An improved method for multi-label classification methods using deep neural language models with ensemble learning

Yuuki Kusumoto[1†] , Makoto Okada[2] and Naoki Mori[3]

[1,2,3] Osaka Metropolitan University, Japan

([1,2,3]Tel: 072-254-7511; E-mail: [1]kusumoto@ss.cs.osakafu-u.ac.jp, [2]okada@omu.ac.jp, [3]mnao@omu.ac.jp)

Abstract: In this paper, we propose a deep language model to improve the conventional method of classifying Japanese review texts published by travel agencies into multi-label. In today's business world, the need to utilize a large amount of text data is increasing, and automating the assignment of multiple labels to each text and improving the accuracy of the assignment is a major challenge. However, it is not easy to classify texts belonging to multi-label, and therefore, improvement of conventional methods is required. Therefore, we proposed a deep language model that improves the classification of text data with multiple labels by using ensemble learning to create specialized classifiers for each label and integrate them and confirmed the effectiveness of the proposed model.

Keywords: natural language processing, multi-label classification, ensemble learning

1. INTRODUCTION

Nowadays, a large amount of text data are written every day, and their effective utilization is required. Especially in business settings, it is a major challenge to automatically classify each text into multiple categories and improve the classification ' s accuracy to save labor and make effective use of the data. Currently, binary classification, which classifies whether a document is positive or negative, is a task that can be solved with high accuracy. However, it is difficult to make effective use of text data with such a simple classification method, so it is necessary to classify data according to the multiple meanings of the text. Therefore, Multi-Label Classification (MLC) task, which assumes that a single document is assigned multiple labels, is becoming increasingly important. In this study, we propose a deep language model that improves the accuracy of multiple labels classification by using ensemble learning. The feature inner product calculation method and the model ' s overall structure are based on the aspect-based sentiment analysis network model[2] proposed by Miura et al. in our previous study. Since our model is capable of extracting document features in addition to features effective for MLC, we believe that it can be used for future research on explainable artificial intelligence in the field of natural language processing. Validation experiments will be conducted to confirm the effectiveness of the proposed method.

2. RELATED TECHNOLOGY

2.1. BERT

BERT (Bidirectional Encoder Representations from Transformers)[1] is a natural language model consisting of bi-directional encoders using transformers. Significant results have been reported, including the best performance at the time of its release on various tasks such as sentence classification, question answering, and named entity recognition. BERT has attracted attention for its versatility in that it can be used for various natural language processing tasks through transfer learning and fine-tuning of pre-trained models. In the experiments of this study, we used the BERT Japanese Pretrained model[1] published by the Inui/Suzuki Laboratory of Tohoku University.

2.2. Transformer

Transformer[3] is a deep language model with an encoder and a decoder with an Attention mechanism proposed in the field of natural language processing. It is a deep language model with an encoder and a decoder equipped with an Attention mechanism. Since Transformer can train and inference on serial data without using a Recurrent Neural Network (RNN), it is expected to improve the training speed by parallelization. It is also used for tasks in the field of natural language processing due to its high performance.

3. ASPECT-BASED SENTIMENT ANALYSIS

Aspect-based Sentiment Analysis is a research in the field of natural language processing that focuses on two tasks: MLC and artificial intelligence explainability. Sentiment analysis is a task that aims to analyze opinions, feelings, and attitudes from texts, and to estimate and classify whether the polarity of certain content is positive or negative. With the development of natural language processing technology, there is a growing demand for quantitative evaluation of users' opinions on social networking services and online review sites. Emotion analysis has also attracted attention. One of the tasks of sentiment analysis is aspect-based sentiment analysis[2], which focuses on the analysis of contextual information. Aspect in aspect-based sentiment analysis is defined by the target of the sentence and its attributes. Aspects in sentences are used to analyze what the sentences are about. In general, aspect-based sentiment analysis involves three steps. First, the sentences are classified into the given aspect categories. Next, we estimate the position of the phrase relative to the aspect categories contained in the sentence.

† Yuuki Kusumoto is the presenter of this paper.

[1]https://github.com/cl-tohoku/bert-japanese

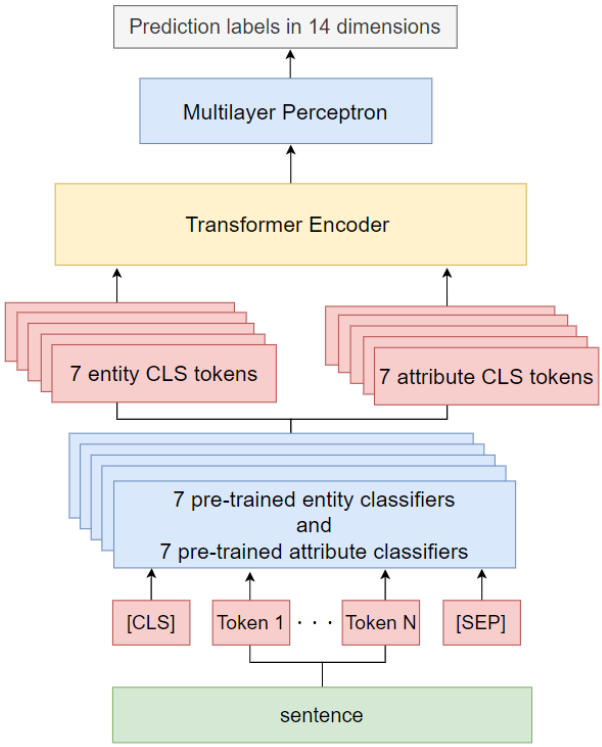

Fig. 1. Schematic diagram of the proposed model Multi pre-trained models Transformer (Mpm+T)

Finally, the polarity of the phrases is analyzed. Finally, we analyze the polarity of the phrases and aim to improve the accuracy of the analysis as a whole.

4. DATASET

In this paper, we used "Rakuten Dataset"[4] provided by Rakuten Group, Inc. Among them, we used the Rakuten Travel Review: Aspects and Sentiment-tagged corpus[5]. The Japanese review sentences from Rakuten Travel and the respective sentences are tagged positively or negatively for seven categories: breakfast, dinner, bath, service, facility, state, and room. From a total of 76623 datasets, we removed noisy datasets, datasets written in languages other than Japanese, and datasets with more than 101 tokens of sentences, resulting in 73461 datasets. The breakdown of the preprocessed dataset is 23429 data that do not belong to any of the 14 categories, and 1858 data that are assigned both positive and negative values to a single category. Excluding these two types of data, the number of data used for model training and evaluation is 48354.

5. PROPOSED MODEL

5.1. Multi pre-trained models Transformer (Mpm+T)

It is not easy to extract features with multi-label classification information from a single sentence in the MLC task. Therefore, we focused on the aspect of aspect-based sentiment analysis. Since the aspect is defined by the target and the attribute of the written text, we devised to create a sub-classifier specialized for the classification of each target and

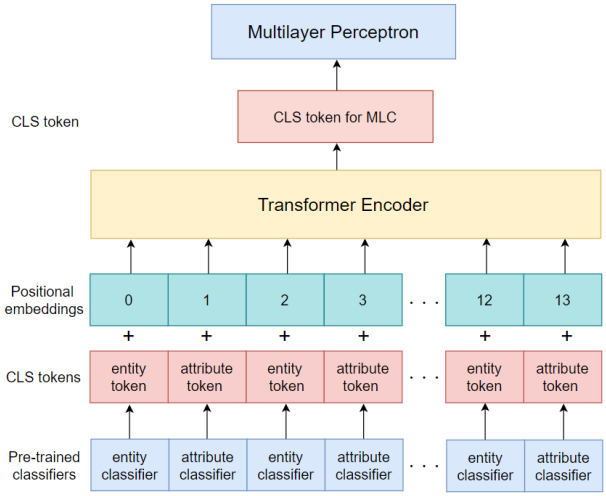

Fig. 2. Learning by Mpm+T

attribute. In this study, we propose the Multi pre-trained models Transformer (Mpm+T) model, which is a deep language model using ensemble learning to improve the multi-label classification method. Figure 1 shows a schematic diagram of the Mpm+T model developed in this study. In the model, CLS is the vector of distributed representations of tags used in BERT.

The overall structure of the proposed model and the use of a Transformer Encoder as a mechanism for computing the inner product of entity CLS tokens and attribute CLS tokens are similar to those in the previous study[2], which follows the aspect-based sentiment analysis network proposed by Miura et al. Two novel points are noted in comparison with this model.

1. The point is that MLC is not done by a single BERT model, but by creating sub-classifiers for objects and attributes in an aspect.
2. Another point is that we do not use the word embeddings of document tokens as input to the Transformer Encoder, but only CLS tokens. In general, document classification tasks using BERT and Transformer use the distributed representation to derive the Attention of each document token. In this study, however, only CLS tokens that retain document classification information from the sub-classifiers are input to the Transformer Encoder to derive the Attention corresponding to each label. This approach significantly reduces the amount of information input to the Transformer Encoder compared to the model of Miura et al. and general multi-label classification.

Next, we describe the overall process flow of the model. Figure 2 shows how the model is trained. First, pre-trained entity classifiers and pre-trained attribute classifiers are populated with common document data. Then, the features of each token are extracted from the input documents. The CLS tokens, which are the first tokens of these tokens, are acquired as entity CLS tokens and attribute CLS tokens in each classifier, and the location information is assigned to these CLS tokens. Next, the Transformer Encoder derives the inner product of the CLS tokens using the self Attention

Table 1. Accuracy and their standard deviation for pre-trained entity classifiers and pre-trained attribute classifiers

	breakfast	dinner	bath	service	state	facility	room
Accuracy for entity classifiers	0.916 ± 0.002	0.923 ± 0.006	0.943 ± 0.006	0.888 ± 0.005	0.951 ± 0.001	0.868 ± 0.01	0.919 ± 0.001
Accuracy for attribute classifiers	0.936 ± 0.001	0.922 ± 0.005	0.917 ± 0.001	0.926 ± 0.001	0.925 ± 0.003	0.908 ± 0.094	0.905 ± 0.007

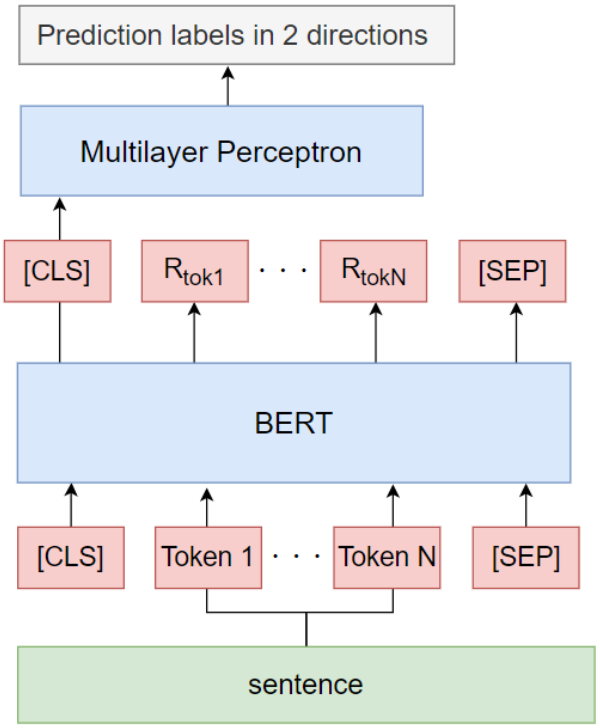

Fig. 3. Schematic diagram of the model of the sub-classifier

mechanism. Finally, the derived results are dimensionally compressed into features for MLC using the Multilayer Perceptron (MLP).

Table 1 shows the Accuracy during the 5 fold cross-validation of entity classifiers and attribute classifiers. The following section describes the details of the model overview diagram. Figure 3 shows a schematic diagram of the model for the sub-classifiers.

5.2. Pre-trained entity classifiers and entity CLS tokens

The pre-trained entity classifiers are binary classification models as shown in Figure 3. It is a set of seven sub-classifiers created by training the Rakuten Travel review dataset on the BERT Japanese pre-trained model, to obtain features for binary classification of whether the data belongs to a category or not. The number of training data is 24000 and the number of validation data is 6000. The entity CLS tokens refer to the CLS tokens that are included in the features transformed from the document data by the classifier. They contain information used for binary classification.

5.3. Pre-trained atribute classifiers and attribute CLS tokens

The pre-trained attribute classifiers are binary classification models as shown in Figure 3. It is created by training the Rakuten Travel review dataset on the BERT Japanese pre-trained model and consists of seven sub-classifiers to obtain features for bivariate classification of whether the review belongs to the positive or negative category. The data set contains 30,000 data sets and was subjected to 5-fold cross-validation. The entity CLS tokens refer to the CLS tokens included in the features transformed from the document data by the classifiers and contain the information used for binary classification.

5.4. Transformer encoder and MLP

In Transformer Encoder, each CLS token obtained by the subclassifies is assigned location information, and the inner product of the tokens is calculated using the Attention mechanism. In MLP, 768-dimensional features obtained by Transformer Encoder are compressed to 14 dimensions and input to the Sigmoid function for normalization.

6. EXPERIMENTS

To validate the effectiveness of the proposed model, we conducted a five-segment cross-validation on 30,000 data sets, where 24,000 training data sets and 6,000 validation data sets were used. In addition to using micro-F1 as an evaluation index as in many related studies of MLC, we define Perfect Accuracy and Partial Accuracy. Perfect accuracy refers to the percentage of all the correct and predicted data among all the data. Partial Accuracy refers to the percentage of partial agreement between the correct and predicted data among all the data. Table 2 shows the parameters of the Mpm+T experiment model.

Table 2. Mpm+T parameters at the time of the experiment

input dimensions for BERT	768
output dimensions for BERT	768
input dimensions for Transformer	768
output dimensions for Transformer	14
Number of Transformer blocks	3
batch size	4
optimization function	Adam
learning rate	7.6×10^{-6}
loss function	BinaryCrossEntropyLoss
loss function	15

6.1. Comparison methods

Two types of comparison methods are presented. We choose a conventional method, BERT+MLP, and a mimetic model of the aspect-based sentiment analysis network by Miura et al. 5 fold cross-validation was performed with the same number of data as for mpm+T.

6.2. Experimental results

Table 2 shows the results of the 5-fold cross-validation. It can be confirmed that Mpm+T outperforms the comparison method in classification accuracy for all evaluation in-

Table 3. Means and standard deviations of evaluation indices of experimental results

Evaluation indices	Precision	Recall	micro-F1
Mpm+T	0.846 ± 0.019	0.872 ± 0.028	0.858 ± 0.017
BERT+MLP	0.683 ± 0.005	0.772 ± 0.009	0.724 ± 0.005
Model of Miura et al.	0.741 ± 0.015	0.804 ± 0.006	0.773 ± 0.013

Table 4. The number of complete/partially correct answers for the test data and the percentage of complete/complete correct answers for multi/single labels.

Evaluation indices	completely correct	partially correct	multi-label completely correct	Single label completely correct
Mpm+T	4311	1689	1298	3013
BERT+MLP	3188	2812	862	2326
Model of Miura et al.	3884	2116	1097	2787

dices. Next, we evaluate 6000 test data with the best micro-F1 model in each model. The number of multi-label data included in the 6000 cases is 2060, and the number of single-label data is 3940. Table 3 shows the number of completely correct and partially correct answers and the number of completely correct answers for multi-label and single-label data. A completely correct answer means that the predicted labels completely match the correct data, while a partially correct answer means that the predicted labels match a part of the correct data. A higher number of complete correct answers and multi-label / single-label complete correct answers is a good value, and Table 3 shows that Mpm+T outperforms the compared methods. In particular, it is confirmed that the number of correct multi-label answers is greatly improved compared to the comparison method. Since multi-label classification requires detailed classification information, we consider that the use of ensemble learning influences this result. We also confirmed the effectiveness of the method of deriving Attention of labels by inputting only CLS tokens, not word embeddings of document tokens, to the Transformer Encoder.

7. SUMMARY AND FUTURE ISSUES

In this study, we proposed an improved method for multi-label classification based on deep language models using ensemble learning. The difference from the previous studies is that we introduce aspect-based sub-classifiers and input only CLS tokens obtained from the sub-classifiers to the Transformer Encoder. As a result, we confirmed that the classification accuracy is higher than that of the conventional method and the imitation models of previous studies.

Future work is to improve the generality of the model by utilizing the data excluded by the cleansing process. In addition, we will confirm the impact of the sub-classifiers on the model as a whole and the extraction of aspect information.

REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language un- derstanding. arXiv preprint arXiv:1810.04805, 2018.

[2] Y. Miura, M. Atsumi. Aspect-Based Sentiment Analysis Using Pre-Learning Language ModAspect-based sentiment analysis neural network using pre-trained language model Estimation of Multiple aspect category polarities and target phrasesels Aspect-Based Sentiment Analysis Neural Networks Using Pre-Learning Language Models. Proceedings of the Annual Conference of JSAI Proceedings of the 35th Annual Conference of Japanese Society for Artificial Intelligence, 2021

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in neural information processing systems, pp. 5998–6008, 2017.

[4] Rakuten Group, Inc. : Rakuten Dataset. Informatics Research Data Repository, National Institute of Informatics. (dataset). https://doi.org/10.32130/idr.2.0, 2014

[5] Rakuten Group, Inc. : Rakuten Travel Review: Aspects and Sentiment-tagged corpus. Informatics Research Data Repository, National Institute of Informatics. (dataset). https://doi.org/10.32130/idr.2.14, 2021