Towards More Generalizable and Accurate Sentence Classification in Medical Abstracts with Less Data

Yan Hu
The University of Texas Health Science Center at Houston
Yong Chen
University of Pennsylvania
Hua Xu (hua.xu@uth.tmc.edu)
The University of Texas Health Science Center at Houston

Research Article

Keywords: Natural Language Processing, Prompt learning, Few-shot learning, Text Classification, Medical Abstracts

Posted Date: September 8th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2026270/v1

Towards More Generalizable and Accurate Sentence Classification in Medical Abstracts with Less Data

Yan Hu[1], Yong Chen[2] and Hua Xu[1*]

[1*]School of Biomedical Informatics, University of Texas Health Science at Houston, 7000 Fannin St, Houston, 77030, TX, USA.
[2]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, 423 Guardian Dr, Philadelphia, 19104, PA, USA.

*Corresponding author(s). E-mail(s): hua.xu@uth.tmc.edu;
Contributing authors: yan.hu@uth.tmc.edu;
ychen123@pennmedicine.upenn.edu;

Abstract

Purpose: With the unprecedented growth of biomedical publications, it is important to have structured abstracts in bibliographic databases (i.e., PubMed), thus, to facilitate the information retrieval and knowledge synthesis in needs of researchers. **Methods:** Here we propose a few-shot prompt learning-based approach to classify sentences in medical abstracts of randomized clinical trials (RCT) and observational studies (OS) to subsections of Introduction, Background, Methods, Results, and Conclusion, using an existing corpus of RCT (PubMed 200k/20k RCT) and a newly built corpus of OS (PubMed 20k OS). 5 manually designed templates in a combination of 4 BERT model variants were tested and compared to a previous Hierarchical Sequential Labeling Network architecture and traditional BERT-based sentence classification method. **Results:** On the PubMed 200k and 20k RCT datasets, we achieved overall F1 scores of 0.9508 and 0.9401 respectively. Under few-shot settings, we demonstrated that only 20% of training data is sufficient to achieve a comparable F1 score by the HSLN model (0.9266 by us and 0.9263 by HSLN). When trained on the RCT dataset, our method

achieved a 0.9065 F1 score on the OS dataset. When trained on the OS dataset, our method achieved a 0.9203 F1 score on the RCT dataset. **Conclusion:** We show that the prompt learning-based method outperformed the existing method, even when fewer training samples were used. Moreover, the proposed method shows better generalizability across two types of medical publications when compared with the existing approach.

Keywords: Natural Language Processing, Prompt learning, Few-shot learning, Text Classification, Medical Abstracts

1 Introduction

The number of biomedical publications keeps growing. As of February 25th, 2022, there are a total of over 40,000,000 publicly available abstracts in PubMed, a widely used bibliographic database that helps researchers find articles of interest. Structural information (e.g., subsections such as Background, Objective, Methods, Results, and Conclusion) of medical abstracts is useful for various Natural Language Processing (NLP) tasks such as information retrieval, information extraction, and text summarization. Nevertheless, many abstracts in PubMed are not organized with subsections. According to Franck , over a half of all the abstracts in PubMed provide their structural information [1], limiting efficient information retrieval and knowledge discovery from those large-scale biomedical bibliographic databases. Therefore, it is valuable to develop automated methods to classify sentences in biomedical abstracts into different subsections.

A number of studies have investigated approaches for classifying sentences in medical abstracts. Dernoncourt et. al. (2017) conducted the first study on this topic [1]. In their study, they developed a corpus named "PubMed 200K RCT", which contains a total of 200,000 abstracts of Randomized Controlled Trials with labeled structure information, as well as a subset of 20K RCT abstracts. The benchmark model achieved an F1 score of 0.900 on the 20K subset and an F1 score of 0.916 on the complete 200K dataset. Later, Jin and Szolovits (2018) reported an F1 score of 0.926 on the 20K subset and an F1 score of 0.939 on the 200K dataset by using the hierarchical sequential labeling network (HSLN) [2]. More recently, Srivastava et al. (2019) have pushed the performance on this task even further, achieving a state-of-the-art (SOTA) performance of F1 scores of 0.928 on the 20K dataset and 0.941 on the 200K dataset [3]. Despite promising results reported by previous studies, two main challenges still exist: (1) current models are based on data from a sub-domain (e.g., RCT) and its generalizability to other types of studies is unknown; and (2) it requires a significant amount of annotated data. Although annotated datasets can be generated using structured abstracts, more annotations from real-world, unstructured abstracts will be ideal for testing its generalizability.

Therefore, methods that can achieve good performance while requiring fewer annotated samples would be valuable.

To address those challenges, we propose to investigate prompt learning (PL) for sentence classification of medical abstracts. PL is an emerging framework that attempts to use the knowledge from the Pretrained Language Model (PLM) without introducing the extra layers for the downstream tasks [4]. Previously, the prevalent way of performing NLP tasks is basically encoding text x by a PLM, then PLM passes the encoded $x_{encoded}$ to a classifier that can predict $P(y|x_{encoded};\ \theta)$, where θ is the parameters of the model, y is the target output [4]. Instead of predicting $P(y|x_{encoded};\ \theta)$ using extra models or extra parameters, PL encodes the candidate outputs y as $y_{encoded}$, then allows the PLM to calculate the probability of $y_{encoded}$ to be in the $x_{encoded}$ and choose the $y_{encoded}$ with highest probability as the final output y. By doing so, there are mainly two advantages : (1) Fewer parameters are needed for deep-learning-based approaches because no extra model or layer is needed [5]; and (2) Fewer data are needed to achieve a comparable performance because of the hidden knowledge in PLM [6].

Researchers have applied prompt learning to the sentence classification task. One approach is to convert the sentence classification task to the cloze question format by PL. More specifically, it puts a [mask] within a text template x, defines a list of vocabulary V, and lets the PLM decide which word $y \in V$ has the highest probability to be in the mask place. And there are several works have been done to illustrate the great potential of PL. For example, Gao et al. (2021) defined a template 'No reason to watch. It was [mask]' and let the V be 'Great', 'Terrible', etc. to perform a sentiment classification in movie reviews, demonstrating that a small amount of training data can lead to comparable results as the whole training data does [7]. Schick and Schutze (2021) combined the results from several templates x to decide on a single output y in text classification, which again shows the ability of PL to reduce the training data [8]. Zhu (2022) tested several templates like 'A [mask] news : [News]' to classify the news, beating the SOTA results in AG news, Snippets, and News title classification tasks [9].

Despite the promising results of PL in text classification, it has not been applied to the sentence classification task for medical abstracts. In this study, we developed a prompt-learning-based sentence classification method for medical abstracts from both randomized clinical trials (RCT) and observational study (OS). Our results show that the proposed PL model not only reduces the number of required training samples, but also shows better generalizability across medical article types, when compared with the existing HSLN method. The main contributions of this work include: (1) a new corpus of OS abstracts for sentence classification in medical abstracts; (2) a new prompt-learning-based sentence classification method that performs well with less training data; (3) a sentence classification model with good generalizability for different types of medical publications.

2 Methods and Materials

2.1 Study Design

Figure 1 shows the overall study design and workflow. First, we designed 5 different templates for PL. Second, we tested 4 different PLMs (BERT, BioBERT, RoBERTa, and PubMedBERT) and decided on the best one for following experiments. Then we compared the PL method with the HSLN method by Jin and Szolovits (2018), using the benchmark datasets of PubMed RCT 200K and 20K. Finally, we conducted a generalizability study of the PL method: we evaluated the model trained using RCT on the OS dataset (and vice versa).

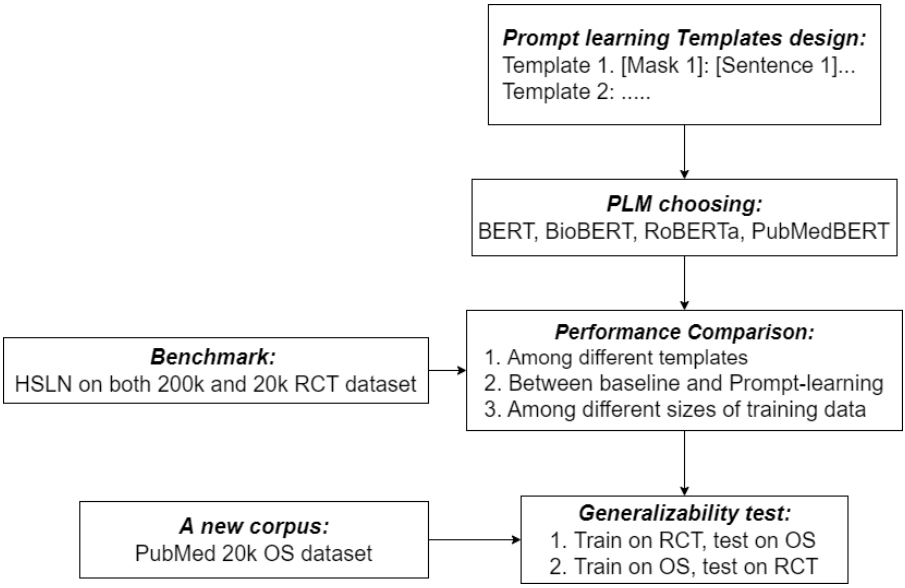

Fig. 1 The overall design and workflow of this study

2.2 Datasets

For RCT, the PubMed 200K RCT dataset and the 20K subset of it were used. These two datasets are directly available at https://github.com/Franck-Dernoncourt/pubmed-rct. To test the generalizability of the sentence classification model, a dataset of Observational Study (OS) called PubMed 20k OS was created from scratch. There is a total number of 122,329 OS abstracts available on PubMed and 39,762 of them are structured. We collected all the structured OS abstracts and randomly picked 20,000 abstracts out of them as a subset. To be noted, even for the structured OS abstracts, they do not have standardized structures. For example, some of the abstracts used OBJECTIVE instead of OBJECTIVES (PMID: 27008686), and some of

them used subtitles like IMPORTANCE (PMID: 29799986) and RESEARCH DESIGN AND METHODS (PMID: 28655740). To standardize all the heterogenous subtitles, National Library of Medicine (NLM) attempted to category them into 5 classes, which is OBJECTIVE, INTRODUCTION, METHODS, RESULTS, and CONCLUSIONS, and published a mapping file for it. The subtitle and corresponding standardized labels is available at https://lhncbc.nlm.nih.gov/ii/areas/structured-abstracts.html. However, the mapping file still does not cover all the subtitles appeared in the OS corpus. For those subtitles that do not exist in the mapping file, we use fuzzy matching to find the most similar one among the mapper. Levenshtein distance was used to find the most similar subtitle existing in the mapping file and the corresponding label was then assigned. After standardization, all the subtitles are converted to one of the five subtitles. We followed the PubMed 200K RCT format to construct the dataset, one of the example abstract (PMID: 31872515) is shown in Figure 2. The PMID is in the first line and in the remaining lines, the first coloumn is the standardized label and the second coloumn is the sentence.

```
###31872515:
OBJECTIVE       To investigate the optimal dual antiplatelet the
BACKGROUND      Diabetic patients are at high risk for ischemic
METHODS This post hoc study analyzed DAPT duration and clinical
METHODS The primary outcome was patient-oriented composite endpo
METHODS The safety outcome was bleeding events.
RESULTS In total, 1,773 diabetic patients with ACS were enrolle
RESULTS Premature DAPT discontinuation before 12 months was an
RESULTS It was associated with a significantly higher risk of P
RESULTS No excess bleeding risk was found in patients who recei
CONCLUSIONS     Premature DAPT discontinuation before 12 months
```

Fig. 2 Example of a standradized abstract

2.3 Design of PL templates

For PL, the design of templates is important, as it may greatly impact the model performance. In this task, we tested several manually-designed templates as shown in Table 1, where the [Mask] and [Sentence] are two variables and <Begin>and <End>are two fixed textual tokens. [Mask n] denotes the standardized subtitle of the [Sentence n], [Title] denotes the Title of that abstract. <Begin>and <End>are two tokens which were added to indicate the beginning or the end of that abstract. Taking PMID 31872515 in Figure 2 as an example, the [Mask 1] denotes [OBJECTIVE] and the [Sentence 1] denotes [To investigate the optimal dual...]. The [Mask 2] denotes [BACKGROUND] and the [Sentence 2] denotes [Diabetic patients are at...]. Under each template in Table 1, we are showing two example inputs respectively in Table 2 for clearer illustration. The standardized subtitles were showed in training, while

they were kept masked in prediction. In our results, all the template indexes are corresponding to the ID column in Table 1.

Here we introduce the idea of how these templates are designed. The first template imitates the way of naturally written structured sentences. The second template is a typical way of how researchers do text classification based on PL. For the first 2 templates, the classification is simply based on the sentences themselves. We noticed that in this task, not only the sentences themselves but also their contextual information should be taken into account. Hence, in the third template, we tried to feed the previous sentence together with the sentence that needs to be classified. In the fourth and fifth templates, we fed the whole abstract to the PLM, trying to give as much contextual information as we can.

Table 1 The templates used in our experiments

ID	Templates
1	[Mask n]: [Sentence n]
2	[Sent n] This sentence belongs to the [Mask n] section.
3	[Sentence n-1] [Mask n]: [Sentence n]
4	In [Sentence 1] [Sentence 2]... [Sentence n]. The sentence "[Sentence n]" belongs to the [Mask n] section.
5	<Begin>[Sentence 1] [Sentence 2] ... [Mask n]: [Sentence n] <End>

In template 3, when [Sentence n] is the first sentence, [Title] was used instead of [Sentence n-1].

2.4 Benchmark architecture

In our experiments, we have two baselines, one of which is Jin and Szolovits's HSLN architecture [2] and the other one is transformer-based sentence classification method.

Despite the SOTA method in this task is the Hierarchical Capsule Based Neural Network Architecture (HCBNN) proposed by Srivastava [3], the code of it is not publicly available. We were not able to repeat their work. Instead, we used Jin and Szolovits's HSLN architecture [2], which has similar performance as Srivastava [3], as a baseline in our study. The Hierarchical Sequential Labeling Network (HSLN), by name, leverages 4 different hierarchical layers. It includes a traditional token-level embedding layer, a sentence-level encoding layer by CNN or bi-RNN, a context enriching layer by bi-LSTM and a sequential labeling layer by CRF. In our experiments, we trained and tested the baseline HSLN model based on both PubMed 200K RCT dataset and 20K subset. All the training settings for HSLN remained the default as

Table 2 The input examples of each template

ID	Input examples
1	OBJECTIVE : To investigate the optimal dual...
	BACKGROUND: Diabetic patients are at...
2	To investigate the optimal dual... This sentence belongs to the OBJECTIVE section.
	Diabetic patients are at... This sentence belongs to the BACKGROUND section.
3	Impact of dual antiplatelet therapy... OBJECTIVE: To investigate the optimal dual...
	To investigate the optimal dual... BACKGROUND: Diabetic patients are at...
4	In To investigate... with high-risk profiles. The sentence "To investigate ..." belongs to the OBJECTIVE section.
	In To investigate... with high-risk profiles. The sentence "Diabetic patients..." belongs to the BACKGROUND section.
5	<Begin>OBJECTIVE : To investigate the optimal dual... Diabetic patients are at... with high-risk profiles. <End>
	<Begin>To investigate the optimal dual... BACKGROUND: Diabetic patients are at... with high-risk profiles. <End>

provided in Jin's Github link (available at https://github.com/jind11/HSLN-Joint-Sentence-Classification).

To demonstrate the advantages of PL compared to Non-PL method, we also tested transformer-based sentence classification method using the same pretrained models as PL methods.

2.5 PLMs

PLMs may play an important role in PL because different PLMs utilize different algorithms and can be trained using different corpora. BERT and RoBERTa were trained on general corpus like books and wiki, which have a good understanding of general English. While BioBERT and PubMedBERT were pre-trained on PubMed articles, which should have a better understanding of PubMed abstracts. Here we tested the performance of BERT, BioBERT, RoBERTa, and PubMedBERT models to evaluate their performance on this task [10] [11] [12] [13].

2.6 Experimental settings

One of the important advantages of PL is to elicit the knowledge of the PLM; thus, the amount of training data needed can be reduced [14]. To examine how

prompt-learning-based methods and HSLN perform in few-shot settings, we evaluated model performance when different percentages of training data were used. We divided both PubMed 200K RCT and PubMed 20K RCT training data into 5 different subsets according to different percentages (5%, 10%, 20%, 50%, and 100%). The validation data and test data remained unmodified.

In our experiments, all the training processes were based on the training set, all the hyper-parameter tunings were based on the validation set and all the displayed F1 scores were based on the test set. The max sequence length for each input instance was set at 512. The excess sequence will be truncated from the head when the length of the input instance exceeded 512. The batch size for each loop is set at 8. The Adaptive Moment Estimation with Decoupled Weight Decay (AdamW) [15] optimizer was leveraged. We set the weight decay at 0.01 and the learning rate at 10^{-6} for AdamW. We stopped the training loop after there is no improvement in the F1 score on the validation data in 2 epochs.

2.7 Evaluation

To make fair comparison with the HSLN, we used the same evaluation metrics as it as well, which is Precision, Recall, and F1 score. In the overall performance calculations, we used weighted F1 score to align with the HSLN also.

The equations of these scores are shown as following:

$$Precision = TP/(TP + FP)$$

$$Recall = TP/(TP + FN)$$

$$F1 = 2 * Precision * Recall/(Precision + Recall)$$

where TP, FP, and FN are True Positive, False Positive and False Negative, respectively.

2.8 Generalizability test

To be applicable to different types of medical abstracts, and to reduce the need for manually annotated data when applying to unstructured abstracts of different publications, a generalizable model is preferred. We compared the generalizability of PL and HSLN. We firstly trained the models using the RCT corpus (with different percentages of training data) and then tested them on the test set of the OS corpus. The same procedure was repeated by using the OS corpus for training and the RCT corpus for testing. By doing so, the models' ability to predict the unseen data was tested.

3 Results and Discussion

3.1 Corpus statistics

The statistics of the PubMed 20k OS and PubMed 20k RCT corpus were shown in Figure 3. As the figure shows, the distribution of 5 classes is not balanced

in 2 corpora. RESULTS is the most frequent class while OBJECTIVE is the least class. In the RCT corpus, the number METHODS and RESULTS are similar. While in the OS corpus, RESULTS are much more than METHODS.

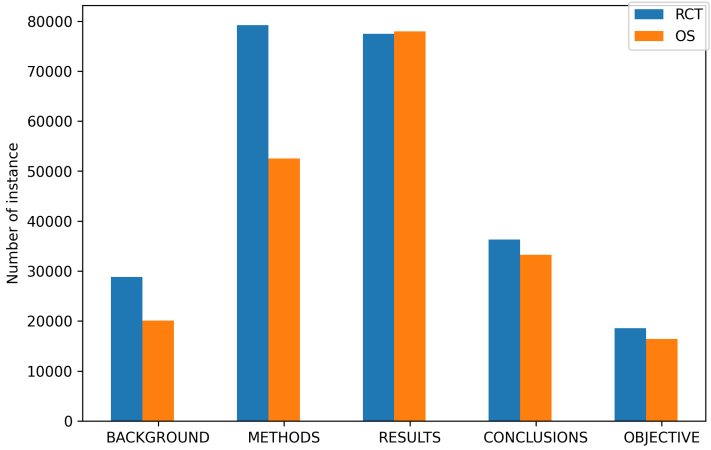

Fig. 3 Distribution of 5 classes in 20k RCT and 20k OS dataset

3.2 Baseline Peformance by HSLN

The average F1 score for HSLN we achieved was 0.9263 on the 20k subset, which aligned with the 0.926 F1 score reported in Jin's original paper. The average F1 score we achieved was the 0.9332 on the 200k dataset, which was lower than (but close to) 0.939, the F1 score reported by Jin. The detailed results are shown in Table 3. In the following experiments, we took the results from this experiment as the baseline F1 score.

Table 3 Baseline performance by HSLN architecture

Dataset	Class	Precision	Recall	F1
200k	OBJECTIVE	0.7293	0.7989	0.7625
	BACKGROUND	0.8009	0.7221	0.7595
	METHODS	0.9615	0.9797	0.9705
	CONCLUSIONS	0.9759	0.9600	0.9679
	RESULTS	0.9712	0.9638	0.9675
	Average	0.9338	0.9334	0.9332
20k	OBJECTIVE	0.7197	0.7218	0.7207
	BACKGROUND	0.8135	0.8048	0.8091
	METHODS	0.9567	0.9727	0.9646
	CONCLUSIONS	0.9744	0.9591	0.9667
	RESULTS	0.9645	0.9583	0.9614
	Average	0.9263	0.9264	0.9263

3.3 Performance by different PLMs

To find out the optimal PLM in this task, we conducted experiments on BERT, BioBERT, RoBERTa, and PubMedBERT models using 5 templates on the PubMed 20K RCT training data. The results are shown in Figure 4. As expected, PubMedBERT outperforms all other PLMs in all different templates and Non-PL method, as it was pre-trained using PubMed articles. Thus, PubMedBERT was chosen as the PLM in the next experiments. But the performance of domain-adapted PLM (BioBERT and PubMedBERT) and general-purpose PLM (BERT and RoBERTa) differ only around 0.02 on F1 score.

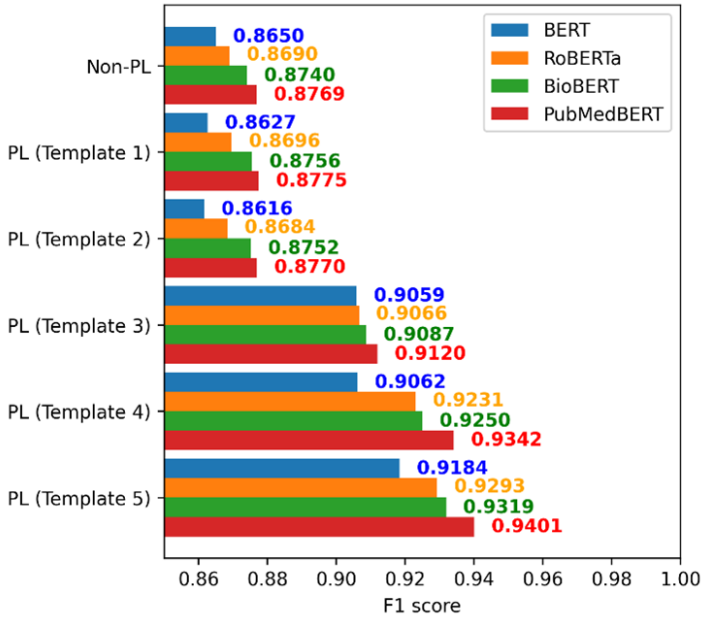

Fig. 4 The F1 scores of different PLMs on the PubMed 20k RCT data

3.4 Few-shot experiments

Based on different percentages of the training data, we tested HSLN and PL based on 5 different templates using the 20k RCT subset. All the results are shown in Table 4 and Figure 5. Among all 5 different templates, the 5th template always outperforms the other templates on each percentage of training data. With only 20% of the training data, it achieved a comparable performance (0.9266) to HSLN (0.9263) which used the complete training data. With over 20% of the training data, the 5th template consistently outperformed HSLN trained on 100% training data. With all the training data, the 5th template achieved 0.9401, which outperformed HSLN by nearly 0.014 on the F1

score, as well as the SOTA results reported in Srivastava's paper by 0.012. These results demonstrated that PL is effective when the training data is limited. Meanwhile, with the same amount of training data, it can outperform HSLN.

The trend is clear that the templates with contextual information are more likely to outperform the templates without contextual information. However, giving more information to the model does not necessarily improve the performance of the model because the template is also an important factor. If the template is more similar to the natural format of how a sentence is written, the model is more likely to perform better.

Table 4 F1 scores on the PubMed 20k RCT test set using different methods at different training sizes

Methods	F1 scores at different percentages of training data				
	5%	10%	20%	50%	100%
Non-PL	0.8564	0.8620	0.8724	0.8733	0.8747
PL (Template 1)	0.8606	0.8686	0.8713	0.8716	0.8775
PL (Template 2)	0.8555	0.8620	0.8682	0.8731	0.8770
PL (Template 3)	0.8928	0.8978	0.9038	0.9081	0.9122
PL (Template 4)	0.8863	0.8996	0.9102	0.9265	0.9342
PL (Template 5)	0.9146	0.9211	0.9266	0.9335	0.9401
HSLN	0.8858	0.9019	0.9117	0.9127	0.9263

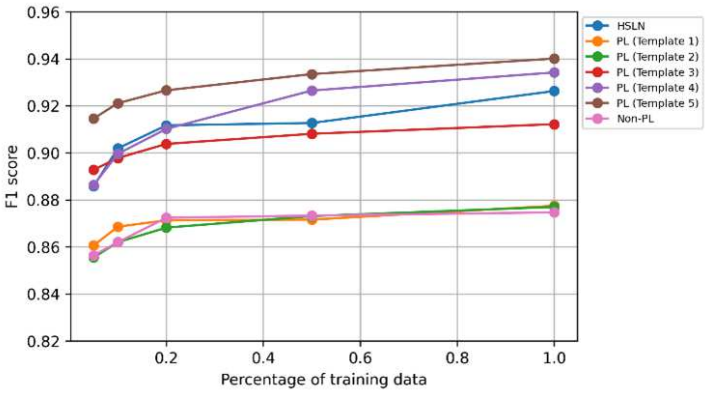

Fig. 5 The trend of how performance increased at different percentages of 20k RCT training data using different methods

To further test the effectiveness of PL under the few-shot setting, we compared it with HSLN on the complete PubMed 200K RCT corpus. Since template 5 outperformed all other templates in the previous experiments, we tested Template 5 only on the 200k dataset. The results are shown in Table 5

and Figure 6. With 5% training data, PL had already outperformed the HSLN trained at 100% training data. With 20% training data, PL achieved a 0.9440, which outperformed the SOTA result reported in Srivastava's paper (0.9407) with 100% training data. With 100% training data, PL beat the SOTA result by 0.01 and beat HSLN by 0.017. The results on the 200k dataset again show the effectiveness of PL when training samples are limited.

Table 5 F1 scores on the PubMed 200k RCT test set using different methods at different training sizes

Methods	F1 scores at different percentages of training data				
	5%	10%	20%	50%	100%
Non-PL	0.8798	0.8822	0.8846	0.8859	0.8909
PL (Template 1)	0.8811	0.8848	0.8855	0.8893	0.8917
PL (Template 2)	0.8796	0.8825	0.8827	0.8909	0.8921
PL (Template 3)	0.9111	0.9129	0.9155	0.9207	0.9224
PL (Template 4)	0.9276	0.9336	0.9396	0.9450	0.9478
PL (Template 5)	0.9369	0.9385	0.9440	0.9469	0.9508
HSLN	0.9255	0.9264	0.9275	0.9332	0.9332

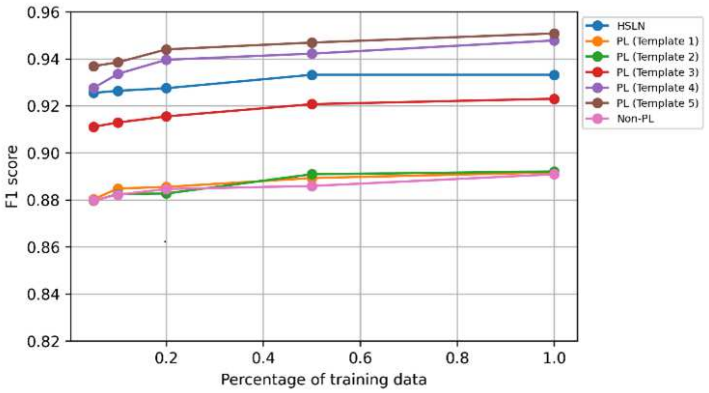

Fig. 6 The trend of how performance increased at different percentages of 200k RCT training data using different methods

3.5 Generalizability

The results of the generalizability test are shown in Tables 6 and 7 and plotted in Figure 7 and 8. The results show that when PL template 5 was trained on either RCT data or OS data, the F1 scores of PL at different sizes of training data were much closer, compared with results from HSLN. The performance of templated 5 ranged from 0.8905 to 0.8971 when trained on RCT and tested

on OS. It ranged from 0.0.9001 to 0.0.9203 when trained on OS and tested on RCT. The results from HSLN showed a more wide range, ranging from 0.8306 to 0.8732 when trained on RCT and tested on OS and ranging from 0.8582 to 0.8921 when trained on OS and tested on RCT. With different percentages of the training data, PL consistently outperformed HSLN. These results demonstrated that the generalizability of the sentence classification model trained by PL template 5 overwhelms the one trained by HSLN. We noticed that when trained on 100% OS data, the F1 score dropped 0.001 compared to trained on 50% OS data, which is probably the overfitting issue.

Table 6 F1 scores on OS test set trained on RCT data

Methods	F1 scores on OS test set				
	5%	10%	20%	50%	100%
Non-PL	0.8378	0.8384	0.8411	0.8441	0.8422
PL (Template 1)	0.8392	0.8396	0.8445	0.8441	0.8510
PL (Template 2)	0.8234	0.8347	0.8416	0.8417	0.8454
PL (Template 3)	0.8681	0.8741	0.8763	0.8781	0.8832
PL (Template 4)	0.8514	0.8723	0.8821	0.8971	0.8999
PL (Template 5)	0.8905	0.8971	0.9025	0.9055	0.9065
HSLN	0.8306	0.8528	0.8643	0.8697	0.8732

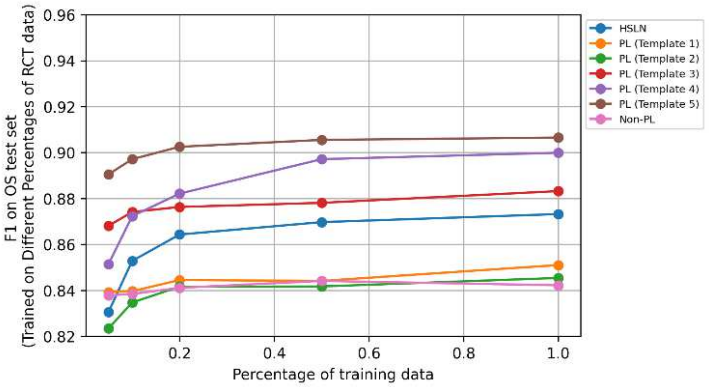

Fig. 7 The trend of how performance increased at different percentages of trained on 20k RCT data tested on 20k OS data using different methods

4 Conclusion

In this paper, we proposed a prompt-learning-based method for sentence classification in medical abstracts. Our results show that PL with designed templates can achieve better performance than the SOTA method. Moreover, much less

Table 7 F1 scores on RCT test set trained on OS data

Methods	F1 scores on RCT test set				
	5%	10%	20%	50%	100%
Non-PL	0.8485	0.8524	0.8583	0.8655	0.8690
PL (Template 1)	0.8497	0.8544	0.8627	0.8667	0.8669
PL (Template 2)	0.8465	0.8512	0.8560	0.8649	0.8655
PL (Template 3)	0.8814	0.8880	0.8899	0.8968	0.9003
PL (Template 4)	0.8652	0.8735	0.8855	0.9060	0.9128
PL (Template 5)	0.9001	0.9021	0.9027	0.9104	0.9203
HSLN	0.8582	0.8668	0.8752	0.8788	0.8921

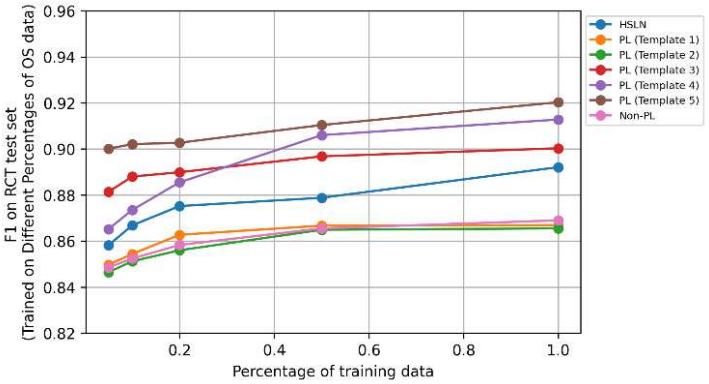

Fig. 8 The trend of how performance increased at different percentages of trained on 20k OS data tested on 20k RCT data using different methods

training data are needed for PL for achieving good performance, indicating the feasibility of few-shot learning for this task. We also demonstrated that our sentence classification model has a better generalizability on the unseen data. In the future, we will continue to explore the influence of different templates on the performance and test the model on other types of articles. In addition, our current method is purely based on manual template and answer engineering, which may not reveal the full potential of prompt learning in this task. Applications of automated template and answer engineering should be explored in the future study.

5 Acknowledgment

We would like to appreciate Dr. Xiaoqian Jiang at the School of Biomedical Informatics, UTHealth for generously sharing computing sources, all the colleagues in the lab for their suggestions and help, and all authors of the open-source codes and dataset involved in our research.

6 Declarations

6.1 Ethical Approval

Not applicable.

6.2 Consent to Participate

Not applicable.

6.3 Consent to Publish

Not applicable.

6.4 Competing interests

Dr. Hua Xu and The University of Health Science Center at Houston have financial related interest in Melax Technologies Inc.

6.5 Authors' contributions

Yan Hu wrote the manuscript text and prepared all the figures and tables. All authors reviewed and revised the manuscript.

6.6 Funding

This work was supported in part by grant R01LM013519 from NLM.

References

[1] Dernoncourt, F., Lee, J.Y.: Pubmed 200k rct: a dataset for sequential sentence classification in medical abstracts. arXiv preprint arXiv:1710.06071 (2017)

[2] Jin, D., Szolovits, P.: Hierarchical neural networks for sequential sentence classification in medical scientific abstracts. arXiv preprint arXiv:1808.06161 (2018)

[3] Srivastava, S., Agarwal, P., Shroff, G., Vig, L.: Hierarchical capsule based neural network architecture for sequence labeling. In: 2019 International Joint Conference on Neural Networks (IJCNN), pp. 1–8 (2019). IEEE

[4] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. arXiv preprint arXiv:2107.13586 (2021)

[5] Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. arXiv preprint arXiv:2104.08691 (2021)

[6] Gu, Y., Han, X., Liu, Z., Huang, M.: Ppt: Pre-trained prompt tuning for few-shot learning. arXiv preprint arXiv:2109.04332 (2021)

[7] Gao, T., Fisch, A., Chen, D.: Making pre-trained language models better few-shot learners. arXiv preprint arXiv:2012.15723 (2020)

[8] Schick, T., Schütze, H.: Exploiting cloze questions for few shot text classification and natural language inference. arXiv preprint arXiv:2001.07676 (2020)

[9] Zhu, Y., Zhou, X., Qiang, J., Li, Y., Yuan, Y., Wu, X.: Prompt-learning for short text classification. arXiv preprint arXiv:2202.11345 (2022)

[10] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

[11] Alrowili, S., Vijay-Shanker, K.: Biom-transformers: building large biomedical language models with bert, albert and electra. In: Proceedings of the 20th Workshop on Biomedical Language Processing, pp. 221–227 (2021)

[12] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H.: Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH) **3**(1), 1–23 (2021)

[13] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020)

[14] Wei, Y., Mo, T., Jiang, Y., Li, W., Zhao, W.: Eliciting knowledge from pretrained language models for prototypical prompt verbalizer. arXiv preprint arXiv:2201.05411 (2022)

[15] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)