An End-to-End Progressive Multi-Task Learning Framework for Medical Named Entity Recognition and Normalization

Baohang Zhou[1,3], **Xiangrui Cai**[2,3], **Ying Zhang**[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
zhoubh@mail.nankai.edu.cn, {yingzhang,caixr,yuanxj}@nankai.edu.cn

Abstract

Medical named entity recognition (NER) and normalization (NEN) are fundamental for constructing knowledge graphs and building QA systems. Existing implementations for medical NER and NEN are suffered from the error propagation between the two tasks. The mispredicted mentions from NER will directly influence the results of NEN. Therefore, the NER module is the bottleneck of the whole system. Besides, the learnable features for both tasks are beneficial to improving the model performance. To avoid the disadvantages of existing models and exploit the generalized representation across the two tasks, we design an end-to-end progressive multi-task learning model for jointly modeling medical NER and NEN in an effective way. There are three level tasks with progressive difficulty in the framework. The progressive tasks can reduce the error propagation with the incremental task settings which implies the lower level tasks gain the supervised signals other than errors from the higher level tasks to improve their performances. Besides, the context features are exploited to enrich the semantic information of entity mentions extracted by NER. The performance of NEN profits from the enhanced entity mention features. The standard entities from knowledge bases are introduced into the NER module for extracting corresponding entity mentions correctly. The empirical results on two publicly available medical literature datasets demonstrate the superiority of our method over nine typical methods.

1 Introduction

To dig into the large amount of electronic medical records, there has been an increasing interest in applying information extraction to them. These techniques can generate tremendous benefit for corresponding research and applications, such as medical knowledge graph (Wu et al., 2019) and QA systems (Lamurias and Couto, 2019). Among the medical text mining tasks, medical named entity recognition and normalization are the most fundamental tasks.

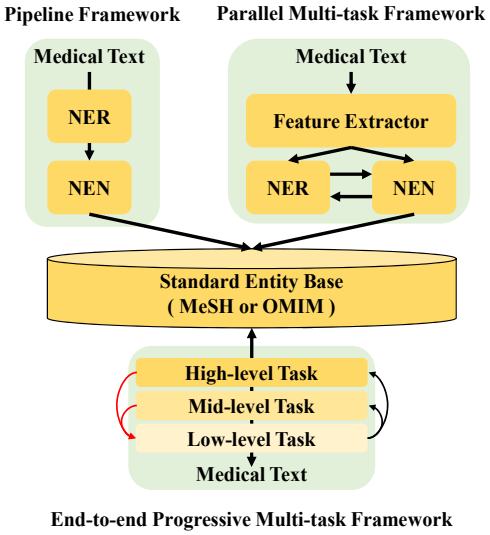

Figure 1: The overall frameworks for medical named entity recognition and normalization.

Named entity recognition tries to find the boundaries of mentions from the medical texts. And named entity normalization maps mentions extracted from the medical text to standard identifiers, such as MeSH and OMIM (Zhao et al., 2019). The initial pipeline implementations for medical NER and NEN have a main limitation: error extractions from NER cascade into NEN which result in normalization errors. Besides, the mutual use between recognition and normalization is not utilized in the pipeline models. To alleviate the limitations and achieve a higher performance, some researchers focused on jointly modeling these two tasks. Leaman and Lu (2016) proposed a joint scoring function for medical NER and NEN. Lou et al. (2017) casted the output construction process of the two tasks

*Corresponding author.

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 6214–6224
August 1–6, 2021.

as a state transition process to perform medical named entity recognition and normalization. To capture the semantic features of two tasks, Zhao et al. (2019) proposed a multi-task learning framework with an explicit feedback strategy for medical NER and NEN.

As shown in Figure 1, there are two common frameworks: pipeline and parallel multi-task framework. The former one is formulated to maximize the posterior probabilities $p(y_{\text{NER}}|x)$ and $p(y_{\text{NEN}}|m, e)$ where x is the medical text, m is the medical mentions extracted by a recognition model, e is the standard entity, y_{NER} and y_{NEN} are the labels. The latter one tries to maximize the posterior probabilities $p(y_{\text{NER}}, y_{\text{NEN}}|x)$ (Zhao et al., 2019). Both of these are struggled with the **bottleneck** that is named entity recognition. In the above frameworks, the NER module is trained to memorize the medial mentions in the training set. However, the medical mentions are various and there is a gap between the training and test set. It is natural that the unseen mentions in training set are hard to recognize during the testing phase. Therefore, the conventional frameworks do not gain more ideal generalization ability.

To overcome the disadvantage mentioned above, we reconsidered the process of medical named entity recognition and normalization. The ultimate goal is to map the extracted medical mentions to the standard entity base. Therefore, the target standard entity base can be regarded as a dictionary. The initial process of NEN and NER can be reconsidered as detecting whether the medical text contains the candidate standard entity and finding the mentions should be replaced. Based on this idea, we propose an **e**nd-to-**e**nd progressive multi-task learning framework for **m**edical named **e**ntity **r**ecognition and **n**ormalization (**E2EMERN**[1]). Compared with ordinary multi-task learning, progressive multi-task learning focuses on the aggregation logic of tasks' specific features (Hong et al., 2020). A difficult target is divided into a few tasks that are interconnected through the combination of features. To take full advantage of the data attributes, we propose the framework including three tasks with progressive difficulty extended from the conventional NER and NEN tasks. The low-level task is the traditional NER which tries to extract all entities in the medical text. The mid-level task is defined to identify whether there exist medical mentions in the text that should be mapped to the candidate standard entity. The high-level task combines the first two level tasks, and targets to extract the mentions which should be mapped to the candidate standard entity.

Unlike the existing frameworks, **E2EMERN** exploits the progressive tasks to learn the fine-grained representations. The mid-level and high-level tasks facilitate the framework learning the corresponding features between the medical mentions and standard entities. The low-level task can gain the supervised signals from the higher level tasks to extract medical mentions corresponded to standard entities in the knowledge bases more exactly. Our contributions in this manuscript can be summarized as follows:

1. We reconsider the process of the NER and NEN tasks, and firstly propose to exploit the three tasks with progressive difficulty to train the end-to-end medical named entity recognition and normalization framework.

2. The experimental results on two medical benchmarks demonstrate that our framework outperforms the existing medical named entity recognition and normalization models. And we conducted detailed analysis on the framework to represent its superiority.

2 Related Work

2.1 Medical Named Entity Recognition and Normalization

Medical named entity recognition and normalization are two basic tasks for the medical text mining. The conventional pipeline frameworks contains the NER model and NEN one separately (Vázquez et al., 2008; Leaman and Lu, 2014; Sahu and Anand, 2016; Zhou et al., 2020). NER models extract medical mentions in texts and then NEN models map these mentions to standard entity identifiers. To reduce the error propagation in the pipeline frameworks, some researchers proposed to model NER and NEN jointly. Leaman et al. (2015) combined two traditional machine learning models as an ensemble NER and NEN model. And to learn the joint probability distribution of the NER and NEN tasks, a semi-markov based model was proposed by Leaman and Lu (2016). However, traditional methods depend on the human-based feature engineering. With the development of the deep

[1]When ready, the code will be published at https://github.com/zhoubaohang/E2EMERN

learning, recurrent neural networks (RNN) have replaced human effort and been utilized to extract features of raw texts. Zhao et al. (2019) designed an RNN-based network architecture with feedback strategy to model the two tasks jointly. Recently, the pre-trained models, such as BERT (Devlin et al., 2019), BioBERT (Lee et al., 2020), make impressive progress in the natural language processing (NLP) area. Xiong et al. (2020) used BERT as the base module and proposed a machine reading comprehension framework to solve the NER and NEN problems jointly.

2.2 Sequence Labeling

Named entity recognition can be regarded as a sequence labeling problem. Sequence labeling was explored extensively as a basic task in NLP. Probabilistic graphical models, such as: hidden markov model (Xiao et al., 2005) and conditional random fields (CRF) (Lafferty et al., 2001) are the typical methods to solve the problem. With deep learning modules gradually replacing manual feature engineering, long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) network stacked with CRF (Xu et al., 2008) has been a benchmark model for sequence labeling (Lample et al., 2016). Some researchers utilized multi-task learning to model relevant NLP tasks and gained better performances on these tasks including sequence labeling (Aguilar et al., 2017; Cao et al., 2018). Besides, the attributes of the data themselves are used to design the multi-task learning model. Considering whether sentences contain entities, Wang et al. (2019) proposed the multi-task learning model to predict whether input data have entities and then extract corresponding entities. Kruengkrai et al. (2020) exploited sentence-level labels and token-level labels to propose a joint model supporting multi-class classification.

2.3 Short Text Matching

Named entity normalization is formulated as a short text matching problem. The information retrieval method, such as: BM25 (Robertson et al., 1994), is a universal model to solve this problem. With the development of neural language model, text semantic is exploited to model the similarity between two short texts. The distributed representations of texts, such as: Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), are utilized to calculate the similarity distance between two texts. Some medical named entity normalization models are based on this method (Leaman and Lu, 2014; Zhou et al., 2020). Considering local texts are more important than global ones, some researchers utilized convolution neural networks (CNN) to extract local features and exploited interactive attention mechanism to match the semantic similarity of two texts (Yin et al., 2016; Chen et al., 2018).

3 Methodology

We introduce the notations about NER and NEN before getting into the details of the framework. For NER task, we denote $\{(\mathbf{X}_i, \mathbf{y}_i)\}_{i=1}^{N^s}$ as a training set with N^s samples, where $\mathbf{X}_i$ is the medical text and $\mathbf{y}_i$ is the NER label. Given a sentence with N^w words, the medical text can be formulated as $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{N^w}\}$ and the NER label is $\mathbf{y} = \{y_1, y_2, \ldots, y_{N^w}\}$. To solve the NER task, we try to maximize the posterior probability $p(\mathbf{y}\,|\mathbf{X})$. According to the NER label, we can extract the medical mentions $\{\mathbf{m}_i\}_{i=1}^{N^m}$ from the medical text, where N^m is the number of the mentions. For NEN task, we need to map each mention $\mathbf{m}$ to a standard entity $\mathbf{e}$ in the entity base $\mathbf{B} = \{\mathbf{e}_i\}_{i=1}^{N^e}$. We formulate the object of NEN task as a posterior probability $p(\mathbf{e}\,|\mathbf{m}, \mathbf{B})$, and $\mathbf{e}$ is the standard entity which the mention $\mathbf{m}$ should be mapped to.

3.1 Progressive Tasks

With the help of NER and NEN, we can map medical mentions in the raw texts to the corresponding standard entities. Traditional pipeline implementations for the two tasks are composed of the individual NER and NEN models. The simple partitioning of the two models leads to the error propagation between them. Considering the correlation between the two tasks, Zhao et al. (2019) proposed the parallel task framework to improve the performance of the model. However, the intuitive feedback strategy for the output layers of two tasks is not beneficial to modeling the fine-grained features between two tasks. The above implementations lack thinking about the learning process. The process of human learning often goes from easy to difficult (Xu et al., 2020). Especially for the correlated tasks, humans can dig into the hidden knowledge and extract them from the easy tasks for completing the hard ones. Based on this idea, we reconsider the process of conventional NER and NEN tasks, and propose three correlated tasks with progressive difficulty. As shown in Figure 2, we take a medical text from the real dataset NCBI (Dogan et al., 2014)

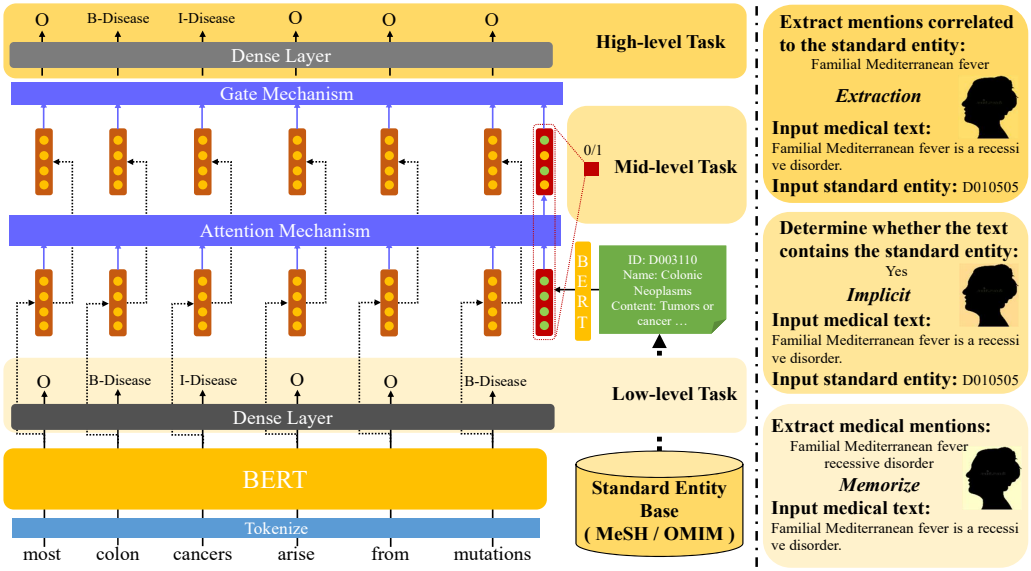

Figure 2: The end-to-end progressive multi-task learning framework for medical named entity recognition and normalization. The left part is the implementation details of the framework. The right part is the real example to describe the three progressive tasks.

as an example to describe the tasks. The medical text is "Familial Mediterranean fever is a recessive disorder" and its corresponding NER label is "B-Disease I-Disease I-Disease O O B-Disease I-Disease". Among the tokens, medical mentions "Familial Mediterranean fever" and "recessive disorder" are mapped to the standard entity identifiers "D010505" and "D030342" respectively.

Low-level task is defined to memorize all medical mentions seen in training set. Given the medical text mentioned above, this task needs to predict the NER label and extract the mentions "Familial Mediterranean fever" and "recessive disorder". Similar to the process of human learning vocabulary, the low-level task forces the framework to learn the medical mentions indiscriminantly. However, the final target is to map mentions to standard entities. We should continue to bridge the gap between medical mentions in raw texts and standard entities in the database.

Mid-level task targets to determine whether medical texts implicit the query standard entities. With the above medical text and the standard entity "D010505" as input, this task should inference the text contains this entity. Through this task, the framework establishes the coarse-grained relationship between the mentions with contexts and the query standard entities. However, the mentions are incomplete correspondence to the query standard entities. Because there is more than one mention in the raw text which should be extracted and mapped to the corresponding standard entities. We need to specify which mention in the text should be mapped to the input standard entity.

High-level task is proposed to extract the mentions which should be mapped to the query standard entity. After acquiring the above medical text and the standard entity "D030342", this task should extract the mention "recessive disorder". If the input text contains no mention which should be mapped to the query entity, the output of this task is empty. The effect of this task is the same as that of NEN, but it is harder than NEN. To accomplish the high-level task, we need to build on the first two tasks. The low-level task provides the representations of the medical mentions with contexts which is beneficial to locating them in raw texts. The mid-level task forces the model to learn the correlated features between mentions with standard entities. With the help of two pre-tasks, the high-level task can be accomplished in an effective way.

3.2 Implementation Details

We build on the progressive tasks to implement the framework E2EMERN as shown in Figure 2. Considering the logic of feature aggregation and the strategies for training different tasks, we need to give detailed explanations by the level of tasks.

For a given sentence $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{N^w}\}$, we need to map it to the dense vector representations. With the impressive performances of pre-trained models, we utilize BERT (Devlin et al.,

2019) as feature extractors to acquire the distributed representations of sentences. The BERT architecture is composed by the transformer networks and its weights are trained with large number of corpus. The feature extraction process is simplified as $\text{BERT}(\mathbf{X}) = \{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_{N^w}\}$, where $\mathbf{h} \in \mathbb{R}^{1024\times 1}$. The low-level task is defined as the same as NER, and we utilize the NER labels as the target. The sentence features $\{\mathbf{h}_i\}_{i=1}^{N^w}$ are fed into the softmax layer, and we can compute the prediction probabilities of low-level task as: $\hat{y}_i = \text{softmax}(\text{W}_l\mathbf{h}_i + \text{b}_l)$ where W_l and b_l are trainable parameters. For training, we utilize the cross-entropy loss as the objective function. The loss function of low-level task is defined as follows:

$$\mathcal{L}_{low} = -\sum_{i=1}^{N^w} y_i \log \hat{y}_i. \quad (1)$$

The sample for the mid-level task is defined as a tuple $(\mathbf{X}, \mathbf{e}, y^m)$. If the text $\mathbf{X}$ contain the mentions which should be mapped to the entity $\mathbf{e}$, y^m is assigned 1 otherwise 0. To bridge the gap between the mentions and standard entities in the mid-level task, we need also to extract the features of standard entities. The standard entity $\mathbf{e}$ is described with the specific name and some medical contents. We feed the name (or contents) of the entity into the BERT and perform the average pooling on the output of BERT. The feature vector of i-th standard entity in the database is defined as $\mathbf{h}_i^e$. Considering the words of mentions in raw texts are more correlated to the standard entity, we adopt the attention mechanism (Zhou et al., 2016) to focus on the local words of sentences. The attention weighted average feature can be calculated as: $\mathbf{h}^a = \sum_{i=1}^{N^w} \alpha_i \mathbf{x}_i$. And the attention score α is defined as: $\alpha_i = \frac{\exp(s(\mathbf{x}_i, \mathbf{h}^e))}{\sum_{i=1}^{N^w} \exp(s(\mathbf{x}_i, \mathbf{h}^e))}$ where $s(\mathbf{x}_i, \mathbf{h}^e) = \text{W}_a[\mathbf{x}_i; \mathbf{h}^e] + \text{b}_a$. W_a and b_a are trainable weights in the attention module. After acquiring the entity-attention feature $\mathbf{h}^a$ and standard entity feature $\mathbf{h}^e$, we can calculate the prediction probabilities $\hat{y}^m = \sigma(\text{W}_m[\mathbf{h}^e; \mathbf{h}^a] + \text{b}_m)$ where σ is the sigmoid function. The loss function for the mid-level task is formulated as the cross-entropy:

$$\mathcal{L}_{mid} = -(y^m \log \hat{y}^m + (1 - y^m) \log(1 - \hat{y}^m)). \quad (2)$$

We define the tuple $(\mathbf{X}, \mathbf{e}, \mathbf{y}^h)$ as the sample for the high-level task where $\mathbf{y}^h = \{y_i^h\}_{i=1}^{N^w}$. Given that the medical text $\mathbf{X}$ is "Familial Mediterranean fever is a recessive disorder." and standard entity $\mathbf{e}$ is "D030342", the label sequence $\mathbf{y}^h$ should be "O O O O O B-Disease I-Disease". To take advantage of the pre-tasks, we propose the gate mechanism to aggregate the different features for solving this task. The sentence feature $\{\mathbf{h}_i\}_{i=1}^{N^w}$ implicit the medical mentions while the entity attention feature $\mathbf{h}^a$ contains clearer locations of the corresponding mentions. Therefore, we propose the gate mechanism to focus on the fine-grained feature dimensions. The formulation of the gate mechanism is $\text{G}(\mathbf{H}, \mathbf{H}^a) = \sigma(\text{W}_g[\mathbf{H}; \mathbf{H}^a] + \text{b}_g)$ where $\mathbf{H} = \{\mathbf{h}_i\}_{i=1}^{N^w}$ and $\mathbf{H}^a = [\mathbf{h}^a; \ldots; \mathbf{h}^a] \in \mathbb{R}^{1024\times N^w}$. Considering the semantic difference between the mentions and corresponding standard entities, we exploit the gate mechanism to fuse the standard entity feature with the sentence feature. The fusion sentence feature is formulated as: $\mathbf{H}^f = \mathbf{H} \odot (1 - \text{G}(\mathbf{H}, \mathbf{H}^a)) + \mathbf{H}^e \odot \text{G}(\mathbf{H}, \mathbf{H}^a)$ where $\odot$ is the element-wise production, $\mathbf{H}^f = \{\mathbf{h}_i^f\}_{i=1}^{N^w}$ and $\mathbf{H}^e = [\mathbf{h}^e; \ldots; \mathbf{h}^e] \in \mathbb{R}^{1024\times N^w}$. We feed the fusion feature into the softmax layer to predict the probabilities $\hat{y}_i^h = \text{softmax}(\text{W}_h\mathbf{h}_i^f + \text{b}_h)$. As the same as the low-level task, we utilize the cross-entropy loss function as follows:

$$\mathcal{L}_{high} = -\sum_{i=1}^{N^w} y_i^h \log \hat{y}_i^h. \quad (3)$$

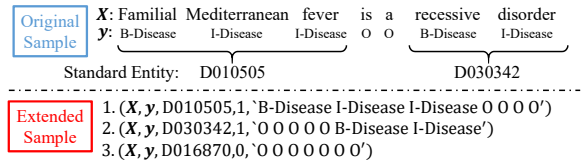

Figure 3: The original sample is from the dataset NCBI. The extended samples are built on the original one and used to train the model. The 3rd sample is generated by negative sampling.

3.3 Training Process

For the framework, we denote the training sample as $(\mathbf{X}, \mathbf{y}, \mathbf{e}, y^m, \mathbf{y}^h)$. According to the definitions of the three tasks, we can generate the task labels corresponding to the input sentence. The example is shown in Figure 3. Given the medical text $\mathbf{X}$, the label $\mathbf{y}$ for the low-level task is the same as the original NER label. We use the standard entities which the mentions $\{\mathbf{m}_i\}_{i=1}^{N^m}$ should be mapped to as the input entity $\mathbf{e}$ respectively. The high-level task label $\mathbf{y}^h$ is based on $\mathbf{y}$, and it only keeps the original labels of $\mathbf{y}$ which are correlated to the input $\mathbf{e}$. Besides, we adopt the negative sampling strategy

to select the standard entity which is not related to the input sentence **X** as the input entity **e**.

To tackle the three level tasks at once, we introduce two hyper-parameters to sum Eqn. 1, Eqn. 2 and Eqn. 3. The overall loss function for the framework is defined as follows:

$$\mathcal{L} = \mathcal{L}_{low} + \lambda \cdot \mathcal{L}_{mid} + \mu \cdot \mathcal{L}_{high} \quad (4)$$

where λ and μ are hyper-parameters for balancing different task losses. After generating samples, we feed them into the model and then calculate the loss according to Eqn. 4. Following the back-propagation method, we update the weights of the networks with the acquired loss. After every epoch of training, we re-sample the training samples for better generalization of the model.

4 Experiments

4.1 Datasets and Experiment Settings

We compare our framework with the existing methods on two medical benchmark datasets. Table 1 presents the detailed statistical information of the two datasets. There are 798 public medical abstracts in the **NCBI** dataset (Dogan et al., 2014). Each medical mention in the text is annotated with MeSH/OMIM identifiers. **BC5CDR** dataset (Li et al., 2016) contains 1500 public medical abstracts which are also annotated with MeSH identifiers. We split each abstract into sentence samples with an average of 40 words according to the ends of sentences. The padding char is used for filling the unequal length samples to the fixed length.

During the training process, we first train the model on the training set and test it on the development set for searching the best hyper-parameters. Then, we fix the best hyper-parameters and train the model on the set composed of the training and development sets. Before the model is trained to the searched maximum number of epochs, we take the F1 score as the reported result when the loss gets the lowest. In our experiments, we set the hyper-parameters λ, μ and learning rate to 0.125, 0.1 and 1e-5 respectively. To train the model, we use the ADAM (Kingma and Ba, 2015) algorithm to update the weights. And all experiments are accelerated by the two NVIDIA GTX 2080Ti devices.

4.2 Compared Methods

To represent the effectiveness of our framework, we adopt the competitive models as the compared methods including traditional machine learning methods and impressive deep learning models.

Item	NCBI	BC5CDR
train set	5424	4560
dev set	923	4581
test set	940	4797
# entities	7025	28545
# NER labels	3	5
# NEN labels	743	2311

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

Dnorm (Leaman et al., 2013) is the pipeline model for medical NER and NEN. It utilizes the TF-IDF feature to learn the bilinear mapping matrix for the normalization task. **LeadMine** (Lowe et al., 2015) considers Wikipedia as dictionary features for normalizing the medical mentions. **TaggerOne** (Leaman and Lu, 2016) is the semi-Markov based model for jointly modeling medical NER and NEN. **Transition-based model** (Lou et al., 2017) consists of the state transformation function for the output of NER and NEN.

To reduce human feature engineering, researchers focus on the deep learning for modeling NER and NEN. **IDCNN** (Strubell et al., 2017) was proposed with an improved CNN module for NER. **MCNN** (Zhao et al., 2017) was composed of the multiple-label CNN modules for better performances on NER. **CollaboNet** (Yoon et al., 2019) exploited the multi-source datasets for training the multi-task model and gained better results on all benchmark datasets. **MTL-MERN** (Zhao et al., 2019) consists of the NER and NEN parallel framework and utilizes the feedback strategy to improve the performances on two tasks.

With the impressive performance of pre-trained models, **BioBERT** (Lee et al., 2020) is built on the BERT (Devlin et al., 2019) and trained with a large medical corpus. And it achieves state-of-the-art results on medical NER datasets. Therefore, we use the **BioBERT** as the feature extractor and compare it with our framework.

4.3 Experimental Results

We compare E2EMERN with the baseline methods on the named entity recognition and normalization. The detailed experiment results on NCBI and BC5CDR are shown in Table 2. The first

Method	NCBI		BC5CDR	
	Recognition	Normalization	Recognition	Normalization
Dnorm (Leaman et al., 2013)	0.7980	0.7820	-	0.8064
LeadMine (Lowe et al., 2015)	-	-	-	0.8612
TaggerOne (Leaman and Lu, 2016)	0.8290	0.8070	0.8260	0.8370
Transition-based Model (Lou et al., 2017)	0.8205	0.8262	0.8382	0.8562
IDCNN (Strubell et al., 2017)	0.7983	0.7425	0.8011	0.8107
MCNN (Zhao et al., 2017)	0.8517	-	0.8783	-
CollaboNet (Yoon et al., 2019)	0.8636	-	0.8818	-
MTL-MERN (Zhao et al., 2019)	0.8743	0.8823	0.8763	0.8645
BioBERT (Lee et al., 2020)	0.8971	-	0.9029	-
E2EMERN	**0.9151**	**0.8901**	**0.9175**	**0.8965**
w/o mid-level task	0.8733	0.8890	0.9073	0.8600
w/o high-level task	0.8862	-	0.9065	-
w/o gate mechanism	0.8885	0.8224	0.9100	0.8681
w/o attention mechanism	0.8767	0.8675	0.9092	0.8676

Table 2: The F1 scores of the models on NCBI and BC5CDR. Our model can outperform the baseline methods. The results for ablation study of E2EMERN are also presented. "w/o gate mechanism" means that the gate mechanism is replaced with the simple feature concatenation strategy in the framework. "w/o attention mechanism" is the same as the above one.

four in the table is the traditional machine learning methods. Among them, the joint models, such as TaggerOne and Transition-based Model, outperform the pipeline ones including Dnorm and LeadMine. When deep learning was introduced into the pipeline frameworks, IDCNN can make a progress over conventional methods, such as Dnorm. Compared with MCNN, CollaboNet utilizes the multi-source dataset as input and performs multi-task learning to improve the performances on NER task. MTL-MERN takes full advantage of multi-task learning and deep semantic representations and outperforms the above methods. By virtue of the dynamic language features, BioBERT can better model the language semantics and outperform the above NER models.

Compared with baseline methods, E2EMERN can always achieve the best results on NER and NEN. The NER results of E2EMERN increase by $1\% \sim 2\%$ over BioBERT. Because our framework takes full advantage of the correlation between NER and NEN. Unlike the simple strategy of MTL-MERN, E2EMERN consists of three progressive tasks that are well-designed for modeling the fine-grained features between medical mentions in raw texts and standard entities. The standard entity information of NEN is introduced into the NER module by the mechanisms in our framework. With the help of the dynamic language features and progressive multi-task learning, the framework can extract the medical mentions more exactly and map them to standard entities. And the semantic correlation between medical mentions and standard entities is built on the three progressive tasks from low to high. The rich semantics captured by the progressive tasks are beneficial to NER and NEN.

4.4 Further Discussion

To dig into the framework, we conduct the detailed analysis for presenting it in different aspects. The ablation study is conducted to present the effectiveness of the mechanisms proposed in the framework. Besides the supervised learning, our framework exploits the standard entity information in the NER task and is potential in a zero-shot scenario compared with BioBERT. We conduct the case study to analyze the prediction results and visualize the attention mechanism to prove its effectiveness.

4.4.1 Ablation Study

As shown in Table 2, we conduct the ablation study to present the effectiveness of the progressive tasks and different mechanisms. When free from completing the mid- or high-level tasks, E2EMERN gains worse results on NER and NEN. The progressive tasks improves the ability of the framework to learn the multi-grained features between original texts and standard entities. Besides, we replace the gate and attention mechanisms with the simple feature concatenation strategy as compared methods. When removed the attention mech-

Text1:	the	von	hippel	-	lindau	tumor	suppressor	gene	is	required	for	cell	cycle	exit	upon	serum	withdrawal	.
Ground Truth:	O	B-Disease	I-Disease	I-Disease	I-Disease	I-Disease	O	O	O	O	O	O	O	O	O	O	O	O
BioBERT:	O	B-Disease	I-Disease	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
E2EMERN: + MeSH:D006623	O	B-Disease	I-Disease	I-Disease	I-Disease	I-Disease	O	O	O	O	O	O	O	O	O	O	O	O

Text2:	genotype	-	phenotype	analyses	in	cowden	disease	and	bannayan	-	zonana	syndrome	,	two	hamartoma	syndromes	with	germline	pten	mutation	.
Ground Truth:	O	O	O	O	O	B-Disease	I-Disease	O	B-Disease	I-Disease	I-Disease	I-Disease	O	O	B-Disease	I-Disease	O	O	O	O	O
BioBERT:	O	O	O	O	O	B-Disease	I-Disease	O	O	O	O	O	O	O	O	O	O	O	O	O	O
E2EMERN: + MeSH:D006223	O	O	O	O	O	B-Disease	I-Disease	O	B-Disease	I-Disease	I-Disease	I-Disease	O	O	B-Disease	I-Disease	O	O	O	O	O

Text3:	reasons	for	seizures	were	ruled	out	and	the	convulsions	stopped	few	hours	after	cessation	of	morphine	and	did	not	reoccur	in	8	months	.
Ground Truth:	O	O	B-Disease	O	O	O	O	O	O	O	O	O	O	O	O	B-Chemical	O	O	O	O	O	O	O	O
BioBERT:	O	O	B-Disease	O	O	O	O	O	O	O	O	O	O	O	O	B-Chemical	O	O	O	O	O	O	O	O
E2EMERN: + MeSH:D009020	O	O	B-Disease	O	O	O	O	O	O	O	O	O	O	O	O	B-Chemical	O	O	O	O	O	O	O	O

Text4:	male	sprague	dawley	rats	were	treated	with	betaine	(	100	,	200	,	and	400	mg	/	kg	)	orally	for	40	days	.
Ground Truth:	O	O	O	O	O	O	O	B-Chemical	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
BioBERT:	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
E2EMERN: + MeSH:D001622	O	O	O	O	O	O	O	B-Chemical	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O

Table 3: The case study results on NCBI and BC5CDR. "Text1" and "Text2" are from NCBI, and the other two are from BD5CDR. "Text2" and "Text4" are the unseen samples from the test set of two datasets. The standard entities coupled with each text are the input of E2EMERN.

anism, E2EMERN achieves worse results on two tasks. It proves that the supervised signals from mid-level task are beneficial to the low-task. And the entity-attention feature generated by the mechanism contributes to the high-level task. E2EMERN without the gate mechanism gains the worse results on NEN. Because the mechanism aggregates the features from lower level tasks which provides the multi-grained information between mentions and standard entities. The ablation study proves the importance of the two mechanisms to E2EMERN.

4.4.2 Results on Unseen Samples

We conduct the statistic analysis on the test set of NCBI and BC5CDR. As shown in Figure 4, there are about $40\% \sim 50\%$ samples contain the words or medial mentions which do not appear in the training set. Therefore, we need to evaluate the generalization ability of models on the unseen samples. We compare E2EMERN with BioBERT on the unseen samples in the test set. To a certain extent, our framework can outperform the existing state-of-the-art NER model. Compared with BioBERT, E2EMERN introduces the standard entity base into the framework. The fine-grained location information of medical mentions from the high-level task is propagated to the low-level task. With the help of standard entity information and progressive multi-task learning, E2EMERN can gain the better generalization ability on unseen samples.

4.4.3 Case Study

We present the case study results in Table 3. Compared with BioBERT, our framework can extract the medical mentions which BioBERT can not extract. We draw the label results of E2EMERN with

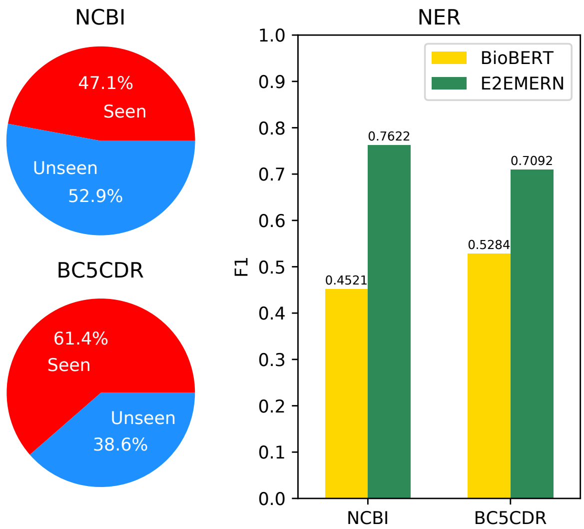

Figure 4: The results on unseen samples. The left part is the proportions of seen and unseen samples in test sets. The unseen samples mean that the words or medical mentions included of them do not appear in the training and development sets. The right part is the NER results of unseen samples in test sets.

the heat map. As the color deepens, the importance of the token in the sentence increases. The visualization results prove that the attention mechanism in E2EMERN focuses on the tokens which make of medical mentions. Although "Text2" and "Text4" are unseen samples, E2EMERN can also extract the mentions in them. The token "convulsions" is paid more attention than "seizures" in "Text3". But convulsion is the symptom of seizures. With the help of medical correlation between them, E2EMERN can extract the token "seizures" as medical mention. To some extent, the effectiveness of E2EMERN can be proved by the case study.

5 Conclusion

In this paper, we reconsider the process of NER and NEN and propose the end-to-end progressive multi-task learning framework for medical named entity recognition and normalization. Compared with existing methods, the framework consists of three tasks with progressive difficulty which contributes to modeling the fine-grained features between medical mentions in raw texts and standard entities. Furthermore, the detailed analysis of E2EMERN proves its effectiveness. Considering the medical area is various, we will try to adapt the framework to the cross domain problem.

Acknowledgments

We would like to thank three anonymous reviewers for their insightful comments. This research is supported by the Chinese Scientific and Technical Innovation Project 2030 (2018AAA0102100), NSFC-General Technology Joint Fund for Basic Research (No. U1936206), NSFC-Xinjiang Joint Fund (No. U1903128), National Natural Science Foundation of China (No. 62002178, No. 62077031), and Natural Science Foundation of Tianjin, China (No. 20JCQNJC01730).

References

Gustavo Aguilar, Suraj Maharjan, Adrian Pastor López-Monroy, and Thamar Solorio. 2017. A multi-task approach for named entity recognition in social media data. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 148–153, Copenhagen, Denmark. Association for Computational Linguistics.

Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, and Shengping Liu. 2018. Adversarial transfer learning for Chinese named entity recognition with self-attention mechanism. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 182–192, Brussels, Belgium. Association for Computational Linguistics.

Haolan Chen, Fred X. Han, Di Niu, Dong Liu, Kunfeng Lai, Chenglin Wu, and Yu Xu. 2018. MIX: multi-channel information crossing for text matching. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 110–119.

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*, pages 4171–4186.

Rezarta Islamaj Dogan, 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.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Yanfei Hong, Benzheng Wei, Zhongyi Han, Xiang Li, Yuanjie Zheng, and Shuo Li. 2020. Mmcl-net: Spinal disease diagnosis in global mode using progressive multi-task joint learning. *Neurocomputing*, 399:307–316.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd ICLR*.

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.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289.

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.

Andre Lamurias and Francisco M. Couto. 2019. Lasigebiotm at MEDIQA 2019: Biomedical question answering using bidirectional transformers and named entity recognition. In *Proceedings of the 18th BioNLP Workshop on ACL*, pages 523–527.

Robert Leaman, Rezarta Islamaj Dogan, and Zhiyong Lu. 2013. Dnorm: disease name normalization with pairwise learning to rank. *Bioinform.*, 29(22):2909–2917.

Robert Leaman and Zhiyong Lu. 2014. Disease named entity recognition and normalization with dnorm. In *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, page 587.

Robert Leaman and Zhiyong Lu. 2016. Taggerone: joint named entity recognition and normalization with semi-markov models. *Bioinform.*, 32(18):2839–2846.

Robert Leaman, Chih-Hsuan Wei, and Zhiyong Lu. 2015. tmchem: a high performance approach for chemical named entity recognition and normalization. *J. Cheminformatics*, 7(S-1):S3.

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. *Bioinform.*, 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*.

Yinxia Lou, Yue Zhang, Tao Qian, Fei Li, Shufeng Xiong, and Donghong Ji. 2017. A transition-based joint model for disease named entity recognition and normalization. *Bioinform.*, 33(15):2363–2371.

Daniel M Lowe, Noel M O'Boyle, and Roger A Sayle. 2015. Leadmine: Disease identification and concept mapping using wikipedia. In *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop*, pages 240–246.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *27th Annual Conference on Neural Information Processing Systems.*, pages 3111–3119.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994*, pages 109–126.

Sunil Sahu and Ashish Anand. 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)*, pages 2216–2225, Berlin, Germany. Association for Computational Linguistics.

Emma Strubell, Patrick Verga, David Belanger, and Andrew McCallum. 2017. Fast and accurate entity recognition with iterated dilated convolutions. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2670–2680, Copenhagen, Denmark. Association for Computational Linguistics.

Miguel Vázquez, Monica Chagoyen, and Alberto D. Pascual-Montano. 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*, pages 147–155.

Yu Wang, Yun Li, Ziye Zhu, Bin Xia, and Zheng Liu. 2019. SC-NER: A sequence-to-sequence model with sentence classification for named entity recognition. In *Advances in Knowledge Discovery and Data Mining - 23rd Pacific-Asia Conference*, pages 198–209.

Yuting Wu, Xiao Liu, Yansong Feng, Zheng Wang, Rui Yan, and Dongyan Zhao. 2019. Relation-aware entity alignment for heterogeneous knowledge graphs. In *Proceedings of the 28th IJCAI*, pages 5278–5284.

JingHui Xiao, Bingquan Liu, and Xiaolong Wang. 2005. Principles of non-stationary hidden markov model and its applications to sequence labeling task. In *Natural Language Processing - IJCNLP 2005, Second International Joint Conference*, pages 827–837.

Ying Xiong, Yuanhang Huang, Qingcai Chen, Xiaolong Wang, Yuan Nic, and Buzhou Tang. 2020. A joint model for medical named entity recognition and normalization. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020)*, pages 499–504.

Benfeng Xu, Licheng Zhang, Zhendong Mao, Quan Wang, Hongtao Xie, and Yongdong Zhang. 2020. Curriculum learning for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6095–6104, Online. Association for Computational Linguistics.

Zhiting Xu, Xian Qian, Yuejie Zhang, and Yaqian Zhou. 2008. Crf-based hybrid model for word segmentation, NER and even POS tagging. In *Third International Joint Conference on Natural Language Processing*, pages 167–170.

Wenpeng Yin, Hinrich Schütze, Bing Xiang, and Bowen Zhou. 2016. ABCNN: Attention-based convolutional neural network for modeling sentence pairs. *Transactions of the Association for Computational Linguistics*, 4:259–272.

Wonjin Yoon, Chan Ho So, Jinhyuk Lee, and Jaewoo Kang. 2019. Collabonet: collaboration of deep neural networks for biomedical named entity recognition. *BMC Bioinform.*, 20-S(10):55–65.

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 33th AAAI*, pages 817–824.

Zhehuan Zhao, Zhihao Yang, Ling Luo, Lei Wang, Yin Zhang, Hongfei Lin, and Jian Wang. 2017. Disease named entity recognition from biomedical literature using a novel convolutional neural network. *BMC Medical Genomics*, 10.

Huiwei Zhou, Shixian Ning, Zhe Liu, Chengkun Lang, Zhuang Liu, and Bizun Lei. 2020. Knowledge-enhanced biomedical named entity recognition and normalization: application to proteins and genes. *Bioinform.*, 21(1):35.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 207–212, Berlin, Germany. Association for Computational Linguistics.