A Multi-Task Approach for Improving Biomedical Named Entity Recognition by Incorporating Multi-Granularity Information

Yiqi Tong, Yidong Chen*, Xiaodong Shi
School of Informatics, Xiamen University, Xiamen, China
yqtong@stu.xmu.edu.cn
{ydchen, mandel}@xmu.edu.cn

Abstract

Neural biomedical named entity recognition (BioNER) methods usually require a large amount of annotated data, while the annotated BioNER datasets are often difficult to obtain and small in scale due to the limitations of privacy, ethics and high degree of specialization. To alleviate the lack of training samples, unlike conventional methods that only use token-level information, this paper proposes a method that simultaneously utilize the latent multi-granularity information in the dataset. Concretely, the proposed model is based on a multi-task approach, which leverages different training objectives by introducing auxiliary tasks, i.e. binary classification, multi-class and multi-token classification. Experimental results over three BioNER datasets show that the proposed model produces better performance over the BioBERT baseline and can get more than 3% improvements of F1-score in low-resource scenarios. Finally, we released our code at https://github.com/zgzjdx/MT-BioNER.

1 Introduction

Biomedical named entity recognition (BioNER) aims to identify entity mentions such as gene/protein, disease and chemicals from unstructured text. Such information is useful for downstream natural language processing (NLP) tasks like relation extraction (Zhou et al., 2014), automatic abstracting (Mishra et al., 2014) and question answering (Athenikos and Han, 2010), etc. Different from those named entity recognition (NER) tasks for general domain like news, BioNER is particular challenge due to the naming complexity (Liu et al., 2015), large variations in same entity names (Jia et al., 2019; Kim et al., 2019), and new entity mentions rapidly reported in scientific publications (Luo et al., 2018). These various factors lead to the small number and size of current BioNER datasets. In recent years, neural BioNER has become a main approach because of its outstanding performance (Lample et al., 2016; Habibi et al., 2017; Yadav and Bethard, 2019). Some researchers have investigated introducing multi-task learning (Crichton et al., 2017; Khan et al., 2020) and pre-training (Peng et al., 2019; Lee et al., 2020) to solve the problem of lacking extensive training data and boost the performance of BioNER model. However, few of them combined these two methods together and tried to transfer sentence-level knowledge to tokens (Rei and Søgaard, 2019; Kruengkrai et al., 2020), which had proven to be effective in

Text: **Peroxydase** reaction stains were negative .
Contain entities: **True**
Number of entities: **One**
Multi-token Label: B-Single O O O O O
NER Label: B-GENE O O O O O

Text: **Brain disease** and molecular analysis in **myotonic dystrophy** .
Contain entities: **True**
Number of entities: **Two**
Multi-token Label: B-Multi I-Multi O O O O B-Multi I-Multi O
NER Label: B-Disease I-Disease O O O O B-Disease I-Disease O

Figure 1: Examples from our constructed dataset. In our work, we designed three auxiliary tasks to help improving the main NER task. Two of them are sentence-level tasks and the other one is a token-level task. Concretely, the first sentence-level task predicts whether or not a sentence contains entities; the second sentence-level task predicts how many entities a sentence contains; and the token-level task predicts whether or not a given token belongs to a multi-token entity. Clearly, to support training the auxiliary tasks, additional labels have been added in our data. However, please note that, the additional labels could be derived from the original NER labels and do not need additional manual annotations. In a word, what we have done in this paper is try to use the multi-granularity information implied in the original dataset to improve the performance of BioNER.

*Corresponding author.

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 4804–4813

other domains (Abhishek et al., 2017).

In this paper, we focus on improving BioNER by exploiting multi-granularity information implied in the dataset, without depending on additional manually annotated data. As shown in Figure 1, Besides the main sequence labeling task, we employ three related classification tasks, i.e. a binary classification task for predicting whether a sentence contains entities or not, a multi-class classification task for predicting how many entities a sentence contains and a multi-token entity classification task (Hu et al., 2020). In the rest of this paper, these three tasks will be named bCLS, mCLS and mtCLS, respectively, while the main task will be named NER. Our primary motivation is to mine useful training signals from coarse-grained classification to guide a more robust and interpretable token-level representation.

Our key contributions can be summarized as follows:

- To take full advantage of the implicit information contained in NER dataset, we present a multi-task model for jointly learning sentence-level and token-level labels, which incorporates BioBERT (Lee et al., 2020) as text encoding layers and shares the hidden states between different tasks. To the best of our knowledge, we are the first to introduce different grained-level information in BioNER domain.

- Experimental results on three datasets show that our proposed method is effective, especially in the low-resource scenarios.

- We performed preliminary pair-wise comparison analysis to investigate the relations between tasks and pointed out that token-level labels are more helpful for sentence-level tasks. While at the same granularity, high-difficulty tasks are more helpful to low-difficulty tasks.

2 Related work

Traditional BioNER methods could be divided into rule- or dictionary-based approaches (Tjong Kim Sang and De Meulder, 2003; Kulick et al., 2004; Gerner et al., 2010). And recent works had shown neural network architecture based BioNER methods achieved promising results. Habibi et al. (2017) used a LSTM-CRF model, which was completely agnostic to entity types. Crichton et al. (2017), on the contrary, used a CNN-based model that takes tokens and their surrounding tokens as input. To solve the label inconsistent problem, Luo et al. (2018) proposed a Att-BiLSTM-CRF model and achieved better performance with little feature engineering.

The neural BioNER system is known to be extremely data-intensive, while the available training datasets are relatively small in scale. To tackle this problem, research has been conducted and language models and multi-task learning have been shown to be effective to deal with this problem (Peters et al., 2018; Liu et al., 2019). Jia et al. (2019) proposed a cross-domain NER model, which extracted knowledge from raw texts through a novel parameter sharing network. Yoon et al. (2019) proposed CollaboNet, which consists of multiple BiLSTM-CRF models where models could send information to one another for more accurate predictions, and got best F1-score at that time. Although these studies have exploited additional token-level information from auxiliary tasks or language models, they do not consider information from other levels that contained in the NER dataset.

More recently, a transformer-based (Vaswani et al., 2017) large-scale pre-training language model, called BERT (Devlin et al., 2018), led to impressive gains on several NLP benchmarks and the domain-specific BERTs, such as blueBERT (Peng et al., 2019), BioBERT (Lee et al., 2020), SciBERT (Beltagy et al., 2019) and PubmedBERT (Gu et al., 2020), have largely outperformed the previous state-of-the-art BioNER systems. But research on multi-task learning based on BERT is still few, and the association between tasks needed to be further explored (Khan et al., 2020; Vu et al., 2020).

The most similar work to ours is the findings of Kruengkrai et al. (2020). However, they only focused on introducing one auxiliary task that requires additional manual annotations, while we attempted to try multiple auxiliary tasks, and our proposed method did not rely on other additional annotations, except for BioNER.

3 The proposed model

3.1 Tasks

As mentioned in Section 1, our model involves four tasks: bCLS, mCLS, mtCLS and NER. The goal is to optimize the token-level representation of BioBERT by introducing auxiliary tasks (bCLS, mCLS, mtCLS) and improve the perfor-

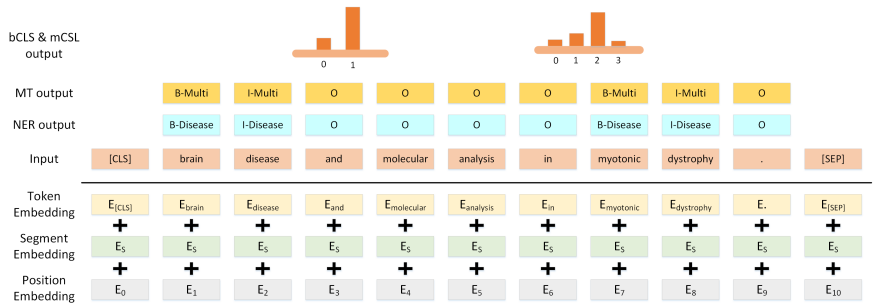

Figure 2: The input and output descriptions of the proposed model. Actually, our model involve three input embeddings and four outputs, and we adopt two special token [CLS] and [SEP] to represent the beginning and the end of the input sentence, respectively.

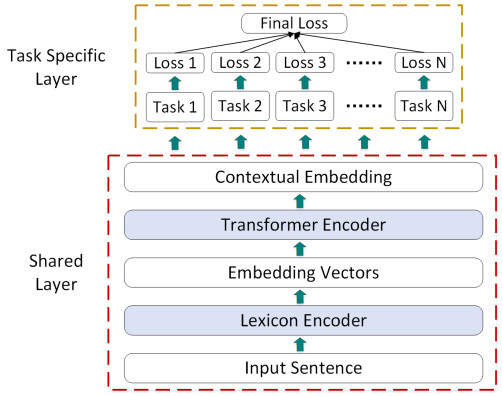

Figure 3: The architecture of our multi-task BioBERT-based model that can jointly learn sentence-level and token-level labels.

mance of the main task (NER). The pre-training model BioBERT are shared across all tasks by hard parameter sharing (Ruder, 2017). The input sequence and output labels of our proposed model are represented in Figure 2. Given a sentence $X = \{x_1, ..., x_i, ..., x_n\}$, where x_i is a token, n is the length of the input sequence. The first token of each X is always a special classification embedding [CLS] and the transformer encoder module maps X into a sequence of input embedding vectors, which are the sum of the token, segment and position embeddings. The detailed description of each task is shown as follows. **bCLS:** This is a sentence-level binary classification task. Given X, the goal is to predict whether it contains entities or not. In some cases, for X that do not contain entities, the model may incorrectly predicts that it contains entities. Or for X that contains entities, the model may incorrectly predicts that it not contain entities. Therefore, we design bCLS task with the hope of solving this problem by introducing a global guidance information.

mCLS: This is a sentence-level multi-classification task. Given X, the goal is to predict how many entities it contains. To balance label numbers, this paper set mCLS as a 4-classification task, which X contains 0, 1 and 2 entities is set to 0, 1 and 2, respectively, while X with more than 2 entities are all set to 3. Compared with bCLS, mCLS is more difficult and we introduce this task to alleviate the under- or over-recognition entity problem.

mtCLS: Multi-token classification is a token-level 3-classification problem. Given x_i in X, the goal is to predict whether it belongs to a multi-token entity like "brain disease" or a single-token entity like "peroxydase" or neither. Our motivation for introducing this task is that if the model knows whether x_i is multi-token entity or single-token entity or neither, it can alleviate the entity boundary problem.

NER: Given X, NER aims to predict corresponding labels $Y = \{y_1, ..., y_i, ..., y_n\}$, where y_i is predefined and differs according to the annotation scheme such as BIO and BIOES. We use this main task to measure the model performance and effectiveness of auxiliary tasks.

3.2 Architecture

The overall architecture of our proposed model is shown in Figure 3, which mainly includes two parts: the shared encoder and task-specific layers. We use multi-task learning to jointly train the main task and auxiliary tasks, which has been shown effective for transferring knowledge among multiple tasks (Yoon et al., 2019). For the shared encoder, we take cased BioBERT-base v1.1[1] as feature extractor and hard shared its parameters. Set X as an input sequence, where x_i denotes the $i-th$ token

[1]https://github.com/naver/biobert-pretrained

in X. We represent each x_i using the pre-trained BioBERT embeddings $h_i \in \mathbb{R}^d$, where d is the dimension of hidden states. And the task-specific layers have independent parameters, which include a project layer and a classifier for generating outputs. We use the output of the shared encoder, i.e. $H = \{h_1, ..., h_i, ..., h_n\}$, as the input of task-specific layers for both sentence- and token-level tasks, as described in detailed as follows.

Sentence-level tasks. As mentioned in subsection 3.1, bCLS and mCLS are two sentence-level classification tasks. Different from the standard BERT-based classification models, which optimize the [CLS] token (Sun et al., 2019) to perform classification. Our model aims at optimizing the token representations of the shared encoder by sentence-level labels. Therefore, we created a fixed size vector by applying mean/max pooling (Reimers and Gurevych, 2019) over H, which encourages the model to capture the most useful local features encoded in hidden states. Finally, the probability of class k is predicted by a linear layer and a logistic regression with softmax.

$$P(m|X) = softmax(Wh + b) \quad (1)$$

where $h \in \mathbb{R}^d$ is the pooling output of model, $W \in \mathbb{R}^{d \times m}$ and $b \in \mathbb{R}^m$ are trainable weight matrix and bias. m denotes the number of category labels, which is 2 for bCLS and 4 for mCLS. Finally, the loss for our sentence-level task is calculated as follows:

$$L_S = -\sum_m \sigma(y_m = \hat{y}) log(P(m|X)) \quad (2)$$

where $\sigma(y_m = \hat{y}) = 1$ if the classification $\hat{y}$ of X is the right ground-truth label for the class m. Otherwise, $\sigma(y_m = \hat{y}) = 0$.

Token-level tasks. As mentioned in subsection 3.1, mtCLS and NER are two token-level classification tasks[2]. Given the dataset D, which consists of N training samples, i.e. $D = (x_j, y_j)_{j=1}^N$, where j denotes the sentence index in D. To train the token-level tasks, we minimize the negative log-likelihood of the correct label sequences over D with the loss function defined as follows:

$$L_T = -\frac{1}{N} \sum_{j=1}^{N} log(P(y_j)|H_j) \quad (3)$$

[2]Generally, NER was treated as a sequence labeling problem. However, for a fair comparison with previous works, instead of using sequence labeling algorithms such as Conditional Random Field (CRF) (Wallach, 2004) in task-specific layers, we still use softmax for token-level tasks.

Algorithm 1 Training a MT-BioBERT model

Initialize: Model parameter of shared layers $\theta_i^{BioBERT}$ by BioBERT and task-specific layer θ_i^{task} randomly. Max epochs, max sequence length, learning rate, etc.

Input: Dataset D

shuffle D
for each epoch in $epoch_{max}$ do
  for each b_t in D do
    # b_t is a mini-batch of D
    Compute Loss: $L(\theta) = \alpha L(\theta)_{ner} + \beta L(\theta)_{mtCLS} + \gamma L(\theta)_{bCLS} + \delta L(\theta)_{mCLS}$
    $L(\theta)_{bCLS}$ = Eq.2 for binary classification
    $L(\theta)_{mCLS}$ = Eq.2 for multi-class classification
    $L(\theta)_{ner}$ = Eq.3 for sequence labeling
    $L(\theta)_{mtCLS}$ = Eq.3 for multi-token classification
    Compute gradient Δ_θ
    Update model $\theta = \theta - \epsilon \Delta_\theta$
  end
end

Dataset	Sentences	Entity Type and Counts
BC2GM	20,131	Gene/Protein (24,583)
BC5CDR	13,938	Chemical(15,935), Disease(12,852)
NCBI	7,287	Disease(6,881), Gene/Protein(35,336)

Table 1: Dataset description. We use BC2GM (Smith et al., 2008), BC5CDR (Li et al., 2016) and NCBI (Doğan et al., 2014) to conduct our experiments.

where $H_j \in \mathbb{R}^{n \times d}$ is the hidden state of x_j.

Algorithm 1 provides the procedure for our cross-task joint training method, where α, β, γ, δ are hyper-parameters. Moreover, the final loss of the proposed model is calculated by weighted summing the losses of different tasks.

4 Experiments

4.1 Datasets

We evaluated the performance of proposed approach on three benchmark datasets[3] used by Wang et al. (2019b); Yoon et al. (2019); Lee et al. (2020); Khan et al. (2020). Table 1 gives the statics of these datasets. Following previous works, we merged the

[3]https://github.com/cambridgeltl/MTL-Bioinformatics-2016

Model	BC2GM			BC5CDR-chem			NCBI-disease		
	P	R	F1	P	R	F1	P	R	F1
Baseline systems									
Habibi et al. (2017)	81.57	79.48	80.51	87.60	86.25	86.92	86.11	85.49	85.80
Sachan et al. (2018)	81.81	81.57	81.69	88.10	90.49	89.28	86.41	88.31	87.34
Devlin et al. (2018)	81.17	82.42	81.79	90.94	91.38	91.16	84.12	87.19	85.63
Wang et al. (2019b)	82.10	79.42	80.74	93.09	89.56	91.29	85.86	86.42	86.14
Yoon et al. (2019)	80.49	78.99	79.73	94.26	92.38	93.31	85.48	87.27	86.36
Khan et al. (2020)	82.10	84.04	83.01	88.46	90.52	89.48	86.73	89.70	88.19
Beltagy et al. (2019)	-	-	83.36	-	-	92.51	-	-	88.25
Gu et al. (2020)	-	-	83.82	-	-	92.85	-	-	89.13
SOTA (Lee et al., 2020)	85.16	83.65	84.40	93.27	93.61	93.44	89.04	89.69	89.36
SOTA*	83.57	85.22	84.38	92.98	94.24	93.60	87.83	90.21	89.00
Our methods									
Ours_CLS	83.21	85.00	84.09	92.95	94.28	93.61	88.29	90.31	89.29
Ours_MEAN	84.31	83.78	84.04	94.08	93.24	93.66	87.75	91.77	89.71
Ours_MAX	84.42	85.14	**84.78**	93.29	94.69	**93.98**	88.90	90.94	**89.91**

Table 2: Model performance comparison on the three benchmark datasets, where SOTA* is our reproduce results of BioBERT, Ours_CLS uses the [CLS] token for sentence-level tasks and Ours_MEAN or Ours_MAX adopts the mean or max pooling strategy for sentence-level tasks, respectively.

Dataset	Metric	FULL-SIZE			50%-SIZE			25%-SIZE			10%-SIZE		
		CS-MTM	SOTA	Ours	CS-MTM	SOTA	Ours	CS-MTM	SOTA	Ours	CS-MTM	SOTA	Ours
BC2GM	P	83.21	85.16	84.42	79.37	82.21	81.78	79.44	80.15	80.82	72.95	75.27	76.60
	R	85.74	83.65	85.14	85.05	83.73	84.52	78.98	81.60	82.56	75.39	79.27	79.32
	F1	84.41	84.40	**84.78**	82.12	82.96	**83.13**	79.21	80.87	**81.68**	74.15	77.22	**77.93**
BC5CDR-chem	P	-	93.27	93.29	-	91.97	92.00	-	89.99	91.38	-	89.78	90.29
	R	-	93.61	94.69	-	92.37	93.96	-	92.48	93.30	-	90.90	91.33
	F1	-	93.44	**93.98**	-	92.17	**92.97**	-	91.22	**92.33**	-	90.34	**90.81**
NCBI-disease	P	86.59	89.04	88.90	84.72	85.77	92.50	81.00	81.22	83.96	79.32	79.69	83.65
	R	86.42	89.69	90.94	84.76	91.67	86.04	81.00	88.33	90.52	74.40	80.52	83.13
	F1	86.50	89.36	**89.91**	84.74	88.62	**89.16**	81.00	85.79	**87.12**	76.68	80.10	**83.39**

Table 3: Impacts of the dataset size. We keep the test set unchanged and only cut the training set.

training and developing sets for the model training. As a part of the data preprocessing step, token labels were encoded using the standard BIO scheme (Reimers and Gurevych, 2017). In this scheme, for example, a token describing a disease entity is tagged with "B-Disease" if it is at the beginning of the entity, and "I-Disease" if it is inside the disease entity. Other tokens that not describing entities of interest are tagged as "O".

4.2 Settings

Following the work of Peng et al. (2020), all datasets are trained with the batch size of 32, maximum sequence length of 256 and a dropout (Srivastava et al., 2014) with the probability of 0.1 after the shared encoder. We use the Adam optimizer (Kingma and Ba, 2014) with a learning rate $5e^{-}5$ for BC2GM, BC5CDR-chem and $1e^{-}5$ for NCBI-disease. The training procedure contains 50 epochs for BC2GM, BC5CDR-chem and 100 epochs for NCBI-disease. A linear learning rate decay schedule with warm-up over 0.1, and a weight decay of 0.01 applied to every epochs of the training by following Liu et al. (2019). Finally, all models were trained on NVIDIA RTX TITAN and used standard F1 metrics[4] to evaluate the overall performance.

4.3 Performance

We compare our model with single-task models, such as LSTM-CRF (Habibi et al., 2017), BiLM-NER (Sachan et al., 2018), domain-specific BERTs (Devlin et al., 2019; Beltagy et al., 2019; Gu et al., 2020) and multi-task models, such as MTM-CW (Wang et al., 2019b), CollaboNET (Yoon et al., 2019), MT-BioNER (Khan et al., 2020).

Table 2 shows the overall performance of our model compared with the existing approaches on

[4]https://github.com/chakki-works/seqeval

Model	Dataset								
	BC2GM			BC5CDR-chem			NCBI-disease		
	P	R	F1	P	R	F1	P	R	F1
Ours	84.42	85.14	84.78	93.29	94.69	93.98	88.90	90.94	89.91
w/o bCLS	83.72	84.71	84.21	92.82	94.06	93.43	87.41	91.14	89.24
w/o mCLS	84.48	84.28	84.39	93.82	93.67	93.75	88.12	91.15	89.61
w/o mtCLS	83.44	84.82	84.12	93.98	93.57	93.78	87.48	91.67	89.52
w/o Joint	83.68	83.26	83.47	93.41	92.88	93.14	87.30	89.48	88.37

Table 4: Ablation study results, where **w/o Joint** means using the training strategy of MT-DNN (Liu et al., 2019) to replace our training algorithm.

Dataset	Model	Result
BC2GM	BioBERT	They are growth – inhibited by TGF - beta1 .
	Ours	They are growth – inhibited by TGF - beta1 .
BC5CDR-chem	BioBERT	Anaesthesia with a propofol infusion and avoidance of serotonin onists provided a nausea - free
	Ours	Anaesthesia with a propofol infusion and avoidance of serotonin onists provided a nausea - free
NCBI-disease	BioBERT	We conclude that paternal transmission of congenital DM is rare and preferentially occurs with onset of DM ...
	Ours	We conclude that paternal transmission of congenital DM is rare and preferentially occurs with onset of DM ...

Table 5: Case study on three datasets, where words in red and in green represent incorrectly and correctly recognized entities, respectively.

the three benchmark datasets, where the current SOTA model is BioBERT. In line with the expectations, Ours_MAX, which uses the max pooling strategy, achieved the best results, with the improvements of 0.40, 0.37 and 0.91 F1-scores for the three datasets, respectively. On the contrary, Ours_CLS and Ours_MEAN achieved negative results from our experiments. The above-mentioned phenomenon is consistent with Reimers and Gurevych (2019); Kruengkrai et al. (2020). Another interesting result is that our best model also achieves higher recall score than all the other approaches expect SOTA* result in BC2GM, which indicates that the introducing of coarse-grained tasks helps the model to predict more positive results.

To simulate low-resource scenarios, we also used the reduced training datasets by randomly removing sentences in training sets, while test sets are not modified. As shown in Table 3, CS-MTM was a multi-task model with cross-sharing structure proposed by Wang et al. (2019a), we record the performance of different situations and the best F1-score for each resource size are bolded. When training sets are reduced and test sets are kept, the missing information in removed sentences make all models produce worse results. However, for 50%-size, 25%-size and 10%-size datasets, our model can get an average of 0.56, 0.79 and 1.72 F1-score improvements over the BioBERT, which demonstrates our designed auxiliary tasks can regularize model to generate more robust token-level representations. For BC2GM, BC5CDR-chem and NCBI-disease in all data size, our model can get an average of 0.47, 0.95 and 1.26 improvements in F1-score, which the largest improvement is observed on NCBI-disease. The smaller the training set is the larger improvement could achived by our model. This finding proves our method is more effective in low-resource scenarios. Specifically, on 10%-size NCBI-disease, our model can get 3.29 F1-score improvements over the BioBERT.

To prove that our joint training algorithm is effective, we plot the performance curve of different tasks, which can be found in Figure 4. Moreover, different task combinations can produce different results in multi-task learning. To measure the impact of our designed auxiliary tasks and training algorithm, we conducted ablation studies and showed in Table 4.

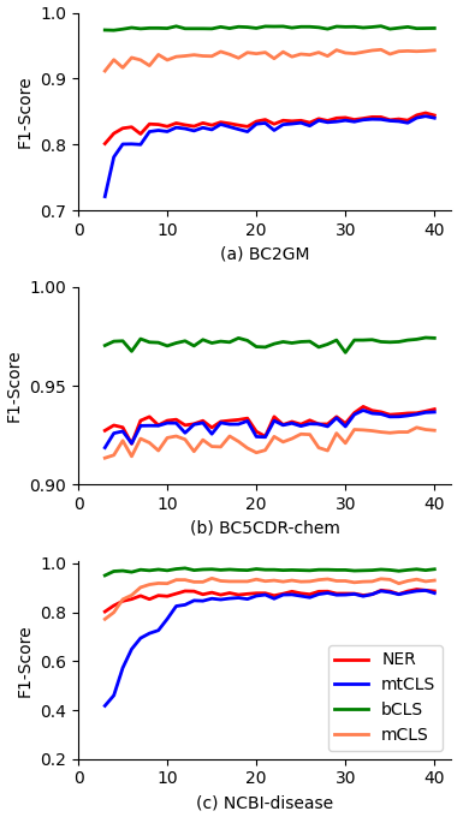

Figure 4: Performance curves for different tasks of our proposed model. Where x-axis is the number of training epoch and y-axis is F1-score. During the model training process, the tasks complement each other and gradually converge.

From the results of the ablation experiment, removing the joint training algorithm leads to a consistent drop in the F1-scores. Compared with the results of Liu et al. (2019), Khan et al. (2020) and Peng et al. (2020), we point out that multi-task learning algorithms such as MT-DNN require a large amount of training data to achieve improvements. Furthermore, all of the auxiliary tasks are helpful to the main task but the impact of different tasks vary. Specifically, mtCLS is the best partner for BC2GM dataset and bCLS can bring the most improvement for BC5CDR-chem and NCBI-disease dataset. This phenomenon shows that different BioNER datasets have different recognition difficulties. For example, the recognition difficulty of BC2GM may mainly related to entity boundary problem, while the recognition difficulty of BC5CDR-chem and NCBI-disease is the entity sparsity problem. Therefore, for BC5CDR-chem and NCBI-disease, the model trends to incorrectly recognize entities in sentences that do not contain entities. This finding is consistent with our statistical results, where such cases are 2.38%, 2.71% and 2.87% on the three datasets, respectively. Compared to bCLS, mCLS is less helpful. This implies that the effect of auxiliary tasks in multi-task leaning is closely related to their performances. In fact, the classification performances of mCLS are lower than that of bCLS due to its higher difficulty.

4.4 Case study

Table 5 shows the case study of three datasets. The BC2GM example showed the effect of bCLS task in that our model could correctly recognize the entity "TGF - beta1" while the BioBERT model fails. In the BC5CDR-chem example, the input sentence contains two entities "propofol" and "serotonin", and the BioBERT model could only identify one of them, while our model could correctly recognize two entities by incorporating the mCLS task. For the NCBI-disease example, "congenital DM" is a multi-token entity and "DM" is a single-token entity. It could be found that without the help of the mtCLS task, the BioBERT model could not capture such difference and incorrectly recognized two "DM". Overall, these examples confirm that supervised objectives at different granularities, i.e. global information and local information, can be combined to help producing better representations.

Although the case study show that our model with auxiliary tasks outperformed the BioBERT model, these tasks can not completely solve the above problems due to their coarser granularities. Take the bCLS task as an example, the model could noticed that current input sentence contains entities by sentence-level label, but still may trapped in the number of entities or entity boundary.

4.5 Impacts of the task relationship

In this subsection, we would like to preliminary study **the relationship between different tasks in the same domain**, such as the interaction between sentence-level tasks and token-level tasks, and whether or not tasks could help one other. Therefore, we conducted pair-wise comparison experiments, as shown in Figure 5, where x-axis is the secondary task and y-axis is the main task.

First, we point out the token-level labels are more helpful for the sentence-level tasks. For mCLS, it can get an average improvement of 0.79%, 0.54% and 0.15% on the three datasets by taking mtCLS, NER and bCLS as auxiliary tasks, respectively. Considering that mtCLS and NER are token-level tasks and bCLS is a sentence-level task,

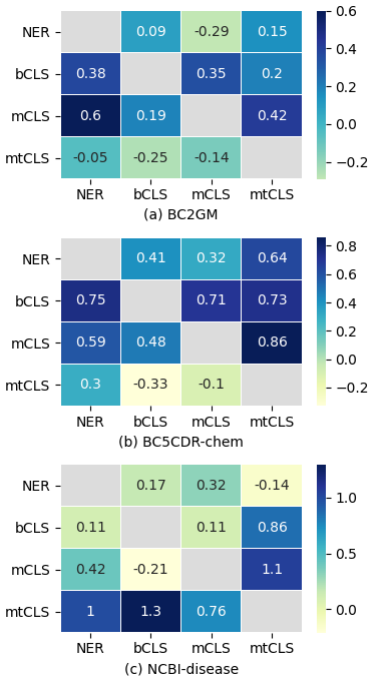

Figure 5: Results of the pair-wise experiments of our model on three datasets.

the results suggest that the coarse-grained tasks can significantly benefit from fine-grained tasks. This finding could be used to guide the choosing of the tasks for multi-task learning.

Second, the same granularity of information also contributes to each other. Concretely, bCLS and mCLS can get an average improvement of 0.39%, 0.15% from mCLS and bCLS, respectively. And mtCLS and NER can get an average improvement of 0.42%, 0.22% from NER and mtCLS, respectively. Meanwhile, the difficulty of task is also a factor that affects the effectiveness of multi-task learning, in that bCLS gets 0.24% more improvements compared to mCLS, and mtCLS gets 0.20% more improvements compared to BioNER.

In addition, the same task combinations performs differently on different datasets. For example, the combinations of mtCLS and mCLS got negative results of -0.25% and -0.33% on the BC2GM and BC5CDR-chem datasets, while achieved 1.3% boost on the NCBI-disease dataset. We guessed it may related to the transferability of specific dataset. So we visualized the task embedding of three datasets, which were generated with the method[5] proposed by Vu et al. (2020), using T-SNE (Belkina et al., 2019) dimension reduction algorithm and showed the results in Figure 6. From the visualization results, we found that the embedding distance between the same tasks (e.g., BC2GM-bCLS, BC5CDR-chem-bCLS, NCBI-disease-bCLS) or the same type of tasks is closer (e.g., NER and mtCLS, bCLS and mCLS). And the embedding distance between different types of tasks is farther (e.g., bCLS and NER), but more specific relations need further exploration.

[5]https://github.com/tuvuumass/task-transferability

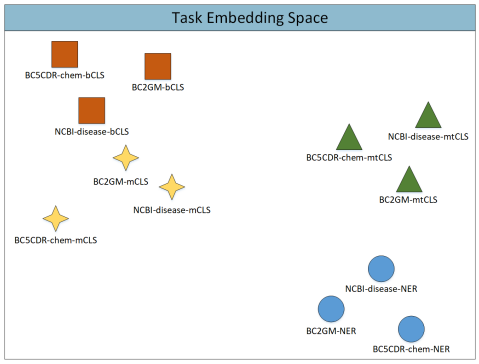

Figure 6: A 2D visualization of the tasks space.

5 Conclusion

In this work, we investigated whether coarse-grained label could benefit the token-level representation for BioNER. We had shown that the proposed BERT-based jointly sentence and token label model was valid without using external data and hand-crafted feature for BioNER in three datasets: BC2GM, BC5CDR-chem, NCBI-disease. Finally, we preliminary discussed the correlation between main task and auxiliary task.

For multi-task learning, describing and reasoning about the relations between tasks through experiments require an amount of computational resources. In future work, with domain related in mind, we will explore efficient methods for generating vectorial representations to measure the relationship between different NLP tasks.

Acknowledgements

The authors would like to thank the three anonymous reviewers for their comments on this paper. This research was supported by the National Natural Science Foundation of China under Grant Nos. 62076211, U1908216 and 61573294.

References

Abhishek Abhishek, Ashish Anand, and Amit Awekar. 2017. Fine-grained entity type classification by

jointly learning representations and label embeddings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 797–807.

Sofia J Athenikos and Hyoil Han. 2010. Biomedical question answering: A survey. *Computer methods and programs in biomedicine*, 99(1):1–24.

Anna C Belkina, Christopher O Ciccolella, Rina Anno, Richard Halpert, Josef Spidlen, and Jennifer E Snyder-Cappione. 2019. Automated optimized parameters for t-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. *Nature communications*, 10(1):1–12.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

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(1):368.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Martin Gerner, Goran Nenadic, and Casey M Bergman. 2010. Linnaeus: a species name identification system for biomedical literature. *BMC bioinformatics*, 11(1):85.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing. *arXiv preprint arXiv:2007.15779*.

Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. 2017. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):i37–i48.

Anwen Hu, Zhicheng Dou, Jian-Yun Nie, and Ji-Rong Wen. 2020. Leveraging multi-token entities in document-level named entity recognition. In *AAAI*, pages 7961–7968.

Chen Jia, Xiaobo Liang, and Yue Zhang. 2019. Cross-domain NER using cross-domain language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2464–2474, Florence, Italy. Association for Computational Linguistics.

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 arXiv:2001.08904*.

Donghyeon Kim, Jinhyuk Lee, Chan Ho So, Hwisang Jeon, Minbyul Jeong, Yonghwa Choi, Wonjin Yoon, Mujeen Sung, and Jaewoo Kang. 2019. A neural named entity recognition and multi-type normalization tool for biomedical text mining. *IEEE Access*, 7:73729–73740.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Canasai Kruengkrai, Thien Hai Nguyen, Sharifah Mahani Aljunied, and Lidong Bing. 2020. Improving low-resource named entity recognition using joint sentence and token labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5898–5905, Online. Association for Computational Linguistics.

Seth Kulick, Ann Bies, Mark Liberman, Mark Mandel, Ryan McDonald, Martha Palmer, Andrew Schein, Lyle Ungar, Scott Winters, and Peter White. 2004. Integrated annotation for biomedical information extraction. In *HLT-NAACL 2004 Workshop: Linking Biological Literature, Ontologies and Databases*, pages 61–68.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

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.

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*, 2016.

Shengyu Liu, Buzhou Tang, Qingcai Chen, and Xiaolong Wang. 2015. Drug name recognition: approaches and resources. *Information*, 6(4):790–810.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*.

Ling Luo, Zhihao Yang, Pei Yang, Yin Zhang, Lei Wang, Hongfei Lin, and Jian Wang. 2018. An attention-based bilstm-crf approach to document-level chemical named entity recognition. *Bioinformatics*, 34(8):1381–1388.

Rashmi Mishra, Jiantao Bian, Marcelo Fiszman, Charlene R Weir, Siddhartha Jonnalagadda, Javed Mostafa, and Guilherme Del Fiol. 2014. Text summarization in the biomedical domain: a systematic review of recent research. *Journal of biomedical informatics*, 52:457–467.

Yifan Peng, Qingyu Chen, and Zhiyong Lu. 2020. An empirical study of multi-task learning on bert for biomedical text mining. *arXiv preprint arXiv:2005.02799*.

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. *arXiv preprint arXiv:1906.05474*.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Marek Rei and Anders Søgaard. 2019. Jointly learning to label sentences and tokens. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6916–6923.

Nils Reimers and Iryna Gurevych. 2017. Optimal hyperparameters for deep lstm-networks for sequence labeling tasks. *arXiv preprint arXiv:1707.06799*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

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

Devendra Singh Sachan, Pengtao Xie, Mrinmaya Sachan, and Eric P Xing. 2018. Effective use of bidirectional language modeling for transfer learning in biomedical named entity recognition. In *Machine learning for healthcare conference*, pages 383–402. PMLR.

Larry Smith, Lorraine K Tanabe, Rie Johnson nee Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M Friedrich, Kuzman Ganchev, et al. 2008. Overview of biocreative ii gene mention recognition. *Genome biology*, 9(S2):S2.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*, pages 194–206. Springer.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Tu Vu, Tong Wang, Tsendsuren Munkhdalai, Alessandro Sordoni, Adam Trischler, Andrew Mattarella-Micke, Subhransu Maji, and Mohit Iyyer. 2020. Exploring and predicting transferability across NLP tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7882–7926, Online. Association for Computational Linguistics.

Hanna M Wallach. 2004. Conditional random fields: An introduction. *Technical Reports (CIS)*, page 22.

Xi Wang, Jiagao Lyu, Li Dong, and Ke Xu. 2019a. Multitask learning for biomedical named entity recognition with cross-sharing structure. *BMC bioinformatics*, 20(1):427.

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*, 35(10):1745–1752.

Vikas Yadav and Steven Bethard. 2019. A survey on recent advances in named entity recognition from deep learning models. *arXiv preprint arXiv:1910.11470*.

Wonjin Yoon, Chan Ho So, Jinhyuk Lee, and Jaewoo Kang. 2019. Collabonet: collaboration of deep neural networks for biomedical named entity recognition. *BMC bioinformatics*, 20(10):249.

Deyu Zhou, Dayou Zhong, and Yulan He. 2014. Biomedical relation extraction: from binary to complex. *Computational and mathematical methods in medicine*, 2014.