

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

Yuuki Kusumoto^{1†} and Makoto Okada² and Naoki Mori³

^{1,2,3} Osaka Metropolitan University, Japan

(^{1,2,3}Tel: 072-254-7511; E-mail: ¹kusumoto@ss.cs.osakafu-u.ac.jp, ²okada@omu.ac.jp, ³mori@omu.ac.jp)

Abstract: abstract

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

1. INTRODUCTION

Nowadays, a large amount of text data are being 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[1] 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. UNDERLYING TECHNOLOGY

2.1. BERT

BERT[2] (Bidirectional Encoder Representations from Transformers) is a natural language model consisting of bidirectional 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 model¹ 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.

2.3. 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. In the task of sentiment analysis, there is an aspect-based sentiment analysis cite Aspect-Based Sentiment Analysis Neural Network Using Yoshiei Miura 2021 Pre-Learning Language Model, 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. 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.

3. DATASET

The dataset is "Rakuten Travel Review: Aspect Sentiment Tagged Corpus" [4] published by Rakuten Group, Inc. Japanese review sentences from Rakuten Travel and their respective sentences are assigned positive or negative tags for seven categories: breakfast, dinner, bath, service, facilities, location, and room. From a total of 76623 datasets, we removed noisy datasets, datasets written in languages other

[†] Yuuki Kusumoto is the presenter of this paper.

¹<https://www.nlp.ecei.tohoku.ac.jp/news-release/3284/>

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

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.

4. PROPOSED MODEL

4.1. 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 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[1], 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 distributed representation 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.

The overall process flow of this model is described below. First, document data are input to pre-trained entity classifiers and pre-trained attribute classifiers. Then, the features of each token are extracted from the input documents, and the first tokens, CLS tokens, are acquired as entity CLS tokens and attribute CLS tokens in each classifier. Next, the Transformer Encoder adds location information to the entity

CLS tokens and attribute CLS tokens, and then derives their inner product using the Attention mechanism. Finally, the resulting derivation is dimensionally compressed into features for MLC using the Multilayer Perceptron (MLP).

Table 1 shows the Accuracy during the five-part cross-validation of entity classifiers and attribute classifiers. The following section describes the details of the model overview diagram.

4.2. pre-trained entity classifiers and entity CLS tokens

The pre-trained entity classifiers are seven types of small classifiers for binary classification of categorization features, which are created by training the Rakuten Travel review dataset on the BERT Japanese pre-trained model. 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.

4.3. pre-trained attribute classifiers and attribute CLS tokens

The pre-trained attribute classifiers are created by training the Rakuten Travel review dataset on the BERT Japanese pre-trained model. It consists of seven sub-classifiers to obtain features related to the binary classification of positive or negative labels. The attribute CLS tokens refer to the CLS tokens that are included in the features transformed from the document data by the classifiers and contain the information used for binary classification.

4.4. 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.

5. EXPERIMENT

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.

REFERENCES

- [1] 三浦義栄, 渥美雅保. 事前学習言語モデルを用いたアスペクトベースセンチメント分析ニューラルネットワーク. 人工知能学会全国大会論文集第 35 回 (2021), pp. 2Yin507–2Yin507. 一般社団法人 人工知能学会, 2021.
- [2]] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
- [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] 楽天グループ株式会社. 楽天データセット (コレクション) , aug 2010