Prompt Combines Paraphrase: Enhancing Biomedical "Pre-training, Prompt and Predicting" Models by Explaining Rare Biomedical Concepts

Anonymous ACL submission

Abstract

Prompt-based fine-tuning for pre-trained models has proven resultful in general domains for few-shot learning in downstream tasks. As to the biomedical domain, rare biomedical entities, which are quite ubiquitous in healthcare contexts, can affect the performance of pre-trained models, especially in low-resource scenarios. We propose a simple yet effective approach to helping models understand rare biomedical words during tuning with prompt. Experiments demonstrate that our method can achieve up to 5% improvement in biomedical tasks without any additional parameters or training steps in few-shot vanilla prompt settings.

1 Introduction

Pre-trained models (PTMs) have achieved a great success in natural language processing (NLP) and become a new paradigm for various tasks (Peters et al., 2018; Devlin et al., 2019; Liu et al., 2019; Qiu et al., 2020). Many studies also have paid attention to PTMs in biomedical NLP tasks (Lee et al., 2020; Lewis et al., 2020; Zhao et al., 2021). However, it is clear that PTMs cannot do very well in biomedical NLP tasks due to its internal characteristics of biomedical texts.

In general, there are two challenges for applying PTMs to biomedical NLP tasks, i.e., 1) **limited data** and 2) **rare biomedical words**. Firstly, it is common that the amount of biomedical labeled data is limited due to data sensitivity (Šuster et al., 2017), high cost and professional requirement for data annotation. PTMs perform poorly with few samples since abundant training samples are essential to optimize task-related parameters. Secondly, biomedical terms are usually low-frequency words and are critical to understanding biomedical texts. As an example of natural language inference (NLI) task in Figure 2 in Appendix A, the model goes wrong when faced with a rare words "*afebrile*[1]" in the premise ,whose meaning is "having no fever". It's hard for PTMs to predict the label right without knowing "afebrile". Thus, PTMs cannot capture the precise semantics of biomedical texts without sufficient information of biomedical rare terms.

With very few annotated samples for a new task, it is hard to fine-tune the PTMs and the new task-specific parameters effectively. Prompt technique has been introduced to smooth the fine-tuning process in few-shot setting by closing the gap between pre-training stage and the downstream task in general domains (Liu et al., 2021), as demonstrated in Figure 1. Similarly, the few-shot setting is also a pervasive challenge in biomedical domain mentioned above. Therefore, it is reasonable to adapt "pre-training, prompt and predicting" framework to biomedical NLP tasks.

Furthermore, the challenge of rare words, which is a critical problem for biomedical PTMs, has not been widely explored. Only a handful of works have studied this issue and they focus on enriching the representation of rare words through pre-training stage (Schick and Schütze, 2020; Yu et al., 2021; Wu et al., 2020). Thus, it's necessary for them to involve a second-round pre-training to enrich specific rare words upon PTMs, which is highly time-consuming and low-efficiency. Alternatively, we emphasize on tuning stage instead of pre-training, leading to an efficient approach. Specifically, we propose to explain biomedical concepts on the basis of "pre-training, prompt and predicting" framework. The new approach could manage to enhance tuning capability in the aspect of understanding biomedical concepts. Besides, our approach is a plug-in module for specific datasets and model-agnostic, which can be easily transferred to other domains and models[2].

[1]There are around 4 billion words in pre-training texts of the BC-RoBERTa-Large and "afebrile" appears only about 100,000 times. For comparison, the frequency of "fever" is 5 times that of "afebrile".

[2]We plan to release our code at http://XXX

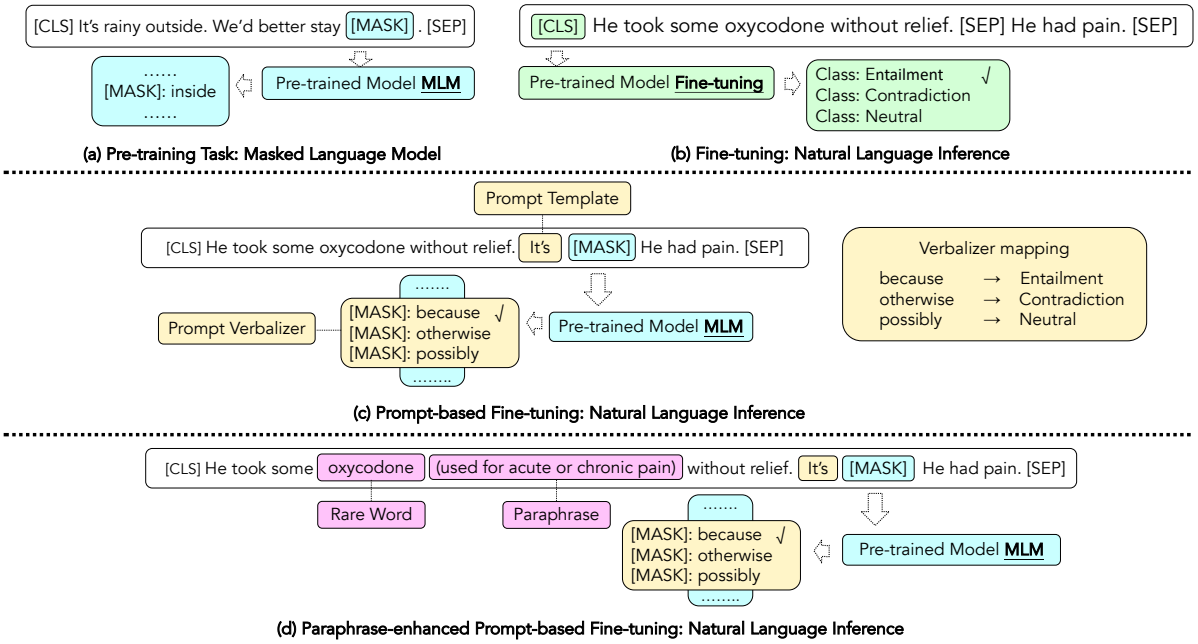

Figure 1: Examples for paradigms of (a) MLM (Masked Language Model) Pre-training. (b) Task-specific fine-tuning. (c) Prompt-based fine-tuning, with same task as pre-training process. (d) Paraphrase-enhanced prompt-based fine-tuning. Best viewed in color.

In summary, our contributions are as follows:

- We investigate a valuable problem of adapting PTMs in scenarios of biomedical text understanding of few samples and rare words, which likely has great impacts on biomedical text mining.
- We propose a novel approach to combine prompt technique and paraphrases of rare words in the PTMs tuning stage to solve the above two challenges.
- We evaluate over two biomedical natural language understanding datasets and our approach can improve the performance by up to 5% in the few-shot setting and 0.6% with a full-size training dataset. Moreover, we discuss how the paraphrases help with the PTMs and provide a perspective about task-related rare words.

2 Related Work

Word frequencies in PTMs Words in the vocabulary list follow a Zipf distribution (Zipf, 2016) by and large. Several previous works have discussed that the word representation space of PTMs is anisotropic and high-frequency words dominates the representation of a sentence inducing a semantic bias (Gao et al., 2019; Li et al., 2020; Yan et al., 2021). Meanwhile, it has been also proven that rare words hamper the PTMs to perform well in which the uncommon words play a decisive role in the sentence understanding (Schick and Schütze, 2020; Wu et al., 2020; Yu et al., 2021). Schick and Schütze (2020) introduces one-token approximation to infer the embedding of arbitrary rare word by a single token. Wu et al. (2020) proposes taking notes on the fly to maintain a note dictionary for rare words to save the contextual information which helps enhance the representation of pre-training.

Biomedical PTMs With the booming trend of PTMs in NLP tasks (Peters et al., 2018; Devlin et al., 2019; Liu et al., 2019), various trials have been made in biomedical domain (Peng et al., 2019; Lee et al., 2020; Huang et al., 2019) by pre-training on biomedical texts. And then, Lewis et al. (2020) and Gu et al. (2021) further get the domain-specific vocabulary list to amend representation of biomedical words. Recently, biomedical PTMs are guided with domain knowledge. Zhang et al. (2021) amplifies the biomedical entities with type semantic information of neighbor entities. Michalopoulos et al. (2021) learns clinical term embedding with relevant meaning and semantic type.

PTMs tuning with prompt Many works are dedicated to applying prompt in fine-tuning by adapting the downstream tasks to the paradigm of pre-training tasks. Prompts that have been employed by now fall into two groups: discrete prompt, described by natural language (Schick and Schütze, 2021; Gao et al., 2021); and continuous prompt, based on trainable vectors (Li and Liang, 2021; Shin et al., 2020b).

3 Rare Biomedical Words and Paraphrases

In this section, we introduce how we find the rare biomedical words and the method we adopt to supplement paraphrases to those words with the prompt-based tuning of PTM.

3.1 Selection of Rare Biomedical Words

"Rare" is a relative concept, which is context-relevant in most cases. We use the RoBERTa-Large model proposed by Lewis et al. (2020) that has been pre-trained adequately on biomedical corpora (details in Appendix B). We download the entire corpora above and loop them through to obtain the frequency of each word in the pre-training phase. In place of involving all rare words, we opt for rare words in biomedical domain for two reasons:
1) Word distribution in general domains differs from that in biomedical domain (Lee et al., 2020).
2) Biomedical rare words can be worth more than general rare words to biomedical tasks. To obtain rare words, we set a threshold on the word frequency in the pre-training corpora empirically as a hyper-parameter similar to Yu et al. (2021). Afterwards, with the help of an online dictionary - Wiktionary[3], we can retrieve the paraphrases of rare words along with the category labels. Only rare words with health-related labels are reserved as rare biomedical words. Full list of selected labels is available in Appendix C.

3.2 Selection of Paraphrases

To avoid introducing noise information from paraphrases, rare biomedical words with more than one paraphrase are eliminated. Also, there should be no additional rare words in the paraphrases. Therefore, we filter out the paraphrases in which frequency of any word is lower than the same threshold mentioned before.

[3]https://en.wiktionary.org/

3.3 Prompt-based Fine-Tuning with Paraphrases

When we read and come across new words, we will consult a proper dictionary for their definitions. Analogously, when the biomedical PTM deal with downstream tasks, we provide the model with paraphrases of biomedical rare words surrounded by brackets attached to the rare words, as Figure 1(d). In this case, given a PTM, paraphrases of biomedical rare words can be considered as a portable plug-in module and generated for any datasets instantly before prompt-based fine-tuning.

4 Experiments

4.1 Setup

Model We use a Biomedical-Clinical-RoBERTa-Large model mentioned in Section 3.1 as a strong baseline to verify our approach.

Datasets Note that rare words hinder the PTMs more in natural language understanding (NLU) (Schick and Schütze, 2020) than in other NLP tasks. However, most biomedical and clinical NLP tasks fall in the category of information extraction (Shin et al., 2020a; Gu et al., 2021). Thus, we evaluate our method over two NLU-relevant biomedical tasks - MedNLI (Romanov and Shivade, 2018) and MedSTS[4] (Wang et al., 2020). Respectively, MedNLI is an NLI dataset in which premises are made up with clinical notes in MIMIC-III and MedSTS is a semantic textual similarity dataset gathered from a clinical corpus at Mayo Clinic. Semantic Textual Similarity is a regression task and we adapt the task following Gao et al. (2021). Statistics of datasets can be found in Appendix D. We sample from 16 up to 256 samples from the original training sets as training and development sets with 10 different random seeds and use full-size testing sets.

Prompt settings We combine the prompt settings from Schick and Schütze (2021) and Gao et al. (2021) for the NLI and STS tasks without further adaption (details in Appendix E) to explore the effectiveness of paraphrases of biomedical rare words rather than the prompt paradigm.

4.2 Main Results and Analysis

We report average accuracy for MedNLI and pearson correlation coefficient for MedSTS along with

[4]We use ClinicalSTS-2018 and 2019 which are sub-datasets of MedSTS provided by the maintainers of MedSTS project.

MedNLI					
Model \ #Samples	16	32	64	128	256
BC-RoBERTa-Large	51.3 (5.9)	60.6 (6.7)	71.0 (3.7)	80.6 (1.3)	83.1 (1.3)
+ paraphrase	**56.6** (5.0)	**62.3** (6.0)	**74.5** (3.0)	**81.1** (1.5)	**83.6** (1.0)
MedSTS: Clinical-2019					
Model \ #Samples	16	32	64	128	256
BC-RoBERTa-Large	41.1 (11.8)	53.9 (6.7)	**67.9** (7.4)	73.1 (5.0)	**80.4** (3.1)
+ paraphrase	**45.2** (9.3)	**57.3** (6.8)	67.2 (7.5)	**74.5** (3.7)	79.6 (2.6)
MedSTS: Clinical-2018					
Model \ #Samples	16	32	64	128	256
BC-RoBERTa-Large	**54.2** (8.1)	63.9 (9.2)	73.3 (3.8)	77.4 (2.7)	81.5 (1.5)
+ paraphrase	53.0 (7.4)	**67.2** (6.6)	**74.5** (2.7)	**79.1** (1.6)	**81.8** (1.2)

Table 1: Our main results on three dataset: MedNLI, MedSTS: Clinical-2018 and Clinical-2019, using BC-RoBERTa-Large (**B**iomedical and **C**linical RoBERTa-Large) (Lewis et al., 2020) with different size of training sets. We report average (and standard deviation) performance (accuracy for MedNLI and pearson correlation coefficient for MedSTS) over 10 different random seeds. + paraphrase: with paraphrases of rare biomedical words.

standard deviation. Table 1 shows results for biomedical natural language inference and semantic textual similarity tasks. Model with paraphrases for rare biomedical words can outperform the baseline in most cases. Paraphrases bring about up to 5% improvement on average for few-shot learning with 16 training samples as to MedNLI task and 0.5% increment with 256 training samples. We can see that PTMs tend to learn more about rare words with more training samples but paraphrases still act well. As to MedSTS, appended paraphrases are also shown as an effective strategy for most cases. In addition, tuning with paraphrases also generally improves model stability and reduces the variance of model prediction in few-shot scenarios.

5 Discussion

Train with more samples Apart from infusing dictionary paraphrases in few-shot scenarios, we also attempt with more training samples, even with full-size training dataset. Table 4 in Appendix F demonstrates that with larger amount of training samples, our method still advances the PTMs for majority cases, implying that paraphrases of rare biomedical words are not only impactful in few-sample situations.

Which to look up? By far, experiment results have attested that paraphrases of rare biomedical words help with PTMs in training with few or more samples. Nevertheless, it may not always work well. We scrutinize the cases that model predicts differently after paraphrases being appended and display several cases in Table 5 in Appendix G. Table 5 shows that paraphrases of rare words which are task-related and decisive in understanding the whole sentence can be beneficial to PTMs. Otherwise, paraphrases can involve more confusion than certainty. When human reads, we probably won't look up a new word until it blocks our understanding. Similarly, it is worthwhile to explore how to attach *helpful* paraphrases or utilize knowledge selectively in future research.

6 Conclusion

Biomedical terms, which are pervasive in biomedical texts, are sometimes rare in the whole corpora and domain-specific rare words understanding remains as a tough challenge for pre-trained models. In this paper, we present a simple yet effective method to help biomedical pre-trained models grasp the semantics of rare biomedical words, that is attaching paraphrases to rare biomedical words as a plug-in approach in the prompt-tuning datasets without additional parameters to train during pre-training and downstream task-related tuning. Experiments show that our method can substantially boost the performance of biomedical pre-trained model in few-shot setting and bring about plausible enhancement with more training data, even full-size of training set.

Ethical Considerations

In this work, we propose an approach to explaining rare biomedical words for biomedical PTMs to help understand sentences with rare biomedical words. We conduct our experiments on the public biomedical datasets MedNLI and MedSTS with the authorization from the respective maintainers of the datasets. All biomedical data involved have been de-identified by dataset providers and only used for research.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tieyan Liu. 2019. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations*.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020. Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130, Online. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen Chen, and Alexander Wong. 2021. Umlsbert: Clinical domain knowledge augmentation of contextual embeddings using the unified medical language system metathesaurus. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1744–1753.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, pages 1–26.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596.

Timo Schick and Hinrich Schütze. 2020. Rare words: A major problem for contextualized embeddings and how to fix it by attentive mimicking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8766–8774.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269.

Hoo-Chang Shin, Yang Zhang, Evelina Bakhturina, Raul Puri, Mostofa Patwary, Mohammad Shoeybi, and Raghav Mani. 2020a. Bio-megatron: Larger biomedical domain language model. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4700–4706.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020b. Eliciting knowledge from language models using automatically generated prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235.

Simon Šuster, Stéphan Tulkens, and Walter Daelemans. 2017. A short review of ethical challenges in clinical natural language processing. *arXiv preprint arXiv:1703.10090*.

Yanshan Wang, Naveed Afzal, Sunyang Fu, Liwei Wang, Feichen Shen, Majid Rastegar-Mojarad, and Hongfang Liu. 2020. Medsts: a resource for clinical semantic textual similarity. *Language Resources and Evaluation*, 54(1):57–72.

Qiyu Wu, Chen Xing, Yatao Li, Guolin Ke, Di He, and Tie-Yan Liu. 2020. Taking notes on the fly helps bert pre-training. *arXiv preprint arXiv:2008.01466*.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A contrastive framework for self-supervised sentence representation transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5065–5075, Online. Association for Computational Linguistics.

Wenhao Yu, Chenguang Zhu, Yuwei Fang, Donghan Yu, Shuohang Wang, Yichong Xu, Michael Zeng, and Meng Jiang. 2021. Dict-bert: Enhancing language model pre-training with dictionary. *arXiv preprint arXiv:2110.06490*.

Taolin Zhang, Zerui Cai, Chengyu Wang, Minghui Qiu, Bite Yang, and Xiaofeng He. 2021. Smedbert: A knowledge-enhanced pre-trained language model with structured semantics for medical text mining. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5882–5893.

Sendong Zhao, Chang Su, Zhiyong Lu, and Fei Wang. 2021. Recent advances in biomedical literature mining. *Briefings in Bioinformatics*, 22(3):bbaa057.

George Kingsley Zipf. 2016. *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio Books.

Appendix

A Wrong Case of Biomedical PTM

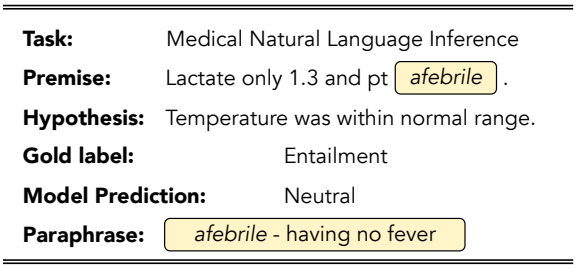

Figure 2: Wrong case of **BC**-RoBERTa-Large model fully pre-trained on **B**iomedical and **C**linical texts (Lewis et al., 2020) and fine-tuned on MedNLI tasks caused by not understanding biomedical rare word - "afebrile".

B Pre-trained Corpora of the Model

We use a biomedical and clinical RoBERTa-Large (Lewis et al., 2020) trained on biomedical corpora, including PubMed abstract[5], PubMed Central[6] (PMC) full-text and MIMIC-III dataset[7].

C Word Labels for Rare Biomedical Words

We focus on the rare words which have been tagged with labels that contain any of following medicine-related strings:

['medical', 'medicine', 'disease', 'symptom', 'pharma']

D Dataset

We conduct our experiments on MedNLI and MedSTS datasets. Specifically, we use the available sub-datasets ClinicalSTS-2018 and ClinicalSTS-2019 for MedSTS provided by the maintainer of MedSTS project. The statistics of involved datasets can be found in Table 2. Note that there is no development set split in MedSTS. Therefore, we sample the development set for MedSTS from its training set with the same quantity as sampled few-shot training set and make sure there is no overlap between training and development set.

[5]https://pubmed.ncbi.nlm.nih.gov
[6]https://www.ncbi.nlm.nih.gov/pmc
[7]https://physionet.org/content/mimiciii/1.4/

Dataset	Train	Dev	Test
MedNLI	11232	1395	1422
MedSTS: ClinicalSTS-2019	1642	/	412
MedSTS: ClinicalSTS-2018	750	/	318

Table 2: Statistics of datasets MedNLI and MedSTS

E Prompt Settings

We adopt the prompt settings empirically from Schick and Schütze (2021) and Gao et al. (2021) for the natural language inference and semantic textual similarity tasks shown in Table 3 since the prompt paradigm is not the core of this work and our method is prompt-agnostic.

Task	Prompt Template	Prompt Verbalizers
MedNLI	<Sent1>. [MASK]. <Sent2>	Yes/No/maybe
MedSTS	<Sent1>. [MASK]. <Sent2>	Yes/No

Table 3: Prompt settings for MedNLI and MedSTS

F Train with More Samples

Besides few-shot scenarios, we also train with more samples for MedNLI since it has 11,232 training samples. Experiment results are shown in Table 4.

G Case Analysis

We display several cases in which model predicts differently with or without paraphrases of rare biomedical words from MedNLI in Table 5. From the cases, we can see that paraphrases of rare biomedical words that are determinant in sentence understanding can be helpful to pre-trained model while paraphrases of those irrelevant rare biomedical words may confuse the model.

MedNLI						
Model \ #Samples	512	1024	2048	4096	8192	full-size
BC-RoBERTa-Large	85.0(0.8)	85.6(0.8)	**86.6**(0.6)	**86.7**(0.7)	86.2(0.6)	86.1(0.6)
+ paraphrases	**85.2**(0.7)	**86.3**(0.9)	86.4(0.7)	86.4(0.5)	**86.7**(0.6)	**86.7**(0.7)

Table 4: Test results on MedNLI dataset with larger size of training sets. We report average (and standard deviation) accuracy over 10 different random seeds. BC-RoBERTa-Large: **B**iomedical and **C**linical RoBERTa-Large (Lewis et al., 2020). + paraphrase: with paraphrases of rare biomedical words.

Sentence Pairs	w/o paraphrases	w/ paraphrases
P: She was found to have **BRBPR** *(bright red blood per rectum)* on rectal exam. H: the patient had bright read blood per rectum	Neutral	Entailment (**right answer**)
P: Antenatal history - pregnancy complicated by chronic hypertension with increased gestational hypertension leading to admission 3 days prior to delivery followed by cesarean section. H: The patient had **proteinuria** *(The presence of protein in the urine)* during pregnancy	Entailment	Neutral (**right answer**)
P: Following this rehab admission she was sent to a different OSH on [**2725-10-26**], for acute **CHF** *(congestive heart failure)* and at least one PEA arrest. H: The patient has a poorly functioning heart.	Contradiction	Entailment (**right answer**)
P: The patient was sent to the HD unit prior to coming to the floor for **workup** *(A general medical examination to assess a persons health and fitness)* of fever. H: The patient has an infection	Neutral (**right answer**)	Contradiction
P: - **COPD** *(chronic obstructive pulmonary disease)* - obesity - unspecified hypoxemia - CNS lymphoma c/b CVAs x3 (posterior circulation) and seizure d/o - history of SAH while on coumadin - diastolic heart failure - coronary artery disease - atrial fibrillation - hypertension - hyperlipidemia - severe OSA (did not tolerate CPAP in the past) - primary hyperparathyroidism/25-vit D deficiency c/b nephrolithiasis - toxic multinodular goiter with **subclinical** *(Less than is needed for clinical reasons)* hyperthyroidism - neovascular glaucoma c/b right eye blindness H: Patient has a history of malignancy	Neutral (**right answer**)	Entailment

Table 5: Cases that BC-RoBERTa-Large predicts differently after the supplement of paraphrases for rare biomedical words in MedNLI. “P" for Premise and “H" for Hypothesis. Words in **bold** are rare biomedical words and expressions in *italic* inside the brackets are the paraphrases of rare words.