ASM Sc. J., 13, Special Issue 3, 2020 for CIIDT2018, 11-16

Data Augmentation for Lyrics Emotion Estimation

Asaki Kataoka[1*], Manabu Sasayama[1] and Kazuyuki Matsumoto[2]

[1]*Department of Information Engineering, National Institute of Technology, Kagawa College, Japan*

[2]*Graduate School of Technology, Industrial and Social Sciences, Tokushima University, Japan*

Lyrics emotion estimation can allow us to realise song retrieval systems and song recommendation systems which are based on not only text retrieval nor melody matching but also emotions in lyrics or transitions of emotions within lyrics of a whole song. This requires lyrics emotion corpora of phrase. However, it is difficult to build large scale lyrics emotion corpora because emotions are labelled manually. In this paper, we propose a method to augment lyrics emotion corpora. As a result, we augmented a corpus consisting of 366 phrases into a larger corpus consisting of 2145 phrases. We also evaluate the proposed method using 2 convolutional neural networks trained on original corpus and augmented corpus respectively. We define the target emotion classes as Joy, Love, Anger, Sorrow and Anxiety. Mean accuracy of the model trained on the augmented corpus was 75.9% whilst the model trained on the original corpus performed 70.7%.

Keywords: lyrics emotion estimation; text data augmentation; word2vec; convolutional neural network

I. INTRODUCTION

Lyrics emotion estimation can realise song retrieval systems and song recommendation systems which are based on the moods of songs; songs can be retrieved by ambiguous queries such as “joyful songs” and “sad songs”. Additionally, automation of matching an image to a relevant song, which can possibly be applied to social networking as Xuelong et al. worked on, can be efficiently achieved by lyrics emotion estimation (Xuelong, Di, & Xiaoqiang, 2017). The traditional approaches have used melodies or texts as queries (Chen & Chen, 1998). However, since they would provide only songs matching with the queries, they can result unexpected songs if the given queries are vaguely-remembered lyrics. Moreover, they can provide too many retrieval results given short texts as queries. For song recommendations, the traditional approaches have been analysing tendencies of users’ preferences by having users give titles of their favourite songs or by making users rate songs that the systems show (Beall, 2008). With this approach, the systems may always provide similar songs or songs whose lyrics are different from users’ preferences even though melodies are not. When lyrics emotion estimation is achieved, refinement of the retrieval results by emotions on song retrievals and emotion specification on song recommendation systems becomes possible.

In this work, we carry out an experimental lyrics emotion estimations using deep neural networks. Training the neural networks requires a large dataset. However, currently available lyrics emotion corpus is small in scale. Additionally, it is difficult to build a large scaled lyrics emotion corpus because emotion labelling is done manually. There is a study in which data augmentation resulted in accuracy increase when training a deep neural network on a small scaled corpus (Silfverberg et al. 2017). In this paper, we propose a method to augment available lyrics emotion corpora using word replacement. In the section 2, related works are introduced, and the proposed method will be described in the section 3. The target corpus and tools used in our experiments are introduced in the section 4. In the section 5, an experiment of augmentation of the lyrics emotion corpus using the proposed method is elaborated. In order to evaluate the proposed method, another experiment, in a cross-validation setup, on performances of two convolutional neural networks (CNNs) respectively trained on the original lyrics emotion corpus and the augmented

*Corresponding author’s e-mail: asaki.kataoka@chiba-u.jp

corpus is shown in the section 6. The section 7 is the conclusion of this work.

II. RELATED WORKS

There are a few studies on sentence emotion estimation. Matsumoto and Sasayama performed lyrics emotion estimation by extracting word embeddings from existing lyrics corpus (Matsumoto & Sasayama, 2018). Their approach involved simple k-measure method, not a complex algorithm such as neural networks with which analysing the results could be complicated. They proposed an efficient way to extract word embeddings which can purely applied to a linear method of emotion estimation. In this paper, the proposed method involves a general word embedding, which is not specified to song lyrics, and evaluate the advantages using CNNs. Homma et al. proposed a method to estimate emotions that occur by reading posts on Twitter. They trained a deep neural network on dataset of posts and the readers' emotions (Homma & Hagiwara, 2013). They reported that deep neural networks performed the task better than baseline methods. Muhammad and Ungar built a large dataset which consists of a lot of phrases tagged emotions using hashtags and they performed estimation on writers' emotions (Muhammad & Ungar, 2017). On the other hand, Zhao and Gao used CNN for emotion detection in daily conversations (Zhao & Gao, 2017). They reported that a deeper CNN performed higher accuracies on emotion detection.

One of the significant studies on text data augmentation is reported by Nishimoto et al (Nishimoto, Noji & Matsumoto, 2017). They augmented a corpus which consists of sentences about some words such as computers and restaurants, and they evaluated changes in accuracies on positive/negative classification. Our target of data augmentation in this work is lyrics. Wang et al, used CNN for image recognition. They reported that accuracy changed by noising the training dataset (Wang & Perez, 2017). We use CNN for evaluating the proposed method.

III. METHODOLOGY

A. Corpus Augmentation

We propose a method to augment lyrics emotion corpora in this section. The proposed method replaces nouns in each phrase with other words of which, similarities calculated involving word distributed representation, are high. In this article, a line on a lyrics card is referred to as a phrase. We focus exclusively on nouns as the target of replacement in order to avoid changes in emotions caused by replacing adjectives or verbs. To quantitate similarity of an arbitrary noun to the noun in the original phrase, we defined *replacement score* which involves both cosine similarity between two words represented as word vectors, and a 3-gram corpus. The procedure of generating a new phrase is:

1. Morphological analysis of the input phrase using MeCab
2. Acquisition of words with neighbour vectors (hereinafter called *similar words*) to the vector of the target words using Word2Vec
3. Calculation of replacement score of each similar word acquired in 2
4. Replacement with top-3 similar words in the replacement scores

Since a noun in the original phrase is to be replaced with top-3 similar words, this method is able to generate three times more phrases than the number of nouns in the phrase.

B. Acquisition of Similar Words using Word2Vec

Similar words are acquired using word distributed representations which *Word2Vec* generates. Word2Vec generates a word vector space on the basis of the distributional hypothesis of word. The distributional hypothesis claims an arbitrary pair of words having relatively large number of shared co-occurring words must be represented as neighbour vectors. Therefore, original words are possibly replaced with antonyms in case the original words and the antonyms appear in the same contexts. Because, especially in adjectives, adverbs, and verbs, massive changes in emotions are often caused from such synonym-antonym disruption, this method limits the target of replacement into only nouns. Such massive changes in emotions require re-labelling on the generated phrases, resulting in vanishment of the advantage of automated data augmentation. The measurement of similarity between a pair of word vectors is cosine similarity.

C. Calculation of Replacement Score

In Japanese Language, simple replacement of nouns in a phrase easily results in grammatically or semantically unnatural phrases. Since the target of this work is emotion estimation, generating phrases as natural as possible is

preferred. Hence, on each of 100 similar words suggested by Word2Vec, this method calculates a replacement score which measures how natural the generated phrases would be. The target noun in the original phrase is referred to as a *pre-replacement word,* and each suggested similar word is referred to as a *post-replacement word*. New phrases are generated by replacing the pre-replacement word in the original phrase with one of the corresponding top-3 post-replacement words. The replacement score is calculated by the equation below:

$$S = sim \cdot (\log_e \frac{k_1 + k_2}{2} + 1),$$

where *sim* corresponds to the cosine similarity between the post-replacement word and the pre-replacement word calculated by Word2Vec. k_1 is the number of a 3-gram pattern in which the first word is the post-replacement word and the remaining 2 words are those which follow the pre-replacement word in the original phrase, and k_2 is the number of a pattern in which the third word is the post-replacement word. Since it seems that, at greater number of pattern occurrence, the increase of the number of pattern occurrence does not affect to the increase of fluency of the generated phrases, logarithm function is involved to restrain the positive gradient of S with respect to the number of patterns.

IV. MATERIALS

A. Lyrics Emotion Corpus

A lyrics emotion corpus which consists of 366 phrases extracted from 15 songs, with five categories of emotions labelled as strengths of each category, was handled for experiments. The five emotion categories are *Joy*, *Love*, *Anger*, *Sorrow*, and *Anxiety*. 9 testers labelled the emotion categories via a questionnaire. As the result of the questionnaire, the numbers of phrases in which each emotion category was got the most votes were respectively, 142 for Joy, 48 for Love, 18 for Anger, 100 for Sorrow, and 55 for Anxiety. In the evaluation experiment, an instruction signal of each phrase is a vector whose value of element corresponding to each of the 3 most voted emotion categories is, 1.0 for the most-voted category, 0.5 for the 2nd most-voted category, and 0.25 for the 3rd most voted category. The values of the two remaining elements are both 0.0.

B. Word2Vec

The proposed method requires acquiring similarity between words on the basis of word-distributed representation. In this work, we use Word2Vec to obtain word similarities. Word2Vec learns word distributed representation from a resource corpus using *skip-gram* model. We use all articles written in Japanese from Wikipedia. A word-concept database which represents each of all vocabulary appeared in the resource corpus as a 200-dimension vector was generated as the result of training the skip-gram Word2Vec model.

C. MeCab

Since, in this work, lyrics written in Japanese are handled, morphological analysis of input phrases is essential. In this study, MeCab (Kudoh, 2006) was used for morphological analysis. Figure 1 shows an example of an input phrase and the result of morphological analysis of the input.

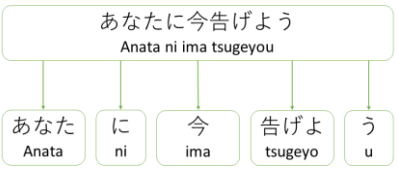

Figure 1. Example of morphological analysis.
An input phrase is divided into morphemes.

D. Convolutional Neural Network

In this study, convolutional neural networks (CNNs) perform emotion estimation in the evaluation experiment. CNN was originally proposed as an advanced architecture of *neocognitron* (Fukushima, 1980), a computational model for a mechanism of visual pattern recognition. CNNs consist of *convolution* layers and *pooling* layers. Each convolution layer processes spatial filtering with a bunch of filters with a certain size, and each pooling layer has a kernel with a certain size which strides over an input pattern taking a mean value or a max value within the kernel at each point of striding. Convolution layers and pooling layers achieve abilities of spatial feature extraction and invariance in response to moving or rotating input patterns. With such abilities, CNNs are reported to be able to perform not only image recognition,

but also text classification accurately (Kim, 2014; Santos & Gatti, 2014).

Since the word vectors have 200 dimensions, CNNs with filters in the first convolution layer having width of 200 and height of 4 is handled in the evaluation experiment.

V. EXPERIMENT OF AUGMENTATION

A. Conditions

An experimental augmentation of the lyrics emotion corpus introduced in *A* of **IV** using the proposed method explained in **III** was conducted. After augmentation, one experimenter classified the generated phrases into '*Correct*', '*Unnatural*', and '*Mismatch*'. A Correct phrase is a generated phrase whose apparent major emotion category seemed to be close to an emotion category which was voted the most for its original phrase. An Unnatural phrase seems grammatically or semantically unnatural, and a Mismatch phrase is a phrase whose emotion seems not to match with its original phrase.

B. Results and Discussions

As the result of the experiment with the lyrics emotion corpus consisting of 366 phrases, 1779 phrases were generated. Table 1 shows the number of generated phrases classified into each of Correct, Unnatural, and Mismatch.

Table 1. Classification of the generated phrases

Class	#Generated phrases
Correct	1341
Unnatural	387
Mismatch	51

The ratio of the number of Mismatch generated phrases to the number of all generated phrases was 0.03%.

Although, in most of the generated phrases, mismatches of the emotions to the emotions of the original phrases were not observed, semantics and vocabulary caused emotion inversion in a number of generated phrases. For example, in a phrase, 'angel' was replaced with 'devil'. This inversion was obviously because of oppositeness between those two words. In another case in which 'regret' was replaced with 'hardship', even though these two words seem to have close meanings, the emotions of the entire phrase were opposed because of Japanese idiom.

The ratio of Mismatch phrases was 0.03%. In machine learning, mislabelling often causes decrease in performance. However, by following a study on effects of noisy label to performance in machine learning, the Mismatch phrases was regarded as acceptable noises and conflated into the training dataset.

VI. EVALUATION EXPERIMENT

A. Conditions

In this experiment, to evaluate the proposed method, a CNN trained on the original lyrics emotion corpus (*alpha*) was compared to the augmented lyrics emotion corpus by the proposed method (*beta*). The evaluation was carried out in a 15-fold cross validation manner. Before training, the original corpus and the augmented corpus were respectively divided into 15 groups. Alpha and beta CNNs were respectively trained on the first group of the original corpus and the augmented corpus and performed estimations on the remaining groups of the original corpus, then were both initialised to be trained on the second group of the corpora. This procedure continues until the fifteenth iteration is finished. Both of the models were trained for 500 epochs. The size of minibatches was 50 for the alpha model, and 100 for the beta model. If index of element with the maximum value of an output vector on a phrase was the same as that of the instruction vector labelled on the phrase, such case was handled as a true positive for the emotion category corresponding to the index. For each emotion category $X_i (i = 1,2, \dots ,5)$, true positives and true negatives were handled as correct outputs. The ratio of the number of correct outputs for X_i to the number of input phrases the maximum element of whose instruction vector corresponded to X_i is referred to as the accuracy on X_i.

B. Results and Discussions

Figure 3 shows accuracy for each emotion category which was performed by the model of each iteration. Each bar shows the mean accuracy within an iteration. Although for five iterations (1, 8, 11, 12, 14), reductions in mean accuracies were observed, there were slight increases which were caused by corpus augmentation in the remaining iterations. Especially, accuracies on Joy and Sorrow which were labelled on many phrases improved greatly. For Love and Anger, even though there were a few cases where accuracies declined, there were more cases where they improved. In the tenth iteration, the

accuracy on Love reached 100% after augmentation.

Figure 4 is two confusion matrices of the maximum index of the instruction vectors and output vectors of the alpha model (left) and the beta model (right). An element on each matrix is the number of phrases which were classified into the emotion category corresponding to the vertical axis and its instruction signal was the emotion category corresponding to the horizontal axis. The numbers on these matrices are sum of all such results over all iterations. Increase in the numbers on the diagonal elements resulting from augmentation was observed, especially for Joy. On the other hand, the beta model did not classify any phrases into Anger.

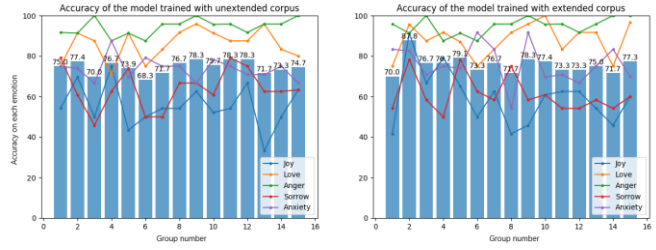

Figure 3. Accuracies of the models on each emotion category.

Lines are accuracies of the model for each emotion category in each iteration, and the bars are mean accuracies in each iteration.

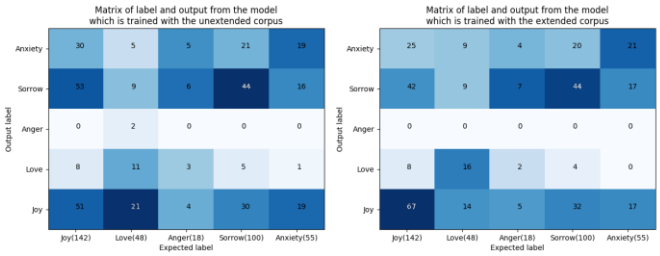

Figure 4. Confusion matrices of instruction labels and outputs.

The numbers on the diagonal lines approximately correspond to accuracies.

For all emotion categories except Anger, accuracies of the beta model were greater than the alpha model. Especially, despite the fact that the number of phrases labelled Love was small, accuracy for Love became much higher. From these observations, corpus augmentation using the proposed method was evaluated to be valid.

By Figure 3 and Figure 4, the beta model seemed to have learnt not to classify any phrases into Anger, labelled on few phrases in the original corpus, in order to perform around 100% accuracies in all iterations. Therefore, accuracies on Anger barely changed. According to the fact that the accuracies for Anxiety, labelled on 55 phrases, and Love, labelled on 48 phrases, increased, it is expectable that, with more Anger phrases in the augmented corpus, the accuracy for Anger should also increase.

In further studies, proposals of methods to augment a corpus with a certain mechanism which can balance the number of phrases on which each emotion category is labelled will be essential. With such mechanism, it may be able to avoid accuracy decline resulting from imbalance of labelling.

VII. CONCLUSION

In this work, a method to augment a small scaled lyrics emotion corpus was proposed. After qualitatively observing generated phrases using the proposed method, evaluation experiment using CNNs was conducted. In this experiment, two CNNs were trained on the original corpus and the augmented corpus respectively, and they performed tests with phrases in the original corpus. As the result of the evaluation, it was found that, for almost all of the five emotion categories, accuracies slightly increased. However, from the fact that, for Anger, which was labelled on few phrases in the original lyrics emotion corpus, it was also found that accuracies for certain emotion categories cannot increase without balancing the number of each label.

VIII. ACKNOWLEDGEMENT

This work is supported by JSPS KAKENHI Grant Number JP15K16077 and JP16K16134.

IX. REFERENCES

Wang, J & Perez, L 2017, 'The Effectiveness of Data Augmentation in Image Classification using Deep Learning', *arXiv:1712.04621*.

Nishimoto, S, Noji, H, & Matsumoto, Y 2017, 'Aspect Estimation of Emotion Analysis by Data Augmentation' [Translated from Japanese], *Proceedings of the Twenty-third Annual Meeting of the Association for Natural Language Processing, pp. 10-14.*

Kim, Y 2014, 'Convolutional Neural Networks for Sentence Classification'. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, pp1746-1751.*

Mikolov, T, Chen, K & Corrado, G 2013, 'Efficient Estimation of Word Representations in Vector Space'. *CoRR, abs/1301.3781.*

Dos Santos, C, & Gatti de Beyser, M 2014, 'Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts', *International Conference on Computational Linguistics, Technical Papers, pp. 69-78.*

Chen, J, & Chen, A 1998, 'Query by Rhythm: An Approach for Song Retrieval in Music Databases', *Research Issues in Data Engineering. pp. 139-146.*

Homma, Y, & Hagiwara, M 2013, 'Estimation of Reader's Emotion to News Articles using Twitter and its Application to Access Analysis', *Journal of Japan Society of Kansei Engineering, vol.12, pp. 167-174.*

Kudoh, T 2006, March, MeCab: Yet Another Part-of-Speech and Morphological Analyzer. http://taku910.github.io/mecab/

Silfverberg, M, Wiemerslage, A, Ling, L, & Mao, LJ 2017, 'Data Augmentation for Morphological Reinflection', *Proceedings of the CoNLL SIGMOPHON 2017 Shared Task: Universal Morphological Reinflection, pp. 90-99.*

Muhammad, AM, & Ungar, L 2017, 'Emonet: Fine-grained emotion detection with gated recurrent neural networks', *ACL, vol 1, pp718-728.*

Zhao, J, & Gao, Q 2017, 'Annotation and Detection of Emotion in Text-based Dialogue Systems with CNN', *arXiv preprint, arXiv: 1710.00987.*

Beall, J 2008, 'The Weakness of Full-Text Searching'. *The Journal of Academic Librarianship, 34(5), pp. 438–444.*

Fukushima, K 1980, 'Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position', *Biol. Cybernetics 36, 193 202 (1980).*

Matsumoto, K, & Sasayama, M 2018, 'Lyric Emotion Estimation Using Word Embedding Learned from Lyric Corpus', *2018 IEEE 4th International Conference on Computer and Communications, pp2295-2301.*

Xuelong, L, Di, H & Xiaoqiang, L 2017, 'Image2song: Song Retrieval via Bridging Image Content and Lyric Words', *arXiv:1708.05851.*