Extracting and Clustering of Story Events from a Story Corpus

Hye-Yeon Yu[1], Yun-Gyung Cheong[2], and Byung-Chull Bae[3*]
[1] Department of Electrical and Computer Engineering, Sungkyunkwan University
Suwon, Korea
[e-mail: yu0529@skku.edu]
[2] Department of AI, Sungkyunkwan University
Suwon, Korea
[e-mail: aimecca@skku.edu]
[3] School of Games, Hongik University
Sejong, Korea
[e-mail: byuc@hongik.ac.kr]
[*]Corresponding author: Byung-Chull Bae

Received February 27, 2021; revised August 27, 2021; accepted August 24, 2021; published October 31, 2021

Abstract

This article describes how events that make up text stories can be represented and extracted. We also address the results from our simple experiment on extracting and clustering events in terms of emotions, under the assumption that different emotional events can be associated with the classified clusters. Each emotion cluster is based on Plutchik's eight basic emotion model, and the attributes of the NLTK-VADER are used for the classification criterion. While comparisons of the results with human raters show less accuracy for certain emotion types, emotion types such as joy and sadness show relatively high accuracy. The evaluation results with NRC Word Emotion Association Lexicon (aka EmoLex) show high accuracy values (more than 90% accuracy in anger, disgust, fear, and surprise), though precision and recall values are relatively low.

Keywords: Event Representation, Emotional Event, Event Clustering, Event Extraction, Sentiment Analysis of Story Event.

1. Introduction

A story is a communication between an author and readers. The author conceives a story in his or her mind, and represents it using various media such as text, sound, or films. While the reader perceives a story that is represented in a specific medium, an abstract form of the story is constructed in the reader's mind. The story that is restored by the reader may not be the same as the story conceived by the author. A skillful author would consider a variety of different mental models of the reader; an ideal (or sentimental) reader would make efforts to grasp the author's underlying messages [1]. While there will be many varying causes for the individual differences of story comprehension, different intelligence capacity can be one reason.

Characters, events, and setting (location and time) are the three main components of a story [2]. In particular, the first two components - characters and events - are crucial to an understanding of a story in respect of plot, which is defined as "rearranging a series of events focusing on causality" [3]). Put differently, characters would take actions, or some events causally happen to them. While experiencing the events in the story, the characters feel various emotions and the reader also feels emotions accordingly [4].

The emotional aspects of story events are crucial to the story understanding. As an approach to understanding of emotions in text, sentiment analysis has been widely used in various social networking services like Twitter [5] or predicting box office success [6]. In this article, we postulate that recognition and classification of story events is critical for building the reader's mental model for story understanding.

The comprehension of events is key to story understanding. As an approach to event-based story comprehension, Chambers and Jurafsky [7] have employed the notion of narrative event chains, where a narrative chain refers to "a partially ordered set of narrative events that share a common actor." As an evaluation metric for the narrative event chain model, they suggested a "narrative cloze test" by slightly modifying the well-known cloze test, where one has to guess a blank word in a given sentence. With the publication of the story text data, collected through crowdsourcing [8], the suggested narrative cloze test has been widely used as one of the standard evaluation measures for the studies of story generation and understanding.

This article depicts how events can be extracted and represented from text stories. We also describe our method to cluster the extracted events through abstraction with similar emotions. This article, in addition to the detailed description, extends our previously published results [9] in two ways. First, 'neutral' class is newly appended to better classify emotions. Second, we conducted further experiments with additional emotion dataset to improve accuracy. The main contributions of this paper are twofold: 1) extracting and identifying emotion events from a text story corpus using a standard NLP parser; 2) presenting a way to map the emotional scores of story events, including neutral emotion, in addition to the eight basic emotions proposed by Plutchik [10].

2. Background and Previous Works

This section describes the previous works to represent and extract events from text stories.

2.1 Representing Events from Text Story

After pre-processing the text into tokenized words and sentences, story events can be represented and extracted using natural language processing tools (such as NLTK [11] and Stanford coreNLP [12]), mainly consisting of three phases:

(i) POS (Part-Of-Speech) tagging
(ii) Dependency parsing (including coreference resolution)
(iii) Text normalization (including lemmatization and named entity resolution)

In general, the extraction and representation of events from a text story can be roughly divided into two categories. One is tuple-based representation, and the other is 5W(what, who, why, where, and when) representation.

2.1.1 Tuple-based Event Representation

When expressing events as tuples, a normalized form of a sentence is required to extract core components from sentences. In early studies, only subjects and verbs are extracted along with the description of the relationship(dependency) between the two. The format is *Character(subject) : (verb, Dependency)*, where dependency refers to the role of a character in the sentence (e.g., subject - subj; object - obj) [7]. For instance, the sentence "Anthony called Laura for support and she helped him" can be transformed to four tuples for each character: *Anthony : (call, subj) (help,obj); Laura : (call, obj) (help, sub)*.

The above approach is simple and efficient. However, it is hard to detect the connections between the two characters. For example, the tuple "Anthony(call, subj)" does not contain information about who the receiver is; the connection between two tuples "Anthony : (call, subj)" and "Laura : (call, obj)" is unknown. To solve this problem, Pichotta and Mooney proposed a detailed event representation to contain more information - such as V(s, o, p), where V refers to Verb, s refers to subject, o refers to object, and p refers to prepositional object [13]. When applying this event representation, the above example is represented as "call(Anthony, Laura, ∅), help (Laura, Anthony,∅)". This representation can also cover related information between the two events. V(s, o, p) is thus widely used as a basic form of tuple-based event representation.

In addition to the basic form of tuples, it may be useful to include more information other than the grammatical components. Martin, et al. [14] employed event representation format V(s, o, p, g), where g is the story genre. They used topic modeling algorithm to classify a corpus of text stories into 100 genres. Each genre is given a number, and this number is the component 'g' of the 4-tuple representation.

2.1.2 5W Event Representation

A well-known "Five Ws and How" is the general event structure of news stories [15], where the "5W event representation" expresses the structure of a text story in who, what, where, when, and why. Dependency parser and an information extraction module from NLTK are exploited for this event extraction method. For instance, information corresponding to subject(1), predicate(2), object(3), location(4), and time(5) can be extracted and matched to "who(1)", "did what(2)", "to whom or what(3)", "where(4)", and "when(5)", respectively [15].

2.2 Inferring Relations among Story Events with Language Models

Inferring the causal and temporal relations among story events is crucial to story comprehension, but it requires huge common sense knowledge base and good inference engines. Previous works on inferring a sequence of events or script learning from (short) text stories can be roughly categorized into two methods - count-based (or probability-based) method or neural language model-based method.

2.2.1 Count-based Approaches

Traditional approaches to comprehending a sequence of events from short text stories are mainly based on probability models by counting the frequencies of events that co-occur in the story. This method can predict the succeeding or contingent event for either a single character [7] or multiple characters [16]. Those studies are often based on pointwise mutual information (PMI), where PMI can be either unordered or order-sensitive. In the order-sensitive PMI models, for instance, $C(e_1, e_2)$ and $C(e_2, e_1)$ are treated differently. In [13], interactions between multiple participants were considered.

2.2.2 Neural Language Model-based Approaches

Various neural net algorithms are also applied to statistical language modeling. Pichotta and Mooney [17][18] presented a novel statistical script model using Long Short Term Memory(LSTM) units. They used tuple event representation (v; e_s; e_o; e_p; p), where v: verb lemma, e_s: subject, e_o: object, e_p: preposition, p: proposition relating to v and e_p). In narrative cloze inference tasks, LSTM has performed better than traditional methods like unigram and bigram.

Martin et al. [14] suggested a modified version of seq2seq model to automatically generate narrative texts. In their approach, an event2event network is first built from a corpus of movie plot summaries, using a recurrent multi-layer encoder-decoder network to predict the following event. The predicted events then are converted to text by using event2sentence neural network.

2.3 Emotion Models

Emotion plays an important role in storytelling. Emotion can be a cause for an event as story characters can take (or not take) certain actions based on their emotional status. Emotion is also the effect the story characters feel from a given event, which is expressed through various manifestations such as subjective feeling, nonverbal signal as in **Fig. 1** [19].

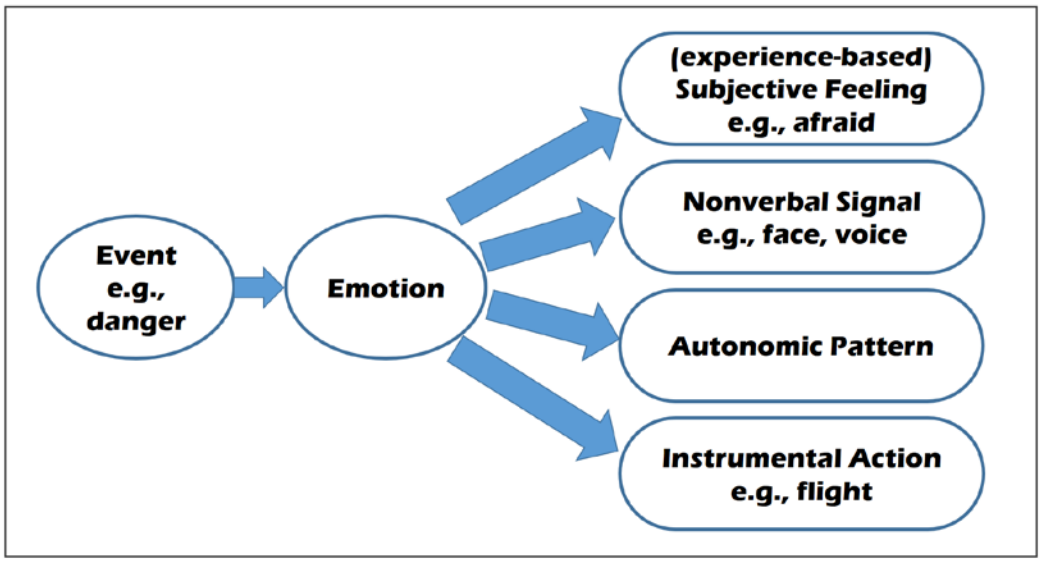

Fig. 1. Events as the causes of emotions which can be represented via various manifestations [19]

Traditional models of emotion fall into two categories - categorical or dimensional. In the categorical models of emotion, emotions are distinct from one another. A typical example of categorical emotion models is Ekman's six basic emotions (happiness, sadness, anger, fear, disgust, and surprise), where each fundamental emotion acts as a separate category [20]. While the categorical classification of emotion is easy to understand and widely used in everyday life, it can not represent many other (similar or complex) emotions outside of the defined emotion categories (e.g., annoying, frustrating, boring, etc.).

In the dimensional models of emotion, all human emotions are represented on a rating scale of either two dimensions (valence and arousal) [21] or three dimensions (pleasantness, arousal, and dominance) [22]. The dimensional emotion models have some advantages over categorical models as they can represent the similarities and differences between various emotions. For instance, anger and fear are both described as negative valence and high arousal in the 2-dimensional emotion model. We can thus consider them as similar emotions, though they are similar but different in terms of 'dominance' in the three dimensional model of emotion. The dimensional emotion models are limited in that they fail to classify discrete emotions, due to ambiguousness of the dimensions, etc.

Plutchik's Wheel of Emotions model [10] incorporates characteristics of both a categorical model and a dimensional model. The Wheel of Emotions model provides eight basic human emotions (joy, sadness, anger, fear, disgust, surprise, anticipation, and trust), and each emotion is represented with three intensity degrees. For example, joy with more intensity is denoted as 'ecstasy,' and joy with less intensity is denoted as 'serenity'; fear with more intensity as 'terror,' and fear with less intensity as 'apprehension.' In the Wheel of Emotions model, opposite emotions (e.g., joy and sadness; anticipation and surprise) are facing each other, and two adjacent basic emotions can be combined into a compound emotion. For example, love is denoted as a combination of joy and trust, contempt as an integration of anger and disgust. Interestingly, fear and anger are placed close together in the 2-dimensional model, but they are in the opposite direction in the Plutchik's Wheel of Emotions model.

3. Experiment

This section describes our simple experiment on emotion-based event modeling and extraction from ROCStories dataset [8]. A detailed hardware specifications for the experiment are i7-6700 CPU, RAM 16GB, GPU NVIDIA GeForece GTX 1080. As for the software versions in the case of sentence structural analysis, stanford-corenlp-3.9.2 and ClausIE [23] are used on Java SE1.8. For sentiment analysis and evaluation, NLTK 3.4.5 module is applied based on Python 3.8.1.

3.1 Dataset

There are a wide diversity of text story corpora from the CMU Movie Summary corpus in which movie plot summaries are collected from Wikipedia [24] to movie scene descriptions/dialogues collected from the IMSDB (Internet Movie Script Database) [25] [26]. There is also a corpus named ROCStories, a text story dataset collected from crowdsourcing [8]. The CMU Movie Summary corpus is suitable for identifying the structure of abstract stories consisting of characters and events; the IMSDB dataset is helpful for understanding the primitive actions and conversations between story characters. The ROCStoreis dataset is useful for the structure analysis in very short stories.

Among the various text story datasets, we employ ROCStories dataset [8]. The ROCStories dataset is practically useful in several senses:

- Each story contains precisely five sentences.
- Each sentence includes no direct speeches and no dialogues.
- Each story keeps a coherent dramatic structure which consists of the three-act structure having beginning, middle, and ending.
- Most story events are commonly occurring ones in our everyday life.

3.2 Event Modeling and Extraction

We follow the method by Pichotta and Mooney [13] using a 4-tuple event representation (e_v, e_s, e_o, e_{pp}), where e_v:verb, e_s:subject, e_o:object, and e_{pp}:prepositional phrase.

$$E = e_v + e(e \in \{e_s, e_o, e_{pp}\}) \tag{1}$$

Of the elements in the 4-tuple event representation, the verb element is the most important. Other items such as subjects, objects, or prepositional phrases may be excluded when extracting events, as they are often omitted in a sentence. All the stories in ROCStoreis dataset comprise five sentences, and most sentences include compound or complex sentences. **Fig. 2** shows an example of the dependency relations of a compound sentence from a story in ROCStories corpus.

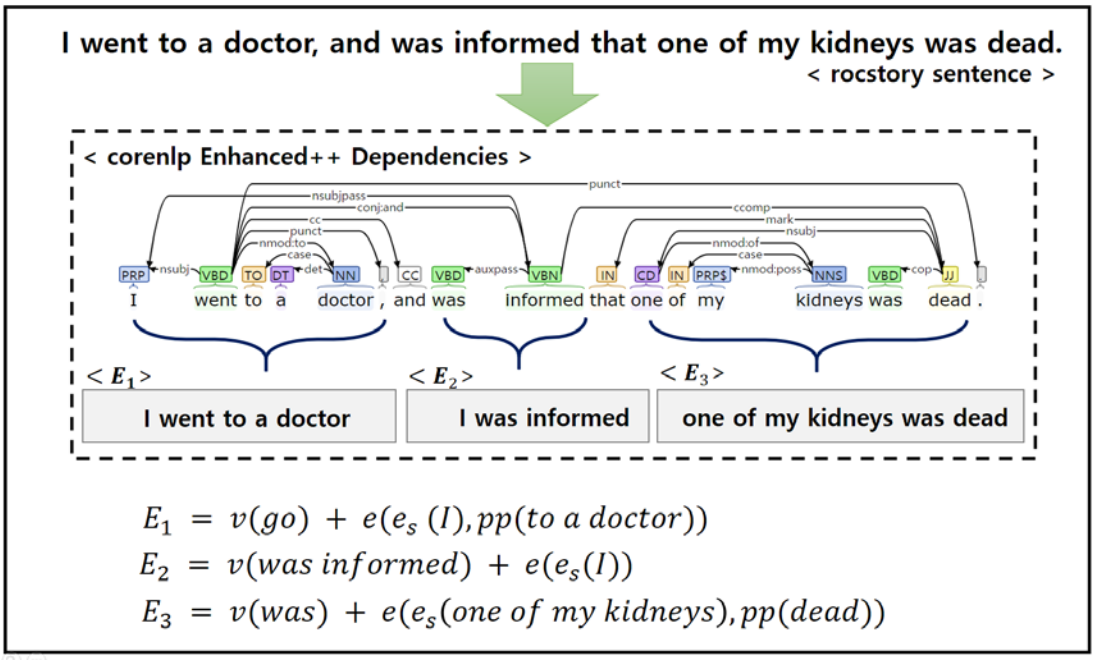

Fig. 2. Example of event extraction from a story sentence

As shown in **Fig. 2**, a several number of events can be included in either one compound or one complex sentence. In the example, one sentence involves three events. Thus all the compound and complex sentences are converted into a multitude of simple sentences, based on the dependency information using Stanford CoreNLP [12]. In total, 496,847 events are extracted from 263,325 sentences in ROCStories, where 7,384 events are classified based on verbs only.

Table 1 shows an example of formatted events, extracted from a story titled "My Diagnosis" in the ROCStories corpus. While converting compound/complex sentences into multiple simple sentences, multiple or overlapped events are extracted. Specifically, when the verb belongs to stop-word verbs (e.g., be, have), objects or prepositional phrases necessarily complement events.

Table 1. Example of structured events extracted from text story

text	When I was 12 years old, my dad got angry and kicked me aggressively. Afterward, I became very ill, and tasted something metallic. I went to a doctor, and was informed that one of my kidneys was dead. Ever since, I've had swelling and hypertension. I started taking medications to combat the symptoms at 13.		
event struct	e_s	v	$e_o(pp)$
	dad	kick	me
	dad	angry	
	I	taste	metallic (something metallic)
	I	become	
	my	have	kidneys
	one (one of my kidneys)	be (dead)	
	I	be (be inform)	
	I	go	(to a doctor)
	I	have	swelling (swelling and hypertension)
	I	start	(taking medications to combat the symptoms at 13)

Table 2 contains top 20 verbs that occur most frequently in ROCStories data sets, except for stop-word verbs that are excluded based on the NLTK Stop Word List. There are some general verbs such as get and take, which could be classified further with including the propositions (e.g., get onoff, take on/off). However, we did not consider it in this paper.

Table 2. Top 20 events (verbs) based on frequency of occurrence in the ROCstories dataset

	verb	count		verb	count
1	go	18,546	11	love	4,521
2	get	13,332	12	start	4,490
3	decide	9,573	13	come	4,415
4	want	9,215	14	look	3,950
5	take	7,593	15	feel	3,750
6	make	6,840	16	play	3,564
7	find	6,310	17	ask	3,460
8	buy	5,617	18	see	3,427
9	try	4,852	19	work	3,301
10	tell	4,677	20	give	3,294

3.3 Emotion-based Event Clustering

The primary goal of our work is to understand a text story by identifying events using a formatted structure. However, extracting events based on verb extraction results in extracting a large number of events, which requires further classification. For example, the extracted events can be grouped using either abstraction or clustering. In this paper we address a simple way of event clustering based on emotions.

For emotion-based event clustering, we selected a rather straightforward approach using the NLTK-VADER (Valence Aware Dictionary and sEntiment Reasoner) sentiment analyzer with which we cluster the extracted events into eight basic emotion types (joy, sadness, anger, fear, disgust, surprise, anticipation, and trust) based on Plutchik's Wheel of Emotions model [10]. The Plutchik's Wheel of Emotions model has several advantages, including easy extension from 8 basic emotions to 24 emotions (with different intensities) and inclusion of other compound emotions such as Love (Joy + Trust) or Contempt (Anger + Disgust). Currently, we focus only on the eight basic emotions.

The output (i.e., the return value) of the NLTK-VADER sentiment analyzer shows corresponding sentiment value to its input (an event structure in this article) with the indication of polarity (positive, negative, neutral). Using the compound values in the return value, we first separated out 'neutral' (i.e., 'unemotional') events that presented either neutral (neu: 1.0) or zero compound value (com:0). Then, with the compound values returned from the NLTK-VADER sentiment analyzer, all the events including numerically expressed sentiment values were clustered using a simple k-Means algorithm in Weka [27]. When grouping the emotions using Weka, three attributes were used: the extracted events, emotion labels (8 basic emotions + neutral), and the emotional values of the NLTK-VADER output (excluding the compound value to minimize the influence of clustering). As a result, we listed 150,955 emotional events in terms of the eight basic emotions in the Wheel of Emotions model and 273,748 neutral events (424,703 events in total).

In **Table 3**, the sentiment polarity (either positive or negative) and compound value (from -1 to +1) is shown for each emotion type, along with some examples of events.

Table 3. Sentiment polarity and NLTK compound values along with sample events for 8 basic emotion types

emotion	sentiment	NLTK-value	(extracted) sample events	events
JOY	positive	0.5859	{his have car own dream car} {he eat a delicious slice of pizza} {John love his trip to Chicago}	46,792
ANTICIPATION	positive	0.1027	{his have mom Bernie mom returned to the store pay for the sandwich} {Julie want her very own phone}	14,082
TRUST	positive	0.5106	{Rick feel proud} {Mary go to this cute cafe for breakfast}	25,972
SADNESS	negative	-0.4404	{It ruin the friendship and relationship} {I end up missing the entire movie} {They be never seem be happy}	21,431
ANGER	negative	-0.5719	{She make lunch the most disgusting lunch} {weather delay have the flight} {parents be furious}	2,905
FEAR	negative	-0.4939	{officer spotted the suspect A short time later} {he be in danger}	9,303
DISGUST	negative	-0.5994	{He be disappoint learn} {he lose the race} {she complain to management} {it be the saddest day of her life}	18,261
SURPRISE	positive	0.2732	{he get the part of Hamlet in the play} {it look like worms}	12,209
NEUTRAL	-	0.0	{Jo be brainstorm} {James bring it} {Charlie be painting garage}	273,748

About 64 % (273,748/424,703) of the events were classified as neutral. Among the eight emotional events (150,955 in total), Joy is the most frequent emotion type (46,792/150,955; 31.0%), and Anger is the least frequent emotion type (2,905/150,955; 1.9%). Based on the NLTK-values (i.e., compound values), two positive emotion types (Joy and Trust) have similar values (0.5859 and 0.5106 respectively); four negative emotion types (Sadness and Fear, Anger and Disgust) also have similar values (-0.4404 and -0.4939, -0.5719 and -0.5994). It is noted that in our study, Surprise emotion is considered a positive emotion, though surprise in some emotion theories is regarded as purely aroused, neither positive nor negative.

Fig. 3 shows the distribution of emotions by cluster, where different colors denote different groups classified by Plutchik's eight basic emotions. For example, the number of events classified as cluster 0 is 14,086 in total, where Anticipation is the representative emotion in orange color. Each cluster (0 to 7) has a representative emotion - 0:ANTICIPATION, 1:SADNESS, 2:JOY, 3:DISGUST, 4:DISGUST (+ FEAR), 5:SURPRISE, 6:TRUST, 7:TRUST. Note that TRUST and DISGUST emotions represent two clusters, respectively (clusters 6 and 7 for TRUST; clusters 3 and 4 for DISGUST). Also note that FEAR emotion is partially represented by cluster 1 (SADNESS) and cluster 4 (DISGUST); ANGER emotion is also represented partially by clusters 1 and 4. It might be due to uneven distribution of emotional events in the dataset. Among all events, excluding the events associated with NEUTRAL emotion, the ANGER events occupy just 1.9% (2,905 out of 151,135); FEAR emotion 6.2% (9,303 out of 151,135). We postulate that it is because the NLTK values of some emotions are only slightly different. For example, SADNESS (-0.4404) and FEAR (-0.4939); ANGER (-0.5719) and DISGUST (-0.5994) (see **Table 3**).

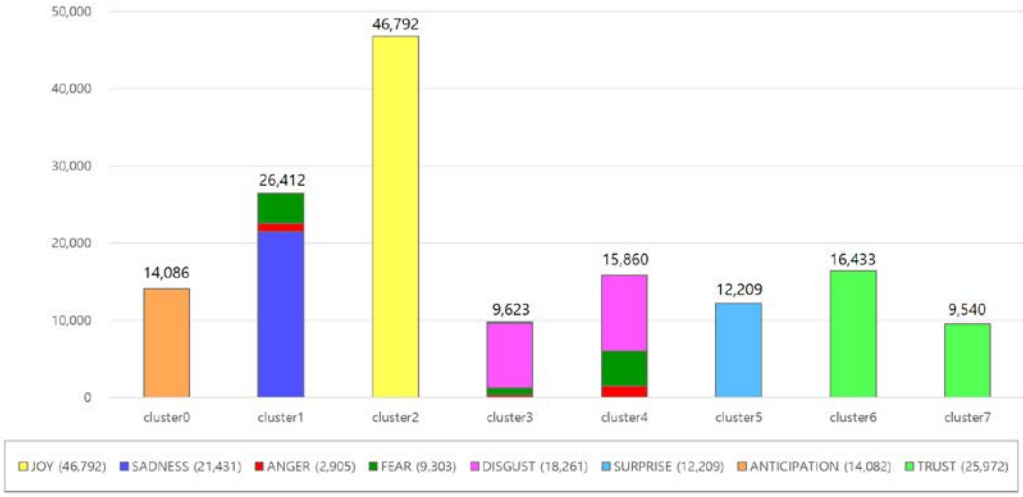

Fig. 3. Distribution of emotions per cluster

4. Experimental Results and Analysis

4.1 Evaluation by Human Raters

For the evaluation of the emotion clustering results, a total of 45 sentences (five sentences for each nine emotion types, including neutral emotion) were randomly selected and assessed by seven human raters who volunteered (two females and five males). The raters were asked to select one primary emotion among 10 choices (including 'neutral' and 'Other' as an open answer in addition to the 8 basic emotions) to represent each emotion cluster.

Fleiss' kappa statistic [28] is used to assess the reliability of agreement among the seven human raters. The overall kappa value is 0.35 for all choices (i.e., eight emotions + neutral + Other) and 0.40 for the choices excluding 'Other' (i.e., eight emotions + neutral). The top two agreed emotions are JOY(k=0.43) and DISGUST(k=0.34), and the least agreed emotions are ANGER(k=0.11) and SURPRISE(0.11). The value of k by sentiment is 0.41 when including 'Other' (i.e., positive, negative, neutral, and other), and 0.51 when excluding 'Other' (i.e., positive, negative, and neutral) as shown in **Table 4**.

Table 4. Fleiss' kappa values for each emotion among seven human raters

Emotions	$\sum_{n=1}^{N} P_i$	k
JOY	3.90	**0.43**
SADNESS	1.91	0.21
ANGER	1.35	0.11
FEAR	2.20	0.26
DISGUST	2.43	0.34
SURPRISE	1.99	0.11
ANTICIPATION	1.33	0.12
TRUST	2.70	0.30
8 Emotions + NEUTRAL + OTHERS	19.72	**0.35**
8 Emotions + NEUTRAL	21.66	**0.40**

As seen in **Table 4**, the agreement level among the different raters is low. As it is hard to directly compare the results of our proposed method with those of the human raters, we make a simple assumption that the emotion clustering is accurate as at least one human rater agreeing with the system. Under this assumption, JOY, SADNESS, and NEUTRAL (1.0, 0.8, and 0.8, respectively) are highly matched with the results of human raters; ANGER and SURPRISE did not match at all.

The overall accuracy for sentiment analysis is 0.8, where four emotions (joy, trust, anticipation, and surprise) are classified as positive, and four emotions (sadness, anger, fear, and disgust) are categorized as negative. **Fig. 4** and **Fig. 5** show the confusion matrices regarding the eight classes of emotion types and the sentiment types including NEUTRAL and OTHERS, respectively.

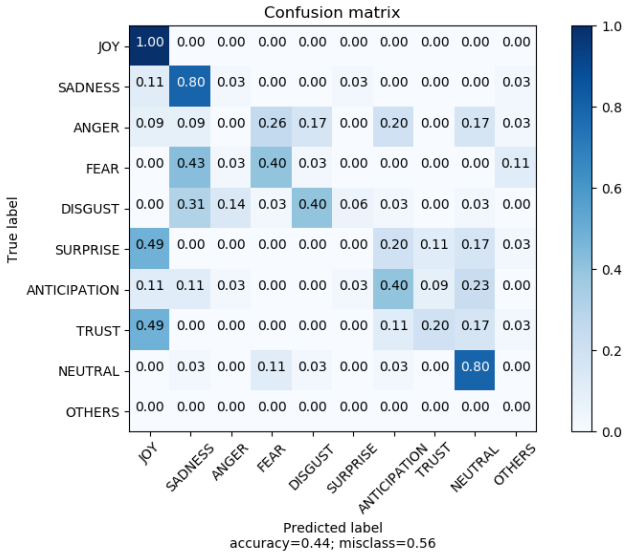

Fig. 4. Confusion matrix for ten-emotion classes, including neutral and others

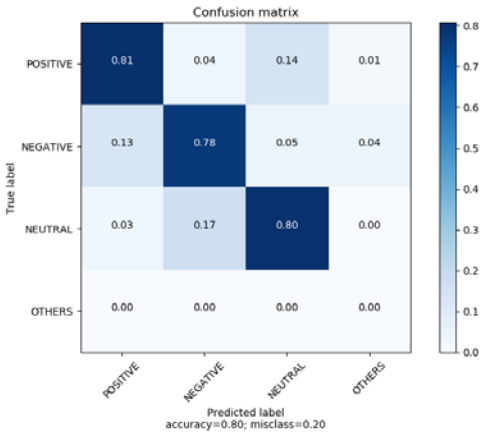

Fig. 5. Confusion matrix for four-sentiment classes, including neutral and others

4.2 Evaluation with EmoLex Dataset

We evaluated our results with the open data NRC Word-Emotion Association Lexicon (a.k.a. EmoLex) [29], where eight emotions and two sentiments (positive and negative) are manually labeled through crowdsourcing. With EmoLex, the degree of each emotion is represented as a value between 0 and 1.

While we employ single-labeling, which assigns a single representative emotion to an individual event, EmoLex is a multi-label dataset where a single word can have multiple emotional labels. For example, the word 'feeling' assigns 1 to all the eight emotion types. Hence, when an event includes the word 'feeling' and contains no emotional words, all of the eight emotions are labeled. As EmoLex labels multiple emotions for each word in an event, we consider a labeled emotion with the largest accumulated value as a representative emotion. In the case where multiple emotion labels receive the same highest scores, those multiple emotions are all recognized as representative emotions. We also label an event with no matching emotions as NEUTRAL, since there is no NEUTRAL class in EmoLex.

Table 5 shows the accuracy, precision, recall, and F1 scores of the event set extracted from the ROCStories dataset, applying the emotion labels in the NRC Emotion Lexicon (EmoLex). While the accuracy of ANGER, DISGUST, FEAR, and SURPRISE emotions are scored high (more than 90% accuracy), their precision and recall values are scored relatively low. It may be due to the differences of the labeling scheme between EmoLex and our approach, where EmoLex employs multi-labeling and our approach employs a single-labeling scheme. It is noted that the sum of the ratio column's values exceeds 43% (143% in total) as the EmoLex allows multiple emotion labels rather than choosing just one representative emotion label.

Table 5. Results of Emotion Classification Using EmoLex dataset

Emotions	Accuracy	Precision	Recall	F1	Ratio
JOY	0.85	0.51	**0.35**	**0.42**	15 %
SADNESS	0.89	0.19	0.12	0.15	8 %
ANGER	**0.93**	**0.58**	0.06	0.11	7 %
FEAR	0.90	0.25	0.06	0.10	9 %
DISGUST	**0.93**	0.26	0.23	0.24	5 %
SURPRISE	0.90	0.16	0.06	0.09	8 %
ANTICIPATION	0.81	0.16	0.03	0.06	16 %
TRUST	0.81	0.42	0.15	0.22	18 %
NEUTRAL	0.66	**0.68**	**0.77**	**0.72**	**57 %**

4.3 Discussion

Emotions are innately subjective, so recognition from a given text can differ depending on human raters and contexts. The kappa values for measuring the inter-reliability among the raters show a diverse range - JOY (0.43) and DISGUST (0.34) are more agreed upon by the raters than ANGER (0.11) and SURPRISE (0.11) in our study. Interestingly, Joy and Disgust have higher rater-agreements than Anger and Surprise. We can assume that Joy and Disgust are more distinguishable than the other emotion types. Among the four positive emotions, Joy (0.43) and Trust (0.3) are more agreed than Anticipation (0.12) and Surprise (0.11). As for the four negative emotions, Disgust (0.34), Fear (0.26), and Sadness (0.21) are more agreed and distinguishable than Anger (0.11). While our approach allows only one representative emotion for an event, multi-emotion labeling can be helpful for a more accurate analysis.

Our approach entails conversion from text to a set of 4-tuple formatted events. While the 4-tuple representation has benefits in terms of conciseness and representation, necessary information on emotion classification might be excluded during the conversion process. Thus using a flexible event format to include words or phrases contributing to the emotion

recognition may enhance the results of classification.

5. Conclusion

This article describes how events can be extracted and analyzed from the given text stories. We conduct an experiment on extracting and clustering events using the ROCStories dataset. After constructing a set of tuple-formatted events form the ROCStories corpus, we first compute the sentiment values of the events using NLTK-VADER sentiment analyzer. And then we apply k-means clustering algorithm to test whether emotionally similar events can be clustered into emotional groups based on Plutchik's eight basic emotions. The evaluation by human raters shows that JOY and SADNESS emotions are highly matched. We also evaluate our approach using the EmoLex dataset which has a different emotion labeling scheme. The evaluation results show high scores in accuracy but relatively low scores in precision and recall. Among the eight emotions, JOY shows the highest F1 score (0.42).

While the NLTK-VADER is a convenient sentiment analysis tool, the results can be improved with other sophisticated methods. As a further study, we will apply the latest sentiment analysis approaches and compare the results. We also plan to extend the eight basic emotions to include complex emotions such as love, jealousy, envy, and remorse, which will help us better understand the emotions underlying story events.

Acknowledgement

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (No.2017-0-01772, Development of QA systems for Video Story Understanding to pass the Video Turing Test) and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education(2019R1A2C1006316).

References

[1] B.-C. Bae, Y.-G. Cheong, and D. Vella, "Modeling foreshadowing in narrative comprehension for sentimental readers," in *Proc. of Interactive Storytelling on Interactive Digital Storytelling*, Cham: Springer International Publishing, pp. 1–12, 2013. Article (CrossRef Link).

[2] S. B. Chatman, *Story and Discourse: Narrative Structure in Fiction and Film*, Ithaca, NY: Cornell University Press, 1980. Article (CrossRef Link).

[3] G. Prince, *A Dictionary of Narratology*, Lincoln, NE: University of Nebraska Press, 2003. Article (CrossRef Link).

[4] K. Oatley, "A taxonomy of the emotions of literary response and a theory of identification in fictional narrative," *Poetics*, vol. 23, no. 1–2, pp. 53–74, 1995. Article (CrossRef Link).

[5] J. Song, K. T. Kim, B. Lee, S. Kim, and H. Y. Youn, "A novel classification approach based on Naïve Bayes for Twitter sentiment analysis," *KSII Trans. Internet Inf. Syst.*, vol. 11, no. 6, pp. 2996–3011, 2017. Article (CrossRef Link).

[6] Y. Kim, M. Kang, and S. R. Jeong, "Text mining and sentiment analysis for predicting box office success," *KSII Trans. Internet Inf. Syst.*, vol. 12, no. 8, pp. 4090–4102, 2018. Article (CrossRef Link).

[7] N. Chambers and D. Jurafsky, "Unsupervised learning of narrative schemas and their participants," in *Proc. of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2*, pp. 602-610, 2009. Article (CrossRef Link).

[8] N. Mostafazadeh et al., "A corpus and cloze evaluation for deeper understanding of commonsense stories," in *Proc. of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 839-849, 2016. Article (CrossRef Link).
[9] H.-Y. Yu, S. Park, Y.-G. Cheong, M.-H. Kim, and B.-C. Bae, "Emotion-based story event clustering," in *Proc. of Interactive Storytelling on Interactive Digital Storytelling*, Springer, pp. 348–353, 2019. Article (CrossRef Link).
[10] R. Plutchik, "The nature of emotions," *American Scientist*, 89(4), 344–350, Jul 2001. Article (CrossRef Link).
[11] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*, Beijing: O'Reilly, 2009. Article (CrossRef Link).
[12] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Proc. of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55-60, 2014. Article (CrossRef Link).
[13] K. Pichotta and R. Mooney, "Statistical script learning with multi-argument events," in *Proc. of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 220-229, 2014. Article (CrossRef Link).
[14] L. J. Martin et al., "Event Representations for Automated Story Generation with Deep Neural Nets," in *Proc. of the Thirty-Second AAAI Conference on Artificial Intelligence*, New Orleans, Louisiana, USA, February 2-7, 2018. Article (CrossRef Link).
[15] A. Tozzo, D. Jovanovic, and M. Amer, "Neural event extraction from movies description," in *Proc. of the First Workshop on Storytelling*, pp. 60-66, 2018. Article (CrossRef Link).
[16] N. Chambers and D. Jurafsky, "Unsupervised learning of narrative event chains," in *Proc. of the 46th Annual Meeting of the Association for Computational Linguistics*, June 15-20, 2008. Article (CrossRef Link).
[17] K. Pichotta and R. J. Mooney, "Learning statistical scripts with LSTM recurrent neural networks," in *Proc. of the 30th AAAI Conf. Artif. Intell. AAAI 2016*, pp. 2800–2806, 2016. Article (CrossRef Link).
[18] K. Pichotta and R. J. Mooney, "Using sentence-level LSTM language models for script inference," in *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 279-289, 2016. Article (CrossRef Link).
[19] J. A. Russell, "Core affect and the psychological construction of emotion," *Psychol. Rev.*, vol. 110, no. 1, pp. 145–172, 2003. Article (CrossRef Link).
[20] P. Ekman, "An argument for basic emotions," *Cogn. Emot.*, vol. 6, no. 3–4, pp. 169–200, 1992. Article (CrossRef Link).
[21] J. A. Russell, "A circumplex model of affect," *Journal of personality and social psychology*, 39(6), 1161–1178, 1980. Article (CrossRef Link).
[22] M. Bradley and P. J. Lang, "Measuring Emotion: The Self-Assessment Manikin and the Semantic Differential," *J. Behav. Ther. Exp. Psychiatry*, vol. 25, pp. 49–59, 1994. Article (CrossRef Link).
[23] Del Corro, L. and Gemulla, R., "ClausIE: clause-based open information extraction," in *Proc. of the 22nd international conference on World Wide Web (WWW '13). Association for Computing Machinery*, New York, NY, USA, pp. 355–366, 2013. Article (CrossRef Link).
[24] D. Bamman, B. O'Connor, and N. A. Smith, "Learning latent personas of film characters," in *Proc. of the 51st Annual Meeting of the Association for Computational Linguistics*, pp. 352–361, 2013. Article (CrossRef Link).
[25] Z. Hu, E. Rahimtoroghi, L. Munishkina, R. Swanson, and M. A. Walker, "Unsupervised induction of contingent event pairs from film scenes," in *Proc. of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 369–379, 2013. Article (CrossRef Link).
[26] M. A. Walker, G. I. Lin, and J. E. Sawyer, "An Annotated Corpus of Film Dialogue for Learning and Characterizing Character Style," in *Proc. of the 8th Int. Conf. Lang. Resour. Eval.*, 1373–1378, 2012. Article (CrossRef Link).

[27] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD*, vol. 11, no. 1, pp. 10–18, 2009. Article (CrossRef Link).
[28] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychol. Bull.*, vol. 76, no. 5, pp. 378–382, 1971. Article (CrossRef Link).
[29] S. M. Mohammad and P. D. Turney, "Crowdsourcing a Word-Emotion Association Lexicon," *Computational Intelligence*, 29(3), 436–465, 2013. Article (CrossRef Link).

Hye-Yeon Yu received the M.S. degree in 2010 from Sungkyunkwan University, South Korea. She is currently a Ph.D. candidate majoring in computer engineering, from Sungkyunkwan University, Suwon, South Korea. Her research interests involve AI, computational narrative, and narrative modeling.

Yun-Gyung Cheong received the B.S. degree and the M.S. degree in information engineering from Sungkyunkwan University, in 1996 and 1998, respectively, and the Ph.D. degree in computer science from North Carolina State University, Raleigh, NC, USA, in 2007. She is currently an Associate Professor with Sungkyunkwan University, South Korea. She was a Postdoctoral Fellow with the Center for Computer Games Research, IT University of Copenhagen, from 2010 to 2014 and a Researcher with the Samsung Advanced Institute of Technology, from 2007 to 2010.

Byung-Chull Bae received the B.S. degree in 1993 and the M.S. degree in 1998 from Korea University, South Korea, and the Ph.D. degree from North Carolina State University, Raleigh, NC, USA, in 2009. He is currently an Assistant Professor at School of Games, Hongik University, Sejong, South Korea. He has worked at LG Electronics and Samsung Electronics as a research engineer, and worked for IT University of Copenhagen, Denmark, as a visiting scholar and a part-time lecturer. His research interests include interactive storytelling, affective computing, and game AI.