Safety Information Mining — What can NLP do in a disaster —

Graham Neubig
Kyoto University
Yoshida Honmachi, Sakyo-ku, Kyoto, Japan

Yuichiroh Matsubayashi
National Institute of Informatics
Chiyoda-ku, Tokyo

Masato Hagiwara, Koji Murakami
Rakuten Institute of Technology
New York

Abstract

This paper describes efforts of NLP researchers to create a system to aid the relief efforts during the 2011 East Japan Earthquake. Specifically, we created a system to mine information regarding the safety of people in the disaster-stricken area from Twitter, a massive yet highly unorganized information source. We describe the large scale collaborative effort to rapidly create robust and effective systems for word segmentation, named entity recognition, and tweet classification. As a result of our efforts, we were able to effectively deliver new information about the safety of over 100 people in the disaster-stricken area to a central repository for safety information.

1 Introduction

On March 11, 2011 at 14:46[1], a massive earthquake of Magnitude 9.0 struck off the coast of Japan. The earthquake and the ensuing Tsunami caused damage across the entire eastern coast of the country, with homes destroyed, roads impassible, and large swaths of the disaster-stricken area losing electricity and the ability to communicate.

Figure 1 shows the death toll and the number of missing people from tsunami and earthquake, presented by the Japanese National Police Agency. The number of missing people reached its peak on March 25th, indicating that the government took 2 weeks to fully grasp the total number of victims, although they began gathering information about the whereabouts of victims and survivors as soon as the earthquake hit. One of the main reasons for this delay was that prefectures could not effectively collect the information from municipal governments or police stations because many of them

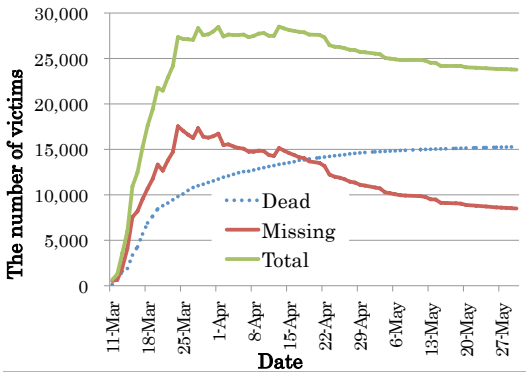

Figure 1: Change of overall death toll and missing people

were extensively damaged or destroyed by the disaster (Anonymous, 2011)[2]. Furthermore, damage to communication equipment, power outages, and inundation of the air waves prevented the use of mobile phones, which are generally the most important tool for communication during a disaster.

As a result, Twitter or social network services (SNS) such as mixi[3] played an important role for propagating safety information among people. However, with SNSs flooded with information about the disaster, it was difficult to ensure that people would be properly connected with the information they were seeking. Soon after the disaster struck, many NLP researchers, engineers, and students from all over Japan (including the authors), spontaneously created a working group called "ANPI_NLP"[4] to try to organize this information into a form that would be useful. In particular, we focused on mining and organizing information on Twitter regarding safety of individuals in the disaster-stricken area, as people are initially more concerned about the safety of their family or

[1] All the dates and times in this paper are in JST.

[2] http://www.asahi.com/special/10005/TKY201103120549.html

[3] http://mixi.jp/

[4] The word "ANPI" means "safety" in Japanese.

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 965–973,
Chiang Mai, Thailand, November 8 – 13, 2011.

friends than anything else.

There were three major elements to this process:

- Separating the tweets that contained safety information from the huge number of irrelevant tweets.
- Extracting important information, such as person names and locations from highly domain-specific text included in the tweets.
- Verifying and delivering this information to the people that need it.

The working group tackled these issues in safety information extraction from Twitter, preparing resources, building tools, and connecting the extracted information with Google Person Finder (GPF)[5], a central location for safety information that was widely publicized through the Japanese news media. Our priorities were not only the accuracy, but also the speed with which we could provide the information.

In this paper, we report the process of building an information extraction system in a disaster-response situation within a matter of days. The challenges involved included organizing volunteers (§2), building resources for out-of-domain, noisy internet data (§3), and applying these volunteers and resources to the construction of NLP tools (§4 and §5). We also describe the final results of the project, the provision of new information about missing persons to GPF (§6), and summarize related work (§7). Finally in §8 we discuss what went right and what went wrong with the project, and provide some insight into what can be done to prepare for similar situations in the future.

2 Organization and Communication

The project began with a single Twitter post soon after the earthquake hit, imploring NLP researchers and engineers to think about what they could do to help in the time of need. In particular, the creation of resources and tools that could be used to process information about the earthquake was cited as one example. In under an hour, the first responders to this call had started creating the language resources described in §3.

Given the public nature of Twitter, word of the efforts spread, and the number of participants quickly increased. With the large number of weakly connected participants, there was no clear power structure in place to delegate authority. Instead, several leaders spontaneously emerged, centered around people who joined the project early, had large Twitter followings, or were experts in a specific area (such as the creation of data or domain adaptation of tools). In addition to the discourse on Twitter, a publicly available Wiki page was created to gather information about the project in a single place[6]. Overall, we believe that the existing online communication framework proved quite effective in rallying and organizing members for the project.

On the other hand, the largest challenge in the project organization was the initial underestimation of the outpouring of support that the project would see. In the end, over 65 volunteers joined the project, and it was often difficult for the few main organizers to rapidly design and delegate tasks to such a large number of volunteers. This resulted in an over-concentration of effort in some areas, and lack of effort in others. We hope that this paper will help provide a road-map for the type of tasks that may be necessary in rapid-response NLP situations, and prompt discussion on what other tasks may be taken on in a disaster.

3 Language Resources and Tweet Corpus

The earliest stage of the project focused on sharing linguistic resources. These included dictionaries, which were used to improve various text analyzers and classifiers. At the same time, we also made efforts towards building a labeled corpus of tweets containing safety information, with the final goal of extracting information from unlabeled tweets.

3.1 Text Analysis Resources

Among the earliest contributions to the project were dictionaries, especially those containing person and place names specific to the Tohoku region. These were generally contributed by people directly familiar with the resources, the creators or maintainers of the dictionaries themselves.

As time was critical, the linguistic analysis tools actually used in the project were based on widely available general domain resources, as well as domain-specific resources gathered within the very early stages of the project. The general domain language resources consisted of the Balanced Corpus of Contemporary Written Japanese

[5] http://japan.person-finder.appspot.com/

[6] http://trans-aid.jp/ANPI_NLP (in Japanese).

(BCCWJ) (Maekawa, 2008) and the UniDic dictionary (Den et al., 2008), which are high quality and annotated with a variety of tags.

The domain-specific resources gathered specifically for the project and used in the analysis tools described in §4 included:

- The dictionary used in the open source Mozc Japanese Input method, which contained 50,848 first names and 26,519 last names and was provided by the maintainer of the project.
- A name list that contains last names specific to northeast Japan[7]. These resources were publicly available on the web.
- A location name list of Iwate, Miyagi, Fukushima, and Ibaraki prefectures created by downloading a database of postal code information and manually checking the data.

A number of other resources were created by volunteers, including a list of all the station names in Japan, station location information, landmark names in Kanto and Tohoku extracted from Wikipedia, a list of geopolitical entities, and a list of abbreviated school names and places. While these resources were certainly significant, it was the resources that were prepared early on in the process that provided the most aid to the project as a whole. This indicates that for similar situations in the future, it is essential to have as many resources as possible immediately available, and preferably familiar to the people in the project to facilitate rapid processing and dispersal.

3.2 Tweet Corpus Construction

As Twitter contained a wide variety of earthquake-related posts including information about the safety of people in the disaster-affected area, we decided to create a corpus of disaster-related tweets to aid our information extraction efforts. The first tweet corpus shared was the collection of 469,504 tweets containing the word "地震" (earthquake) from March 11th 16:09 to March 13th 8:59. We also collected tweets with the hash tags "#anpi" (safety information) such as "#hinan" (evacuation), "#j_j_helpme" (help request), and "#save_"+ location names. Tweets containing RT (re-tweets) and QT (quote tweets) were removed to eliminate duplication. As a result, 61,376 tweets were collected from March 13th 1:37am until March 14th 16:45pm.

A typical tweet containing safety information looked like the following [8]:

> 気仙沼市の田中太郎・花子さんと連絡が取れません！どなたか消息をご存知ありませんでしょうか？
> TANAKA Taro and Hanako who lived in Kesennuma City can't be reached. Does anybody know where they are?

From the large corpus of tweets, we hope to discover two pieces of information. First, we must recognize the *topic* of the tweet, in this case that the tweet is asking for information about missing people (tweet classification). Second, we need to recognize that the people in question are "TANAKA, Taro" and "Hanako," both of whom live in "Kesennuma City" (NE recognition).

In order to create data to train tools to perform both tweet classification and NE recognition, volunteers began to perform tagging as soon as the tweet collection was finished. First, we defined nine kinds of tweet labels specifying the tweet topics, which are listed in Table 1. The distinction among S/O/U was sometimes unclear and went under extensive discussion among volunteers. In practice, the distinction of I/L/P/M from others (S/O/U) is of the highest importance.

One example is shown below:

```
<person type="M">TANAKA Taro</person> and
<person type="M">Hanako</person> living in
<location>Kesennuma city</location> can't
be reached.
```

Safety information is optionally added to every person tag, specifying the object of the safety information, as shown in the above example. Also multiple tags are permitted when multiple types of safety information are contained in a single tweet.

Utilizing Twitter, word was spread about the tagging task force, gathering a large number of volunteers, most of whom were NLP researchers, engineers, or students. In order to facilitate the task-force, several people with annotation experience responded to questions in real-time, and various tools were created and shared among the annotators. Over 65 volunteers gathered to tag the corpus with class and NE tags over the course of two days, resulting in a total of 33,242 tweet annotated.

[7] http://platinum22000.fc2web.com/{miyagi,fukushima,iwate,tochigi}.htm

[8] The actual person and place names have been replaced with pseudonyms to preserve anonymity.

Label	Definition	Example	Count
I	Him/Herself is alive	I'm XXX in YYY City. I'm all right.	405
L	Alive	Mr./Ms. XXX in YYY City is at ZZZ Shelter. He/She is alive.	1,154
P	Passed away	—	93
M	Missing	The safety of Mr./Ms. XXX living in YYY City is unknown.	4,438
H	Help Request	Mr./Ms. XXX is left in YYY and needs help!	280
		My relatives/parents/... staying in South XXX City are missing.	
S	Information request	Refugees at XXX School are provided enough daily supplies? (Safety information of unspecific individual, region, etc.)	1,903
O	Not safety information	You can post safety information on this site!	24,035
R	External link	Survivor list of XXX City: http://...	773
U	Unknown	(Non-Japanese or nonsense postings)	1,235

Table 1: Safety information tags on tweets

While the annotators were generally more skilled and motivated than those in previous attempts to create language resources using crowdsourcing (Callison-Burch and Dredze, 2010; Finin et al., 2010), given the rapid nature by which the project developed, annotation started before tagging standards were put in place, leading to some inconsistency in the tagged corpus. In retrospect, despite the speed of the project, it would have been helpful to spend some more time thinking about what information was really necessary for the task, and have more experienced annotators do a quick test run before opening annotation to the broader volunteer base. In addition, as many of the annotators were less experienced, explicitly allowing the annotators to "pass" on difficult instances would have reduced the amount of time wasted on difficult or ambiguous cases.

4 Text Analysis

Before mining the tweets for useful information, it was necessary to create text analysis tools that were capable of handling this very specialized data set. The main objective of text analysis was named entity recognition (NER), which would allow for the identification of names and locations.

4.1 Morphological Analysis and NER

The first step in Japanese text processing is morphological analysis (MA), which splits raw Japanese text into words and assigns POS tags to each word. While previous research in tagging for Twitter has profited by building a custom NLP pipelines over the course of several months (Gimpel et al., 2011; Ritter et al., 2011) or by building semi-supervised learning systems (Liu et al., 2011), it was necessary to create a morphological analyzer in the course of several hours. This is due to the fact that all other components of the system depend on MA, and cannot be developed until MA is in place. Thus, we utilized existing general domain resources, and added new resources as they were collected in a domain-adaptation framework.

For this task we used KyTea[9], an open source morphological analysis tool notable for being relatively robust to out-of-domain data, and being able to flexibly incorporate a variety of language resources (Neubig et al., 2011).

We trained a word segmentation (WS) and POS tagging model for KyTea using the BCCWJ and UniDic as a base. We trained the POS tagging model, but in order to facilitate NE recognition farther down the pipeline, we replaced all proper nouns with their subcategory tag ("first name," "place name," etc.). We also added a corpus of conversational and news text (CN Corpus) that was only annotated with word boundaries, and a large list of Japanese first and last names. We indicate the model trained with all of these resources as ORIG.

While the POS tagger works on a word-by-word basis, most named entities consist of multiple words. Previous work has developed linguistic resources for English NE tagging on Twitter (Finin et al., 2010), but again considering the short time frame, we developed a simple rule-based system to connect multiple words into single named entities. Rules scanned the corpus in order, finding the first of three POS tags: "first name (FNAME)," "last name (LNAME)," or "place name (PNAME)." These words are labeled with PERSON, PERSON, or LOCATION NE tags respectively. Continuing in the order of the corpus, all words directly following a marked NE are merged if marked with one of the three previously mentioned POS tags, or as a "suffix (SUF)[10]". An example of the three step

[9] Available at `http://www.phontron.com/kytea`.

[10] A small set of suffixes that are used in conjunction with person names such as "*-san*" or "*-kun*" were omitted.

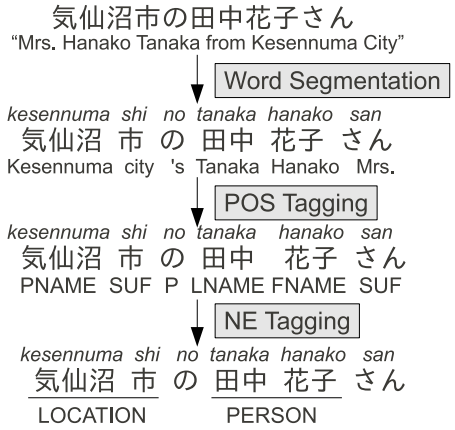

Figure 2: An example of the three steps of named entity recognition. Pronunciations and English translations are also provided for reference.

Name	Words	WS	POS	Group
BCCWJ	997k	○	○	ORIG
CN Text	503k	○		ORIG
UniDic	217k	○	○	ORIG
Names	144k	○	○	ORIG
Address	73.4k	○	○	+DICT
+Names	29.4k	○	○	+DICT
Active	10.2k	○		+ACTIVE

Table 2: Resources used in building the model with names, word counts, whether each corpus is annotated with WS and POS information, and which group the resource was added in.

NE tagging process is shown in Figure 2.

4.2 Domain Adaptation

While this classifier worked well on general domain data, it is known that accuracy greatly decreases for text in different domains or styles than the training data (Finin et al., 2010; Neubig et al., 2011; Ritter et al., 2011). In the tweet data there were a large number of place and person names specific to the disaster-stricken region, as well as a large number of linguistic phenomena specific to tweets, and thus it was necessary to add a number of language resources (summarized in Table 2) to adapt the text processing tools to the new domain.

To improve the accuracy on person and place names, we added the language resources previously described in section 3. The combination of these dictionaries is indicated by “+Names” in Table 2, and a model trained adding these resources is indicated with +DICT.

Finally, to handle the the large number of Twitter-related linguistic phenomena such as the word “tweet” or hash tags, we annotated word boundaries using the active-learning/partial-annotation method described by Neubig et al. (2011). For each round of active learning, we chose the 100 words for which the morphological analyzer was least confident, and had a single human annotator correct the word boundaries for these words. This was repeated for 4 rounds (400 words), which took approximately 1.5 hours total. The model trained by adding this data to +DICT is indicated as +ACTIVE.

		ORIG	+DICT	+ACTIVE
WS	F	97.3%	97.3%	97.7%
Lab. NE	P	53.9%	69.6%	69.2%
	R	55.6%	71.7%	72.7%
Unlab. NE	P	70.6%	80.4%	80.8%
	R	72.7%	82.8%	84.9%

Table 3: Word segmentation F-measure and NE precision (P) and recall (R) for both labeled and unlabeled evaluation.

4.3 Result Analysis

In order to demonstrate the necessity and effectiveness of domain adaptation, we present results of a quantitative and qualitative analysis of WS and NER for ORIG, +DICT, and +ACTIVE.

Table 3 shows results for a quantitative analysis of WS and NE tagging results on a manually annotated corpus of 50 tweets. Labeled NE accuracy indicates that the NE is only considered correct if both its span and its label (PERSON/LOCATION) are correct, and unlabeled NE accuracy indicates that the NE is considered correct if the span is matched, regardless of the label. It can be seen that the addition of dictionaries for the target domain greatly increased the NE accuracy, up to 16% for labeled tagging. The active-learning-based annotation further improved the WS accuracy by 0.42%, resulting in gains of NE recall of 1-2%.

The main reason for the improvement between the ORIG and +DICT was the proper handling of place names in the disaster-stricken area. ORIG would split long addresses into several NEs, some of which were mistakenly recognized as as person names. For example, “気仙沼_市” (Kesennuma city) was mistakenly split into “気仙_沼_市” (Kesen swamp city) and “気仙” was recognized as a person name. Further, the active learning for WS fixed segmentation errors such as correcting the improperly segmented “平浄_水場” to the properly segmenting “平_浄水_場” (Taira water purify-

ing plant). This, in turn, allowed for "平" (Taira) to be properly recognized as a location name.

The largest remaining challenge for the fully adapted system was the distinction between last names and place names, which are highly ambiguous in Japanese. The use of the address information corpus greatly improved the NE tagging results, but also biased the classifier heavily towards place names. In a rapid-response situation it is necessary to use any and all of the resources available, even if they are known to be biased. This sample-bias problem can likely be ameliorated by recent progress in machine learning techniques (Huang et al., 2007).

5 Safety Information Classifier

Next, we describe the classifier that we built to find tweets containing safety information out of all the tweets in the corpus.

5.1 Model 1

Our first classifier, which we will call Model 1, was developed in several hours by over-simplifying the classification problem and using a very limited feature set.

While, as explained in §3.2, a single tweet could be assigned multiple labels, for Model 1 we simply took the first label that each annotator assigned to be true, and trained a one-vs.-rest classifier on this data. We used two types of features for the classification, bag-of-character-n-gram features, and NE features. The bag-of-character-n-gram features created a separate feature for each character n-gram from unigram to trigrams. As most Japanese words are relatively short, this is able to capture the most important words while not requiring word segmentation and thus being relatively robust. Second, we used the counts of each type of named entity tag, which are generally larger in tweets that contained useful information.

We trained all models using the default SVM solver of LIBLINEAR (Fan et al., 2008). 10-fold cross validation on the test set showed that the classifier achieved a tag accuracy of 86.13%.

5.2 Model 2

After few days, we created Model 2 as an extension of Model 1. Model 2 used a multi-class and multi-label classifier which output all the labels each of whose score exceeds a threshold t.

We extended the feature set for Model 2. Model 2 includes the same features as Model 1, but for n-grams, Model 2 does not count duplicated sequences. As NE features, co-occurrence information of PERSON and LOCATION in a single tweet was added as tweets for safety confirmation should generally include both of them. Combination of character 3-grams and the appearance of a PERSON was also added to capture a tweet's intention for that person. We also used existence of the hash tag "#anpi" and the total number of appearance of hash tags except for "#anpi" as clues for dividing tweets not related to safety confirmation (label O, U, S) from the others. Finally, we used the existence of the string "http(s)" and combination of existence of "http(s)" and NE as clues for the label R (external lists).

5.3 Evaluation

We evaluated the Model 2 classifier in detail using the corpus from the 17th of March, which contained 9,812 tweets. We divided the corpus into three sets: 80% for training, 10% for development, and the other 10% for test. For comparison, we also developed Model $1'$ which is a multi-label model using the same feature set as Model 1.

Table 4 shows the results of the two models. The performance is not significantly different in terms of accuracy, but Model 2 obtained higher recall than Model $1'$. The thresholds here were empirically determined to the one that maximize the F1-scores on the development set.

We described the result for each label with Model 2 in Table 5. Good performance was obtained for the O and M that comparatively have a larger number of instances. However, the recalls for the labels having few instances were lower. Especially, most of the label L were still missing. The most frequent mis-prediction was occurred between U, S and O. This is not a big problem since all these three labels are unrelated to safety confirmation. However, by analyzing errors, we found that there are a significant number of inconsistencies for these three labels depending on the annotators due to the inconsistent annotation standards mentioned in Section 2. A promising solution for situations like these is the recent development of methods and publicly available tools for extremely efficient active learning for text classification (Tong and Koller, 2002; Settles, 2011). These would allow for similar final results to be

		Dev				Test			
Model	t	P	R	F	Acc.	P	R	F	Acc.
Model 1′	0.55	87.8%	84.3%	86.0%	96.8%	87.9%	84.9%	86.4%	96.9%
Model 2	0.45	86.1%	85.7%	85.9%	96.7%	87.1%	86.4%	86.8%	97.0%

Table 4: Micro-averaged precision (P), recall (R), F-score (F), and accuracy (Acc.) achieved by Model 1′ and Model 2. For the evaluation, we transformed the gold and predicted data from multi-labeled instances to nine sets of binary-labeled instances.

Label	#samples	P	R	F
O	719	88.4%	98.3%	93.1%
M	134	91.2%	93.3%	92.3%
S	51	72.5%	56.9%	63.7%
L	45	50.0%	11.1%	18.2%
R	32	76.9%	31.3%	44.4%
U	22	0.0%	0.0%	0.0%
I	12	50.0%	58.3%	53.9%
H	6	100.0%	50.0%	66.7%
P	4	0.0%	0.0%	0.0%

Table 5: Precision (P), recall (R) and F-score (F) for each labels using Model 2 ($t = 0.45$).

Gold	Predict	#err.	Gold	Predict	#err.
U	O	22	LR	OR	2
S	O	17	R	LR	2
L	O	13	O	LR	2
LR	O	11	S	OS	2
R	O	9	MS	S	2
O	S	7	LPR	O	2
L	M	5	LM	M	2
M	IM	5	LM	IM	2
M	O	4	IM	M	2
R	OR	3			
H	O	3	Others		16
M	S	3	Total		136

Table 6: Label sequences where the errors occurred with Model 2 ($t = 0.45$).

achieved by a fewer number of annotators, reducing inconsistency issues.

6 Application of the System

The final step was verifying that the information was in fact reliable, and then matching PERSON and LOCATION tokens from tweets, which were classified as "L" (the person is alive) to the information of Google Person Finder (GPF). We chose to verify and match the information by hand to prevent the provision of misinformation on a sensitive topic such as safety of earthquake victims.

GPF contained several columns, such as the first name and last name of the person, home neighborhood, home city and home prefecture. The manually corrected NE information was matched with GPF data, and we were able to update the personal information of more than a hundred individuals that were confirmed to be alive in tweets.

However, many tokens extracted from tweets had problems, including:

- **Incomplete** LOCATION **information:** The tokens lacked a couple of information such as city, or neighborhood. "I was able to contact Taro Tanaka, who lives in Miyagi prefecture! [宮城県の田中太郎と無事に連絡が取れました！]". In this case, there may be many Taro Tanakas living in the prefecture, so a single individual cannot be identified.
- **Incomplete** PERSON **names:** a) Last or first names were be omitted when describing people of the same family. "I found Taro, Jiro, and Hanako Tanaka of Sendai at an evacuation shelter. [仙台市に住む田中太郎、次郎、花子が避難所にいました。]" or "The Tanaka family who live in Sendai has been reached! [仙台市に住む田中一家の安否が確認できました。]" b) Many person names that are likely written in logographic *kanji* in normal text were instead written phonetically (using *katakana* or *hiragana*).

While we focused mainly on extracting information about missing people, it has also been noted that Twitter is a source of other information in disaster situations (Corvey et al., 2010; Vieweg et al., 2010). For example, "50 people in Kesennuma city have evacuated to a hill behind the city hall. [気仙沼で被災し、市民会館の裏山に 50 人避難しています]" includes the number of people (50) and a concrete location (a hill behind the city hall), and could be a good indicator of where to concentrate rescue efforts. This could be an area of very practical application for recent research in verifying reliability of information on Twitter (Kawahara et al., 2008; Qazvinian et al., 2011).

7 Related Work

Besides the safety information mining task which we described in the previous sections, we observed many other efforts to help the earthquake

victims via NLP technologies as sub-projects of ANPI_NLP. Prominent ones include:

- **Association of tweets to evacuation shelter locations:** This includes geo-coding location names extracted from tweets and associating them with the evaluation shelter lists, which are also geo-coded.
- **Visualization of tweets on a map:** This includes mapping geo-tagged tweets on a map so that they can be searched easily.
- **Translation of information to foreign languages:** This includes compilation and sharing of technical term multilingual dictionaries, automatic translation of earthquake information, and provision of the translated information in four major languages spoken by foreigners in Japan.

We are definitely not the first to focus on the disaster-related natural language processing. Lewis (2010) reports the development project of Haitian Creole translation system in rapid response to the Haiti earthquake in 2010. While their motivations have a lot in common with ours, such as the necessity to set up a deployable system in a very short time span. Corvey et al. (2010) describe the annotation of a corpus about the Oklahoma wildfires, aiming at provision of broad-scale information as opposed to safety information mining about individuals. There have also been a number of works on detecting general trends from twitter, including work by Sakaki et al. (2010), who detect earthquakes based on Twitter postings. There are many other systems designed for information sharing in emergency, including the one described in (Hasegawa et al., 2005).

8 Conclusion

In this paper we presented a description of the efforts of a group of volunteers to create a useful NLP system in a short time frame in response to a natural disaster.

On the whole, we believe the project was a success. In an extremely short time-frame, we were able to design, implement, and run a system for extracting useful information from a highly unorganized and difficult-to-process information source. In the short term, this allowed us to provide information about the safety of over 100 people. In addition, this allowed us to clarify the framework required, and develop tools that can be used to allow for provision of information on an even larger scale in future disaster situations. On a more abstract level, we were able to show that NLP has the potential to make a contribution in a disaster-response situation, which we hope will provide an impetus for similar future projects.

However, a number of challenges remain, and we conclude by summarizing the major lessons learned in the project.

- **Speed is everything:** Gathering data, making analysis tools, creating a classifier, and providing information to the public all needed to be performed on a limited time frame. As each of these steps must be completed in order, a delay in any part would result in delays for the overall process. Thus, it was necessary to create working tools as fast as possible, even if this meant making sacrifices in accuracy and refining later.
- **A better annotation framework is needed:** To resolve the challenges posed by annotation in such a short time frame, we must focus on create of better standards considering only the necessary information, do test annotation runs to work out the kinks, and allow annotators to "pass" on difficult annotations.
- **Human resources must be used effectively:** The project summoned an outpouring of support much larger than originally expected. To harness all of the people that are willing to help, it is important to have an overall project vision, and be able to quickly identify new projects for volunteers to work on.

Disasters such as earthquakes are tragic, overwhelming times, with too much information coming in for any individual to handle. We hope that our work has demonstrated that there is a need and a means for processing of this information, and will motivate similar projects in the future.

Acknowledgments

While too numerous to list here, the authors would like to sincerely thank all of the over 65 participants in the project. Without their generous contributions of time, resources, and expertise, the work described here would have never been possible. Finally, we thank Taiichi Hashimoto, Atsushi Fujita, Shinsuke Mori, and anonymous reviewers for their helpful comments on this manuscript.

References

Anonymous. 2011. Yakusho/keisatsu kinou daun, anpi kakunin susumazu, higashinihon daishinsai (In Japanese). In *the Asahi Shimbun*, page 20. March 13.

Chris Callison-Burch and Mark Dredze. 2010. Creating speech and language data with amazon's mechanical turk. In *NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 1–12.

William J. Corvey, Sarah Vieweg, Travis Rood, and Martha Palmer. 2010. Twitter in mass emergency: What NLP can contribute. In *NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, pages 23–24, Los Angeles, California, USA, June. Association for Computational Linguistics.

Yasuharu Den, Junpei Nakamura, Toshinobu Ogiso, and Hideki Ogura. 2008. A proper approach to Japanese morphological analysis: Dictionary, model, and evaluation. In *6th International Conference on Language Resources and Evaluation (LREC)*.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: a library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874.

Tim Finin, William Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. 2010. Annotating named entities in Twitter data with crowdsourcing. In *NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 80–88, Los Angeles, June. Association for Computational Linguistics.

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

Satoshi Hasegawa, Kumi Sato, Shohei Matsunuma, Masaru Miyao, and Kohei Okamoto. 2005. Multilingual disaster information system: information delivery using graphic text for mobile phones. *AI & Society*, 19(3):265–278.

Jiayuan Huang, Alexander J. Smola, Arthur Gretton, Karsten M. Borgwardt, and Bernhard Scholkopf. 2007. Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems*, 19:601.

Daisuke Kawahara, Sadao Kurohashi, and Kentaro Inui. 2008. Grasping major statements and their contradictions toward information credibility analysis of web contents. In *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 393–397. IEEE.

William D. Lewis. 2010. Haitian Creole: how to build and ship an MT engine from scratch in 4 days, 17 hours, & 30 minutes. In *14th Annual Conference of the European Association for Machine Translation*.

Xiaohua Liu, Shaodian Zhang, Furu Wei, and Ming Zhou. 2011. Recognizing named entities in tweets. In *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 359–367, Portland, Oregon, USA, June.

Kikuo Maekawa. 2008. Balanced corpus of contemporary written Japanese. In *6th Workshop on Asian Language Resources*, pages 101–102.

Graham Neubig, Yosuke Nakata, and Shinsuke Mori. 2011. Pointwise prediction for robust, adaptable japanese morphological analysis. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT) Short Paper Track*, Portland, Oregon, USA, 6.

Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *2011 Conference on Empirical Methods in Natural Language Processing*, pages 1589–1599, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Conference on Empirical Methods in Natural Language Processing*.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *19th international conference on World wide web*, pages 851–860. ACM.

Burr Settles. 2011. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In *2011 Conference on Empirical Methods in Natural Language Processing*, pages 1467–1478, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Simon Tong and Daphne Koller. 2002. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 2:45–66, March.

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 *28th international conference on Human factors in computing systems*, pages 1079–1088. ACM.