Visualization of the occurrence trend of infectious diseases using Twitter

Ryusei Matsumoto, Minoru Yoshida, Kazuyuki Matsumoto, Hironobu Matsuda, Kenji Kita
Institute of Technology and Science, University of Tokushima
2-1, Minami-josanjima, Tokushima, 770-8506, Japan
c501306048@tokushima-u.ac.jp, mino@is.tokushima-u.ac.jp, matumoto@is.tokushima-u.ac.jp,
c501637026@tokushima-u.ac.jp, kita@is.tokushima-u.ac.jp

Abstract

We propose a system for visualizing the epidemics of infectious diseases. We apply factuality analysis both to disease detection and location estimation for accurate visualization. We tested our methods for several infectious diseases, and show that our method performs well on various diseases.

Keywords: Twitter, disease surveillance, location estimation

1. Introduction

Information related to human life such as illness and infectious diseases is important for many people. However, people have limited kinds of sources of information on diseases such as TV news and newspapers (including Internet news.) On the other hand, SNS including Twitter is a promising source of another type of information about diseases. It has a highly real-time nature and can obtain a large amount of information about the event occurring just now. The final goal of this research is to provide users with real-time infection maps automatically created from Twitter posts. To this end, we propose a system to obtain how many people are infected by the given disease in each area. The proposed system extracts the tweets that contain the disease name, and estimates the location of the user. The system performs factuality analysis of mentions both for infection events and user locations for more accurate estimation. Finally, an infectious disease map is created using the result of the location estimation.

2. Existing Research

Several researches that combine Twitter and illness have been done so far (Charles-Smith et al., 2015). For example, the system by Aramaki et al. (Aramaki et al., 2011; Iso et al., 2016; Kanouchi et al., 2015; Kitagawa et al., 2015) extracts tweets related to influenza posted on Twitter, conducted factuality analysis using SVMs on extracted tweets for detecting epidemic influenza. They also provide the "influ kun" service[1] which extracts tweets related to influenza from Twitter and provides a epidemic map of influenza overlaid on a map of Japan. Other existing research on disease surveillance with location estimation used simple geocoding (Li and Cardie:, 2013), geotagged tweets (Sadilek et al., 2012) and user profiles (Dredze et al., 2013; Paul and Dredze, 2011). Our system is different from this service in that it is aimed at various kinds of diseases other than influenza, and estimate users location even if the user location is not explicitly provided in the form of geotag or user profiles.

There have also been many researches on estimating user locations from tweets (contents) (Chandra et al., 2011; Cheng et al., 2010; Han et al., 2012; Sugiya et al., 2013). Most of the research use the geotagged tweets as training data and estimate the positions of users/tweets based on the word-geotag collocations. The goal of those previous location estimation researches is to estimate the locations vaguely with best-effort approaches. On the other hand, our goal is to provide the positions of infected people as accurately as possible. For that reason, our location estimation algorithm is different from the previous ones: we only focus on the location names found in tweets, and *do the factuality analysis on them* to determine whether the user actually is in the position. To the best of our knowledge, this is the first research to take factuality analysis approach to location estimation for Twitter users.

Some of the researches take time-aware approaches, which estimate the given location name is for the current position, or future position, etc. (Li and Sun, 2014; Suzuki et al., 2015) Our problem setting is similar to them, but different from in that our goal is to estimate where the user live in by excluding not only the non-current-position-indicating tweets but also non-living-position-indicating tweets, which is difficult only by lexical approaches such as '"now" indicates the current position', etc.

Our contribution is two-fold. First, we report how the factuality analysis proposed for influenza generalize to other diseases by testing it on several disease names with their accuracies. Second, we propose a new task of factuality analysis of location indication when the location name itself is clearly provided.

3. Proposed method

We define our problem as the task to perform random search on Twitter with a disease name, and determine the living place of the user by looking at tweets of the user as few as possible. We assume this setting because in the real system it will be needed to estimate user location as lightly as possible.

3.1. Acquisition of tweets related to infectious diseases

In this research, we use TwitterAPI to acquire tweets. We use names or synonyms of seven kinds of infectious diseases that are prevalent in the summer as the search target.

[1] http://mednlp.jp/influ_map/

・As a result of inspection at the hospital, it was Helpangina
・Suspicion of Helpangina
・it's prevalent, Helpangina
・@XXX Is it the Helpangina that is in fashion now?

Figure 1: Examples of Helpangina tweets (translated from Japanese)

As a result of inspection at the hospital, it was Helpangina. 1
It may be Helpangina. -1
I am scared of Helpangina. -1

Figure 2: Examples of training data on infectious diseases

The targeted infections are described in detail in Section 4. As an example, we explain an experimental method of conducting a search using an infection term "Helpangina". Figure 1 shows an example of the acquired tweets. The reason for using synonyms is that many of the infectious diseases targeted in this research were more likely to contain synonyms in tweets than the official names. As an example, for the infectious diseases "epidemic parotitis", another name for this infection is "mumps". Generally, "mumps" is used more often than "epidemic parotitis". Using such synonyms thus increases the number of retrieved tweets.

3.2. Factuality analysis on infectious diseases

Factuality analysis of disease events are performed by using SVMs, which is the same way as the method by Aramaki et al. (Aramaki et al., 2011) The tweet which indicates infections is a positive example, and taken as a negative example otherwise. The training data for each disease consist of 1,000 tweets (500 positive and 500 negative) and the test data consists of 400 tweets for each disease. Figure 2 shows an example of training data. As a meaning of labels, "1" is infected, "-1" is not infected

3.3. Location estimation

We also performed the factuality analysis for location indications. As mentioned above, our problem setting is that we estimate the user's living area given *the tweets as few as possible*. This is because in the real situation it takes much cost to obtain whole list of tweets from one user, which makes it difficult to use user's whole tweets for location estimation. We assume the situation in which we search random tweet collections, collect disease-related tweets, look at the user's timeline and find one tweet with location name, and determine if the location tweet is valid. This makes it possible to construct the system which are more comprehensive for surveillance by being able to refer to more users.

3.3.1. Tweet Acquisition

We constructed the test data for location estimation for other 100 users, extracting his/her representative location tweet by obtaining the latest tweets that contain location name in his/her timeline[2]. We used the location information database for geotagging which consists of words expressing location information such as prefecture name, city name, spot name such as station name. We constructed this database by compling the data from *the Japanese location data download service*[3], which provides a list of city/street names with their location metadata such as its latitude, longitude, and corresponding prefecture name. The database also includes a list of major train station names and famous leisure spots with their corresponding prefecture names, which was manually added by the authors. The estimation is a prefecture-level, which means that we estimate the prefecture the user lives in by matching location names found in the tweets with our database.

We manually exclude the tweet if the tweet is travel-related one by extracting and checking the tweets before and after the target tweet. (Figure 3 shows an example.) [4]

It's a learning trip from tomorrow!
↓
End day 1
↓
End day 2
↓
I forgot to buy a Tokyo banana. let's buy(Detect「Tokyo」)
↓
It seems long and it was a short learning trip!

Figure 3: Examples of tweets representing travel

It seems that I can return to Tokyo earlier than I expected. 1
I will go to Tokyo on my own weekend. -1
I should buy it. Should I get off at Hakata? -1

Figure 4: Example of training data on location information.

3.3.2. Factuality analysis on location estimation

We constructed the training data that consist of 1,000 tweets (500 positive and 500 negative ones.) A positive example means a tweet that can be determined to live in a place of a word representing location information, a negative example means a tweet that can be determined that the user does not live in that location. Figure 4 shows an example of training data. As a meaning of labels, "1" is what the contributor tweets in the location where the per-son lives, "-1" is other things.

[2] The users whose timeline did not include any location names were excluded from the test data.

[3] http://nlftp.mlit.go.jp/isj/

[4] We are currently trying to automate this process. Determination is done by the same way as other classification, i.e., using SVMs. Again, our problem setting is to estimate if the given one tweet is the one during travel or not. We constructed the training data consisting of 200 positive and 200 negative tweets, and 50 test tweets. Positive tweets are the tweets that indicate the user is in travel, and negative tweets are other ones. We obtained 62% accuracy (31/50).

Infectious disease	Accuracy[%]
Helpangina	83.75
Mumps	66.25
Infectious gastroenteritis	79.75
Epidemic conjunctivitis	72.00
Mycoplasma	84.00
Pool heat	68.50
Hand-foot-and-mouth disease	67.75

Table 1: Experimental results of factuality analysis on infectious diseases

4. Experiments

4.1. Factuality analysis for infectious disease events

In this research, we focus on the following seven infectious diseases, for which the results of accuracy evaluation of factuality analysis are shown.

1. Helpangina
2. Mumps
3. Infectious gastroenteritis
4. Epidemic conjunctivitis
5. Mycoplasma
6. Pool heat
7. Hand-foot-and-mouth disease

Because it targets infectious diseases prevailing in the summer, tweets collected from July to August are used as experimental data. Regarding location estimation, it shows the result of position estimation for the newly acquired twitter account. We used word unigrams as features to represent each tweet for SVM training/classification.
Table 1 shows the evaluation results of the accuracy of factuality analysis using SVMs for the seven targeted infectious diseases. The accuracy was 66-84[%].

4.2. Factuality analysis for location estimation

As described in previous sections, 100 user accounts are newly selected for evaluation of location estimation. The acquisition conditions are as follows.

1. Location information is described on the profile
2. Does not include "BOT", "Bot", "bot" in one or more of Twitter client name, account name, display name, and profile

(1) of the above conditions is aimed at easily judging whether or not the result of estimating the position information from the tweet sentence is correct. Also, users who describe location information on their profile tend to have more tweets containing words that describe the area in which they live, than tweets that do not describe location information. (2) is aimed to exclude Bot accounts.
Table 2 shows the evaluation results. The accuracy was 78[%].

	Accuracy[%]
Factuality analysis on location estimation	78.00

Table 2: Experimental results of factuality analysis on location estimation

	Accuracy[%]
Helpangina	81.00
Mumps	79.00
Infectious gastroenteritis	76.00
Epidemic conjunctivitis	62.00
Mycoplasma	78.00
Pool heat	71.00
Hand-foot-and-mouth disease	65.00

Table 3: Experimental results of factuality analysis on location estimation

We also evaluated the accuracy of factuality analysis for locations for users infected with diseases used in Sec. 4.1. Table 3 shows the results. The accuracy was 62-81[%].

5. Error Analysis

5.1. Infectious diseases detection

We observed the variation of accuracy across the diseases. For example, infectious diseases such as "mumps" are fundamentally children's infections. In this research, we treat it as a positive case even if the person related to the user is infected, not the person who posted the tweet sentence. As an example, when there is a tweet sentence "My son has caught a mumps", that tweet sentence is a positive example. Also, There was a tendency that there were a lot of tweets showing that they are prevalent in schools where their children attend, such as "The mumps in my school where my children are attending are popular", "The child of the kindergarten the daughter passes through is suffering from mumps". For this reason, accuracy is considered to be low with respect to infectious diseases prevailing in children such as mumps and hand-foot-and-mouth disease. In order to improve the accuracy more, it is considered effective to increase the learning data or reduce the deviation.

5.2. Location estimation

As a factor of misclassification for location estimation, some tweets contained two or more location words, many of which were misclassified. Although such tweets are not majority of the data, we need a method to choose the appropriate word among several location-indicating words for further improvement of accuracy. In order to improve the accuracy more, it is considered effective to increase training data, improvement of noise removal method, increase of location information database. Currently, our noise removal method uses only special symbols such as musical notes. Removing more special symbols and emoticons will contribute to improving accuracy. Currently the location database contains only nationwide known locations or events, such as "Tokyo Sky Tree" and "Tottori Dune Hill" for the spot name, "Tokyo Game Show" for the event name,

Figure 5: Example of Infectious Disease Map.

so it is thought that accuracy will improve if we add even a little popular locations/events.

6. Visualization of location information

Figure 5 shows the infectious disease map created based on the result of estimating the location of Helpangina infected person. The mark shown on the map shows the following meaning.

Red mark: One infected person,

Blue mark and number: 2 to 10 people infected,

Yellow mark and number: More than 10 people infected.

7. Conclusions

We proposed a method to create an infectious disease map automatically from Twitter. We perform factuality analysis both for disease infection events and location indication for improved accuracy. Experimental results showed that our method perform well on various kinds of diseases. Future work includes the increase of training data and location databases to improve the system accuracy.

8. Bibliographical References

Aramaki, E., Maskawa, S., and Morita, M. (2011). Twitter catches the flu: Detecting influenza ep-idemics using twitter. In *Proceedings of EMNLP 2011*, pages 1568–1576.

Chandra, S., Khan, L., and Muha-ya, F. B. (2011). Estimating twitter user location using social interactions-a content based approach. In *Proceedings of SocialCom/PASSAT 2011*, pages 838–843.

Charles-Smith, L. E., Reynolds, T. L., Cameron, M. A., Conway, M., Lau, E. H., Olsen, J. M., Pavlin, J. A., Shigematsu, M., Streichert, L. C., Suda, K. J., and Corley, C. D. (2015). Using social media for actionable disease surveillance and outbreak management: a systematic literature review. *PloS one*, 10(10).

Cheng, Z., Caverlee, J., and Lee, K. (2010). You are where you tweet: a content-based ap-proach to geo-locating twitter users. In *Proceedings of CIKM 2010*, pages 759–768.

Dredze, M., Paul, M. J., Bergsma, S., and Tran, H. (2013). Carmen: A twitter geolocation system with applications to public health. In *AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI (HIAI)*.

Han, B., Cook, P., and Baldwin, T. (2012). : Geolocation prediction in social media data by finding location indicative words. In *Proceedings of COLING 2012*, pages 1045–1062.

Iso, H., Wakamiya, S., and Aramaki, E. (2016). Forecasting word model: Twitter-based influenza surveillance and prediction. In *Proceedings of COLING 2016*, pages 76–86.

Kanouchi, S., Komachi, M., Oka-zaki, N., Aramaki, E., and Ishikawa, H. (2015). Who caught a cold ? - identifying the subject of a symptom. In *Proceedings of ACL (1) 2015*, pages 1660–1670.

Kitagawa, Y., Komachi, M., Ara-maki, E., Okazaki, N., and Ishikawa, H. (2015). Disease event detection based on deep modality analy-sis. In *Proceedings of ACL (Student Research Workshop) 2015*, pages 28–34.

Li, J. and Cardie:, C. (2013). Early stage influenza detection from twitter. In *CoRR abs/1309.7340*.

Li, C. and Sun, A. (2014). Fine-grained location extraction from tweets with temporal awareness. In *Proceedings of SIGIR 2014*, pages 43–52.

Paul, M. J. and Dredze, M. (2011). You are what you tweet: Analyzing twitter for public health. In *Proceedings of ICWSM 2011*.

Sadilek, A., Kautz, H. A., and Silenzio, V. (2012). Predicting disease transmission from geo-tagged micro-blog data. In *Proceedings of AAAI 2012*.

Sugiya, T., Sirakawa, M., Hara, T., and Nishio, S. (2013). User position estimation method when posting tweets with supervised reinforcement learning. In *Proceedings of IPSJ SIG Technical Reports*.

Suzuki, Y., Kaji, N., Yoshinaga, N., and Toyoda, M. (2015). Geolocation estimation of mi-crobloggers utilizing their recent posts (in japanese). In *Proceedings of the 7th forum on Data Engineering and Information Management (DEIM)*.