NHK STRL at TREC 2018 Incident Streams track

Taro Miyazaki* **Kiminobu Makino*** **Yuka Takei** **Hiroki Okamoto** **Jun Goto**
NHK STRL (Science and Technology Research Laboratories)
{miyazaki.t-jw, makino.k-gg, takei.y-ek, okamoto.h-iw, goto.j-fw}
@nhk.or.jp

Abstract

We describe NHK STRL's models for the TREC 2018 Incident Streams track. The goal of this track is classifying incident related Tweets into information types such as InformationWanted and EmergingThreats. The number of provided pieces of training data is about 2,000, which is not enough for current machine learning methods. We propose two models to overcome this small amount of data scenario: a knowledge base-based model and a model that considers meta-information. In addition, we used two bag-of-words baseline models, a multi-layer perceptron-based one and a support vector machine-based one, for comparison. Evaluation results show that our models can classify Tweets with a rather high F1 score.

1 Introduction

Twitter has been playing an important role in getting to know what is occurring in the real world. There are many applications that use Twitter information, such as disaster monitoring (Ashktorab et al., 2014; Mizuno et al., 2016) and news material gathering (Vosecky et al., 2013; Hayashi et al., 2015). NHK has also been studying news material gathering targeted at disasters and societal accidents/incidents (Takei et al., 2017; Makino et al., 2018; Goto et al., 2018). Our models judge Tweets on the basis of whether they are able to be used as news material or not and classifies the Tweets into news genres, such as fires, floods, and car accidents. The basis of the models can be adopted for various applications. Therefore, we can adopt our models with little customization for the Incident Streams (IS) track of the Text REtrieval Conference (TREC) 2018.

The task of the IS track for TREC 2018 is classifying incident related Tweets along with their information type. In this shared task, we have around 2,000 Tweets as training/development data, and more than 20,000 Tweets as test data. Each data set includes Tweets related to several kinds of incident, as listed in Table 1. The Tweets are classified along with the information type, as listed with the number of Tweets belonging to each respective class in Table 2.

As shown in Table 2, a training set does not include much data, and it is unbalanced for classes, so we developed models by taking the following strategies into account.

- We put a high priority both on micro and macro F1 scores when choosing a fine-grained model, even though the IS track measures models by using only the micro F1 score. This is because data sets are unbalanced, so the micro F1 score may not show

Table 1: List of incidents.

Event types	Train/development set	Test set
Earthquake	2012 Costa Rica Earthquake	2012 Guatemala Earthquake 2012 Italy Earthquake 2014 Chile Earthquake 2015 Nepal Earthquake
Flood	2013 Flood Colorado	2012 Philippines Floods 2013 Alberta Floods 2013 Manila Floods 2013 Queensland Floods
Typhoon	2012 Typhoon Pablo	2011 Joplin Tornado 2013 Typhoon Yolanda 2014 Typhoon Hagupit
Bombing	2013 West Texas Explosion	2013 Boston Bombing 2015 Paris Attacks
Wildfire	2012 Fire Colorado	2013 Australia Bushfire
Shooting	2013 LA Airport Shooting	2018 FL School Shooting

Table 2: List of classes and number of pieces of data in training/development set.

Class	#	Class	#
Request-SearchAndRescue	0	Request-GoodsServices	0
Request-InformationWanted	10	CallToAction-Volunteer	2
CallToAction-MovePeople	26	CallToAction-Donations	15
Report-FirstPartyObservation	28	Report-Weather	41
Report-ThirdPartyObservation	15	Report-EmergingThreats	36
Report-SignificantEventChange	34	Report-MultimediaShare	127
Report-ServiceAvailable	15	Report-Factoid	140
Report-Official	52	Report-CleanUp	2
Report-Hashtags	4	Other-PastNews	12
Other-ContinuingNews	250	Other-Advice	39
Other-Sentiment	132	Other-Discussion	51
Other-Irrelevant	163	Other-Unknown	26
Other-KnownAlready	112		

*These authors are equally contributed to this work.

the actual accuracy.

- To overcome the small size of the training data, we use a knowledge base (KB) or meta-information such as timestamp to expand data.

- We use only Tweets provided as training/development data: we do not aggregate Tweets for training. This is because we have only a few pieces of data, so we cannot evaluate the effects of the aggregated data precisely.

We developed two models for this task: a KB-based model and a model that considers meta-information. We also use multi-layer perceptron (MLP)- and support vector machine (SVM)-based bag-of-words (BoW) baseline models for comparison. We describe our models in detail in this paper.

2 Related Work

There are many studies on classifying Tweets by information type. Toriumi and Baba (2016) focus on retweets – one important user behavior – to classify Tweets that are related to disasters into information types. Stowe et al. (2016) propose methods that use meta-information – timestamps, whether a tweet is a retweet or not, and so on – to classify disaster-related Tweets into information types. Kanouchi et al. (2015) classify Tweets according to the people who are mentioned in the Tweet by using meta-information in addition to bag-of-words as input features.

Also, many methods for extracting and identifying Tweets for certain tasks are reported. Vosecky et al. (2013) propose a novel multi-faceted topic model for discovering topics on Twitter. Hayashi et al. (2015) use streaming NMF (non-negative matrix factorization) with filter for "hijacking topics," which are pseudo-topics caused by advertisements and automatic messages, to detect topics. Li et al. (2018) use a naive Bayes classifier with an iterative self-training strategy to learn from unlabelled data and extracts disaster-related Tweets. Caragea et al. (2016) adopt a convolutional neural network (CNN) to identify informative tweets during disaster events.

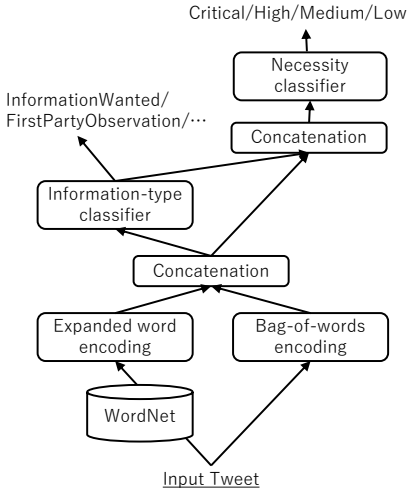

Figure 1: Overall architecture of our KB-based model.

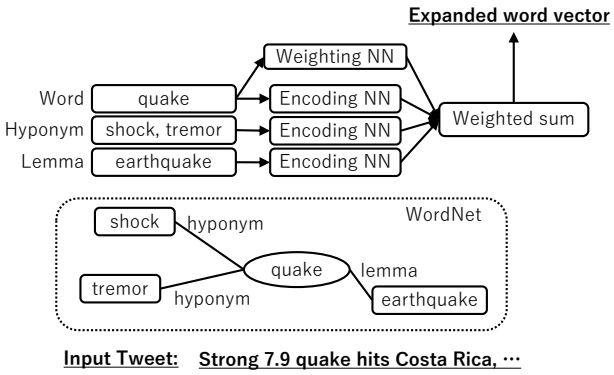

Figure 2: Word expansion using WordNet..

3 Models

3.1 KB-based model (run1)

The KB-based model is based on the model proposed in (Miyazaki et al., 2018), which is inspired by relational graph convolutional networks (R-GCN) (Schlichtkrull et al., 2018). An overview of the model is given in Figure 1. The model expands each word in a Tweet using WordNet (Miller, 1995) as a KB to encode texts (Figure 2). Then, the encoded vector is fed into a feed-forward neural network to classify the information type and give a necessity score. We give details on the methods used to do this below.

We use the following notation to describe the methods in this section; E is a set of entry words for a KB, $\mathcal{R}$ is a set of relation in the KB, T is a set of terms in the data set, and d_{KB} and d_{BoW} are the size of the dimensions for KB- and BoW-based embedding respectively.

Text encoding Consider a Tweet containing n entry words that mentions $e_1, e_2, ..., e_n$, each of which is contained in a KB, $e_i \in E$. The vector $m_{e_i r} \in \mathbb{1}^{|d_{KB}|}$ represents the entry word e_i based on the set of other entities connected through directed relation r:

$$m_{e_i r} = \sum_{e' \in \mathcal{N}_r(e_i)} W^{(1)}_{e'}, \quad (1)$$

where, $W^{(1)}_{e'} \in \mathbb{1}^{d_{KB}}$ is an embedding of entry word e' from embedding matrix $W^{(1)} \in \mathbb{R}^{|E| \times d_{KB}}$, and $\mathcal{N}_r(e)$ is the neighborhood function, which returns all nodes e' connected to e by directed relation r.

Then, $m_{e_i r}$ for all r are transformed by using a weighted sum:

$$\begin{aligned} v_{e_i} &= \sum_{r \in \mathcal{R}} a_{ir} \, \mathrm{ReLU}(m_{e_i r}) \\ a_i &= \sigma(W^{(2)} \cdot \vec{e}_i), \end{aligned} \quad (2)$$

where, $a_i \in \mathbb{1}^{|\mathcal{R}|}$ is the attention that entry word e_i represented by one-hot vector $\vec{e}_i$ pays to all relations using weight matrix $W^{(2)} \in \mathbb{R}^{|E| \times |\mathcal{R}|}$, and σ and ReLU are sigmoid and the rectified linear unit activation functions, respectively. Here, we obtain embedded vector v_{e_i} for entry word e_i.

Since the number of entry words in Tweets is sparse, we also encode, and use all the terms in Tweets regardless of if they are entry words or not. We represent each term by:

$$v_{w_j} = W^{(3)} \cdot \vec{w}_j, \quad (3)$$

where $\vec{w}_j$ is a one-hot vector of size $|T|$ where the value j represents the frequency of w_j in a Tweet, and $W^{(3)} \in \mathbb{R}^{|T| \times d_{BoW}}$ is a weight matrix.

Overall, a Tweet representing vector v is obtained by concatenating mean vectors of KB- and BoW-based encoding:

$$v = \left[\frac{1}{n} \sum_{i=1}^{n} v_{e_i}, \frac{1}{m} \sum_{j=1}^{m} v_{w_j} \right], \quad (4)$$

where m is the number of words that a Tweet includes.

The model is almost the same as that of (Miyazaki et al., 2018), but we do not share the weight matrix for KB- and BoW-based encoding because $|T|$ is too small, so if the weight matrix is shared, the effect of BoW embedding may be too small. Also, we concatenate KB- and BoW-based encoding vectors instead of adding them together. This is because the dimensions of input for KB- and BoW-based encoding is far different, so the embedding dimensions (d_{KB} and d_{BoW}) should be different. Therefore, we cannot add these vectors.

Classifying To estimate the information type of a given Tweet, we use 1-layer feed-forward neural network with classification output layers:

$$o = \mathrm{softmax}\, W^{(4)} v, \quad (5)$$

where $W^{(4)} \in \mathbb{R}^{class \times d_{KB} + d_{BoW}}$ is a weight matrix.

Then, an importance score is also obtained as:

$$\begin{aligned} h &= \mathrm{softmax}\, W^{(5)} [o, v] \\ score &= h_0 \times 1.0 + h_1 \times 0.75 \\ &\quad + h_2 \times 0.5 + h_3 \times 0.25, \end{aligned} \quad (6)$$

where $W^{(5)} \in \mathbb{R}^{4 \times d_{KB} + d_{BoW} + class}$ is a weight matrix. The importance is classified into one of four classes in the training data, "Critical," "High," "Medium," and "Low." Therefore, we use a weighted sum by using the classification score h as the weight, and obtain an importance score.

3.2 Meta-information considering model (run3)

The model that considers meta-information is based on a simple MLP model. An overview of the model is given in Figure 3. In addition to texts, this model uses date/time categories and event type information as the input of MLP. It encodes each input to each vector. Then, encoded vectors are concatenated and fed into a feed-forward neural network to classify the information type, and give a necessity score. We give details of the method below. d_{BoW} and d_{Meta} are the size of the dimensions for text BoW- and meta-information-based embedding.

Each input encoding Tweets are arranged in chronological order on the basis of the time at which they were created. The frequency and cumulative distribution function (CDF) of Tweets regarding elapsed time from the first Tweet of each event is shown in Figure 4. We divide Tweets into three classes along with their time difference from an event that has occurred so that each class has the same number of Tweets for each event. Then, those classes are date/time categories. The number

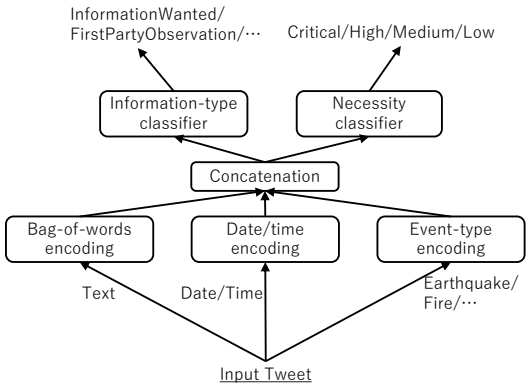

Figure 3: Overall architecture of our model that considers meta-information

of event types is six, as shown in Figure 1. $\vec{d} \in \mathbb{1}^3$ and $\vec{t} \in \mathbb{1}^6$ are one-hot vectors for date/time categories and event types respectively. We represent each term by:

$$v_w = W^{(6)} \cdot \sum_{j=1}^{m} \vec{w}_j + b^{(6)}, \quad (7)$$

$$v_d = W^{(7)} \cdot \vec{d}, \quad (8)$$

$$v_t = W^{(8)} \cdot \vec{t}, \quad (9)$$

where $W^{(6)} \in \mathbb{R}^{|T| \times d_{BoW}}$, $W^{(7)} \in \mathbb{R}^{3 \times d_{Meta}}$, and $W^{(8)} \in \mathbb{R}^{6 \times d_{Meta}}$ are weight matrices and $b^{(7)} \in \mathbb{1}^{d_{BoW}}$ is a bias.

Overall, a Tweet representing vector $\vec{v}_{all}$ is obtained by concatenating vectors of text, date/time, and event type encoding:

$$v_{all} = \mathrm{ReLU}([v_w, v_d, v_t]). \quad (10)$$

Classifying To estimate the information type and importance score of a given Tweet, we use each 1-layer feed-forward neural network with a classification output layer:

$$o = \mathrm{softmax}\, W^* \cdot v_{all} + b^*, \quad (11)$$

where $W^* \in \mathbb{R}^{k \times (d_{BoW} + 2 \cdot d_{Meta})}$ is a weight matrix, and $b^* \in \mathbb{1}^k$ is a bias, for which k is a $class$ for information types and 4 is for the importance score. Then, the importance score is calculated the same as eq. (6).

3.3 MLP- and SVM-based baseline models (run2 and run4)

The MLP-based baseline model is a simple MLP model. The model uses eq. (11) by inputting text BoW vector v_w as v_{all}, where $W^* \in \mathbb{R}^{k \times d_{BoW}}$ is a weight matrix.

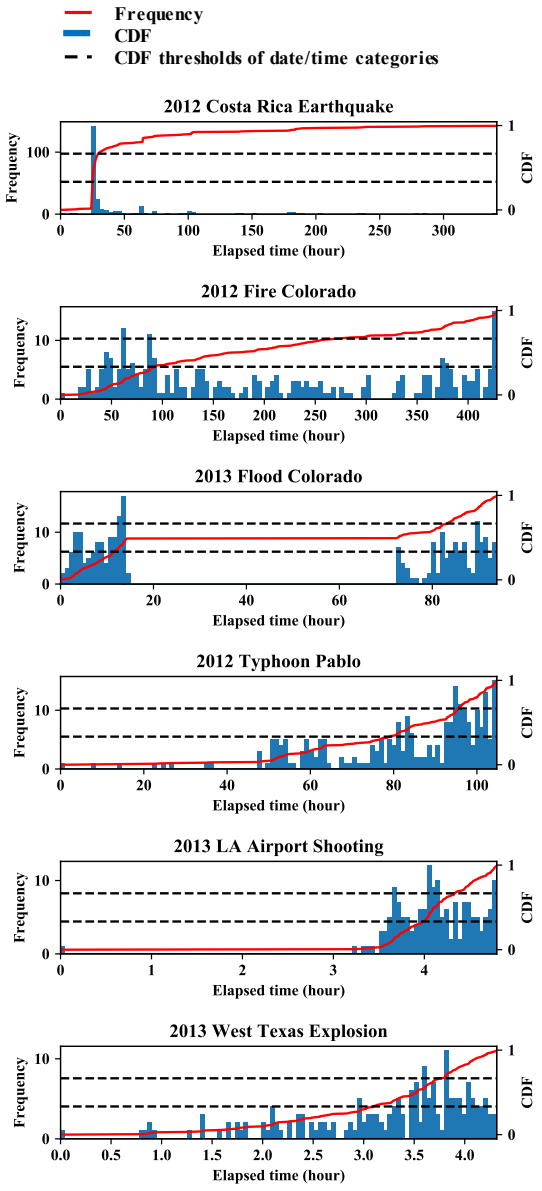

Figure 4: Frequency of Tweets regarding elapsed time from first Tweet of each event

The SVM baseline model is inputted with concatenated vector $[v_w, v_d, v_t]$, and it uses a linear kernel.

4 Experiments

4.1 Data set and settings

Our experiments were based on the data set provided for the TREC 2018 IS track, which was a Twitter data set with original json data of Tweet (including text, user information, timestamp, and so on) related to incidents. The data set contained approximately 2,000 Tweets for training/development and more than 20,000 Tweets for testing. Each Tweet in the training/development

Table 3: Relations that we used for KB-based model.

Lemmas, Hypernyms, Hyponyms, PartMeronyms, SubstanceMeronyms, MemberHolonyms, Entailments

set had labels that indicate the incident name (as shown in Table 1), information type (as shown in Table 2), information importance ("Critical," "High," "Medium," and "Low"), and indicator terms that human annotators selected when choosing an information type. The models were trained with 10-fold cross validation to find the best setting and all models were used as ensemble models for test data. We excluded words appeared fewer than five times in training sets.

All neural network-based models were learned with the Adam optimizer (Kingma and Ba, 2014), based on categorical cross-entropy loss, and models were implemented in Chainer (Tokui et al., 2015).

The SVM-based model was implemented by using scikit-learn (Pedregosa et al., 2011).

The hyperparameters used were as follows.

KB-based model The minibatch size was 10; the hidden layer size for KB-based encoding was 1,500, for BoW-based encoding is 500, classifier for information type is 500, and classifier for importance score is 250; There were 100 training iterations, with early stopping based on development performance; WordNet 3.1 (Miller, 1995) with the nltk toolkit (Bird and Loper, 2004) was used as the KB; The relations shown in Table 3 were used.

We used channel weights $W_c = \frac{|c_{max}|}{|c|}$, where $|c|$ is the number of information types c appearing in the training data, and $|c_{max}|$ is that of the most-frequent class, for calculating losses in model training.

Meta-inforation considering model The hidden layer size for text encoding d_{BoW} was 200, and the other's encoding d_{Meta} was 10. The hidden layer size of the classifier was 200. One hundred training iterations, with early stopping based on development performance, were used.

MLP- and SVM-based baseline model All of the hidden layer sizes for the MLP-based baseline model were 200; One hundred training iterations, with early stopping based on development performance, were used. The LinearSVC module was used for the SVM-based baseline model.

Table 4: Results of the classifying by information type using training data as 10-fold cross validation.

Model	Micro F1	Macro F1
KB-based	0.557	0.328
Meta-information considering	0.598	0.369
MLP-based baseline	0.597	0.375
SVM-based baseline	0.546	0.304

4.2 Results

Table 4 presents the results for our models. Each scores in the table is the mean average of each of the 10-fold cross validations using trainig data. We can see that Meta-information considering model is the bast result in the micro F1 score, and MLP-based baseline is the best in the macro F1 score.

Table 5 shows the results using test data. This is the official results of TREC 2018 IS track. Values in the brackets shows the rank in the all methods submitted to the track[1] . The target metric of the main task of the track is the macro F1 score, and that of the sub task is the information priority. Information priority is measured with the mean squared error between the output and the gold data that obtained by human annotators, so the lower is the better.

In the table, MLP-based baseline model is better in both micro and macro F1 scores in our methods. Our KB-based method achieved the best result in the sub task of the track.

4.3 Discussion

Our Meta-information considering model and MLP-based baseline model achieve rather better scores in both Micro and Macro F1 scores. This is because we have only a small training data, so it is better to have parameters need to be learned. Meta-information considering model and MLP-based baseline model have rather smaller number of parameters, so these methods fit for the task.

On the other hand, KB-based model has rather smaller differences between the two results – using training data and test data. As we mentioned, we can use only a small data for training, so test data includes many OOVs (see Table 6). This is one of the big reasons of that all our models drop the F1 scores when using test data for evaluation. Our KB-based model can consider OOVs, so the effect of OOVs is rather small.

[1] TREC 2018 IS track accepts 39 methods from 12 research groups, so the rank has the range of 1 to 39,

Table 5: Results of the classifying by information type using test data.

Model	Micro F1	Macro F1	Information Priority (Lower is better)
KB-based	0.120	0.497 (13)	0.060 (1)
Meta-information considering	0.114	0.542 (7)	0.061 (2)
MLP-based baseline	0.119	0.551 (5)	0.062 (3)
SVM-based baseline	0.088	0.465 (21)	0.066 (4)
TREC Median	0.083	0.478	0.093

Table 6: Comparing vocabulary size of data sets. Training (all) means all vocabulary, and training (5+) means vocabuary of words that appeared more than 5 times in training data.

	Vocabulary size	# of OOVs
Test	43,117	–
Training (all)	4,476	40,413
Training (5+)	619	42,540

Meta-information considering model is the best in Micro F1 score in the experiment using training data, but the third place in that using test data. We used timestamp information as Meta-information, which is affected by the physical distance between the incident occurred and the Tweet posted. This is because if the distance is far, the Twitter user only can know about the incident from TV, Web news and some other sources. Therefore, the post may be made very after the incident occurred. Incidents included in test set are occurred in various places including non-English countries/regions. Tweets included in data set are written in English, so most of Tweet for these incidents may delay. This may make our Meta-information considering model work with limited improvement for test set.

Our models achieved the best results in TREC 2018 IS track in information priority estimation sub task. We regard estimating information priority as one of the tasks of multi-task learning, which may work well in this task.

5 Conclusion and Future Work

In this paper, we described models for classifing incident-related Tweets along with information type. We used four models, a KB-based model, model considers meta-information, and MLP- and SVM-based baseline models. We showed that our models – the KB-based model and model considers Meta-information – outperformed baseline methods.

Using our models in combination is left as our future work.

Acknowledgements The KB-based method is based on the method that the first author studied when visiting the University of Melbourne. The authors greatly appreciate Professor Timothy Baldwin, Associate Professor Trevor Cohn, and Dr. Afshin Rahimi for their useful advices.

References

Zahra Ashktorab, Christopher Brown, Manojit Nandi, and Aron Culotta. 2014. Tweedr: Mining twitter to inform disaster response. In *ISCRAM*.

Steven Bird and Edward Loper. 2004. Nltk: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 31. Association for Computational Linguistics.

Cornelia Caragea, Adrian Silvescu, and Andrea H Tapia. 2016. Identifying informative messages in disaster events using convolutional neural networks. In *International Conference on Information Systems for Crisis Response and Management*, pages 137–147.

Jun Goto, Taro Miyazaki, Yuka Takei, and Kiminobu Makino. 2018. Automatic tweet detection based on data specified through news production. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion, Tokyo, Japan, March 07-11, 2018*, pages 1:1–1:2.

Kohei Hayashi, Takanori Maehara, Masashi Toyoda, and Ken-ichi Kawarabayashi. 2015. Real-time top-r topic detection on twitter with topic hijack filtering. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 417–426. ACM.

Shin Kanouchi, Mamoru Komachi, Naoaki Okazaki, Eiji Aramaki, and Hiroshi Ishikawa. 2015. Who caught a cold?-identifying the subject of a symptom. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1660–1670.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Hongmin Li, Doina Caragea, Cornelia Caragea, and Nic Herndon. 2018. Disaster response aided by tweet classification with a domain adaptation approach. *Journal of Contingencies and Crisis Management*, 26(1):16–27.

Kiminobu Makino, Yuka Takei, Taro Miyazaki, and Jun Goto. 2018. Classification of tweets about reported events using neural networks. In *Proceedings of the 4th Workshop on Noisy User-generated Text (W-NUT)*, pages 153–163.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Taro Miyazaki, Afshin Rahimi, Cohn Trevor, and Timothy Baldwin. 2018. Twitter geolocation using knowledge-based methods. In *Proceedings of the 4th Workshop on Noisy User-generated Text (W-NUT)*, pages 7–16.

Junta Mizuno, Masahiro Tanaka, Kiyonori Ohtake, Jong-Hoon Oh, Julien Kloetzer, Chikara Hashimoto, and Kentaro Torisawa. 2016. Wisdom x, disaana and d-summ: Large-scale nlp systems for analyzing textual big data. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 263–267.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*, pages 593–607. Springer.

Kevin Stowe, Michael J Paul, Martha Palmer, Leysia Palen, and Kenneth Anderson. 2016. Identifying and categorizing disaster-related tweets. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, pages 1–6.

Yuka Takei, Taro Miyazaki, Ichiro Yamada, and Jun Goto. 2017. Tweet extraction for news production considering unreality. In *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*, pages 370–375. The National University (Phillippines).

Seiya Tokui, Kenta Oono, Shohei Hido, and Justin Clayton. 2015. Chainer: a next-generation open source framework for deep learning. In *Proceedings of workshop on machine learning systems (LearningSys) in the twenty-ninth annual conference on neural information processing systems (NIPS)*, volume 5, pages 1–6.

Fujio Toriumi and Seigo Baba. 2016. Real-time tweet classification in disaster situation. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 117–118. International World Wide Web Conferences Steering Committee.

Jan Vosecky, Di Jiang, Kenneth Wai-Ting Leung, and Wilfred Ng. 2013. Dynamic multi-faceted topic discovery in twitter. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 879–884. ACM.