

Exploring Knowledge Transfer in Clinical Natural Language Processing Tasks

Stanford CS224N Custom Project

Wenji Zhang

Department of Computer Science
Stanford University
wenjiz@stanford.edu

Abstract

Natural language processing (NLP) tasks in clinical domain generally involve information extraction, text classification, and text summary from clinical notes or electronic health records. However, due to the limited data resource, many NLP tasks in clinical domain have not been exhaustively studied as of the general domain NLP tasks. Transfer learning has achieved great success in many NLP applications and it especially benefits in applications where there is shared knowledge between the sub-tasks while data is limited. Therefore, this project aims to gain insights on knowledge transfer in multiple clinical NLP tasks and analyze the impact of joint learning on the performance of individual task. Specifically, our main contributions are twofold: (1) we train a multi-task model based on clinical-BERT on varieties of NLP tasks including named-entity recognition (NER), sentence entailment, and text classification, and analyze the performance of the model with different tasks settings; (2) we train a NER model with different entity label annotations and investigate whether there exists knowledge transfers between different entity labels within the same dataset. Our results show that multi-task model achieves improved result on tasks that have shared knowledge (e.g., same task type or similar data distribution) and adding different entity annotations can benefit model performance on named-entity extraction. These fundamental findings provide directions on how to utilize transfer learning to improve clinical domain NLP applications.

1 Introduction

Since BERT [1] was first introduced in 2018, research works on contextual word embedding flourished and has been extended to many other domains - for example, SciBERT [2] is based on scientific literature corpus and FinBERT [3] is for financial service corpus. There are several research works have explored contextual embedding models in clinical and biomedical domain. BioBERT [4] trains a BERT model on a corpus of biomedical research articles from PubMed and PubMed Central and found improved performance on several biomedical natural language processing (NLP) tasks. As clinical narratives have known differences in linguistic characteristics from both general text and non-clinical biomedical text, [5] proposes a pretrained BERT on clinical domain corpus including clinical notes and patient discharge summaries and it has achieved performance improvement on many clinical NLP tasks.

Transfer learning has been widely adopted in NLP and other deep learning research fields. It is based on the idea that utilizing knowledge learned from one task to another related task and it especially helps improve model performance on tasks with limited data annotations. The most common approach for transfer learning is fine-tuning - e.g., BERT model is pretrained on language modeling task, and then it can be fine-tuned for another specific task such as sentiment analysis, question answering, etc. The knowledge acquired during pretraining is ‘transferred’ to the downstream tasks via pretrained weights and the downstream tasks generally benefit from transfer learning. Another approach of

transfer learning is multi-task learning (MTL). Instead of sequentially transfer knowledge as the fine-tuning method, it aims to learn multiple tasks in parallel while keeping a shared representation [6]. It has achieved success in many NLP applications (e.g., [7]) and it particularly benefits in applications where there are shared knowledge between sub-tasks but data is limited for each task.

In clinical field, due to the restricted access of patient data and domain expert required labeling process, annotated data resource is extremely limited compared to general domain NLP tasks. There are several community shared annotated datasets such as i2b2¹ and PhysioNet² but most of the datasets are in modest size and they are annotated with a different focus. With the aforementioned characteristics of transfer learning and the existing challenges in clinical NLP field, this work intends to experiment and analyze different types of knowledge transfers in several clinical domain NLP tasks. In summary, our contributions are the following:

- We train a multi-task model based on Clinical BERT [5] and the model architecture is inspired by [8]. We perform extensive experiments on several clinical tasks and analyze the impact of transfer learning across different task types. The experimented tasks includes: (1) named-entity recognition (NER) with i2b2 2010 [9], i2b2 2012 [10], i2b2 2014 [11], n2c2 2018 [12] and Quaero dataset [13]; (2) sentence entailments with MedNLI [14] and MedRQE [15]; (3) text classification with MIMIC-CXR [16] and annotated by CheXBert³ [17].
- We train a NER model based on Clinical BERT and investigate knowledge transferring between different entity annotations. The datasets used is i2b2 2010 [9] which originally is labeled for three entities: *problem*, *treatment*, *test*. We then experiment and analyze whether having *treatment* and *test* labels can benefit the performance of detecting *problem* entity and vice versa. This type of knowledge transfer has not been investigated in the aforementioned literature and it can provide valuable guidance on practical applications as annotations are costly to obtain.

The rest of the paper is organized as follows: we present related work on transfer learning and MTL in clinical domain in Section 2; model architecture is described in Section 3; we present our experiments, results and results analysis in Section 4. Finally, we conclude the work and discuss future directions in Section 5.

2 Related Work

There are several prior research works on transfer learning in the biomedical and clinical domain. Some earlier works such as [18, 19] mainly focus on knowledge transfer between different data distribution for the same task. However, since those works are conducted before large pretrained language models such BERT was released, the result comparison is no longer compatible with current state-of-the-art (SOTA) results. A more recent work [20] focuses on transfer learning from pretraining and fine-tuning. They release a BERT model that is pretrained on PubMed and MIMIC-III and evaluated their model on ten benchmarking datasets. Another research area is on MTL for biomedical and clinical domain NLP tasks [21, 8, 22, 23]. [21] proposes BioNER which is a NER model based on BiLSTM-CNN-CRF model with shared Bi-LSTM unit that learns shared features from different datasets. [23] proposes MT-BioNER which is initialized from BERT base model and then fine-tuned on several NER tasks. Specifically, [23] considers the training of different datasets as a multi-task problem. [22, 8] both present an MTL frame which uses BERT model as shared layer and then add a task-specific layer on top of it. Specifically, [8] focus on clinical domain tasks, and the authors propose to use BERT that is initialized from Clinical BERT [5] whereas [22] performs experiments on both clinical and biomedical domain with different BERT variants. The main focus of [8] is to provide an MTL framework rather than extensively experiment impact of different tasks. Inspired by [8], we propose to study the impact of joint learning on different clinical domain NLP tasks. Besides tasks that have been evaluated in [8], we also introduce a new task MIMIC-CXR [16] which emphasizes on disease classification based on clinical text. Additionally, we also investigate knowledge transferring between different entity annotations which has not been covered in the previously mentioned research literature.

¹<https://www.i2b2.org/NLP/DataSets/>

²<https://physionet.org/about/database/>

³We would like to extend our thanks to Akshay (akshaysm@stanford.edu) for sharing the annotated data.

3 Approach

We leverage the MTL model architecture proposed in [8]. [8] develops a multi-task model based on clinical BERT with eight task-specific heads. Specifically, they build 5 NER heads, 2 entailment heads, and 1 semantic textual similarity (STS) head. In this work, due to the data accessibility restriction, we remove STS head and add a head for text classification task. Our model architecture is shown in Figure 1.

- **NER head:** The NER head contains a fully connected layer with softmax, the output is similar as multi-class classification output with classes defined as $B - entity$, $I - entity$ for beginning and inside of each labeled entity, and O for outside word. The NER head is applied on hidden state for all tokens and they share the same weights. The loss used here is cross-entropy loss.
- **Entailment head:** The textual entailment head is a multi-class classification layer on the hidden state corresponding to the [CLS] token. It is used to determine if sentence pairs are logically entailed. The output classes are *positive*, *neutral* or *negative* and the loss function is cross-entropy loss.
- **Classification head:** The text classification is used to predict if a disease evidence presents in the given texts. Similar as the entailment head, it adds a classification layer on the hidden state corresponding to [CLS] token. The output classes are *blank*, *positive*, *negative* and *uncertain*.

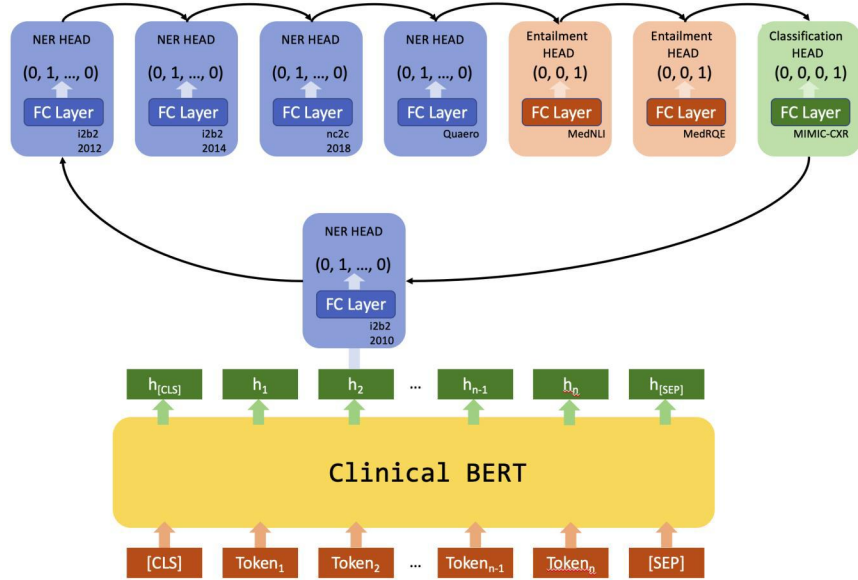


Figure 1: Eight headed MT Clinical BERT with a round robin training schedule.

In order to incorporate loss function from different task head, [8] propose a training scheme that can perform gradient update without additional proportional calculations. At each epoch training, all the heads get updates in a round-robin fashion until the tasks with the largest dataset are all sampled. A detailed algorithm implementation can be found in the paper [8]. For our experiment on the entity labels, we leverage the same model architecture with only one NER head. In such case, the model is equivalent to a fine-tuned Clinical BERT model.

4 Experiments

As mentioned before, we will explore two types of knowledge transfer in clinical NLP tasks. In this section, we will first present our experiment on knowledge transfers between entity annotations

in a NER task, then explain our second experiment on MTL for eight different NLP tasks. Both experiments are build upon an open codebase⁴ released by [8].

4.1 Experiment 1: Knowledge Transfer in NER task

4.1.1 Data

For the first experiment, we use a NER dataset i2b2 2010 [9]. The training set contains 170 de-identified discharge summaries from two health facilities and testing data contains 256 discharge summaries. The data is annotated with three entity types: *treatment*, *test* and *problem*. In order to investigate knowledge transfer between different entity annotations, we create three sub-dataset i2b2 2010_treatment, i2b2 2010_test, i2b2 2010_problem with only *treatment*, *test* and *problem* annotation, respectively. For example, for the following text:

She was then started on Heparin with transition to Coumadin (goal INR of 2-3 secondary to h/o bilateral DVT 's).

The original i2b2 2010 dataset annotates "Heparin" and "Coumadin" as *treatment*, "INR" as *test*, "bilateral DVT" as *problem* and all the other words as *O* (representing outside words). In our i2b2 2010_treatment dataset, we only keep the *treatment* label and replace *test* and *problem* with *O*. We repeat the similar process to create i2b2 2010_test and i2b2 2010_problem datasets.

4.1.2 Experimental Details

To study whether there exists knowledge transfer between different entity annotations, we train four NER models where each model is trained on one of the datasets: i2b2 2010, i2b2 2010_treatment, i2b2 2010_test, i2b2 2010_problem. We keep the training hyperparameters the same to make sure the only variance is on the data annotations. We train the model with *learning_rate* 5e-5, *batch_size* 8, *dropout_prop* 0.1, *max_sequence_len* 512, *num_epoch* 20 (we then choose the best epoch results for each task) and *optimizer* Adam. Since the dataset is relatively small, we repeat the experiment three times and take the average evaluation metrics to reduce the impact of randomness.

4.1.3 Evaluation Method and Results

To evaluate the performance of our NER model, we use exact F1 per entity as our evaluation metrics. We denote the four NER models as $NER_{i2b2\ 2010}$, $NER_{i2b2\ 2010_treatment}$, $NER_{i2b2\ 2010_test}$ and $NER_{i2b2\ 2010_problem}$. The results are shown in Table 1.

Entity	$NER_{i2b2\ 2010}$	$NER_{i2b2\ 2010_treatment}$	$NER_{i2b2\ 2010_test}$	$NER_{i2b2\ 2010_problem}$
treatment	88.0 (+0.7)	87.3	—	—
test	89.8 (+1.2)	—	88.6	—
problem	91.0 (+0.2)	—	—	90.8

Table 1: Exact F1 per entity for four NER models

4.1.4 Analysis

Based on the results reported in Table 1, we found that all the three entity annotations get benefits from other tags. Specially, *test* entity received highest improvement (+1.2) while *problem* benefits the least (+0.2). After examining the dataset, we conjecture that one possible reason is due to the distribution of the entities in this dataset. There are only 4608 annotations for *test* whereas the entity *problem* has 7073 labels. Therefore, *problem* entity gain the least information from the other two. This finding provides us some guidance on labeling design for practical application - For example, in order to build a NER model to detect *test* performed from patient charts, instead of only annotating *test* as the entity, it would be helpful to add annotations of other entities such as *problem* if resource allows.

⁴https://github.com/NLPatVCU/multitasking_transformers/

4.2 Experiment 2: Knowledge Transfer in MTL

4.2.1 Data

As mentioned earlier, we train an MTL for eight clinical NLP tasks. The dataset for each task is described below:

- i2b2 2010 (NER task): Please refer to Section 4.1.1 for a full description of the original i2b2 2010 task;
- i2b2 2012 (NER task): i2b2 2012 dataset consists of 310 (190 for training set and 120 for test set) de-identified discharge summaries. The data is annotated for six entity types: *occurrence, problem, clinical department, evidential, treatment and test*.
- i2b2 2014 (NER task): The dataset contains 1,304 (790 for training set and the 514 for test set) de-identified medical record from 296 patients. The data is annotated with 23 entities including patient information (such as name, occupation, age, state, city etc.), doctor and hospital information and other common entities such as date, medical record id, health plan, url, etc.
- n2c2 2018 (NER task): The dataset contains 505 (303 for training set and 202 for test set) de-identified discharge summaries from MIMIC-III clinical care data base. The original dataset is annotated with entities and relations between entities. In this experiments, only entity annotations are used. The data is annotated with nine entity labels: *drug, frequency, reason, ADE, dosage, duration, form, route, strength*.
- Quaero (NER task): The dataset consists of 2,536 (1688 in training set and 848 in test set) french clinical corpus covering medication, patent and medical research articles. The dataset is annotated with 10 entity types including drugs, disorders, chemical, objects, etc.
- MedNLI (sentence entailment task): The dataset consists of 14,049 (12,627 in training set and 1,422 in test set) sentence pairs from MIMIC-III clinical notes. The sentence pair is annotated as *entailment* if the second sentence is entailed by the first one, *contradiction* if the two statements are contradict with each other or *neutral* if none of the above.
- MedRQE (sentence entailment task): The dataset contains 8,890 (8,588 for training and 302 for testing) sentence pairs. One sentence is a clinical related questions and the other sentence is a potential answer. The data is annotated as *true* if the answer is paired with the question or *false* otherwise.
- MIMIC-CXR (classification task): The original MIMIC-CXR dataset contains chest X-rays and free-text radiology reports. In our experiment, we utilize the impression section from radiology reports to predict whether the patient has Cardiomegaly. The data is annotated with four labels: *blank* if the medical condition is not mentioned, *positive* if the report supports that the condition is present, *negative* if report explicitly says condition is not present, or *uncertain* if the report is unsure on the condition. The annotations are generated by CheXBert [17]. The dataset contains 187,674 reports and we split them into 178,290 for training and 9,384 for testing.

4.2.2 Experimental Details

In this experiment, we trained multiple multi-task models with different selections of tasks and datasets. Since the goal of this experiment is to explore knowledge transfer among different tasks through MTL, we also trained eight models for each task individually which can be used as the baseline for comparison. As we have three types of tasks (NER, sentence entailment, and classification) in our collection, we also experimented knowledge transfer within the same task for different datasets and annotation types. For all the experiments, we set *learning_rate* as $5e - 5$, *batch_size* as 8 for NER task and 40 for other tasks, *dropout_prop* 0.1, *max_sequence_len* 512 and *optimizer* Adam. To better explain our experiment setup and results, we group our discussion by each task type below -

NER Tasks As mentioned earlier, we include five NER datasets in our experiments. We trained one model on all five NER tasks and compare it with model trained for each individual task. Due to time constraints, we did not conduct an exhausting pairwise datasets training to demonstrate the impact of each task pairs. However, since we observe that i2b2 2010 and i2b2 2012 dataset contains

similar corpus type (discharge summaries) with certain shared entity types (*problem*, *treatment* and *test*), it worth experimenting whether MTL on these two tasks can improve the model performance.

Text Entailment Tasks There are two sentence entailment datasets included in our experiments. We trained one multi-task model on these two tasks, two separate models for each individual task and compare it with the multi-task model trained on all eight tasks. We use F1 score as the evaluation metrics.

Classification Tasks Since there is only one classification dataset, the experiment is to compare the single task optimized model with the multi-task model that trained on all eight tasks. The evaluation metrics is F1 score.

4.2.3 Evaluation and Results

We use F1 as evaluation metrics for all the tasks. The results for the aforementioned experiments are shown in Table 2. The notation $ST_{optimized}$ represents for optimized single task model and $MT_{some\ tasks}$ denotes a multi-task model with subscript of the tasks that it trained on.

Tasks	$ST_{optimized}$	$MT_{i2b2\ 2010\ \&\ i2b2\ 2012}$	$MT_{5-NER-tasks}$	$MT_{2-entailment-tasks}$	$MT_{8-tasks}$
i2b2 2010	89.8	90.4 (+0.5)	90.4 (+0.5)	—	90.1 (+0.3)
i2b2 2012	83.5	85.1 (+1.6)	85.4 (+1.9)	—	84.4 (+0.9)
i2b2 2014	95.6	—	94.7 (-0.9)	—	93.0 (-2.6)
nc2c 2018	87.6	—	87.7 (+0.1)	—	87.6 (+0)
Quaero	39.1	—	35.7 (-3.4)	—	27.8 (-11.3)
MedNLI	83.5	—	—	83.1 (-0.4)	79.5 (-4.0)
MedRQE	77.8	—	—	79.1 (+1.3)	76.2 (-1.6)
MIMIC-CXR	98.6	—	—	—	98.5 (-0.1)

Table 2: Experimental results: F1 score for all tasks.

4.2.4 Analysis

There are several interesting findings from the MTL results. In [8], they trained a multi-task model with eight tasks and found that the single-task model generally always outperforms the multi-task model. Intuitively, this should be true since the single-task model is solely optimized to solve that exact problem. While we also observe performance degradation in our multi-task model with all eight tasks, our results demonstrate knowledge gains from MTL with carefully selected tasks.

Based on Table 2, there is only one task Quaero that gets significant benefit from single-task training. This is as expected since Quaero is the only task in French that brings a lingual orthogonality with the rest seven English tasks during MTL. For the three tasks i2b2 2014, MedNLI and MIMIC-CXR, multi-task training shows only a minor performance decreasing and the rest four tasks (i2b2 2010, i2b2 2012, nc2c 2018, MedRQE) receive minor or significant improvement from MTL. As expected, we find that MTL works well for tasks that have shared “similarities”. For example, for the multi-task model that trained on i2b2 2010 and i2b2 2012, both tasks gain significantly compared with single-task model. As mentioned in the previous section, we believe this is due to the fact that the two tasks have common data distributions and shared entity types. Therefore, features induced by training on one task are generalizable to the other task. Comparing the results of $MT_{5-NER-tasks}$ and $MT_{2-entailment-tasks}$ with $MT_{8-tasks}$, we find that MTL generally benefits on each individual task if the tasks are carefully selected. Among the five tasks trained in $MT_{5-NER-tasks}$, three of them receive an improvement and two tasks (i2b2 2010 and i2b2 2012) even beat the state-of-the-art results reported in [8]. Similarly, $MT_{2-entailment-tasks}$ contributes a performance boost on MedRQE task with a minor sacrifice on MedNLI. For the MIMIC-CXR dataset, we do not observe a direct impact on its performance via MTL. Since the dataset of this task is relatively large (compared with other tasks) and data distribution is different, a pairwise experiment may help us have a better understanding on how it reacts with different tasks. In terms of the performance degradation on $MT_{8-tasks}$, we suspect the main reason is on the task orthogonality - as discussed in [8], the NER tasks are token-level

classification while MedNLI and MedRQE requires segment level understanding. The comparison between $MT_{5-NER-tasks}$, $MT_{2-entailment-tasks}$ and $MT_{8-tasks}$ also validates the claim.

5 Conclusion and Future Work

The main focus of this work is to experiment and analyze knowledge transfer in multiple clinical domain NLP tasks. We studied two types of transfer learning - one is from entity annotations in NER tasks and the other is from MTL. Based on our experimental results, we conclude that adding additional entity labels generally improves the model performance on NER tasks especially for entities that are less common in the dataset. We also observe that MTL generally helps the performance on individual tasks if the chose tasks have some common properties, e.g., the same task type (for example, NER or sentence entailment, etc.) or similar data distribution. Our multi-task model contributes new state-of-the-art results on several clinical NLP tasks.

There are several limitations of this work that we hope to address here as a potential future improvement direction - first is on the entity annotation experiment, since i2b2 2010 is a relatively small dataset, adding experiments on other NER datasets with more entity labels would lead to a more concrete conclusion. Secondly, for the MTL tasks, we do not run extensive pairwise tasks experiment as [22] to further demonstrates which task benefits or impairs other tasks. Additionally, we believe a comprehensive error analysis on each multi-task model is necessary for a comprehensive insight on the model performance. Despite the existing limitations, and given the limited data resources in clinical domain, we hope this work sheds some light on different types of knowledge transfer present in clinical NLP tasks and how to utilize them in practical applications to improve model performance.

References

- [1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.
- [2] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text, 2019.
- [3] Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models. *CoRR*, abs/1908.10063, 2019.
- [4] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *CoRR*, abs/1901.08746, 2019.
- [5] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.
- [6] Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.
- [7] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 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, Florence, Italy, July 2019. Association for Computational Linguistics.
- [8] Andriy Mulyar and Bridget T. McInnes. Mt-clinical bert: Scaling clinical information extraction with multitask learning, 2020.
- [9] Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, 06 2011.
- [10] Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813, 04 2013.

- [11] Amber Stubbs, Christopher Kotfila, and Özlem Uzuner. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/uthealth shared task track 1. *Journal of Biomedical Informatics*, 58:S11–S19, 2015. Proceedings of the 2014 i2b2/UTHealth Shared-Tasks and Workshop on Challenges in Natural Language Processing for Clinical Data.
- [12] Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12, 10 2019.
- [13] Aurélie Névéal, Cyril Grouin, Jeremy Leixa, Sophie Rosset, and Pierre Zweigenbaum. The quaero french medical corpus: A ressource for medical entity recognition and normalization. In *In Proc BioTextM, Reykjavik*, 2014.
- [14] Alexey Romanov and Chaitanya Shivade. Lessons from natural language inference in the clinical domain. *CoRR*, abs/1808.06752, 2018.
- [15] Asma Ben Abacha and Dina Demner-Fushman. Recognizing question entailment for medical question answering. In *AMIA 2016, American Medical Informatics Association Annual Symposium, Chicago, IL, USA, November 12-16, 2016*, 2016.
- [16] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, 2019.
- [17] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y. Ng, and Matthew P. Lungren. Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using bert, 2020.
- [18] Xinbo Lv, Yi Guan, and Benyang Deng. Transfer learning based clinical concept extraction on data from multiple sources. *Journal of Biomedical Informatics*, 52:55–64, 2014. Special Section: Methods in Clinical Research Informatics.
- [19] Philip Ganchev, David Malehorn, William L. Bigbee, and Vanathi Gopalakrishnan. Transfer learning of classification rules for biomarker discovery and verification from molecular profiling studies. *Journal of Biomedical Informatics*, 44:S17–S23, 2011. AMIA Joint Summits on Translational Science – 2011.
- [20] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: An evaluation of BERT and elmo on ten benchmarking datasets. *CoRR*, abs/1906.05474, 2019.
- [21] Xi Wang, Jiagao Lyu, Li Dong, and Ke Xu. Multitask learning for biomedical named entity recognition with cross-sharing structure. *BMC Bioinformatics*, 20(1):427, 2019.
- [22] Yifan Peng, Qingyu Chen, and Zhiyong Lu. An empirical study of multi-task learning on bert for biomedical text mining, 2020.
- [23] Muhammad Raza Khan, Morteza Ziyadi, and Mohamed AbdelHady. Mt-bioner: Multi-task learning for biomedical named entity recognition using deep bidirectional transformers, 2020.