A Review of Recent Work in Transfer Learning and Domain Adaptation for Natural Language Processing of Electronic Health Records

Egoitz Laparra¹, Aurelie Mascio², Sumithra Velupillai³, Timothy Miller⁴,⁵
¹ School of Information, University of Arizona, Tucson, USA
² Department of Biostatistics and Health Informatics, King’s College London, London, United Kingdom
³ Institute of Psychiatry, Psychology & Neuroscience, King’s College London, London, United Kingdom
⁴ Computational Health Informatics Program, Boston Children’s Hospital, Boston, USA
⁵ Department of Pediatrics, Harvard Medical School, Boston, USA

1 Introduction

The text in electronic health records (EHRs) contains a wealth of information about the status of patients that is not contained in any other source. Natural language processing (NLP) is the sub-field of artificial intelligence concerned with machine understanding of language, and NLP methods have long been promised as a solution to making text information in EHRs usable for downstream tasks.

Most modern NLP methods take advantage of supervised machine learning, where representative datasets must be manually labeled with medico-linguistic annotations in order to train NLP systems. Recent years have seen an increase in the availability of clinical texts annotated for such information. Clinical datasets have been publicly released for standard NLP tasks such as named entity recognition (NER) and relation extraction [3, 4], coreference resolution [5], as well as datasets for directly addressing tasks of clinical interest such as disease classification [6], heart disease risk factors [7], and text de-identification [8]. In addition, the Medical Information Mart for Intensive Care - III (MIMIC) [9] project has enabled accessing a large and continually growing set of de-identified EHR notes from an intensive care unit, creating a resource suitable for methods that require “big data” (e.g., self-supervised pre-training).

While this increase in availability has encouraged clinical NLP methods development, a key question is whether the reported gains in performance reflect true improvements that will generalize to new data. This question is difficult to answer because it seemingly requires that we have multiple datasets for each problem of interest, when it is already difficult to create even a single dataset. However, the alternative is that we do not know if these systems generalize until we attempt to apply them to real problems. If they do not generalize well, we still end up needing to do additional annotations and method development for each dataset.

These are the issues that we address in this survey. Specifically, we delve into the topics of generalizability and adaptability. Generalizability refers to the ability of a method to extrapolate from limited training data in a way that allows it to perform well on diverse test data that may differ in non-trivial ways from the training data. Adaptability, on the other hand, refers to the potential of methods to take some initially trained model...
and make it especially suited for the type of data it will run on at test time. They are not mutually exclusive — training a more generalizable model is desirable even if one adapts it to test data eventually — but they may represent competing priorities in research directions.

Transfer learning is the foundation of many of the recent developments in both adaptable and generalizable methods. In this paradigm (Figure 1), knowledge from tasks, domains or even languages where more data is available is applied to tasks, domains, or languages where data is scarce. We review the 2018-2020 literature on the clinical application of different transfer learning settings, including domain-to-domain (domain adaptation), task-to-task (inductive transfer learning) and language-to-language (cross-lingual learning) (see Figure 2 for a graphical depiction of this taxonomy). Due to their unprecedented achievements in NLP [10], we give special attention to pre-trained transformers, a family of models that are first trained to solve general problems in vast amounts of unlabeled data and subsequently fine-tuned in downstream tasks.

Our findings are that, although pre-trained transformers start to dominate clinical NLP, they are still not optimized for biomedical data, and applying domain adaptation techniques on top of these models is still relatively unexplored. The field benefits from the large MIMIC-III [9] dataset but is limited by all the largest methods being trained on one particular source of data. Finally, there is an encouraging amount of work on non-English languages, but more could be done to leverage knowledge gained on English to other languages.

2 Methods

We searched the digital libraries of PubMed, the Institute of Electrical and Electronics Engineers (IEEE), the Association for Computational Linguistics (ACL) anthology, and the Association for the Advancement of Artificial Intelligence (AAAI) proceedings for publications from 2018-2020 whose titles or abstracts matched the following query:

(domain adaptation OR transfer learning OR generalizability) AND ((medical OR biomedical OR clinical) AND (text OR language)) OR electronic health record

For the cross-lingual work, we extracted a seed set of articles by using Google Scholar’s “Cited by” feature to look up articles that cited a recent survey of clinical NLP works in non-English languages [11]. We reviewed titles and abstracts for relevance, then read full-text of selected articles to understand the work and pull out certain pieces of information (Methods, Tasks, Domains, Languages, Results, and Reproducibility). We also reviewed citations in the full-text articles and
added any relevant work that was not already covered by our earlier searches. We referred to arXiv preprints for those cases where a peer-reviewed publication was not available. Our search resulted in 87 references that were reduced to 55 after the manual filtering.

3 Results

Table 1 shows a high-level quantitative summary of the results of our literature survey.

### 3.1 Domain Adaptation

A major concern of supervised machine learning (ML) is the lack of robustness under domain shift, especially when labelled data is scarce or difficult to obtain. Two recent works, on biomedical relation extraction [12], and psychiatric salient risk indicator prediction [13], showed large drops in out-of-domain performance and concluded that the in-domain data was insufficient. Domain adaptation, or transductive transfer learning, provides a framework to address this problem by transferring the knowledge acquired from a source domain to a target domain for a particular task.

Among the variety of domain adaptation approaches, some focus more on selecting or augmenting target-domain related data. For example, one work [14] pruned and weighted instances from the source domain to adapt a conditional random field (CRF) for the de-identification of psychiatric notes. However, a larger number of works focus on transferring or combining the model parameters trained in different domains. One approach applied an ensemble of classifiers for auxiliary diagnosis trained on multiple domains that were combined using mutual information [15], while another trained a CRF for NER on nursing handover data by adapting the outputs of another CRF trained on the general medical domain [16]. The models can also be trained jointly, as shown in the work that developed an architecture for NER on EHRs with a shared Bi-LSTM and domain specific CRFs [17]. Another method, called adversarial domain adaptation [18], has been one of the best-performing techniques for deep learning architectures. In this approach, an additional domain discriminator is trained together with the target task. In one application of this approach, it was applied on a Bi-LSTM for disease phrase matching [19]. Domain adaptation has also been applied on speech recognition (SR) for doctor-patient conversations [20], by approaching the problem as a machine-translation task, from source to target domain, to correct errors made by off-the-shelf systems.

### 3.2 Multi-Task

Inductive transfer is an alternative family of techniques that share learned representations between different, although usually related, tasks. In the multi-task setting, models are trained in multiple tasks at the same time, generally with specific loss functions. This approach was applied in the clinical domain [21], by training a Bi-LSTM jointly in Speech Tagging and NER to improve the latter in Chinese EHRs. One participant in the MediQA 2019 challenge [22] combined sentence classification, pairwise text classification, text similarity and relevance ranking along with the challenge’s natural language inference task [23]. Multi-task transfer can also be applied for domain adaptation. One approach developed a Bi-LSTM for word segmentation on Chinese medical text, where the main task was trained jointly with an adaptive loss to minimize the distance between the hidden representations of the different domains [24].

### 3.3 Sequential Transfer

In the clinical domain, inductive transfer has been applied by training different tasks sequentially. A system pre-trained a convolutional neural network (CNN) for medical subject heading identification on PubMed indexed biomedical articles and transferred this model to the prediction of International Classification of Diseases (ICD) codes in EHRs [25]. A related approach started from a small set of labeled data, and combined self-training and transfer learning for radiology report classification, leveraging unlabeled data across three different institutions [26].

A common practice consists in transferring pre-trained word embeddings to downstream tasks, for example, by training medical-specific embeddings and applying them to NER [27]. Several techniques, from concatenation to fine-tuning, have been explored to adapt embeddings, trained on both general and medical domains [28]. One approach pre-trained embeddings on the relation extraction task of the Informatics for Integrating Biology & the Bedside (i2b2) 2009 challenge [29] and fed them to neural networks (NNs) for medical term extraction in the same corpus [30].

### 3.4 Approaches Using Pre-trained Transformers

Contextual word embeddings like Embeddings from Language Model (ELMo) [31] or Bidirectional Encoder Representations from Transformers (BERT) [10] have dramatically improved the performance of NLP tasks. While rule-based or classic statistical approaches still remain prevalent in clinical NLP [32], general domain transformers have recently been applied for various tasks including concept extraction, question answering, or relation extraction [33–36].
Many studies use off-the-shelf BERT models [10] pre-trained on general corpora with BooksCorpus [37] and English Wikipedia. Domain adaptation is then carried out by fine-tuning the pre-trained model on the task-specific dataset. This effective transfer learning method, which does not involve any model pre-training, achieved results on par with state-of-the-art at the time of publication.

In an effort to tackle linguistic characteristic differences between general and biomedical domains, several contextual models such as BioBERT [38] were pre-trained on medical literature (PubMed and PMC articles) atop BERT, and made publicly available. When fine-tuned on a downstream task, these generally showed in-line or improved performance compared to general domain models, albeit not across all tasks (38,39). Going one step further to incorporate the specificities of EHR language (misspellings, abbreviations), various clinically-oriented BERT models, such as clinicalBERT [39], medBERT [40] or BEHR [41], pre-trained on clinical records, were recently released. These models were shown to outperform non-clinical ones on a variety of shared clinical NLP tasks [33, 42].

However, EHR models do not always outperform biomedical ones, notably for de-identification tasks [39]. Furthermore, combining biomedical and EHR texts to pre-train contextual models can increase performance, specifically when tested on out-domain data [43]. Consequently, enriching BERT with specific as well as less specific data could potentially improve the generalizability and adaptability of such models (e.g., across different hospital settings), on top of limiting the amount of EHRs required for training. In specific cases, such as clinical negation detection, a version of BERT adapted to the clinical domain with domain adversarial training [18] underperforms BERT-base [44], implying domain adaptation may be harmful if it moves the model parameters too far away from their starting point.

Finally, more “elaborate” methods have been used to extend the fine-tuning process and push benchmark performances further. Examples include the use of active learning on top of pre-trained BERT models [45], complementing the base model with a transfer learning framework [46], or a graph NN architecture [47].

3.5 Cross-lingual Adaptation

A special case of adaptation is in the development of NLP systems for new languages. Researchers developing systems in lower-resourced languages may be able to take advantage of advances made in English. A recent survey looked more generally at work in developing clinical NLP systems for non-English languages [11]; we refer readers to that work for a broader look. In addition to including some work which has been published since that review, we focus on systems that explicitly used some kind of cross-lingual adaptation and describe several dimensions of cross-lingual adaptation that each work uses some subset of. In particular, these include leveraging methods developed on English into new languages, building on English open-source software, using automatic translation methods, leveraging annotation guidelines to create clinical language resources, and using or extending other knowledge resources (e.g., the Unified Medical Language System [48]) in languages other than English.

Several approaches used similar methods as in work on English, but also explicitly mentioned taking advantage of guidelines and standards developed on English in order to create datasets for tasks in other languages. One work leveraged the THYME temporal annotation guidelines [3] to create a dataset of Italian cardiology documents, then trained and evaluated recurrent NN (RNN)-based methods to extract temporal events [49]. Work on de-identification of Dutch clinical text [50] used guidelines from the i2b2/UTHealth shared task [8], then applied RNN-based methods and showed them superior to feature-based methods.

Other work has leveraged software resources, sometimes including model building but mostly focused on the software architecture. Work in Spanish [51] and German [52] has created modules mirroring those in Apache cTAKES and OpenNLP for some important foundational NLP tasks.

One interesting approach was the use of machine translation methods where models between English and a lower-resourced language pair could be leveraged to build resources in a new language. One work used machine translated death certificates from other languages to complement existing data resources, and showed that for the task of coding these certificates with ICD-10 codes, the augmented data resource was superior to just using the one language [53]. Another approach translated radiology reports in Spanish into English in order to process with the English MetaMap [54]. Interestingly, this approach still used language-specific knowledge resources, as they found that translation was improved if they pre-processed the data by expanding Spanish medical abbreviations.

Finally, some of the most valuable cross-lingual efforts relate to the development of new data resources in non-English languages, including knowledge resources and labeled training sets. One effort created a multilingual corpus (German and Spanish) of clinical text by scraping biomedical publications in those languages for clinical case reports [55]. Another effort scraped journal articles, blog posts, and books for biomedical text in Romanian topically related to three medical specialties -- cardiology, diabetes, and endocrinology -- and also added layers of linguistic annotation to facilitate model training [56].

4 Discussion and Conclusion

The advent of pre-trained transformer models has affected clinical NLP by enabling large performance gains, and in some cases, these gains dwarf the painstaking progress made over the previous decade. However, the most recent works have shown that general domain pre-training does not always transfer adequately to the clinical domain due to its highly specialized language. As shown in some of the work mentioned above, this hurdle is being addressed by either incorporating additional adaptation techniques or pre-training domain specific transformers. It is expected that in-domain versions of some of the newest transformer architectures will appear soon, like those learning long-distance dependencies or from multilingual data.

There is still much work to be done in showing that the gains obtained by pre-trained transformers are meaningful to real world use cases. One major concern is that running these large models is computation-
ally expensive and often prohibitive for many institutions. Approaches to obtain smaller models and faster fine-tuning, like distillation [57] or adapter modules [58], should be explored. In addition, advances in domain adaptation and transfer learning should show that they make measurable impact on the kinds of performance that matter (e.g., time savings for clinical researchers, better clinical trial accrual). Finally, it is unclear whether clinical transformers will still require further adaptation to some specialties with large numbers of rare words and tasks lacking training data. How to transfer these models in the zero- or few-shot scenarios is an open research question.

Overall, the amount of work in domain adaptation and transfer learning is limited by dataset availability. Besides the high costs of creating and distributing clinical datasets, the incentives around creating new datasets (e.g., citation metrics) favor creating the first dataset for a new task rather than the n\textsuperscript{th} dataset for an existing task. Therefore, to enable further research, new dataset creation should prioritize the inclusion of heterogeneous data, so that generalizability can be assessed from the start.

Several studies use MIMIC-III [9] as part of either model development or evaluation, and it has proven to be an important resource for providing accessible evaluation benchmarks. Looking forward, more varied types of benchmark datasets and evaluation frameworks will be needed. In particular, MIMIC is often used for pre-training since it is large, while also being used as a benchmark for outcome prediction [59], and this overlap in data likely leads to overestimated performance. New datasets, methods, or resources around developing shareable pre-trained models that do not rely on a single data source would have a major impact.

Despite advances in the use of transfer learning and domain adaptation techniques for clinical NLP, the majority of studies still report work on English. However, the growing body of research in other languages is encouraging, and further work on new languages is made more feasible thanks to these advances. More collaboration between researchers across the language divide would likely accelerate progress in non-English clinical NLP to prevent reinventing the wheel.

References

19. Liu M, Han J, Zhang H, Song Y. Domain Adapta
20. Mani A, Palaskar S, Konam S. Towards Understand-
ing ASR Error Correction for Medical Conversa-
tions. In: Proceedings of the First Workshop on Natural Language Processing for Medical Conversa-


