Exploiting Narrative Context and A Priori Knowledge of Categories in Textual Emotion Classification

Hikari Tanabe **Tetsuji Ogawa** **Tetsunori Kobayashi** **Yoshihiko Hayashi**
Faculty of Science and Engineering, Waseda University
Waseda-machi 27, Shinjuku, Tokyo 1620042, Japan
tanabe@pcl.cs.waseda.ac.jp ogawa.tetsuji@waseda.jp
koba@waseda.jp yshk.hayashi@aoni.waseda.jp

Abstract

Recognition of the mental state of a human character in text is a major challenge in natural language processing. In this study, we investigate the efficacy of the narrative context in recognizing the emotional states of human characters in text and discuss an approach to make use of a priori knowledge regarding the employed emotion category system. Specifically, we experimentally show that the accuracy of emotion classification is substantially increased by encoding the preceding context of the target sentence using a BERT-based text encoder. We also compare ways to incorporate a priori knowledge of emotion categories by altering the loss function used in training, in which our proposal of multi-task learning that jointly learns to classify positive/negative polarity of emotions is included. The experimental results suggest that, when using Plutchik's Wheel of Emotions, it is better to jointly classify the basic emotion categories with positive/negative polarity rather than directly exploiting its characteristic structure in which eight basic categories are arranged in a wheel.

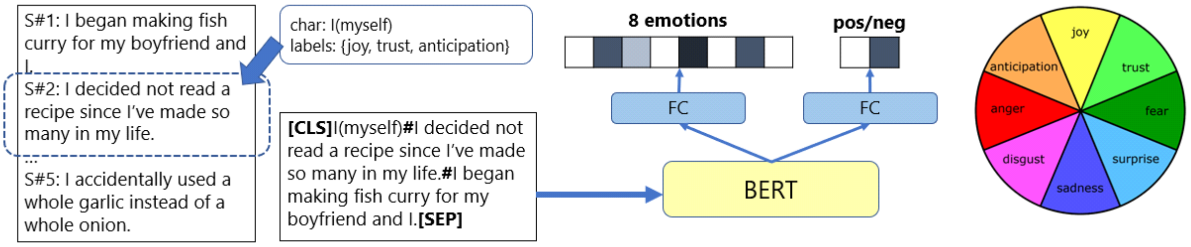
1 Introduction

Understanding of narrative text requires the ability to read between the lines and determine changes in the characters' emotional states concerning the ongoing situation described in the text. Given this motivation, Rashkin et al. (2018) proposed a framework for annotating fully-specified chains of mental states with respect to characters' motivations and emotional reactions. They publicized the resulting annotated corpus, called the Story Commonsense dataset.

(Rashkin et al., 2018) further demonstrated that the mental state of a character in a narrative can be more accurately classified by using the preceding context of a target sentence. Although the results indicate the efficacy of the narrative context, the method might have certain limitations. The approach proposed by (Rashkin et al., 2018) focuses on the character-specific context that selectively concatenates the sub-set of sentences in which the target character appears. Although this account might be effective in excluding apparently irrelevant sentences, it may fail to extract implicit yet useful information from other parts of the context. Thus, we adopt a simple yet presumably effective method that encodes the entire preceding context along with the target sentence using BERT (Devlin et al., 2019).

The task of emotion classification is generally formulated as a multi-label classification task (Zhou et al., 2016; He and Xia, 2018; Yu et al., 2018; Alswaidan and Menai, 2020) due to the innate complexity of human emotions: a sentence can express multiple emotions that an annotator may be unable to disassociate. Among the issues linked to multi-label classification, a crucial one is the handling of mutual dependencies among emotion categories. He and Xia (2018) proposed incorporating a term into the loss function so that it captures the interactions between a pair of Plutchik's basic emotion categories (Plutchik, 1980). However, the reported evaluation results concentrated on a comparison with existing methods and did not disclose any outcomes that demonstrate the efficacy of the proposed loss function. We propose a multi-task learning method that jointly learns to classify Plutchik's eight basic emotion categories and the positive/negative polarity of emotions.

Proceedings of the 28th International Conference on Computational Linguistics, pages 5535–5540
Barcelona, Spain (Online), December 8-13, 2020


S#1: I began making fish curry for my boyfriend and I.
S#2: I decided not read a recipe since I've made so many in my life.
...
S#5: I accidentally used a whole garlic instead of a whole onion.
char: I(myself)
labels: {joy, trust, anticipation}
[CLS]I(myself)#I decided not read a recipe since I've made so many in my life.#I began making fish curry for my boyfriend and I.[SEP]
8 emotions
pos/neg
FC
FC
BERT
joy
trust
anticipation
anger
fear
disgust
surprise
sadness

(a) Story Commonsense dataset (b) Multi-task classification (c) Wheel of Emotions

Figure 1: (a) Story Commonsense dataset, (b) emotion classification model, and (c) Plutchik's Wheel of Emotions.

Our overall experimental results show that both the narrative context and a priori knowledge regarding the emotion category system are effective for increasing classification accuracy. However, the latter is not very significant, suggesting that further exploration into the nature of human emotions that might be latent in written text is necessary.

2 Problem Formulation

In the present work, we use the Story Commonsense dataset (Rashkin et al., 2018) and compare several methods to identify the possibly multiple emotions of a human character that are explicitly or implicitly expressed by target sentences in narrative text. Note that each classification method is applied sentence-by-sentence.

2.1 Story Commonsense Dataset and the Emotion Categories

The Story Commonsense dataset[1] includes around 150,000 annotations of the mental states of characters for approximately 15,000 narrative stories presented in the ROCStories corpus (Mostafazadeh et al., 2016). Although the annotated mental states are of motivation and emotional reaction types, we only consider the latter in the present work. Multiple emotional categories based on Plutchik's basic emotion categories (Plutchik, 1980) may be assigned to each of the sentence–character pairs, as exemplified in Fig. 1 (a). Notice that Plutchik's emotional category system presents eight basic emotion categories that are arranged in a wheel as illustrated in Fig. 1 (c), where each bipolar category indicates a pair of opposite emotions and the categories exhibiting some proximity are adjacent. Among the eight basic categories, *joy* and *trust* are considered positive, whereas *fear, sadness, disgust*, and *anger* are assumed negative in the present work. There has been a debate among psychologists that surprise-like categories are not emotional categories (Ortony et al., 1987). We thus excluded *surprise* and *anticipate* in the positive/negative polarity classification.

2.2 Multi-label Emotion Classification Methods

An emotion classification method is specified by the types of information that are used as the input and the architecture of the classification model. The inputs to the classification method are threefold: a target sentence, a human character associated with the target sentence, and the narrative context. The first two items are explicitly specified in the dataset, and we could devise a way to use them in a model. We needed to consider a technique to incorporate narrative context. One solution proposed in (Rashkin et al., 2018) was to extract sentences in which the target character appears. We instead simply concatenate the preceding sentences regardless of the appearing characters. We formulate a data packet that fed into BERT as follows, where the [CLS] vector is then employed as the input representation and fed into the classifier layer.

```
[CLS]CHARACTER#SENTENCE#CONTEXT[SEP]
```

[1] https://uwnlp.github.io/storycommonsense/

In the present work, the classification models are first divided into single-task models and a multi-task model. The single-task models are characterized by the loss function used during training. The general form of the loss functions can be formulated as follows (He and Xia, 2018). Here, C and p_c respectively indicate the number of categories and the class probability of category c, which is calculated using the sigmoid function. Our baseline model (mnimonic:`BL`) is defined by the formula with $\lambda_1 = \lambda_2 = 0$.

$$L_{emo} = -\sum_{c=1}^{C}(y_c \mathrm{log} p_c + (1-y_c)\mathrm{log}(1-p_c)) + \lambda_1 \sum_{s,t} w_{s,t}(p_s - p_t)^2 + \lambda_2 ||\theta||^2 \tag{1}$$

Notably, the weight $w_{s,t}$ measures the degree of proximity between categories s and t. We compare two methods to define the weights. One, (`He`), is the method presented in (He and Xia, 2018), which uses pre-defined constants by observing the angle between a pair of categories (0.5 for $\frac{\pi}{4}$, 0 for $\frac{\pi}{2}$, -0.5 for $\frac{3\pi}{4}$, and -1 for π). The other, (`Cooc`), calculates the co-occurrence probability between a pair of categories that are pre-computed using the training portion of the dataset. These methods respectively exploit theoretical (`He`) and data-driven (`Cooc`) knowledge of the categories.

The total loss function L_{multi} for a multi-task model (`Multi`) is defined as follows, where the loss term for the binary positive/negative classification L_{pol} is added. The λ parameters balance the impact of two classification tasks. Note that we set loss for the *surprise* and *anticipation* categories to 0 during training as we deemed they were neither positive nor negative.

$$L_{multi} = \lambda_3 L_{emo} + \lambda_4 L_{pol} \tag{2}$$

$$L_{pol} = \{y^+ \mathrm{log} p^+ + (1-y^+)\mathrm{log}(1-p^+)\} + \{y^- \mathrm{log} p^- + (1-y^-)\mathrm{log}(1-p^-)\} \tag{3}$$

3 Experiments

3.1 Experimental Settings

To investigate the efficacy of the narrative context and knowledge about emotion categories in textual emotion classification, we conducted emotion classification experiments using the Story Commonsense dataset. We used the dev and test portions of the dataset for training and testing, respectively, as the dataset does not provide any annotations for the training portion. The dataset contains 13,004 and 11,859 sentence–character pairs for training and testing, respectively. The size of the preceding context was limited to 80 words. We used BERT-Large as the text encoder, whose hidden layer's dimensionality is 768. We fine-tuned the parameters during training by using the Adam optimizer. Classification accuracy is summarized with the micro-averaged precision/recall/F1-score, which is compatible with the results reported in (Rashkin et al., 2018).

Model	Precision	Recall	F1
BiLSTM*	20.3	30.4	24.3
CNN*	21.2	23.4	22.2
REN*	26.2	33.3	29.3
NPN*	22.0	37.3	27.7
`BL` (w/o char.)	59.2	56.9	58.1
`BL` (w char.)	**58.4**	**58.8**	**58.6**

Table 1: Comparison with the existing models reported in (Rashkin et al., 2018).

3.2 Major Results

Table 1 shows a comparison of our baseline (`BL`) performance with that of the existing complicated models presented in (Rashkin et al., 2018). The results show that the `BL` method outperformed the reported methods. The relatively large margins may be attributed to the simple yet adequate incorporation of narrative context supported by the powerful mechanism of the BERT-based text encoder. It can be noted that the performance gains achieved by incorporation of character specification are less prominent compared to that reported in (Rashkin et al., 2018). This outcome may again imply that the BERT-based encoder can be used to capture implicit clues of character identities.

#	Model	Character	Context	Precision	Recall	F1
1	`BL`	✓		58.4	58.8	58.6
2	`He`	✓		58.4	58.1	58.2
3	`Cooc`	✓		58.9	58.1	58.5
4	`Multi`	✓		59.4	57.9	58.7
5	`BL`	✓	✓	61.2	**61.5**	61.3
6	`Multi`	✓	✓	**62.6**	60.3	**61.4**

Table 2: Comparison among the classification methods.

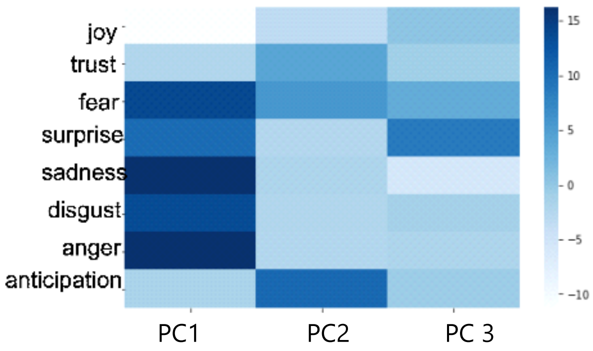


Figure 2: Averaged sentence vectors projected into a three-dimensional space by PCA.

Table 2 compares the classification methods, where the most significant outcome is the incorporation of narrative context, which is highly beneficial (lines #5 and #6). These two lines also show that the proposed multi-task method provided high accuracy but low recall, resulting in a slight increase in the F1-score. The same applies to the comparison of the results in #4 with those in #1 through #3. These results suggest that multi-tasking with positive/negative polarity classification would be beneficial but has scope for improvement. However, the results presented in #2 (a theoretical approach) and #3 (a data-driven approach) imply that Plutchik's Wheel of Emotions cannot be directly exploited in emotion classification. We should consider the internal cognitive mechanisms associated with human emotion.

3.3 Discussion: Reconsidering Plutchik's Wheel of Emotions

Figure 2 presents a heatmap of the averaged sentence vectors projected into a three-dimensional space by applying principle component analysis (PCA). To create an average vector for each emotion category, we extracted singleton sentences that are uniquely labeled with one category and used the [CLS] vectors obtained by the `BL` method. The following trends can be observed in the figure: (1) The first component discriminates positive emotions (*joy, trust, anticipation*) from negative emotions (*fear, surprise, sadness, disgust, anger*). In particular, *joy* can be clearly distinguished from any other category; (2) The second and third components jointly show that surprise-like categories, *anticipation* and *surprise*, are discriminated; and (3) The third component suggests the negative categories can be further grouped into two subsets: {*sadness, disgust, anger*} and {*surprise, fear*}. In summary, these results support the use of Plutchik's Wheel of Emotions to sufficiently capture the nature of human emotions.

4 Related Work

The Story Commonsense dataset provides the annotations for characters' mental states, which include emotions as well as human needs. Thus, this dataset provides an opportunity to investigate into a method to infer a character's underlying motivation to perform an action and the resulting emotional statuses. As these mental states are often implicit in text, the Story Commonsense dataset has facilitated research into the effective exploitation of commonsense knowledge (Paul and Frank, 2019).

The prediction of a character's emotional state is usually formulated as a classification problem (Gaonkar et al., 2020), but Bosselut and Choi (2019) framed it as a zero-shot commonsense question-answering problem, where the use of dynamically generated commonsense knowledge graphs was specifically focused. We thus highlight the former work in the following, which shares the research motivations with the present work.

Similar to the present work, Gaonkar et al. (2020) exploit information or knowledge obtained from the emotion category in the process of emotion classification, but in different ways. They firstly introduced emotion label embeddings which are made from emotion label name to incorporate relevant semantic information. They further devised a clever way to incorporate semantics from the "emotion-label sentences" by using a pre-trained BERT model. These devices improved both precision and recall. Besides, they modeled the transition of emotion by observing label correlations, which significantly contributed to improving the recall. As a result of these efforts, they successfully achieved state-of-the-art emotion classification performances on the Story Commonsense dataset. However, unlike the present work, they did not present any experimental discussions that try to explore the emotion classification system, in particular the wheel structure of Plutchik's basic emotion categories.

5 Concluding Remarks

The contributions of the present work are threefold. (1) Emotion classification accuracy can be considerably increased by using a BERT-based text encoder that properly incorporates narrative context; (2) Joint classification with positive/negative polarity would be promising for achieving higher classification accuracy; (3) It is partly demonstrated that the use of Plutchik's Wheel of Emotions sufficiently captures the nature of human emotions.

However, the experimental results suggest that a priori knowledge regarding the emotion category system has not been fully captured and incorporated into the classification model, showing that further investigations are necessary. There could be two ways to accomplish this. One way is to incorporate the ontologically organized knowledge about the cognitive structure of emotions with respect to events, agents, and objects (Ortony et al., 1990). Such ontological knowledge could be injected into a classification model, but it should be represented as a computable structure such as a knowledge graph. Another, presumably more direct, way is to exploit a set of inferential knowledge of everyday commonsense that could directly connect an event with the participants' emotional states. In this regard, we will seek an effective way to incorporate commonsense knowledge resources such as ATOMIC (Sap et al., 2019).

Acknowledgments

The present work was supported in part by JSPS KAKENHI Grants number 17H01831.

References

Nourah Alswaidan and Mohamed El Bachir Menai. 2020. A survey of state-of-the-art approaches for emotion recognition in text. *Knowledge and Information Systems*, pages 1 – 51.

Antoine Bosselut and Yejin Choi. 2019. Dynamic knowledge graph construction for zero-shot commonsense question answering.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Radhika Gaonkar, Heeyoung Kwon, Mohaddeseh Bastan, Niranjan Balasubramanian, and Nathanael Chambers. 2020. Modeling label semantics for predicting emotional reactions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4687–4692, Online, July. Association for Computational Linguistics.

Huihui He and Rui Xia. 2018. Joint binary neural network for multi-label learning with applications to emotion classification. In Min Zhang, Vincent Ng, Dongyan Zhao, Sujian Li, and Hongying Zan, editors, *Natural Language Processing and Chinese Computing*, pages 250–259, Cham. Springer International Publishing.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California, June. Association for Computational Linguistics.

Andrew Ortony, Gerald L. Clore, and Mark A. Foss. 1987. The referential structure of the affective lexicon. *Cognitive Science*, 11(3):341–364.

Andrew Ortony, Gerald L. Clore, and Allan Collins. 1990. *The Cognitive Structure of Emotions*. Cambridge University Press.

Debjit Paul and Anette Frank. 2019. Ranking and selecting multi-hop knowledge paths to better predict human needs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3671–3681, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In Robert Plutchik and Henry Kellerman, editors, *Theories of Emotion*, pages 3 – 33. Academic Press.

Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi. 2018. Modeling naive psychology of characters in simple commonsense stories. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2289–2299, Melbourne, Australia, July. Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3027–3035. AAAI Press.

Jianfei Yu, Luís Marujo, Jing Jiang, Pradeep Karuturi, and William Brendel. 2018. Improving multi-label emotion classification via sentiment classification with dual attention transfer network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1097–1102, Brussels, Belgium, October-November. Association for Computational Linguistics.

Deyu Zhou, Xuan Zhang, Yin Zhou, Quan Zhao, and Xin Geng. 2016. Emotion distribution learning from texts. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 638–647, Austin, Texas, November. Association for Computational Linguistics.