MTAAL: Multi-Task Adversarial Active Learning for Medical Named Entity Recognition and Normalization

Baohang Zhou,[1,3] **Xiangrui Cai,** [2,3] **Ying Zhang,** [1,3*] **Wenya Guo,** [1,3] **Xiaojie Yuan** [1,3]

[1] College of Computer Science, Nankai University, Tianjin 300350, China
[2] College of Cyber Science, Nankai University, Tianjin 300350, China
[3] Tianjin Key Laboratory of Network and Data Security Technology, Tianjin 300350, China
{zhoubaohang, guowenya}@dbis.nankai.edu.cn, {yingzhang, caixr, yuanxj}@nankai.edu.cn

Abstract

Automated medical named entity recognition and normalization are fundamental for constructing knowledge graphs and building QA systems. When it comes to medical text, the annotation demands a foundation of expertise and professionalism. Existing methods utilize active learning to reduce costs in corpus annotation, as well as the multi-task learning strategy to model the correlations between different tasks. However, existing models do not take task-specific features for different tasks and diversity of query samples into account. To address these limitations, this paper proposes a multi-task adversarial active learning model for medical named entity recognition and normalization. In our model, the adversarial learning keeps the effectiveness of multi-task learning module and active learning module. The task discriminator eliminates the influence of irregular task-specific features. And the diversity discriminator exploits the heterogeneity between samples to meet the diversity constraint. The empirical results on two medical benchmarks demonstrate the effectiveness of our model against the existing methods.

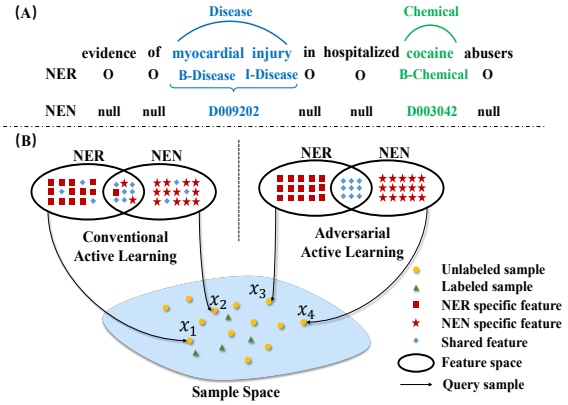

Figure 1: (A) shows a medical text sample. The labels from the NER task are strongly correlated to that of the NEN task. (B) shows two active learning schemes for the two tasks, i.e., NEN and NER. Different feature spaces influence the procedure of query samples in multi-task active learning.

Introduction

Named entity recognition (NER) and named entity normalization (NEN) are the fundamental tasks for constructing the medical knowledge graph (Wu et al. 2019) and building QA systems (Lamurias and Couto 2019). The former tries to find the boundaries of mentions from the medical text, and the latter maps mentions extracted from the medical text to standard identifiers, such as MeSH and OMIM (Zhao et al. 2019).

As shown in Figure 1(A), the locations of the NER labels correspond to those of NEN. When predicting the label of NER or NEN, the location information of one task favors the other one. To make better use of the relevance between label locations in two tasks, a multi-task learning model (Zhao et al. 2019) is proposed for the medical NER and NEN. However, it is based on supervised learning which requires a large number of labeled corpus. It is expensive to achieve, especially in the medical domain.

To reduce labeling costs, active learning is widely used and trained in a semi-supervised manner. It exploits task models to estimate the uncertainty of unlabeled samples to query more valuable ones (Zhu and Ma 2012). The effectiveness of active learning is influenced by the performance of task models. Considering the relevance of different tasks, some researchers have proposed multi-task active learning for linguistic annotations (Reichart et al. 2008; Ikhwantri et al. 2018). The model in (Ikhwantri et al. 2018) exploits the encoder-decoder framework with soft-shared parameters. The encoder is trained to learn the shared features beneficial to all tasks. The decoder extracts task-specific features for predicting sequence labels.

However, the existing multi-task active learning models do not take the influence of task-specific features and the diversity constraint into account. Firstly, the performance of encoder-decoder framework is influenced by the task-specific features (Liu, Qiu, and Huang 2017). The models utilize task-specific features to estimate uncertainty degree for selecting the most uncertain unlabeled samples for labeling. As shown in Figure 1(B), the task-specific features mix up in conventional active learning models which influences the performance of NER and NEN task models. Without effectively fitting on labeled data, the conventional model se-

*Corresponding author.

lects the unlabeled samples x_1, x_2 close to the labeled ones. And the combination of x_1, x_2 and labeled data does not satisfy the diversity well. Therefore, they are not the optimal selections in the respective task models to reduce the expected risk. With the expectations for regular feature space in the adversarial active learning, x_3, x_4 are added to the labeled pool for generalizing task models. Because x_3, x_4 are better coverage of the dataset shown in Figure 1(B). They are the more valuable samples than x_1, x_2 to improve the performance of models.

Secondly, the existing models do not take the diversity of query samples into account. Without considering diversity as an explicit target, the active learning model cannot acquire valuable unlabeled samples for labeling effectively (Deng et al. 2018). We evaluate the diversity of NER and NEN by the words and entities contained in query samples. And the original methods lack a quantitative analysis of the diversity of query samples.

To solve the above problems in multi-task active learning, we propose a **M**ulti-**T**ask **A**dversarial **A**ctive **L**earning (MTAAL[1]) model based on task and diversity adversarial learning. The contributions of this paper can be summarized as follows:

1. We propose the MTAAL consisting of the task discriminator and diversity discriminator. The former eliminates the influence of irregular task-specific features. And the latter exploits the heterogeneity between samples to meet the diversity constraint.
2. We evaluate the MTAAL model on the two common medical NER and NEN benchmarks. The performance of active learning models and the quantitative analysis of diversity on query samples demonstrate the effectiveness of our model against the existing methods.

Related Work

Medical Named Entity Recognition and Normalization. A few existing studies used pipeline models to implement NER and NEN separately (Sahu and Anand 2016; Lou et al. 2017; Vázquez, Chagoyen, and Pascual-Montano 2008). Entity mentions are extracted from medical texts by a named entity recognition model firstly and then a named entity normalization model maps these mentions to standard medical identifiers. Due to the error propagation in the pipeline models, some researchers have proposed to jointly model NER and NEN. An ensemble model that contains two traditional machine learning models is developed as a named entity recognizer and normalizer (Leaman, Wei, and Lu 2015). With the development of deep learning, RNN modules are used in the multi-task sequence learning models. Zhao et al. (2019) proposed a feedback strategy for jointly modeling medical NER and NEN and gained the state-of-the-art results in supervised learning. Our model tries to complete medical NER and NEN with the premise of reducing labeling work.

Recurrent Models for Sequence Labeling. As a basic task in natural language processing (NLP), sequence labeling was explored extensively. NER and NEN can be formulated as sequence labeling problems. With the development of deep learning, traditional CRF based models (Xu et al. 2008) have been combined with RNN modules (Hochreiter and Schmidhuber 1997) to become a classic model on NER (Lample et al. 2016). The relevance between NLP tasks has been used in social media NER (Aguilar et al. 2017). To exploit the shared information between tasks, an adversarial transfer learning algorithm was proposed for Chinese NER (Cao et al. 2018) with the self-attention mechanism (Vaswani et al. 2017).

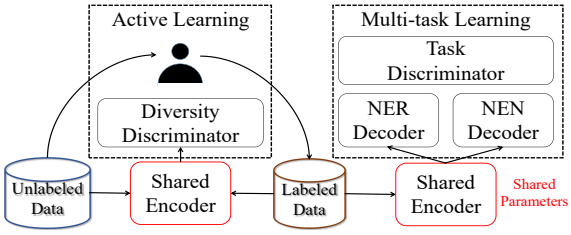

Figure 2: The multi-task adversarial active learning (MTAAL) model for NER and NEN on the medical text. It mainly contains four components: the shared encoder, the task private decoders, the task discriminator ,and the diversity discriminator.

Active Learning. Active learning is a semi-supervised algorithm that estimates unlabeled samples to query valuable ones. It is an effective way to reduce labeling work. Shen et al. (2018) firstly combined deep neural networks with active learning for NER. To adapt active learning in a multi-task scenario, Reichart et al. (2008) proposed a multi-task active learning framework for linguistic annotations with CRF-based models. With the advantages of the RNN model, a multi-task active learning framework was put forward for neural semantic role labeling on low resource conversational corpus (Ikhwantri et al. 2018). The above methods are uncertainty-based active learning without demanding diversity explicitly. Deng et al. (2018) proposed adversarial active learning with a diversity target in a single task scenario. The existing multi-task active learning models do not consider the influence of task-specific features and diversity constraint. Therefore, we introduce task and diversity discriminators into a normal model for solving the above problems.

Model

Before getting into the details of our model, we introduce some notations about NER and NEN in the active learning scenario. We denote $\left(x^L, y^L_{\text{NER}}, y^L_{\text{NEN}}\right) \sim \left(X^L, Y^L_{\text{NER}}, Y^L_{\text{NEN}}\right)$ as a tuple of labeled sample, where x^L is the medical text, y^L_{NER} and y^L_{NEN} the labels of NER and NEN respectively. Given a sentence with N_w words, $y^L_{\text{NER}} = \{y^{(i)}_{\text{NER}}\}^{N_w}_{i=1}$, $y^L_{\text{NEN}} = \{y^{(i)}_{\text{NEN}}\}^{N_w}_{i=1}$, where $y^{(i)}_{\text{NER}}$ and $y^{(i)}_{\text{NEN}}$ are the NER and NEN labels for the i-th word. In the active learning scenario, $x^U \sim X^U$ is denoted as an unlabeled sample in the pool. We perform an active learning algorithm on the unlabeled

[1]We have released the code of our model at: https://github.com/zhoubaohang/MTAAL.

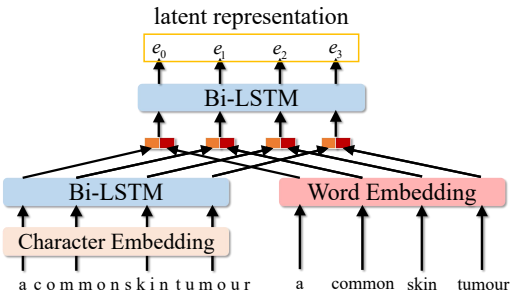

Figure 3: The internal details of the shared encoder. It maps the character- and word-level features of a sentence into the latent representation for prediction and active learning.

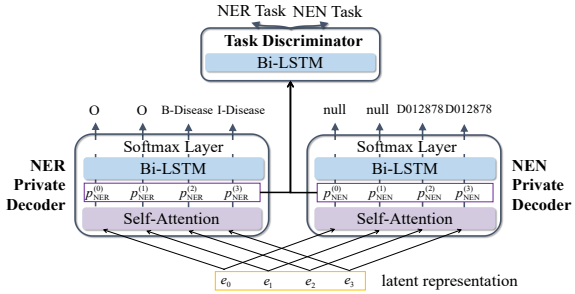

Figure 4: The implementation details of the multi-task learning module. It exploits the latent representation from the shared encoder for predicting task labels. And the task private features are fed into the task discriminator for learning regular feature space.

pool X^U for querying the most effective unlabeled samples $x^S \sim X^S$. The selected samples are manually labeled and then are added to the labeled set $\left(X^L, Y^L_{\text{NER}}, Y^L_{\text{NEN}}\right)$ in the next round.

The MTAAL model is shown in Figure 2. The shared encoder is responsible for extracting latent representations of input sentences. The multi-task learning module contains the task private decoders and discriminator. The task private decoders for NER and NEN are trained to acquire prediction sequence labels of respective tasks. The task discriminator is responsible for learning regular feature space. The active learning module controls the procedure of query unlabeled samples. And the diversity discriminator exploits the heterogeneity between samples to select unlabeled ones for labeling in this module. Each module contains a specific sequence feature extractor. And the whole training procedure is shown in algorithm 1.

Shared Encoder

The shared encoder E is established to map the input sentence x to a latent representation: $e = E(x)$. We consider the character-level features and the word-level features of a sentence as shown in Figure 3. Similar to other neural language models, we firstly map discrete characters and words into the distributed representations. For a given sentence $x = \{x_1, x_2, \ldots, x_{N_w}\}$, the i-th word x_i is represented by a pretrained word embedding vector $w_i \in \mathbb{R}^{d_w}$, where $i = 1, 2, \ldots, N_w$. And for each character in a word, we use a pre-defined char embedding matrix to map it to a vector $c_i \in \mathbb{R}^{d_c}$. After getting the multi-level features of a sentence, we need to extract composite features further. To extract contextual representations, we feed the multi-level features into bidirectional LSTM (BiLSTM) layers respectively. For character-level features of word i, we extract its contextual representations u^c_i from a BiLSTM layer. Then, we get the fusion representation by concatenation $u_i = [w_i; u^c_i]$. Furthermore, the hidden states of the fusion feature extractors are obtains by $e_i = \text{BiLSTM}\,(u_i, e_{i-1}; \theta_w)$, where θ_w is the parameters of the BiLSTM.

Multi-task Learning Module

We adopt the soft-shared parameter schema as other multi-task learning models (Cao et al. 2018; Zhao et al. 2019; Ikhwantri et al. 2018). The shared encoder E is responsible for extracting the shared features of the two tasks. The latent representation from the shared encoder should be transformed into task-specific features that are beneficial for predicting the label sequences. Therefore, we build task private decoders D_{NER} and D_{NEN} for the NER and NEN tasks respectively as shown in Figure 4. Each decoder is composed of the self-attention layer (Vaswani et al. 2017) and the BiLSTM layer.

The tasks $k \in \{\text{NER}, \text{NEN}\}$ focuses on the specific inner structure information of the same sentence because of their label space. To explicitly learn the relationship between two words in a sentence, we apply self-attention to the latent representations extracted from the shared encoder. The self-attention mechanism can be formulated as: $\text{Attention}(\text{Q}, \text{K}, \text{V}) = \text{softmax}(\frac{\text{QK}^T}{\sqrt{\text{d}}})\text{V}$. The task private features could be expressed as: $p_k = \text{Attention}(\text{Q}, \text{K}, \text{V}; \theta_k)$ where θ_k are the parameters of the respective decoders. To get the prediction results, we also exploit BiLSTM to map the task private features $p_{\text{NER}} = \{p^{(i)}_{\text{NER}}\}^{N_w}_{i=1}$ and $p_{\text{NEN}} = \{p^{(i)}_{\text{NEN}}\}^{N_w}_{i=1}$ to label sequences. The hidden states of BiLSTM layer in this module can be computed as: $h^{(i)}_k = \text{BiLSTM}(p^{(i)}_k, h^{(i-1)}_k; \theta_k)$. Furthermore, we apply the softmax layer stacked on the BiLSTM layer. For each token, we can compute the prediction probabilities as: $\hat{y}^{(i)}_k = \text{softmax}(\text{W}_k h^{(i)}_k + \text{b}_k)$ where W_k, b_k are trainable parameters. $\hat{y}_{\text{NER}} = \{\hat{y}^{(i)}_{\text{NER}}\}^{N_w}_{i=1}$ and $\hat{y}_{\text{NEN}} = \{\hat{y}^{(i)}_{\text{NEN}}\}^{N_w}_{i=1}$ denote the predicted probabilities of respective task labels.

For training, we exploit the cross-entropy loss as the objective function. Given M training examples $\left(x^L, y^L_{\text{NER}}, y^L_{\text{NEN}}\right)$, the single-task loss functions can be defined as: $\mathcal{L}_{\text{NER}} = -\sum_M y^L_{\text{NER}} \log(\hat{y}^L_{\text{NER}})$ and $\mathcal{L}_{\text{NEN}} = -\sum_M y^L_{\text{NEN}} \log(\hat{y}^L_{\text{NEN}})$. To train our multi-task model, we sum the above single-task loss functions up as our loss function. The multi-task loss can be formulated as follows:

$$\mathcal{L}_{\text{Task}} = \mathcal{L}_{\text{NER}} + \mathcal{L}_{\text{NEN}}. \quad (1)$$

In conventional multi-task active learning, the performance of task models is influenced by the task-specific

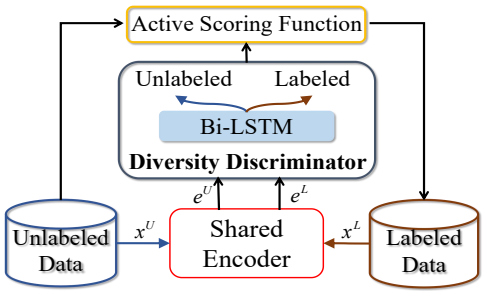

Figure 5: The internal details of the active learning module. The diversity discriminator is trained to select the unlabeled samples that are the least similar to the labeled ones. Our model explicitly targets diversity over conventional active learning models.

features. Because the shared features can exist in the task private space and the task-specific features creep into the shared space (Liu, Qiu, and Huang 2017). Adversarial learning can be applied to multi-task active learning that the task models should learn a regular feature space as shown in Figure 1. To address this problem, we introduce the task discriminator (TD) into our model. We exploit the min-max game between the shared encoder and the task private decoder with task discriminator (Goodfellow et al. 2014). TD is composed of a BiLSTM layer stacked with a softmax layer to estimate what kind of tasks the task private feature p_k comes from. It is used to map the task private representations into a probability distribution. Besides, we need a task adversarial loss $\mathcal{L}_{\text{Task}}^{Adv}$ to limit the task-specific features into their private feature space. The task-based adversarial loss can be defined as follows:

$$\mathcal{L}_{\text{Task}}^{Adv} = \min_{\theta_E} \max_{\theta_k, \theta_{TD}} \sum_k d_k \log(TD(p_k)), \tag{2}$$

where d_k denotes the ground-truth label indicating the task type where the private features come from.

Active Learning Module

In conventional multi-task active learning, the diversity of query samples is not analyzed as a target. And the uncertainty-based methods are strongly correlated to the training of task models (Culotta and McCallum 2005). Although the diversity adversarial learning was exploited in the single task scenario (Deng et al. 2018), the multi-task active learning still has not taken the diversity into account. To overcome this, we further develop it to the multi-task scenario which is beneficial to medical NER and NEN. We introduce the diversity discriminator (DD) into our model to estimate whether the input sentence x is from the labeled or the unlabeled as shown in Figure 5.

The diversity discriminator is also composed of a BiLSTM layer stacked with a softmax layer. It is responsible to map the latent representation e to the probability distribution, estimating which data set the input sentence comes from. Given the input sentence x^L and x^U, we firstly compute their latent representations $e^L = E(x^L)$ and $e^U = E(x^U)$. The discriminator is encouraged to assign e^L to the labeled ($DD(e^L) = 1$) and e^U to the unlabeled ($DD(e^U) = 0$). The objective function can be defined as follows:

$$\begin{aligned}\mathcal{L}_{\text{DD}}^{Adv} = &- \mathbb{E}_{x^L \sim X^L} \log(DD(E(x^L))) \\ &- \mathbb{E}_{x^U \sim X^U} \log(1 - DD(E(x^U))).\end{aligned} \tag{3}$$

To learn the fine-grained representations, the shared encoder is trained to fool the discriminator to regard all latent representations ($\mathbf{e}^L$ and $\mathbf{e}^U$) as labeled ones. We can define the corresponding loss as follows:

$$\begin{aligned}\mathcal{L}_{\text{E}}^{Adv} = &- \mathbb{E}_{x^L \sim X^L} \log(DD(E(x^L))) \\ &- \mathbb{E}_{x^U \sim X^U} \log(DD(E(x^U))).\end{aligned} \tag{4}$$

After training the shared encoder and diversity discriminator, the latter can estimate the unlabeled samples that are the least similar to the labeled ones. Then, we define the active scoring function for selecting the unlabeled samples.

$$\begin{aligned}\psi_{\text{Diversity}}(x^U) &= 1 - DD(E(x^U)) \in (0, 1), \\ x^S &= \max_{x^U \sim X^U} \psi_{\text{Diversity}}(x^U).\end{aligned} \tag{5}$$

Algorithm 1 Multi-Task Adversarial Active Learning

Input: Labeled data X^L with label sequences Y_{NER}^L and Y_{NEN}^L; Unlabeled data X^U;
Initialized parameters: shared encoder θ_{E}, task discriminator θ_{TD}, diversity discriminator θ_{DD}, task private decoders $\{\theta_k | k \in \{\text{NER}, \text{NEN}\}\}$

for $e = 1$ to query-times **do**
  Sample batches $x^L \sim X^L$ and $x^U \sim X^U$
  Minimize $\mathcal{L}_{\text{DD}}^{Adv}$ in (3) to update θ_{DD}
  Minimize $\mathcal{L}_{\text{E}}^{Adv}$ in (4) to update θ_{E}
  Query samples X^S according to (5)
  $(Y_{\text{NER}}^S, Y_{\text{NEN}}^S) \leftarrow \mathcal{ORACLE}(X^S)$
  $(X^L, Y_{\text{NER}}^L, Y_{\text{NEN}}^L) \leftarrow (X^L, Y_{\text{NER}}^L, Y_{\text{NEN}}^L) \cup (X^S, Y_{\text{NER}}^S, Y_{\text{NEN}}^S)$
  $X^U \leftarrow X^U - X^S$
  for mini-batches $x^L \sim X^L$ and $x^U \sim X^U$ **do**
    Minimize $\mathcal{L}_{\text{Task}}^{Adv}$ in (2) to update θ_{E}
    Maximize $\mathcal{L}_{\text{Task}}^{Adv}$ in (2) to update θ_{TD} and $\{\theta_k | k \in \{\text{NER}, \text{NEN}\}\}$
  end for
  for mini-batches in $(X^L, Y_{\text{NER}}^L, Y_{\text{NEN}}^L)$ **do**
    Minimize $\mathcal{L}_{\text{Task}}$ in (1) to update θ_{E}, $\{\theta_k | k \in \{\text{NER}, \text{NEN}\}\}$
  end for
end for

Output: The well trained θ_{E}, θ_{TD}, θ_{DD} and $\{\theta_k | k \in \{\text{NER}, \text{NEN}\}\}$

Experiments

Datasets

We compare our model against the existing methods (Ikhwantri et al. 2018; Shen et al. 2018) on two medical text datasets. The detailed statistics of the two datasets are

Item	NCBI	BC5CDR
vocabulary size	9839	15380
# medical abstracts	793	1500
# entities	7025	28545
# NER labels	3	5
# NEN labels	743	2311
# initial labeled samples	1439	3198
# unlabeled samples	5037	11193
# test samples	719	1599

Table 1: The statistical information of the NCBI dataset and the BC5CDR dataset in our experimental settings.

shown in Table 1. **NCBI** dataset (Dogan, Leaman, and Lu 2014) contains 793 public medical abstracts. There are 7025 disease entities annotated with MeSH/OMIM ids. **BC5CDR** dataset (Li et al. 2016) consists of 1500 public medical abstracts. And 28454 disease and medical entities are manually annotated with MeSH identifiers. We split each abstract into samples in sentence-level according to the ends of sentences. And each sentence contains 40 words averagely. To handle sentences with unequal lengths, we use padding chars to fill them to a fixed length. In the active learning scenario, the whole dataset is partitioned into three parts: the labeled set, the unlabeled set, and the test set. The initial labeled set usually contains a few samples in the real world scenario.

Experiment Settings

The number of hidden units of all the BiLSTM layers is 64. We initialize the character embeddings with a uniform distribution from $\left[-\sqrt{\frac{3}{dim}}, \sqrt{\frac{3}{dim}}\right]$ where dim is set to 30 (Zhao et al. 2019). Considering the influence of different pre-trained word embeddings, we apply the GloVe vectors (Pennington, Socher, and Manning 2014) and the Word2Vec vectors (Mikolov et al. 2013) as comparison respectively. We use the ADAM algorithm (Kingma and Ba 2015) to train the networks and the learning rate of task private decoder is set to 0.001. During every epoch of training, we query for a fixed number of unlabeled samples and add them to the labeled set, which is the same behavior as the existing active models.

We do 70 queries and the model is fine-tuned for 1 epoch after every query (Shen et al. 2018). There are 64 unlabeled samples selected from NCBI and 128 unlabeled ones from BC5CDR at every query respectively. To evaluate the performance of models on NER and NEN, we apply the F1 score as the metric (Zhao et al. 2019). All experiments are accelerated by a single NVIDIA GTX 2080Ti.

Compared Methods

Uncertainty-based multi-task active learning exploits the prediction probabilities from task models. They define different active scoring functions to query unlabeled samples with those probabilities. Therefore, these methods take the current training situation of task models into account in the single- (ST) and multi-task (MT) learning. The uncertainty scores are calculated by active scoring functions with the expectation for the unlabeled samples which are the most beneficial to improve the models.

(1) **Random** (Erdmann et al. 2019): This is an intuitive baseline. And this strategy queries unlabeled samples with the same probability.

(2) **Entropy** (Ikhwantri et al. 2018; Reichart et al. 2008): After getting the prediction probabilities of a sentence, the entropy term is widely used as the uncertainty score: $\psi^{ST}_{\text{Entropy}}(x^U) = -\sum_{N_w} \hat{y}_k \log(\hat{y}_k)$. In the multi-task learning, the outputs from different task models should be taken into account. It is intuitive to sum up the entropy scores from all task models. The entropy score for multi-task learning can be formulated as: $\psi^{MT}_{\text{Entropy}}(x^U) = -\sum_k \sum_{N_w} \hat{y}_k \log(\hat{y}_k)$. We sort the unlabeled samples in ascending order and query the ones with the highest scores.

(3) **Least Confidence** (LC) (Shen et al. 2018): This method is also correlated with the prediction probability of task models. The least confidence score in single-task learning is defined as: $\psi^{ST}_{\text{LC}}(x^U) = 1 - \max_{\hat{y}_k} P(\hat{y}_k|x^U)$. Just like what we did for the entropy score, the least confidence score for multi-task learning can be formulated as: $\psi^{MT}_{\text{LC}}(x^U) = \sum_k 1 - \max_{\hat{y}_k} P(\hat{y}_k|x^U)$. We rank the unlabeled samples according to the above score, then select the samples with the highest scores.

Results

To compare the performances of different multi-task active learning models, we apply the methods to two datasets and record the F1 scores after specific query times. In the single-task (ST) scenario, we complete two tasks with active learning on every dataset respectively. Meanwhile, we apply the baseline models without the task discriminator (TD) to two datasets in the multi-task (MT) scenario.

Comparison with Existing Methods. As shown in Table 2 and 3, it can be observed that the results of two tasks with multi-task active learning are worse than the results in the single-task scenario. Because the task-specific features creep into the shared feature space in the multi-task scenario. We perform experiments on NCBI and BC5CDR for NER and NEN. And the task decoder and the procedure of query samples are influenced by the shared feature space. However, our model can guarantee that the task-specific features do not interfere with respective task decoders. And the diversity constraint of query samples is also included by our modules. These advantages are beneficial to the performances of downstream tasks and the query procedure of active learning. Therefore, our model can always gain the best results in all situations. Furthermore, the pre-trained word embeddings have little influence on the results. Although different word embeddings vary in vocabulary sizes, the character-level features make up for the out-of-vocabulary problem (Chiu and Nichols 2016).

Influence of Task-specific Features. We can see that the results of two tasks with multi-task active learning are worse than the results in the single-task scenario shown in Table

Method	ST		MT	
	NER	NEN	NER	NEN
LC+w	0.8408	0.8862	0.8107	0.8869
LC+g	0.8402	0.9066	0.8338	0.9040
Entropy+w	0.8334	0.8864	0.7868	0.8931
Entropy+g	0.8381	0.9008	0.8292	0.9015
Random+w	0.8137	0.8856	0.7651	0.8836
Random+g	0.8188	0.8936	0.8005	0.8923
MTAAL+w	0.8411	0.8873	0.8462	0.9091
MTAAL+g	0.8492	0.9103	**0.8600**	**0.9152**

Table 2: The F1 scores of the models on BC5CDR. "+g" denotes the model using the Glove word embeddings, while "+w" the Word2Vec word embeddings. The reported results of different methods are after 55 queries.

Method	ST		MT	
	NER	NEN	NER	NEN
LC+w	0.7608	0.9158	0.6736	0.9151
LC+g	0.7452	0.9200	0.6752	0.9151
Entropy+w	0.7394	0.9154	0.7257	0.9139
Entropy+g	0.7462	0.9201	0.7291	0.9137
Random+w	0.7099	0.9151	0.6763	0.9138
Random+g	0.7185	0.9194	0.6749	0.9137
MTAAL+w	0.7688	0.9284	0.7682	0.9267
MTAAL+g	0.7542	0.9267	**0.7744**	**0.9287**

Table 3: The F1 scores of the models on NCBI. "+g" denotes the model using the Glove word embeddings, while "+w" the Word2Vec word embeddings. The reported results of different methods are after 25 queries.

2 and 3. The reason for these comparison results is that the task-specific features can creep into the shared feature space. And the query samples are not strongly correlated to the task-specific features in the normal multi-task active learning. Therefore, the results of LC and Entropy strategies are worse than those of Random. To solve the problem in LC, Entropy, and Random, we introduce the task discriminator (TD) into them.

With the help of the task discriminator, we limit the task-specific features into their feature spaces. The MT with TD gains appreciable improvements on both tasks. To intuitively view the influence of the task discriminator, we show the results of the normal multi-task active learning methods based on TD with the increase of query times. In Figure 6, the dotted lines denote the normal multi-task active learning methods. And the full lines represent the original methods based on the task discriminator. It is significant that the methods based on the task discriminator gain better results than the normal ones. When the task-specific features could creep into the shared feature space, the results of LC and Entropy strategies are worse than that of Random. With the advantage of the task discriminator, the LC and Entropy can gain better results than Random.

Further Discussions

Influence of Query Times. To visually view the influence of query times, we show the results of different multi-task active learning methods with the ascending query times in Figure 7. We can see that our model has a rapid increase in the F1 score on the test set compared with the baseline methods. Although the results of NER on the BC5CDR dataset are not very significant, the performance of NEN is appreciable. And our model gains the best performance on the NCBI dataset. Because the complexity of datasets affects the effectiveness of active learning models. Besides, we can also see that the results of the LC strategy on the NCBI dataset are worse than that of Random strategy. In conventional multi-task active learning models, the task-specific features mix up in the shared feature space. And the procedure of query

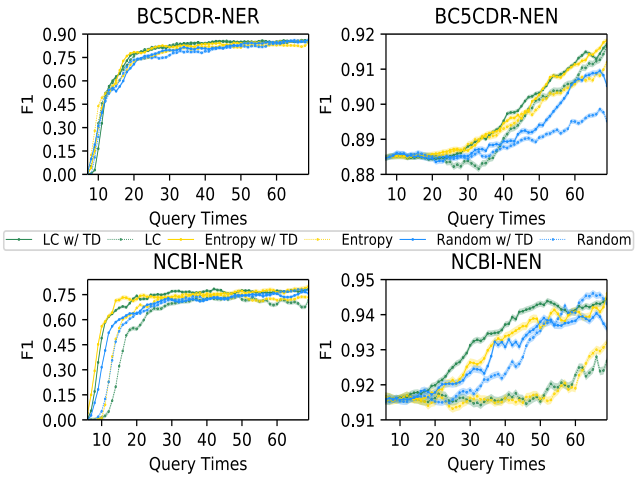

Figure 6: The influence of the task discriminator (TD) on the original multi-task active learning methods. We can see that the results of the methods with TD are better than the normal (w/o TD) ones.

samples is influenced by the task-specific features.

Ablation Study. To investigate the influences of the task discriminator (TD) and the diversity discriminator (DD), we perform an ablation study on MTAAL. During every query, we do not optimize the corresponding modules and record the F1 score on the test set.

The corresponding results are shown in Figure 8. It can be seen that the diversity discriminator is more important than the task discriminator in our model. Because the diversity discriminator is the active learning module in our model. It makes use of the heterogeneity between samples. During every query, it selects the unlabeled samples that are the least similar to the labeled ones. Besides, we can also see that the results of MTAAL are very close to that of the model without TD. Because the diversity discriminator is free of task models. This proves from the side that the task-free active learning methods are not influenced by the task private fea-

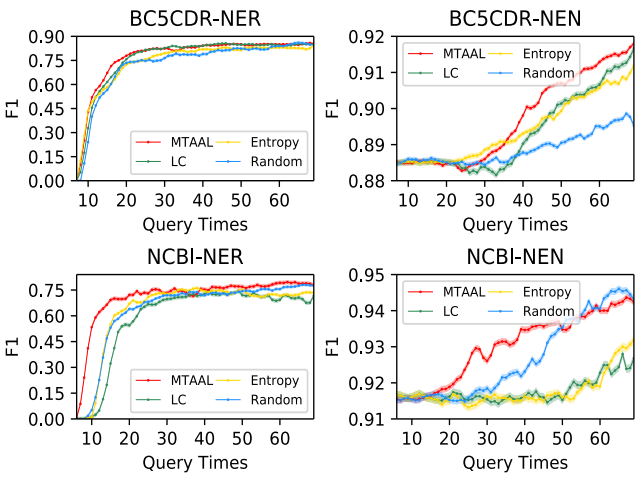

Figure 7: The comparison of different multi-task active learning models with the incremental of query times.

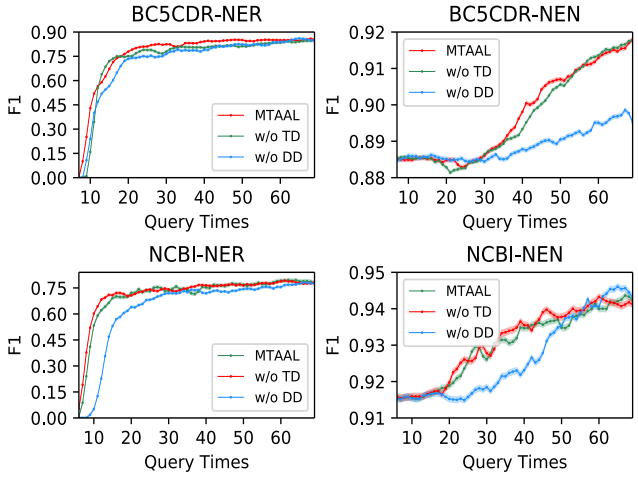

Figure 8: The ablation study of MTAAL. We show the influence of the task discriminator (TD) and the diversity discriminator (DD) to our model respectively.

tures. Although the task discriminator leads to less improvement than the diversity discriminator, decoupling features between tasks also makes sense for MTAAL.

Case Study. The diversity of labeled data means that the trained model can reduce the expected risk effectively (Deng et al. 2018). Therefore, we should analyze the diversity of query samples to intuitively show the effectiveness of different multi-task active learning models. We do statistics on the number of different words, the number of entities and, the percentage of samples with the entity. We select 100 unlabeled samples during every query. After 10 queries, the statistical analysis is conducted on 1000 samples. We repeat the above process for 5 times and calculate the mean and standard deviation of results.

As shown in Table 4, different methods vary in the diversity of query samples. Random strategy performs worst in the diversity of query samples. Because this naive method does not consider the performance improvements that query samples bring to task models. Entropy and LC strategy gain a better diversity of query samples compared with Random one. Although LC and Entropy start from the perspective of training models, the diversity of samples can also be guaranteed. LC and Entropy exploit different ways to handle predicted probabilities for estimating the uncertainty of unlabeled samples. They directly bridge the gap between unlabeled samples with current models. MTAAL exploits the heterogeneity between samples for extending the diversity of the labeled pool. It gains better diverse results in the word- and entity-level.

Dataset	Method	# words	# ent.	% sample w/ ent.
NCBI	LC	3973±16	1528±33	66.23±0.64
	Entropy	3081±32	1295±188	59.22±6.13
	Random	3181±33	915±26	53.16±0.70
	MTAAL	3980±17	1530±21	67.46±0.47
BC5CDR	LC	5454±50	2601±37	90.40±0.37
	Entropy	5020±276	2232±195	80.40±6.70
	Random	4160±55	1761±49	76.98±1.16
	MTAAL	5465±58	2614±44	90.17±0.62

Table 4: The diversity statistical results of query samples on NCBI and BC5CDR. “ent.” is short for “entity”.

Conclusion

In this paper, we propose a multi-task adversarial active learning model and achieve state-of-the-art results on two datasets. With the advantages of the task discriminator and the diversity discriminator, our model can eliminate the influence of irregular task-specific features and select the most valuable unlabeled samples to improve the performance of respective task models. The task-based adversarial learning makes the multi-task active learning model learn the regular feature space. And task models gain better performance on NER and NEN tasks. We also gain appreciable improvement on the compared methods when introducing the task discriminator into them. The diversity-based adversarial learning exploits the heterogeneity between samples for effective procedure of active learning.

In future work, we will design controllable diversity-based multi-task active learning models to avoid selecting unlabeled samples that are abnormal and far away from labeled data. And it is also a good choice to extend the model to more than two tasks.

Acknowledgments

We thank Zhenglu Yang, Ya Guo and Chang Che for their valuable suggestions regarding this manuscript. This work is supported by the Chinese Scientific and Technical Innovation Project 2030 (2018AAA0102100), NSFC-General Technology Joint Fund for Basic Research (No. U1936206, No. U1836109), Joint Funds of NSFC-Xinjiang of China (No. U1903128), National Natural Science Foundation of China (No. 62002178), and Natural Science Foundation of Tianjin, China (No. 20JCQNJC01730).

References

Aguilar, G.; Maharjan, S.; López-Monroy, A. P.; and Solorio, T. 2017. A Multi-task Approach for Named Entity Recognition in Social Media Data. In *Proceedings of the 3rd Workshop on EMNLP*, 148–153.

Cao, P.; Chen, Y.; Liu, K.; Zhao, J.; and Liu, S. 2018. Adversarial Transfer Learning for Chinese Named Entity Recognition with Self-Attention Mechanism. In *Proceedings of the 2018 conference on EMNLP*, 182–192.

Chiu, J. P. C.; and Nichols, E. 2016. Named Entity Recognition with Bidirectional LSTM-CNNs. *TACL* 4: 357–370.

Culotta, A.; and McCallum, A. 2005. Reducing Labeling Effort for Structured Prediction Tasks. In *Proceedings of the 20th AAAI*, 746–751.

Deng, Y.; Chen, K.; Shen, Y.; and Jin, H. 2018. Adversarial Active Learning for Sequences Labeling and Generation. In *Proceedings of the 27th IJCAI*, 4012–4018.

Dogan, R. I.; Leaman, R.; and Lu, Z. 2014. NCBI disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics* 47: 1–10.

Erdmann, A.; Wrisley, D. J.; Allen, B.; Brown, C.; Cohen-Bodénès, S.; Elsner, M.; Feng, Y.; Joseph, B.; Joyeux-Prunel, B.; and de Marneffe, M. 2019. Practical, Efficient, and Customizable Active Learning for Named Entity Recognition in the Digital Humanities. In *Proceedings of the 2019 Conference of NAACL-HLT*, volume 1, 2223–2234.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A. C.; and Bengio, Y. 2014. Generative Adversarial Nets. In *Advances in NIPS 27*, 2672–2680.

Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Computation* 9(8): 1735–1780.

Ikhwantri, F.; Louvan, S.; Kurniawan, K.; Abisena, B.; Rachman, V.; Wicaksono, A. F.; and Mahendra, R. 2018. Multi-Task Active Learning for Neural Semantic Role Labeling on Low Resource Conversational Corpus. In *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*, 43–50. Melbourne: Association for Computational Linguistics. doi:10.18653/v1/W18-3406. URL https://www.aclweb.org/anthology/W18-3406.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *3rd ICLR*.

Lample, G.; Ballesteros, M.; Subramanian, S.; Kawakami, K.; and Dyer, C. 2016. Neural Architectures for Named Entity Recognition. In *Proceedings of the 56th ACL*, 260–270.

Lamurias, A.; and Couto, F. M. 2019. LasigeBioTM at MEDIQA 2019: Biomedical Question Answering using Bidirectional Transformers and Named Entity Recognition. In *Proceedings of the 18th BioNLP Workshop on ACL*, 523–527.

Leaman, R.; Wei, C.; and Lu, Z. 2015. tmChem: a high performance approach for chemical named entity recognition and normalization. *J. Cheminformatics* 7(S-1): S3.

Li, J.; Sun, Y.; Johnson, R. J.; Sciaky, D.; Wei, C.; Leaman, R.; Davis, A. P.; Mattingly, C. J.; Wiegers, T. C.; and Lu, Z. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database* .

Liu, P.; Qiu, X.; and Huang, X. 2017. Adversarial Multi-task Learning for Text Classification. In *Proceedings of the 55th ACL*, 1–10.

Lou, Y.; Zhang, Y.; Qian, T.; Li, F.; Xiong, S.; and Ji, D. 2017. A transition-based joint model for disease named entity recognition and normalization. *Bioinform.* 33(15): 2363–2371.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient Estimation of Word Representations in Vector Space. In *1st ICLR Workshop Track Proceedings*.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on EMNLP*, 1532–1543.

Reichart, R.; Tomanek, K.; Hahn, U.; and Rappoport, A. 2008. Multi-Task Active Learning for Linguistic Annotations. In *Proceedings of the 46th ACL*, 861–869.

Sahu, S.; and Anand, A. 2016. Recurrent neural network models for disease name recognition using domain invariant features. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2216–2225. Berlin, Germany: Association for Computational Linguistics. doi:10.18653/v1/P16-1209. URL https://www.aclweb.org/anthology/P16-1209.

Shen, Y.; Yun, H.; Lipton, Z. C.; Kronrod, Y.; and Anandkumar, A. 2018. Deep Active Learning for Named Entity Recognition. In *6th ICLR*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *Advances in NIPS 30*, 5998–6008.

Vázquez, M.; Chagoyen, M.; and Pascual-Montano, A. D. 2008. Named Entity Recognition and Normalization: A Domain-Specific Language Approach. In *2nd International Workshop on Practical Applications of Computational Biology and Bioinformatics, IWPACBB 2008, Salamanca, Spain, 22th-24th October 2008*, 147–155.

Wu, Y.; Liu, X.; Feng, Y.; Wang, Z.; Yan, R.; and Zhao, D. 2019. Relation-Aware Entity Alignment for Heterogeneous Knowledge Graphs. In *Proceedings of the 28th IJCAI*, 5278–5284.

Xu, Z.; Qian, X.; Zhang, Y.; and Zhou, Y. 2008. CRF-based Hybrid Model for Word Segmentation, NER and even POS Tagging. In *Proceedings of the 3th IJCNLP*, 167–170.

Zhao, S.; Liu, T.; Zhao, S.; and Wang, F. 2019. A Neural Multi-Task Learning Framework to Jointly Model Medical Named Entity Recognition and Normalization. In *Proceedings of the 33th AAAI*, 817–824.

Zhu, J.; and Ma, M. Y. 2012. Uncertainty-based active learning with instability estimation for text classification. *ACM Trans. Speech Lang. Process.* 8(4): 5:1–5:21.