Fuzzy Systems and Data Mining VI
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA200714

Keyword Extraction from TV Program Viewers' Tweet Based on Neural Embedding Model

Taiga KIRIHARA[a,1], Kazuyuki MATSUMOTO[b] , Minoru YOSHIDA[b], Kenji KITA[b]
[a] *Graduate Schools of Science and Technology for Innovation Division of Science and Technology, Tokushima University, Japan*
[b] *Graduate School of Technology, Industrial and Social Sciences, Tokushima University, Japan*

Abstract. In recent years, young people have not been watching television (TV) as much as they used to. This is mainly because a number of TV programs are very long and/or have limited viewing times. Recently, individuals have been actively posting live-action tweets on Twitter to comment on TV content while watching programs in real time. In this study, we propose a method for extracting key phrases related to the event scenes of TV programs using live tweets, and we propose a scene search system that aims at efficient TV program viewing. The experimental results indicated that the program contents were estimated with an error of approximately 5% to 10% with respect to the program time. In addition, the extracted key phrases were visualized for each event scene category using the t-SNE algorithm.

Keywords. Natural Language Processing, TV Program, Key Phrase Extraction, Twitter Analysis

1. Introduction

The number of people who watch television (TV) programs by recording them has increased in recent years due to the diversification of TV viewing methods. The main reasons in regard to why a person records TV programs are the following: preferring to watch at a favorite time, not wanting to be tied to the broadcast time, and wanting to use time effectively. Even if there is a TV program that an individual does want to watch, it may be difficult to watch in real time if the program is long or has limited available viewing times; therefore, the number of people who record TV programs to watch them at a later time is increasing.

With the spread of social networking sites (SNS) such as Twitter in recent years, individuals are actively posting live-text tweets, which are tweets sent in real time while the person watches TV programs. These live tweets contain information on the TV program's events and content as well as the poster's impressions and opinions on the program. When a person posts to Twitter in real time in regard to the TV program

[1] Graduate Schools of Science and Technology for Innovation Division of Science and Technology, Tokushima University. E-mail: c612035024@tokushima-u.ac.jp.

they are currently viewing, the title and the name of the performer are included in hash tags. We refer to this as the act of performing a live TV program on Twitter.
In this research, we propose an event scene retrieval method for supporting the viewing of TV programs by focusing on live TV program tweets on Twitter. We expected that the detection of a specific scene in a TV program could be automated by collecting tweets for a specific TV program and extracting information about the program in the tweet text. We propose a method for extracting key phrases related to event scenes of TV programs using tweets. In addition, we consider the possibility of scene retrieval by comparing the categorization of the scene with the extracted keywords.

2. Related research

2.1. Detection and labeling of significant scenes from a TV program based on Twitter analysis

In Nakazawa et al.'s [1] study, the researchers were able to automatically detect an important scene by collecting data on the fluctuation of the number of tweets about a TV program. The tweet contents were used to determine the main character involved and the event contents in the TV scene, and the result was taken as the scene. They proposed a method that included adding a label. In Shamma et al. [2], the researchers detected essential scenes, and TF-IDF was used to ascertain the main character in the important scene. These authors aimed to develop a system that is able to search all the scenes in a program to determine which ones are important scenes rather than focusing on the peak number of tweets.

2.2. A scene explorer for TV programs based on Twitter emotion analysis

In Yamauchi et al. [3], the authors collected data on tweeted sentences for specific TV programs from Twitter, analyzed the tweets, estimated the viewer's feelings on the recorded program, and visualized the data to search for a program. They proposed a support method. Yamauchi et al. used the emotion analysis method and adopted Plutchik's Wheel of Emotions[4]. In addition, the researchers identified the emotions for emoticons and are used as an index of emotions.

In our research, keywords were extracted based on the specific content of the event scene as described in the tweet texts, rather than using sensitizing information such as how the viewer felt.

3. Proposed method

The present study followed these steps: first, we collected data on the tweets about a TV program; next, the tweets were vectorized to extract the important expressions that directly illustrated the scene content of the program. We refer to the expression that represents the scene content of the program for each tweet as a key phrase, and we searched the event scenes based on these key phrases. In addition, we classified the scenes via categorizing the key phrases.

3.1. Tweet collection and time division

We used Twitter's API to collect data on the tweets. The hash tag ("#[program name]") was used as a specific search keyword when collecting the data. After collecting the tweets, the tweet set for each program was divided into one-minute units. Taking this step allowed us to search for programs in one-minute units. We separated by one-minute units based on this present study's purpose, which was to search TV program scenes.

Conducting these time-divided scene searches made it easier to compare errors with actual scenes and extracted key phrases. We used the one-minute mark since there were a sufficient number of tweets to perform.

3.2. Vectorization

We vectorized the tweet sentence sets that were divided into one-minute units using the bidirectional encoder representations from transformers (BERT) method[5]. BERT is a deep learning model for natural language processing that was first announced by Google in 2018. BERT has 12 stages of transformers and can be used to understand the context.

The BERT used in this study is a pre-learned model based on the Japanese Wikipedia article corpus[6]. We did not use the pre-learned model that is based on the SNS text because we expected that many of the important keywords that characterize the program scenes were Wikipedia-registered words. Wikipedia offers rich resources, which allowed us to collect accurate information, such as events that occur in TV programs and character names. On the other hand, such information may be missing in SNS-based models, increasing the possibility that incorrect information would be obtained. Therefore, this study used the Wikipedia-based BERT model.

In addition, BERT requires a large-scale, high-quality dataset to learn the distributed expressions. Since the tweet data collected in the present study are biased and small, this dataset was not suitable for fine tuning with BERT. Therefore, we did not train the BERT model using my own dataset as fine tuning, and we use a pre-trained BERT model.

3.3. Key phrase extraction

Regarding key phrase extraction, there were many sentences in the live-action tweets that included a "noun + adjective" part-of-speech pattern. For example, many collected tweets included the following: "Pikachu cute" and "Pennywise scary." There was such a thing. We argue that such tweets characterize the TV program scene; therefore, we used the EmbedRank algorithm[7] to extract key phrases by combining proper nouns and adjectives. Among the nouns, proper nouns often characterize the TV program, while adjectives are closely related to the content of the scene. In addition, the adjectives can represent the emotions of the person.

EmbedRank is an algorithm that extracts important key phrases that do not require training data. In this study, we obtained a distributed expression vector for candidate phrases and entire sentences (a set of tweets per unit time) that were extracted based on a part-of-speech pattern. The important key phrases were decided based on the degree of similarity. The Japanese morphological analyzer MeCab[8] and the morphological analysis dictionary NEologd[9] were used for part-of-speech determination. The figure

below shows the following: "The giant Snorlax will appear in the next week's Anipoke!" (Raishu no Anipoke wa Kabigon kyodaika desu.). This flow is explained using a concrete example of this tweeted sentence. In this case, the proper nouns are "Anipoke" and "Snorlax" (Kabigon, with "Anipoke" indicating Pokémon anime. Fig.1 shows the flow of keyphrase extraction.

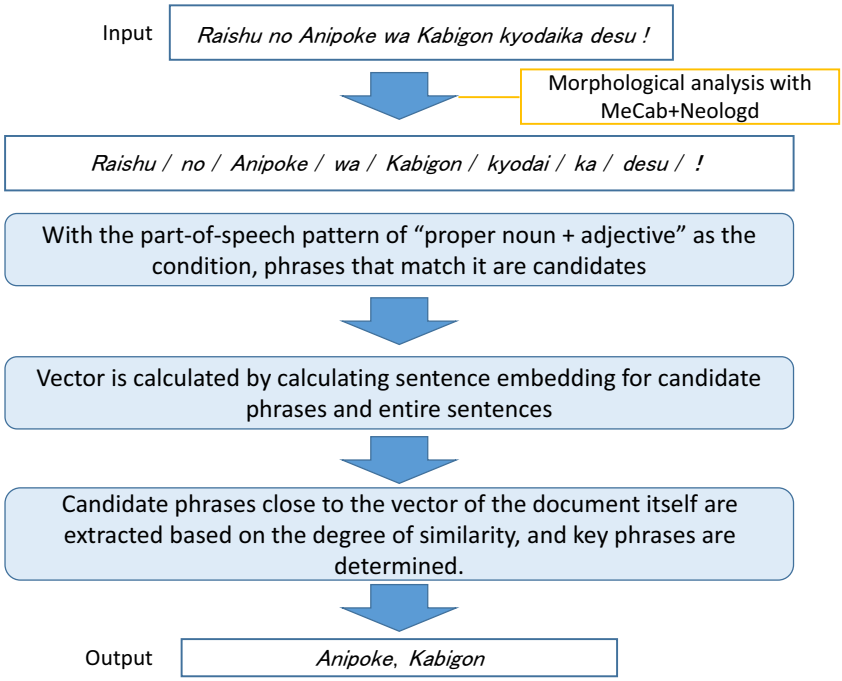

Figure 1. Flow of key phrase extraction

By simply extracting key phrases based on the degree of similarity, we were able to avoid including redundant expressions with similar meanings by calculating the maximal marginal relevance (MMR). Equation (1) shows the MMR calculation formula.

$$MMR \coloneqq arg_{D_i \in R \setminus S} \max \left[\lambda \cdot Sim_1(D_i, Q) - (1 - \lambda) max_{D_i \in S} Sim_2(D_i, D_j) \right] \quad (1)$$

Using the MMR equation allowed us to exclude expressions with similar meanings and extract more important key phrases. We assumed that this would improve the accuracy of scene retrieval.

3.4. Difference between this method and conventional method

Aside from the EmbedRank and MMR methods used in the present study, there are several other methods that can be used to extract important words. The TF-IDF has been used in previous research and is considered to be a general method. This method calculates a word's importance according to the frequency with which the word appears. The TF-IDF formula is shown as Equation (2).

$$tfidf(t_i, d_j) = tf(t_i, d_j) \cdot idf(t_i) = \frac{f(t_i, d_j)}{\sum_{t_k \in d_j} f(t_k, d_j)} \cdot \log\left(\frac{N}{df(t_i)} + 1\right) \quad (2)$$

The disadvantage of the TF-IDF method is that the term frequency (TF) value increases as the number of words in the document decrease (and vice versa). Since the total number of words in a document is calculated during the TF value estimation process, the difference in the number of words in each document affects the importance of words. As explained previously, we divided the tweet set every minute. When the set of tweets is divided by time, we can assume that the number of tweets in time increases in a scene that is live on a TV program, and that the number of tweets decreases during other parts of the program. In fact, even in the tweet data used in the present study, there was a difference in the number of tweets for each time period. When we used TF-IDF, we observed that the difference in tweets in each time zone had a large effect on the extraction of key phrases. Therefore, we extracted the key phrases using the BERT and EmbedRank algorithms and MMR.

3.5. Classification of key phrases by scene category

In this step, we manually classified the scenes into categories. We expected that using the relationship between key phrases and scene categories would lead to a rough visualization of similar event scenes. In this study, the following six categories were defined:

- People: Scenes in which people appear
- Action: A scene in which a person takes action
- Emotion: A scene in which a person shows emotion
- Summary: Rough description of a scene without humans
- Scenery: A scene where the location and time of day can be observed
- Conversion: A change of scene from the main program to other images (e.g., a commercial)

4. Evaluation experiment

In order to observe how the TV program genre may lead to differences in the results, we targeted the following program types: "variety," "animation," and "movie." In this experiment, we evaluated the scene search performance by searching the event scene based on the extracted key phrases and calculating the error from the actual time zone. In addition, we verified the effectiveness by comparing cases in which the MMR method was used for key phrase extraction to cases in which it was not used.

4.1. Data

In this section, we describe the collected data— the TV programs and the related tweets. Three programs in total were recorded— "Getsuyou Kara Yofukashi" (Yofukashi) that

was broadcast on the "Nihon TV" series on October 14, 2019, "Friday Road SHOW – It" (Friday) that was broadcast on the "Nihon TV" series on November 8, 2019, and "Pokémon" (Pokémon) that was broadcast on the "TV Tokyo" series on December 8, 2019. In addition, we collected live tweets posted by viewers in each time zone based on hashtags using the Twitter API. We collected a total of 37,804 tweets. All three programs were programs of the broadcasting station that was on the same net as the production station. This is because the event scene search uses the time slot in which the live tweet was posted as the search result. We assumed that it was necessary to post the tweet in the same time zone regardless of where (i.e., which region) the program was viewed.

For evaluation, we created a key with the correct answers for each program. Including multiple keywords for each scene, this data describes the time zone in which the event scene occurred, the category that the scene belongs to, the character who appeared in the event scene, what happened in the scene, etc. For this step, we used the keyword of the correct answer data as the search query. The time zone of the event scene was estimated based on the similarity to the key phrase that was automatically extracted from the live tweet in each scene. The time zone was outputted as the search result. Table 1 shows the number of scenes by category in each program.

Table 1. Number of scenes in each program category

Category	Yofukashi	Friday	Pokemon	Total
Person	5	13	7	25
Action	8	70	4	82
Emotion	1	1	1	3
Summary	7	2	0	9
Scenery	2	19	2	23
Conversion	26	17	9	52
Total	49	122	23	194

4.2. Result

In this section, the experimental results are presented. By investigating the error between the output time zone and the correct answer time zone, we compare the search performance with and without MMR and confirm the difference depending on the program genre. Furthermore, by visualizing the vector of the key phrases in the two-dimensional coordinate space, the similarity between scenes is roughly determined.

4.2.1. Event scene search results

Table 2 shows the average error when the smallest error is selected from the top five search results.

Table 2. Average error of each program

	With MMR	Without MMR
Yofukashi	9 min. 41 sec.	14 min. 43 sec.
Friday	11 min. 3 sec.	16 min. 3 sec.
Pokemon	3 min. 33 sec.	5 min. 38 sec.

As a result, there was an error of approximately 5% to 10% with respect to the broadcast time of each program. The error was smaller when MMR was used than

when it was not used. This is considered to be the result of MMR being able to suppress the duplication of similar key phrases. The difference between the correct answer data and the tweet is considered to be the cause of the error. For example, even though the correct answer data says "XX (person name) appeared," the tweet text was written as "XX Kita" using Japanese slang. The context is important for vectorization by BERT to absorb the notational fluctuation of such expressions. There is a difference in the amount of information between the key phrase vector and the sentence vector, and it is considered that there is a large difference between the two in the obtained vector. In addition, we used the Wikipedia-based pre-learning model this time, but this model did not handle the internet slang words such as "Kita" (come out) well. In addition, since the correct answer data itself is a different type of expression (keyword rather than phrase) from the tweeted sentence, it is necessary to improve the method by converting a search query into a key phrase and reviewing the experimental setting itself.

4.2.2. Key phrase visualization for each scene category

Figures 1 to 6 represent two-dimensional coordinates obtained by dimensionally compressing the BERT vector of the key phrase extracted per unit time for each scene category of the program "Friday Road SHOW" using t-distributed stochastic neighbor embedding (t-SNE) [10]. This is visualized in space.

In the correct data created this time, most event scenes are classified as "action." Generally, in an event scene related to a motion, when a live tweet about how a viewer feels about the motion of a character, an adjective and a proper noun (personal name that is used to indicate a person) are often included in the tweet. For this reason, the key phrases include not only motions but also various types, such as emotional expressions and person names. In addition, a number of key phrases extracted in the "feelings" and "summary" scenes (of which relatively few were classified) were found to match the emotions and summaries; however, in this method, proper nouns and adjectives were used. Since the combination of was used as a key phrase candidate, the verbs, adverbs, and general nouns that are necessary to express emotions and abstract scenes could not be extracted. Thus, we assume that the divergence from the search query was large.

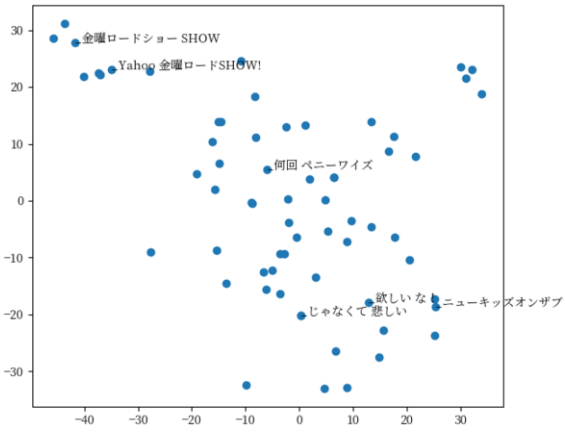

Figure 2. Key phrases in the "person" scene

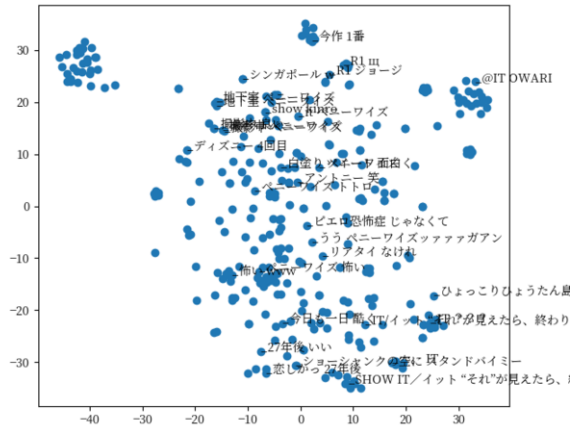

Figure 3. Key phrases in the "action" scene

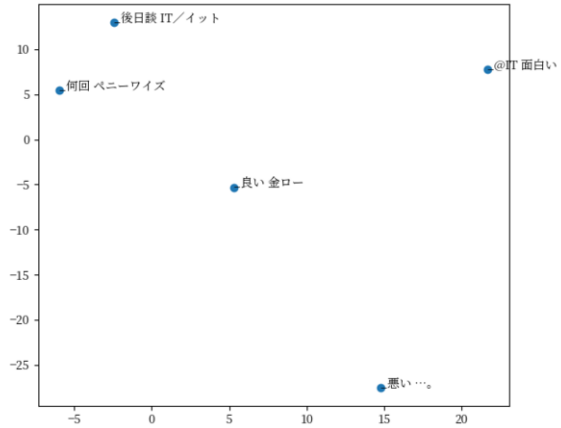

Figure 4. Key phrases in the "emotion" scene

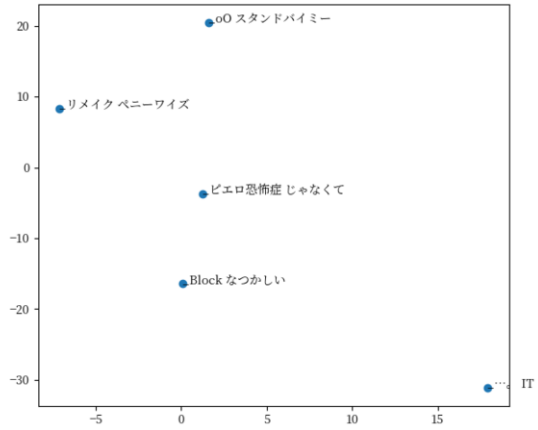

Figure 5. Key phrases in the "summary" scene

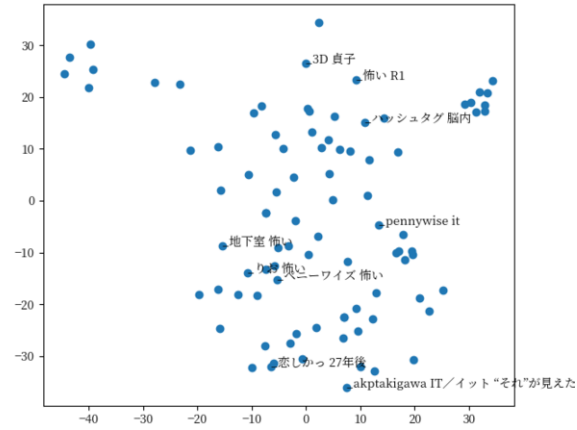

Figure 6. Key phrases in the "scenery" scene

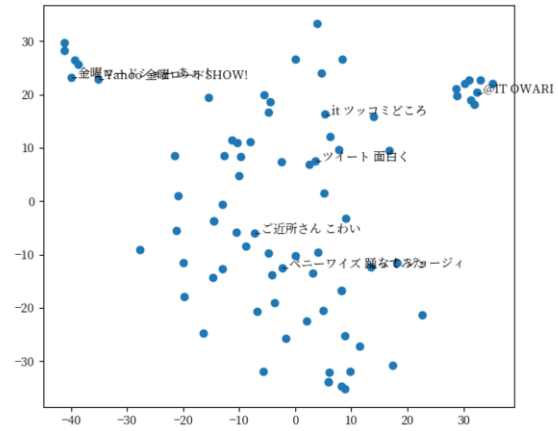

Figure 7. Key phrases in the "conversion" scene

5. Conclusions

In this research, we focused on TV program live tweets and devised a system that retrieves and classifies the scenes by extracting the key phrases that represent the event scenes of the program from the tweeted sentences. As a result, approximately 5% to 10% of the total error that was observed in the performance evaluation was due to the time zone error.

We also succeeded in extracting more important key phrases by mining these phrases based on the vector using the EmbedRank algorithm and MMR. In addition, we simultaneously extracted key phrases with part-of-speech patterns (e.g., proper nouns and adjectives), "scene estimation," and "emotion estimation." Conventional research has extracted these separately. We expect that it will be easier to provide feedback on the results (e.g., the type of influence it has on the viewers) and to more simply provide scene information when deploying services using this method. Furthermore, previous research has manually determined the emotional polarity value (an index of emotion)

for each emotional word. However, we succeeded in significantly reducing the required human labor by integrating emotional words as adjectives.

On the other hand, we determined whether the outline of the scene could be confirmed by scene classification and visualization of the key phrase; however, in each category, there was a tendency for many key phrases to be extracted that did not describe the characteristics of the scene well. Therefore, it was found that the key phrase extraction condition needs to be improved in order to use the visualization of the key phrases to confirm the general tendency of the scene.

In the future, in order to improve the accuracy of key phrase extraction, we plan to use the BERT pre-learned model based on SNS. We also plan to improve it so that the context can be properly acquired, and we intend to examine the conversion method from the input query to the key phrase. In addition, we aim to construct a TV program event scene search system from more intuitive keyword input by examining the weighting of key phrases corresponding to the frequency of the occurrence of words, and we hope to expand key phrase candidates by optimizing the part-of-speech conditions.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers JP20K12027, JP18K11549.

References

[1] Masami Nakazawa, Keiichiro Hoshida, Chihiro Ono: Detection and Labeling of Significant Scenes from TV program based on Twitter Analysis. DEIM Forum 2011 F5-6.
[2] David A. Shamma, Lyndon Kennedy, Elizabeth F. Churchill: Tweet the Debates: Understanding Community Annotation of Uncollected Sources, In *Proceedings of the first SIGMM workshop on Social media*, pp. 3-10, 2009.
[3] Takashi Yamauchi, Yukiko Nakano: A Scene Explorer for TV Programs based on Twitter Emotion Analysis, In *Proceedings of The 26th Annual Conference of the Japanese Society for Artificial Intelligence*, 2012
[4] Plutchik, R: The Emotions, University Press of America, Lanham, 1991.
[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1, pp. 4171–4186, 2019.
[6] https://github.com/yoheikikuta/bert-japanese. Last viewed date: 2020/2/5
[7] Kamil Bennani-Smires, Claudiu Musat, Andreea Hossmann, Michael Baeriswyl, Martin Jaggi: Simple Unsupervised Keyphrase Extraction using Sentence Embeddings, In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pp. 221–229, 2018.
[8] MeCab: Yet Another Part-of-Speech and Morphological Analyzer, https://taku910.github.io/mecab/. Last viewed date: 2020/2/5
[9] Neologism dictionary for MeCab, https://github.com/neologd/mecab-ipadic-neologd. Last viewed date: 2020/2/5.
[10] Laurens van der Maaten, Geoffrey Hinton: Visualizing Data using t-SNE, *Journal of Machine Learning Research*, Vol. 9, pp. 2579-2605, 2008.