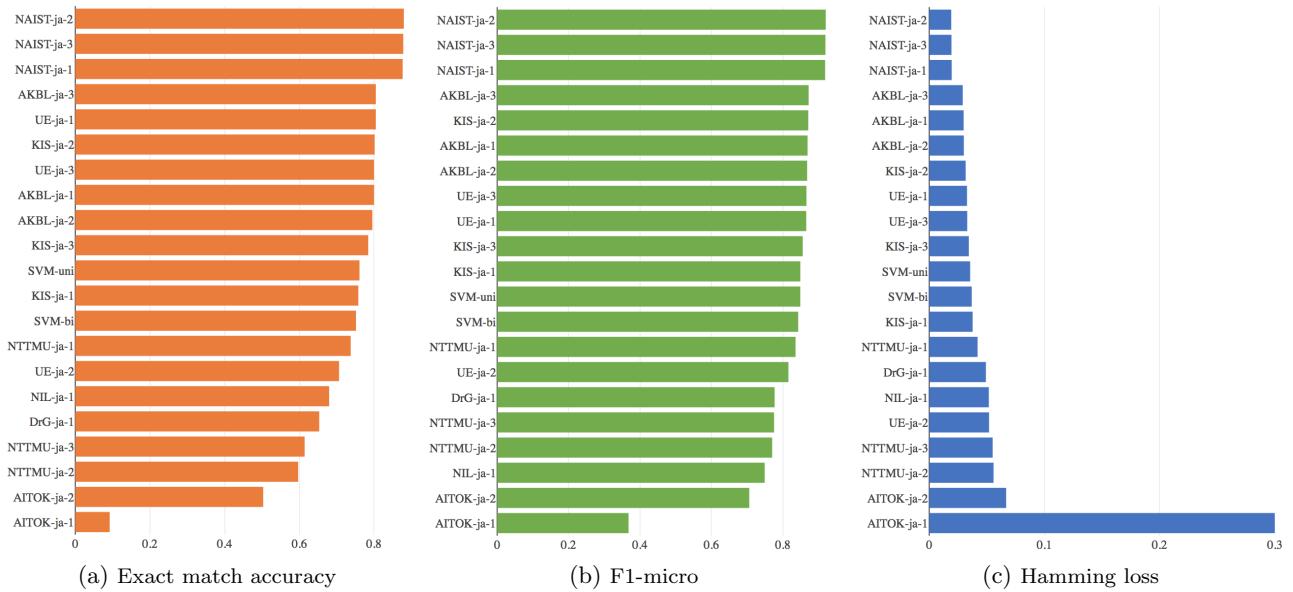


Figure 1: Performance in the Japanese subtask (19 participating systems and two baseline systems). (a) Exact match accuracy, (b) F1-micro, and (c) Hamming loss. Higher scores are better in (a) and (b), and lower scores are better in (c).

system is constructed using the same method as that used in the Japanese and English subtasks. The averages across the participating groups and the baseline systems were 0.81, 0.88, and 0.032, respectively. Only the top four of the six participating systems showed better performance than the baseline system (SVM unigram) in exact match accuracy, as shown in Figure 3.

6. DISCUSSION

6.1 Machine learning advantage

One of characteristics of the MedWeb task is to use a multi-label corpus. Because the multi-label classification is a complex task, the performance of straightforward approaches, such as rule-based and keyword-based methods, is relatively lower than that of other approaches. In contrast, we found that machine learning (e.g., an SVM) achieved better performance. Of the participant systems, the ensemble of a hierarchical attention network (HAN) and a deep convolutional neural network (CNN) with loss functions, employed by the NAIST group, achieved the best performance in all subtasks.

Note that previous NTCIR Medical tasks and MedNLP workshops [13, 3, 4] have shown that the rule-based approach is still competitive with the machine learning approaches. One of the reasons for this was the small size of the corpus they used. Although the size of the corpus is also limited in this task, this result shows the advantage of the complex machine learning, indicating the advancement of machine learning techniques.

6.2 Language comparison

The MedWeb task provided a cross-language corpus. Although this is another characteristic of this task, only one group (NAIST) challenged all subtasks, which was fewer than we expected. The Japanese subtask had the highest

participation (19 systems from eight groups) and the Chinese subtask had the lowest participation (six systems from only two groups), which was also lower than expected. The performance varied depending on the subtasks.

Figure 4 shows the distribution of the three metric scores of the systems in each subtask. For the Japanese subtask, the performance varied widely, relative to that of the other subtasks. Although the Chinese subtask had the lowest participation, their performance was relatively high. The four groups that participated in the Japanese subtask also challenged the English subtask, with most achieving worse results in the English subtask. This indicates that the difficulty of classification is Japanese, English, and Chinese, in increasing order of difficulty. This is a surprising result, because most of the groups come from Japan, which means they are familiar with the Japanese NLP.

This might indicate that the Chinese language has less ambiguity in clinical factuality analyses. Another possibility is that the process we used to generate the corpora had a language bias. For example, the translations from Japanese to English and Chinese may have reduced the ambiguity of the language in each case. In order to test for a language bias, experiments based on different directions of translation are necessary. This is left for future work.

Note that the baseline systems performed best in the English subtask. This indicates that the standard settings for the SVM are effective in terms of classifying English tweets.

6.3 Limitations

The corpora provided by the MedWeb task have limitations. The first is the generating process. For example, the translation process might bias the results, as described above. In addition, our pseudo-tweets do not include several tweet-specific features such as reply, retweet, hashtag, url, and so on.

Another limitation is the size of each corpus (1,920 mes-

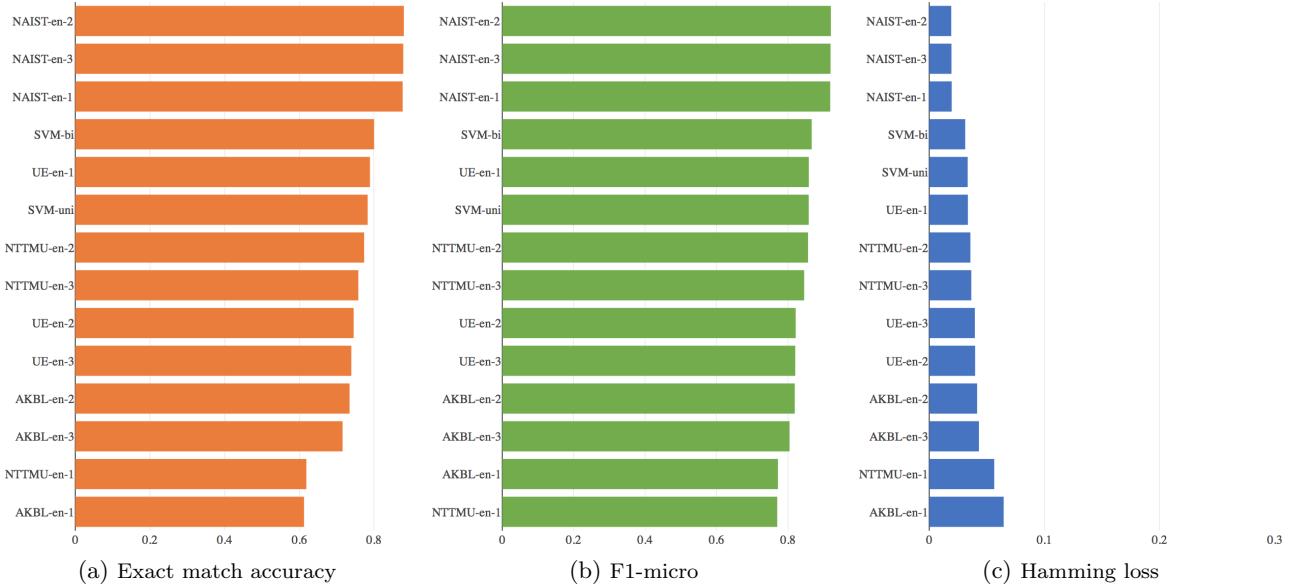


Figure 2: Performance in the English subtask (12 participating systems and two baseline systems). (a) Exact match accuracy, (b) F1-micro, and (c) Hamming loss. Higher scores are better in (a) and (b), and lower scores are better in (c).

sages are used as training data, and 640 messages are used as test data). Regardless of these limitations, we believe this is a valuable attempt to generate and share a cross-language corpus consisting of multi-label pseudo-tweets.

Even though our corpus has some limitations, we still believe it is helpful as a benchmark for tweet-based applications, because it is freely available and covers multiple languages.

Acknowledgements

This work was supported by Japan Agency for Medical Research and Development (Grant Number: 16768699) and JST ACT-I. We appreciate annotators in Social Computing labo. at Nara Institute of Science and Technology for their efforts on generating the corpus. We also greatly appreciate the NTCIR-13 chairs for their efforts on organizing the NTCIR-13 workshop. Lastly we thank all the participants for their contributions to the NTCIR-13 MedWeb task.

7. CONCLUSION

This paper provided an overview of the NTCIR-13 MedWeb task. This task is designed as a more generalized task for public surveillance, focusing on social media (e.g., Twitter). In particular, the task's goal is to classify symptom-related messages. This task has two characteristics: (1) multi-label (cold, cough, diarrhea, fever, hay fever, headache, flu, and runny nose) and (2) cross-language (Japanese, English, and Chinese). In total, nine groups (37 systems) participated in the MedWeb task. Specifically, eight groups (19 systems) participated in the Japanese subtask, four groups (12 systems) participated in the English subtask, and two groups (six systems) participated in the Chinese subtask. The results empirically demonstrate that a machine learning approach is effective in terms of tweet classification, providing a foundation for future, deeper approaches.

8. REFERENCES

- [1] H. Achrekar, A. Gandhe, R. Lazarus, S. Yu, and B. Liu. Twitter Improves Seasonal Influenza Prediction. In *Proc. of the International Conference on Health Informatics*, pages 61–70, 2012.
- [2] E. Aramaki, S. Maskawa, and M. Morita. Twitter Catches the Flu: Detecting Influenza Epidemics Using Twitter. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1568–1576, 2011.
- [3] E. Aramaki, M. Morita, Y. Kano, and T. Ohkuma. Overview of the ntcir-11 mednlp-2 task. In *Proc. of the 11th NTCIR Conference*, pages 147–154, 2014.
- [4] E. Aramaki, M. Morita, Y. Kano, and T. Ohkuma. Overview of the ntcir-12 mednlpdoc task. In *Proc. of the 12th NTCIR Conference on Evaluation of Information Access Technologies*, pages 71–75, 2016.
- [5] S. Bird. NLTK: The Natural Language Toolkit. In *Proc. of the COLING/ACL on Interactive Presentation Sessions (COLING-ACL)*, pages 69–72, 2006.
- [6] D. A. Broniatowski, M. J. Paul, and M. Dredze. National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic. *PLOS ONE*, 8:e83672, 2013.
- [7] L. E. Charles-Smith, T. L. Reynolds, M. A. Cameron, M. Conway, E. H. Y. Lau, J. M. Olsen, J. A. Pavlin, M. Shigematsu, L. C. Streichert, K. J. Suda, and C. D. Corley. Using Social Media for Actionable Disease Surveillance and Outbreak Management: A Systematic Literature Review. *PLOS ONE*, 10:e0139701, 2015.
- [8] A. Culotta. Towards Detecting Influenza Epidemics by Analyzing Twitter Messages. In *Proc. of the First Workshop on Social Media Analytics (SOMA)*, pages 115–122, 2010.

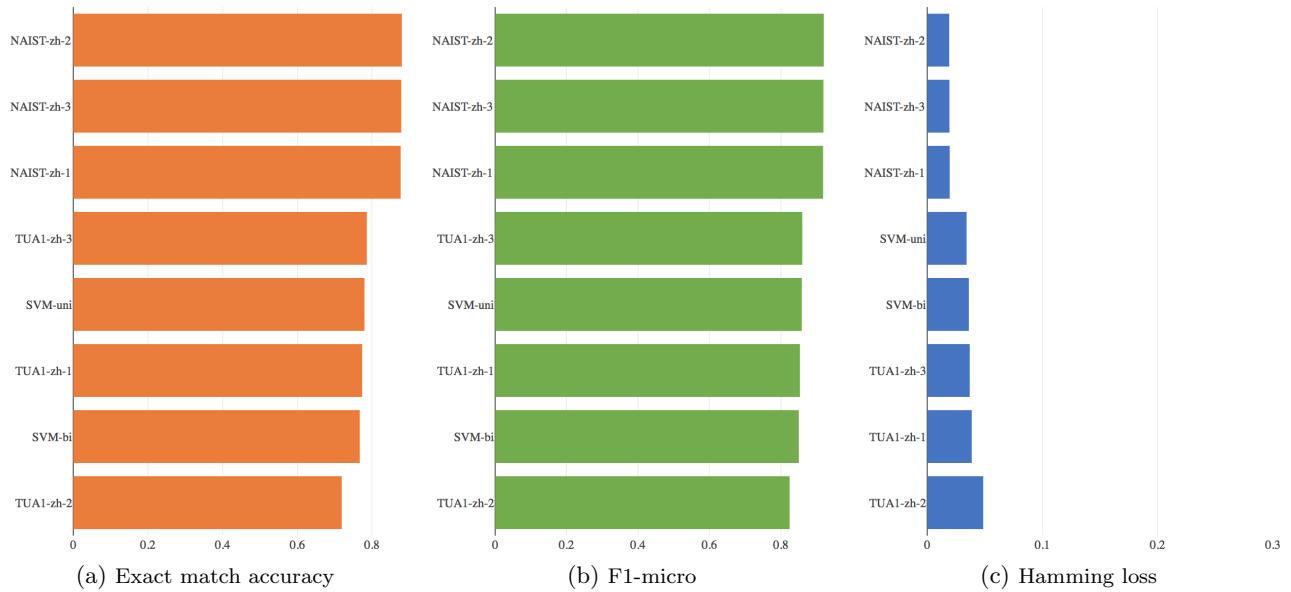


Figure 3: Performance in the Chinese subtask (six participating systems and two baseline systems). (a) Exact match accuracy, (b) F1-micro, and (c) Hamming loss. Higher scores are better in (a) and (b), and lower scores are better in (c).

- [9] F. Gesualdo, G. Stilo, E. Agricola, M. V. Gonfiantini, E. Pandolfi, P. Velardi, and A. E. Tozzi. Influenza-Like Illness Surveillance on Twitter through Automated Learning of Naïve Language. *PLOS ONE*, 8:e82489, 2013.
- [10] H. Iso, S. Wakamiya, and E. Aramaki. Forecasting Word Model: Twitter-based Influenza Surveillance and Prediction. In *Proc. of the International Conference on Computational Linguistics (COLING)*, pages 76–86, 2016.
- [11] T. Kudo, K. Yamamoto, and Y. Matsumoto. Applying Conditional Random Fields to Japanese Morphological Analysis. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 230–237, 2004.
- [12] A. Lamb, M. J. Paul, and M. Dredze. Separating fact from fear: Tracking flu infections on Twitter. In *Proc. of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2013.
- [13] M. Morita, Y. Kano, T. Ohkuma, and E. Aramaki. Overview of the NTCIR-10 MedNLP task. In *Proc. of the 10th NTCIR Conference*, pages 696–701, 2013.
- [14] U. Ozlem. Second i2b2 workshop on natural language processing challenges for clinical records. In *AMIA Annu Symp Proc.*, pages 1252–1253, 2008.
- [15] M. J. Paul, M. Dredze, and D. Broniatowski. Twitter Improves Influenza Forecasting. *PLOS Currents*, 6, 2014.
- [16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, pages 2825–2830, 2011.
- [17] P. M. Polgreen, Y. Chen, D. M. Pennock, F. D. Nelson, and R. A. Weinstein. Using Internet Searches for Influenza Surveillance. *Clinical Infectious Diseases*, 47:1443–1448, 2008.
- [18] A. Signorini, A. Segre, and P. Polgreen. The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic. *PLOS ONE*, 6:e19467, 2011.
- [19] E. Voorhees and W. Hersh. Overview of the TREC 2012 Medical Records Track. In *Proc. of the Twentieth Text REtrieval Conference*, 2012.
- [20] S. Wang, M. J. Paul, and M. Dredze. Social media as a sensor of air quality and public response in China. *Journal of medical Internet research*, 17:e22, 2015.
- [21] M. L. Zhang and Z. H. Zhou. A Review on Multi-Label Learning Algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, 2014.

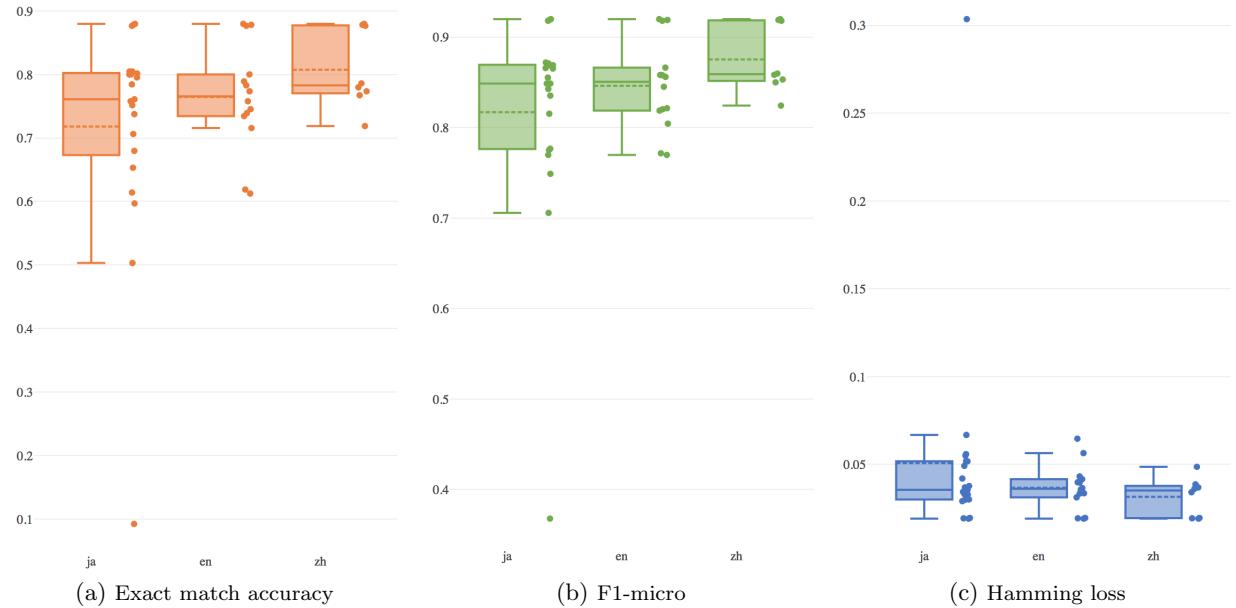


Figure 4: Statistical summary of the performance in each of the subtasks (ja: Japanese, en: English, and zh: Chinese). (a) Exact match accuracy, (b) F1-micro, and (c) Hamming loss. Higher scores are better in (a) and (b), and lower scores are better in (c). The bottom and top of a box are the first and third quartiles, the band inside the box is the median, and the dotted band inside the box is the mean. Dots on the right side of the box represent the distribution of values of participating systems.