Intermediate Domain Finetuning for Weakly Supervised Domain-adaptive Clinical NER

Shilpa Suresh , Nazgol Tavabi , Shahriar Golchin ,
Leah Gilreath , Rafael Garcia-Andujar , Alexander Kim ,
Joseph Murray , Blake Bacevich , and Ata M. Kiapour

Musculoskeletal Informatics Group, Boston Children's Hospital-Harvard Medical School
{shilpa.suresh, nazgol.tavabi, ata.kiapour}@childrens.harvard.edu

Abstract

Accurate human-annotated data for real-world use cases can be scarce and expensive to obtain. In the clinical domain, obtaining such data is even more difficult due to privacy concerns which not only restrict open access to quality data but also require that the annotation be done by domain experts. In this paper, we propose a novel framework - *InterDAPT* - that leverages Intermediate Domain Finetuning to allow language models to adapt to narrow domains with small, noisy datasets. By making use of peripherally-related, unlabeled datasets, this framework circumvents domain-specific data scarcity issues. Our results show that this weakly supervised framework provides performance improvements in downstream clinical named entity recognition tasks.

1 Introduction

Domain adaptation is a vast topic that is central to the field of biomedical NLP largely due to the limited number of high-quality, task and domain-specific biomedical and clinical corpora available. In such situations, it would be advantageous if we could leverage any available poorly-annotated, unlabeled datasets to boost model performance as opposed to entirely relying on one small, noisy dataset. In this work, we propose an ensemble training strategy combining multiple machine learning training methods to mitigate the issues impacting downstream named entity recognition (NER) performance caused by incorrect and missing labels in highly domain-specific datasets.

Biomedical applications of NLP often require high-quality annotation of task-specific datasets which can be very expensive. In addition, such data may be protected by privacy concerns which would make human annotation harder to perform. Another issue that is commonly observed with human annotations is that there may be variability in labels from different annotators. For example, NER labels in the Beginning-Inside-Outside (BIO) format only allow annotators to assign one entity label to each span. However, depending on the sentence structure and the definitions of each entity category, the same span might be considered to be a candidate for two or more entity labels. In such situations, different annotators might assign labels differently regardless of the standardized guidelines provided to all annotators. Such issues may be difficult to identify and expensive to rectify.

Due to these issues, leveraging weak supervision techniques which use unlabeled data to buttress the impact of noisy labels would be of interest. Our proposed training framework uses noisy, machine-labeled data to help create a transferable middle layer model which, when combined with DAPT (Domain Adaptive Pre-Training) (Gururangan et al., 2020) would allow a pre-trained model to better adapt to noisy clinical data.

Our generalized training framework which we refer to as the *Intermediate Finetuning for Weakly Supervised Sub-domain Adaptation (InterDAPT)* Framework is described in Figure 1. In the first phase, we perform continual Domain Adaptive Pre-training (DAPT) (Gururangan et al., 2020) on BioClinicalBERT (Alsentzer et al., 2019) using generalized patient data. In parallel, we label orthopedics-related operative notes using Radiology NER models provided by John Snow (Kocaman and Talby, 2021). This machine-annotated dataset is used to train the intermediate model which is then used to finetune downstream NER tasks. We apply this approach on two different datasets from orthopedics-related clinical notes - Spine and Hip. Our preliminary results indicate that our proposed framework achieves similar or better performance for the same dataset with noisy labels as can be achieved with clean labels.

The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 320–325

2 Related Work

2.1 Domain Adaptation

Domain adaptation is a discipline that is central to Biomedical and Clinical Applications of NLP and hence has been widely studied. One of the dominant methods of domain adaptation is domain adaptive pre-training which involves continued pre-training of transformer-based language models using domain-specific data. Domain Adaptive Pre-training (DAPT) (Gururangan et al., 2020) has been shown to be an effective method in Clinical NLP through the introduction of ClinicalBERT, BioClinicalBERT (Alsentzer et al., 2019), BioBERT (Lee et al., 2019) and BCH-BERT (Tavabi et al., 2022).

2.2 Intermediate Finetuning

Intermediate Task finetuning has been shown to be an effective method to improve task transferability by (Phang et al., 2019), (Chang and Lu, 2021) and (Pruksachatkun et al., 2020). In these works, the authors observe that the best improvements in performance are seen in NLP tasks that are the most closely related. In order to leverage these performance boosts, we perform *intermediate domain finetuning* whereby we perform intermediate finetuning using datasets from the same NLP task in different domains.

2.3 Weak Supervision

Weakly Supervised Learning is a machine learning discipline that studies the impact of training models using noisy labels where labels are either incorrect or absent altogether (Violeta et al., 2022). In industry applications, training frameworks are often employed as a way of dealing with weak labels. COSINE (Yu et al., 2020) is a weak supervision framework that uses a feedback mechanism to correct weak labels on the fly during the training process.

3 Disambiguating Domain and Task

As there is little domain variability in most of the task-specific corpora that are commonly used for benchmarking in NLP research, many works tend to conflate tasks with domains. In this paper, we seek to disambiguate these terms as the goal of InterDAPT is to improve the performance of entity recognition models in the absence of high-quality data in some areas by leveraging relatively large datasets in other areas with the assumption that there is some overlap of implicit information among these datasets by virtue of their domains being related. Hereafter, we refer to NLP tasks such as named entity recognition as *tasks*. We break down categorizations of domains further into meta-domains, tangential domains, and sub-domains. We define the following terms:

1. *Meta-domain:* A domain that is one abstraction level away from our set of target sub-domains.

2. *Tangential Domain:* A domain that has some overlap with either the meta-domain or the target sub-domains

3. *Target Sub-domain:* A domain that can be categorized under the meta-domain but has specific characteristics that distinguish it from other sub-domains as well as the tangential domain.

As the main motivation for this work is to improve model performance in highly specific clinical domains with inconsistent or incorrect labeling, we evaluate InterDAPT's performance on the named entity recognition (NER) task. We keep the task stable for the purposes of our experiments in order to assess the impact of the different types of domain-specific datasets while ruling out possible performance variations brought on by variations in the type of task.

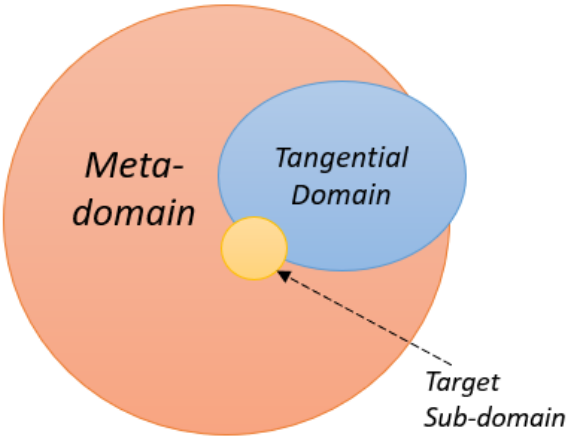

Figure 1: Relationship between types of domain-specific data used to train InterDAPT

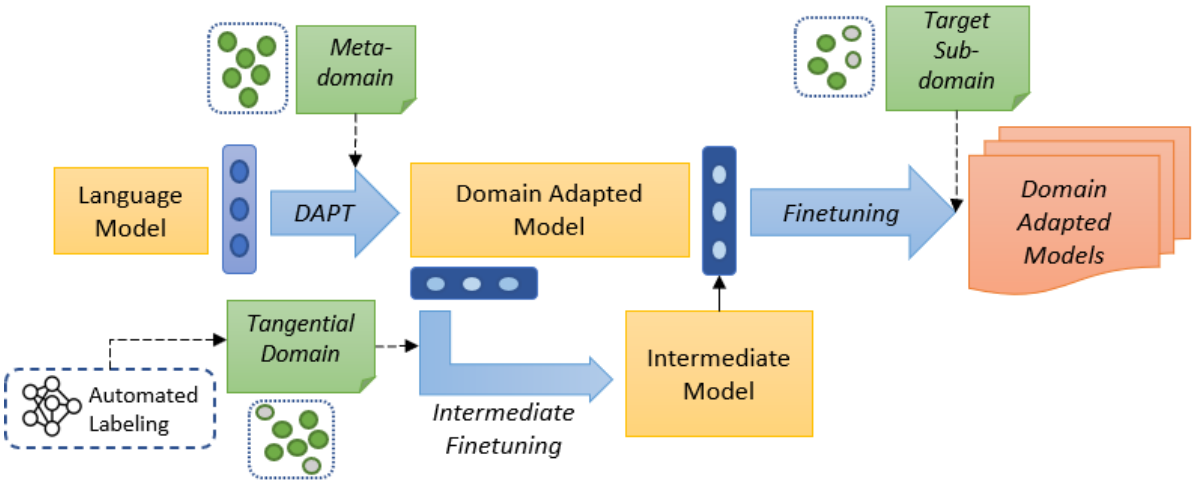

Figure 2: Generalized Intermediate Finetuning for Weakly Supervised Sub-domain Adaptation (InterDAPT) Framework

4 Method

4.1 Intermediate Finetuning for Weakly Supervised Sub-domain Adaptation (InterDAPT)

The *Intermediate Finetuning for Weakly Supervised Sub-domain Adaptation (InterDAPT) Framework* is a training paradigm that allows pre-trained language models to adapt to small, weakly labeled domain-specific datasets without requiring further human annotation to correct or add labels. InterDAPT is comprised of three training components - DAPT for Clinical Domain Adaptation on Meta-domain dataset, Intermediate Domain Finetuning using Weak Supervision to allow the model to learn task-specific information on Tangential-domain, and finally downstream finetuning to a variety of Target Sub-domains.

4.1.1 Domain Adaptive Pretraining (DAPT)

The first stage of InterDAPT is to perform Domain Adaptive Pretraining in order to allow a generalized pre-trained model to adapt to the domain of interest. In this work, we make use of BCH-BERT introduced in (Tavabi et al., 2022) which performs DAPT from BioClinical BERT (Alsentzer et al., 2019) using pediatric patient data which allows the system to make use of linguistic information that is unique to pediatric clinical cases. As downstream sub-domain datasets may have very few examples that can reliably represent entity relationships of interest, the more domain-specific the base linguistic model is, the more entity-related information can be learned from the sub-domain finetuning stage.

4.1.2 Weakly Supervised Intermediate Domain Finetuning

This stage leverages any available unlabeled data in a tangentially related domain by annotating them using an off-the-shelf, publicly available NER model. The entity labels generated must be similarly tangentially related domain to the entities to be predicted for the target sub-domains. In our experiments, we generate Radiology labels using SparkNLP (Kocaman and Talby, 2021) as Radiology-related entities are similar to the target sub-domain entities under study. This labeling process produces a noisy dataset with labels of uncertain quality which we refer to hereafter as the OrthoRad dataset. We finetune BCH-BERT using the OrthoRad dataset which creates the intermediate domain finetuning model which we refer to as the InterNER-BERT model.

4.1.3 Target Sub-domain finetuning

In the final training stage of InterDAPT, we continue training the domain-adapted pre-trained BCH-BERT model with weights from InterNER-BERT and use this to fine-tune the noisily labeled target sub-domains.

5 Experimental Setup

In order to evaluate the efficacy of InterDAPT we run our preliminary experiments on datasets of clinical notes described below. As there is significant variability in the baseline performance of target tasks due to variation in the nature of the task and factors such as the number and types of entities, we compare the performance of InterDAPT against

direct finetuning from BioClinicalBERT to the target sub-domain where the test set is cleaned and human-annotated. We report results on datasets for which data cleaned and vetted by domain experts is available at this time.

5.1 Data

To evaluate the applicability of InterDAPT on narrow clinical domains, we use clinical notes from the Orthopedics and Sports Medicine department at Boston Children's Hospital for our experiments.

The meta-domain dataset (used to pretrain BCH-BERT) is comprised of clinical notes from patients with at least one visit to any of the Orthopedics and Sports Medicine clinics in Boston Children's Hospital from 2000-2020 (Tavabi et al., 2022). Our intermediate model is trained using operative notes that are labeled using John Snow Lab's SparkNLP Platform (Kocaman and Talby, 2021). The resulting machine-labeled dataset contains 81 Radiology-related Entities and is comprised of roughly 500,000 notes.

The sub-domain datasets are created using scoliosis operative notes (referred to hereafter as the Spine dataset) and pre-operative notes of hip joint-related surgeries (referred to hereafter as the Hip dataset). The Spine dataset is the most robust dataset in our study with roughly 16,000 notes and 13 entities and so we primarily center our results on this dataset. The entities were designed to identify surgical details as well as information about patient status and diagnosis. More details on these entities are included in Table 1. Notably, there is a wide variation amongst entity counts in the dataset. The most common entity appears in 11,800 tokens while the least common one appears in 129 tokens. While some entities such as Level are highly specific to spine-related surgeries, others overlap more broadly with other types of surgical notes (eg. Estimated Blood Loss, Fluid Amount). Entities such as Procedure and Diagnosis are also likely to appear in many other types of clinical notes outside of surgical notes.

The Hip dataset included in our experiments (in a limited capacity) contains some entities that are difficult to differentiate between due to a variety of reasons - the entity definition is broad or ambiguous (eg. Symptom Status), many notes do not contain qualifying contextual information that would indicate that the span should be identified as one among two similar entities (eg. Left Anatomy Value vs. Right Anatomy Value), or two or more entity definitions are overlapping (eg. Symptom, Symptom Status).

5.2 Intermediate Task Size

We evaluate the impact of the size of the intermediate model by producing three variations of InterNER-BERT - InterNER-BERT-1M, InterNER-BERT-3M, and InterNER-BERT-5M - which are trained by limiting the intermediate task size, i.e. the size of the OrthoRad training dataset to 1,000,000, 3,000,000 and 5,000,000 examples respectively. As noted in Table 2 as well as in Section 6, We observe that a larger intermediate task size yields better results.

5.3 Baselines

We set our baseline model as BCH-BERT which is finetuned directly onto our target sub-domain datasets with a cleaned, human-vetted test dataset.

6 Results

Results on the Spine dataset show that the InterDAPT Framework is effective at improving domain adaptation especially as we approach an intermediate task size of 5,000,000. Interestingly, noisy labels seem to provide slightly better baseline results than baseline results with a clean dataset. Notably, we observe a strong initial improvement in results for noisy labels with InterDAPT and further improvements are correlated with increasing intermediate task size.

Our preliminary experiments with the Hip Dataset (Noisy) notes provide an F1 score improvement of 10 points from 70.91 on the baseline BCH-BERT model to 80.23 on a model configuration of BCH-BERT + InterNER-BERT-1M. This increase in performance could indicate that intermediate domain finetuning is much better at distinguishing between ambiguous entities. During our error analysis, we found that the baseline model without InterDAPT had relatively more issues with distinguishing between entities that were very similar presentations. However, further experiments with cleaned datasets are required to confirm this hypothesis.

7 Future Work

While our experiments show that InterDAPT is an effective framework that helps ameliorate issues stemming from noisy labeling, more exper-

Entity	Description	Frequency
Procedure	Name of the surgical procedure	11800
Diagnosis	Description or name of diagnosis	9765
IONM Tech	Intraoperative Neuro-monitoring technique	3986
IONM Outcome	Outcomes of intraoperative neuro-monitoring procedure	1675
Level	Vertbrral level being evaluated	4048
Bone graft	Bone graft type used for the surgery	3821
Intraop Imaging	Intraoperative imaging technique used to confirm implant position	1401
Estimated Blood Loss	Surgeon's estimated blood lost during surgery	583
Navigation	Navigation technique used during the surgery to guide implantation	993
Fluid	Type of IV fluid used during the surgery	620
Fluid Amount	Amount of the IV fluid used during the surgery	331
Complications	Any surgical complications	559
Pelvic Fixation	Type of pelvic fixation used during the surgery	129

Table 1: Entities in the Spine Dataset. Frequencies refer to the number of tokens labeled as the corresponding entity.

Model	Target Sub-domain	F1
BCH-BERT (base)	Spine-Clean	94.36
BCH-BERT + InterNER-BERT-1M	Spine-Clean	94.88
BCH-BERT + InterNER-BERT-3M	Spine-Clean	94.96
BCH-BERT + InterNER-BERT-5M	Spine-Clean	**95.02**
BCH-BERT (base)	Spine-Noisy	94.74
BCH-BERT + InterNER-BERT-1M	Spine-Noisy	95.48
BCH-BERT + InterNER-BERT-3M	Spine-Noisy	95.42
BCH-BERT + InterNER-BERT-5M	Spine-Noisy	**96.37**

Table 2: ***Preliminary Results for InterDAPT Framework on Spinal Operative notes:*** The baseline models are denoted by BCH-BERT (base). We experiment using the InterDAPT Framework with BCH-BERT as the base DAPT model combined with InterNER-BERT models as the Intermediate Domain Fine-tuning models. The variations in InterNER-BERT models are produced by training using varying intermediate task sizes - 1,000,000, 3,000,000, and 5,000,000. These models are fine-tuned further on the Spine Target Sub-domain which is sub-categorized as Spine-Clean and Spine-Noisy which refer to versions of the same dataset which are cleaned and noisy respectively. The data cleaning is performed by human domain experts.

iments are needed to show that this framework is applicable to a multitude of different domains. In continuing work, we aim to expand to more sub-domains as we continue to obtain high-quality human-annotated data to compare the relative performance of InterDAPT against.

Limitations

While InterDAPT can be used as a strategy to reduce the negative impact of weak labeling in real-world use cases, it is difficult to understand the magnitude of performance improvements that can be achieved using InterDAPT as these improvements are highly dependent on how noisy the target sub-domain datasets are and how robust the target entity labels are. Despite potentially reducing data annotation costs, InterDAPT still has data requirements that are domain-specific to some extent. While such data can be obtained from publicly available datasets, those tend to be less noisy than real-world data. As this work does not explore the impact of the amount and nature of noise in data, it is unclear at this time how this framework would perform when cleaner datasets are used in prior stages of training.

Acknowledgements

We would like to acknowledge funding support from the Children's Orthopedic Surgery Foundation and hardware support from NVIDIA.

Ethics Statement

This work was performed with approval from the IRB of Boston Children's Hospital.

References

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. Publicly available clinical bert embeddings.

Ting-Yun Chang and Chi-Jen Lu. 2021. Rethinking why intermediate-task fine-tuning works.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks.

Veysel Kocaman and David Talby. 2021. Spark nlp: Natural language understanding at scale. *Software Impacts*, page 100058.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Jason Phang, Thibault Févry, and Samuel R. Bowman. 2019. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. Intermediate-task transfer learning with pretrained models for natural language understanding: When and why does it work?

Nazgol Tavabi, Marium Raza, Mallika Singh, Shahriar Golchin, Harsev Singh, Grant D Hogue, and Ata M Kiapour. 2022. A natural language processing pipeline to study disparities in cannabis use and documentation among children and young adults a survey of 21 years of electronic health records. *medRxiv*, pages 2022–10.

Lester Phillip Violeta, Ding Ma, Wen-Chin Huang, and Tomoki Toda. 2022. Intermediate fine-tuning using imperfect synthetic speech for improving electrolaryngeal speech recognition.

Yue Yu, Simiao Zuo, Haoming Jiang, Wendi Ren, Tuo Zhao, and Chao Zhang. 2020. Fine-tuning pretrained language model with weak supervision: A contrastive-regularized self-training approach. *ArXiv*, abs/2010.07835.