# Extracting and Aggregating False Information from Microblogs

**Naoaki Okazaki**[†‡]     **Keita Nabeshima**[†]     **Kento Watanabe**[†]

**Junta Mizuno**[§]     **Kentaro Inui**[†]

{okazaki, nabeshima, kento.w, junta-m, inui}@ecei.tohoku.ac.jp
[†]Graduate School of Information Sciences, Tohoku University
6-3-09 Aramakiaza-Aoba, Aoba-ku, Sendai, 980-8579 Japan
[‡]Precursory Research for Embryonic Science and Technology (PREST), Japan Science and Technology Agency (JST)
[§]National Institute of Information and Communications Technology (NICT)

## Abstract

During the 2011 East Japan Earthquake and Tsunami Disaster, we had found a number of false information spread on Twitter, e.g., "The Cosmo Oil explosion causes toxic rain." This paper extracts pieces of false information exhaustively from all the tweets within one week after the earthquake. Designing a set of linguistic patterns that correct false information, this paper proposes a method for detecting false information. More specifically, the method extracts text passages that match to the correction patterns, clusters the passages into topics of false information, and selects, for each topic, a passage explaining the false information the most suitably. In the experiment, we report the performance of the proposed method on the data set extracted manually from Web sites that are specialized in collecting false information.

## 1 Introduction

In the aftermath of the Tohoku Earthquake (also known as the Great East Japan Earthquake) in March 2011, social media, such as the Twitter social networking and microblogging service, served as highly active and beneficial sources of information. Among Internet users, 18.3% referred to social media as information sources, 18.6% referred to Internet newspapers, and 23.1% referred to national and regional government websites (Nomura Research Institute, 2011). This indicates that social media rivaled the other two in influence. It has also been noted that the Internet and social media has accelerated the dissemination of disinformation and other types of misinformation, e.g., "Toxic rain will follow the explosion at the Cosmo Oil petrochemical complex".

Misinformation such as this regarding safety and danger spread quickly in the aftermath of the Tohoku Earthquake and the related accident at the Fukushima Dai-Ichi Nuclear Power Plant, which threatened the lives and welfare of numerous people. Other themes of the misinformation included the admonition, "Drink Isodine (povidone iodine) to protect your thyroid from radiation". One tweet consolidation site dedicated to collecting/correcting information on the Tohoku Earthquake[1] found that during the month of January 2012, even ten months after the event, more than ten pieces of misinformation related to the earthquake were posted. This indicates a strong need for misinformation alerts in normal times as well as in times of disaster.

In this study, we aim at automatic collection of misinformation disseminated on Twitter. More concretely, we focus on corrective patterns (CPs), such as *It is incorrect that ...*, which are commonly used to correct or refute misinformation, and propose a method incorporating such CPs into a system for automatic collection of misinformation. We then describe the experimental application of this method to tweets posted during the week following the Tohoku Earthquake. The results of this experiment showed that our method could detect approximately half of the 60 misinformative tweets identified by the existing misinformation consolidation sites, as well as 22 other misinformative tweets that had not been recorded on those sites.

## 2 Related work

Twitter has been the subject of a number of studies. Here, we review those relating to the truth or credibility of information posted on Twitter.

Qazvinian et al. (2011) proposed a method for classifying a group of tweets related to misinfor-

---

[1]https://twitter.com/#!/jishin_dema

mation (such as a group of tweets containing the terms "Barack Obama" and "Muslim") into those including explicit expression of misinformation (e.g., "Barack Obama is a Muslim"), and those that do not (e.g., "Barack Obama met with Muslim leaders"), then further classifying the latter into those that support the misinformation and those that oppose it. Unlike the present study, it assumed that mining misinformation from large volumes of tweet and obtaining a group of misinformation-related tweets was outside its scope.

In Japan, numerous studies have been performed on misinformation dissemination via Twitter, prompted by a strong awareness of this problem following the Tohoku Earthquake. For example, Fujikawa et al. (2012) proposed a method for assessing the truth or falsehood of information by classifying user reactions based on the number of specific responses, such as those expressing doubt or presenting well-grounded arguments that the topic is misinformative. Toriumi et al. (2012) proposed a method of investigating the co-occurrence of words and terms such as *disinformation*, *lie*, and *false report* in arguments related to tweet content in order to determine whether the content comprises misinformation.

To analyze the trends in misinformation dissemination and correction on Twitter following the Tohoku Earthquake, Umejima et al. (2011) tested hypotheses such as "the probability of tweet text containing a URL being misinformation is low", "many information tweets contain content that urges action, is negative, or fans unrest", and "a tweet containing any of these three features is apt to be retweeted". In subsequent studies (Umejima et al., 2012; Miyabe et al., 2012), their group showed that words and terms that clearly indicate an intention to correct, such as *disinformation* and *mistaken*, provide a useful feature for recognizing corrective tweets during the construction of a misinformation database. They collected tweets containing such terms and built a binary classifier to assess whether the tweets were correcting specific information.

In all of these studies, each tweet text is taken as a unit with the focus on determining whether it contains misinformation or corrects by providing specific information[2], without precisely identify-

---

[2]For example, the tweet "Be careful! All kinds of misinformation are circulating on Twitter" contains the expression "misinformation" but does not correct any specific information.

ing the region of the misinformation in the tweet text. Therefore, to our knowledge, the present study represents the first investigation of a comprehensive collection of misinformation extracted from large volumes of tweet data.

# 3 Proposed method

In this study, we assume that misinformation disseminated on Twitter is corrected or refuted by other users. For example, we found tweets correcting information (*corrective tweets* hereafter) for the misinformation, "Toxic rain is falling because of the Cosmo Oil explosion."

- It is counterfactual that toxic rain will fall due to the Cosmo Oil explosion.

- Be aware of the false rumors that rain contaminated by the toxic substances produced by the Cosmo Oil explosion will fall.

A corrective tweet consists of a corrective expression (e.g., the underlined parts in the above examples) and misinformation where the corrective expression corrects or refutes. Thus, we can locate misinformation by finding corrective expressions in tweets. The goal of the proposed method presented in this section is to collect phrases of misinformation by using corrective patterns (CPs) and aggregate them into a small number of descriptions of misinformation.

Figure 1 shows the flow of the proposed method, which is essentially comprised of the four steps. In Step 1, the proposed method searches for occurrences of CPs in tweets and extracts their targets of correction (*corrected phrases* hereafter). Step 2 chooses keywords that appear frequently in the corrected phrases. In order to merge keywords referring to the same misinformation, we cluster keywords (Step 3). Finally in Step 4, the proposed method chooses the small number of phrases that describe misinformation the most suitably. We will explain the detail of these steps in the subsequent subsections.

## 3.1 Step 1: Extraction of corrected phrases

Here, we search for tweets with corrected phrases. In corrective tweets, the search determines the presence of the misinformation that is being corrected or refuted via terms such as *misinformation* or *mistaken*, as in the statement, "It is disinformation that Isodine provides protection against radiation", in
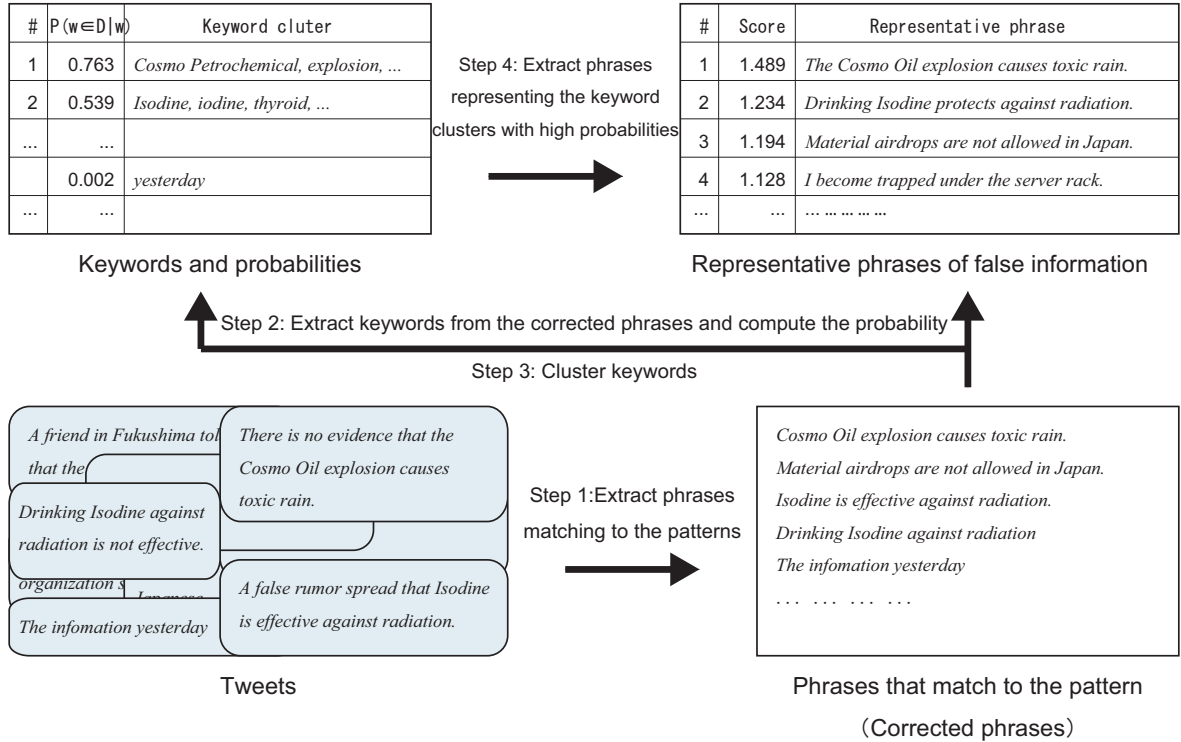
Figure 1: Overview of the proposed method

which the underlined portion is the misinformative phrase being corrected. A Japanese translation of this sentence is, "Isojin ha hibaku wo fusegeru to iu no wa dema da".

In the Japanese sentence that corrects or refutes misinformation, the corrected phrase (e.g., in transliteration, the underlined portion of the sentence "Isojin ha hibaku wo fusegeru", which corresponds to the underlined portion of the above English example) is followed by a functional attributive particle expression, such as the above "to iu no wa" or some other functionally similar term such as "no yo na", and the corrective expression ("wa dema da" in the above example).

We manually formulated 368 CPs to recognize the corrected phrases. We obtained these CPs by examining tweets containing keywords that correspond to 15 kinds of well-known misinformation. If a region of a tweet text matches any of these CPs, the corrected phrase is deemed to comprise the portion of the Japanese sentence ranging from the first word in the sentence to just before the CP. We applied this process to the whole of the tweets under investigation, and the set of corrected phrases extracted in this way is denoted as $D$.

## 3.2 Step 2: Keyword extraction

Some of the corrected phrases extracted in this way simply refer to the misinformation rather than stating it, as in "kino no are" (literally "That thing yesterday") in a sentence such as "kino no are wa dema da" ("That thing yesterday was misinformation"). Such phrases cannot be considered misinformative and must therefore be excluded. This is done by determining whether the words in the corrected phrase co-occur prominently with the CPs. For this purpose, the conditional probability that a word $w$ used in the tweet is among those in the corrected phrase set $D$ is computed as,

$$P(w \in D|w) = \frac{\text{\# tweets where } w \text{ co-occurs with CPs}}{\text{\# tweets containing } w}.$$
(1)

We extract the top-500 words yielding the highest probability as misinformation keywords.

## 3.3 Step 3: Keyword clustering

Misinformative phrases pertaining to the same information may differ considerably in wording and information quantity, as in "Rain containing hazardous substances from the Cosmo Oil fire will occur" and "The Cosmo Oil explosion is toxic", which must be consolidated to avoid redundancy when extracting misinformation. For this rea-

son, we perform clustering of the keywords extracted in Step 2. As the inter-keyword distance (i.e., similarity), we use the cosine similarity on context vectors whose elements correspond to co-occurrence counts between the keywords and the content words (nouns, verbs, and adjectives) in sentences. For the feature value of the context vector, we use the pointwise mutual information (PMI), which provides a measure of the co-occurrence of the keywords and the content words. Performing the complete-link clustering method (furthest neighbor method), we choose the cluster keywords as those yielding high conditional probabilities in Step 2.

### 3.4 Step 4: Representative phrase selection

For each cluster obtained in Step 3, we select representative phrases from among those containing the keywords, and output them as identified misinformation. To select corrected phrases of suitable length that can provide a sufficient description of the misinformation, we compute the score,

$$\text{Score}_p(s,t) = \text{hist}_t(\text{len}_s) \sum_{w \in C_s} \text{PMI}(t,w), \quad (2)$$

where $s$ denotes the corrected phrase, $t$ denotes the representative keyword of the misinformation cluster, $C_s$ indicates the set of content words in $s$, and $\text{len}_s$ is the number of words in $s$. The term $\text{hist}_t(\text{len}_s)$ represents the ratio (relative frequency) of occurrences of sentences consisting of $\text{len}_s$ words with the keyword $t$. $\text{PMI}(t,w)$ represents the pointwise mutual information of the cooccurrence $t$ and $w$.

Equation 2 is designed to yield a high score for corrected phrases that contain numerous content words that co-occur frequently with the keywords and are of a standard length. In essence, $\text{hist}_t(\text{len}_s)$ is a compensatory term that yields a high score for phrases of typical length among those containing the keyword $t$. For each cluster obtained in Step 3, we choose the phrase $\hat{s}$ yielding the highest score as the representative description for the keyword $t$.

## 4 Experiment 1 — CP evaluation

For misinformation acquisition by the proposed method, it is essential to identify CPs that can effectively represent the misinformation. Our first experiment was to evaluate the performance of our CPs.

Table 1: Precision and recall of correction patterns

| Precision | Recall |
|---|---|
| 0.79 (118/150) | 0.83 (50/60) |

### 4.1 Experimental setting

The corpus that was used as the source of information for the misinformation extraction evaluation comprised 179,286,297 tweets posted between 9:00 JST on March 11 and 9:00 on March 18, 2011, which were provided by Twitter Japan at the Great East Japan Earthquake Big Data Workshop[3]. To create a reference data set, we collected all the instances of misinformation from four misinformation consolidation websites[4] and chose from them 60 instances of misinformation that were determined to have been posted during the week following the Tohoku Earthquake. During the CP performance evaluation, these 60 misinformation instances were compared with approximately 20,000 corrected phrases that were automatically extracted by our CPs. In this evaluation, these 60 instances were denoted as "valid (or gold) instances".

The CPs were evaluated for precision and recall. For precision evaluation, we took 150 samples selected at random from the approximately 20,000 instances of corrected phrases. Precision was defined as the proportion of those samples that were recognized by the CPs as instances of information correction or refutation made by their posters. Recall was defined as the proportion of the 60 valid instances that were recognized from the set of approximately 20,000 instances of corrected phrases.

### 4.2 Results and analysis

As shown by the values found for the CP precision and recall values in Table 1, the precision and recall values of the misinformation extraction were both approximately 80%. The corrected phrases that were extracted were found to be of four types, as shown in Table 2.

Those in types (a) and (b) were identified as

---

Table 2: Types of corrected phrases extracted

| Phrase type | # |
|---|---|
| (a) Having sufficient content for recognition as phrases with corrected information | 76 |
| (b) Lacking sufficient content for recognition as phrases with corrected information | 42 |
| (c) Phrases erroneously extracted that represent instances of ambiguous patterns | 24 |
| (d) Phrases erroneously extracted that represent instances of unclear author intent | 8 |
| Total | 150 |

Table 3: Causes of failure to extract misinformation

| Cause | # |
|---|---|
| (e) New correction pattern | 3 |
| (f) Evidence present in corrective tweet | 4 |
| (g) No corrective tweet | 3 |
| Total | 10 |

Table 4: Accuracy and recall of extracted misinformation

| N | Acc (4-sites) | Acc (manual) | Recall |
|---|---|---|---|
| 25 | 0.44(11/25) | 0.64(16/25) | 0.18(11/60) |
| 50 | 0.34(17/50) | 0.58(29/50) | 0.28(17/60) |
| 75 | 0.33(25/75) | 0.56(42/75) | 0.42(25/60) |
| 100 | 0.30(30/100) | 0.52(52/100) | 0.50(30/60) |

valid in the evaluation that yielded the results shown in Table 1. Type (b) is of special interest because it comprises phrasing instances in which the misinformation is either not explicitly stated (e.g. " (That thing yesterday)" in "

(That thing yesterday was a piece of disinformation)", where the CP underlined) or insufficiently expressed (e.g. "that Isodine affair" in "I heard that Isodine affair was a case of disinformation"). Presumably, corrected phrases of type (b) can be eliminated by the conditional probability ranking and representative phrase selection performed by Steps 2 and 4, respectively.

Those of types (c) and (d) were both mistakenly extracted in the evaluation. Type (c) comprises instances of phrasing in which the corrected phrase was extracted by erroneous CP application (e.g., " (In times of disaster such as this)" in "

(In times of disaster such as this, disinformation flows freely)"). Type (d) comprises phrases in which the attitude of the writer toward the CP (in regards to the correction) is ambiguous or vague (e.g., "

(Fundraising will make you popular - spreading that rumor will have an effect)").

We also examined the 10 instances of failure to extract misinformation and, as shown in Table 6, found the following three types.

Type (e) involved corrective phrasing that was not covered by the existing CPs, such as the underlined portion of the statement, " 24

(No information source is given to show that the Emperor actually performed 24 hours of prayer)". Extraction of this type will be possible with the addition of new CPs. Type (f) comprises types of misinformation correction or refutation that are outside the scope of the CP forms considered in this study. One instance of this is the following corrective tweet, which opposes the disinformation stating that, "

(South Korea requests loan from Japan. And (P.M.) Kan readily agrees.)"

"

RT @xxx RT

!

(This looks like a fabrication. No source given. RT@xxx. RT. In the present state of emergency, South Korea asks Japan for a loan. And Kan readily agrees!)".

Several tweets intended to correct misinformation were found to take the form of commentary on an original tweet, as in this example.

Type (g) comprises several instances in which tweets purveying misinformation were present among the tweet collection used in this study, but related corrective tweets were not. Extraction of such misinformative tweets by the proposed method would be difficult at best, as our method assumes the occurrence of correction tweets, but such instances were small in number.

## 5 Experiment 2 — Evaluation of misinformation consolidation

Table 5: Types of errors that lowered accuracy

| Error type | # | % |
|---|---|---|
| (a) Errors in topic extraction | 12 | 25.0 |
| (b) Errors in clustering 1 | 20 | 41.7 |
| (c) Information of uncertain content | 5 | 10.4 |
| (d) Extraction of correct information | 1 | 2.1 |
| (e) Prediction of future events | 5 | 10.4 |
| (f) Validity unclear | 5 | 10.4 |
| Total | 48 | 100.0 |

Table 6: Types of errors that lowered recall

| Error type | # | % |
|---|---|---|
| (g) Errors in clustering 2 | 2 | 10.0 |
| (h) Low ranking | 18 | 90.0 |
| Total | 20 | 100.0 |

We next evaluated Steps 2 to 4 of Section 3. This evaluation essentially consisted of determining whether these two steps, when applied to the corrected phrases extracted in Section 4, effectively excluded the type (b) corrected phrases from the extracted phrase set (lacking a statement of the specific information) and whether the selected representative phrases contained appropriate descriptions of misinformation.

## 5.1 Experimental setting

We assessed the misinformation extracted by the proposed method by manually examining each instance to determine whether it was equivalent in content to any of the 60 gold instances from the four consolidation websites. For some of the misinformation extracted by the proposed method, no similar instances were found in the gold set. In those cases, we manually investigated the information with Web search engines to determine whether it actually was a case of misinformation. Additionally, as the objective in the present study is a comprehensive extraction of misinformation, in cases where the content two or more instances of extracted misinformation were deemed to be essentially the same, we counted them as one correct instance of extraction. Ultimately, the accuracy and recall in this investigation were determined using various values of N, which is the predetermined number of information instances output in order of decreasing score, in the proposed method.

## 5.2 Experimental results and analysis

Table 4 shows the results of the evaluation. With N as 100, approximately 30% of the information

instances extracted by the proposed method were found to be present in the gold set. In addition, approximately 20% of the extracted instances were found to be actual instances of information, and thus correct, even though they were not present in the gold set. Therefore, it can be said that the proposed method extracted misinformation with a precision of approximately 50%. Among the incorrect answers, approximately half involved redundant expressions of essentially the same misinformation phrased differently. In summary, approximately 70% of the misinformation extracted by the proposed method represented a correct answer.

Investigation of the causes of the inaccuracy in output represented by the 48 incorrect answers present among the top 100 extracted misinformation instances showed that they could be classified into six types. These are listed in Table 5, together with the number of incorrect answers attributable to each type. Types (a) to (d) involve instances that were easily judged as errors, but types (e) and (f) involve instances that would be difficult for humans to characterize as either true information or misinformation. The six cause types, and potential means of avoiding them, are as follows:

(a) **Errors in keyword extraction** In some instances, unsuitable keywords such as " (watchamacallit)", " (mess)", and " " (a symbol used to mean "a certain", as in "a certain person") were extracted as misinformation keywords. It may be possible to eliminate this source of error in Step 2 by excluding extraction terms that are written entirely in hiragana (the Japanese cursive syllabary) and/or terms composed in large degree of symbols, such as " " above.

(b) **Errors in clustering** Among the top 100 instances of information extraction, some involved redundancies in the form of different phrases that have essentially the same content, as in the following examples, in which the terms in parentheses were theme terms used in the selection process.

LPG

(Due to the explosion of the Cosmo

Oil Chiba Refinery LPG tank in Ichihara City, residents of Chiba Prefecture and its neighboring regions will be subjected to toxic rain. (Cosmo Oil Chiba Refinery)

Due to the Chiba Prefecture petrochemical complex explosion, the substances adversely affecting human health will mix in the air and fall as acidic rain. (petrochemical complex explosion)

Because these two instances of misinformation were not assigned to the same cluster in Step 3, they gave rise to apparent redundancy. While the current method takes words that co-occur in corrected phrases as their features, it may be possible to reduce this type of redundancy by adding surface information of the keywords themselves to the feature set.

**(c) Information of uncertain content** This involves instances in which the selected representative phrase states the misinformation inadequately, as in the following example:

Death by starvation and freezing has occurred.

The gold set included the sentence "  
(In Iwaki City, death by starvation and freezing have occurred)", but the above representative statement is less specific and was therefore considered to be uncertain in content. Tweets containing such phrases were small in number, and may therefore be excluded by setting a threshold number for this purpose.

**(d) Erroneous extraction of true information** The following was extracted as misinformation, but when checked against reality was found to be true:

The tip of Tokyo Tower has been bent.

When people saw this, many also considered this to be misinformation, as its content

seems wildly implausible. However, in the present evaluation, it was the only instance of this type detected among 100 instances of extracted information, and is therefore not considered to be a substantial problem.

**(e) Prediction of future events** In some instances, expressions comprising the prediction of a future event were extracted, such as the following:

A nuclear explosion will occur in Fukushima.

**(f) Unclear validity** We found some instances in which a search of several websites yielded no indication of whether they involved misinformation, as in the following example:

Suntory opens vending machines to dispense products free of charge.

Among the 60 gold instances of misinformation, 20 were included in the corrected phrase set but were not extracted as misinformative. Our investigation into the causes showed that they were of the following two types, which are listed in Table 6 together with the number that occurred in each type.

**(g) Errors in clustering** In some instances, the candidates were extracted by the CPs but were mistakenly merged with other misinformation instances during the clustering process. However, they apparently do not pose a substantial problem because their number was small in comparison with the total quantity of extracted misinformation.

**(h) Unduly low ranking** In some instances, candidates were extracted by the CPs but were not extracted as keywords because of their low conditional probability. One example of this is in the misinformation, "  
(A man pretending to be from Tokyo Electric Power appeared on the scene)". The keyword "Tokyo Electric Power" frequently occurs in statements that do not involve misinformation, and its conditional probability of exhibiting misinformation was therefore estimated to be low. Accordingly, a means of scoring for corrected

phrases themselves, rather than for independent keywords, is necessary to eliminate this problem.

## 6 Conclusion

In this study, we focused on expressions that correct or refute misinformation, and proposed a method for automatic collection of misinformation. The method was evaluated in an experiment during which entries extracted from misinformation consolidation websites that had been manually classified as misinformation were taken as gold instances and used as a basis for comparison with information extracted by the proposed method as misinformation. Some of this extracted misinformation was not listed as misinformation in the consolidation websites, which, together with the other results, showed that the proposed method could be useful for automatic collection of misinformation or, at least, for helping people create and update a comprehensive list of misinformation.

In our future studies, we intend to work to expand the set of CPs used in this method and to improve the corrected phrase scoring, thereby enhancing its performance in misinformation extraction, together with the development of a complete system for real-time misinformation acquisition.

## References

Tomohide Fujikawa, Nobuhiro Kaji, Naoki Yoshinaga, and Masaru Kitsuregawa. 2012. Classsification of users' attitudes toward rumors on microblogs. In *Technical Report of the Institute of Electronics, Information and Communication Engineers*.

Mai Miyabe, Ayana Umejima, Akiyo Nadamoto, and Eiji Aramaki. 2012. Ryugenjoho cloud: Collecting false rumors by extracting correction information from humans. In *Proceedings of the 18th Annual Meeting of the Association for Natural Language Processing*, pages 891–894.

Ltd. Nomura Research Institute. 2011. Survey on "trends in people's use and views of media in the wake of the tohoku - pacific ocean earthquake". `http://www.nri.co.jp/english/news/2011/110329.html`.

Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. 2011. Rumor has it: identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1589–1599, Stroudsburg, PA, USA. Association for Computational Linguistics.

Fujio Toriumi, Kosuke Shinoda, and Genta Kaneyama. 2012. Evaluating a system that judges false rumors in social media. *Journal of Digital Practice*, 3(3):201–208.

Ayana Umejima, Mai Miyabe, Akiyo Nadamoto, and Eiji Aramaki. 2011. Tendency of rumor and correction re-tweet on the twitter during disasters. In *IPSJ SIG Technical Report*, volume 2011, pages 1–6.

Ayana Umejima, Mai Miyabe, Akiyo Nadamoto, and Eiji Aramaki. 2012. Analysis for extracting rumor markers in microblogs. In *Proceedings of DEIM Forum 2012*, pages F3–2.