Multi-Label Text Classification Using Only Label Names

Rongbo ZHANG[†] Zhewei XU[‡] and Mizuho IWAIHARA[‡]

Graduate School of Information, Production and Systems, Waseda University
2-7 Hibikino, Wakamatsu-ku, Kitakyushu-shi, Fukuoka, 808-0135 Japan

E-mail: †zhangrongbo@toki.waseda.jp, ‡xuzhewei@toki.waseda.jp, ‡iwaihara@waseda.jp

Abstract Multi-label text classification (MLTC) aims to tag each document with multiple labels by a set of classifiers. Most of the existing MLTC methods train classifiers using a large number of labeled documents, which usually require costly annotation work in the real world. Class-name only classification is an approach to discover a list of replaceable words from the category names and then use these replaceable words to tag documents. However, class name-only classification has not been discussed for multi-label text classification. Also, only single words are considered as label names. In this paper, we propose a multi-label text classification method that only uses label names to train classifiers. Specifically, (1) a masked language model is used to search replaceable phrases for each label name (one word or phrase), to construct a replaceable phrase list for each category. (2) Identifying documents having high semantic similarity with the replaceable phrase list of each label to train a classifier. (3) Finetuning the classifier through multi-label self-training. Our experimental results show that our approach achieves higher micro-F1 scores than semi-supervised models trained by 5% and 10% of labeled documents.

Keyword Multi-label classification, masked language model, fine-tuning, self-training

1. Introduction

Text categorization is a fundamental task in information retrieval and natural language processing. It is widely used in various applications, such as sentiment analysis, topic classification and news categorization. Traditional text classification usually assumes that each document is related to one topic, which could not fit well for documents in the real world as they may cover two or more topics simultaneously.

Multi-label text classification (MLTC) allows assigning multiple labels to one document. Such as the following sentence, "The FIFA World Cup 2018 official match ball – is the first innovative football, which steps in IoT era with NFC technology." It is easily found that this sentence has multi labels "sports" and "technology". MLTC usually needs a large collection of labeled documents to train the classifier, requiring a large amount of manual work by domain experts or workers.

In recent years, several approaches were proposed to semi-supervised MLTC [10][18], which requires less labeled data. The success of semi-supervised methods stems from the use of large amounts of unlabeled data. Although mitigating the human annotation burden, these methods still require a labeled dataset that covers all labels, which could be too expensive to obtain when we have a large number of labels in MLTC. Meanwhile, label-name only classification (LOTClass) has also been proposed [11]. However, LOTClass in [11] is a multi-class classification assigning one class label to each document. LOTClass has not yet been discussed on MLTC. Also, only single words, not phrases, are considered as replaceable words.

In this paper, we study the problem of multi-label text classification where only label surface names and unlabeled corpus are available for model training. We propose an approach named Label Name Only for Multi-Label Text Classification (LNO-MLTC), consisting of the following four major steps. First, we search replaceable words for each label name, and construct a replaceable phrase list for each category. Second, we identify documents having high semantic similarity with the replaceable phrase list. Third, documents having high semantic similarity with the replace words are used for training an initial text classifier, where the pre-trained language model BERT [5] is used as the document encoder. Finally, we generalize this text classifier using multi-label self-training on unlabeled documents, where documents with pseudo labels are used for the next round of classifier training. We evaluate our model on two datasets [1][17], our model outperforms all compared methods on the Reuters dataset and achieves comparative performance on the Arxiv dataset.

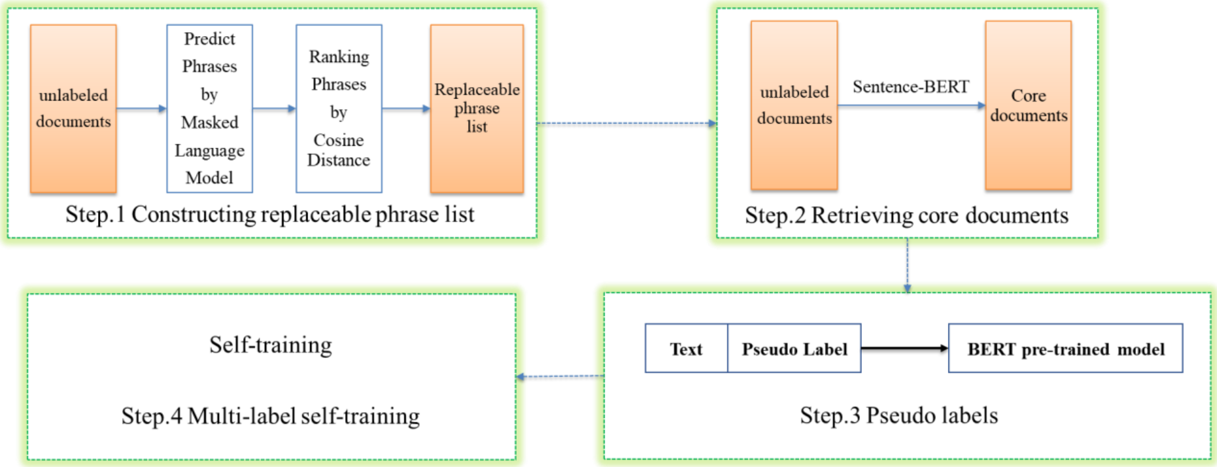

unlabeled documents
Predict Phrases by Masked Language Model
Ranking Phrases by Cosine Distance
Replaceable phrase list
Step.1 Constructing replaceable phrase list
unlabeled documents
Sentence-BERT
Core documents
Step.2 Retrieving core documents
Text
Pseudo Label
BERT pre-trained model
Step.3 Pseudo labels
Self-training
Step.4 Multi-label self-training

Figure 1. Overview of the Label Name Only Multi-label Text Classification

2. Related work

2.1 Multi-label text classification

Multi-label text classification (MLTC) is a basic task in natural language processing. MLTC is often solved by transforming the multi-label text classification task into a collection of binary classification tasks [3]. Several approaches take advantage of pair associations or mutexes between labels. Pairwise Comparison [6] transforms the multi-label task into a label ranking task by using a natural extension of pairwise comparison. Classifier Chains [14] transforms the multi-label task into a chain of binary classification tasks. With the development of deep learning, SGM [17] uses a sequence learning model to solve the MLTC.

In recently years, several studies use small labeled documents as the signals for the semi-supervised multi-label text classification. TRAM [10] transforms the classification problem into an optimization problem of estimating the label composition for each labeled instance. COINS [18] adapts the co-training strategy in the multi-label context. In each co-training round, a dichotomy over the feature space is learned by maximizing the diversity between the two classifiers.

Even though these methods achieved greatly improvement on MLTC tasks, the number of given labeled samples heavily affects classification accuracy. On the other hand, our label name-only classification assumes only category names of the target task. Pretrained language models trained over general corpora are exploited to discover phrases replaceable with category labels.

2.2 BERT

BERT [5] has a pre-trained bidirectional encoder architecture based on a Transformer encoder, setting new state-of-the-art results on various NLP tasks. With its self-attention mechanism, and training over large corpora by the masked language model and next sentence prediction training, BERT captures deep contextual relations better than RNN-based models. The masked language model (MLM) randomly masks some of the tokens from the input. The objective is to predict the original vocabulary id of the masked word based only on its context.

Pre-trained word embeddings are an integral part of modern NLP systems. Unlike the left-to-right language model [13] or objectives to discriminate correct from incorrect words in left and right contexts [12], the MLM objective enables the presentation to fuse the left and right contexts, through a deep bidirectional Transformer.

In our work, we try to use extended BERT models for both multi-label classification and generation of phrases replaceable with category labels. We introduce an extended MLM to predict replaceable phrases, because words that are semantically related to category labels in the form of phrases, and the BERT encoder divides the phrases into multiple tags. The traditional MLM of BERT can only be used for prediction of one masked token. In this case, we utilize the encoding of SpanBERT [7], which can mask out contiguous sequences of tokens for improved span representations.

2.3 Semantic similarity

The main objective of semantic similarity is to measure the distance between the semantic meanings of a pair of words,

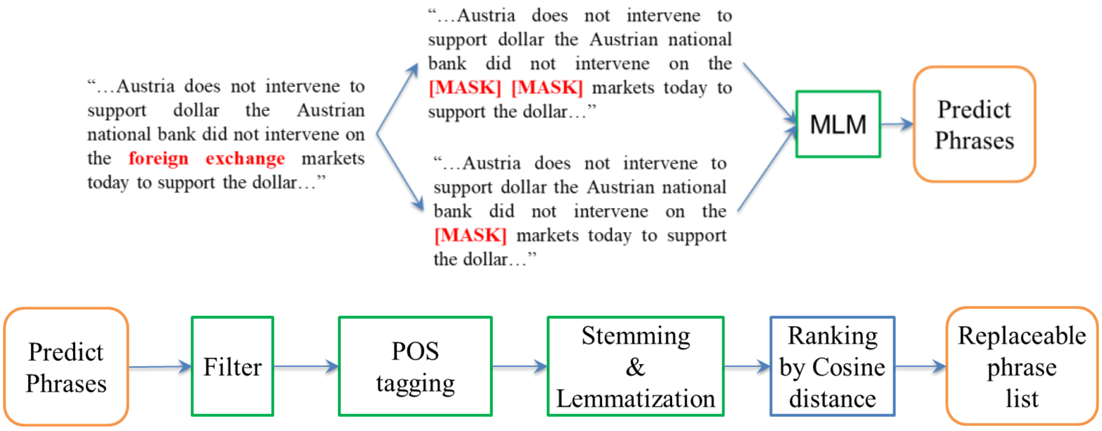

"…Austria does not intervene to support dollar the Austrian national bank did not intervene on the foreign exchange markets today to support the dollar…"
"…Austria does not intervene to support dollar the Austrian national bank did not intervene on the [MASK] [MASK] markets today to support the dollar…"
"…Austria does not intervene to support dollar the Austrian national bank did not intervene on the [MASK] markets today to support the dollar…"
MLM
Predict Phrases
Predict Phrases
Filter
POS tagging
Stemming & Lemmatization
Ranking by Cosine distance
Replaceable phrase list

Figure 2. The process of constructing the replaceable phrase list Text Classification

phrases, sentences, or documents. For example, the word "car" is more similar to "bus" than "cat".

Word embeddings are widely used, because of their ability to represent a word in the form of a low-dimensional vector. Recently, Siamese networks and contrastive learning methods are proposed for sentence embeddings, which can be used to learn semantic similarity between sentences. Sentence-BERT [15] is specialized for semantic textual similarity tasks and achieves better results than the BERT baseline. In Sentence-BERT, a bi-encoder of shared BERT layers is forming a Siamese network, generating sentence representations, by which semantic similarities are calculated as the cosine similarity between them.

We utilize Sentence-BERT for calculating cosine similarities between phrase lists and sentences and between sentences.

3. Preliminaries

3.1 Major Mathematical Notations

A corpus $D = \{D_1, D_2, \ldots, D_{n-1}, D_n\}$ is a document collection where each document $D_i \in D$ is a sequence of words.

Let the list $N = [N_1, N_2, \ldots, N_m]$ denotes label names, where m is the number of the labels.

3.2 Problem Definition

Given an unlabeled document set and label names N, multi-label text classification is to predict a label vector $L = [L_1, L_2, \ldots, L_m]$, $L \in \{0,1\}$ on each document in D, where $L_i = 1$ indicates the document belongs to N_i, otherwise $L_i = 0$.

4. Methodology

In this paper, we propose a new method named Label Name Only Multi-label Text Classification (**LNO-MLTC**), which trains a multi-label text classifier from label names and unlabeled documents. Figure 1 shows the overview of our proposed approach. In the beginning, we search replaceable phrases for each label name, and use the masked language model to generate replaceable phrases for each label name. The pos-tagging and cosine distance score are used to select phrases while reducing irrelevant phrases. The top-50 phrases are concatenated as the replaceable phrase list for each label. Then, by the semantic similarity with the replaceable phrase list, highly similar documents are retrieved from the unlabeled documents. These similar documents are used to train an initial multi-label text classifier. Finally, we apply self-training to improve the classifier.

4.1 Constructing replaceable phrase list

In this section, we introduce our method for generating a replaceable phrase list from the name of each label with a pre-trained language model.

4.1.1 Masked language model prediction

As Figure 2 shows, when a label name in a document is found, we replace the label name with one special token "[MASK]" and two tokens "[MASK] [MASK]". The

POS tagging	Descriptions	Examples
NN	Noun, singular	Corn
NNS	Noun, plural	Corns
NNP	Proper noun, singular	USA
NNPS	Proper noun, plural	Johns

Table 1. POS tags for extracting noun phrases

Method	money-fx (foreign exchange)
MLM Prediction	'foreign exchange', 'international exchange', 'international increase', '~~financial transfer~~', 'foreign increase', 'world', 'gold increase', ..., '~~import~~', '~~value~~', 'national exchange', 'rate', 'import', 'wealth', 'global exchange', '~~job~~', '~~food~~', '~~balance~~', 'new exchange', 'european export', 'international account', 'financial change', 'new change'…
MLM + Ranking by similarity to label	'foreign exchange', 'currency', 'international exchange', 'international market', 'foreign market', 'international increase', 'global exchange', 'financial change', 'banking', 'foreign increase', 'national exchange', 'country investment', 'large exchange', 'treasury', 'global investment', 'euro', 'gdp', 'bank', 'inflation', 'german market', 'asian investment', 'finance', 'foreign value', ...
	cs.ai (artificial intelligence)
MLM Prediction	'artificial intelligence', '~~human computer~~', 'human ai', 'computational ai', 'automated intelligent', 'cognitive ai', '~~robust brain~~', 'ai intelligence', 'computational intelligence', '~~visual brain~~', ..., 'human attention', 'ai', 'computational knowledge', '~~biology~~', 'automated mind', 'intelligent ai', '~~mind~~', '~~design~~', 'adaptive ai', 'ai intel', 'intelligent knowledge', '~~data~~', '~~vision~~', 'adaptive knowledge', 'adaptive intelligent', '~~game~~', 'adaptive intelligence'…
MLM + Ranking by similarity to label	'artificial intelligence', 'human ai', 'computational ai', 'automated intelligent', 'computational intelligent', 'cognitive ai', 'ai intelligence', 'computational intelligence', 'automated intelligence', 'ai', 'computational knowledge', 'automated mind', 'intelligent ai', 'adaptive ai', 'ai intel', 'adaptive intelligent', 'adaptive intelligence', 'intelligent learning', 'automated information', 'computational', 'computational human', 'computational learning', ...

sentences containing the label name n constitute the labeled documents D^n, which is given to the BERT encoder to obtain the contextualized embedding for D^n. Then the masked language model (MLM) is applied to predict words that can be inserted at each masked position. As used in [11], the original MLM is only to predict for one masked position and cannot be directly used to predict consecutive positions. All results are single words, and it is easy to select valuable results.

For dealing with phrases, we first utilize the encoder of SpanBERT [7], which can perform prediction on consecutive positions. Two words assigned on two consecutive masked positions form a two-word phrase. Words and phrases are regarded as phrases that are likely to occur in the context of the position and can replace the label name. Note that three-word phrases can be generated from three consecutive masked positions, our evaluations indicate that key phrases are rarely found in this situation. Then cleaning the predictions by removing stop-words using NLTK [2] and phrases with symbols, we can get the final MLM predictions, these interchangeable phrases are expected to have similar meanings or are related to the label name.

4.1.2 Select phrases

Even after stop-words are removed, there still exist undesirable items in the phrase list, and part of the predicted phrases are not nouns. To further select quality phrases, we apply the following methods.

Part-of-speech (POS) Tagging: First, we apply part-of-speech (POS) tagging [3] to choose noun phrases, based on the idea that noun phrases often consist of zero or more adjectives followed by one or more nouns. Table 1 shows the type of extracted phrases by POS tags. After POS tagging, the noun phrases are selected from the results. Then stemming and lemmatization are applied to reduce inflectional forms and related forms, such as "corns" for "corn".

Cosine distance to labels: A phrase that is more semantically relevant to a category label name should be better. We calculate the cosine distance between each phrase and the documents in which a label name l is occurring, for evaluating semantic relatedness to l:

$$Sim(p, D_i) = \frac{p \cdot em(D_i)}{|p||em(D_i)|} \quad (1)$$

$$Cosine\_Distance(p, l) = \sum_{D_i \in D^l} \left(1 - Sim(p, em(D_i))\right) / |D^l| \quad (2)$$

Here, p is the embedding of the given phrase, $em(D_i)$ is the embedding of a document D_i, and D^l is the set of documents in which label name l occurs. Phrases with lower cosine distance are selected.

We can obtain the final top-50 phrases for each label name using these methods. As an example, Table 2 shows the replaceable phrase list for the label "money-fx" and "cs.ai" on "Reuters" and "Arxiv" datasets. We observe these phrases can express the meaning of the label name.

4.2 Retrieving core documents

When people are asked to tag documents, they will first identify important phrases based on relatedness to the category label.

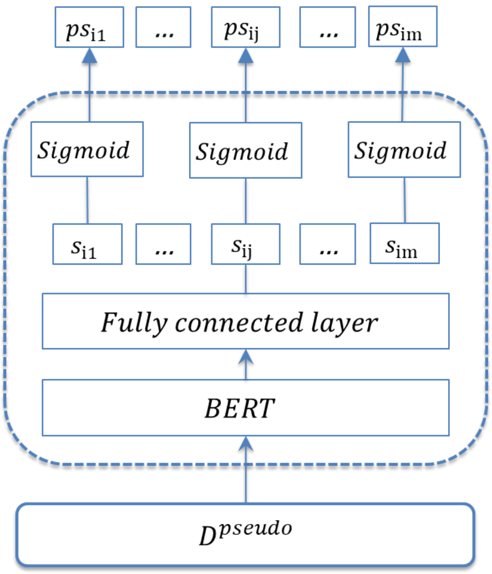


Figure 3. Architecture for fine-tuning BERT

For each label, we identify the top-K_i documents from the selected corpus D^n, called core documents D_i^{core} for label i, which should have high similarity with label i's replaceable phrase list. Because of the long-tailed distributions, the number of core documents for different labels are diverse. When documents containing a label name are little in number, we try to find more documents related to the label from the unlabeled set, by calculating cosine similarity scores between each document in the unlabeled documents D with the top-10 core documents C_i in D_i^{core}. After scoring the documents, we rank them by the decreasing order of the scores and select top- K_i unlabeled documents, and add these documents to core documents D_i^{core} for label i.

In this stage, we utilize Sentence-BERT for generating document embeddings, on which cosine distance is calculated as semantic similarity. We use sentence-transformers/all-MiniLM-L6-v2 [20] as the pre-trained language model, a MiniLM model fine-tuned on a large dataset of over 1 billion training pairs.

4.3 Pseudo labels

After retrieving core documents D_i^{core} for each label, we assign pseudo labels to the core documents. For core documents of label i, each document d is similar to the label, and we give the hard label 1 as the pseudo label.

Because the rest of the labels of document d are unknown, we cannot give the hard label 0 on other labels. For this situation, we first check whether document d appears in the core documents of another label, and we give pseudo label 1 if it appears. Otherwise, we calculate the cosine similarity score between the document d and D_j^{core} of other labels j as follow:

$$L_{d_j} = 1, \quad if \;\; Sim\left(d, D_j^{core}\right) > \alpha \qquad (3)$$

If the cosine scores larger than $\alpha = 0.6$, the hard label 1 is assigned to L_{d_j}, otherwise the hard label 0 is assigned. Let D^{pseudo} be the pseudo-labeled corpus which is the union of D_i^{core}, $i = 1, \dots, m$.

As shown in Figure 3 we construct the multi-label classifier by adding a fully connected layer on top of the BERT pre-trained model. The fully connected layer produce logit vectors $s_i = [s_{i1}, s_{i2}, \dots, s_{im}]$. And train the classifier using the multi-label categorical cross-entropy loss (MCE) to obtain a multi-label classifier, which is proposed by [19] for the multi-label scenario, based on the categorical cross-entropy loss function, derived from the idea of label ranking:

$$L_{MCE}(s_i, y) = \log\left(1 + \sum_{j\in\Omega_i^{neg}} e^{s_{ij}}\right) + \log\left(1 + \sum_{j\in\Omega_i^{pos}} e^{-s_{ij}}\right) \quad (4)$$

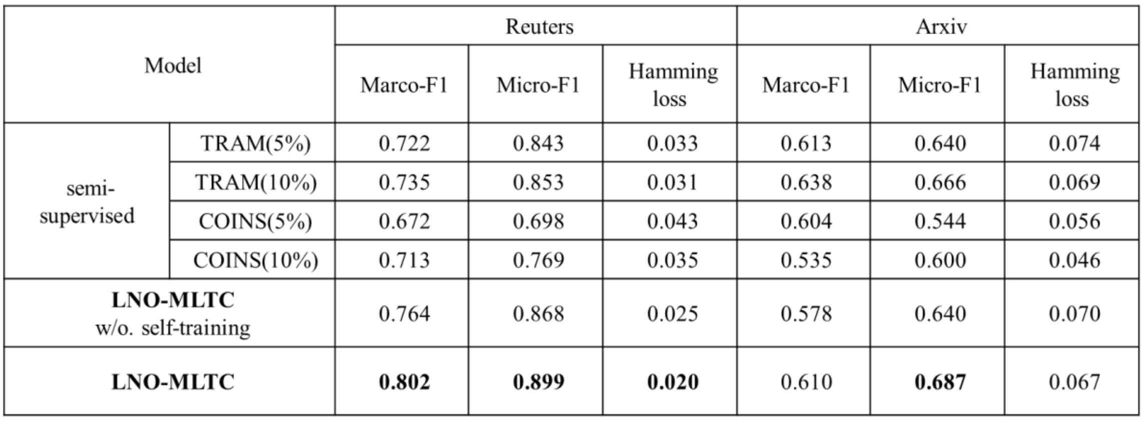
where s_i is the i-th row of the logit matrix of the input data, Ω_i^{neg} and Ω_i^{pos} refer to the sets of the positive and negative labels of the i-th sample.

Then, we use the pseudo-labeled corpus D^{pseudo} as the labeled samples train the multi-label text classifier.

4.4 Multi-label self-training

After training the multi-label classifier based on pseudo-labeled documents D^{pseudo}, there still remain unlabeled documents $D' = D - D^{pseudo}$. We propose to further refine the classifier via self-training on the unlabeled documents D' for better generalization. Self-training is a classic technique in semi-supervised learning, widely used in multi-class classification as well as multi-label classification. The idea has been extended to multi-label classifications [16], in which a confidence estimation is carried out to assign pseudo labels.

In this section, we check the prediction scores of unlabeled documents and introduce pseudo label threshold θ for pseudo label assignment. If the prediction score is close to 1 or 0, it is regarded that the prediction is confident. Based on this assumption, the confident prediction is used to assign pseudo labels for the next round of classifier training. Denote prediction score of each unlabeled document produced by the classifier as $ps = [ps_1, ps_2, \dots, ps_m]$. We select the high confident predictions when $ps_i > \gamma$, $i \in \{1,2,\dots,m\}$. A fixed γ cannot always



Model		Reuters			Arxiv		
		Marco-F1	Micro-F1	Hamming loss	Marco-F1	Micro-F1	Hamming loss
semi-supervised	TRAM(5%)	0.722	0.843	0.033	0.613	0.640	0.074
	TRAM(10%)	0.735	0.853	0.031	0.638	0.666	0.069
	COINS(5%)	0.672	0.698	0.043	0.604	0.544	0.056
	COINS(10%)	0.713	0.769	0.035	0.535	0.600	0.046
LNO-MLTC w/o. self-training		0.764	0.868	0.025	0.578	0.640	0.070
LNO-MLTC		**0.802**	**0.899**	**0.020**	0.610	**0.687**	0.067

Table 3. Results on two datasets

precisely select satisfactory high confidence predictions, the threshold γ will change at each iteration, decreasing by 0.25 from 0.975 in the first iteration. For each high confident prediction, we assign a pseudo label L_i as follow:

$$f(L_i) = \begin{cases} 1, & ps_i \geq \theta \\ 0, & ps_i < \theta \end{cases}, \ \theta = 0.5 \tag{5}$$

When over δ% of unlabeled documents obtain pseudo labels, the BERT classifier is finetuned with the expanded pseudo labels. This self-training process terminates when the iteration reaches the maximum time T = 3.

5. Experiments

5.1 Dataset and Evaluation Metrics

We use two representative multi-label document collections to evaluate our proposed approach.

Reuters [1]: Reuters-21578 dataset collects documents published on the Reuters newswire. We extract 8311 news articles for the topics from the ModApte subset of the Reuters-21578 benchmark, and the first 8000 samples are used as our evaluation dataset. The topics are ‘acq’ (acquisitions), ‘corn’, ‘crude’, ‘earn’, ‘grain’, ‘interest’ (interest rates), ‘money-fx’ (foreign exchange), ‘ship’, ‘trade’ and ‘wheat’.

Arxiv Academic Paper Dataset [17]: A collection of scientific paper abstracts derived from Arxiv. We extract 21000 abstracts from the well-known dataset for 10 subfields under the computer science field. The 10 subfields are ‘cs.ai’ (artificial intelligence), ‘cs.cv’ (computer vision), ‘cs.db’ (databases), ‘cs.ds’ (data structures), ‘cs.lg’ (machine learning), ‘cs.lo’ (logic), ‘cs.it’ (information theory), ‘cs.ro’ (robotic), ‘cs.se’ (software engineering) and ‘cs.pl’ (programming languages).

For multi-label classification, a single metric is difficult to evaluate a model's performance comprehensively. We employ three classical metrics used in extensive previous research on multi-label classification tasks to evaluate the performance, marco-F1, micro-F1 and hamming loss.

Macro-F1 and **Micro-F1:** F1-score is the harmonic mean of precision and recall. Macro-F1 takes the average of each label’s F1-score, Micro-F1 calculates the F1-score overall sample-label pairs.

Hamming loss directly calculates the proportion of misclassified labels. A value of 0 means that all labels for each sample have been assigned the correct label.

5.2 Compared Methods

The performance of our proposed approach is compared against two semi-supervised methods, TRAM and COINS, described as follows:

TRAM [10]: This model estimates the label sets of the unlabeled instances by utilizing the information from labeled and unlabeled data under the transductive setting.

COINS [18]: This model learns from labeled and unlabeled data by adapting the well-known co-training strategy, which naturally works under the inductive setting.

For both of the semi-supervised methods, we conduct experiments under different label rates, and the labeled samples account for exactly 5% and 10% of the size of the dataset.

5.3 Experiment Settings

In the experiments, we finetune the model with Adam [8] as the optimizer and learning rare {5e-5, 2e-5}, training batch size 24, and the test batch size 32. The parameter λ is set to 0.1 and the maximum time T is set to 3. We implement the experiments on a single GPU with NVIDIA GeForce GTX 3090. Our programming environment is based on Python 3.8 and Pytorch 1.8.1 + cu111.

We use the BERT-base-uncased model for the pre-trained language models for both training and self-training step. Moreover, we use the 8-fold cross-validation method [9] on the Reuters dataset to evaluate each model's performance.

5.4 Performance Comparison

Table 3 presents the overall results of all compared methods on the Reuters and Arxiv datasets. Our proposed LNO-MLTC shows the best results, outperforming the semi-supervised methods in most cases, achieving better performance by nearly 5% at the Micro-F1 score on the Reuters dataset. For the Arxiv dataset, our proposed also achieve higher Micro-F1 scores. However, because the documents are related to academic papers and more precisely about computer science, may cause noises at train and self-train steps, the significant improvement is not achieved. Moreover, LNO-MLTC without the self-training shows the effectiveness of our proposed method via the label name to discover core documents. Also, the self-training enhances the performance of the multi-label classifier and demonstrates the effectiveness of our multi-label self-training.

6. Conclusion

In this paper, we propose a model named LNO-MLTC for multi-label document classification under the situation that only label names and unlabeled documents are available. The masked language model is exploited to generate replaceable phrases from each label name. Then the BERT-based multi-label classifier is trained via self-training from core documents which give weak supervisions. Our experimental results demonstrate that our model can alleviate the burden of data labeling to a certain extent and improve the performance of multi-label text classification on unlabeled documents. We believe that further improvements are possible by improve self-training method.

References

[1] Apté, C., et al. "Towards Language Independent Automated Learning of Text Categorization Models." Proceedings of the 17th Annual ACM/SIGIR Conference, 1994.

[2] Bird, Steven, et al. "Natural Language Processing with Python: Analyzing Text with the Natural Language Tooltik." 2009.

[3] Boutell, Matthew R., et al. "Learning Multi-Label Scene Classification." Pattern Recognition, vol. 37, no. 9, 2004, pp. 1757–71.

[4] Brill, Eric. "A Simple Rule-Based Part of Speech Tagger." ANLC '92 Proceedings of the Third Conference on Applied Natural Language Processing, 1992.

[5] Devlin, Jacob., et al. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." Proceedings of the 2019 Conference of the North, 2019.

[6] Hüllermeier, Eyke., et al. "Label Ranking by Learning Pairwise Preferences." Artificial Intelligence, vol. 172, no. 16-17, 2008, pp. 1897–916.

[7] Joshi, Mandar., et al. "SpanBERT: Improving Pre-Training by Representing and Predicting Spans." Transactions of the Association for Computational Linguistics, vol. 8, 2020, pp. 64–77.

[8] Kingma, Diederik P., and Jimmy Ba. "Adam: A Method for Stochastic Optimization." ICLR, 2015.

[9] Kohavi, R. "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection." IJCAI, 1995.

[10] Kong, Xiangnan., et al. "Transductive Multilabel Learning via Label Set Propagation." IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 3, 2013, pp. 704–19.

[11] Meng, Yu., et al. "Text Classification Using Label Names Only: A Language Model Self-Training Approach." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020, pp. 9006–17.

[12] Mikolov, Tomas., et al. "Efficient Estimation of Word Representations in Vector Space." Proceedings of ICLR Workshops Track, 2013.

[13] Mnih, Andriy., and Geoffrey E. Hinton. "A Scalable Hierarchical Distributed Language Model." Neural Information Processing Systems, vol. 21, Curran Associates, Inc., 2009.

[14] Read, Jesse., et al. "Classifier Chains for Multi-Label Classification." Machine Learning, vol. 85, no. 3, 2011, pp. 333–59.

[15] Reimers, Nils., and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks." ArXiv.org, 2019.

[16] Xu, Zhewei., and Mizuho Iwaihara. "Semantic Space-Based Self-Training for Semi-Supervised Multi-Label Text Classification." DEIM Forum E24-2, 2021.

[17] Yang, Pengcheng., et al. "SGM: Sequence Generation Model for Multi-Label Classification." Proceedings of the 27th International Conference, 2018.

[18] Zhan, Wang., and Min-Ling Zhang. "Inductive Semi-Supervised Multi-Label Learning with Co-Training." Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017.

[19] https://www.spaces.ac.cn/archives/7359

[20] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2