International Journal of
Computer & Software Engineering
Research Article Open Access

Construction and Expansion of Dictionary of Idiomatic Emotional Expressions and Idiomatic Emotional Expression Corpus

Kazuyuki Matsumoto*[1], Seiji Tsuchiya[2], Minoru Yoshida[1] and Kenji Kita[1]
[1]*Faculty of Science and Technology, Tokushima University, Japan*
[2]*Faculty of Science and Engineering, Department of Intelligent Information Engineering and Sciences, Doshisha University, Japan*

Abstract

Objective: In the study of sentiment estimation from language, methods focusing on words, phrases, sentence patterns, and sentence-final expressions have been proposed. However, it is difficult to deal with a wide variety of emotional expressions by only assigning emotions to words and phrases. In particular, it is difficult to analyze metaphorical expressions and idiomatic expressions on a word-by-word basis, and it is impossible to register all expressions in a dictionary because new expressions can be created by flexibly replacing words. However, it is difficult to determine the constraints on the words to be replaced, and not all expressions can be registered in the dictionary as sentence patterns.

Methods: In this paper, we construct a dictionary of idiomatic sentiment expressions, which contains idioms expressing emotions. In this paper, we construct a pseudo-emotional corpus by collecting utterances containing emotional idioms from social media and automatically assigning emotions expressed by the idioms.

Results: This corpus includes expressions other than idioms, and can be an effective resource for estimating emotions in sentences that do not contain idioms. In this study, we create an emotion estimation model for utterances based on the constructed corpus, and conduct evaluation experiments to explore the problems of the idiomatic emotion corpus. In addition, using the constructed sentiment corpus, we investigate how to expand the dictionary of sentiment expressions in idiomatic phrases by using deep learning methods.

Conclusion: Using the corpus of idiomatic sentiments constructed by the proposed method as training data, models with and without idioms were constructed by machine learning models. The results show that the F-values of all emotions with idioms exceed 0.8. On the other hand, when idioms were not included, the F-values tended to decrease overall. However, the F-values of emotions such as "shame" and "excitement" were around 0.7, indicating that the characteristics of emotional expressions other than idioms were expressed.

Publication History:

Received: November 25, 2021
Accepted: December 13, 2021
Published: December 15, 2021

Keywords:

Idiomatic expressions, Emotion dictionary, Emotion corpus

Introduction

Conventional methods have been proposed for estimating emotions from language, using words, phrases, sentence patterns, and sentence endings as cues, either dictionary-based or rule-based [1-6]. However, it is difficult to deal with a wide variety of emotions that are frequently expressed in colloquial sentences, such as unknown words and slang expressions, based on emotion labels assigned to words and phrases in advance. In particular, metaphorical expressions and idiomatic expressions have been considered difficult to analyze on a word-by-word basis in previous studies. Since new expressions can be created by replacing words, it is almost impossible to register all these newly created expressions in a dictionary. However, it is difficult to represent all the expressions as a fixed pattern or concept because the decision of words to be replaced depends on the context.

In this study, we focus on these idiomatic expressions and define the idioms that express emotions as emotional idioms. In this study, we focus on these idiomatic expressions and define them as emotional idioms. We collect utterances containing emotion idioms from social media and automatically assign emotions expressed by the idioms to them. The corpus thus constructed includes many expressions other than idioms. Therefore, we believe that this corpus can be used to estimate the sentiment of sentences that do not contain idioms.

In this paper, we create an emotion estimation model for utterances based on the automatically constructed emotion corpus. In this paper, we develop a model of sentiment estimation for utterances based on the automatically constructed sentiment corpus, and conduct evaluation experiments to explore the problems of constructing an emotion corpus using idioms. In addition, based on the constructed sentiment corpus, we investigate the extension of the idiomatic sentiment dictionary by using deep learning techniques.

Literature Survey

Citron et al. [7] compared idioms and literal expressions using fMRI and found that idioms tend to elicit more emotional resonance. However, their study did not mention the linguistic features of idioms.

Williams et al. [8] analyze the role of idioms in emotion analysis. They collected a dataset of 580 idioms by mapping them to emotions using a crowdsourcing approach. Based on this dataset, they created a corpus of idioms tagged with emotions in the context of the idioms, and observed a score of 64% in the three emotion polarities (positive,

***Corresponding Author:** Prof. Kazuyuki Matsumoto, Faculty of Science and Technology, Tokushima University, Japan; E-mail: daitoyouji@gmail.com

Citation: Matsumoto K, Tsuchiya S, Yoshida M, Kita K (2021) Construction and Expansion of Dictionary of Idiomatic Emotional Expressions and Idiomatic Emotional Expression Corpus. Int J Comput Softw Eng 6: 174. doi: https://doi.org/10.15344/2456-4451/2021/174

Citation: Matsumoto K, Tsuchiya S, Yoshida M, Kita K (2021) Construction and Expansion of Dictionary of Idiomatic Emotional Expressions and Idiomatic Emotional Expression Corpus. Int J Comput Softw Eng 6: 174. doi: https://doi.org/10.15344/2456-4451/2021/174

negative, neutral). However, their study did not consider more complex emotion categories such as joy, anger, and disgust.

Spasić [9] proposed rules for identifying idioms in text and a method for automatically constructing lexical semantic resources for emotion polarity classification. Their work also addressed emotion polarity, but did not mention complex emotion categories.

In another study, Briskilal et al. [10] proposed an ensemble model for identifying idioms and literals (literal meaning) by using fine-tuned BERT and RoBERTa. Although their main goal is to detect idioms, they show that language models such as BERT can be useful in acquiring semantic features of idioms through fine tuning.

Chen et al. [11] proposed a method for identifying Chinese metaphors by means of a neural network-based labeling method for metaphor identification, which is useful for improving the performance of word sense disambiguation and sentiment analysis tasks.

Shudo et al. [12] have constructed a dictionary of Japanese multi-word expressions (JDMWE). This dictionary contains not only typical idioms and clichés, but also quasi-idioms. The dictionary can be used in various situations because it covers variant notations and derived forms. However, because it is a dictionary describing a semantic conceptual system, it is not a resource that can be used directly for sentiment analysis.

Our research differs from the above-mentioned researches in that we aim to build a corpus of sentiment expressions including idiomatic phrases by creating a sentiment dictionary for Japanese idiomatic phrases, and utilize it as a resource for example collection and machine learning.

This research also differs from other research in that it considers complex emotion categories such as joy, anger, surprise, sadness, etc., rather than emotion polarity.

Methods

Most of the dictionaries studied and constructed in previous research were dictionaries of evaluative expressions that were categorized and registered by polarity of evaluation per word or phrase [13-15]. Many of these dictionaries are problematic in that they are not limited to phraseological expressions with special word-to-word connections such as idioms, and their comprehensiveness is low.

In this study, we will create an idiomatic phrase sentiment dictionary using the sentiment classification of idiomatic phrases in "Example Dictionary of Idioms" [16], which has relatively high coverage and detailed sentiment classification among existing dictionaries. The target idioms are those classified into 55 sub-items in the "idioms expressing senses and emotions" section. Based on the headings of these 55 sub-items, the emotional categories to be registered in the dictionary are determined. Although previous research on emotion polarity classification of multi-word expressions has been done by [17], their research was limited to noun-predicate combinations, and not to idiomatic phrases with special semantic relations between words, such as the one treated in this study.

In the case of multi-word expressions, the problem is that emotion polarity cannot be determined by a simple sum of the polarities of the constituent words. In this study, we address this problem by treating idiomatic phrases expressing emotions as if they were a single word, without considering the polarity of the constituent words.

We used 10 basic emotion categories (joy, anger, hate, sorrow, surprise, shame, fear, relief, like, excitement) defined in Nakamura's Dictionary of Emotional Expressions as the types of emotion categories to be assigned to idioms.

The number of idioms for each emotion category is shown in Table 1. Idioms were counted in a form that takes each conjugation into account. For example, the idiom "mimi ni sawaru" (English meaning: jar on the ear) can be registered as "mimi ni sawara" or "mimi ni sawari".

Emotion	# of Idioms	Emotion	# of Idioms
Joy	68	Excitement	28
Anger	103	Fear	179
Sorrow	154	Relief	114
Like	171	Hate	136
Surprise	77	Shame	44

Table 1: Number of idioms for each emotion category.

Construction of the Corpus

In this section, we describe a method for constructing a sentiment corpus by collecting tweets that contain idioms registered in the idiomatic sentiment dictionary. In this section, we describe a method to construct a sentiment corpus without manually annotating sentiment tags.

In the past, manual annotation was required to construct a sentiment corpus. The problem with manual annotation is that there is no standardized specification and the tagging criteria depends on the corpus builder. One of the biggest problems is the high cost of tagging. Even for tagging short sentences, it is necessary to understand the intent of the sentence and then understand the emotion. It is very difficult to build a large corpus while maintaining the quality.

In this study, we automatically extract sentences containing idioms expressing emotions from a large set of documents, and assign the emotions expressed by the idioms as the sentiments of the sentences. We aim to automate the construction of a large-scale emotion corpus by automatically extracting sentences containing idioms expressing emotions from a large set of documents and assigning the emotions expressed by the idioms as the sentiments of the sentences. For this purpose, we use relatively short sentences, such as Twitter tweets, because the target sentences need to be short (Figure 1).

In addition, there is a problem of cancellation by negative words. If a negative word appears before or after a negative word, the opposite emotion will be expressed. Depending on the usage of the idiomatic expression, the emotion of the idiom may not be the same as the emotion of the sentence.

However, in this paper, we have constructed a corpus that collects a large number of examples of idiomatic phrases, even if they contain some noise, and have developed a dictionary of idiomatic sentiments. In this paper, we aim to build a corpus of a large number of examples of idiomatic phrases, even if they contain some noise, and to assign emotions to those idiomatic phrases that are not registered in the idiomatic phrase sentiment dictionary and have not been assigned emotions.

Citation: Matsumoto K, Tsuchiya S, Yoshida M, Kita K (2021) Construction and Expansion of Dictionary of Idiomatic Emotional Expressions and Idiomatic Emotional Expression Corpus. Int J Comput Softw Eng 6: 174. doi: https://doi.org/10.15344/2456-4451/2021/174

Emotion Estimation from Idioms

Next, we describe a method for estimating the sentiment of idioms. We use the Twitter API to collect tweets that contain idioms that are not registered in the idiomatic sentiment dictionary constructed in this study, and estimate the sentiment using the sentiment estimation model based on the idiomatic sentiment corpus.

Figure 2 shows an overview of the emotion estimation model for idioms. First, we extract features from the tweets containing the unknown idioms collected from Twitter. We use Bag of Words for the features and TF-IDF [18] for the feature weighting. The purpose of using the Bag of Words feature is to clarify the problem of the corpus by using a feature that is easier to analyze than a word variance representation.

We train an emotion estimation model based on logistic regression on the features obtained from the "Emotion Tagged Idiom Corpus" without idioms and only from the idioms. Using this model, we can tag unknown idioms with emotions. By collecting multiple tweets

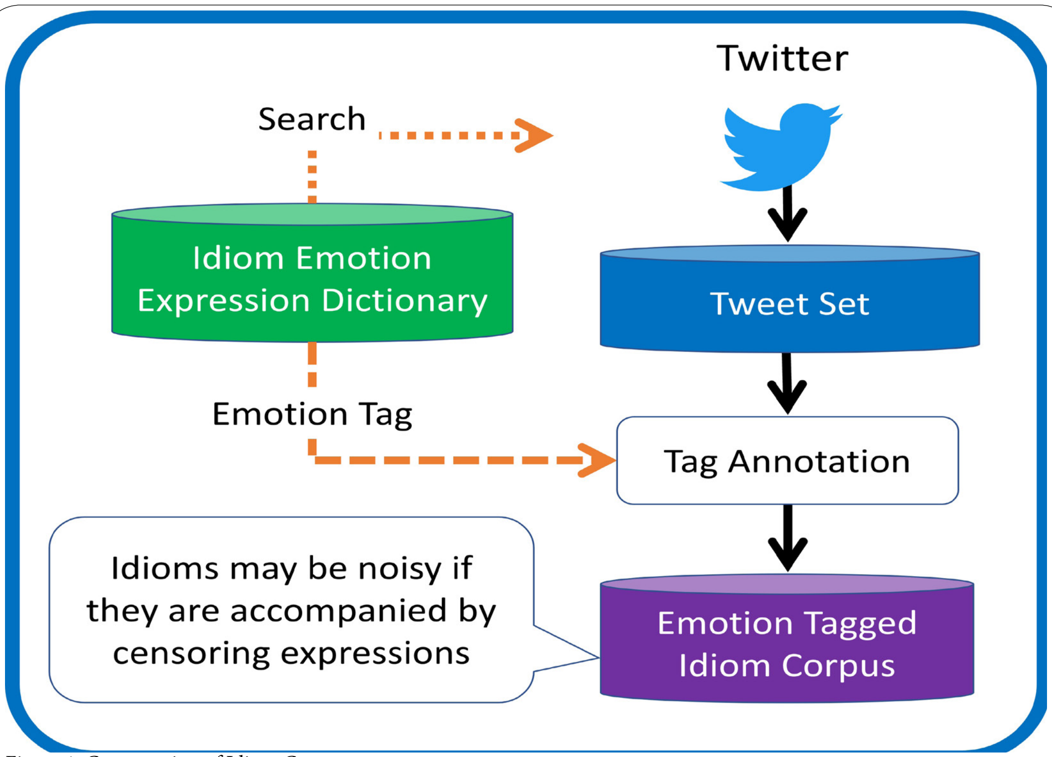

Figure 1: Construction of Idiom Corpus.

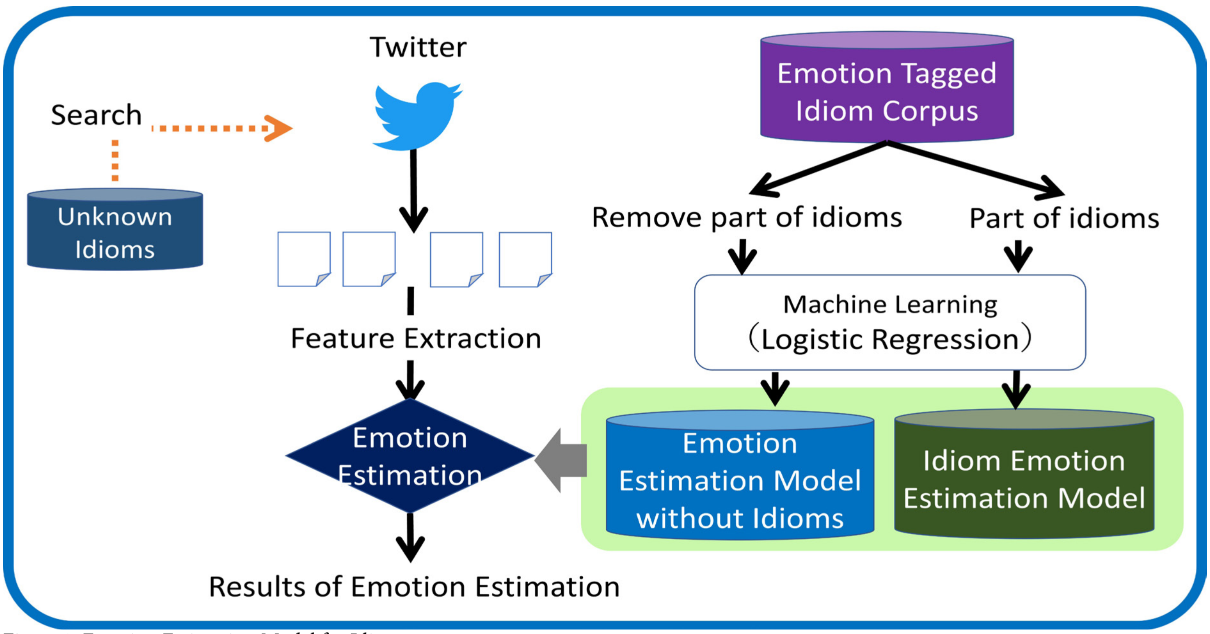

Figure 2: Emotion Estimation Model for Idioms.

for each idiom, we can obtain multiple sentiment estimation results, which are then integrated to determine the final sentiment type.

There are many examples of idiomatic phrases, and there are some idiomatic phrases in which the same emotion is estimated for more than a certain number of tweets, even if the contexts are different. In such cases, it is likely that the idiom itself, rather than the whole sentence, expresses the emotion strongly, so it may be possible to add the idiom to the dictionary of emotional expressions.

When adding an unknown idiom to the dictionary, we register the finally determined emotion category along with the idiom, and then add only the example sentences estimated as the emotion category to the idiomatic sentiment corpus to exclude noise and prevent quality deterioration.

Preliminary experiments

As a preliminary experiment, we will compare the accuracy of the sentiment estimation model with and without excluding the corresponding part of the idiom from the training data. To construct the emotion estimation model, we use logistic regression as a machine learning method. We used the Logistic Regression class[1] of scikit-learn as a library for logistic regression. Default values are used for training parameters.

To evaluate the emotion estimation model, we used the 10-split cross-validation method, and calculated the F-values of the reproduction rate and the fit rate for each emotion.

The data used were randomly balanced so that the number of sentences in each emotion category was 100 in each test set to avoid differences among emotion categories, and a total of 1000 sentences were evaluated for each emotion category.

[1]https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

The results of the experiment with the idioms left out and the F-values for each emotional category with the idioms excluded are shown in Figure 3.

The results of the experiment showed that the reproduction and fit rates of more than 90% were obtained for almost all emotion categories when the examples included idioms directly in the sentences, which is reasonable considering that idioms are always included in each sentence.

On the other hand, the results for the examples in which idioms were excluded from the sentences showed that the F-values for some emotion categories were relatively high.

Figure 4 shows the confusion matrix between the emotion estimation results and the correct answer when idioms are excluded from the tweets.

The results show that the correct answer rates for "fear," "sorrow," and "relief" are low, and they are often estimated as emotions with different emotional polarity.

The reason for this is that the negative expression before and after the idiom cancels out the positive/negative polarity, or when a tweet contains multiple sentences. The reason for this may be the cancellation of the negative expression before and after the idiom, or the reversal of the positive/negative polarity when a tweet contains multiple sentences. In addition, when the idioms themselves express emotions, there is still a problem that the features cannot be extracted well when the idioms are excluded.

Here, we analyzed the following points that may be the cause of the reversal of emotional polarity described above.

1. Number of occurrences of negative expressions before and after the idiom
2. Ratio of emotion polarity in words other than idioms

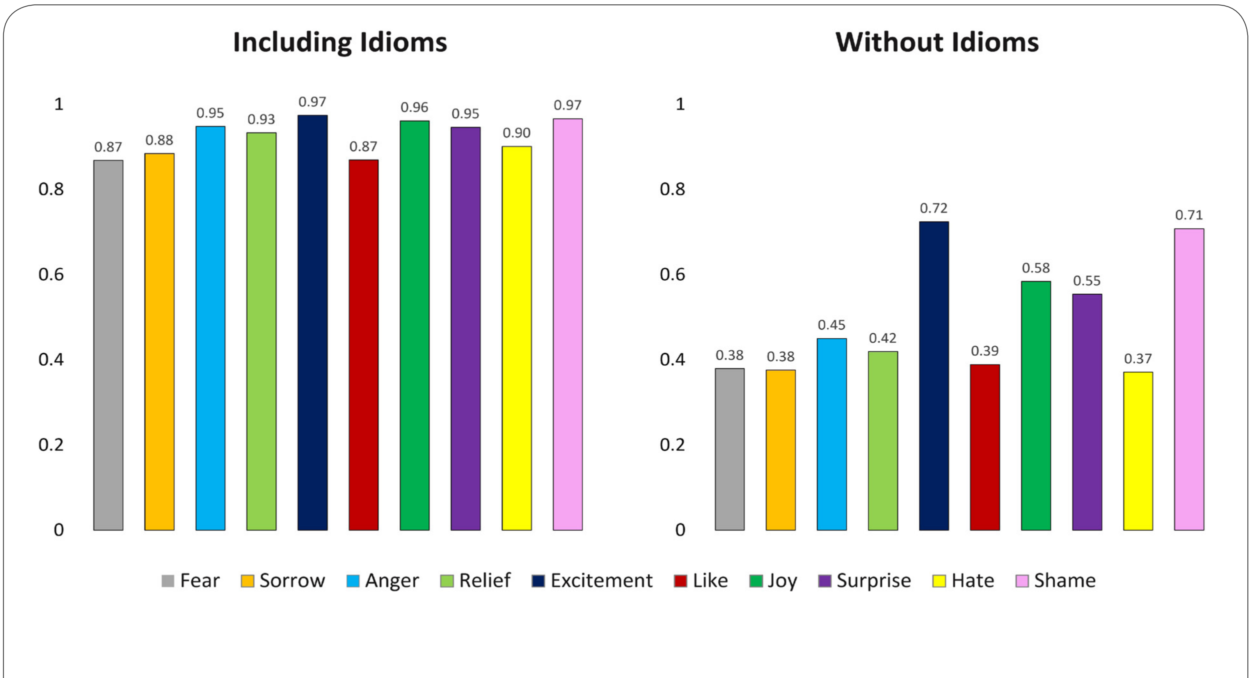

Figure 3: Result of Preliminary Experiments (F-score).

Citation: Matsumoto K, Tsuchiya S, Yoshida M, Kita K (2021) Construction and Expansion of Dictionary of Idiomatic Emotional Expressions and Idiomatic Emotional Expression Corpus. Int J Comput Softw Eng 6: 174. doi: https://doi.org/10.15344/2456-4451/2021/174

In this analysis, the target negative expressions were limited to "*nai*", "*nai*", "*nu*", and "*zu*", and only those occurring within the backward three words of the idiom were counted. The reason for this is that although there are other negative expressions, these four patterns are the most commonly used, and the other expressions are often mis-segmented in morphological analysis, making them difficult to detect.

Some of the results are shown in Table 2.

Idiom	f_{idiom}	f_{neg}	Breakdown	*rate*
Fukaku wo tora	34	34	*nu*:34	1
Ki nikakara	4	4	*nai*:4	1
Ma ga sasa	4	4	*nai*:4	1
Kata ga kora	324	322	*nai*:322	0.99
Kokoro wo yurusa	716	705	*nu*:676, *nai*:28 , *zu*:1	0.98
Ki wo otosa	117	113	*nai*:82, *zu*:29, *nu*:2	0.97
Kokoronikakeru	22	21	*zu*:21	0.95

Table 2: Analysis Result of Negative Expression of Idiom.

For example, the negative expression "*nai*" following "*kini suru*" appears in the form "*kini suru kotoha nai*", which cancels out the emotion "*fear*" expressed by "*kini suru*". In this case, the emotion expressed by "concern", "fear", is canceled out.

Therefore, it is highly likely that the sentences in which idioms are uttered with backward negative expressions are not correctly assigned to the sentences' emotions.

Table 3 shows some of the results of the ratio of emotion polarity in words other than idioms.pn_{idiom} is the emotional polarity of the idiom, and p_{total}, n_{total} are the number of words with positive/negative emotional polarity that co-occur in the same sentence as the idiom. The C_{match} indicates the number of co-occurring words whose emotional polarity matches the emotional polarity of the idiom.

Idiom	pn_{idiom}	p_{total}	n_{total}	C_{match}
Mega nai	1	1372	2161	485
Meno hoyou	1	906	929	450
Chiga sawagu	1	591	419	407
Waraiga tomaranai	1	741	2095	383
YuYu jiteki	1	581	849	323
Mega hanasenai	1	648	1216	256
Kiga aru	1	802	2035	250
Kiwo tsukeru	1	938	4098	235
Kokoro yukumade	1	604	782	214
Kiwo kubaru	1	650	1206	199

Table 3: Aggregate results of co-occurring emotional polarity (partial).

The larger the value of C_{match}, the higher the reliability of the sentiment of the given sentence. It is considered that these idioms are useful for sentiment estimation.

In addition, the sign of the sum of the emotion polarity values of the co-occurring words is different from the polarity of the idiom. Based on the results of this analysis, it is possible to construct a reliable sentiment corpus by excluding sentences whose length exceeds a certain

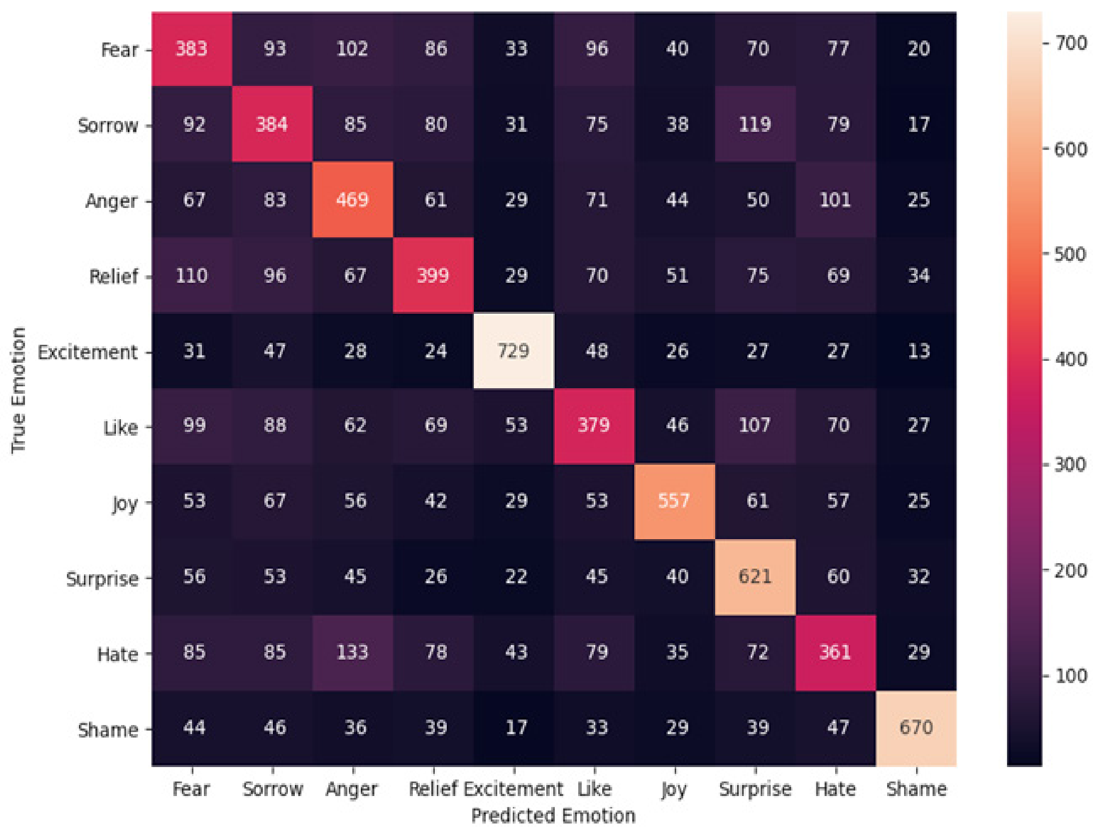

Figure 4: Confusion Matrix of Emotion Estimation

Citation: Matsumoto K, Tsuchiya S, Yoshida M, Kita K (2021) Construction and Expansion of Dictionary of Idiomatic Emotional Expressions and Idiomatic Emotional Expression Corpus. Int J Comput Softw Eng 6: 174. doi: https://doi.org/10.15344/2456-4451/2021/174

level and sentences whose sign is different from that of idioms. In addition, sentences that have a certain bias in the ratio of emotion polarity of words other than idioms are not included in the training data. In addition, we attempted to construct a model that can estimate emotion by not including sentences with a certain length and sentences with a certain bias in the ratio of emotion polarity of words other than idioms.

Evaluation experiments

Using the emotion estimation model learned from the corpus of idiomatic sentiments, we can estimate sentences containing idioms and register them in the dictionary. We conducted an experiment to estimate sentences containing idioms and register them in the dictionary. The results of the experiment are manually checked and evaluated on a three-point scale (T: Correct, N: May be correct, F: Incorrect) to determine the validity of the registered idioms and their emotional categories.

The following is the equation (1), which calculates the weight of each emotional category from the estimation results.

$$weight_{e_i^x} = \frac{freq_{e_i^x}}{\sum_{j=1}^{N} freq_{e_j^x}} \quad (1)$$

$freq_{e_i^x}$, $freq_{e_j^x}$ indicates the frequency of assigning the emotion category e_i^x , e_j^x to the examples of an idiom *x*. The *N* indicates the number of types of emotion categories.

In this study, only those sentiment categories with more than 100 examples and $weight_{e_i^x}$ greater than 0.2 were output as results and evaluated. Out of a total of 29,801 example sentences (791 idioms), 34 idioms satisfied the above conditions. The 34 idioms, their assigned emotions, and the evaluation results are shown in Table 4.

The value of $weight_{e_i^x}$ is appended after each emotion category. Even though the contexts of the sentences differed from each other, the co-occurrence of the features necessary for emotion estimation resulted in reasonable estimation results.

However, tweets sometimes contain multiple sentences, and the results were erroneous in such cases, and also for idioms that tend

Idiom	Outputs	Eval.	Idiom	Outputs	Eval.
"Tewo awaseru" put one's hands together	Relief(0.209), Like(0.203)	T	"Te ga nai" have no choice	Sorrow (0.243)	T
"Tega hanareru" leave one's hands	Relief (0.219)	T	"Iki wo tsuku" draw a breath	Sorrow (0.228)	F
"Munega tsumaru" "get a catch in the throat"	Sorrow (0.242)	T	"Doro wo nuru" spot	Shame (0.288)	T
"Te wo hirogeru" extend (one's) business	Fear (0.211)	F	"Me wo nuku" get one over on	Surprise (0.236)	T
"Te wo utsu" shake on it	Relief (0.203)	N	"Me wo oou" hideous	Surprise (0.309)	T
"Te ni noru" be deceived [fooled] by someone's trick	Surprise (0.208)	F	"Amai shiru wo suu" be onto a good thing	Hate (0.227)	T
"Iji wo haru" obstinate	Fear (0.217)	N	"Me ga sameru" awake from illusion	Relief (0.202)	N
"Me wo samasu" snap out of it	Relief (0.208)	N	"Iki wo ireru" stop for a breath	Relief (0.209)	T
"Ato ni suru" desert	Sorrow (0.208)	N	"Iroke yori kuike" dumpling rather than flowers	Sorrow (0.214)	F
"Baton wo watasu" pass the torch	Sorrow (0.201)	N	"Iki wo nuku" let go of one's breath	Fear (0.220)	F
"Atama wo hiyasu" cool out	Sorrow (0.231)	N	"Ate ga hazureru" be disappointed of one's purpose	Fear (0.236)	T
"Te ga todoku" reach	Relief (0.207)	T	"Inochi nosentaku" get away from it all	Sorrow (0.256)	F
"Anaume wo suru" make up for	Sorrow (0.208)	F	"Mi ni tsukeru" adopt	Like (0.292), Relief (0.260)	T
"Netsu ga sameru" fever goes down	Fear (0.228)	N	"Ate ni suru" depend	Fear (0.208)	F
"Ashi ga omoi" have leaden feet	Sorrow (0.221)	N	"Nami ni noru" catch a wave	Like (0.226)	T
"Atama ga hikui" humble	Sorrow (0.243)	N	"Ikkan no owari" the end of the line	Sorrow (0.207)	T
"Ashi wo ireru" get one's feet wet in	Sorrow (0.205)	F	"Atokata mo nai" bugger off	Sorrow (0.258)	T

Table 4: A Part of Result of Idiom emotion estimation.

to be used with negative expressions, such as "*Iki wo tsuku hima mo nai* (I don't have time to breathe out)" or that partially match other expressions, such as "*Iki wo tsuku*" in addition, the evaluation result was "N".

In addition, many of the idioms that received a "N" rating had ambiguities, such as expressing different emotions depending on who was the subject of the action indicated by the idiom.

In addition, the results of this experiment revealed that there were cases where idioms did not express emotions. The results of this experiment showed that there were some cases in which the idioms did not express emotions, and the emotion categories of the examples tended to be dispersed.

In this experiment, the idioms used for sentiment estimation mainly consisted of three morphemes, but the number of morphemes in an idiom can be relatively large if the sentiment is determined by the idiom itself. Therefore, in order to register an idiom as an emotional expression in a dictionary, it is considered appropriate to estimate the emotion as a single expression, including the context of the idiom, and register it in the dictionary.

Discussions and analysis

As a result of our experiments, we found that the sentiment of idioms is ambiguous, and it is difficult to estimate the sentiment of an idiom by itself without considering the context. It is necessary to develop a highly accurate sentiment estimation method for sentences containing idioms, rather than for idioms alone.

One of the problems with the idiomatic sentiment corpus constructed in this study is that idiomatic phrases are not used with the meaning of an idiom in a sentence. In this case, if we use the idioms in the idiomatic sentiment dictionary as features for sentiment estimation, it will lead to false estimation. Therefore, it is necessary to improve the idiomatic sentiment dictionary by organizing contextual information and registering sentiments in different cases.

In order to investigate the features of an idiom by itself, we visualize the features of BERT (bidirectional encoder representations form transformer) [19] for idioms. The visualization is based on the acquisition of distributed representations from idioms based on the pre-trained SentenceBERT [20] model, and dimensional compression using a self-encoder and UMAP [21,22].

The visualization procedure is shown below.

1. The idioms are transformed into 768-dimensional feature vector with pre-trained SentenceBERT model.
2. The 786-dimensional features vectors are trained with a self-autoencoder composed of Feed Forward Neural Networks (FFNNs), and convert the idioms into 100-dimensional feature vector using an encoder (output from one layer before the output layer). The 100-dimensional feature vectors are standardized by Standard Scaler class[2], and compressed into two dimensions using UMAP algorithm.
3. The 100-dimensional feature vectors are standardized, dimensionally compressed by UMAP to two dimensions, and then plotted on a two-dimensional plane.
4. The two-dimensional feature vectors created in the previous step are divided by into five clusters using K-means, an unsupervised clustering method.

Figure 5 shows the architecture of FFNNs used for the autoencoder.

The visualization results are shown in Figure 6. The colors of the points in the figure indicate the type of emotion, and the shapes of the markers indicate the clusters. From this figure, there are no clusters that are biased toward any particular emotion.

SentenceBERT is a method for sentence clustering. Therefore, in the pre-training process, sentence pairs selected from the same topic or corpus of the same type are mainly trained as examples of similar

[2]https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

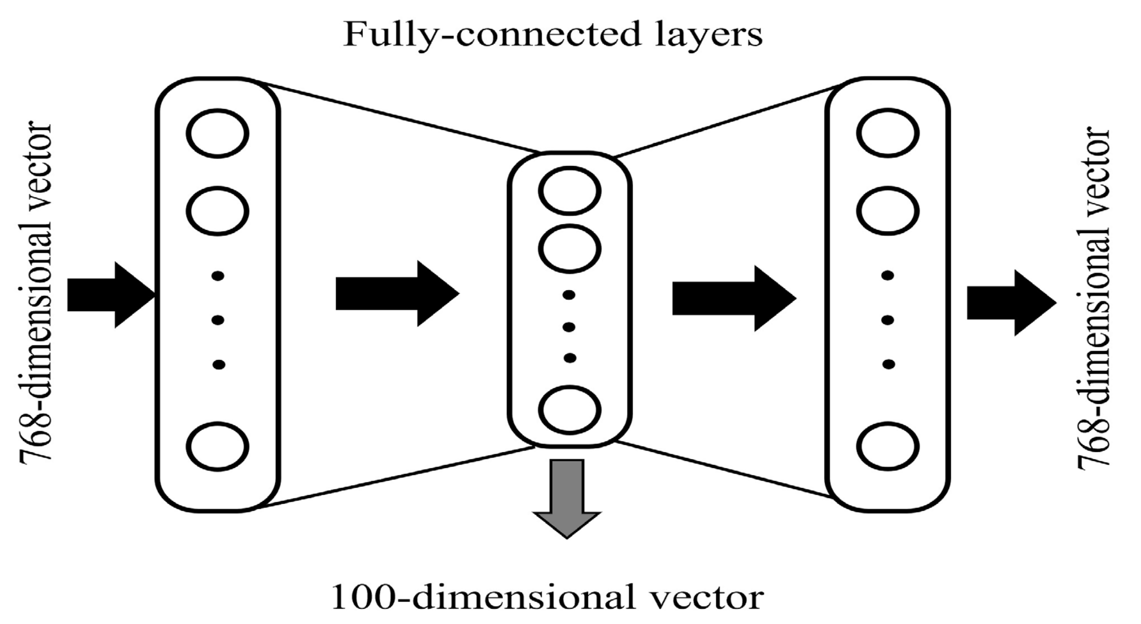

Figure 5: Neural Network Architecture for Self-autoencoder.

Citation: Matsumoto K, Tsuchiya S, Yoshida M, Kita K (2021) Construction and Expansion of Dictionary of Idiomatic Emotional Expressions and Idiomatic Emotional Expression Corpus. Int J Comput Softw Eng 6: 174. doi: https://doi.org/10.15344/2456-4451/2021/174

sentence pairs, while sentence pairs selected from a different document corpus or topic are trained as examples of dissimilar sentence pairs. This suggests that the system is likely to convert feature expressions to feature expressions that take into account similarities other than meaning, such as sentence style and sentence length.

In general, the corpus does not contain many cases where an idiom itself is used as a sentence. By using SentenceBERT, the context of each word in the idiom is combined, and thus the original meaning of the idiom may not be expressed. Therefore, we can see that it is difficult to obtain the features of idioms from the idioms themselves in a BERT-based model of feature extraction like SentenceBERT.

We will also visualize the distributed representation of idiomatic phrases using the method described above. In addition, we visualize the concatenation or composite vector (vector sum) with the distributed representation of the idiom itself. Figures 7(a) to 7(c) show the visualization results.

These results indicate that the introduction of distributed expressions of misinterpretations creates a more diverse cluster than the use of idioms alone. However, there is a lack of cohesion for each emotion, and it is possible that the misinterpretations themselves do not contain emotional information. Thus, the misinterpretations themselves may not contain emotional information. This suggests that the misinterpretations alone cannot explain all the uses of the idioms, and that it is necessary to understand the emotions more precisely by looking at the examples.

Conclusions

In this paper, we describe a dictionary of idiomatic sentiment expressions, the construction of a sentiment corpus using the dictionary, and the extension of the dictionary. In the evaluation experiment, we estimated the sentiment of idiomatic phrases, and evaluated only those with more than 100 example sentences and sentiment categories with weights of 0.2 or more. As a result, we found that about half of the idioms we evaluated showed reasonable sentiment estimation results. On the other hand, it is difficult to assign a sentiment to an idiom by itself, and it is necessary to construct a dictionary that includes the context of the idiom.

In addition, the SentenceBERT model, which is an application of BERT that has been used in the field of natural language processing in recent years, was used to convert idioms and misinterpretations of idioms into distributed expressions, and visualization and clustering were performed. Therefore, in order to use idioms effectively in tasks such as sentiment analysis, it is necessary to collect as many examples as possible and classify them in a systematic way.

In the future, we plan to conduct research not only on example sentences of idioms in tweets, but also on example sentences in other media (news articles, bulletin boards).

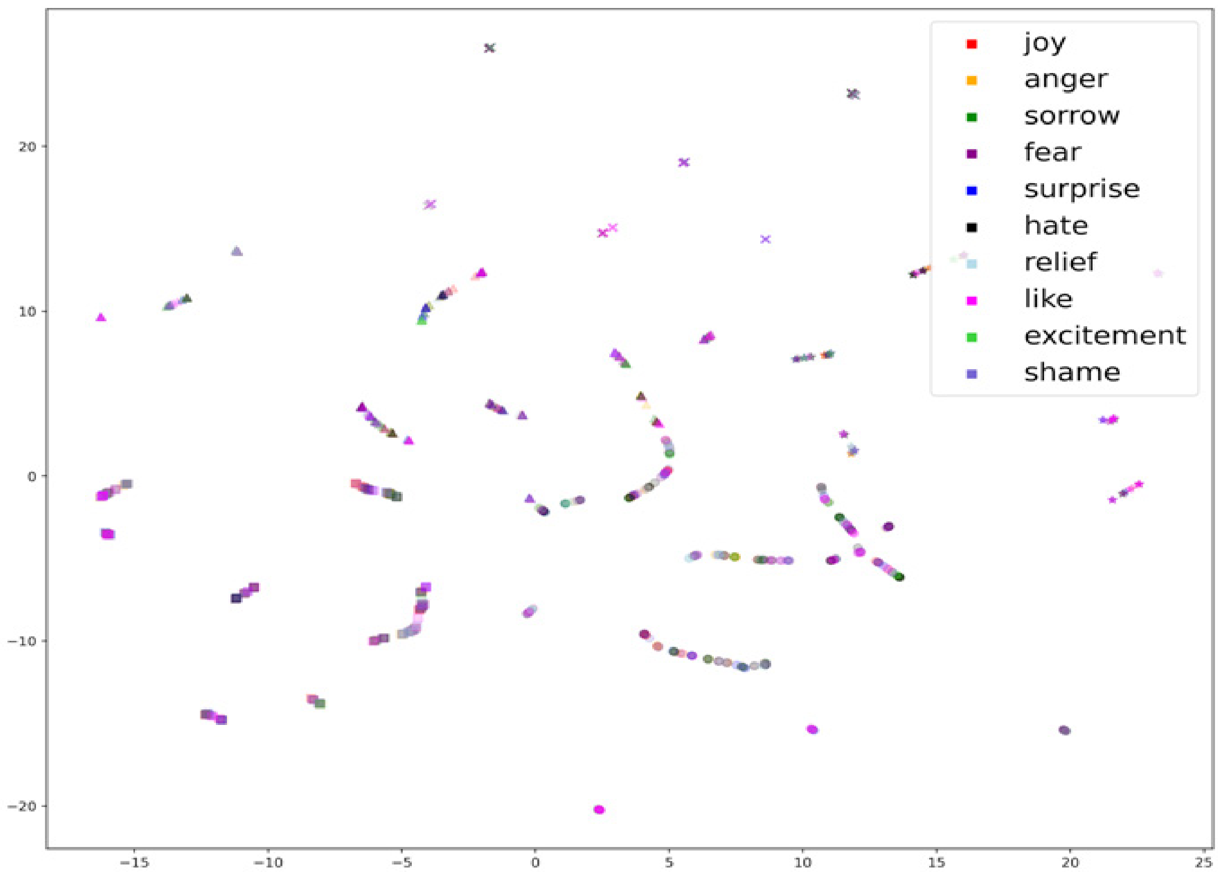

Figure 6: Visualizing of SentenceBERT Vector of Idioms.

Citation: Matsumoto K, Tsuchiya S, Yoshida M, Kita K (2021) Construction and Expansion of Dictionary of Idiomatic Emotional Expressions and Idiomatic Emotional Expression Corpus. Int J Comput Softw Eng 6: 174. doi: https://doi.org/10.15344/2456-4451/2021/174

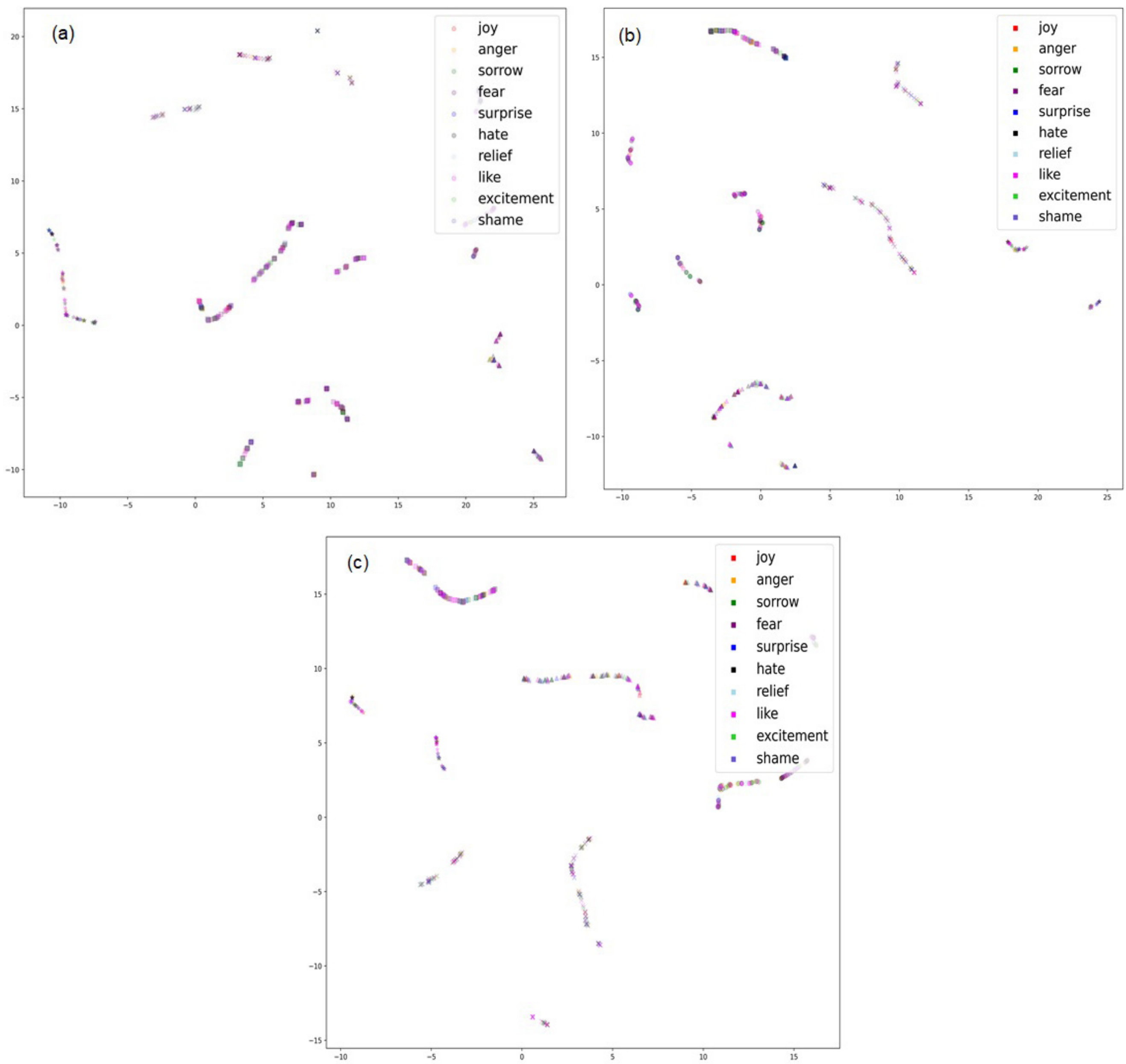

Figure 7: (a) Visualization of SentenceBERT Vector from Interpretation of Idioms. (b)Vector Visualization that Concatenates the SentenceBERT Vectors of Idiom Itself and Interpretation of Idiom. (c) Vector Visualization that Synthesizes the SentenceBERT Vector of Idiom Itself and Interpretation of Idiom.

Competing Interests

The author declare that there is no competing interests regarding the publication of this article.

Funding

This work was supported by the 2021 SCAT Research Grant and JSPS KAKENHI Grant NumberJP20K12027.

References

1. Matsumoto K, Ren F (2011) Estimation of word emotions based on part of speech and positional information. Computers in Human Behavior 27: 1553-1564.
2. Minato J, Matsumoto K, Ren F, Tsuchiya S, Kuroiwa S, et al. (2008) Evaluation of emotion estimation methods based on statistic features of emotion tagged corpus. International Journal of Innovative Computing, Information and Control 4: 1931-1941.
3. Ren F, Matsumoto K (2016) Semi-automatic creation of youth slang corpus and its application to affective computing. IEEE Transactions on Affective Computing.
4. Quan C, Ren F (2016) Weighted high-order hidden Markov models for compound emotions recognition in text. Information Sciences 329: 581-596.
5. Ren F, Wu Y (2013) Predicting user-topic opinions in Twitter with social and topical context. IEEE Transactions on Affective Computing 4: 412-424.
6. Ren F, Quan C, Matsumoto K (2013) Enriching mental engineering. International Journal of Innovative Computing, Information and Control 9: 3271-3286.
7. Citron F, Cacciari C, Funcke J, Thsu C, Jacobs A (2019) Idiomatic expressions evoke stronger emotional responses in the brain than literal sentences. Neuropsychologia 131: 233-248.
8. Williams L, Bannister C, Ayllon MA, Preece A, Spasic I (2015) The role of idioms in sentiment analysis. Expert Systems with Applications 42: 7375-7385.
9. Spasic I, Williams L, Buerki A (2020) Idiom-based features in sentiment analysis: cutting the Gordian Knot. IEEE Transactions on Affective Computing.
10. Briskilal J, Subalalitha CN (2022) An ensemble model for classifying idioms and literal texts using BERT and RoBERTa. Information Processing & Management.
11. Chen X, Hai Z, Wang S, Li D, Wang C, Luan H (2021) Metaphor identification: A contextual inconsistency based neural sequence labeling approach. Neurocomputing 428: 268-279.
12. Shudo K, Tanabe T (2010) JDMWE: A Japanese dictionary of multi-word expressions. Journal of Natural Language Processing 17: 51-74.
13. Takamura H, Inui K, Okumura M (2006) Extracting semantic orientations using spin model. IPSJ journal 47: 627-637.
14. Nakamura A (1993) Emotional expression dictionary. Tokyodo publishers.
15. Sano M (2011) Classification of evaluative expressions in Japanese: an Appraisal perspective. IEICE technical report 110: 19-24.
16. Inoue M (1990) Reikai Kanyoku Jiten. Soutakusha.
17. Takamura H, Inui K, Okumura M (2006) Latent variable models for semantic orientations of phrases. IPSJ journal 47: 3021-3031.
18. Salton G, McGill MJ (1983) Introduction to modern information retrieval. McGraw-Hill Book Company.
19. Devlin J, Chang MW, Lee K, Toutanova K (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 1: 4171-4186.
20. Reimers N, Gurevych I (2019) Sentence-BERT: sentence embeddings using Siamese BERT-networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).
21. McInnes L, Healy J, Saul N, Großberger L (2018) UMAP: uniform manifold approximation and projection. The Journal of Open Source Software 3: 861.
22. McInnes L, Healy J (2018) UMAP: uniform manifold approximation and projection for dimension reduction. ArXiv e-prints.