An Open-Set Semi-Supervised Multi-Task Learning Framework for Context Classification in Biomedical Texts

Difei Tang[1], Thomas Yu Chow Tam[1], Natasa Miskov-Zivanov[1]

[1]University of Pittsburgh, Pittsburgh, PA.
[1]*{dit18, tht42, nmzivanov}@pitt.edu*

Abstract

In biomedical research, knowledge about the relationship between entities, including genes, proteins, and drugs, is vital for unraveling the complexities of biological processes and mechanisms. Although text mining methods have recently demonstrated great success in biomedical relation extraction. However, such an extraction process often ignores context information like cell type, species, and anatomy, which are crucial components of biological knowledge. Moreover, existing methods addressing this problem as a text classification task are limited by the lack of labeled examples due to costly manual context annotations, which, although can achieve high precision, they perform poorly in unseen contexts. Additionally, despite some attempts to generate more examples automatically from the literature, these methods are often restricted to a fixed generation pattern. This study introduces an open-set semi-supervised multi-task learning framework for biomedical context classification in a practical setting. The proposed scheme assumes that the unlabeled data contains both in-distribution (ID) and out-of-distribution (OOD) examples. The main challenge in context classification is the limited data with sparse distribution across different context types. Therefore, we first build a large-scale context classification dataset using an automatic span annotation method by grounding two manually curated corpora. Next, we develop an outlier detector to properly distinguish the ID and OOD data. Moreover, to capture the inherent relationships between biomedical relations and their associated contexts, the context classification is treated as an individual task, and we design a multi-task (MTL) learning architecture that seamlessly integrates with the semi-supervised learning strategies during training. Extensive experiments on the context classification dataset demonstrate that the proposed method outperforms baselines and efficiently extracts context without requiring many manually annotated data for training.

1 Introduction

In recent years, there has been a significant increase in the number of biomedical texts and scientific articles, providing a rich source for related research and treatment discovery. Spurred by this research boom, researchers have developed natural language processing (NLP) and various text-mining methods to extract knowledge from biomedical text and enhance their understanding of specific problems in the biomedical field. Biological relationships between entities, such as genes, proteins, and drugs, are crucial for unraveling the complexities of biological processes and mechanisms. These relationships can be retrieved from the literature using extraction, a vital part of text mining. However, a common issue with these extracted relations is the absence of pertinent biological context information. Context information, such as cell type, cell line, and species, is vital in interpreting extracted relations in the biomedical domain, as it provides the necessary details to understand the specific biological relevance of extracted relations because interactions can vary depending on the context. For example, a protein-protein interaction may not be observed in a different context of cell type, and the location of which interaction occurs within the cell can also influence the extraction, as different cellular compartments may have distinct molecular environments and regulatory mechanisms. Thus, context-specific relation extraction can help understand more profound, more accurate mechanistic knowledge about extracting findings.

Despite a few studies exploring associating context information with annotations such as recognized entities, events, and protein-protein interactions, several limitations and obstacles remain regarding extracting context information related to

biomedical relations. Firstly, a lack of a golden annotated corpus for this problem significantly hinders the development of relevant text mining methods or tools, especially deep learning methods, which require many manual annotations in the training process. Secondly, not all the context information is explicitly stated in evidence where a biomedical relationship is identified. Current methods can only recognize the associated contexts as they are identified as entities. In addition, the fact that prior research has traditionally conceptualized this issue within the pipeline framework (i.e., identifying context associated with extracted relations) suffers from neglecting the inherent relationship between relation extraction and context identification tasks. Consequently, errors may propagate to the context identification models in practical applications.

While finding a public context association corpus is challenging, [1] present a small corpus of annotated biomedical texts to highlight context-event relations and analyze the context-event inter-sentential relation extraction problem as anaphora resolution. Various classifiers, such as Logistic Regression and Support Vector Machines, have been applied to extract and evaluate the linguistic features using this corpus. In this corpus, an event extraction tool named REACH [2] is used to identify the text spans of event and context entities. Negative example context-event pairs are built assuming human annotators do not annotate the context types. Such NLP tools and classifiers based on traditional features may lead to error propagation, affecting the accuracy of context information extraction.

Thus, [3] attempts to address the problem of associating context with protein-protein interactions (PPIs) using intuitive and interpretable features. Specifically, the authors automatically generate an Insider corpus for training classifiers by extracting sentence-level protein-protein interactions associated with identified context mentions, including cell types and tissues. However, their Insider corpora are limited as they only focus on the protein-protein interactions (single entity pair) extracted from relatively simple sentences, and two context types are considered. Additionally, their study assumes that contexts with the prepositional phrase “in” in the lexicon only imply the association with an interaction, which may not always be accurate. However, several critical restrictions remain as further research and precise context extraction are difficult.

Regarding the last limitation, [4] frame this problem as a text classification task. The authors introduce a multi-modal model by concatenating the token embeddings and graph embeddings to capture the interactions of biomedical relations and contexts, which is evaluated on four context classification tasks for cell line, disease, location, and species. However, in the context annotation classification tasks, the proposed model shows marginal improvement over a baseline NLP model trained solely on textual evidence. Moreover, the overall performance of these models remains relatively low, suggesting that context classification in biomedical texts is a challenging task that requires further research and development to improve the accuracy and robustness of the solutions.

To address the abovementioned problems, we propose an open-set semi-supervised multi-task learning framework for biomedical context classification. Unlike the existing methods, the proposed approach harnesses the advantages of various techniques and seamlessly integrates them into a unified framework. Precisely, to capture the inherent relationships between biomedical relations and their associated contexts, we design a multi-task learning (MTL) architecture by adding two task-specific classification heads on top of a BERT-based backbone model. Furthermore, we utilize semi-supervised learning (SSL) strategies to leverage many unlabeled data when training the framework. Furthermore, to demonstrate the effectiveness of our approach, the overall framework is trained on a dataset constructed from two existing corpora. The low-quality annotations are discarded, and the remaining are further annotated using the developed automatic span annotation method. Finally, extensive experiments on both open-set and close-set scenarios examine the impact of the out-of-distribution data, providing a comprehensive analysis of the proposed framework.

The contributions of this work are as follows:

1. A large-scale context classification dataset is created from existing corpora. This dataset is then annotated by a novel automatic span annotation

```
SET STATEMENT_GROUP = "Group 1"

SET Disease = atherosclerosis
SET Anatomy = artery
SET TextLocation = Review
SET Evidence = "Oxidation and nitration of macromolecules, such as proteins, DNA and lipids, are prominent
in atherosclerotic arteries."
SET Citation = {"PubMed","Trends in molecular medicine","12928037","","de Nigris F|Lerman A|Ignarro LJ|Williams-Ignarro
S|Sica V|Baker AH|Lerman LO|Geng YJ|Napoli C","""}
```

Figure 1: An example of BEL small corpus.

method to reduce labor in annotating text spans of entities involved in relation extraction.

2. A multi-task learning (MTL) architecture is designed to capture the inherent relationships between biomedical relations and their associated contexts. This framework effectively leverages the dataset annotated using the developed automatic span annotation method.

3. SSL strategies are proposed for training the MTL framework to leverage unlabeled data extracted from existing corpora.

4. An out-of-distribution detection method is incorporated to adapt to open-set semi-supervised text classification (OSTC) tasks to compare and evaluate our framework's effectiveness and accuracy in various experimental settings.

2 Background

2.1 Biological Relationships and Context Annotations

Biological relationships are crucial for discovering and representing knowledge from literature. These relationships can be extracted and stored in structured formats such as INDRA statements [5] and BEL language [6]. INDRA (Integrated Network and Dynamical Reasoning Assembler) is integrated with multiple resources that utilize domain-specific statements to represent detailed information about biological relations. BEL (Biological Expression Language) describes scientific findings by representing causal and correlative relationships within relevant contexts, such as the biological and experimental systems, the supporting publications, and the curation process. BEL has made several public knowledge bases available, with the BEL small and large corpora being the most extensive open-access resources. In BEL, contexts are expressed through annotations in the form of “SET <keyword>”, where the keyword can be a prefix of identifiers or a specific category from defined namespaces (e.g., Disease, CellLine). Fig.1 illustrates an example of a BEL small corpus.

2.2 Open-Set Semi-Supervised Learning in Text Classification

SSL is a method that incorporates unlabeled data into the training phase to improve learning performance without extensive manual labeling work. In NLP, SSL text classification (STC) has proven advantageous when obtaining labeled data, which is labor-intensive and time-consuming, while a wealth of unlabeled data is available. One part of STC is based on consistency regularization, which aims to improve the generalization ability of the classification model. For example, Virtual Adversarial Training (VAT) [7] adds a small perturbation to a model, forcing it to make consistent predictions. Unsupervised Data Augmentation (UDA) [8] generates augmented data from unlabeled samples by applying diverse augmentation techniques (e.g., back translation, random insertion). UDA calculates the consistency loss based on the difference in predicted outputs, and the total loss is a sum of the supervised loss on labeled texts and the consistency loss on unlabeled texts. Another approach, Fixmatch [9], combines the advantages of consistency regularization and pseudo-labeling by training the model on strongly-augmented data using the pseudo-labels generated from weakly-augmented data as targets.

Conventionally, STC often makes an unrealistic assumption that all the unlabeled data belong to a known distribution. However, in real-world applications, many unlabeled data may belong to out-of-distribution (OOD) categories not pre-defined in the labeled dataset. Therefore, open-set semi-supervised learning (OSSL) has been proposed to address this challenge by incorporating open-set recognition into a semi-supervised learning framework, i.e., recognizing the data from unseen categories. In the OSSL setting, the most straightforward pipeline detects and discards

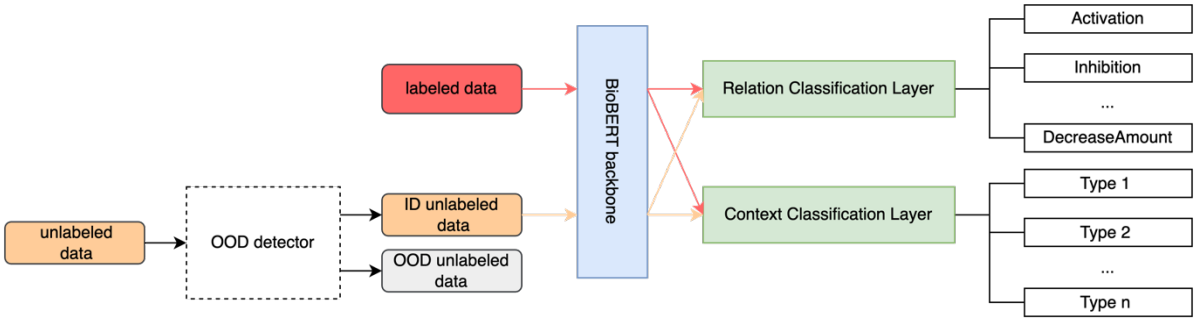

Figure 2: Structure of our open-set semi-supervised multi-task learning framework for biomedical context classification task.

samples from unseen categories before performing classification for known categories. Various OOD detection methods [10] have been proven effective in identifying outliers in text classification tasks, allowing the language model to focus on the in-distribution (ID) data and improve its classification performance.

2.3 Multi-Task Learning

Multi-task learning (MTL) is a technique that has shown significant potential in tackling the problem of overfitting and data scarcity in natural language processing. It has been widely adopted in specialized domains, including biomedical text mining, to leverage the inherent relationships between related tasks through joint training. MTL is usually implemented in two methods: hard and soft parameter sharing [11]. Hard parameter sharing refers to sharing the same model parameters before the task-specific layers, while soft parameter applies constraints to model parameters for different tasks. Recent studies have applied MTL on language models such as BERT and its variants for various biomedical text-mining tasks, demonstrating notable improvements over single-task models. For instance, [12] investigated a multi-task learning model with multiple decoders for tasks such as text similarity and relation extraction, named entity recognition, and text inference and found that MTL fine-tuned models outperformed state-of-the-art transformer models trained on single tasks. Furthermore, [13] proposes a transformer-based framework for few-shot biomedical relation extraction, combined with MTL and a knowledge distillation strategy, that outperforms state-of-the-art few-shot learning methods.

3 Methods

This section describes the proposed method from two perspectives: the dataset used for experiments, which is built on top of existing corpora with an automatic span annotation method, and the proposed context classification methods, including an open-set semi-supervised learning pipeline, a multi-task learning architecture, and the corresponding loss function.

3.1 Dataset

Automatic Span Annotation

In biomedical relation extraction tasks, a model is often trained on a corpus to determine the specific type of relationship between pairs of biomedical entities from unstructured text. Hence, annotators must annotate the particular position of biomedical entities from texts and relation types for each entity pair. However, due to the intricate nature of biomedical ontology and the difficulties of identifying variants of named entities, data annotation for biological relationships requires expertise, resulting in a significant investment of time and effort. Thus, this study introduces an automatic span annotation method that leverages the generalization ability of large language models (LLMs) and existing pre-trained NER models to annotate entity spans in a corpus.

Initially, we manually annotate five cases as few-shot examples for LLMs (GPT-3.5-turbo) and devise an appropriate prompt template. Subsequently, we employ a pre-trained biomedical NER model (en_core_sci_sm) from scispacy [14] to identify the biomedical entities and their corresponding spans within the text. Combining the manually annotated examples with the entity information obtained from the NER model prompted LLMs to automatically annotate the text spans of biomedical entities associated with their relation words. Considering the presence of inaccurately labeled or grounded entities in the original corpus, we exclude instances that could

not be automatically annotated using this method, constituting our biomedical context classification dataset. The following section explains the process of constructing the context classification dataset.

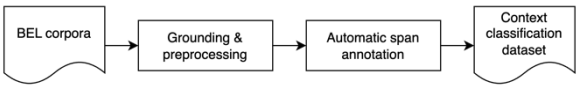

Figure 3: Workflow of constructing our context classification dataset.

Dataset Construction

Previous studies often consider the problem of context extraction as a binary text classification task. Given an identified biomedical relation, a machine learning model should predict the instance to be either positive or negative, indicating whether there is an association between context words and the biomedical relation. However, annotating association pairs from the literature is still challenging, especially the negative ones, due to the low co-occurrence frequency of relations and their context. Thus, we treat the context extraction as a multi-class text classification task.

To our knowledge, relation extraction datasets lack annotating associated context information. Therefore, we create a large-scale context classification dataset based on two manually annotated BEL corpora to address this limitation. Fig.1 presents every piece of evidence (a text sequence of varying lengths, from sentence fragments to entire paragraphs). The same article is bound to a statement group, and the annotated biomedical relations are expressed in BEL and associated contexts. Thus, we extend the context classification tasks [4], aiming to predict which context class a given case of data belongs to. Moreover, we add two more contexts, cell type and organ, to establish the biomedical settings in our dataset while retaining three contexts (i.e., location, cell line, and disease) [4]. Furthermore, we remove the species task from this study, as this information can be readily accessed from public databases such as UniProt [15]. Fig. 3 illustrates a series of preprocessing measures we implement to improve the quality of our context classification dataset. The first stage involves discarding relevant examples once multiple context annotations are present in any text. Next, we use INDRA to normalize these two BEL corpora to eliminate annotations where identified relations and involved entities cannot be mapped to any INDRA ontology and link them to unique identifiers. Lastly, we concatenate these two processed corpora and selected the most frequently annotated contexts within each type to construct our dataset. The dataset comprises five context classification tasks, each with varying numbers of label categories ranging from three to ten. The number of categories depends on the frequency of classes appearing in the corpus. Given the prevalence of no labels in some contexts and to ensure the train and test set has a similar distribution, the dataset was divided into five groups based on the presence of labels in the five contexts. Each row of the data belongs to one of five groups, from having labels across all five contexts to having only one label in one of the contexts (and no labels in other contexts). After the grouping, within each group, the data was randomly split into training and test sets (70:30 ratio) and concatenated to form the final train and test dataset.

3.2 Open-Set Semi-Supervised Learning Pipeline

Unlike previous work, we further transform the context extraction problem into an OSTC task, which is seldom explored in NLP. The reasoning is that context annotations are scarcely distributed in the dataset, and unlabeled texts might include out-of-distribution examples from known categories that cannot be ignored. Hence, to integrate the outlier detection method with semi-supervised learning, we build the pipeline model described in [16], which is implemented in two steps. First, we develop an outlier detector in text classification to recognize the OOD text set for each context. Then, we train an SSL model using the remaining unlabeled text set. Our pipeline model uses the maximum softmax probability (MSP) method [17] for outlier detection. To analyze the effect of data augmentation in STC, we select the VAT and Fixmatch algorithms as our SSL strategies during training.

Fixmatch [9] requires two augmentations (weak and strong). For our adaption in the biomedical domain, we utilize a UMLS-based data augmentation tool [18], named UMLS-eda, to generate weakly and strongly augmented data for our unlabeled texts. Besides, we apply weak augmentation by replacing biomedical terms with

synonyms in the UMLS database. For substantial augmentation, we randomly choose a technique (e.g., random insertion, deletion) from straightforward data augmentation (EDA) [19] to augment our unlabeled texts. Then, the augmented and unlabeled texts are input into the model with the labeled ones in the training phase.

3.3 Multi-Task Learning Architecture and Loss Function

The developed dataset has five context classification tasks, and for each C context classification problem, the training set contains a labeled text set X and an unlabeled text set U. Each context is associated with the identified biomedical relation annotated in the dataset. As a result, we design a multi-task learning architecture consisting of a Bert-based backbone and two multi-class classification heads to capture the relationship between the biomedical relation and its associated context information. This work utilizes the BioBERT model (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) [20] as our backbone, which employs sequence encoders and pre-trained on an extensive collection of biomedical literature. Since we treat each context as an individual task, we adopt the default workflow of multi-class text classification. For all the task-specific labeled examples, we first use the pre-trained tokenizer of BioBERT to obtain the token sequence with padding and truncation strategies equal to the input length (maximum 512 tokens) of BERT models. Then, the token sequence is mapped into contextual embedding vectors as a shared representation between the relation extraction task and context classification task. After that, the first [CLS] token of returned embedding vectors, i.e., the hidden representation of each sequence, is used for the latter classification. Besides, the standard cross-entropy loss function computes the logits and true labels to get the class probabilities of each context. Fig. 2 depicts an overview of the proposed architecture.

In our dataset, the biomedical relations between pairs of entities are annotated for all data, while the associated context annotations are sparsely distributed. For the kth context classification task, let the labeled text set be S_k and the unlabeled text be U_k, where $k \in \{1, \dots, 5\}$ due to having five context types. Note that that our "unlabeled" data U_k refers to the text examples that only have relation labels. In an open-set scenario, we use an outlier detector to split U_k into two subsets U_k^{OOD} and U_k^{ID}, representing the OOD and ID unlabeled data, respectively. Then, we mix the labeled data and ID unlabeled data in a minibatch regarding our selected SSL algorithms during the training. Let $s = \{(x_i,\ y_i,\ z_i)\}_{i=1}^{n_s}$ be the labeled samples in a batch. x_i represent the ith text input, and it has two corresponding labels $y_i \in \{0, \dots, C-1\}$ belonging to one of the task-specific classes and $z_i \in \{0, \dots, R-1\}$, where R is determined by the grounded relation types from INDRA. For convenience, we denote the ID unlabeled text set in a batch as $u^{ID} = \{u_i : i \in \{1, \dots, n_u^{ID}\}\}$. Next, we will elucidate the integration of our chosen SSL algorithms into the MTL architecture in an open-set scenario and compare them with our baselines. To integrate SSL algorithms into a standard multi-task learning architecture (hard parameter sharing), the loss of each data should comprise a context-specific loss $l_{context}$ and a relation classification loss $l_{relation}$. Therefore, the loss of the labeled data is a weighted sum of the two supervised losses:

$$L_s = \lambda_a \frac{1}{n_s} \sum_{i=1}^{n_s} H\,(y_i, p_i) + \lambda_b \frac{1}{n_s} \sum_{i=1}^{n_s} H\,(z_i, p_i')$$

where H is the typical cross-entropy loss function and p_i, p_i' are the predicted results for context and relation classification given labeled examples.

For the unlabeled data u_k^{ID}, according to the consistency regularization in STC, when context annotations are absent, the context-specific loss comprises a supervised loss l_{sup} and an unsupervised loss l_{unsup}, formulated as follows:

$$l_{context} = \lambda_c \frac{1}{n_u^{ID}} \sum_{i=1}^{n_u^{ID}} l_{sup} + \lambda_d \frac{1}{n_u^{ID}} \sum_{i=1}^{n_u^{ID}} l_{unsup}$$

$$L_{u^{ID}} = \lambda_c l_{context} + \lambda_d l_{relation}$$

where l_{sup}, $l_{relation}$ are computed using a cross-entropy function and l_{unsup} depends on the SSL algorithm described below.

VAT: The unsupervised loss is:

$$l_{unsup} = \frac{1}{n_u^{ID}} \sum\nolimits_i^{n_u^{ID}} KL\,(p_i, p_i(y|u_i + r_{adv}))$$

where KL denotes the Kullback-Leibler Divergence, p_i is the predicted probability distribution of input logits and r_{adv} is the adversarial perturbation added to the unlabeled data u_i.

Fixmatch: Inspired by Fixmatch, the loss function comprises a supervised loss l_s for labeled texts and an unsupervised loss l_u for ID texts. Data augmentation is performed using UMLS-eda, and we denote the weak and robust augmentation as α and A, respectively. To obtain the pseudo-label of U_k^{in}, the class probability distribution q_i is first computed using the weakly augmented data given unlabeled ID texts, and τ is the threshold. Then, we take $\hat{q}_i = argmax(q_i)$ as the pseudo-label. The unsupervised loss is another cross-entropy loss between pseudo-labels and the output of the strongly augmented u_i. Consequently, the context-specific loss function is as follows:

$$l_{unsup} = \frac{1}{n_u^{ID}} \sum_{i=1}^{n_u^{ID}} \mathbb{I}(max(q_i) > \tau)\, H(\hat{q}_i, p_m(y|A(u_i)))$$

Finally, the total loss for the backpropagation during training is calculated by combining the losses for the labeled and unlabeled data.

$$L_{total} = L_s + \lambda_u L_{u^{ID}}$$

where λ_u is a scalar parameter representing the weight of loss for unlabeled data.

3.4 Implementation

As described in Section 3.1, the dataset created comprises five context classification tasks, where two BEL corpora are filtered and grounded using INDRA to remove the inaccurate associations between annotated relations and contexts. Additionally, we annotate the entity spans for each biomedical relation within the texts using the proposed automatic span annotation method. The baseline model uses the huggingface [21] package under the default BioBERT parameter settings suggested for sequence classification tasks. The BioBERT model is fine-tuned with 20 epochs as our outlier detector for the MSP method and set the threshold as 0.9. The proposed models are trained with the Adam optimizer [22] for 50 epochs, with a training batch size of eight and a learning rate of 5e-5. Early stopping is applied, and the training is terminated if there is no improvement in the evaluation score for 10 consecutive epochs. The experiments for all SSL algorithms adopt the default parameters presented in [23]. Regarding the MTL settings, we manually tune the relative weights of two tasks to maximize the predicted score of our model for context classification. All experiments are conducted on the Center of Research Computing (CRC) server using a Nvidia a100 GPU.

The predicted evaluation results are evaluated on the test set using the most common metric, micro f1. The Results section exclusively reports the F1 scores of the context classification task, evaluating the models' efficiency in comprehending and categorizing contexts. The relation extraction task provides additional insights for predicting the association between them.

Table 1: Statistics of context classification dataset on the close-set scenario.

Dataset	Close-Set # of instance				
	Location	Cell Line	Cell Type	Disease	Organ
Train-unlabeled	16952	16390	15196	16790	15352
Train-labeled	230	792	1986	392	1830
Test	103	337	853	168	786

Table 2: Statistics of context classification dataset on the open-set scenario.

Dataset	Open-Set # of instance				
	Location	Cell Line	Cell Type	Disease	Organ
Train-OOD	1809	3220	1124	3212	1745
Train-ID	15143	13170	14072	13578	13607
Train-labeled	230	792	1986	392	1830

4 Results

This section briefly introduces the experimental results and the performance of SSL models on our context classification dataset. Then, we report the scores of the proposed MTL-SSL models using different strategies and discuss the discrepancies with SSL models. Finally, this section comprehensively analyzes and compares the results of various models and performs a simple ablation study.

4.1 Dataset Characteristics and Statistics

The context classification dataset contains 19004 examples from 50 biomedical articles adapted from BEL corpora and is available on our GitHub repository. Table 1 reports the general statistics of our dataset on close-set scenario. In Table 2, we show the number of labeled, OOD and ID data in the training set.

Table 3: Performance comparison of baseline, MTL, SSL, OSSL, and the proposed models on the context classification dataset.

Method	Context Classification Task				
	Location	Cell Line	Cell Type	Disease	Organ
Baseline	0.7087	0.7211	0.7233	0.7917	0.6921
MTL	0.7488	0.7400	0.7238	0.7900	0.6737
VAT	0.7312	0.7450	0.7262	0.8037	0.6863
Fixmatch	0.7025	0.7438	0.7362	0.8087	0.6850
MSP+VAT	0.7087	0.7488	0.7412	0.8063	0.6875
MSP+Fixmatch	0.7113	0.7262	0.7325	0.8050	0.6850
MTL+VAT	0.7650	**0.7562**	**0.7438**	0.8137	**0.7113**
MTL+Fixmatch	0.7688	0.7325	0.7188	0.8237	0.6913
MTL+MSP+VAT	**0.7775**	0.7450	0.7375	**0.8287**	0.6875
MTL+MSP+fixmatch	0.7625	0.7262	0.7250	0.8113	0.6825

4.2 Performance of Semi-Supervised Learning Models

Except for the organ classification task, the SSL models outperform the baseline by at least 1%, demonstrating their effectiveness in scenarios where supervised data is scarce. The performance of the Fixmatch models is relatively inferior to that of the VAT models. Interestingly, there is a slight improvement in VAT models with the out-of-distribution detection method, while this method seems detrimental to the Fixmatch models across most classification tasks. Table 3 reveals that most of the unlabeled data are classified as In-distribution data with a threshold of 0.9, which indicates that such an approach may not substantially impact context classification tasks in our circumstances. This also implies that the unlabeled data is high quality and should be annotated with the same context categories as the labeled data.

4.3 Performance of Multi-Task Semi-Supervised Learning Models

Table 3 highlights that the MTL models are superior to SSL and fully supervised models in single-task experiments for location and disease classification. In contrast, SSL models achieve competitive cell line and type classification performance. We find significant improvements in MTL+VAT models in both close-set and open-set scenarios. For example, the MTL+MSP+VAT model performs best in location and disease classification tasks, attaining 77.75% and 82.87% F1 scores, respectively. Additionally, the MTL+VAT model surpasses other cell line and type classification models. Notably, a slight improvement is observed only for the location and disease classification tasks using the MSP method

5 Conclusion

This work combines multi-task and semi-supervised learning approaches for context classification in the biomedical domain. Additionally, we constructed a large-scale biomedical dataset from existing corpora to reduce labor effort, and entity spans are further annotated using an automatic span annotation method to improve the performance of relation extraction. The problem of identifying context associated with biomedical relations is constrained as a context classification task. Thus, we design an MTL architecture to capture the interaction between relation extraction and context classification tasks. Furthermore, we utilize the SSL approach to use labeled and unlabeled data from the dataset.

The experimental results demonstrate that the proposed approach performs better than the baseline model, single MTL model, and SSL models in most tasks. Besides, the performance of different models is compared in both open-set and close-set settings, demonstrating that in a fixed domain, the OOD method may negatively impact the model's performance. Finally, a detailed analysis for each context classification task is conducted by selecting representative examples for manual evaluation. Future work includes adapting the proposed approach to other information extraction tasks in the biomedical domain and exploring more advanced SSL and OOD methods to improve prediction accuracy.

Acknowledgments

This research was supported in part by the University of Pittsburgh Center for Research Computing, RRID:SCR_022735, through the resources provided. Specifically, this work used the H2P cluster, which is supported by NSF award number OAC-2117681.

Conflict of Interest: none declared.

References

1. Noriega-Atala, E., et al., *Extracting inter-sentence relations for associating biological context with events in biomedical texts.* IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2019. **17**(6): p. 1895-1906.
2. Valenzuela-Escárcega, M.A., et al., *Large-scale automated machine reading discovers new cancer-driving mechanisms.* Database, 2018. **2018**: p. bay098.
3. Sosa, D.N., et al., *Associating biological context with protein-protein interactions through text mining at PubMed scale.* Journal of Biomedical Informatics, 2023. **145**: p. 104474.
4. Balabin, H., et al., *STonKGs: a sophisticated transformer trained on biomedical text and knowledge graphs.* Bioinformatics, 2022. **38**(6): p. 1648-1656.
5. Gyori, B.M., et al., *From word models to executable models of signaling networks using automated assembly.* Molecular systems biology, 2017. **13**(11): p. 954.
6. Slater, T., *Recent advances in modeling languages for pathway maps and computable biological networks.* Drug discovery today, 2014. **19**(2): p. 193-198.
7. Miyato, T., et al., *Virtual adversarial training: a regularization method for supervised and semi-supervised learning.* IEEE transactions on pattern analysis and machine intelligence, 2018. **41**(8): p. 1979-1993.
8. Xie, Q., et al., *Unsupervised data augmentation for consistency training.* Advances in neural information processing systems, 2020. **33**: p. 6256-6268.
9. Sohn, K., et al., *Fixmatch: Simplifying semi-supervised learning with consistency and confidence.* Advances in neural information processing systems, 2020. **33**: p. 596-608.
10. Baran, M., et al., *Classical Out-of-Distribution Detection Methods Benchmark in Text Classification Tasks.* arXiv preprint arXiv:2307.07002, 2023.
11. Chen, S., Y. Zhang, and Q. Yang, *Multi-task learning in natural language processing: An overview.* ACM Computing Surveys, 2021.
12. Peng, Y., Q. Chen, and Z. Lu, *An empirical study of multi-task learning on BERT for biomedical text mining.* arXiv preprint arXiv:2005.02799, 2020.
13. Moscato, V., et al., *Multi-task learning for few-shot biomedical relation extraction.* Artificial Intelligence Review, 2023. **56**(11): p. 13743-13763.
14. Neumann, M., et al., *ScispaCy: fast and robust models for biomedical natural language processing.* arXiv preprint arXiv:1902.07669, 2019.
15. Consortium, U., *UniProt: a hub for protein information.* Nucleic acids research, 2015. **43**(D1): p. D204-D212.
16. Chen, J., et al. *Open-Set Semi-Supervised Text Classification with Latent Outlier Softening*. in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2023.
17. Hendrycks, D. and K. Gimpel, *A baseline for detecting misclassified and out-of-distribution examples in neural networks.* arXiv preprint arXiv:1610.02136, 2016.
18. Kang, T., et al., *UMLS-based data augmentation for natural language processing of clinical research literature.* Journal of the American Medical Informatics Association, 2021. **28**(4): p. 812-823.
19. Wei, J. and K. Zou, *Eda: Easy data augmentation techniques for boosting performance on text classification tasks.* arXiv preprint arXiv:1901.11196, 2019.
20. Lee, J., et al., *BioBERT: a pre-trained biomedical language representation model for biomedical text mining.* Bioinformatics, 2020. **36**(4): p. 1234-1240.
21. Wolf, T., et al., *Huggingface's transformers: State-of-the-art natural language processing.* arXiv preprint arXiv:1910.03771, 2019.
22. Kingma, D.P. and J. Ba, *Adam: A method for stochastic optimization.* arXiv preprint arXiv:1412.6980, 2014.
23. Wang, Y., et al., *Usb: A unified semi-supervised learning benchmark for classification.* Advances in Neural Information Processing Systems, 2022. **35**: p. 3938-3961.