Leveraging Task Transferability to Meta-learning for Clinical Section Classification with Limited Data

Zhuohao Chen[1]*, **Jangwon Kim**[2], **Ram Bhakta**[2]†, **Mustafa Y. Sir**[2]
[1]University of Southern California, Los Angeles, USA
[2]Amazon, Seattle, USA
zhuohaoc@usc.edu {jangwok, simustaf}@amazon.com
hiten.bhakta@gmail.com

Abstract

Identifying sections is one of the critical components of understanding medical information from unstructured clinical notes and developing assistive technologies for clinical note-writing tasks. Most state-of-the-art text classification systems require thousands of in-domain text data to achieve high performance. However, collecting in-domain and recent clinical note data with section labels is challenging given the high level of privacy and sensitivity. This paper proposes an algorithmic way to improve the *task transferability* of meta-learning-based text classification in order to address the issue of low-resource target data. Specifically, we explore how to make the best use of the source dataset and propose a unique task transferability measure named *Normalized Negative Conditional Entropy* (NNCE). Leveraging the NNCE, we develop strategies for selecting clinical categories and sections from source task data to boost cross-domain meta-learning accuracy. Experimental results show that our task selection strategies improve section classification accuracy significantly compared to meta-learning algorithms.

1 Introduction

An important part of Electronic Health Records (EHRs) is the digitized clinical notes that contain the medical and treatment histories of patients. The section of clinical notes can be defined as a text segment that clusters consecutive sentences with relevant content of one dimension of a patient's health encounter (Pomares-Quimbaya et al., 2019). Clinical note sections, labeled with either headings or subheadings, make the notes well organized and offer improved clinical information extraction (Wang et al., 2018b). However, many clinical notes contain narratives that are in an unstructured free-text format, (e.g., History of Present Illnesses described in paragraph form), which makes it challenging to retrieve and utilize this information. In the United States, physicians generally spend an excessive amount of time interfacing with EHRs and computerized physician order entry (CPOE) workflows in their aftercare work, resulting in burnout, low job satisfaction, and system-wise inefficiencies (Patel et al., 2018). An automated section classifier can play a key role in mitigating this problem. In some cases, section classification serves as an end task of automatic report segmentation. For example, according to an internal survey we conducted with Amazon Care providers, we found evidence that classifying sentences related to the History of Present Illness from medical encounters can greatly assist providers with their documentation. For computer-assisted report generation, understanding clinical notes from an unstructured format is an important data pre-processing (Gopinath et al., 2020).

There are some challenges for clinical note section classification in practice. First, it is difficult to collect and access a large amount of in-domain data. Second, section types and medical contents within a section substantially vary depending on care providers, which makes it hard to utilize open-source datasets. Even though some sections exist in multiple different sources, their contents vary across clinical categories. For example, the Diagnosis section for Nutrition specialty and Rehabilitation Service specialty vary in types of content.

Recently developed neural network language technologies capture rich contextual information in sentences. Among them, Bidirectional Encoder Representations from Transformers (BERT) achieved significant improvements in multiple Natural Language Processing (NLP) tasks, establishing strong baselines in low-resource scenarios (Devlin et al., 2019). However, there remains room for performance improvement because BERT uses

*Work done during an internship at Amazon Care
†Work done while Ram Bhakta was a researcher at Amazon Care in 2021, and Ram is now affiliated to Oath Care, Austin, USA

source data – data outside of in-domain or target-domain data – in an unsupervised training fashion only. Another approach for low-resource in-domain NLP tasks is Multi-Task Learning (MTL). The MTL adopts shared text encoding layers across all tasks while the top layers are task-specific for each dataset (Liu et al., 2015, 2019). The target task with limited data benefits from the knowledge learned from source tasks. Instead of MTL, which minimizes the loss of the source tasks, Dou et al. (2019) proposed a model-agnostic meta-learning algorithm that finds optimal model parameters for better adaptation capability to new tasks. In classification tasks, Nichol et al. (2018) proposed Reptile, an optimization-based meta-learning algorithm for section classification, and achieved comparable accuracy on well-established benchmarks on low resourced image datasets. In the present paper, we adopted these methods as strong baselines in our experiments and computed the relative performance improvement of our method.

Task transferability denotes how easy it is to transfer the representation learned from one task to another task (Tran et al., 2019; Nguyen et al., 2020b). It helps discover the relationship between two types of tasks and provides supporting evidence for developing transfer learning strategies. Task transferability becomes more useful in realistic situations where the assumption of the meta-learning, which is that data of the target task can be drawn from the distribution of the source tasks, does not hold. One common example is that there are 'outlier tasks' in the training (source) tasks, which are dissimilar from the testing (target) ones (Venkitaraman et al., 2020). For this problem, good selection of relevant source tasks can benefit knowledge transfer to unseen tasks (Zamir et al., 2018; Achille et al., 2019; Nguyen et al., 2020a).

In clinical section classification, we suppose how close a source task is toward the target task is determined by its specialty and the section types included. However, few studies of task transferability estimation have discussed the function of each label. Thus we propose an information-theoretic metric for task transferability, namely Normalized Negative Conditional Entropy (NNCE). The NNCE score is calculated by the classifier of a source task and target data samples without training on the target task, thus saving expensive computation for model optimization. We hypothesize that this score correlates with how well the source data labels (sections) distinguish the target labels.

Leveraging the NNCE, we explore strategies of source task selection to improve the performance of meta-learning. The goal is to make the best use of available data from various clinical specialties for any target tasks. Specifically, we explore two strategies: 1) *category selection* - we select a subset of clinical categories that are relevant to the target task; 2) *section selection* - for a clinical category, we filter out the samples of certain section types which are not relevant to the target task and merge similar sections by assigning the same label. The category selection is informed directly by the best NNCE scores. For section selection, however, there are too many combinations, and it is time-consuming to train models for every possible task and find optimal ones. To handle that, we apply a backward selection method for heuristic search. The experiment results show that our task selection strategies improve the meta-transfer learning of section classification in low-resource scenarios.

Our work has the following contributions:

- We apply the meta-learning for clinical section classification at sentence level in low-resource scenarios utilizing out-of-domain datasets.
- We propose a task transferability metric for selecting the source tasks relevant to the target tasks by category and section selection, which improves meta-learning performance.
- We evaluate a computationally efficient backward selection method for section selection and show that it leads to a better knowledge transfer. To the best of our knowledge, this is the first attempt to apply class subset selection to improve the task transferability in the NLP field.

2 Related Work

In this section, we briefly discuss several areas in machine learning that are related to our work.

2.1 Clinical Section Classification

The goal of this paper is to address the automated clinical section classification task in low-resource scenarios. Notable early work focused on the extraction of frequency-based features and classified the sections of the clinical narratives with traditional machine learning approaches, including Support Vector Machines (Apostolova et al., 2009),

Maximum Entropy (MaxEnt) models (Tepper et al., 2012) and Bayesian models (Ganesan and Subotin, 2014). Li et al. (2010) framed section mapping as a sequence-labeling problem and adopted a Hidden Markov Model (HMM). Dai et al. (2015) formulated the task as a token-based classification using the conditional random fields (CRF) model. Ni et al. (2015) applied active learning and distant supervision to the section classification. In the study of Tran et al. (2015), the tasks were performed by an object-based section annotator using an ontology to describes the relationship among the section concepts. However, most of the studies above investigate the section classification task for a single domain without exploring how to transfer knowledge from the source dataset to an unseen target domain with limited data.

Recently, Rosenthal et al. (2019) leveraged the data from medical literature and performed section classification at the sentence level via transfer learning, recurrent neural networks (RNNs), and BERT in scenarios where a limited amount of in-domain training data was available. This work performs simple transfer learning and only predicts the shared sections across different clinical categories, and in practice, most section labels are domain-specific. This paper applies meta-learning and task transferability to transfer information learned from the source category to the target category with a new section classification task.

2.2 Meta-learning

Meta-learning aims at fast adaptation to new tasks with small amounts of data through learning knowledge from multiple source tasks. Among different approaches to meta-learning, one proposal is learning the initialization of a network that is good at adapting to new jobs. Dou et al. (2019) applied this proposal to the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2018a) and explored the model-agnostic meta-learning (MAML) (Finn et al., 2017) and its variants called first-order MAML (FO-MAML) and Reptile. In this paper, we adopted the Reptile algorithm that achieved the best performance in (Dou et al., 2019).

2.3 Task Transferability

Previous work explores the relationship between classification tasks on task similarity using traditional machine learning algorithms (Thrun and O'Sullivan, 1996; Bakker and Heskes, 2003; Xue et al., 2007; Zhang and Yeung, 2010). Other recent work mapped the functions into a vector space (Achille et al., 2019, 2021) to estimate the transferability using a non-symmetric distance. Vu et al. (2020) further developed the task embeddings approach and applied it to the NLP field to predict the most transferable source tasks. Zamir et al. (2018) modeled the underlying structure among different tasks to reduce the number of labeled training data. However, the common theme in all these approaches is that they require fine-tuning the target task and exhaustive optimization of parameters. The transferability estimation, unfortunately, is not robust if there are insufficient training samples. Moreover, none of these algorithms have discussed label selection which is crucial for task selection in clinical section classification. Tran et al. (2019) investigated the correlation of the label distributions between those tasks and proposed a negative conditional entropy (NCE) measure to estimate the task transferability. This algorithm only requires the source model and the labeled target samples without fine-tuning the in-domain data. Nguyen et al. (2020b) developed a variant of NCE measure called the Log Expected Empirical Prediction (LEEP) that denotes the average log-likelihood of the expected empirical predictor. Our proposed NNCE is similar in concept to NCE and LEEP. However, we apply the class subset selection to improve the knowledge transfer. Unlike previous work (Manjunatha et al., 2018), which does not use knowledge about the target task while finding the subset, our approach incorporates how the decision boundary of each source label distinguishes the labels of the target task.

3 Dataset

We conduct experiments on the Medical Information Mart for Intensive Care III (MIMIC-III) database (Johnson et al., 2016), a large open-access dataset of de-identified patient records. We collected data from 9 different clinical categories of MIMIC-III and randomly picked 200 clinical notes for each. There are nearly 1,000 section labels of these categories, and most of them contain very few sentence instances. To handle the sparsity, we only keep the section types of each category satisfying the following conditions:

- The section is among the ten most frequent ones.

Category	Nb. of Instance	Section labels
Discharge Summary Addendum	2.2K	addendum,discharge medications, service, dictated by, hospital course, medications on discharge, discharge diagnosis, discharge instructions, tablet sig, history of present illness
Discharge Summary Reports	8.8K	history of present illness, past medical history, hospital course, discharge instructions, tablet sig, impression, discharge medications, social history,allergies, medications on admission
Echo	6.0K	conclusions,mitral valve, left ventricle, aortic valve, tricuspid valve, general comments,aorta, right ventricle, right atrium/interatrial septum, impression
Nutrition	2.4K	Specifics,labs, current diet order / nutrition support, gi,pertinent medications, ptat risk due to, tube feeding / tpn recommendations, comments, diagnosis, protein
Nursing Generic	5.4K	plan, assessment, action, response, vs, chief complaint:
Nursing Progress	2.8K	plan, assessment, action, response
Recab Service Evaluation	4.3K	clinical impression/prognosis, time frame, diagnosis,history of present illness / subjective complaint,arousal/attention/cognition/communication, pulmonary status, education /communication, prior functional status/activity level, frequency/duration,posture:
Recab Service Progress	2.5K	assessment,balance, updated medical status, education / communication, gait, plan,anticipated discharge, aerobic activity response, rolling, follow up ptvisit to address goals of
Social Work	3.0K	patient/ family assessment, continuing issues to be addressed, employment status, previous living situation, previous level of functioning, assessment, past addictions history, plan / follow up, past psychiatric history, healthcare proxy appointed

Table 1: Sentence and section lists in each MIMIC-III category.

- The number of sentences with this section label is more than 2% of the total instances.

Table 1 shows the number of sentence instances and the lists of selected section types. The section list varies across categories, with only a few section labels in more than one domain. However, some sections in different categories are still related to each other. For example, sentences in the *social history* section of 'Discharge Summary Reports' category are similar to the instances in the *employment status* and *previous living situation* section of 'Social Work'.

4 Methods

4.1 Meta-learning Approach

We adopt Reptile, an optimization-based meta-learning algorithm, to be our baseline approach. Assume we have a set of source tasks $\{T_1, T_2,...,T_N\}$ from multiple open-resource clinical datasets. We perform the Reptile with these source tasks to learn the BERT model parameters ϕ to provide a good initialization for fine-tuning the target task. For sampling batches of tasks, we use the same strategy proposed in Dou et al. (2019) that the probability of selecting a task is proportional to the size of its dataset. The training procedure of Reptile is described in Algorithm 1 where β denotes learning rate. In the baseline meta-learning approach, we train the model with all the available datasets without data selection which might suffer from 'outlier' tasks. In the next step, we leverage the task transferability estimation for selecting the sources tasks bettering transferring knowledge to the target task.

4.2 Normalized Negative Conditional Entropy

Fig. 1 shows the general framework of NNCE. The motivation of the NNCE for estimating the task transferability is the idea of evaluating how well the decision boundaries of source labels distinguish the target labels.

Algorithm 1 Reptile Approach

Initialize model parameters ϕ with the pre-trained BERT
for iteration in 1,2,... **do**
  Sample batch of tasks $\{T_i\}$ proportional to the size of its dataset
  for all T_i **do**
    Compute ϕ_i^k: k steps of gradient descent
    Update $\phi = \phi + \beta \frac{1}{|T_i|} \sum_{T_i} (\phi_i^k - \phi)$

Consider a source task defined on $\mathcal{X} \times \mathcal{Y}$ and a target task on $\mathcal{X} \times \mathcal{Z}$. We denote the target samples as $D = \{(x_1, z_1), (x_1, z_2), ..., (x_n, z_n)\}$ and use $y \in \mathcal{Y} = \{1, 2, ..., L_S\}$ and $z \in \mathcal{Z} = \{1, 2, ..., L_T\}$ to represent the label variables of source and target data respectively. We train a classifier f on the source task which maps the space $\mathcal{X}$ to $\mathcal{Y}$. By feeding the target samples into the source model f, we assign the predicted source labels for the target samples so that $\hat{Y} = \{\hat{y}_1, \hat{y}_2, ..., \hat{y}_n\}$. Thus, every target sample is attached with a 'true

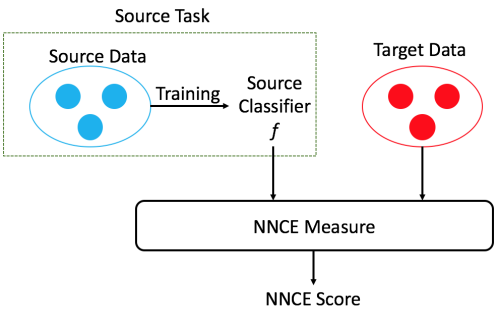

Figure 1: NNCE measure.

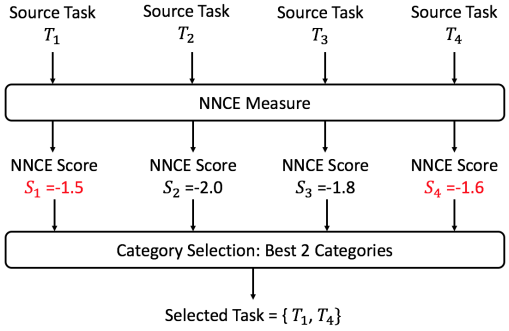

Figure 2: Category selection example.

label' from $\mathcal{Z}$ and a predicted label from $\mathcal{Y}$ that can be denoted as $(x_i, \hat{y}_i, z_i)$.

We compute the empirical joint distribution and the empirical marginal distribution by

$$\begin{aligned}\hat{P}(y) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{\hat{y}_i = y\},\\ \hat{P}(z) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{z_i = z\},\\ \hat{P}(y,z) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{\hat{y}_i = y, z_i = z\}.\end{aligned} \tag{1}$$

To measure how the source and task labels are related, we handle the class imbalance issue of the target dataset by normalizing the target class frequency:

$$\widetilde{P}(y,z) = \hat{P}(y|z) = \frac{\hat{P}(y,z)}{\hat{P}(z)}. \tag{2}$$

The value of $\widetilde{P}(y,z)$ represents the ratio of the target samples in class z that are assigned with the predicted label y. Then we compute:

$$\widetilde{P}(z|y) = \frac{\widetilde{P}(y,z)}{\sum_{z}^{L_T}\widetilde{P}(y,z)}. \tag{3}$$

so that $\sum_z \widetilde{P}(z|y) = 1$. We suppose that a good source label $y = l$ that distinguishes the target labels well should have large values of $\widetilde{P}(z = l|y)$ for some target classes as well as small values for other target classes. On the contrary, if the values of $\widetilde{P}(z = l|y)$ for different target class z are approximately equal, this label is useless for classifying the target labels. Based on that, we define the NNCE to estimate the task transferability by:

$$\begin{aligned}NNCE &= \sum_{y\in\mathcal{Y}}^{L_S}\hat{P}(y)\sum_{z\in\mathcal{Z}}^{L_T}\widetilde{P}(z|y)log\widetilde{P}(z|y)\\ &= \sum_{y\in\mathcal{Y}}^{L_S}\hat{P}(y)E(y)\end{aligned} \tag{4}$$

where we use $E(y) = \sum_{z\in\mathcal{Z}}^{L_T}\widetilde{P}(z|y)log\widetilde{P}(z|y)$ to estimate how well the decision boundary of a source label classifies the target classes and NNCE is the overall measurement weighted by the prior $\hat{P}(y)$. NNCE score is always negative. For a determined target task, a larger score indicates better transferability between the source and target tasks. The advantage of NNCE over some other label correlation methods like LEEP is that it allows us to select the source labels better distinguishing the target class with respect to$E(y)$. The NNCE is related to the NCE proposed by Tran et al. (2019), and it is equal to NCE if we do not normalize the target class frequency in Equation (2). The proof is in the Appendix A.

4.3 Task Selection for Clinical Section Classification

We suppose that selecting the source tasks with good task transferability can benefit the meta-learning of the low-resource target task. In clinical section classification tasks, the pattern of the data and the section types vary across categories. So we propose two approaches for choosing the source tasks - category selection and section selection.

4.3.1 Category Selection

The procedure of category selection is direct. Fig. 2 shows a simple example of category selection. We compute the NNCE score for each of the source tasks from different clinical categories. Then we pick the N 'best' categories whose task achieves

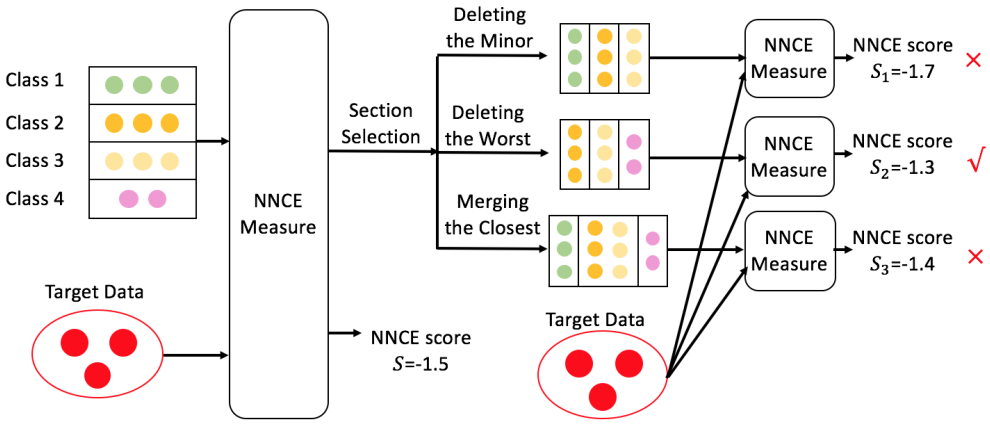

Figure 3: A single step for section selection.

the highest NNCE scores. This approach helps filter out the 'outlier' tasks by removing the clinical categories irrelevant to the target task.

4.3.2 Section Selection

Section selection is a process of searching for the optimal task for each of the clinical categories. It aims to make the best use of the section labels to benefit transferring knowledge to the target task. We modify the list of the section classes by deleting the instances from the useless sections and merge similar ones. However, there are too many combinations for partitioning that lead to high computational costs. To reduce the computational complexity, we propose a backward selection method with three operations for heuristic search.

We perform a section selection procedure with NNCE measure and the following three operations that delete or merge sections to generate new tasks:

Deleting the Minor

We delete the section l^* of the source dataset with the smallest value of empirical marginal distribution $\hat{P}(y)$:

$$l^* = \underset{l}{\operatorname{argmin}}\, \hat{P}(y = l) \tag{5}$$

The motivation behind this operation is that the fewest target samples are tagged with source label l^* representing this section is unrelated to the target category.

Deleting the Worst

We delete the section l^* satisfying:

$$l^* = \underset{l}{\operatorname{argmin}}\, E(y = l) \tag{6}$$

From the demonstration in Section 4.2 we can conclude l^* has the smallest value of $E(y)$, which indicates the source section l^* is worst at distinguishing the target sections.

Merging the Closest

This operation aims to find the 'closest' pair of the source sections and merge them into one. To find such sections i^*, j^*, we adopt the following equation:

$$i^*, j^* = \underset{i,j\ i \neq j}{\operatorname{argmin}}\, JSD(\widetilde{P}(z|y = i) \parallel \widetilde{P}(z|y = j)) \tag{7}$$

where $JSD(\cdot)$ presents the Jensen–Shannon divergence (Lin, 1991). A small value of $JSD(\cdot)$ indicates that the $\widetilde{P}(z|y = i^*)$ and $\widetilde{P}(z|y = j^*)$ distribute closely and the source sections i^* and j^* are similar. In this case, the decision boundary between the source label i^* and j^* are trivially helpful for discriminating the target labels.

Backward Selection

We initialize the source task by including all the samples and sections labels, and perform a backward selection algorithm to reduce the section numbers iteratively. Fig. 3 shows a single step of this process. We apply the NNCE measure with three operations introduced before to generate NNCE scores and produce no more than three new tasks [1]. Then we compute the NNCE score for each of the new tasks. The final picked task at this step is the one that achieves the highest scores among the original one and the newly generated ones. We keep

[1] Different operations may result in the same task.

Target category	Sample size	By chance	BERT	MTL	Reptile
Discharge Summary Report	200	0.216	0.542	0.552	**0.558(p<0.01)***
	500	0.216	0.632	0.640	**0645(p=0.06)**
	1000	0.216	0.673	0.678	**0.680(p=0.16)**
Nursing Progress	200	0.332	0.645	0.649	**0.650(p=0.22)**
	500	0.332	0.742	0.745	**0.747(p=0.23)**
	1000	0.332	0.785	0.787	**0.788(p=0.36)**
Rehab Service Progress	200	0.254	0.848	0.855	**0.857(p=0.04)**
	500	0.254	0.923	0.926	**0.927(p=0.10)**
	1000	0.254	0.948	**0.950**	**0.950(p=0.19)**
Social Work	200	0.580	0.818	0.828	**0.834(p<0.01)**
	500	0.580	0.896	0.904	**0.907(p=0.01)**
	1000	0.580	0.934	0.936	**0.938(p=0.04)**

*Comparing with BERT

Table 2: The Classification accuracy results of baseline approaches.

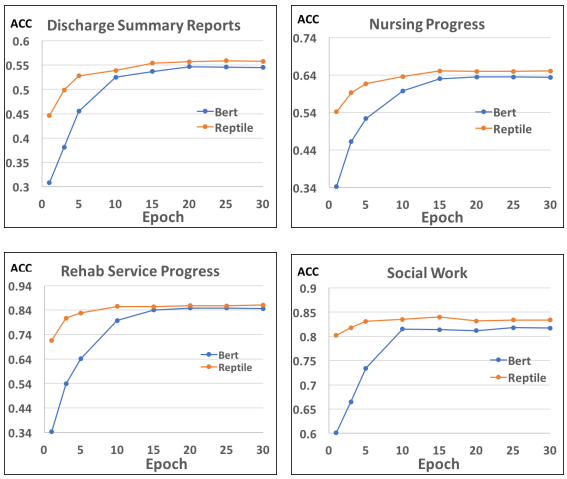

Figure 4: Convergence of accuracy of fine-tuning for sample size=200.

performing this process until none of the produced tasks improves the NNCE score anymore.

5 Experiment Results and Discussion

We carry out the experiments with four target tasks of different clinical categories 'Discharge Summary Report', 'Nursing Progress', 'Recab Service Progress' and 'Social Work' presented in Table 1. For the target task of 'Social Work', we utilize all the other eight categories for pre-training. For 'Discharge Summary Report', 'Nursing Progress' and 'Recab Service Progress', we remove their close categories - 'Discharge Summary Addendum', 'Nursing Generic' and 'Recab Service Evaluation' categories, respectively, and the pre-training is performed by the remaining seven categories.

For each target categories, we split the samples into the training and testing set with a roughly 3:1 ratio across the 'SUBJECT_ID' referring to a unique patient. We randomly pick 200/500/1000 samples from each target datasets to simulate low-resource scenarios and perform BERT, MTL, and Reptile for the clinical section classification.

5.1 Implementation Details

We adopt the PyTorch (version 1.3.0) implementation of BERT[2] for our tasks and the model is initialized with **BERT-base**. The settings of MTL and Reptile are same as the ones described in (Dou et al., 2019). We threshold the word sequence length to 80, which covers more than 99% of the sentences. We use Adam (Kingman and Ba, 2015) for optimization and a batch size of 32 for all the experiments. For both MTL and Reptile, the learning rate is 5e-5, and the number of pre-training epoch is 5. We set the inner update step k to be 5, the inner learning rate to be 5e-5 and the number of sampled tasks in each step to be 8 for Reptile. For BERT fine-tuning, we train the model with the learning rate of 2e-5 for 25 epochs.

[2]https://github.com/huggingface/pytorch-pretrained-BERT

5.2 Results of Baseline Approaches

The classification accuracy results of BERT, MTL, and Meta-learning for different tasks are shown in Table 2. From the table, we find that both MTL and Reptile improve the performance of the low-resource target task while Reptile outperforms multi-task learning and achieves the best results. The comparison between BERT and Reptile demonstrates that the meta-learning approach can benefit the fine-tuning of the target task. The improvement is more significant when we perform the classification task with fewer target samples.

Fig. 4 shows the convergence of accuracy of BERT fine-tuning with and without Reptile pre-training. The curves in these figures suggest that meta-learning has the advantage of fast convergence and adapts to the new task more quickly. We also discover that after 15 epochs of fine-tuning, the performance is not sensitive to the epoch number.

5.3 Evaluation of NNCE

For any selected target category, we pre-train the model for each of the remaining categories and fine-tune with 200 target samples to obtain the transfer learning accuracies. We compute the NNCE scores for different source tasks and evaluate the NNCE by

Target category	Correlation coefficients	
	NCE	NNCE
Discharge Summary Report	0.671	**0.676**
Nursing Progress	0.772	**0.807***
Rehab Service Progress	0.918	**0.922**
Social Work	0.479	**0.703***

*The correlations between the NNCE scores and transfer learning accuracy are statistically significant with $p < 0.05$.

Table 3: Comparison of Pearson correlation coefficients of NCE and NNCE(Tran et al., 2019).

the Pearson correlation coefficients between these scores and their accuracies of adaptation. We also report the correlations using the NCE scores for comparison. By comparing the correlation coefficients presented in Table 3, we find that NNCE receives higher correlations over NCE for all the tasks and is better at task transferability estimation.

5.4 Results of Task Selection

We set the target sample size to be 200 and explore how task selection strategies - category selection and section selection benefit meta-learning.

Table 4 shows the results of meta-learning approach with category selection. We report the classification accuracies of picking N = 2/4/6 categories with the highest NNCE scores and compare with including all the source categories. The results reveal that the category selection improves the meta-learning performance, and there is an optimal value of N for each task. If N is too large, it might include 'outlier' tasks that degrade the performance. If N is too small, it loses the benefit of utilizing large amounts of source data. We also perform the category selection with NCE to compare it with NNCE. The underlined tasks in Table 4 indicate that different subsets of categories are selected if we replace NNCE with NCE. For all these tasks, NNCE achieves higher accuracies. Please see Appendix C for detailed results for different target categories.

We discuss whether the section selection benefits the meta-learning in two scenarios. First, we compare the performances of Reptile with and without section selection using all the source categories. In the second scenario, we repeat the first procedure but only use the best subset of the source categories determined in Table 4, and repeat the comparison method in the first scenario. The comparisons presented in Table 5 indicate that adopting section selection can improve the classification performance of Reptile in both scenarios. However, the improvement is not statistically significant for most tasks. The average relative gains to Reptile brought by the category selection and section selection are 1.5% and 0.8%, which indicates that category selection contributes more to improving the meta-learning. We also find that combining both category and section selection results in better performance than using each of them independently for most tasks.

Task	Nb. selected categories			
	2	4	6	All**
Discharge Summary Report	0.556	**0.569***	0.559	0.558$^+$
Nursing Progress	0.660	**0.666***	0.645	0.650$^+$
Rehab Service Progress	0.859	**0.862**	0.861	0.857
Social Work	0.835	0.838	**0.843**	0.834

* is significantly higher than $^+$ at $p < 0.05$.

Table 4: The classification accuracy of Reptile with category selection. The categories are selected with the highest NNCE scores. **'All' denotes all the original source tasks are included

Task	Reptile	Reptile + SS	Reptile + BCS	Reptile + BCS + SS
Discharge Summary Report	0.558	0.562	0.569$^+$	**0.579***
Nursing Progress	0.650	0.655	**0.666**	0.665
Rehab Service Progress	0.857$^+$	0.866*	0.862	**0.867**
Social Work	0.834	0.840	0.843	**0.846**

* is significantly higher than $^+$ at $p < 0.05$.

Table 5: The classification accuracy of Reptile with and without section selection. SS: section selection, BCS: Best category subset

We show an example in Table 6 to further illustrate section selection. The source and target categories are 'Rehab Service Progress' and 'Nursing Progress', and the original section types are presented. The labels in blue are the selected sections, and the merged ones are displayed inside the brackets. We observe that the common section types - *plan* and *assessment* are kept. Although the content of the same section type is different across categories, there are similarities between their utterance patterns. The source sections in black are irrelevant to any of the target sections, so they are removed. The merged sections *balance* and *gait* are of close concepts, both of which describe the patient's progress of mobility. This example shows

that the selection procedure extracts information of the source sections related to the target sections, which benefits the knowledge transfer.

	Target Category: Nursing Progress	Source Category: Rehab Service Progress
Section Labels	**plan, assessment, action, response**	**plan, assessment, {balance, gait}, updated medical status, education / communication, aerobic activity response, anticipated discharge rolling, follow up pt visit to address goals of**

Table 6: An Example of Section Selection

6 Conclusion and Future Work

In this paper, we explored the clinical section classification with limited in-domain data. We applied a meta-learning algorithm utilizing multiple out-of-domain clinical datasets, improving the classification accuracy and adaptation speed. We proposed a *Normalized Negative Conditional Entropy* measure to estimate the task transferability and leverage it to select the clinical categories and sections related to the target task that best improves knowledge transfer. In addition, we examined a backward selection method to reduce the computational complexity of section selection. Our study suggests that both category selection and section selection outperform the baseline meta-learning approach, and combining two strategies results in better performance than adopting each of them independently.

Future work will look to develop a joint optimization of category selection and section selection. We also plan to apply our approach to other styles of text data. For example, section classification on spoken utterances of doctor-patient conversations is an exciting extension of the present work, which we plan to explore (Krishna et al., 2021). Finally, we will continue to apply the proposed method to other text processing applications, e.g., medical information retrieval (Goeuriot et al., 2016).

References

Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless C Fowlkes, Stefano Soatto, and Pietro Perona. 2019. Task2vec: Task embedding for meta-learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6430–6439.

Alessandro Achille, Giovanni Paolini, Glen Mbeng, and Stefano Soatto. 2021. The information complexity of learning tasks, their structure and their distance. *Information and Inference: A Journal of the IMA*, 10(1):51–72.

Emilia Apostolova, David S Channin, Dina Demner-Fushman, Jacob Furst, Steven Lytinen, and Daniela Raicu. 2009. Automatic segmentation of clinical texts. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5905–5908. IEEE.

Bart Bakker and Tom Heskes. 2003. Task clustering and gating for bayesian multitask learning. *Journal of Machine Learning Research*, 4:83–99.

Hong-Jie Dai, Shabbir Syed-Abdul, Chih-Wei Chen, and Chieh-Chen Wu. 2015. Recognition and evaluation of clinical section headings in clinical documents using token-based formulation with conditional random fields. *BioMed research international*, 2015.

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.

Zi-Yi Dou, Keyi Yu, and Antonios Anastasopoulos. 2019. Investigating meta-learning algorithms for low-resource natural language understanding tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1192–1197.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR.

Kavita Ganesan and Michael Subotin. 2014. A general supervised approach to segmentation of clinical texts. In *2014 IEEE International Conference on Big Data (Big Data)*, pages 33–40. IEEE.

Lorraine Goeuriot, Gareth JF Jones, Liadh Kelly, Henning Müller, and Justin Zobel. 2016. Medical information retrieval: introduction to the special issue. *Information Retrieval Journal*, 19(1-2):1–5.

Divya Gopinath, Monica Agrawal, Luke Murray, Steven Horng, David Karger, and David Sontag. 2020. Fast, structured clinical documentation via contextual autocomplete. In *Machine Learning for Healthcare Conference*, pages 842–870. PMLR.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits,

Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

DP Kingman and J Ba. 2015. Adam: A method for stochastic optimization. conference paper. In *Proceedings of the 3rd International Conference on Learning Representations*.

Kundan Krishna, Sopan Khosla, Jeffrey P. Bigham, and Zachary C. Lipton. 2021. Generating SOAP notes from doctor-patient conversations using modular summarization techniques.

Ying Li, Sharon Lipsky Gorman, and Noémie Elhadad. 2010. Section classification in clinical notes using supervised hidden markov model. In *Proceedings of the 1st ACM International Health Informatics Symposium*, pages 744–750.

Jianhua Lin. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng, Kevin Duh, and Ye-yi Wang. 2015. Representation learning using multi-task deep neural networks for semantic classification and information retrieval. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 912–921.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496.

Varun Manjunatha, Srikumar Ramalingam, Tim K Marks, and Larry Davis. 2018. Class subset selection for transfer learning using submodularity. *arXiv preprint arXiv:1804.00060*.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. Scispacy: Fast and robust models for biomedical natural language processing. *arXiv preprint arXiv:1902.07669*.

Cuong Nguyen, Thanh-Toan Do, and Gustavo Carneiro. 2020a. Similarity of classification tasks. In *4th Workshop on Meta-Learning at NeurIPS*.

Cuong Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. 2020b. Leep: A new measure to evaluate transferability of learned representations. In *International Conference on Machine Learning*, pages 7294–7305. PMLR.

Jian Ni, Brian Delaney, and Radu Florian. 2015. Fast model adaptation for automated section classification in electronic medical records. *MedInfo*, 216:35–9.

Alex Nichol, Joshua Achiam, and John Schulman. 2018. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*.

Rikinkumar S Patel, Ramya Bachu, Archana Adikey, Meryem Malik, and Mansi Shah. 2018. Factors related to physician burnout and its consequences: a review. *Behavioral sciences*, 8(11):98.

Alexandra Pomares-Quimbaya, Markus Kreuzthaler, and Stefan Schulz. 2019. Current approaches to identify sections within clinical narratives from electronic health records: a systematic review. *BMC medical research methodology*, 19(1):1–20.

Sara Rosenthal, Ken Barker, and Zhicheng Liang. 2019. Leveraging medical literature for section prediction in electronic health records. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4864–4873.

Michael Tepper, Daniel Capurro, Fei Xia, Lucy Vanderwende, and Meliha Yetisgen-Yildiz. 2012. Statistical section segmentation in free-text clinical records. In *Lrec*, pages 2001–2008.

Sebastian Thrun and Joseph O'Sullivan. 1996. Discovering structure in multiple learning tasks: The tc algorithm. In *ICML*, volume 96, pages 489–497.

Anh T Tran, Cuong V Nguyen, and Tal Hassner. 2019. Transferability and hardness of supervised classification tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1395–1405.

Le-Thuy T Tran, Guy Divita, Andrew Redd, Marjorie E Carter, Matthew Samore, and Adi V Gundlapalli. 2015. Scaling out and evaluation of obsecan, an automated section annotator for semi-structured clinical documents, on a large va clinical corpus. In *AMIA Annual Symposium Proceedings*, volume 2015, page 1204. American Medical Informatics Association.

Arun Venkitaraman, Anders Hansson, and Bo Wahlberg. 2020. Task-similarity aware meta-learning through nonparametric kernel regression. *arXiv preprint arXiv:2006.07212*.

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.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018a. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, et al. 2018b. Clinical information extraction applications: a literature review. *Journal of biomedical informatics*, 77:34–49.

Ya Xue, Xuejun Liao, Lawrence Carin, and Balaji Krishnapuram. 2007. Multi-task learning for classification with dirichlet process priors. *Journal of Machine Learning Research*, 8(1).

Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. 2018. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3712–3722.

Yu Zhang and Dit-Yan Yeung. 2010. A convex formulation for learning task relationships in multi-task learning. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pages 733–742.

Appendix

A The Relationship between NNCE and NCE

Proposition: NNCE is equal to NCE if we do not normalize the target class frequency in Equation (2).

Without normalizing the target class frequency, we modify the empirical distributions in Equation (2) and Equation (3) by

$$\begin{aligned} \widetilde{P}(y,z) &= \hat{P}(y,z), \\ \widetilde{P}(y|z) &= \hat{P}(y|z). \end{aligned} \tag{8}$$

Based on Equation (8) and the definition of NNCE in Equation (4) , we can achieve the new formula of NNCE:

$$\begin{aligned} NNCE &= \sum_{y\in\mathcal{Y}}^{L_S} \hat{P}(y) \sum_{z\in\mathcal{Z}}^{L_T} \widetilde{P}(z|y) log \widetilde{P}(z|y) \\ &= \sum_{y\in\mathcal{Y}}^{L_S} \hat{P}(y) \sum_{z\in\mathcal{Z}}^{L_T} \hat{P}(z|y) log \hat{P}(z|y) \\ &= \sum_{y\in\mathcal{Y}}^{L_S} \sum_{z\in\mathcal{Z}}^{L_T} \hat{P}(y)\hat{P}(z|y) log \hat{P}(z|y) \\ &= \sum_{z\in\mathcal{Z}}^{L_T} \sum_{y\in\mathcal{Y}}^{L_S} \hat{P}(y,z) log \frac{\hat{P}(y,z)}{\hat{P}(y)}. \end{aligned} \tag{9}$$

which is equal to the definition of NCE in (Tran et al., 2019).

B Data Preprocessing

We considered a new line starting with '[A-Z][a-zA-Z]+:' as the first line of a new label. Then, '[A-Z][a-zA-Z]+:' in the line became the new label, while text after ':' until another new label became the text data of the label. Then, we split the text data into sentence-level data by two sequential processes: (Step 1) Splitting it if starting with uppercase at the beginning of the newline or if it is an empty line, and then (Step 2) Splitting it further by SciSpacy sentencizer with en_core_sci_sm model (Neumann et al., 2019). The multi-label sentences (1.4% of 38326 instances) are filtered out.

For each collected sentences, we remove the punctuation marks and special characters like ==, –,*. We replace the de-identified brackets and time phrases like hh:mm:ss with the symbols "[phi]" and "[num]" that are added into the BERT vocabulary.

C Category Selection Details

Tables 7, 8, 9, and 10 show the category selection details for different target categories. We compare the NCE and NNCE by their selected categories and the classification accuracy of Reptile. For all these tasks, we observe that NNCE achieves higher accuracies. However, the difference between the NCE and NNCE is not very evident, presumably because the number of the total source categories is small, making their subset of the selected categories similar. A more standard way to compare these two metrics is the correlation coefficients between the NNCE scores and transfer learning accuracy, shown in Table 3 in the main body.

Nb. Selected Categories	Selected Categories		Accuracy	
	NCE	NNCE	NCE	NNCE
2	Social Work, Nutrition	-	0.556	-
4	Social Work, Nutrition, Nursing Generic, Rehab Service Progress	-	0.569	-
6	Social Work, Nutrition, Nursing Generic, Rehab Service Progress, Echo, Rehab Evaluation	Social Work, Nutrition, Nursing Generic, Rehab Service Progress, Echo, Nursing Progress	0.556	**0.559**

Table 7: The category selection details for the task category Discharge Summary Report. - denotes that the NCE and NNCE select the same subset of the categories

Nb. Selected Categories	Selected Categories		Accuracy	
	NCE	NNCE	NCE	NNCE
2	Discharge Summary Report, Rehab Service Evaluation	Discharge Summary Report, Rehab Service Progress	0.653	**0.660**
4	Discharge Summary Report, Rehab Service Progress, Rehab Service Evaluation, Echo	Discharge Summary Report, Rehab Service Progress, Rehab Service Evaluation, Nutrition	0.654	**0.666**
6	Discharge Summary Report, Rehab Service Progress, Rehab Service Evaluation, Nutrition, Social Work, Echo	-	0.645	-

Table 8: The category selection details for the task category Nursing Progress. - denotes that the NCE and NNCE select the same subset of the categories

Nb. Selected Categories	Selected Categories		Accuracy	
	NCE	NNCE	NCE	NNCE
2	Echo, Nursing Progress	-	0.859	-
4	Echo, Nursing Progress Nursing Generic, Social Work	-	0.862	-
6	Echo, Nursing Progress Nursing Generic, Social Work, Discharge Summary Addendum, Nutrition	-	0.556	-

Table 9: The category selection details for the task category Rehab Service Progress. - denotes that the NCE and NNCE select the same subset of the categories

Nb. Selected Categories	Selected Categories		Accuracy	
	NCE	NNCE	NCE	NNCE
2	Rehab Service Progress, Rehab Service Evaluation	Rehab Service Progress, Discharge Summary Report	0.829	**834**
4	Rehab Service Progress, Rehab Service Evaluation, Discharge Summary Report, Nursing Generic	Rehab Service Progress, Rehab Service Evaluation, Discharge Summary Report, Discharge Summary Addendum	0.834	**838**
6	Rehab Service Progress, Rehab Service Evaluation, Discharge Summary Report, Discharge Summary Addendum, Nutrition, Echo	Rehab Service Progress, Rehab Service Evaluation, Discharge Summary Report, Discharge Summary Addendum, Nursing Generic, Nursing Progress	0.831	**843**

Table 10: The category selection details for the task category Social Work.