Who caught a cold? — Identifying the subject of a symptom

Shin Kanouchi [†], Mamoru Komachi [†], Naoaki Okazaki [‡],
Eiji Aramaki[§], and Hiroshi Ishikawa [†]

[†] Tokyo Metropolitan University, {kanouchi-shin at ed., komachi at, ishikawh at}tmu.ac.jp
[‡] Tohoku University, okazaki at ecei.tohoku.ac.jp
[§]Kyoto University, eiji.aramaki at gmail.com

Abstract

The development and proliferation of social media services has led to the emergence of new approaches for surveying the population and addressing social issues. One popular application of social media data is health surveillance, e.g., predicting the outbreak of an epidemic by recognizing diseases and symptoms from text messages posted on social media platforms. In this paper, we propose a novel task that is crucial and generic from the viewpoint of health surveillance: estimating a subject (carrier) of a disease or symptom mentioned in a Japanese tweet. By designing an annotation guideline for labeling the subject of a disease/symptom in a tweet, we perform annotations on an existing corpus for public surveillance. In addition, we present a supervised approach for predicting the subject of a disease/symptom. The results of our experiments demonstrate the impact of subject identification on the effective detection of an episode of a disease/symptom. Moreover, the results suggest that our task is independent of the type of disease/symptom.

1 Introduction

Social media services, including Twitter and Facebook, provide opportunities for individuals to share their experiences, thoughts, and opinions. The wide use of social media services has led to the emergence of new approaches for surveying the population and addressing social issues. One popular application of social media data is flu surveillance, i.e., predicting the outbreak of influenza epidemics by detecting mentions of flu infections on social media platforms (Culotta, 2010; Lampos and Cristianini, 2010; Aramaki et al., 2011; Paul and Dredze, 2011; Signorini et al., 2011; Collier, 2012; Dredze et al., 2013; Gesualdo et al., 2013; Stoové and Pedrana, 2014).

Previous studies mainly relied on shallow textual clues in Twitter posts in order to predict the number of flu infections, e.g., the number of occurrences of specific keywords (such as "flu" or "influenza") on Twitter. However, such a simple approach can lead to incorrect predictions. Broniatowski et al. (2013) argued that media attention increases chatter, i.e., the number of tweets that mention the flu without the poster being actually infected. Examples include, "I don't wish the flu on anyone" and "A Harry Potter actor hospitalised after severe flu-like syndromes." Lazer et al. (2014) reported large errors in Google Flu Trends (Carneiro and Mylonakis, 2009) on the basis of a comparison with the proportion of doctor visits for influenza-like illnesses.

Lamb et al. (2013) aimed to improve the accuracy of detecting mentions of flu infections. Their method trains a binary classifier to distinguish tweets reporting flu infections from those expressing concern or awareness about the flu, e.g., "Starting to get worried about swine flu." Accordingly, they reported encouraging results (e.g., better correlations with CDC trends), but their approach requires supervision data and a lexicon (word class features) specially designed for the flu. Moreover, even though this method is a reasonable choice for improving the accuracy, it is not readily applicable to other types of diseases (e.g., dengue fever) and symptoms (e.g., runny nose), which are also important for public health (Velardi et al., 2014).

In this paper, we propose a more generalized task setting for public surveillance. In other words, our objective is to *estimate the subject (carrier) of a disease or symptom mentioned in a Japanese tweet.* More specifically, we are interested in determining who has a disease/symptom

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pages 1660–1670,
Beijing, China, July 26-31, 2015.

(if any) in order to examine whether the poster suffers from the disease or symptom. For example, given the sentence "I caught a cold," we would predict that the first person ("I," i.e., the poster) is the subject (carrier) of the cold. On the other hand, we can ignore the sentence, "The TV presenter caught a cold" only if we predict that the subject of the cold is the third person, who is at a different location from the poster.

Although the task setting is simple and intuitive, we identify several key challenges in this study.

1. **Novel task setting.** The task of identifying the subject of a disease/symptom is similar to predicate-argument structure (PAS) analysis for nominal predicates (Meyers et al., 2004; Sasano et al., 2004; Komachi et al., 2007; Gerber and Chai, 2010). However, these studies do not treat diseases (e.g., "influenza") and symptoms (e.g., "headache") as nominal predicates. To the best of our knowledge, this task has not been explored in natural language processing (NLP) thus far.

2. **Identifying whether the subject has a disease/symptom.** Besides the work on PAS analysis for nominal predicates, the most relevant work is PAS analysis for verb predicates. However, our task is not as simple as predicting the subject of the verb governing a disease/symptom-related noun. For example, the subject of the verb "beat" is the first person "I" in the sentence "I beat the flu," but this does not imply that the poster has the flu. At the same time, we can use a variety of expressions for indicating an infection, e.g., "I'm still sick!! This flu is just incredible...," "I can feel the flu bug in me," and "I tested positive for the flu."

3. **Omitted subjects.** We often come across tweets with omitted subjects, e.g., "Down with the flu feel" and "Thanks the flu for striking in hard this week" even in English tweets. Because the first person is omitted frequently, it is important to predict omitted subjects from the viewpoint of the application (public surveillance).

In this paper, we present an approach for identifying the subjects of various types of diseases and symptoms. The contributions of this paper are three-fold.

1. In order to explore a novel and general task setting, we design an annotation guideline for labeling a subject of a disease/symptom in a tweet, and we deliver annotations in an existing corpus for public surveillance. Further, we propose a method for predicting the subject of a disease/symptom by using the annotated corpus.

2. The experimental results show that the task of identifying subjects is independent of the type of diseases/symptom. We verify the possibility of transferring supervision data to different targets of diseases and symptoms. In other words, we verify that it is possible to utilize the supervision data for a particular disease/symptom to improve the accuracy of predicting subjects of another disease/symptom.

3. In addition, the experimental results demonstrate the impact of identifying subjects on improving the accuracy of the downstream application (identification of an episode of a disease/symptom).

The remainder of this paper is organized as follows. Section 2 describes the corpus used in this study as well as our annotation work for identifying subjects of diseases and symptoms. Section 3.1 presents our method for predicting subjects on the basis of the annotated corpus. Sections 3.2 and 3.3 report the performance of the proposed method. Section 3.4 describes the contributions of this study toward identifying episodes of diseases and symptoms. Section 4 reviews some related studies. Finally, Section 5 summarizes our findings and concludes the paper with a brief discussion on the scope for future work.

2 Corpus

2.1 Target corpus

We used a Japanese corpus for public surveillance of diseases and symptoms (Aramaki et al., 2011). The corpus targets seven types of diseases and symptoms: *cold*, *cough*, *headache*, *chill*, *runny nose*, *fever*, and *sore throat*. Tweets containing keywords for each disease/symptom were collected using the Twitter Search API: for example, tweets about *sore throat* were collected using the query "(sore OR pain) AND throat". Further,

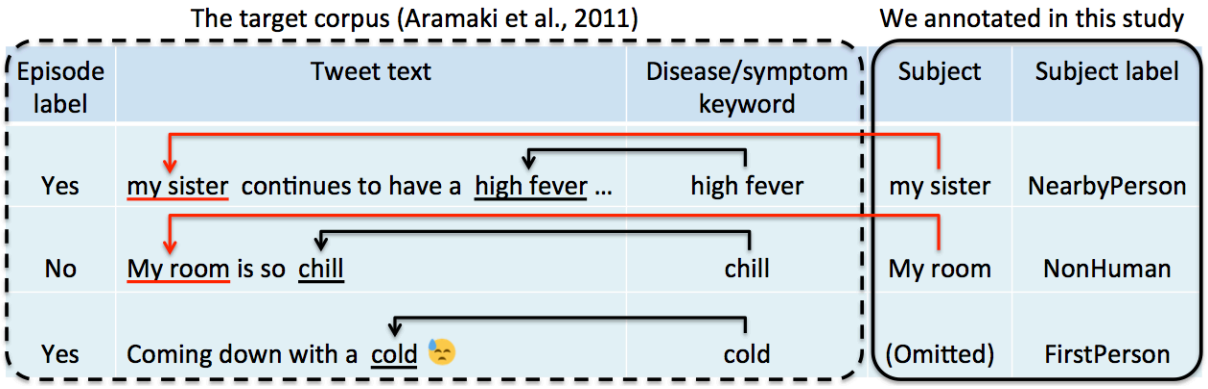

Figure 1: Examples of annotations of subject labels.

Subject label	Definition	Example
FIRSTPERSON	The subject of the disease/symptom is the poster of the tweet.	I wish I have fever or something so that I don't have to go to school.
NEARBYPERSON	The subject of the disease/symptom is a person whom the poster can directly see or hear.	my sister continues to have a high fever...
FARAWAYPERSON	The subject of the disease/symptom is a person who is at a different location from the poster.	@***** does sour stuff give you a headache?
NONHUMAN	The subject of the disease/symptom is not a person. Alternatively, the sentence does not describe a disease/symptom but a phenomenon or event related to the disease/symptom.	My room is so chill. But I like it.
NONE	The subject of the disease/symptom does not exist. Alternatively, the sentence does not mention an occurrence of a disease/symptom.	I hate buyin cold medicine cuz I never know which one to buy

Table 1: Definitions of subject labels and example tweets.

the corpus consists of 1,000 tweets for each disease/symptom besides *cold*, and 5,000 tweets for *cold*. The corpus was collected through whole years 2007-2008. This period was not in the A/H1N1 flu pandemic season.

An instance in this corpus consists of a tweet text (in Japanese) and a binary label (*episode label*, hereafter) indicating whether someone near the poster has the target disease/symptom[1]. A positive episode indicates an occurrence of the disease/symptom. In this study, we disregarded instances of *sore throat* in the experiments because most such instances were positive episodes[2].

[1]This label is positive if someone mentioned in the tweet is in the same prefecture as the poster. This is because the corpus was designed to survey the spread of a disease/symptom in every prefecture.

[2]In Japanese tweets, *sore throat* or *throat pain* mostly describes the health condition of the poster.

2.2 Annotating subjects

In this study, we annotated the subjects of diseases and symptoms in the corpus described in Section 2.1. Specifically, we annotated the subjects in 500 tweets for each disease/symptom (except for *sore throat*). Thus, our corpus includes a total of 3,000 tweets in which the subjects of diseases and symptoms are annotated.

Figure 1 shows examples of annotations in this study. Episode labels, tweet texts, and disease/symptom keywords were annotated by Aramaki et al. (2011) in the corpus.

We annotated the subject labels of the diseases/symptoms in each tweet and identified those who had the target disease/symptom. The subject labels indicate those who have the corresponding disease/symptom; they are described in detail

Label	FIRSTPERSON	NEARBYPERSON	FARAWAYPERSON	NONHUMAN	NONE	Total
# tweets	2,153	129	201	40	401	2,924
# explicit subjects	70 (3.3%)	112 (86.8%)	175 (87.1%)	38 (95.0%)	0 (0.0%)	395
# positive episodes	1,833	99	2	0	16	1,950
# negative episodes	320	30	199	40	385	974
Positive ratio	85.1%	76.7%	1.0%	0.0%	4.0%	66.7%

Table 2: Associations between subject labels and positive/negative episodes of diseases and symptoms.

herein.

In addition to the subject labels, we annotated the text span that indicates a subject. However, the subjects of diseases/symptoms are often omitted in tweet texts. Example 3 in Figure 1 shows a case in which the subject is omitted. The information as to whether the subject is omitted is useful for analyzing the difficulty in predicting the subject of a disease/symptom.

Table 1 lists the definitions of the subject labels with tweeted examples. Because it is important to distinguish the primary information (information that is observed and experienced by the poster) from the secondary information (information that is broadcasted by the media) for the application of public surveillance, we introduced five labels: FIRSTPERSON, NEARBYPERSON, FARAWAYPERSON, NONHUMAN, and NONE.

FIRSTPERSON is assigned when the subject of the disease/symptom is the poster of the tweet. When annotating this label, we ignore the modality or factuality of the event of acquiring the disease/symptom. For example, the example tweet corresponding to FIRSTPERSON in Table 1 does not state that the poster has a fever but only that the poster has a desire to have a fever. Although such tweets may be inappropriate for identifying a disease/symptom, this study focuses on identifying the possessive relation between a subject and a disease/symptom. The concept underlying this decision is to divide the task of public surveillance into several sub-tasks that are sufficiently generalized for use in other NLP applications. Therefore, the task of analyzing the modality lies beyond of scope of this study (Kitagawa et al.,). We apply the same criterion to the labels NEARBYPERSON, FARAWAYPERSON, and NONHUMAN.

NEARBYPERSON is assigned when the subject of the disease/symptom is a person whom the poster can directly see or hear. In the original corpus (Aramaki et al., 2011), a tweet is labeled as positive if the person having a disease/symptom is in the same prefecture as the poster. However, it is extremely difficult for annotators to judge from a tweet whether the person mentioned in the tweet is in the same prefecture as the poster. Nevertheless, we would like to determine from a tweet whether the poster can directly see or hear a patient. For these reasons, we introduced the label NEARBYPERSON in this study.

FARAWAYPERSON applies to all cases in which the subject is a human, but not classified as FIRSTPERSON or NEARBYPERSON. This category frequently includes tweeted replies, as in the case of the example corresponding to FARAWAYPERSON in Table 1. We assign FARAWAYPERSON to such sentences because we are unsure whether the subject of the symptom is a person whom the poster can physically see or hear.

NONHUMAN applies to cases in which the subject is not a human but an object or a concept. For example, a sentence with the phrase "My room is so chill" is annotated with this label.

NONE indicates that the sentence does not mention a target disease or symptom even though it includes a keyword for the disease/symptom.

In order to investigate the inter-annotator agreement, we sampled 100 tweets of *cold* at random, and examined the Cohen's κ statistic by two annotators. The κ statistic is 0.83, indicating a high level agreement (Carletta, 1996).

Table 2 reports the distribution of subject labels in the corpus annotated in this study. When the subject of a disease/symptom is FIRSTPERSON, only 3.3% of the tweets have explicit textual clues for the first person[3]. In other words, when the subject of a disease/symptom is FIRSTPERSON, we rarely find textual clues in tweets. In contrast, there is a greater likelihood of finding explicit clues for NEARBYPERSON, FARAWAYPERSON, and NONHUMAN subjects.

Table 2 also lists the probability of positive episodes given a subject label, i.e., the positive ratio. The likelihood of a positive episode

[3]This ratio may appear to be extremely low, but it is very common to omit first person pronouns in Japanese sentences.

is extremely high when the subject label of a disease/symptom is FIRSTPERSON (85.1%) or NEARBYPERSON (76.7%). In contrast, FARAWAYPERSON, NONHUMAN, and NONE subjects represent negative episodes (less than 5.0%). These facts suggest that identifying subject labels can improve the accuracy of predicting patient labels for diseases and symptoms.

3 Experiment

3.1 Subject classifier

We built a classifier to predict a subject label for a disease/symptom mentioned in a sentence by using the corpus described in the previous section. In our experiment, we merged training instances having the label NONHUMAN with those having the label NONE because the number of NONHUMAN instances was small and we did not need to distinguish the label NONHUMAN from the label NONE in the final episode detection task. Thus, the classifier was trained to choose a subject label from among FIRSTPERSON, NEARBYPERSON, FARAWAYPERSON, and NONE. We discarded instances in which multiple diseases or symptoms are mentioned in a tweet as well as those in which multiple subjects are associated with a disease/symptom in a tweet. In addition, we removed text spans corresponding to retweets, replies, and URLs; the existence of these spans was retained for firing features. We trained an L2-regularized logistic regression model using Classias 1.1[4]. The following features were used.

Bag-of-Words (BoW). Nine words included before and after a disease/symptom keyword. We split a Japanese sentence into a sequence of words using a Japanese morphological analyzer, MeCab (ver.0.98) with IPADic (ver.2.7.0)[5].

Disease/symptom word (Keyword). The surface form of the disease/symptom keyword (e.g. "cold" and "headache").

2,3-gram. Character-based bigrams and trigrams before and after the disease/symptom keyword within a window of six letters.

URL. A boolean feature indicating whether the tweet includes a URL.

[4] http://www.chokkan.org/software/classias/

[5] http://taku910.github.io/mecab/

Feature	Micro F1	Macro F1
BoW (baseline)	77.2	42.2
BoW + Keyword	81.9	53.6
BoW + 2,3-gram	79.1	46.1
BoW + URL	77.3	42.7
BoW + RT & reply	80.0	47.1
BoW + NearWord	77.6	46.8
BoW + FarWord	77.3	42.7
BoW + Title word	77.1	42.7
BoW + Tweet length	77.4	43.3
BoW + Is-head	77.6	43.5
All features	**84.0**	**61.8**

Table 3: Performance of the subject classifier.

RT & reply. Boolean features indicating whether the tweet is a reply or a retweet.

Word list for NEARBYPERSON (NearWord). A boolean feature indicating whether the tweet contains a word that is included in the lexicon for NEARBYPERSON. We manually collected words that may refer to a person who is near the poster, e.g., "girlfriend," "sister," and "staff." The NearWord list includes 97 words.

Word list for FARAWAYPERSON (FarWord). A boolean feature indicating whether the tweet contains a word that is included in the lexicon for FARAWAYPERSON. Similarly to the NearWord list, we manually collected 50 words (e.g., "infant") for compiling this list.

Title word. A boolean feature indicating whether the tweet contains a title word accompanied by a proper noun. The list of title words includes expressions such as "さん" and "くん" (roughly corresponding to "Ms" and "Mr") that describe the title of a person.

Tweet length. Three types of boolean features that fire when the tweet has less than 11 words, 11 to 30 words, and more than 30 words, respectively.

Is-head. A boolean feature indicating whether the word following a disease/symptom keyword is a noun. In Japanese, when the word following a disease/symptom keyword is a noun, the disease/symptom keyword is unlikely to be the head of the noun phrase.

Correct/predicted label	FIRSTPERSON		NEARBY.		FARAWAY.		NONE		Total
FIRSTPERSON	**2,084**	(−15)	6	(+1)	25	(+21)	38	(−7)	2,153
NEARBYPERSON	80	(−20)	**41**	(+29)	4	(−5)	4	(−4)	129
FARAWAYPERSON	88	(−49)	8	(+2)	**89**	(+46)	16	(+1)	201
NONE	174	(−158)	2	(+1)	10	(+4)	**255**	(+153)	441
Total predictions	2,426	(−237)	57	(+33)	128	(+66)	313	(+137)	2,924

Table 4: Confusion matrix between predicted and correct subject labels.

3.2 Evaluation of the subject classifier

Table 3 reports the performance of the subject classifier measured via five-fold cross validation. We used 3,000 tweets corresponding to six types of diseases and symptoms for this experiment. The Bag-of-Words (BoW) feature achieved micro and macro F1 scores of 77.2 and 42.2, respectively. When all the features were used, the performance was boosted, i.e., micro and macro F1 scores of 84.0 and 61.8 were achieved. Features such as disease/symptom keywords, retweet & reply, and the lexicon for NEARBYPERSON were particularly effective in improving the performance.

The surface form of the disease/symptom keyword was found to be the most effective feature in this task, the reasons for which are discussed in Section 3.3.

A retweet or reply tweet provides evidence that the poster has interacted with another person. Such meta-linguistic features may facilitate semantic and discourse analysis in web texts. However, this feature is mainly limited to tweets.

The lexicon for NEARBYPERSON provided an improvement of 4.6 points in terms of the macro F1 score. This is because (i) around 90% of the subjects for NEARBYPERSON were explicitly stated in the tweets and (ii) the vocabulary of people near the poster was limited.

Table 4 shows the confusion matrix between the correct labels and the predicted labels. The diagonal elements (in bold face) represent the number of correct predictions. The figures in parentheses denote the number of instances for which the baseline feature set made incorrect predictions, but the full feature set made correct predictions. For example, the classifier predicted NEARBYPERSON subjects 48 times; 34 out of 48 predictions were correct. The full feature set increased the number of correct predictions by 22.

From the diagonal elements (in bold face), we can confirm that the number of correct predictions increased significantly from the baseline case, except for FIRSTPERSON. One of the reasons for the improved accuracy of NONE prediction is the imbalanced label ratio of each disease/symptom. NONE accounts for 14% of the entire corpus, but only 5% of the *runny nose* corpus. On the other hand, NONE accounts for more than 30% of the *chill* corpus. The disease/symptom keyword feature adjusts the ratio of the subject labels for each disease/symptom, and the accuracy of subject identification is improved.

As compared to the baseline case, the number of FIRSTPERSON cases that were predicted as FARAWAYPERSON increased. Such errors may be attributed to the reply feature. According to our annotation scheme, FARAWAYPERSON contains many reply tweets. Because the reply & retweet features make the second-largest contribution in our experiment, the subject classifier tends to output FARAWAYPERSON if the tweet is a reply.

Table 5 summarizes the subject classification results comparing the case in which the subject of a disease/symptom exists in the tweet with that in which the subject does not exist. The prediction of FIRSTPERSON is not affected by the presence of the subject because FIRSTPERSON subjects are often omitted (especially in Japanese tweets). The prediction of NEARBYPERSON and FARAWAYPERSON is difficult if the subject is not stated explicitly. In contrast, it is easy to correctly predict NONE even though the subject is not expressed explicitly. This is because it is not easy to capture a variety of human-related subjects using Bag-of-Words, N-gram, or other simple features used in this experiment.

3.3 Dependency on diseases/symptoms

The experiments described in Section 3.2 use training instances for all types of diseases and symptoms. However, each disease/symptom may have a set of special expressions for describing the state of an episode. For example, even though "catch a cold" is a common expression, we cannot

Subject	FIRSTPERSON	NEARBYPERSON	FARAWAYPERSON	NONE
# Explicit	66/69 (95.7%)	40/112 (35.7%)	79/174 (45.4%)	1/26 (3.8%)
# Omitted	2,018/2,084 (96.8%)	1/17 (5.9%)	10/27 (37.0%)	254/415 (61.2%)
# Total	2,084/2,153 (96.8%)	41/129 (31.8%)	89/201 (44.3%)	255/441 (57.8%)

Table 5: Subject classification results comparing explicit subjects with omitted subjects.

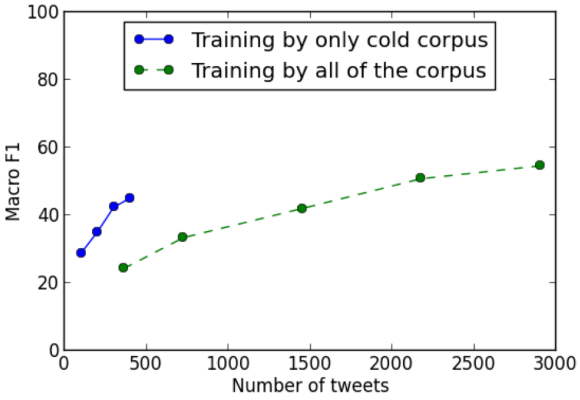

Figure 2: F1 scores for predicting subjects of *cold* with different types and sizes of training data.

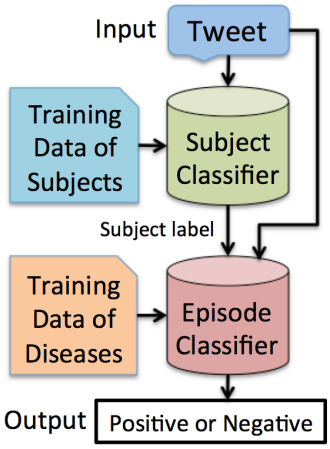

Figure 3: Overall structure of the system.

say "catch a fever" by combining the verb "catch" and the disease "fever." The corpus developed in Section 2.2 can be considered as the supervision data for weighting linguistic patterns that connect diseases/symptoms with their subjects. This viewpoint raises another question: how strongly does the subject classifier depend on specific diseases and symptoms?

In order to answer this question, we compare the performance of recognizing subjects of *cold* when using the training instances for all types of diseases and symptoms with that when using only the training instances for the target disease/symptom. Figure 2 shows the macro F1 scores with all training instances (dotted line) and with only *cold* training instances (solid line)[6].

In this case, training with *cold* instances is naturally more efficient than training with other types of diseases/symptoms. When trained with 400 instances only for *cold*, the classifier achieved an F1 score of 45.2. Moreover, we confirmed that adding training instances for other types of diseases/symptoms improved the F1 score: the maximum F1 score was 54.6 with 2,900 instances. These results indicate the possibility of building a subject classifier that is independent of specific diseases/symptoms but applicable to a variety of diseases/symptoms. We observed a similar tendency for other types of diseases/symptoms.

3.4 Contributions to the episode classifier

The ultimate objective of this study is to detect outbreaks of epidemics by recognizing diseases and symptoms. In order to demonstrate the contributions of this study, we built an *episode classifier* that judges whether the poster or a person close to the poster suffers from a target disease/symptom. Figure 3 shows the overall structure of the system. Given a tweet, the system predicts the subject label for a disease/symptom, and integrates the predicted subject label as a feature for the episode classifier. In addition to the features used in Aramaki et al. (2011), we included binary features, each of which corresponds to a subject label predicted by the proposed method. We trained an L2-regularized logistic regression model using Classias 1.1.

Table 6 summarizes the performance of the episode classifier with different settings: without subject labels (baseline), with predicted subject la-

[6]For the solid line, we used 500 instances of "cold" as a test set, and we plotted the learning curve by increasing the number of training instances for other diseases/symptoms. For the dotted line, we fixed 100 instances for a test set, and we plotted the learning curve by increasing the number of training instances (100, 200, 300, and 400).

Setting	Cold	Cough	Headache	Chill	Runny nose	Fever	Macro F1
Baseline (BL)	84.4	88.5	90.8	75.9	89.2	78.1	84.5
BL + predicted subjects	85.0	88.3	90.7	81.4	89.4	80.2	85.8
BL + gold-standard subjects	87.7	92.6	93.5	88.5	91.4	88.6	90.4

Table 6: Performance of the episode classifier.

bels , and with gold-standard subject labels. We measured the F1 scores via five-fold cross validation[7]. Further, we confirmed the contribution of subject label prediction, which achieved an improvement of 1.3 points over the baseline method (85.8 vs. 84.5). When using the gold-standard subject labels, the episode classifier achieved an improvement of 5.9 points. These results highlight the importance of recognizing a subject who has a disease/symptom using the episode classifier.

Considering the F1 score for each disease/symptom, we observed the largest improvement for *chill*. This is because the Japanese word for "chill" has another meaning *a cold air mass*. When the word "chill" stands for *a cold air mass* in a tweet, the subject for "chill" is NONE. Therefore, the episode classifier can disambiguate the meaning of "chill' on the basis of the subject labels. Similarly, the subject labels improved the performance for "fever".

In contrast, the subject labels did not improve the performance for *headache* and *runny nose* considerably. This is because the subjects for these symptoms are mostly FIRSTPERSON, as we seldom mention the symptoms of another person in such cases. In other words, the episode classifier can predict a positive label for these symptoms without knowing the subjects of these symptoms.

4 Related Work

4.1 Twitter and NLP

NLP researchers have addressed two major directions for Twitter: adapting existing NLP technologies to noisy texts and extracting useful knowledge from Twitter. The former includes improving the accuracy of part-of-speech tagging (Gimpel et al., 2011) and named entity recognition (Plank et al., 2014), as well as normalizing ill-formed words into canonical forms (Han and Baldwin, 2011; Chrupała, 2014). Even though we did not incorporate the findings of these studies, they could be beneficial to our work in the future.

The latter has led to the development of several interesting applications besides health surveillance. These include prediction of future revenue (Asur and Huberman, 2010) and stock market trends (Si et al., 2013), mining of public opinion (O'Connor et al., 2010), event extraction and summarization (Sakaki et al., 2010; Thelwall et al., 2011; Marchetti-Bowick and Chambers, 2012; Shen et al., 2013; Li et al., 2014a), user profiling (Bergsma et al., 2013; Han et al., 2013; Li et al., 2014b; Zhou et al., 2014), disaster management (Varga et al., 2013), and extraction of common-sense knowledge (Williams and Katz, 2012). Our work can directly contribute to these applications, e.g., sentiment analysis, user profiling, event extraction, and disaster management.

4.2 Semantic analysis for nouns

Our work can be considered as a semantic analysis that identifies an argument (subject) for a disease/symptom-related noun. NomBank (Meyers et al., 2004) provides annotations of noun arguments in a similar manner to PropBank (Palmer et al., 2005), which provides annotations of verbs. In NomBank, nominal predicates and their arguments are identified: for example, ARG0 (typically, subject or agent) is "customer" and ARG1 (typically, objects, patients, themes) is "issue" for the nominal predicate "complaints" in the sentence "There have been no customer complaints about that issue." Gerber and Chai (2010) improved the coverage of NomBank by handling implicit arguments. Some studies have addressed the task of identifying implicit and omitted arguments for nominal predicates in Japanese (Komachi et al., 2007; Sasano et al., 2008).

Our work shares a similar goal with the abovementioned studies, i.e., identifying an implicit ARG0 for a disease and symptom. However, these studies do not regard a disease/symptom as a nominal predicate because they consider verb nominalizations as nominal predicates. In addition,

[7]For the "predicted" setting, first, we predicted the subject labels in a similar manner to five-fold cross validation, and we used the predicted labels as features for the episode classifier.

they use a corpus that consists of newswire text, the writing style and word usage of which differ considerably from those of tweets. For these reasons, we proposed a novel task setting for identifying subjects of diseases and symptoms, and we built an annotated corpus for developing the subject classifier and analyzing the challenges of this task.

5 Conclusion

In this paper, we presented a novel approach to the identification of subjects of various types of diseases and symptoms. First, we constructed an annotated corpus based on an existing corpus for public surveillance. Then, we trained a classifier for predicting the subject of a disease/symptom. The results of our experiments showed that the task of identifying the subjects is independent of the type of disease/symptom. In addition, the results demonstrated the contributions of our work toward identifying an episode of a disease/symptom from a tweet.

In the future, we plan to consider a greater variety of diseases and symptoms in order to develop applications for public health, e.g., monitoring the mental condition of individuals. Thus, we can not only improve the accuracy of subject identification but also enhance the generality of this task.

Acknowledgments

This study was partly supported by Japan Science and Technology Agency (JST). We are grateful to the anonymous referees for their constructive reviews. We are also grateful to Takayuki Sato and Yasunobu Asakura for their annotation efforts. This study was inspired by Project Next NLP[8], a workshop for error analysis on various NLP tasks. We appreciate Takenobu Tokunaga, Satoshi Sekine, and Kentaro Inui for their helpful comments.

References

Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. 2011. Twitter catches the flu: Detecting influenza epidemics using twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1576.

Sitaram Asur and Bernardo A. Huberman. 2010. Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '10, pages 492–499, Washington, DC, USA. IEEE Computer Society.

Shane Bergsma, Mark Dredze, Benjamin Van Durme, Theresa Wilson, and David Yarowsky. 2013. Broadly improving user classification via communication-based name and location clustering on Twitter. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1010–1019.

David Broniatowski, Michael J. Paul, and Mark Dredze. 2013. National and local influenza surveillance through Twitter: An analysis of the 2012-2013 influenza epidemic. *PLoS ONE*, 8(12):e83672.

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, 22(2):249–254.

Herman Anthony Carneiro and Eleftherios Mylonakis. 2009. Google trends: a web-based tool for real-time surveillance of disease outbreaks. *Clinical Infectious Diseases*, 49(10):1557–1564.

Grzegorz Chrupała. 2014. Normalizing tweets with edit scripts and recurrent neural embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 680–686.

Nigel Collier. 2012. Uncovering text mining: a survey of current work on web-based epidemic intelligence. *Global Public Health: An International Journal for Research, Policy and Practice*, 7(7):731–749.

Aron Culotta. 2010. Towards detecting influenza epidemics by analyzing Twitter messages. In *Proceedings of the Workshop on Social Media Analytics (SOMA)*, pages 115–122.

Mark Dredze, Michael J. Paul, Shane Bergsma, and Hieu Tran. 2013. Carmen: A Twitter geolocation system with applications to public health. In *Proceedings of the AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI (HIAI)*, pages 20–24.

Matthew Gerber and Joyce Y. Chai. 2010. Beyond NomBank: A study of implicit arguments for nominal predicates. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1583–1592.

Francesco Gesualdo, Giovanni Stilo, Eleonora Agricola, Michaela V. Gonfiantini, Elisabetta Pandolfi, Paola Velardi, and Alberto E. Tozzi. 2013. Influenza-like illness surveillance on Twitter through automated learning of naïve language. *PLoS One*, 8(12):e82489.

[8] https://sites.google.com/site/projectnextnlp/english-page

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 42–47.

Bo Han and Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 368–378.

Bo Han, Paul Cook, and Timothy Baldwin. 2013. A stacking-based approach to Twitter user geolocation prediction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 7–12.

Yoshiaki Kitagawa, Mamoru Komachi, Eiji Aramaki, Naoaki Okazaki, and Hiroshi Ishikawa. Disease event detection based on deep modality analysis. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP) 2015 Student Research Workshop*.

Mamoru Komachi, Ryu Iida, Kentaro Inui, and Yuji Matsumoto. 2007. Learning based argument structure analysis of event-nouns in Japanese. In *Proceedings of the Conference of the Pacific Association for Computational Linguistics (PACLING)*, pages 120–128.

Alex Lamb, Michael J. Paul, and Mark Dredze. 2013. Separating fact from fear: Tracking flu infections on Twitter. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 789–795.

Vasileios Lampos and Nello Cristianini. 2010. Tracking the flu pandemic by monitoring the social web. In *2nd IAPR Workshop on Cognitive Information Processing (CIP 2010)*, pages 411–416.

David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. The parable of Google flu: Traps in big data analysis. *Science*, 343(6176):1203–1205.

Jiwei Li, Alan Ritter, Claire Cardie, and Eduard Hovy. 2014a. Major life event extraction from Twitter based on congratulations/condolences speech acts. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1997–2007.

Jiwei Li, Alan Ritter, and Eduard Hovy. 2014b. Weakly supervised user profile extraction from Twitter. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 165–174.

Micol Marchetti-Bowick and Nathanael Chambers. 2012. Learning for microblogs with distant supervision: Political forecasting with Twitter. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 603–612.

Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. 2004. The NomBank Project: An interim report. In *Proceedings of the NAACL/HLT Workshop on Frontiers in Corpus Annotation*, pages 24–31.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, , and Noah A. Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 122–129.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Michael J. Paul and Mark Dredze. 2011. You are what you tweet: Analyzing Twitter for public health. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 265–272.

Barbara Plank, Dirk Hovy, Ryan McDonald, and Anders Søgaard. 2014. Adapting taggers to Twitter with not-so-distant supervision. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1783–1792.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World Wide Web (WWW)*, pages 851–860.

Ryohei Sasano, Daisuke Kawahara, and Sadao Kurohashi. 2004. Automatic construction of nominal case frames and its application to indirect anaphora resolution. In *Proceedings of the 20th international conference on Computational Linguistics*, pages 1201–1207.

Ryohei Sasano, Daisuke Kawahara, and Sadao Kurohashi. 2008. A fully-lexicalized probabilistic model for Japanese zero anaphora resolution. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 769–776.

Chao Shen, Fei Liu, Fuliang Weng, and Tao Li. 2013. A participant-based approach for event summarization using Twitter streams. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1152–1162.

Jianfeng Si, Arjun Mukherjee, Bing Liu, Qing Li, Huayi Li, and Xiaotie Deng. 2013. Exploiting topic based Twitter sentiment for stock prediction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 24–29.

Alessio Signorini, Alberto Maria Segre, and Philip M. Polgreen. 2011. 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(5):e19467.

Mark A. Stoové and Alisa E. Pedrana. 2014. Making the most of a brave new world: Opportunities and considerations for using Twitter as a public health monitoring tool. *Preventive Medicine*, 63:109–111.

Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. 2011. Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology*, 62(2):406–418.

István Varga, Motoki Sano, Kentaro Torisawa, Chikara Hashimoto, Kiyonori Ohtake, Takao Kawai, Jong-Hoon Oh, and Stijn De Saeger. 2013. Aid is out there: Looking for help from tweets during a large scale disaster. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1619–1629.

Paola Velardi, Giovanni Stilo, Alberto E. Tozzi, and Francesco Gesualdo. 2014. Twitter mining for fine-grained syndromic surveillance. *Artificial Intelligence in Medicine*, 61(3):153–163.

Jennifer Williams and Graham Katz. 2012. Extracting and modeling durations for habits and events from Twitter. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 223–227.

Deyu Zhou, Liangyu Chen, and Yulan He. 2014. A simple bayesian modelling approach to event extraction from Twitter. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 700–705.