Aid is Out There: Looking for Help from Tweets during a Large Scale Disaster

István Varga[†] **Motoki Sano**[†] **Kentaro Torisawa**[†] **Chikara Hashimoto**[†]
Kiyonori Ohtake[†] **Takao Kawai**[§] **Jong-Hoon Oh**[†] **Stijn De Saeger**[†]
[†]Information Analysis Laboratory,
National Institute of Information and Communications Technology (NICT), Japan
{istvan, msano, torisawa, ch, kiyonori.ohtake, rovellia, stijn}@nict.go.jp
[§]Knowledge Discovery Research Laboratories, NEC Corporation, Japan
t-kawai@bx.jp.nec.com

Abstract

The 2011 Great East Japan Earthquake caused a wide range of problems, and as countermeasures, many aid activities were carried out. Many of these problems and aid activities were reported via Twitter. However, most problem reports and corresponding aid messages were not successfully exchanged between victims and local governments or humanitarian organizations, overwhelmed by the vast amount of information. As a result, victims could not receive necessary aid and humanitarian organizations wasted resources on redundant efforts. In this paper, we propose a method for discovering matches between problem reports and aid messages. Our system contributes to problem-solving in a large scale disaster situation by facilitating communication between victims and humanitarian organizations.

1 Introduction

The 2011 Great East Japan Earthquake in March 11, 2011 killed 15,883 people and destroyed over 260,000 households (National Police Agency of Japan, 2013). Accustomed way of living suddenly became unmanageable and people found themselves in extreme conditions for months.

Just after the disaster, many people used Twitter for posting problem reports and aid messages as it functioned while most communication channels suffered disruptions (Winn, 2011; Acar and Muraki, 2011; Sano et al., 2012). Examples of such problem reports and aid messages, translated from Japanese tweets, are given below (P1, A1).

P1 *My friend said* **infant formula is sold out**. *If somebody knows shops in Sendai-city where they still have it in stock, please let us know.*

A1 *At Jusco supermarket in Sendai, you can still* **buy** *water and* **infant formula**.

If A1 would have been forwarded to the sender of P1, it could have helped since it would help the "friend" to obtain infant formula. But in reality, the majority of such reports/messages, especially unforeseen ones went unnoticed amongst the mass of information (Ohtake et al., 2013). In addition, there were cases where many humanitarian organizations responded to the same problems and wasted precious resources. For instance, many volunteers responded to problems which were heavily reported by public media, leading to oversupply (Saijo, 2012). Such waste of resources could have been avoided if the organizations would have successfully shared the aid messages for the same problems.

Such observations motivated this work. We developed methods for recognizing problem reports and aid messages in tweets and finding proper matches between them. By browsing the discovered matches, victims can be assisted to overcome their problems, and humanitarian organizations can avoid redundant relief efforts. We define problem reports, aid messages and their successful matches as follows.

Problem report: A tweet that informs about the possibility or emergence of a problem that requires a treatment or countermeasure.

Aid message: A tweet that (1) informs about situations or actions that can be a remedy or solution for a problem, or (2) informs that the problem is solved or is about to be solved.

Problem-aid tweet match: A tweet pair is a problem-aid tweet match (1) if the aid message informs how to overcome the problem, (2) if the aid message informs about the set-

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 1619–1629,
Sofia, Bulgaria, August 4-9 2013. ©2013 Association for Computational Linguistics

tlement of the problem, or (3) if the aid message provides information which contributes to the settlement of the problem.

In this work we excluded *direct requests*, such as “Send us food!”, from problem reports. This is because it is relatively easy to recognize such direct requests by checking mood types (i.e., imperative) and their behavior is quite different from problem reports like “People in Sendai are starving”. Problem reports in this work do not directly state which actions are required, only implying the necessity of a countermeasure through claiming the existence of problems.

An underlying assumption of our method is that we can find a noun-predicate dependency relation that works as an *indicator* of problems and aids in problem reports and aid messages, which we refer to as *problem nucleus* and *aid nucleus*.[1] An example of problem nucleus is “infant formula is sold out” in P1, and that of aid nucleus is “(can) buy infant formula” in A1. Many problem-aid tweet matches can be recognized through problem and aid nuclei pairs.

We also assume that if the problem and aid nuclei match, they share the same noun. Then, the semantics of predicates in the nuclei is the main factor that decides whether the nuclei constitute a match. We introduce a semantic classification of predicates according to the framework of excitation polarities proposed in Hashimoto et al. (2012). Our hypothesis is that excitation polarities along with trouble expressions can characterize problem reports, aid messages and their matches. We developed a supervised method encoding such information into its features.

An evident alternative to this approach is to use sentiment analysis (Mandel et al., 2012; Tsagkalidou et al., 2011) assuming that problem reports should include something ‘bad’ while aid messages describe something ‘good’. However, we will show that this does not work well in our experiments. We think this is due to mismatch between the concepts of problem/aid and sentiment polarity. Note that previous work on ‘demand’ recognition also found similar tendencies (Kanayama and Nasukawa, 2008).

Another issue in this task is, of course, the context surrounding problem/aid nuclei. The following (imaginary) tweets exemplify the problems caused by contexts.

FP1 *I do not believe* **infant formula is sold out** *in Sendai.*

FA1 *At Jusco supermarket in Iwaki, you can still* **buy infant formula**.

The problem nuclei of FP1 and P1 are the same but FP1 is not a problem report because of the expression “I do not believe”. The aid nuclei of FA1 and A1 are the same but FA1 does not constitute a proper match with P1 because FA1 and P1 refer to different cities, “Iwaki” and “Sendai”. In this work, the problems concerning the modality and other semantic modifications to problem/aid nuclei by context are dealt with by the introduction of features representing the text surrounding the nuclei in machine learning. As for the location problem, we apply a location recognizer to all tweets and restrict the matching candidates to the tweet pairs referring to the same location.

2 Approach

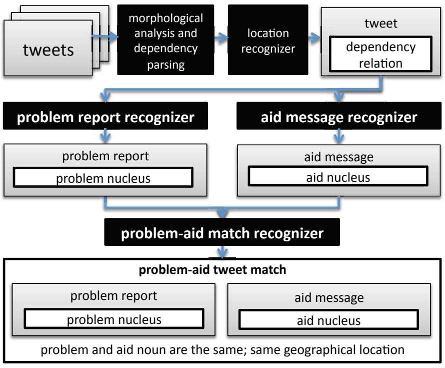

Figure 1: Problem-aid matching system overview.

We developed machine learning based systems to recognize problem reports, aid messages and problem-aid tweet matches. Figure 1 illustrates the whole system. First, location names in tweets are identified by matching tweets against our location dictionary, described in Section 3. Then, each tweet is paired with *each* dependency relation in the tweet, which is a candidate of problem/aid nuclei and given to the problem report and aid message recognizers. A tweet-nucleus-candidate pair judged as problem report is combined with another tweet-nucleus-candidate pair recognized as an aid message if the two nuclei share the same noun and the tweets share the same location name, and given to the problem-aid match recognizer.

[1] We found that out of 500 random tweets only 4.5% of problem reports and 9.1% of aid messages did not contain any problem report/aid message nuclei.

In the following, problem and aid nuclei are denoted by a noun-template pair. A *template* is composed of a predicate and its argument position. For instance, "water supply stopped" in P2 is a problem nucleus, "water supply recovered" in A2 is an aid nucleus and they are denoted by the noun-template pairs ⟨water supply, X stopped⟩ and ⟨water supply, X recovered⟩.

P2 *In Sendai city, water supply stopped.*

A2 *In Sendai city, water supply recovered.*

Roughly speaking, we regard the tasks of problem report recognition and aid message recognition as the tasks of finding proper problem/aid nuclei in tweets and our method performs these tasks based on the semantic properties of nouns and templates in problem/aid nucleus candidates and their surrounding contexts.

The basic intuition behind this approach can be explained using excitation polarity proposed in Hashimoto et al. (2012). Excitation polarity differentiates *templates* into 'excitatory' or 'inhibitory' with regard to the main function or effect of entities referred to by their argument noun. While excitatory templates (e.g., cause X, buy X, suffer from X) entail that the main function or effect is activated or enhanced, inhibitory templates (e.g., ruin X, prevent X, X runs out) entail that the main function or effect is deactivated or suppressed. The templates that do not fit into the above categorization are classified as 'neutral'.

We observed that problem reports in general included either of (A) a dependency relation between a noun referring to some trouble and an excitatory template or (B) a dependency relation between a noun not referring to any trouble and an inhibitory template. Examples of (A) include ⟨carbon monoxide poisoning, suffer from X⟩, ⟨false rumor, spread X⟩. They refer to events that activate troubles. On the other hand, (B) is exemplified by ⟨school, X is collapsed⟩, ⟨battery, X runs out⟩, which imply that some non-trouble objects such as resources, appliances and facilities are dysfunctional. We assume that if we can find such dependency relations in tweets, the tweets are likely to be problem reports.

Contrary, a tweet is more likely to be an aid message when it includes either (C) a dependency relation between a noun referring to some trouble and an inhibitory template or (D) a dependency relation between a noun not referring to any trouble and an excitatory template. Examples of (C) are ⟨flu, X was eradicated (in some shelter)⟩ and ⟨debris, remove X⟩. They represent the *dysfunction* of troubles and can mean the solution or the settlement of troubles. On the other hand, examples of (D) include ⟨school, X re-build⟩ and ⟨baby formula, buy X⟩. They entail that some resources function properly or become available. These formulations are summarized in Table 1.

	trouble	**non-trouble**
excitatory	(A) problem nucleus	(D) aid nucleus
inhibitory	(C) aid nucleus	(B) problem nucleus

Table 1: Problem/aid-excitation matrix.

As an interesting consequence of such a view on problem/aid nucleus, we can say the following regarding problem-aid tweet matchings: when a problem nucleus and an aid nucleus are an adequate match, the excitation polarities of their templates are opposite. Consider the following tweets.

P3 *Some people were going back to Iwaki, but the water system has not come back yet. It's terrible that* **bath is unusable***.*

A3 *We* **open the bath** *for the public, located on the 2F of Iwaki Kuhon temple. If you're staying at a relief shelter and would like to take a bath, you can use it.*

"Bath is unusable" in P3 is a problem nucleus while "open the bath" in A3 is an aid nucleus. Since the problem reported in P3 can be solved with A3, they are a successful match. The inhibitory template "X is unusable" indicates that the function of "bath", a non-trouble expression, is suppressed. The excitatory template "open X" indicates that the function of "bath" is activated.

The same holds when we consider the noun referring to troubles like "flu". The polarity of the template in a problem nucleus should be excitatory like "flu is raging" while that of an aid nucleus should be inhibitory like ⟨flu, X was eradicated⟩. These examples keep the constraint that the problem and aid nucleus should have opposite polarities when they constitute a match.

Note that the formulations of problem report, aid message and their matches or the excitation matrix (Table 1) were not presented to our annotators and our test/training data may contain data that contradict with the formulations. These formulations constitute the hypothesis to be validated in this work.

An important point to be stressed here is that there are problem-aid tweet matches that do not fit into our formulations. For instance, we assume that the problem nucleus and aid nucleus in a proper match share the same noun. However, tweet pairs such as “There are many injured people in Sendai city” and “We are sending ambulances to Sendai” can constitute a proper match, but there is no proper problem-aid nuclei pair that share the same noun in these tweets. (We can find the dependency relations sharing “Sendai” but they do not express anything about the contents of problem and aid.) The point is that the tweet pairs can be judged because people know ambulances can be a *countermeasure* to injured people as world knowledge. Introducing such world knowledge is beyond the scope of this current study.

Also, we exclude direct requests from problem reports. As mentioned in the introduction, identifying direct requests is relatively easy, hence we excluded them from our target.

3 Problem Report and Aid Message Recognizers

We recognize problem reports and aid messages in given tweets using a supervised classifier, SVMs with linear kernel, which worked best in our preliminary experiments. The feature set given to the SVMs are summarized in the top part of Table 2. Note that we used a common feature set for both the problem report recognizer and aid message recognizer and that it is categorized into several types: features concerning trouble expressions (TR), excitation polarity (EX), their combination (TREX1) and word sentiment polarity (WSP), features expressing morphological and syntactic structures of nuclei and their context surrounding problem/aid nuclei (MSA), features concerning semantic word classes (SWC) appearing in nuclei and their context, request phrases, such as “Please help us”, appearing in tweets (REQ), and geographical locations in tweets recognized by our location recognizer (GL). MSA is used to express the modality of nuclei and other contextual information surrounding nuclei. REQ was introduced based on our observation that if there are some requests in tweets, problem nuclei tend to appear as justification for the requests.

We also attempted to represent nucleus template IDs, noun IDs and their combinations directly in our feature set to capture typical templates frequently appearing in problem and aid nuclei, but since there was no improvement we omit them.

TR	Whether the nucleus noun is a trouble/non-trouble expression.
EX1	The excitation polarity and the value of the excitation score of the nucleus template.
TREX1	All possible combinations of trouble/non-trouble of TR and excitation polarities of EX1.
WSP1	Whether the nucleus noun is positive/negative/not in the Word Sentiment Polarity (WSP) dictionary.
WSP2	Whether the nucleus template is positive/negative/not in the WSP dictionary.
WSP3	Whether the nucleus template is followed by a positive/negative word within the tweet.
MSA1	Morpheme n-grams, syntactic dependency n-grams in the tweet and morpheme n-grams before and after the nucleus template. $(1 \leq n \leq 3)$
MSA2	Character n-grams of the nucleus template to capture conjugation and modality variations. $(1 \leq n \leq 3)$
MSA3	Morpheme and part-of-speech n-grams within the *bunsetsu* containing the nucleus template to capture conjugation and modality variations. $(1 \leq n \leq 3)$ (A bunsetsu is a syntactic constituent composed of a content word and several function words, the smallest unit of syntactic analysis in Japanese.)
MSA4	The part-of-speech of the nucleus template’s head to capture modality variations outside the nucleus template’s bunsetsu.
MSA5	The number of bunsetsu between the nucleus noun and the nucleus template. We found that a long distance between the noun and the template suggests parsing errors.
MSA6	Re-occurrence of the nucleus noun’s postpositional particle between the nucleus noun and the nucleus template. We found that the re-occurrence of the same postpositional particle within a clause suggests parsing errors.
SWC1	The semantic class n-grams in the tweet.
SWC2	The semantic class(es) of the nucleus noun.
REQ	Presence of a request phrase in the tweet, identified from within 426 manually collected request phrases.
GL	Geographical locations in the tweet identified using our location recognizer. Existence/non-existence of locations in tweets are also encoded.

EX2	Whether the problem and aid nucleus templates have the same or opposite excitation polarities.
EX3	Product of the values of the excitation scores for the problem and the aid nucleus template.
TREX2	All possible combinations of trouble/non-trouble of TR, excitation polarity EX1 of the problem nucleus template and excitation polarity EX1 of the aid nucleus template.
SIM1	Common semantic word classes of the problem report and aid message.
SIM2	Whether there are common nouns modifying the common nucleus noun or not in the problem report and aid message.
SIM3	Whether the words in the same word class modify the common nucleus noun or not in the problem report and aid message.
SIM4	The semantic similarity score between the problem nucleus template and the aid nucleus template.
CTP	Whether the problem nucleus template and the aid nucleus template are in contradiction relation dictionary or not.
SSR1	Problem report recognizer’s SVM score of problem nucleus template.
SSR2	Problem report recognizer’s SVM score of aid nucleus template.
SSR3	Aid message recognizer’s SVM score of the problem nucleus template.
SSR4	Aid message recognizer’s SVM score of the aid nucleus template.

Table 2: Features used with the problem report recognizer and the aid message recognizer (*above*); additional features used in training the problem-aid match recognizer (*below*).

The other feature types need some non-trivial dictionaries. In the following, we explain how we created the dictionaries for each feature type along with the motivation behind their introduction.

Trouble Expressions (TR) As mentioned previously, trouble expressions work as good evidence for recognizing problem reports and aid messages. The TR feature indicates whether the noun in the problem/aid nucleus candidate is a trouble expression or not. For this purpose, we created a list of trouble expressions following the semi-supervised procedure presented in De Saeger et al. (2008). After manual validation of the list, we obtained 20,249 expressions referring to some troubles, such as "tsunami" and "flu". The value of the TR feature is determined by checking whether the nucleus noun is contained in the list.

Excitation Polarities (EX) The excitation polarities are also important in recognizing problem reports and aid messages as mentioned before. For constructing the dictionary for excitation polarities of templates, we applied the bootstrapping procedure in Hashimoto et al. (2012) to 600 million Web pages. Hashimoto's method provides the value of the *excitation score* in $[-1, 1]$ for each template indicating the polarities and their *strength*. Positive value indicates excitatory, negative value inhibitory and small absolute value neutral. After manual checking of the results by the majority vote of three human annotators (other than the authors), we limited the templates to the ones that have score values consistent with the majority vote of the annotators, obtaining a dictionary consisting of 7,848 excitatory, 836 inhibitory and 7,230 neutral templates. The Fleiss' (1971) kappa-score was 0.48 (moderate agreement). We used the excitation score values as feature values. Excitation has already been used in many works, such as causality and contradiction extraction (Hashimoto et al., 2012) or Why-QA (Oh et al., 2013).

Word Sentiment Polarity (WSP) As we suggested before, full-fledged sentiment analysis to recognize the expressions, including clauses and phrases, that refer to something good or bad was not effective in our task. However, the sentiment polarity, assigned to single words turned out to be effective. To identify the sentiment polarity of words, we employed the word sentiment polarity dictionary used with a sentiment analysis tool for Japanese, the Opinion Extraction Tool software[2], which is an implementation of Nakagawa et al. (2010). The dictionary includes 9,030 positive and 27,951 negative words. Note that we used the Opinion Extraction Tool in the experiments to check the effectiveness of the full-fledged sentiment analysis in this task.

Semantic Word Class (SWC) We assume that nouns in the same semantic class behave similarly in crisis situations. For example, if "infection" appears in a problem report, the tweets including "pulmonary embolism" are also likely to be problem reports. Semantic word class features are used to capture such tendencies. We applied an EM-style word clustering algorithm in Kazama and Torisawa (2008) to 600 million Web pages and clustered 1 million nouns into 500 classes. This algorithm has been used in many works, such as relation extraction (De Saeger et al., 2011) and Why-QA (Oh et al., 2012), and can generate various kinds of semantically clean word classes, such as *foods*, *disease names*, and *natural disasters*. We used the word classes in tweets as features.[3]

Geographical Locations (GL) Our location recognizer matches tweets against our location dictionary. Location names and their existence/non-existence in tweets constitute evidence, thus we encoded such information into our features. The location dictionary was created from the Japan Post code data[4] and Wikipedia, containing 2.7 million location names including cities, schools and other facilities (Kazama et al., 2013).

4 Problem-Aid Match Recognizer

After problem report and aid message recognition, the positive outputs of the respective classifiers are used as input in this step. The problem-aid match recognizer classifies an aid message-nucleus pair together with the problem report-nucleus pair employing SVMs with linear kernel, which performed best in this task again. The problem-aid match recognizer uses all the features used in the problem report recognizer and the aid message recognizer along with additional features regarding: excitation polarity (EX) and trouble expressions (TR), distributional similarity (SIM), contradiction (CTP) and SVM-scores of the problem report and aid message recognizers (SSR). Here also we attempted to capture typical or frequent matches of nuclei using template and noun IDs and their combinations, but we did not observe any improvement so we omit them from the feature set. The bottom part of Table 2 summarizes the additional feature set, some of which are described below in more detail.

[2] Provided at the ALAGIN Forum (http://www.alagin.jp/).

[3] There is a slight complication here. For each noun n, EM clustering estimates a probability distribution $P(n|c^*)$ for n and semantic class c^*. From this distribution we obtained discrete semantic word classes by assigning each noun n to semantic class $c = argmax_{c^*}\ p(c^*|n)$.

[4] http://www.post.japanpost.jp/zipcode/download.html

As for TR and EX, our intuition is that if a problem nucleus and an aid nucleus are an adequate match, their excitation polarities are opposite, as described in Section 2. We encode whether the excitation polarities of nuclei templates are the same or not in our features. Also, the excitation polarities of problem and aid nuclei and TR are combined (TREX1, TREX2) so that the classifier can know whether the nuclei follow the constraint for adequate matches described in Section 2.

As for SIM, if an aid message matches a problem report, besides the common nucleus noun, it is reasonable to assume that certain contexts are semantically similar. We capture this characteristic in three ways. SIM1 looks for common semantic word classes in the problem report and aid message. SIM2 and SIM3 target the modifiers of the common nucleus noun if they exist.

We also observed that if an aid message matches a problem report, the problem nucleus template and aid nucleus template are often distributionally similar. A typical example is "X is sold out" and "buy X". SIM4 captures this tendency. As the distributional similarity between templates, we used a Bayesian distributional similarity measure proposed by Kazama et al. (2010).[5]

CTP indicates whether the problem and aid nuclei are in contradiction relation or not. This feature was implemented based on the observation that when problem and aid nuclei are in contradiction relation, they are often proper matches (e.g., ⟨blackout, "X starts"⟩ and ⟨blackout, "X ends"⟩). CTP indicates whether nucleus pairs are in the one million contradiction phrase pairs[6] automatically obtained by applying a method proposed by Hashimoto et al. (2012) to 600 million Web pages.

5 Experiments

We evaluated our problem report recognizer and problem-aid match recognizer. For the sake of space, we give only the performance figures of the aid message recognizer at the end of Section 5.1.

We collected tweets posted during and after the 2011 Great East Japan Earthquake, between March 10 and April 4, 2011. After applying keyword-based filtering with a list of over 300 disaster related keywords, we obtained 55 million tweets. After dependency parsing[7], we used them in our evaluation.

5.1 Problem Report Recognition

Firstly, we evaluated our problem report recognizer. Particularly, we assessed the effect of excitation polarities and trouble expressions in two settings. The first is against a naturally distributed gold standard data. The second targets problem reports with problem nuclei unseen in the training data.

In both experiments we observed that the performance drops when excitation polarities and trouble expressions are removed from the feature set. The performance drop was larger in the second experiment which suggests that the excitation polarities and trouble expressions are more effective against unseen problem reports.

Training and test data for problem report recognition consist of tweet-nucleus candidate pairs randomly sampled from our 55 million tweet data. The training data (R) and test data (T) consist of 13,000 and 1,000 pairs, respectively, manually labeled by three annotators (other than the authors) as problem or other. Final judgment was made by majority vote. The Fleiss' kappa score for training and test data for annotation judgement is 0.74 (substantial agreement).

Our problem report recognizer and its variants are listed in Table 3. Table 4 shows the evaluation results. The proposed method achieved about 44% recall and nearly 80% precision, outperforming all other systems in terms of precision, F-score and average precision[8]. The improvement in precision when using TR&EX is statistically significant ($p < 0.05$).[9] Note that F-measure dropped

PROPOSED: Our proposed method with all features used.
PROPOSED-*: The proposed method without the feature set denoted by "*". Here EX and TR denote all excitation polarity and trouble expression related features, respectively, including their combinations (TREX1).
PROPOSED+OET: The proposed method incorporating the classification results of problem nucleus candidates by the Opinion Extraction Tool as additional binary features.
RULE-BASED: The method that regards only nuclei satisfying the constraint in Table 1 as problem nuclei.

Table 3: Evaluated problem report recognizers.

[5] The original similarity was defined over noun pairs and it was estimated from dependency relations. Obtaining similarity between template pairs, not noun pairs, is straightforward given the same dependency relations. We used 600 million Web pages for this similarity estimation.

[6] The precision of the pairs was reported as around 70%.

[7] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?KNP

[8] We calculate average precision using the formula: $aP = \frac{\sum_{k=1}^{n}(Prec(k) \times rel(k))}{n}$, where $Prec(k)$ is the precision at cut-off k and $rel(k)$ is an indicator function equaling 1 if the item at rank k is relevant, zero otherwise.

[9] Throughout this paper we performed two-tailed test of

Recognition system	**R** (%)	**P** (%)	**F** (%)	**aP** (%)
PROPOSED	44.26	**79.41**	**56.83**	**71.82**
PROPOSED-TR&EX	**45.08**	74.83	56.26	69.67
PROPOSED-EX	44.67	74.66	55.89	69.90
PROPOSED-TR	43.85	74.31	55.15	69.44
PROPOSED-MSA	28.69	70.71	40.81	57.74
PROPOSED-SWC	43.42	75.97	55.25	70.61
PROPOSED-WSP	43.14	77.83	55.50	70.45
PROPOSED-REQ	42.64	76.16	55.50	54.67
PROPOSED-GL	44.14	78.34	55.50	56.46
PROPOSED+OET	44.24	**79.41**	56.82	71.81
RULE-BASED	30.32	67.96	41.93	n/a

Table 4: Recall (R), precision (P), F-score (F) and average precision (aP) of the problem report recognizers.

whenever each type of feature was removed, implying that each type of feature is effective in this task. Especially note the performance drop if we remove excitation polarities (EX), trouble expression (TR) and both excitation and trouble expression features (TR&EX), confirming that they are crucial in recognizing problem reports with high accuracy. Also note that the performance of PROPOSED+OET was actually slightly worse than that of the proposed method. This suggests that full-fledged sentiment analysis is not effective at least in this setting. The rule-based method achieved relatively high precision despite of the low recall, demonstrating the importance of problem and aid nuclei formulations described in Section 1.

The second experiment assessed the efficiency of our problem report recognizer against unseen problem nuclei under the condition that every template in nuclei has excitation polarity. We sampled the training and test data so that the problem nucleus nouns and templates in the training and test data are disjoint. First we created a subset of the test data by selecting the samples which had nuclei with excitation templates. We call this subset T'. Next, we removed samples from training data R if either of their problem nouns or templates appeared in the nuclei of T'. The resulting new training data (called R') and test data (T') consist of 6,484 and 407 tweet-nucleus candidate pairs, respectively. We trained our problem report recognizer using R' and tested its performance using T'. Figure 2 shows the precision-recall curves obtained by changing the threshold on the SVM scores. The effectiveness of excitation polarities and trouble expressions was more evident in this setting. The PROPOSED's performance was actually better in this setting (almost 50% recall at more that 80% precision), than the previous setting, showing that excitation templates and trouble expressions are crucial in achieving high performance especially for unseen problem nuclei. The same was confirmed when we removed excitation polarity and trouble expression related features, with performance dropping by 7.43 points in terms of average precision. The improvement in precision when using TR&EX is statistically significant ($p < 0.01$). This implies, assuming that we have a wide-coverage dictionary of templates with excitation polarities, that excitation polarities are important in dealing with unexpected problems in disaster situations.

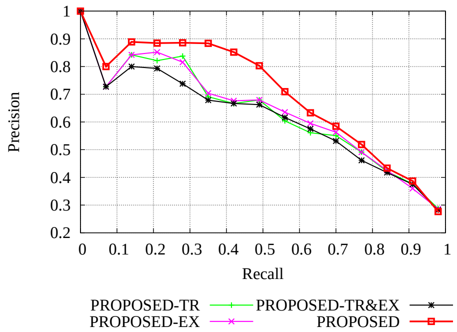

Figure 2: Precision-recall curves of problem report recognizers against unseen problem nuclei.

We also evaluated the aid-message recognizer, using tweet-nucleus pairs in R and T as training and test data and the annotation scheme was also the same. The average Fleiss' kappa score was 0.55 (moderate agreement). Our recognizer achieved 53.82% recall and 65.67% precision and showed similar tendencies with the problem report recognizer, with the excitation polarities and trouble expressions contributing to higher accuracy.

We can conclude that excitation polarities and trouble expressions are important in identifying problem reports and aid messages during disaster situations.

5.2 Problem-Aid Matching

Next, we evaluated the performance of the problem-aid match recognizer. We applied our problem report recognizer and aid message recognizer to all 55 million tweets and combined the tweet-nucleus pairs judged as problem reports and aid messages, respectively, to create the training and test data.

The training data consists of two parts ($M1$ and $M2$). $M1$ includes many variations of the aid messages for each problem report, while $M2$ en-

population proportion (Ott and Longnecker, 2010) using SVM-threshold=0.

sures diversity in nouns and templates in problem nuclei. For $M1$, we randomly picked up problem reports from the output of the problem report recognizer and to each we attached up to 30 randomly picked, distinct aid messages that have the same nucleus noun. Building $M2$ follows the construction method of $M1$, except that: (1) we used up to 30 distinct problem nuclei for each noun; (2) for each problem report we attached only *one* randomly picked aid message.

In creating the test data $T2$, we followed the construction method used for $M2$ to assess the performance of our proposal with a large variety of problems. $M1$, $M2$ and $T2$ consist of 3,000, 6,000 and 1,000 samples, respectively. The annotation was done by majority vote of three human annotators (other than the authors), the average Fleiss' kappa-score for training and test data was 0.63 (substantial agreement).

We trained the problem-aid match recognizers of Table 5 with $M1$ and $M2$. The evaluation results performed on $T2$ are shown in Table 6. We can observe that, among the nuclei related features, the trouble expression (TR) and excitation polarity (EX) features and their combination (TR&EX) contribute most to the performance, although the contribution of nuclei related features is less in comparison to the problem report and aid message recognition. The improvement in precision when using TR&EX is marginally significant ($p = 0.056$). Instead, morphological and syntactic analysis (MSA) and semantic word class (SWC) features greatly improved performance.

As the final experiments, we evaluated top-ranking matches of our problem-aid match recognizer, where the recognizer classified all the possible combinations of tweet-nuclei pairs taken from 55 million tweets. In addition, we assessed the effectiveness of excitation polarities and trouble expressions by comparing all positive matches produced by our full problem-aid match recognizer (PROPOSED) and those produced by the problem-aid match recognizer (PROPOSED-TR&EX) that did not use excitation polarities and trouble expressions in its feature set. Note that PROPOSED-TR&EX was fed by the problem report and aid message recognizers that didn't use excitation polarities and trouble expressions. For both systems' training data we used R for the problem report and aid message recognizers; $M1$ and $M2$ for the problem-aid matching recognizers.

PROPOSED: Our proposed method with all features used.
PROPOSED-*: The proposed method without the feature set denoted by "*". Here also EX and TR denote all excitation polarity and trouble expression related features, respectively, including their combinations (TREX1 and TREX2).
RULE-BASED: The method that judges only problem-aid nuclei combinations with opposite excitation polarities as proper matches.

Table 5: Evaluated problem-aid match recognizers.

Matching system	R (%)	P (%)	F (%)	aP (%)
PROPOSED	30.67	**70.42**	**42.92**	**55.16**
PROPOSED-TR&EX	28.83	67.14	40.33	53.99
PROPOSED-EX	**31.29**	67.11	42.68	54.19
PROPOSED-TR	30.56	69.33	42.42	54.85
PROPOSED-MSA	13.50	53.66	21.57	44.52
PROPOSED-SWC	26.99	67.69	38.59	52.23
PROPOSED-WSP	30.61	69.51	42.50	54.81
PROPOSED-CTP	30.06	70.00	42.05	54.94
PROPOSED-SIM	29.95	70.11	41.97	54.98
PROPOSED-REQ	30.58	70.25	42.61	54.67
PROPOSED-GL	30.61	70.31	42.65	55.02
PROPOSED-SSR	30.67	69.44	42.72	54.91
RULE-BASED	15.33	17.36	16.28	n/a

Table 6: Recall (R), precision (P), F-score (F) and average precision (aP) of the problem-aid match recognizers.

PROPOSED and PROPOSED-TR&EX output 15.2 million and 13.4 million positive matches, covering 1,691 and 1,442 nucleus nouns, respectively. Table 7 shows match samples identified with PROPOSED. We observed that the output of each system was dominated by just a handful of frequent nucleus nouns, such as "water" or "gasoline". We preferred to assess the performance of our system against a large variation of problem-aid nuclei, thus we restricted the number of matches to 10 for each noun[10]. After this restriction the number of matches found by PROPOSED and PROPOSED-

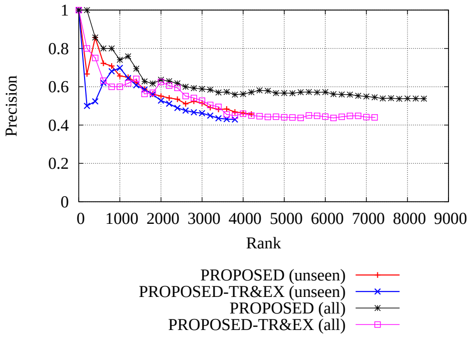

Figure 3: Problem-aid match recognition performance for 'all' and 'unseen' problem reports.

[10] Note that this setting is a pessimistic estimation of our system's overall performance, since according to our observations problem reports with very frequent nucleus nouns had proper matches with a higher accuracy than problem reports with less frequent nucleus nouns.

Problem report: いわきの常磐病院、いわき泌尿器科病院、竹林貞吉記念クリニック、泉中央クリニックは、１７日から透析を中止します。患者の方は至急連絡してください。 (*Starting from the 17th, the Iwaki Joban Hospital, the Iwaki Urology Clinique, the Takebayashi Sadakichi Memorial Clinique and the Izumi Central Clinique have all* ***suspended dialysis sessions****. Patients are advised to urgently make contact.*)
Aid message: いわき泌尿器科病院で短時間透析が可能です。受付時間は９時から１６時までです。（透析の再開） (*Restart of dialysis sessions: short* ***dialysis sessions are available*** *at the Iwaki Urology Clinique between 9 AM and 4 PM.*)
Problem report: ごめんなさい拡散をお願いしてもいいですか。仙台の父親の話ですと携帯の充電がもうない人が続出しているそうです。携帯充電器の支援が必要かと思われます。 (*Please spread this message. According to my father in Sendai, there are more and more people whose phones* ***ran out of battery****. We need phone chargers!)*
Aid message: 【拡散希望】仙台若林区役所で携帯電話の充電ができるそうです。 (*[Please spread] At the City Hall of Wakabayashi-ku, Sendai, you* ***can recharge*** *your phone* ***battery****.*)

Table 7: Examples from the output of the proposed method in the 'all' setting. Problem report and aid message nuclei are boldfaced in the English translations.

TR&EX was 8,484 and 7,363, respectively.

The performance of PROPOSED and PROPOSED-TR&EX were assessed in two settings: 'all' and 'unseen'. For 'all', we selected 400 problem-aid matches from the outputs of the respective systems after applying the 10-match restriction. For 'unseen', first we removed the samples from the systems' outputs if either the nucleus noun or template pair appear in the nuclei of the problem-aid match recognizers' training data. Next we applied the same sampling process as with 'all'. Three annotators (other than the authors) manually labeled the sample sets, final judgment being made by majority vote. The Fleiss' kappa score for all test data was 0.73 (substantial agreement).

Figure 3 shows the systems' precision curves, drawn from the samples whose X-axis positions represent the ranks according to SVM scores. In both scenarios we can confirm that excitation polarity and trouble expression related features contribute to this task. In the 'all' setting in terms of average precision calculated over the top 7,200 matches, PROPOSED's 62.36% is 10.48 points higher than that of PROPOSED-TR&EX. For unseen problem/aid nuclei PROPOSED method's average precision of 58.57% calculated at the top 3,800 matches is 5.47 points higher than that of PROPOSED-TR&EX at the same data point. The improvement in precision when using TR&EX is statistically significant in both settings ($p < 0.01$).

6 Related Work

Twitter has been observed as a platform for situational awareness during various crisis situations (Starbird et al., 2010; Vieweg et al., 2010), as sensors for an earthquake reporting system (Sakaki et al., 2010; Okazaki and Matsuo, 2010) or to detect epidemics (Aramaki et al., 2011). Besides Twitter, blogs or forums have also been the target of community response analysis (Qu et al., 2009; Torrey et al., 2007). Similar to our work are the ones of Neubig et al. (2011) and Ishino et al. (2012), who tackle specific problems that occur during disasters (i.e., *safety information* and *transportation information*, respectively); and Munro (2011), who extracted "actionable messages" (requests and aids, indiscriminately), matching being performed manually. Our work differs from (Neubig et al., 2011) and (Ishino et al., 2012) in that we do not restrict the range of problem reports, and as opposed to (Munro, 2011), matching is automatic.

Systems such as that of Seki (2011)[11] or Munro (2013)[12] are successful examples of crisis crowdsourcing, but these require extensive human intervention to coordinate useful information.

Another category of related work relevant to our task is troubleshooting. Baldwin et al. (2007) and Raghavan et al. (2010) use discussion forums to solve technical problems using supervised learning methods, but these approaches presume that the solution of a specific problem is within the same thread. In our work we do not employ structural characteristics of tweets as restrictions (e.g., a problem report and its aid message need to be in the same tweet chain).

7 Conclusions

In this paper, we proposed a method to discover matches between problem reports and aid messages from tweets in large-scale disasters. Through a series of experiments, we demonstrated that the performance of the problem-aid matching can be improved with the usage of semantic orientation of excitation polarities, proposed in (Hashimoto et al., 2012), and trouble expressions.

We are planning to deploy our system and release model files of the classifiers to assist relief efforts in future crisis scenarios.

[11] http://www.sinsai.info/

[12] http://www.mission4636.org/

References

Adam Acar and Yuya Muraki. 2011. Twitter for crisis communication: Lessons learned from Japan's tsunami disaster. *International Journal of Web Based Communities*, 7(3):392–402.

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

Timothy Baldwin, David Martinez, and Richard B. Penman. 2007. Automatic thread classification for Linux user forum information access. In *Proceedings of the 12th Australasian Document Computing Symposium (ADCS 2007)*, pages 72–79.

Stijn De Saeger, Kentaro Torisawa, and Jun'ichi Kazama. 2008. Looking for trouble. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, pages 185–192.

Stijn De Saeger, Kentaro Torisawa, Masaaki Tsuchida, Jun'ichi Kazama, Chikara Hashimoto, Ichiro Yamada, Jong-Hoon Oh, István Varga, and Yulan Yan. 2011. Relation acquisition using word classes and partial patterns. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, pages 825–835.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 5:378–382.

Chikara Hashimoto, Kentaro Torisawa, Stijn De Saeger, Jong-Hoon Oh, and Jun'ichi Kazama. 2012. Excitatory or inhibitory: A new semantic orientation extracts contradiction and causality from the web. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012)*, pages 619–630.

Aya Ishino, Shuhei Odawara, Hidetsugu Nanba, and Toshiyuki Takezawa. 2012. Extracting transportation information and traffic problems from tweets during a disaster: Where do you evacuate to? In *Proceedings of the Second International Conference on Advances in Information Mining and Management (IMMM 2012)*, pages 91–96.

Hiroshi Kanayama and Tetsuya Nasukawa. 2008. Textual demand analysis: Detection of users' wants and needs from opinions. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, pages 409–416.

Jun'ichi Kazama and Kentaro Torisawa. 2008. Inducing gazetteers for named entity recognition by large-scale clustering of dependency relations. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-08: HLT)*, pages 407–415.

Jun'ichi Kazama, Stijn De Saeger, Kow Kuroda, Masaki Murata, and Kentaro Torisawa. 2010. A Bayesian method for robust estimation of distributional similarities. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, pages 247–256.

Jun'ichi Kazama, Stijn De Saeger, Kentaro Torisawa, Jun Goto, and István Varga. 2013. Saigaiji jouhou e no shitsumon outo shisutemu no tekiyou no kokoromi. (An attempt for applying question-answering system on disaster related information). In *Proceeding of the Nineteenth Annual Meeting of The Association for Natural Language Processing*. (in Japanese).

Benjamin Mandel, Aron Culotta, John Boulahanis, Danielle Stark, Bonnie Lewis, and Jeremy Rodrigue. 2012. A demographic analysis of online sentiment during Hurricane Irene. In *Proceedings of the Second Workshop on Language Analysis in Social Media (LASM 2012)*, pages 27–36.

Robert Munro. 2011. Subword and spatiotemporal models for identifying actionable information in Haitian Kreyol. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL-2011)*, pages 68–77.

Robert Munro. 2013. Crowdsourcing and the crisis-affected community. *Information Retrieval*, 16(2):210–266.

Tetsuji Nakagawa, Kentaro Inui, and Sadao Kurohashi. 2010. Dependency tree-based sentiment classification using CRFs with hidden variables. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010)*, pages 786–794.

National Police Agency of Japan. 2013. Damage situation and public countermeasures associated with 2011 Tohoku district – off the Pacific Ocean Earthquake. `http://www.npa.go.jp/archive/keibi/biki/higaijokyo_e.pdf`. (accessed on 30 April, 2013).

Graham Neubig, Yuichiroh Matsubayashi, Masato Hagiwara, and Koji Murakami. 2011. Safety information mining — what can NLP do in a disaster —. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, pages 965–973.

Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Takuya Kawada, Stijn De Saeger, Jun'ichi Kazama, and Yiou Wang. 2012. Why question answering using sentiment analysis and word classes. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012)*, pages 368–378.

Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Motoki Sano, Stijn De Saeger, and Kiyonori Ohtake. 2013. Why-question answering using intra- and inter-sentential causal relations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*.

Kiyonori Ohtake, Kentaro Torisawa, Jun Goto, and Stijn De Saeger. 2013. Saigaiji ni okeru hisaisha to kyuuen kyuujosha kan no souhoko komyunikeeshon. (Bi-directional communication between victims and rescures during a crisis). In *Proceeding of the Nineteenth Annual Meeting of The Association for Natural Language Processing*. (in Japanese).

Makoto Okazaki and Yutaka Matsuo. 2010. Semantic Twitter: Analyzing tweets for real-time event notification. In *Proceedings of the 2008/2009 international conference on Social software: Recent trends and developments in social software (BlogTalk 2008)*, pages 63–74.

R. Lyman Ott and Michael T. Longnecker, 2010. *An Introduction to Statistical Methods and Data Analysis*, chapter 10.2. Brooks Cole, 6th edition.

Yan Qu, Philip Fei Wu, and Xiaoqing Wang. 2009. Online community response to major disaster: A study of Tianya forum in the 2008 Sichuan Earthquake. In *42st Hawaii International International Conference on Systems Science (HICSS-42)*, pages 1–11.

Preethi Raghavan, Rose Catherine, Shajith Ikbal, Nanda Kambhatla, and Debapriyo Majumdar. 2010. Extracting problem and resolution information from online discussion forums. In *Proceedings of the 16th International Conference on Management of Data (COMAD 2010)*.

Takeo Saijo. 2012. *Hito-o tasukeru sungoi shikumi. (A stunning system that saves people)*. Diamond Inc. (in Japanese).

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 2010)*, pages 851–860.

Motoki Sano, István Varga, Jun'ichi Kazama, and Kentaro Torisawa. 2012. Requests in tweets during a crisis: A systemic functional analysis of tweets on the Great East Japan Earthquake and the Fukushima Daiichi nuclear disaster. In *Papers from the 39th International Systemic Functional Congress (ISFC39)*, pages 135–140.

Haruyuki Seki. 2011. Higashi-nihon daishinsai fukkou shien platform sinsai.info no naritachi to kongo no kadai. (The organizational structure of sinsai.info restoration support platform for the 2011 Great East Japan Earthquake and future challenges). *Journal of digital practices*, 2(4):237–241. (in Japanese).

Kate Starbird, Leysia Palen, Amanda L. Hughes, and Sarah Vieweg. 2010. Chatter on the red: What hazards threat reveals about the social life of microblogged information. In *Proceedings of The 2010 ACM Conference on Computer Supported Cooperative Work (CSCW 2010)*, pages 241–250.

Cristen Torrey, Moira Burke, Matthew L. Lee, Anind K. Dey, Susan R. Fussell, and Sara B. Kiesler. 2007. Connected giving: Ordinary people coordinating disaster relief on the Internet. In *Proceedings of the 40th Annual Hawaii International Conference on System Sciences (HICSS-40)*, pages 179–188.

Katerina Tsagkalidou, Vassiliki Koutsonikola, Athena Vakali, and Konstantinos Kafetsios. 2011. Emotional aware clustering on micro-blogging sources. In *Proceedings of the 4th international conference on Affective computing and intelligent interaction (ACII 2011)*, pages 387–396.

Sarah Vieweg, Amanda L. Hughes, Kate Starbird, and Leysia Palen. 2010. Microblogging during two natural hazards events: What Twitter may contribute to situational awareness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2010)*, pages 1079–1088.

Patrick Winn. 2011. Japan tsunami disaster: As Japan scrambles, Twitter reigns. *GlobalPost*, 18 March.