Tweet Extraction for News Production Considering Unreality

Yuka Takei, Taro Miyazaki, Ichiro Yamada, Jun Goto
NHK Science & Technology Research Laboratories
1-10-11 Kinuta, Setagaya-ku, Tokyo, Japan
{takei.y-ek, miyazaki.t-jw, yamada.i-hy, goto.j-fw}@nhk.or.jp

Abstract

Acquiring information on incidents and accidents from social media can be useful for broadcasters to report news faster. However, many tweets including words related to incidents and accidents are actually irrelevant to real events, for example, "Backdraft's explosion scene was impressive!!!" Social media contains many comments on events in unreal worlds such as movies, animations and dramas, and it is time-consuming to discriminate these tweets manually. This work presents a method for automatically extracting useful tweets for news reports by focusing on "unreal" information. We first prepare unreal tweets as learning data and use a distributed representation and features that can determine if a tweet is real or unreal. By adding the features of a neural network, we generate a learning model that can effectively discriminate whether a tweet includes information on actual incidents or accidents. Results of evaluations revealed that the proposed method achieved a 3.8-point higher F-measure than the baseline method.

1. Introduction

Social networking services (SNSs) enable us to easily transmit information anywhere in real-time. The large amount of information transmitted on SNS, known as "Social Big Data," is a valuable information source for grasping newsworthy occurrence (Vieweg et al., 2010; Kanouchi et al., 2015), and broadcasters monitor social media such as Twitter to collect information about incidents and accidents. By obtaining information directly from witnesses of such events, broadcasters can report news more quickly and effectively. They use various tools to manually search for tweets that have potential news value, using keywords to find tweets indicating incidents and accidents. However, a lot of effort is required to find valuable information from among the large number of tweets sent every day.

Methods have been reported for automatically extracting tweets with potential news value by using machine learning (Freitas et al., 2016; Mizuno et al., 2016; Doggett et al., 2016). However, many tweets irrelevant to actual incidents or accidents include relevant words, which worsen the extraction results. Examples include tweets about events in current animations and TV programs, such as "ドラえもん「のび太の家火事になる・前編」(Doraemon - Nobita's House Fire・Part 1)." Many viewers tweet while watching TV to share their opinions with other people. Therefore, many tweets include names of animations (which we call "virtual proper nouns") and TV programs. In addition, words in Japanese idioms could also suggest incidents or accidents, such as "火の無いところに煙は立たない (Where there's smoke, there's fire)." Furthermore, there are tweets that include hypothetical expressions that assume an incident or accident occurring, such as "火事になったら、どこに逃げるべきだろう (If a fire occurs, where should I escape to?)."

All three tweets include the word "fire" but do not indicate the occurrence of a real fire. The conventional method extracts information from tweets that include words related to incidents and accidents, regardless of whether one has actually occurred. Therefore, to utilize the extracted tweet as a news source, more work is required to determine

whether it is a "real event" or "unreal event." In this paper, virtual proper nouns (movies and animations), TV program titles, and idiomatic phrases are defined as "characteristic phrases." In addition, phrases including expressions of hypothesized situations are defined as "hypothesis expressions." By adding the presence or absence of "characteristic phrases" and "hypothesis expressions" to the input of a neural network as features, we generate a learning model that can efficiently discriminate whether a tweet includes information on actual incidents or accidents. Extending the input dimension like this, improved the F-measure by 3.8 points, revealing the effectiveness of the proposed method.

2. Related Work

During large-scale disasters, such as the 2011 Great East Japan Earthquake, SNSs such as Twitter are effective for transmitting information (Aida et al., 2012). On the basis of information on SNSs, public officials and emergency workers can grasp what is happening in the disaster area in real-time. However, on Twitter, unreliable and unnecessary information is also diffused excessively, requiring more effort to discriminate relevant information.

To extract relevant information during a disaster, Neubig et al. developed a semiautomatic information extraction method (Neubig et al., 2011, 2013) that efficiently filters information by using active learning. In the process of active learning, an annotator labels each tweet presented by the system as positive or negative. Conventional active learning labels sequentially from the tweets near the boundary of positive and negative samples. On the other hand, their method presents tweets that have the highest possibility of being positive samples, making it possible to minimize the number of negative samples labeled by annotators and improving work efficiency. However, when large-scale incidents or accidents occur, secondary tweets such as retweets often occur, so the absolute number of tweets judged to be positive samples increases. Continually presenting tweets with high scores as positive samples will increase the accuracy, but less information will be covered, causing tweets judged to be positive samples to be overlooked.

Broadcasters must acquire a wide variety of information, not only information about large-scale disasters. By limiting negative samples to the minimum, we aim to improve information gathering efficiency, and by maintaining the diversity of the positive samples, we reduce the risk of information being missed.

3. Methodology

In this section, we describe a method for extracting tweets for news reporting. In the proposed method, we generate a model that learns by focusing on unreal negative samples. We use a feed forward neural network as a learning algorithm to automatically extract tweets that have potential news value. The input to the neural network uses the distributed representation of tweets. By adding a feature of whether a characteristic phrase or hypothesis expression is included in a tweet, learning models are generated. The configuration of the neural network is shown in Figure. 1.

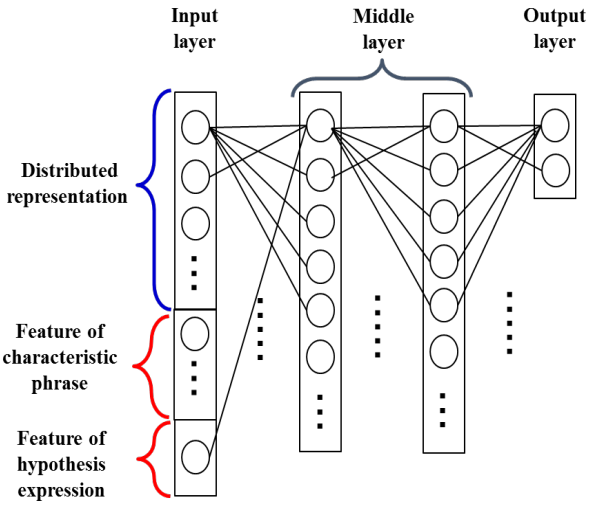

Figure 1. Configuration of neural network

In this paper, our target for news production is to extract tweets related to "fire," which is the most frequently occurring topic in Japanese news.

3.1 Features based on Distributed Representation

First, a tweeted sentence is divided into morpheme units using the morphological analyzer MeCab (Kudo et al., 2004). Then, by using Word2Vec (Mikolov et al., 2013), each unit is converted into a 200-dimensional distributed representation. The

average of all vectors for words included in a sentence is regarded as the sentence vector and is used for an input for a neural network. We used the Wikipedia dump data of September 2016 to generate distributed expressions using Word2Vec.

3.2 Features of Characteristic Phrases

As described in the introduction, information about broadcast content such as dramas and animations is often sent to SNSs, and some tweets includes idiomatic phrases. We prepared three kinds of characteristic phrases: TV program names, virtual proper nouns, and idiomatic phrases. If a tweet includes them, we put "1" in the corresponding dimension of the phrase and "0" if not.

TV program names
We gathered 9,473 titles, mainly of dramas, using the program guide application programming interface (API) of broadcasting stations and Wikipedia.

Virtual proper nouns
12,310 proper nouns such as animation, movie, and video game titles were gathered from Wikipedia.

Idiomatic phrases
We gathered 32 phrases that contained "fire" from published dictionaries[1].

In characteristic phrases, we exclude titles that contained common verbs or adjectives such as "生きる (live)" and single-character titles such as "江 (Gou)."

Table 1. Examples of characteristic phrases

Feature type	Example
TV program names	ひよっこ (Hiyokko), べっぴんさん (Beppinnsann), あさイチ(Asaichi)
Virtual proper nouns	スーパーマン (Superman), マリオパーティ (Mario-Party), スラムダンク (Slam-Dunk)
Idiomatic phrases	対岸の火事 (taiganno-kazi), 火事場の馬鹿力 (kajibano-bakadikara)

The features of characteristic phrases are set as follows. The example sentence "海外の事例を対岸の火事と楽観視できない (Foreign cases cannot be optimistic about the fire on the other side,)" includes the Japanese idiomatic phrase "対岸の火事 (the fire on the other side)." The idiomatic phrases dimension corresponding to it is "1." Since the sentence does not include any TV program names or virtual proper nouns, their values are set to "0."

3.3 Features of Hypothesis Expressions

Due to the effect of recent news of terrorism overseas, tweets expressing worries about terrorism have been posted such as "近くで爆発が起きたら怖いな (If an explosion occurs nearby, I'll be scared)" We extract this kind of assumption from a sentence and use it as a feature for tweet extraction. We analyze the relationship between words that include expressions related to fire such as "爆発 (explosion)" and include assumptions such as "たら (if)." A tweeted sentence is divided into clauses by using the parser CaboCha (Kudo and Matsumoto, 2002). If the tweet includes (1) or (2), it is determined to include a "hypothesis expression."

(1) A dependency relationship between an expression related to fire and an assumption

(2) An Expression related to fire and an assumption in the same clause

In the above example, since "if" has a dependency relationship with "explosion," the feature of the "hypothesis expressions" is set to "1." Specific examples are shown in Figure. 2.

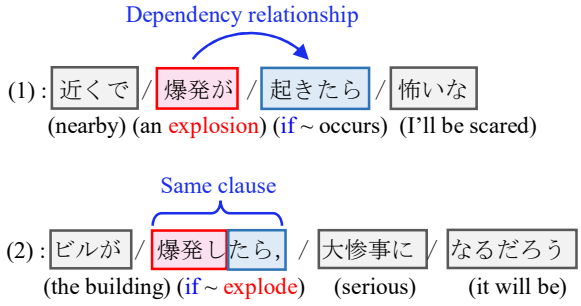

Figure 2. Examples of hypothesis expressions

4. Evaluation Experiment

We conducted two experiments to evaluate the effectiveness of our method. The first was to evaluate the effect of learning data using tweets that include "characteristic phrases." The second is to evaluate the effect of the features including "characteristic phrases" and "hypothesis expressions."

[1] http://www.jlogos.com/

4.1 Experimental Settings

Dataset

For the training data, we gathered 5,065 tweets used in actual news reports as positive samples, which included information related to "fire" from March 2014 to August 2015. For comparison, we prepared two kinds of negative samples.

(A) A random sample of 5,065 tweets randomly selected from all tweets in September 2016. This random sample did not include news source.

(B) A mixed sample of 5,065 tweets randomly selected from a dataset that mixed tweets in (A) and tweets including characteristic phrases.

The evaluation data were narrowed down to 8,154 tweets from about 7,700,000 from October 23rd, 2016. These were selected by keyword matching concerning fire-based events. The keywords are devised by the news production section of our broadcasting station[2]. There are 61 keywords related to fire, and broadcasters combine them to search for newsworthy information. Then a positive sample label was given to tweets with content related to actual fires or explosions, and a negative sample label was given to tweets with content not related to fire. For example, if a fire is happening in an unreal world or someone's imagination, this tweet is a negative sample. All the tweets are annotated by one annotator.

Implementation

We use Chainer (Tokui et al., 2015) to implement our method. The input layer uses 204 dimensions (1 to 200 dimensions indicate the distributed representation and 201 to 204 dimensions respectively indicate presence or absence of TV program names / virtual proper nouns / idiomatic phrases / hypothesis expressions). The output layer is two-dimensional, and the middle layer has two layers. The middle layers contain 500 nodes and 250 nodes from the nearest to the input layer. In addition, exponential linear units (ELUs) (Clevert et al., 2015) were used as an activation function, and batch normalization was performed in each layer. The number of learning sessions was set to 30.

[2]NHK (Japan Broadcasting Corporation) has a social media analysis team, that looks for news on the internet.

4.2 Experimental Results

Comparison of training data

The experimental results of the training data are shown in Table 2. The random sample uses negative samples (A) of the learning data as the baseline. The mixed sample uses negative samples (B) of the learning data.

Table 2. Experimental results for each training dataset

Method	Recall	Precision	F-measure
Random sample	84.1	79.7	81.9
Mixed sample (MS)	85.4	83.4	84.4

Comparing the training data, the mixed sample including the characteristic phrases performs better than the random sample. Therefore, we used the mixed sample as training data in the next experiment and experimental results with various features.

Effects of features

Table 3 shows the experimental results for using each feature. Mixed sample (MS) is the method that learned only distributed representation as described in Section 3.1. We added the features of TV program names, virtual proper nouns, and idiomatic phrases described in Section 3.2. Characteristic (1d) indicates the results of summarizing three features expressing characteristic phrases into one dimension, and Characteristic (3d) indicates the results of simultaneously adding three features to different dimensions. Furthermore, as a result of adding hypothesis expressions described in Section 3.3 as a feature to the MS method, the results obtained by adding all the features are shown.

Table 3. Experimental results for each method

Method	Recall	Precision	F-measure
Mixed sample (MS)	85.4	83.4	84.4
MS + TV program names	84.5	84.6	84.6
MS + Virtual proper nouns	**89.7**	80.1	84.7
MS + idiomatic phrases	83.9	**85.2**	84.5
MS + Characteristic (1d)	82.7	83.7	83.2
MS + Characteristic (3d)	88.9	82.8	**85.7**
MS + Hypothesis Expression (HE)	82.5	84.4	83.4
MS + All feature (3d+HE)	83.5	84.4	84.0

Among the three types of features of characteristic phrases, using virtual proper nouns achieves the highest F-measure. Performance was improved more by dividing each feature into three dimensions rather than putting each feature together. In addition, even when the features of hypothesis expressions were added, the F-measure did not improve.

4.3 Discussion

Training data

As a negative sample of the training data, the mixed sample that included characteristic phrases performed better than the random sample. By including these mixed tweets, our method can learn negative samples including news-related words precisely. It can also learn combinations of news-related words and other words. Therefore, a characteristic phrase is a clue to select effective training data from among a large number of tweets.

Effects of features

The results of adding features of characteristic phrases to different dimensions (3d) is better than those of other methods. The proposed method has a 1.3-point higher F-measure and 3.5-point higher recall than the MS method. This result shows that we can acquire many positive samples as well as excluding tweets about unreal worlds. Examples of improvements by the proposed method are shown in case-A and case-B.

Case-A

MS method: Positive → proposed method: Negative

「火の鳥」の最終回が炎上
(The last round of "Fire Bird" is flaming.)

Case-B

MS method: Negative → proposed method: Positive

せっかく特急乗ったのに, 沿線火災で電車が
止まっている
(Even though I got on a limited express, the train stopped due to a fire along the railroad.)

In Case-A, words related to fires such as "fire" and "flame" were included, so the MS method judged it as a positive. However, "Fire Bird" is the name of Japanese animation. Therefore, by adding proposed features, the proposed method can judge it as negative.

The MS method sometimes judged tweets including words related to fire as negative such as Case-B, because the method learned mixed sample including characteristic phrase without adding proposed features. For example, the method learned tweet including phrases related to fire like a "対岸の火事 (the fire on the other side)" as a negative example. Thus, words related to fire are included in negative examples as well as positive examples. When the features were added, the positive and negative criteria were clarified. Therefore, our proposed method can maintain the diversity of the positive samples. In addition, features of characteristic phrases were improved more by dividing each feature into three dimensions rather than using each feature as one dimensions. By dividing TV program names, virtual proper nouns, and idiomatic phrases into features, the proposed method can learn patterns of notation when phrases appear in sentences.

The features of hypothesis expressions could not improve the F-measure because recall decreased. From results of error analysis, our method judged positive samples in the evaluation data as negative. For example, the negative results included tweets attributing causality to fire such as "たくさんの煙が見える, 火事だったらこまるなぁ (There is a lot of smoke over there, I'm in trouble if it's a fire)". In order not to miss such a tweet expressing the possibility of an incident or accident, a detailed analysis method needs to be developed to analyze causality.

5. Conclusion

In this paper, we presented a method to automatically extract tweets with potential news value by adding new features focusing on "unreal" events to a neural network. The proposed method achieved a highest F-measure of 85.7, a 3.5-point increase over the baseline method, by focusing on "characteristic phrases" (TV program names, virtual proper nouns, and idiomatic phrases). This method is expected to reduce the workload of broadcasters who acquire information from social media.

In the future, we aim to further improve the performance by acquiring more characteristic phrases such as "cast of a TV program" and "TV program-related information" from real-time data.

References

Shin Aida, Yasutaka Shindoh, and Masao Utiyama. 2013. Rescue Activity for the Great East Japan Earthquake Based on a Website that Extracts Rescue Requests from the Net. *Proceedings of the Workshop on Language Processing and Crisis Information 2013*, pages 19-25.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Ochreiter. 2015. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). arXiv:1511.07289.

Erika Doggett and Alejandro Cantarero. 2016. Identifying Eyewitness News-Worthy Events on Twitter. *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, pages 7-13.

Jesse Freitas and Heng Ji, 2016. Identifying News from Tweets. *Proceedings of 2016 EMNLP Workshop on Natural Language Processing and Computational Social Science*, pages 11-16.

Shin Kanouchi, Mamoru Komachi, Naoaki Okazaki, Eiji Aramaki, and Hiroshi Ishikawa. 2015. Who caught a cold? — Identifying the subject of a symptom. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 1660-1670.

Taku Kudo and Yuji Matsumoto. 2002. Japanese dependency analysis using cascaded chunking. *In Proceedings of the 6th Conference on Natural Language Learning 2002*, pages 1-7.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying Conditional Random Fields to Japanese Morphological Analysis. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 230-237.

Tomas Mikolov, Kai Chen, Greg Corradoet, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv*:1301.3781.

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 the 26th International Conference on Computational Linguistics, pages 263–267.*

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

Graham Neubig, Shinsuke Mori, and Masahiro Mizukami. 2013. A Framework and Tool for Collaborative Extraction of Reliable Information. *In Proceedings of the Workshop on Language Processing and Crisis Information*, pages 26-35.

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*.

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*, pages 1079-1088.