Empirical Study on Intermediate Fine-tuning for Biomedical Domain Tasks

Yuna Jeong
Korea Institute of Science and
Technology Information
Korea
jeongyuna@kisti.re.kr

Myunggwon Hwang
Korea Institute of Science and
Technology Information
Korea
mgh@kisti.re.kr

ABSTRACT

Recently, the Supplementary Training on Intermediate Labeled-data Task (STILT) was proposed for more effective fine-tuning by performing fine-tuning on the intermediate task before fine-tuning of target task. STILT is a simple but effective way to improve performance on some task pairs. To find beneficial task pairs for STILT, previous studies have shown many broad experimental results, but most of these are general language domain tasks. In this paper, we performed experiments to verify the effectiveness of STILT in specialized domain. From the experimental results, it was observed that the pre-trained language model on general domain can learn some domain knowledge from intermediate task fine-tuning. Also, it observed that STILT can achieve performance improvement even in a specialized domain.

KEYWORDS

Intermediate task, Fine-tuning, Deep Learning

1 INTRODUCTION

Pre-trained Language Model (PTLM) has contributed greatly to the growth of the natural language processing. PTLM pre-trains general language knowledge based on the unsupervised learning, and then optimizes it to each downstream task by fine-tuning [1]. Recently, the *Supplementary Training on Intermediate Labeled-data Task (STILT)* was proposed for more effective fine-tuning by performing fine-tuning on the intermediate task before fine-tuning of target task [2]. It is generally expected that the performance of the target task is improved by learning additional domain or task knowledge from training of intermediate task. However, STILT does not always guarantee better performance. Also, it is not yet clear when and why the STILT will perform well. To answer this question, Pruksachatkun et al. [3] performed broad experiments with 110 task pairs based on RoBERTa [4]. Similarly, Vu et al. [5] experimented with more than 3000 task pairs and tried to predict the most transferable intermediate task to the target task based on task embedding. From the previous studies, it has been revealed that the performance of STILT is affected by various factors, such as the size of target data, domain difference between tasks, task similarity, and task complexity [3,5].

Table 1 Overview of the tasks in our experiments.

Dataset	Train	Dev	Metrics
Task: Named Entity Recognition			
BC5-chem	5203	5347	F1 entity-level
BC5-disease	4182	4244	
NCBI-disease	5134	787	
BC2GM	15197	3061	
JNLPBA	46750	4551	
Task: PICO Extraction			
EBM PICO	339167	85321	Macro F1 word-level
Task: Relation Extraction			
ChemProt	18035	11268	Micro F1
DDI	25296	2496	
GAD	4261	535	
Task: Question Answering			
PubMedQA	450	50	Accuracy
BioASQ	670	75	

In this paper, we perform experiments and analyzes of STILT in the specialized domain, not the general language domain, which has been mainly focused in previous studies. Specifically, we aim to answer the following questions:

- Can STILT also work beneficially in domain-specific tasks such as biomedical?
- Does fine-tuning of the domain-specific intermediate task to general domain PTLM contribute to domain knowledge learning?

To answer the questions, we experimented 11 datasets of 4 tasks from BLURB, a representative benchmark on biomedical natural language processing [6].

2 EXPERIMENTAL DETAILES

2.1 Tasks and Datasets

We used 11 datasets from four types of tasks of the BLURB benchmark [6]: named entity recognition (NER), PICO extraction, relation extraction (RE), and question answering (QA). We constructed 100 pairs of intermediate and target task from 11 datasets for the experiment. Table 1 shows the dataset, data size, and evaluation metrics used in the experiment.

2.2 PTLMs and Fine-tuning

We used $\mathrm{BERT}_{\mathrm{base}}$ [1] and $\mathrm{PubMedBERT}_{\mathrm{base}}$ [6] as PTLMs for fine-tuning. BERT is a representative language model that learned language knowledge in the general domain from Wikipedia corpus. Similarly, PubMedBERT pre-trained BERT as a corpus in the biomedical domain. The two models have the same model architecture and only show differences in the domain of the data used for training. Fine-tuning of intermediate and target task of PTLM is the same as the method proposed in BERT [1]. STILT first fine-tunes the PTLM to the intermediate task, and subsequently fine-tunes the resulting model to the target task. The hyperparameters of fine-tuning are the same as those of PubMedBERT.

3 EXPERIMENTAL RESULTS

We experimented our work on an Intel Core i9 3.0-GHz machine with two NVIDIA TITAN RTX.

Tables 2 and 3 show the experimental results of STILT based on the PubMedBERT and BERT, respectively. In the tables, the column indicates the intermediate task and the row indicates the target task. Baseline is the performance of fine-tuned PTLM on target task without intermediate task. Basically, PubMedBERT, a biomedical-specific model, shows better performance in all biomedical tasks.

In the performance of STILT, a blue background color is used when performance is improved, and an orange background color is used when performance is degraded. Higher saturation means relatively more performance change; the whiter the color, the less the performance differs from the baseline. Comparing the overall STILT performance in Table 1 and Table 2, the performance improvement was more frequent when using the intermediate task in the case of the BERT than the PubMedBERT. It is considered that the performance of the target task was improved as BERT, which had no biomedical domain knowledge, learned biomedical domain knowledge by fine-tuning the intermediate task. On the other hand, PubMedBERT will have little or no such benefit because biomedical domain knowledge has already been

Table 3 Experimental results of the STILT method in PubMedBERT. The resulting value is the average of three runs.

			Intermediate Task										
		Baseline	BC5-chem	BC5-disease	NCBI-disease	BC2GM	JNLPBA	EBM PICO	ChemProt	DDI	GAD	PubMedQA	BioASQ
Target task	BC5-chem	92.97	-	92.99	93.11	93.46	93.20	93.22	92.88	92.88	93.05	93.28	93.00
	BC5-disease	84.61	85.41	-	85.91	85.73	85.71	85.52	85.24	85.38	85.63	85.38	85.29
	NCBI-disease	87.73	87.46	88.10	-	87.57	88.01	88.27	88.06	87.99	87.77	88.81	87.72
	BC2GM	83.91	83.86	84.27	84.17	-	84.00	84.13	84.24	83.87	84.52	83.98	84.25
	JNLPBA	79.02	78.93	78.94	78.77	78.84	-	79.25	79.25	79.18	79.26	79.00	78.99
	EBM PICO	73.20	74.00	73.86	73.90	73.58	73.66	-	73.66	73.85	73.96	73.88	73.83
	ChemProt	77.28	77.29	77.11	77.44	75.14	76.65	76.97	-	76.80	76.70	77.08	77.42
	DDI	82.96	82.72	81.67	82.28	79.50	82.21	81.47	82.40	-	82.59	82.70	83.22
	GAD	82.17	81.63	81.73	82.22	83.62	81.98	82.42	82.62	83.02	-	81.53	83.57
	PubMedQA	55.00	57.80	51.60	54.40	55.00	50.80	52.90	62.30	61.80	50.40	-	65.40
	BioASQ	84.29	88.22	83.93	82.86	78.22	86.07	80.72	80.00	92.15	83.22	86.79	-

Table 2 Experimental results of the STILT method in BERT. The resulting value is the average of three runs.

			Intermediate Task										
		Baseline	BC5-chem	BC5-disease	NCBI-disease	BC2GM	JNLPBA	EBM PICO	ChemProt	DDI	GAD	PubMedQA	BioASQ
Target task	BC5-chem	88.92	-	89.37	89.46	89.47	89.66	89.40	89.38	89.54	89.23	89.27	89.27
	BC5-disease	80.74	80.69	-	81.35	81.32	80.93	80.33	80.32	80.38	80.80	80.58	80.70
	NCBI-disease	85.33	86.19	85.93	-	86.37	85.32	85.64	85.76	86.05	86.12	86.60	85.53
	BC2GM	80.83	80.64	81.02	81.43	-	81.28	81.44	80.80	80.95	81.27	81.65	80.97
	JNLPBA	77.32	77.16	77.61	77.52	77.63	-	77.58	77.51	77.62	77.63	77.64	77.84
	EBM PICO	72.15	72.55	72.32	72.44	72.16	72.49	-	71.95	72.49	72.55	72.26	72.26
	ChemProt	71.23	71.00	71.06	70.86	69.82	71.05	71.11	-	70.96	71.47	71.82	71.68
	DDI	77.01	77.32	78.38	77.87	74.57	77.71	78.57	79.03	-	78.91	78.29	78.23
	GAD	78.93	77.21	78.40	78.79	80.56	79.12	80.21	78.91	78.59	-	79.70	77.97
	PubMedQA	50.6	50.20	51.10	50.80	50.90	51.80	49.20	58.30	58.50	51.10	-	53.40
	BioASQ	61.43	66.91	68.33	71.90	62.86	67.38	71.67	75.48	76.66	70.48	66.66	-

sufficiently learned through pre-training. That is, if there is a LM pre-trained in the same domain as the target task, it is generally best to use the PTLM, but if not, the domain gab can be reduced and performance can be improved by fine-tuning of intermediate task. Interestingly, when PubMedQA is a target task, some BERT experiments using STILT outperformed the baseline performance of PubMedBERT; nevertheless, PubMedBERT with STILT still performs best.

In the results of Table 2, we could observe the results excluding the benefit from domain knowledge transfer. NER and PICO extraction have improved performance as an intermediate task in most cases, if not significantly. On the other hand, RE showed slight performance degradation in most cases. In the case of QA, it showed a great performance improvement in the same kind of intermediate task. The intermediate task of RE also significantly changed the performance of QA, but both the performance improvement and the decrease appeared. From the experimental results, it was observed that STILT can be also applied to some extent effectively in the biomedical domain as well.

4 CONCLUSIONS

In this paper, we conducted an experiment to investigate the effect of STILT on specialized domain tasks. We tested the performance of two PTLMs for 110 pairs of intermediate and target tasks in the biomedical field. From the experimental results, it was observed that the general domain PTLM can learn some domain knowledge from intermediate task fine-tuning. Although it has been observed from some experimental results that STILT can achieve performance improvement even in a specialized domain, it is difficult to confirm which task combination will work with a benefit yet. Additional experiments and analyzes should be performed considering many other factors such as task complexity, task similarity, and data size.

ACKNOWLEDGMENTS

This research was supported by Korea Institute of Science and Technology Information (KISTI). Correspondence concerning this paper should be addressed to Myunggwon Hwang.

REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT.

[2] Jason Phang, Thibault Fevry, and Samuel R. Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. arXiv preprint arXiv:1811.01088.

[3] Yada Pruksachatkun et al.. 2020. Intermediate-task transfer learning with pretrained models for natural language understanding: When and why does it work?. arXiv preprint arXiv:2005.00628.

[4] Yinhan Liu et al.. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

[5] Tu Vu et al.. 2020. Exploring and Predicting Transferability across NLP Tasks. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).

[6] Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH) 3, 1, 1-23.