An Empirical Study of Multi-Task Learning on BERT for Biomedical Text Mining

Yifan Peng **Qingyu Chen** **Zhiyong Lu**
National Center for Biotechnology Information
National Library of Medicine, National Institutes of Health
Bethesda, MD, USA
{yifan.peng, qingyu.chen, zhiyong.lu}@nih.gov

Abstract

Multi-task learning (MTL) has achieved remarkable success in natural language processing applications. In this work, we study a multi-task learning model with multiple decoders on varieties of biomedical and clinical natural language processing tasks such as text similarity, relation extraction, named entity recognition, and text inference. Our empirical results demonstrate that the MTL fine-tuned models outperform state-of-the-art transformer models (e.g., BERT and its variants) by 2.0% and 1.3% in biomedical and clinical domains, respectively. Pairwise MTL further demonstrates more details about which tasks can improve or decrease others. This is particularly helpful in the context that researchers are in the hassle of choosing a suitable model for new problems. The code and models are publicly available at https://github.com/ncbi-nlp/bluebert.

1 Introduction

Multi-task learning (MTL) is a field of machine learning where multiple tasks are learned in parallel while using a shared representation (Caruana, 1997). Compared with learning multiple tasks individually, this joint learning effectively increases the sample size for training the model, thus leads to performance improvement by increasing the generalization of the model (Zhang and Yang, 2017). This is particularly helpful in some applications such as medical informatics where (labeled) datasets are hard to collect to fulfill the data-hungry needs of deep learning.

MTL has long been studied in machine learning (Ruder, 2017) and has been used successfully across different applications, from natural language processing (Collobert and Weston, 2008; Luong et al., 2016; Liu et al., 2019c), computer vision (Wang et al., 2009; Liu et al., 2019a; Chen et al., 2019), to health informatics (Zhou et al., 2011; He et al., 2016; Harutyunyan et al., 2019). MTL has also been studied in biomedical and clinical natural language processing (NLP) such as named entity recognition and normalization and the relation extraction. However, most of these studies focus on either one task with multi corpora (Khan et al., 2020; Wang et al., 2019b) or multi-tasks on a single corpus (Xue et al., 2019; Li et al., 2017; Zhao et al., 2019).

To bridge this gap, we investigate the use of MTL with transformer-based models (BERT) on multiple biomedical and clinical NLP tasks. We hypothesize the performance of the models on individual tasks (especially in the same domain) can be improved via joint learning. Specifically, we compare three models: the independent single-task model (BERT), the model refined via MTL (called MT-BERT-Refinement), and the model fine-tuned for each task using MT-BERT-Refinement (called MT-BERT-Fine-Tune). We conduct extensive empirical studies on the Biomedical Language Understanding Evaluation (BLUE) benchmark (Peng et al., 2019), which offers a diverse range of text genres (biomedical and clinical text) and NLP tasks (such as text similarity, relation extraction, and named entity recognition). When learned and fine-tuned on biomedical and clinical domains separately, we find that MTL achieved over 2% performance on average, created new state-of-the-art results on four BLUE benchmark tasks. We also demonstrate the use of multi-task learning to obtain a single model that still produces state-of-the-art performance on all tasks. This positive answer will be very helpful in the context that researchers are in the hassle of choosing a suitable model for new problems where training resources are limited.

Our contribution in this work is three-fold: (1) We conduct extensive empirical studies on 8 tasks from a diverse range of text genres. (2) We

Proceedings of the BioNLP 2020 workshop, pages 205–214

demonstrate that the MTL fine-tuned model (MT-BERT-Fine-Tune) achieved state-of-the-art performance on average and there is still a benefit to utilizing the MTL refinement model (MT-BERT-Refinement). Pairwise MTL, where two tasks were trained jointly, further demonstrates which tasks can improve or decrease other tasks. (3) We make codes and pre-trained MT models publicly available.

The rest of the paper is organized as follows. We first present related work in Section 2. Then, we describe the multi-task learning in Section 3, followed by our experimental setup, results, and discussion in Section 4. We conclude with future work in the last section.

2 Related work

Multi-tasking learning (MTL) aims to improve the learning of a model for task t by using the knowledge contained in the tasks where all or a subset of tasks are related (Zhang and Yang, 2017). It has long been studied and has applications on neural networks in the natural language processing domain (Caruana, 1997). Collobert and Weston (2008) proposed to jointly learn six tasks such as part-of-speech tagging and language modeling in a time-decay neural network. Changpinyo et al. (2018) summarized recent studies on applying MTL in sequence tagging tasks. Bingel and Søgaard (2017) and Martínez Alonso and Plank (2017) focused on conditions under which MTL leads to gain in NLP, and suggest that certain data features such as learning curve and entropy distribution are probably better predictors of MTL gains.

In the biomedical and clinical domains, MTL has been studied mostly in two directions. One is to apply MTL on a single task with multiple corpora. For example, many studies focused on named entity recognition (NER) tasks (Crichton et al., 2017; Wang et al., 2019a,b). Zhang et al. (2018), Khan et al. (2020), and Mehmood et al. (2019) integrated MTL in the transformer-based networks (BERT), which is the state-of-the-art language representation model and demonstrated promising results to extract biomedical entities from literature. Yang et al. (2019) extracted clinical named entity from Electronic Medical Records using LSTM-CRF based model. Besides NER, Li et al. (2018) and Li and Ji (2019) proposed to use MTL on relation classification task and Du et al. (2017) on biomedical semantic indexing. Xing et al. (2018) exploited domain-invariant knowledge to segment Chinese word in medical text.

The other direction is to apply MTL on different tasks, but the annotations are from a single corpus. Li et al. (2017) proposed a joint model extract biomedical entities as well as their relations simultaneously and carried out experiments on either the adverse drug event corpus (Gurulingappa et al., 2012) or the bacteria biotope corpus (Deléger et al., 2016). Shi et al. (2019) also jointly extract entities and relations but focused on the BioCreative/OHNLP 2018 challenge regarding family history extraction (Liu et al., 2018). Xue et al. (2019) integrated the BERT language model into joint learning through dynamic range attention mechanism and fine-tuned NER and relation extraction tasks jointly on one in-house dataset of coronary arteriography reports.

Different from these works, we studied to jointly learn 8 different corpora from 4 different types of tasks. While MTL has brought significant improvements in medicine tasks, no (or mixed) results have been reported when pre-training MTL models in different tasks on different corpora. To this end, we deem that our model can provide more insights about conditions under which MTL leads to gains in BioNLP and clinical NLP, and sheds light on the specific task relations that can lead to gains from MTL models over single-task setups.

3 Multi-task model

The architecture of the MT-BERT model is shown in Figure 1. The shared layers are based on BERT (Devlin et al., 2018). The input X can be either a sentence or a pair of sentences packed together by a special token `[SEP]`. If X is longer than the allowed maximum length (e.g., 128 tokens in the BERT's base configuration), we truncate X to the maximum length. When X is packed by a sequence pair, we truncate the longer sequence one token at a time. Similar to (Devlin et al., 2018), two additional tokens are added at the start (`[CLS]`) and end (`[SEP]`) of X, respectively. Similar to (Lee et al., 2020; Peng et al., 2019), in the sequence tagging tasks, we split one sentence into several sub-sentences if it is longer than 30 words.

In the shared layers, the BERT model first converts the input sequence to a sequence of embedding vectors. Then, it applies attention mechanisms to gather contextual information. This se-

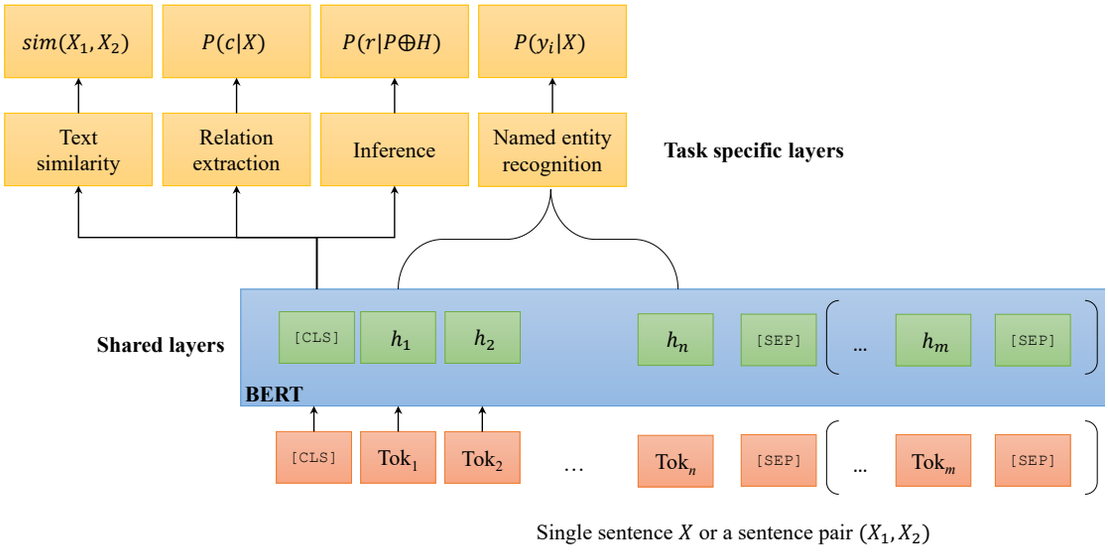

Figure 1: The architecture of the MT-BERT model.

mantic representation is shared across all tasks and is trained by our multi-task objectives. Finally, the BERT model encodes that information in a vector for each token $(h_0, \ldots, h_n)$.

On top of the shared BERT layers, the task-specific layer uses a fully-connected layer for each task. We fine-tune the BERT model and the task-specific layers using multi-task objectives during the training phase. More details of the multi-task objectives in the BLUE benchmark are described below.

3.1 Sentence similarity

Suppose that h_0 is the BERT's output of the token `[CLS]` in the input sentence pair (X_1, X_2). We use a fully connected layer to compute the similarity score $sim(X_1, X_2) = ah_0 + b$, where $sim(X_1, X_2)$ is a real value. This task is trained using the Mean Squared Error (MSE) loss: $(y - sim(X_1, X_2))^2$, where y is the real-value similarity score of the sentence pair.

3.2 Relation extraction

This task extracts binary relations (two arguments) from sentences. After replacing two arguments of interest in the sentence with pre-defined tags (e.g., GENE, or DRUG), this task can be treated as a classification problem of a single sentence X. Suppose that h_0 is the output embedding of the token `[CLS]`, the probability that a relation is labeled as class c is predicted by a fully connected layer and a logistic regression with softmax: $P(c|X) = softmax(ah_0 + b)$. This approach is widely used in the transformer-based models (Devlin et al., 2018; Peng et al., 2019; Liu et al., 2019c). This task is trained using the categorical cross-entropy loss: $-\sum_c \delta(y_c = \hat{y}) \log(P(c|X))$, where $\delta(y_c = \hat{y}) = 1$ if the classification $\hat{y}$ of X is the correct ground-truth for the class $c \in C$; otherwise $\delta(y_c = \hat{y}) = 0$.

3.3 Inference

After packing the pair of premise sentences with hypothesis into one sequence, this task can also be treated as a single sentence classification problem. The aim is to find logical relation R between premise P and hypothesis H. Suppose that that h_0 is the output embedding of the token `[CLS]` in $X = P \oplus H$, $P(R|P \oplus H) = softmax(ah_0 + b)$. This task is trained using the categorical cross-entropy loss as above.

3.4 Named entity recognition

The output of the BERT model produces a feature vector sequence $\{h_i\}_{i=0}^n$ with the same length as the input sequence X. The MTL model predicts the label sequence by using a softmax output layer, which scales the output for a label $l \in \{1, 2, \ldots, L\}$ as follows: $P(\hat{y}_i = j|x) = \frac{\exp(h_i W_j)}{\sum_{l=1}^{L} \exp(h_i W_j)}$, where L is the total number of tags. This task is trained using the categorical cross-entropy loss: $-\sum_i \sum_{y_i} \delta(y_i = \hat{y}_i) \log P(y_i|X)$.

3.5 The training procedure

The training procedure for MT-BERT consists of three stages: (1) pretraining the BERT model,

(2) refining it via multi-task learning (MT-BERT-Refinement), and (3) fine-tuning the model using the task-specific data (MT-BERT-Fine-Tune).

3.5.1 Pretraining

The pretraining stage follows that of the BERT using the masked language modeling technique (Devlin et al., 2018). Here we used the base version. The maximum length of the input sequences is thus 128.

3.5.2 Refining via Multi-task learning

In this step, we refine all layers in the model. Algorithm 1 demonstrates the process of multi-task learning (Liu et al., 2019c). We first initialize the shared layers with the pre-trained BERT model and randomly initialize the task-specific layer parameters. Then we create the dataset by merging mini-batches of all the datasets. In each epoch, we randomly select a mini-batch b_t of task t from all datasets D. Then we update the model according to the task-specific objective of the task t. Same as in (Liu et al., 2019c), we use the mini-batch based stochastic gradient descent to learn the parameters.

Algorithm 1: Multi-task learning.

Initialize *model parameters* θ
  Shared layer parameters by BERT;
  Task-specific layer parameters randomly;
end
Create D by merging mini-batches for each dataset;
for $epoch$ *in* $1, 2, ..., epoch_{max}$ **do**
  Shuffle D;
  for b_t *in* D **do**
    Compute loss: $L(\theta)$ based on task t;
    Compute gradient: $\nabla(\theta)$
    Update model: $\theta = \theta - \eta\nabla(\theta)$
  end
end

3.5.3 Fine-tuning MT-BERT

We fine-tune existing MT-BERT that are trained in the previous stage by continue training all layers on each specific task. Provided that the dataset is not drastically different in context to other datasets, the MT-BERT model will already have learned general features that are relevant to a specific problem. Specifically, we truncate the last layer (softmax and linear layers) of the MT-BERT and replace it with a new one, then we use a smaller learning rate to train the network.

4 Experiments

We evaluate the proposed MT-BERT on 8 tasks in BLUE benchmarks. We compare three types of models: (1) existing start-of-the-art BERT models fine-tuned directly on each task, respectively; (2) refinement MT-BERT with multi-task training (MT-BERT-Refinement); and (3) MT-BERT with fine-tuning (MT-BERT-Fine-Tune).

4.1 Datasets

We evaluate the performance of the models on 8 datasets in the BLUE benchmark used by (Peng et al., 2019). Table 1 gives a summary of these datasets. Briefly, ClinicalSTS is a corpus of sentence pairs selected from Mayo Clinics's clinical data warehouse (Wang et al., 2018). The i2b2 2010 dataset was collected from three different hospitals and was annotated by medical practitioners for eight types of relations between problems and treatments (Uzuner et al., 2011). MedNLI is a collection of sentence pairs selected from MIMIC-III (Shivade, 2017). For a fair comparison, we use the same training, development and test sets to train and evaluate the models. ShARe/CLEF is a collection of 299 de-identified clinical free-text notes from the MIMIC-II database (Suominen et al., 2013). This corpus is for disease entity recognition.

In the biomedical domain, the ChemProt consists of 1,820 PubMed abstracts with chemical-protein interactions (Krallinger et al., 2017). The DDI corpus is a collection of 792 texts selected from the DrugBank database and other 233 Medline abstracts (Herrero-Zazo et al., 2013). These two datasets were used in the relation extraction task for various types of relations. BC5CDR is a collection of 1,500 PubMed titles and abstracts selected from the CTD-Pfizer corpus and was used in the named entity recognition task for chemical and disease entities (Li et al., 2016).

4.2 Training

Our implementation of MT-BERT is based on the work of (Liu et al., 2019c).[1] We trained the model on one NVIDIA® V100 GPU using the PyTorch framework. We used the Adamax

[1] https://github.com/namisan/mt-dnn

Corpus	Task	Metrics	Domain	Train	Dev	Test
ClinicalSTS	Sentence similarity	Pearson	Clinical	675	75	318
ShARe/CLEFE	NER	F1	Clinical	4,628	1,075	5,195
i2b2 2010	Relation extraction	F1	Clinical	3,110	11	6,293
MedNLI	Inference	Accuracy	Clinical	11,232	1,395	1,422
BC5CDR disease	NER	F1	Biomedical	4,182	4,244	4,424
BC5CDR chemical	NER	F1	Biomedical	5,203	5,347	5,385
DDI	Relation extraction	F1	Biomedical	2,937	1,004	979
ChemProt	Relation extraction	F1	Biomedical	4,154	2,416	3,458

Table 1: Summary of eight tasks in the BLUE benchmark. More details can be found in (Peng et al., 2019).

Model	ClinicalSTS	i2b2 2010 re	MedNLI	ShARe/CLEFE	*Avg*
BlueBERT$_{clinical}$	0.848	0.764	0.840	0.771	*0.806*
MT-BlueBERT-Refinement$_{clinical}$	0.822	0.745	0.835	0.826	*0.807*
MT-BlueBERT-Fine-Tune$_{clinical}$	0.840	0.760	**0.846**	**0.831**	***0.819***

Table 2: Test results on clinical tasks.

Model	ChemProt	DDI	BC5CDR disease	BC5CDR chemical	*Avg*
BlueBERT$_{biomedical}$	0.725	0.739	0.866	0.935	*0.816*
MT-BlueBERT-Refinement$_{biomedical}$	0.714	0.792	0.824	0.930	*0.815*
MT-BlueBERT-Fine-Tune$_{biomedical}$	**0.729**	**0.820**	0.865	0.931	***0.836***

Table 3: Test results on biomedical tasks.

optimizer (Kingma and Ba, 2015) with a learning rate of $5e^{-5}$, a batch size of 32, a linear learning rate decay schedule with warm-up over 0.1, and a weight decay of 0.01 applied to every epoch of training by following (Liu et al., 2019c). We use the BioBERT (Lee et al., 2020), BlueBERT base model (Peng et al., 2019), and ClinicalBERT (Alsentzer et al., 2019) as the domain-specific language model[2]. As a result, all the tokenized texts using wordpieces were chopped to spans no longer than 128 tokens. We set the maximum number of epochs to 100. We also set the dropout rate of all the task-specific layers as 0.1. To avoid the exploding gradient problem, we clipped the gradient norm within 1. To fine-tune the MT-BERT on specific tasks, we set the maximum number of epochs to 10 and learning rate e^{-5}.

4.3 Results

One of the most important criteria of building practical systems is fast adaptation to new domains. To evaluate the models on different domains, we multi-task learned various MT-BERT on BLUE biomedical tasks and clinical tasks, respectively. BlueBERT$_{clinical}$ is the base BlueBERT model pretrained on PubMed abstracts and MIMIC-III clinical notes, and fine-tuned for each BLUE task on task-specific data. MT- model are the proposed models described in Section 3. We used the pre-trained BlueBERT$_{clinical}$ to initialize its shared layers, refined the model via MTL on the BLUE tasks (MT-BlueBERT-Refinement$_{clinical}$). We keep fine-tuning the model for each BLUE task using task-specific data, then got MT-BlueBERT-Fine-Tune$_{clinical}$.

Table 2 shows the results on clinical tasks. MT-BlueBERT-Fine-Tune$_{clinical}$ created new state-of-the-art results on 2 tasks and pushing the benchmark to 81.9%, which amounts to 1.3% absolution improvement over BlueBERT$_{clinical}$ and 1.2% absolute improvement over MT-BlueBERT-Refinement$_{clinical}$. On the ShAReCLEFE task, the model gained the largest improvement by 6%. On

[2] https://github.com/ncbi-nlp/bluebert

Figure 2: Pairwise MTL relationships in clinical (left) and biomedical (right) domains.

Model	ClinicalSTS	i2b2 2010 re	MedNLI	ShARe/CLEFE	*Avg*
MT-ClinicalBERT-Fine-Tune	0.816	0.746	0.834	0.817	*0.803*
MT-BioBERT-Fine-Tune	0.837	0.741	0.832	0.818	*0.807*
MT-BlueBERT-Fine-Tune$_{biomedical}$	0.824	0.738	0.824	0.825	*0.803*
MT-BlueBERT-Fine-Tune$_{clinical}$	**0.840**	**0.760**	**0.846**	**0.831**	***0.819***

Table 4: Test results of MT-BERT-Fine-Tune models on clinical tasks.

Model	ChemProt	DDI	BC5CDR disease	BC5CDR chemical	*Avg*
MT-BioBERT-Fine-Tune	**0.729**	0.812	0.851	0.928	*0.830*
MT-BlueBERT-Fine-Tune$_{biomedical}$	**0.729**	**0.820**	**0.865**	**0.931**	***0.836***
MT-BlueBERT-Fine-Tune$_{clinical}$	0.714	0.792	0.824	0.930	*0.815*

Table 5: Test results of MT-BERT-Fine-Tune models on biomedical tasks.

the MedNLI task, the MT model gained improvement by 2.4%. On the remaining tasks, the MT model also performed well by reaching the state-of-the-art performance with less than 1% differences. When compared the models with and without fine-tuning on single datasets, Table 2 shows that the multi-task refinement model is similar to single baselines on average. Consider that MT-BlueBERT-Refinement$_{clinical}$ is one model while BlueBERT$_{clinical}$ are 4 individual models, we believe the MT refinement model would bring the benefit when researchers are in the hassle of choosing the suitable model for new problems or problems with limited training data.

In biomedical tasks, we used BlueBERT$_{biomedical}$ as the baseline because it achieved the best performance on the BLUE benchmark. Table 3 shows the similar results as in the clinical tasks. MT-BlueBERT-Fine-Tune$_{biomedical}$ created new state-of-the-art results on 2 tasks and pushing the benchmark to 83.6%, which amounts to 2.0% absolute improvement over BlueBERT$_{biomedical}$ and 2.1% absolute improvement over MT-BlueBERT-Refinement$_{biomedical}$. On the DDI task, the model gained the largest improvement by 8.1%.

4.4 Discussion

4.4.1 Pairwise MTL

To investigate which tasks are beneficial or harmful to others, we train on two tasks jointly using MT-BlueBERT-Refinement$_{biomedical}$ and MT-BlueBERT-Refinement$_{clinical}$. Figure 2 gives pairwise relationships. The directed green (or red and grey) edge from s to t means s improves (or decreases and has no effect on) t.

In the clinical tasks, ShARe/CLEFE always gets benefits from multi-task learning the remaining 3 tasks as the incoming edges are green. One factor might be that ShARe/CLEFE is an NER task

Model	BlueBERT $_{biomedical}$	BlueBERT $_{clinical}$	MT-BioBERT Fine-Tune	MT-BlueBERT Fine-Tune$_{biomedical}$	MT-BlueBERT Fine-Tune$_{clinical}$
ClinicalSTS	0.845	**0.848**	0.807	0.820	0.807
i2b2 2010 re	0.744	**0.764**	0.740	0.738	0.748
MedNLI	0.822	0.840	0.831	0.814	**0.842**
ChemProt	0.725	0.692	**0.735**	0.724	0.686
DDI	0.739	0.760	**0.810**	0.808	0.779
BC5CDR disease	**0.866**	0.854	0.849	0.853	0.848
BC5CDR chemical	**0.935**	0.924	0.928	0.928	0.914
ShARe/CLEFE	0.754	0.771	0.812	0.814	**0.830**
Avg	*0.804*	*0.807*	***0.814***	*0.812*	*0.807*

Table 6: Test results on eight BLUE tasks.

that generally requires more training data to fulfill the data-hungry need of the BERT model. ClinicalSTS helps MedNLI because the nature of both are related and their inputs are a pair of sentences. MedNLI can help other tasks except ClinicalSTS partially because the test set of ClinialSTS is too small to reflect the changes. We also note that i2b2 2010 re can be both beneficial and harmful, depending on which other tasks they are trained with. One potential cause is i2b2 2010 re was collected from three different hospitals and have the largest label size of 8.

In the biomedical tasks, both DDI and ChemProt tasks can be improved by MTL on other tasks, potentially because they are harder with largest size of label thus require more training data. In the meanwhile, BC5CDR chemical and disease can barely be improved potentially because they have already got large dataset to fit the model.

4.4.2 MTL on BERT variants

First, we would like to compare multi-task learning on BERT variants: BioBERT, ClinicalBERT, and BlueBERT. In the clinical tasks (Table 4), MT-BlueBERT-Fine-Tune$_{clinical}$ outperforms other models on all tasks. When compared the MTL models using BERT model pretrained on PubMed only (rows 2 and 3) and on the combination of PubMed and clinical notes (row 4), it shows the impact of using clinical notes during the pretraining process. This observation is consistently as shown in (Peng et al., 2019). On the other hand, MT-ClinicalBERT-Fine-Tune, which used ClinicalBERT during the pretraining, drops ∼1.6% across the tasks. The differences between ClinicalBERT and BlueBERT are at least in 2-fold. (1) ClinicalBERT used "cased" text while BlueBERT used "uncased" text; and (2) the number of epochs to continuously pretrained the model. Given that there are limited details of pretraining ClinicalBERT, further investigation may be necessary.

In the biomedical tasks, Table 5 shows that MT-BioBERT-Fine-Tune and MT-BlueBERT-Fine-Tune$_{biomedical}$ reached comparable results and pretraining on clinical notes has a negligible impact.

4.4.3 Results on all BLUE tasks

Next, we also compare MT-BERT with its variants on all BLUE tasks. Table 6 shows that MT-BioBERT-Fine-Tune reached the best performance on average and MT-BlueBERT-Fine-Tune$_{biomedical}$ stays closely. While confusing results were obtained when combing variety of tasks in both biomedical and clinical domains, we observed again that MTL models pretrained on biomedical literature perform better in biomedical tasks; and MTL models pretrained on both biomedical literature and clinical notes perform better in clinical tasks. These observations may suggest that it might be helpful to train separate deep neural networks on different types of text genres in BioNLP.

5 Conclusions and future work

In this work, we conduct an empirical study on MTL for biomedical and clinical tasks, which so far has been mostly studied with one or two tasks. Our results provide insights regarding domain adaptation and show benefits of the MTL refinement and fine-tuning. We recommend a combination of the MTL refinement and task-specific fine-tuning approach based on the evaluation results. When learned and fine-tuned on a different domain, MT-

BERT achieved improvements by 2.0% and 1.3% in biomedical and clinical domains, respectively. Specifically, it has brought significant improvements in 4 tasks.

There are two limitations to this work. First, our results on MTL training across all BLUE benchmark show that MTL is not always effective. We are interested in exploring further the characterization of task relationships. For example, it is not clear whether there are data characteristics that help to determine its success (Martínez Alonso and Plank, 2017; Changpinyo et al., 2018). In addition, our results suggest that the model could benefit more from some specific examples of some of the tasks in Table 1. For example, it might be of interest to not using the BC5CDR corpus in the relation extraction task in future. Second, we studied one approach to MTL by sharing the encoder between all tasks while keeping several task-specific decoders. Other approaches, such as fine-tuning only the task specific layers, soft parameter sharing (Ruder, 2017), knowledge distillation (Liu et al., 2019b), need to be investigated in the future.

While our work only scratches the surface of MTL in the medical domain, we hope it will shed light on the development of generalizable NLP models and task relations that can lead to gains from MTL models over single-task setups.

Acknowledgments

This work was supported by the Intramural Research Programs of the NIH National Library of Medicine. This work was also supported by the National Library of Medicine of the National Institutes of Health under award number K99LM013001. We are also grateful to the authors of mt-dnn (https://github.com/namisan/mt-dnn) to make the codes publicly available.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.

Joachim Bingel and Anders Søgaard. 2017. Identifying beneficial task relations for multi-task learning in deep neural networks. In *EACL*, pages 164–169.

Rich Caruana. 1997. Multitask learning. *Machine Learning*, 28(1):41–75.

Soravit Changpinyo, Hexiang Hu, and Fei Sha. 2018. Multi-task learning for sequence tagging: an empirical study. In *COLING*, pages 2965–2977.

Qingyu Chen, Yifan Peng, Tiarnan Keenan, Shazia Dharssi, Elvira Agro N, Wai T. Wong, Emily Y. Chew, and Zhiyong Lu. 2019. A multi-task deep learning model for the classification of age-related macular degeneration. *AMIA 2019 Informatics Summit*, 2019:505–514.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing. In *ICML*, pages 160–167.

Gamal Crichton, Sampo Pyysalo, Billy Chiu, and Anna Korhonen. 2017. A neural network multi-task learning approach to biomedical named entity recognition. *BMC Bioinformatics*, 18:368.

Louise Deléger, Robert Bossy, Estelle Chaix, Mouhamadou Ba, Arnaud Ferré, Philippe Bessières, and Claire Nédellec. 2016. Overview of the bacteria biotope task at BioNLP shared task 2016. In *Proceedings of BioNLP Shared Task Workshop*, pages 12–22.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint: 1810.04805*.

Yongping Du, Yunpeng Pan, and Junzhong Ji. 2017. A novel serial deep multi-task learning model for large scale biomedical semantic indexing. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 533–537.

Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics*, 45:885–892.

Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. 2019. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6:96.

Dan He, David Kuhn, and Laxmi Parida. 2016. Novel applications of multitask learning and multiple output regression to multiple genetic trait prediction. *Bioinformatics (Oxford, England)*, 32:i37–i43.

María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. 2013. The DDI corpus: an annotated corpus with pharmacological substances and drug-drug interactions. *Journal of Biomedical Informatics*, 46:914–920.

Muhammad Raza Khan, Morteza Ziyadi, and Mohamed AbdelHady. 2020. MT-BioNER: multi-task learning for biomedical named entity recognition using deep bidirectional transformers. *arXiv preprint: 2001.08904*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: a method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, pages 1–15.

Martin Krallinger, Obdulia Rabal, Saber A. Akhondi, Martín Pérez Pérez, Jesús Santamaría, Gael Pérez Rodríguez, Georgios Tsatsaronis, Ander Intxaurrondo, José Antonio López1 Umesh Nandal, Erin Van Buel, Akileshwari Chandrasekhar, Marleen Rodenburg, Astrid Laegreid, Marius Doornenbal, Julen Oyarzabal, Analia Lourenço, and Alfonso Valencia. 2017. Overview of the BioCreative VI chemical-protein interaction track. In *Proceedings of the BioCreative workshop*, pages 141–146.

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 (Oxford, England)*, 36:1234–1240.

Diya Li and Heng Ji. 2019. Syntax-aware multi-task graph convolutional networks for biomedical relation extraction. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 28–33.

Fei Li, Meishan Zhang, Guohong Fu, and Donghong Ji. 2017. A neural joint model for entity and relation extraction from biomedical text. *BMC Bioinformatics*, 18:198.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database (Oxford)*, 2016.

Qingqing Li, Zhihao Yang, Ling Luo, Lei Wang, Yin Zhang, Hongfei Lin, Jian Wang, Liang Yang, Kan Xu, and Yijia Zhang. 2018. A multi-task learning based approach to biomedical entity relation extraction. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 680–682.

Shikun Liu, Edward Johns, and Andrew J. Davison. 2019a. End-to-end multi-task learning with attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1871–1880.

Sijia Liu, Majid Rastegar Mojarad, Yanshan Wang, Liwei Wang, Feichen Shen, Sunyang Fu, and Hongfang Liu. 2018. Overview of the BioCreative/OHNLP 2018 family history extraction task. In *Proceedings of the BioCreative Workshop*, pages 1–5.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019b. Improving multi-task deep neural networks via knowledge distillation for natural language understanding. *arXiv preprint: 1904.09482*.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019c. Multi-task deep neural networks for natural language understanding. In *ACL*, pages 4487–4496.

Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. In *ICLR*.

Héctor Martínez Alonso and Barbara Plank. 2017. When is multitask learning effective? semantic sequence prediction under varying data conditions. In *EACL*, pages 44–53.

Tahir Mehmood, Alfonso E Gerevini, Alberto Lavelli, and Ivan Serina. 2019. Multi-task learning applied to biomedical named entity recognition task. In *Italian Conference on Computational Linguistics*.

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 Workshop on Biomedical Natural Language Processing (BioNLP)*, pages 58–65.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint: 1706.05098*.

Xue Shi, Dehuan Jiang, Yuanhang Huang, Xiaolong Wang, Qingcai Chen, Jun Yan, and Buzhou Tang. 2019. Family history information extraction via deep joint learning. *BMC medical informatics and decision making*, 19:277.

Chaitanya Shivade. 2017. Mednli – a natural language inference dataset for the clinical domain.

Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, Wendy W. Chapman, Guergana Savova, Noemie Elhadad, Sameer Pradhan, Brett R. South, Danielle L. Mowery, Gareth J. F. Jones, et al. 2013. Overview of the ShARe/CLEF eHealth evaluation lab 2013. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 212–231.

Özlem Uzuner, Brett R. South, Shuying Shen, and Scott L. DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association : JAMIA*, 18:552–556.

Xi Wang, Jiagao Lyu, Li Dong, and Ke Xu. 2019a. Multitask learning for biomedical named entity recognition with cross-sharing structure. *BMC Bioinformatics*, 20:427.

Xiaogang Wang, Cha Zhang, and Zhengyou Zhang. 2009. Boosted multi-task learning for face verification with applications to web image and video search. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 142–149.

Xuan Wang, Yu Zhang, Xiang Ren, Yuhao Zhang, Marinka Zitnik, Jingbo Shang, Curtis Langlotz, and Jiawei Han. 2019b. Cross-type biomedical named entity recognition with deep multi-task learning. *Bioinformatics (Oxford, England)*, 35:1745–1752.

Yanshan Wang, Naveed Afzal, Sunyang Fu, Liwei Wang, Feichen Shen, Majid Rastegar-Mojarad, and Hongfang Liu. 2018. MedSTS: a resource for clinical semantic textual similarity. *Language Resources and Evaluation*, pages 1–16.

Junjie Xing, Kenny Zhu, and Shaodian Zhang. 2018. Adaptive multi-task transfer learning for Chinese word segmentation in medical text. In *COLING*, pages 3619–3630.

Kui Xue, Yangming Zhou, Zhiyuan Ma, Tong Ruan, Huanhuan Zhang, and Ping He. 2019. Fine-tuning BERT for joint entity and relation extraction in Chinese medical text. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 892–897.

Jianliang Yang, Yuenan Liu, Minghui Qian, Chenghua Guan, and Xiangfei Yuan. 2019. Information extraction from electronic medical records using multitask recurrent neural network with contextual word embedding. *Applied Sciences*, 9(18):3658.

Qun Zhang, Zhenzhen Li, Dawei Feng, Dongsheng Li, Zhen Huang, and Yuxing Peng. 2018. Multitask learning for Chinese named entity recognition. In *Advances in Multimedia Information Processing – PCM*, pages 653–662.

Yu Zhang and Qiang Yang. 2017. A survey on multi-task learning. *arXiv preprint: 1707.08114*.

Sendong Zhao, Ting Liu, Sicheng Zhao, and Fei Wang. 2019. A neural multi-task learning framework to jointly model medical named entity recognition and normalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 817–824.

Jiayu Zhou, Lei Yuan, Jun Liu, and Jieping Ye. 2011. A multi-task learning formulation for predicting disease progression. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 814–822.