Leveraging Medical Literature for Section Prediction in Electronic Health Records

Sara Rosenthal
IBM Research
Yorktown Heights, NY, USA
sjrosenthal@us.ibm.com

Ken Barker
IBM Research
Yorktown Heights, NY, USA
kjbarker@us.ibm.com

Jason Zhicheng Liang*
Rensselaer Polytechnic Institute
Troy, New York, USA
liangz4@rpi.edu

Abstract

Electronic Health Records (EHRs) contain both structured content and unstructured (text) content about a patient's medical history. In the unstructured text parts, there are common sections such as *Assessment and Plan*, *Social History*, and *Medications*. These sections help physicians find information easily and can be used by an information retrieval system to return specific information sought by a user. However, it is common that the exact format of sections in a particular EHR does not adhere to known patterns. Therefore, being able to predict sections and headers in EHRs automatically is beneficial to physicians. Prior approaches in EHR section prediction have only used text data from EHRs and have required significant manual annotation. We propose using sections from medical literature (e.g., textbooks, journals, web content) that contain content similar to that found in EHR sections. Our approach uses data from a different kind of source where labels are provided without the need of a time-consuming annotation effort. We use this data to train two models: an RNN and a BERT-based model. We apply the learned models along with source data via transfer learning to predict sections in EHRs. Our results show that medical literature can provide helpful supervision signal for this classification task.

1 Introduction

Electronic Health Records (EHRs) are an important tool used by physicians that contain an abundance of information about each patient. In addition to aiding the physician in providing patient care, EHRs can be used for downstream tasks such as predicting patient outcome, representation learning, and information extraction (Shickel et al., 2017). All of these tasks can be applied to the unstructured text parts of an EHR. Furthermore, they would benefit from the section structure to pinpoint the likely location where the information should be found (e.g. drug codes are more likely to be in the *Medication* section). However, when physicians edit the unstructured text parts of an EHR, there are no set rules for indicating sections and format is not strictly enforced, nor is there any markup (e.g. XML). An added difficulty is that the formatting is not uniform across EHRs. This is even more common when the EHRs come from different hospitals. For example, one EHR may have the section heading "Assessment and Plan" where another may use "A&P". As an added challenge there may also be nested sections, such as information about a specific medication. Being able to predict sections and headers in EHRs automatically is beneficial to physicians. It allows them to find information easily as well as discover errors and omissions in an EHR.

Consider the publicly available EHR note[1] in Figure 1. The note is divided into the 10 sections found in that EHR (*Problems*, *Medications*, *History*, etc.). This example provides insight into why section prediction can be a difficult task. Although most of the headers appear to be bold, there is also plenty of bold text which is not the main header (e.g. see the *History* section). Additionally, in some cases (e.g. *Allergies* section) there is no text under the header at all. This makes it difficult to segment the data appropriately. Finally, although

*Work completed during internship at IBM Research

[1] http://www.ahrq.gov/professionals/prevention-chronic-care/improve/system/pfhandbook/mod8appbmonicalatte.html

Problems
DIABETES MELLITUS (ICD-250.)
HYPERTENSION, BENIGN ESSENTIAL (ICD-401.1)

Medications
PRINIVIL TABS 20 MG (LISINOPRIL) 1 po qd
Last Refill: #30 x 2 : Carl Savem MD (08/27/2010)
HUMULIN INJ 70/30 (INSULIN REG & ISOPHANE (HUMAN)) 20 units ac breakfast
Last Refill: #600 u x 0 : Carl Savem MD (08/27/2010)

Directives

Allergies and Adverse Reactions (! = critical)

OFFICE VISIT

History of Present Illness
Reason for visit: Routine follow up
Chief Complaint: No complaints

History
Diabetes Management
Hyperglycemic Symptoms
Polyuria: no
. . .

Review of Systems
General: denies fatigue, malaise, fever, weight loss
Eyes: denies blurring, diplopia, irritation, discharge
Ear/Nose/Throat: denies ear pain or discharge, nasal obstruction or discharge, sore throat
Cardiovascular: denies chest pain, palpitations, paroxysmal nocturnal dyspnea, orthopnea, edema Respiratory: denies coughing, wheezing, dyspnea, hemoptysis
Gastrointestinal: denies abdominal pain, dysphagia, nausea, vomiting, diarrhea
. . .

Vital Signs
Ht: **64 in.** Wt: **140 lbs.**T: **98.0** degF. T site: **oral** P: **72** R: **16** BP: **158/90**

Physical Exam
General Appearance: well developed, well nourished, no acute distress
Eyes: conjunctiva and lids normal, PERRLA, EOMI, fundi WNL
Ears, Nose, Mouth, Throat: TM clear, nares clear, oral exam WNL
Respiratory: clear to auscultation and percussion, respiratory effort normal
Cardiovascular: regular rate and rhythm, S1-S2, no murmur, rub or gallop, no bruits, peripheral pulses normal and symmetric, no cyanosis, clubbing, edema or varicosities
Skin: clear, good turgor, color WNL, no rashes, lesions, or ulcerations
. . .

Plan
Medications:
HUMULIN INJ 70/30 20 u ac breakfast
PRINIVIL TABS 20 MG 1 qd
Treatment: Will have annual foot exam at next visit.
. . .

Figure 1: Shortened version of a public EHR[1]. Sections that we classify are highlighted in different colors.

Review of systems

a psychiatric review of systems may include screening questions directed at identifying or exploring co morbid psychiatric illnesses or issues e g sigecaps mnemonic or phq 9 for depression generalized anxiety disorder 7 for anxiety digfast mnemonic for mania or specific questioning around psychoses or other psychiatric complaints a full review of systems should attempt to identify and list all of the relevant stressors that may be impacting a patient s function and overall health

Diagnosis

ultrasound remains as one of the only effective ways of prenatally diagnosing larsen syndrome prenatal diagnosis is extremely important as it can help families prepare for the arrival of an infant with several defects ultrasound can capture prenatal images of multiple joint dislocations abnormal positioning of legs and knees depressed nasal bridge prominent forehead and club feet these symptoms are all associated with larsen syndrome so they can be used to confirm that a fetus has the disorder

Medical History

kniest dysplasia is an autosomal dominant condition this means that the person only needs to have one copy of the mutated gene in order to have the condition people with a family history are at a higher risk of having the disease than people with no family history a random mutation in the gene can cause a person with no family history to also have the condition

Physical Examination

the general appearance of patients may vary according to the experienced symptoms the patient may be comfortable or restless and in severe distress with an increased respiratory rate a cool and pale skin is common and points to vasoconstriction some patients have low grade fever 38 39°c blood pressure may be elevated or decreased and the pulse can become irregular if heart failure ensues elevated jugular venous pressure and hepatojugular reflux or swelling of the legs due to peripheral edema may be found on inspection rarely a cardiac bulge with a pace different from the pulse rhythm can be felt on precordial examination various abnormalities can be found on auscultation such as a 3rd and 4th heart sound systolic murmurs paradoxical splitting of the second heart sound a pericardial friction rub and rales over the lung

Figure 2: Sections from Wikipedia articles (*Psychiatric history*, *Larsen syndrome*, *Kniest dysplasia*, and *Myocardial infarction diagnosis*) that are similar to sections that occur in EHRs. Colors match sections types from Figure 1. The four headers map to the classes *ROS, Find, PMH, and Exam.*

medications have their own section, they also appear in the *Plan* section. Other issues that are not exposed in this example include: 1) Section order is not consistent across EHRs, 2) Headers may be missing, 3) Common features of headers (e.g. bold or colon) are not guaranteed to appear.

Prior approaches in EHR section prediction have been specific to one source and have required significant annotation effort. We propose to reduce the annotation burden by augmenting training data using sections in medical literature (journals, textbooks, web content). Section headers in medical literature are often much more consistent than in EHRs, allowing us to identify with high precision a large number of training example sections using a small number of simple patterns. And although the style and content of sections is different from EHRs, our hypothesis is that there is enough similarity in language to help train models to detect sections in EHRs. Figure 2 shows paragraphs from Wikipedia articles whose sections are similar to those found in EHRs. The text tends to be more descriptive of the kinds of things that could be found in an EHR, but there is semantic overlap.

Our approach is not specific to one hospital or software system and uses data where labels are easy to infer without the need of a time-consuming annotation effort. Furthermore, the large amount of data allows us to explore state of the art neural classification approaches. Finally, training from medical literature can aid in identification of common sections in EHRs from different sources, making it possible to share EHR information more easily across medical facilities and insurance companies which may have access to EHRs from multiple providers.

Our work has two main contributions, demonstrating that augmenting automatically labeled

data from medical literature with:

1. a *small amount* of labeled in-domain EHR training data significantly improves prediction in the EHR dataset.

2. labeled EHR data from a *different source* (out-of-domain) significantly improves the transferability of models trained when there is *no labeled* data in the in-domain EHR dataset.

2 Related Work

Prior work in EHR section prediction has focused on the following tasks:

1. *Section detection:* detecting the boundaries of sections; detecting section header text.

2. *Section classification:* assigning a class label to a section or to sentences (a section class label on a sentence indicates the section to which a sentence likely belongs).

Research may concentrate on section detection only (Ganesan and Subotin, 2014; Dai et al., 2015), section classification (with section boundaries assumed to be known) (Li et al., 2010; Haug et al., 2014) or both (Apostolova et al., 2009; Denny et al., 2009; Tepper et al., 2012). In this paper we focus on section-level classification and section classification at the sentence level.

Prior approaches to section prediction include Support Vector Machines leveraging features computed by bi-gram tf-idf vector representations (Apostolova et al., 2009), Hidden Markov Models (HMM) with sections regarded as part of a sequence (Li et al., 2010), Maximum Entropy Classifiers (Tepper et al., 2012), ℓ1-Regularized Logistic Regression (Ganesan and Subotin, 2014), Bayesian models using N-gram features (Haug et al., 2014), and linear-chain Conditional Random Fields (CRF) to determine section headers (Dai et al., 2015). Most of these approaches rely heavily on hand-crafted features that are time consuming to develop and may not easily generalize across EHRs from different sources. To the best of our knowledge no prior work has explored deep learning or transfer learning. Given enough data, deep neural networks can extract useful features automatically. Some prior work is able to capture regularities in section ordering, either by using the Viterbi Algorithm (Ganesan and Subotin, 2014; Li et al., 2010) or beam search (Tepper et al., 2012) during the final section labeling phase, or by incorporating section ordering features (e.g. the class of the section preceding the section to be classified) into the model (Tepper et al., 2012). In this paper, since we mainly focus on transfer learning based on deep learning models, we predict labels for each section or sentence independently, leaving section sequence considerations to future work.

Two previous efforts have explored more than one dataset (Tepper et al., 2012; Ganesan and Subotin, 2014). Both evaluated domain adaptability and found significant reduction in performance across domains. The largest dataset, *i2b2*, has 13,962 expert-labeled sections (Dai et al., 2015). We use this dataset in our experiments as well. Because of the difficulty of annotating training data, some prior work adopts a semi-automated labeling approach (Apostolova et al., 2009; Denny et al., 2009; Li et al., 2010; Ganesan and Subotin, 2014) as we do for our medical literature dataset.

With the exception of the approaches using i2b2, all prior work uses proprietary datasets that are not available publicly making it difficult to compare to earlier approaches. We make our MedLit dataset as well as our extension of the labels for the i2b2 dataset available for research.

3 Data

We chose a set of 11 section class labels based on those used in the prior work discussed in Section 2 and on the most common sections from EHRs in our sources. The 11 classes are:

Allergies (Alrg)
Assessment and Plan (A&P)
Chief Complaint (CC)
Examination (Exam)
Family History (FHX)
Diagnostic Findings (Find)
Medications (Meds)
Past Medical History (PMH)
Personal and Social History (PSH)
Procedures (Proc)
Review of Systems (ROS)

For our experiments we use three datasets: Medical Literature (MedLit), Cleveland Clinic EHRs (ClvC), and i2b2 EHRs (i2b2).

Source	# Sections	# Sentences
Textbooks	807	7,407
Guidelines	4,781	53,013
Wikipedia	2,658	24,830

Table 1: Sections and sentences in each MedLit source.

3.1 Medical Literature (MedLit)

The medical literature dataset consists of passages from textbooks, guidelines, and a subset of medically relevant Wikipedia articles[2]. The number of sentences per source type is shown in Table 1. In total there are four sources in this dataset: Wikipedia and licensed content from DynaMed, Elsevier, and Wiley publishers.

To generate training data, we created a short list of common phrases for each of the 11 section classes by looking at labels used in prior work as well as section headers labeled in our annotated EHR datasets. We extracted sections in the MedLit corpus whose headers (indicated by XML markup) matched these phrases. For example, any section whose header matched "Chief Complaint", "Reason for Admission", "CC", "Signs and Symptoms", "History of Present Illness", or "SUBJ" was extracted as a positive example of the *Chief Complaint* class[3]. We experimented with partial phrase matches but ultimately decided to use exact match with stemming for higher precision on the common classes. For less frequent classes (e.g. *Allergies, Family History, Personal and Social History*), we allowed additional words to be present in the title. We make the 2,658 sections and labels for the Medical Wikipedia dataset available publicly for research purposes. Examples of sections and their headers are shown in Figure 2, with corresponding class labels indicated in the caption. In the discussion section (Section 5.3) we analyze the quality of the MedLit dataset as a cheaply labeled resource.

3.2 I2b2 EHRs (i2b2)

We use the i2b2 Risk Factors dataset (Stubbs and Uzuner, 2015; Stubbs et al., 2015) which was annotated for section header boundaries (Dai et al., 2015). The annotations do not indicate section class so we matched the headers to our 11 classes. We used partial matches for this data due to the variability of EHR header text. There were 743 unique headers in the Dai et al. (2015) dataset that map to our 11 headers.

	Source	# Sections	# Sentences	Ratio
Train	MedLit	7042	72191	10
	i2b2	1680	10093	6
	ClvC	294	3467	12
Dev	MedLit	1204	13059	11
	i2b2	1591	9373	6
	ClvC	404	3282	8
Test	i2b2	3098	19110	6
	ClvC	404	4046	10

Table 2: The number of sections and sentences as well as the ratio of sentences to sections in each of the three datasets for train, dev, and test sets.

3.3 Cleveland Clinic EHRs (ClvC)

The Cleveland Clinic dataset consists of 178 deidentified patient notes from 54 patients acquired through a research collaboration agreement with Cleveland Clinic. The notes were annotated by two medical students in prior work. Inter-annotator agreement was computed on the first 34 notes (containing 106 sections) annotated by both annotators. The κ score was 0.86 for the sections and 0.80 at the note level. Due to the high agreement between the annotators, the remaining notes were annotated by one annotator each. We manually mapped the section class labels from those annotations to our 11.

3.4 Data Splits

The MedLit dataset was split 80/20 by section for training and tuning. We do not test on MedLit as our goal is finding sections in EHRs. The ClvC EHR dataset was split by patient into 60% for training and tuning, and 40% for testing. For i2b2 (Stubbs and Uzuner, 2015; Stubbs et al., 2015) we use Set 1 and 2 for training and development, and the i2b2 Test Set for testing. The distribution of the train, development/tuning (dev), and test is shown in Table 2. MedLit is the largest dataset. It has more than twice as many sentences as i2b2 and 8 times as many as ClvC. For some of the classes (e.g. *Allergies* and *Personal and Social History*) MedLit data is harder to find. *Procedures* is the smallest category in i2b2 (102 sentences) and *Allergies* is the smallest category in ClvC with

[2]Articles under the 'Clinical Medicine' category (https://en.wikipedia.org/wiki/Category:Clinical_medicine).

[3]It is common for *History of Present Illness* to be its own section but it is often part of or interchangeable with *Chief Complaint* so we opted to combine these two sections.

only 35 sentences in the training data.

4 Method and Results

Our work addresses the scenario of an EHR dataset (a *target* dataset) with little or no training data for section classification. We would like to measure how well models trained on a different EHR dataset and/or on medical literature (*source* datasets) can be transferred to classify the target dataset. We use each of the ClvC and i2b2 datasets in turn as target, with the other as source. We also experiment with using the MedLit dataset as source, both alone and together with data from a source EHR dataset. Specifically, for each target EHR dataset we compare the following models:

- ***ClvC:*** train on labeled data from ClvC EHRs
- ***i2b2:*** train on labeled data from i2b2 EHRs
- ***MedLit:*** train on medical literature data only
- ***MedLit + TR ClvC:*** take the MedLit model, then continue to train (transfer) on labeled ClvC data
- ***MedLit TR i2b2:*** take the MedLit model, then continue to train (transfer) on labeled i2b2 data

In all cases we evaluate our model on our two test sets: ClvC and i2b2. The data used in the test set is considered the **target**, or **in-domain** dataset. The other dataset(s) are considered **source**, or **out-of-domain**, for that experiment and will differ depending on the experiment. We consider the model that trains and tests on the same dataset to be an upper bound (UB). Given enough data[4], we would expect this configuration to perform the best.

We show our results with two different approaches. The first is a Recurrent Neural Network (RNN) using Gated Recurrent Unit (GRU) cell (Cho et al., 2014) and attention mechanism (Bahdanau et al., 2015). GRU aims to solve the vanishing gradient problem revealed in standard RNNs. In our model, attention is used to generate a weighted sum of GRU cell outputs for each word in the input text, for predicting the classification label (rather than only using the output from the last cell). The motivation is to let the model focus on those words that are the most useful for prediction, especially for long input text. The weights are computed by the soft alignment scores between each of the outputs and the last output of the RNN.

The second approach is based on BERT (Devlin et al., 2019), a state-of-the-art language representation model that pre-trains bidirectional representations by jointly conditioning on both left and right context. Once pre-trained, a BERT model can be fine-tuned to specific tasks. In our setting, we take the output of the transformer for the first token in the input, i.e., the special `[CLS]` word embedding, as the representation of the input, which is then used for label prediction by feeding into a classification layer.

We also experimented with a Convolutional Neural Network text classification model (Kim, 2014) as well as traditional machine learning models (Naive Bayes and SVM) using n-gram features, but all performed worse than our GRU RNN and BERT.

For the GRU RNN, we use the Adam optimizer, a batch size of 32, dropout of 0.2, and embedding size 300. We experimented with other parameter values on the development set, but these worked best. We ran each model for 50 epochs—enough for the training loss to converge. For our BERT experiments we use a PyTorch implementation[5] with the bert-base-uncased model. We use the default BERT parameters including the BERT Adam optimizer, a batch size of 32, dropout of 0.1, and embedding size 768. All text is cut off to the first 128 word-pieces. We experimented with different numbers of epochs, and chose the model that performed best on the dev set (usually one tuned at 10 epochs or fewer). Statistical significance was computed using McNemar's test. We experimented with both section classification and sentence classification as described in the following subsections.

Our transfer setting follows the pre-training and fine-tuning approach: after training on the large source domain (MedLit), we continue to tune the model on a small amount of labeled data from the target domain. We follow this approach using our RNN as well as the BERT model where we first tune BERT to the MedLit data and then continue to tune that model on the EHR data. We also experimented with using subsets of the MedLit dataset, downsampling, and class balancing. These exper-

[4] In the case of ClvC we do not have enough annotated data to train a model that would be considered a good upper bound

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

Experiment	RNN F1	BERT F1
MedLit	0.47	0.55
ClvC (UB)	0.59^{+}	0.89
MedLit + TR ClvC	**0.78***	**0.90**
i2b2	0.71	0.83
MedLit + TR i2b2	0.65	0.78

(a) Testing on ClvC

Experiment	RNN F1	BERT F1
MedLit	0.68	0.76
i2b2 (UB)	**0.95**	**0.99**
MedLit + TR i2b2	0.93	**0.99**
ClvC	0.53^{+}	0.84^{+}
MedLit + TR ClvC	0.81^{*}	0.92^{*}

(b) Testing on i2b2

Table 3: Section-Level results for testing on the (a) ClvC and (b) i2b2 EHRs. F1-scores are micro averages across all classes. The best results in each column are highlighted in bold. The results are grouped by MedLit, in-domain ((a) ClvC, (b) i2b2), and out-of-domain ((a) i2b2, (b) ClvC) training. * is significantly higher than + at $p <= .01$.

iments resulted in no significant change or a negative change to the results[6], so we report results using the complete, raw MedLit dataset. We expect that the models learn which data sources are more useful.

4.1 Section Classification

Our first set of experiments follow the approach of prior work (Li et al., 2010; Haug et al., 2014) in classifying sections of the EHR. The section header is included in the data if it exists. As noted in prior work, this task is easier than sentence classification as it is common for the header to be discoverable with accuracy $> 90\%$ for most categories. We find similar trends in the i2b2 data with the upper bound (train on i2b2, test on i2b2) achieving an average 95% F-score with the RNN and near perfect F-score of 99% with BERT.

Results of experiments using the Cleveland Clinic (ClvC) test set as target are shown in Table 3a. The ClvC upper bound using the RNN (59% average F-score) is much lower than is typical in prior work, underlining the difficulty of even section-level classification with small amounts of training data (294 sections). Using the RNN model trained on MedLit alone does quite poorly, but tuning on MedLit prior to training on ClvC improves results very significantly over ClvC training alone (average F-score of 78%). BERT improves the results significantly here with 89% F-score, but the 90% F-score when tuning on the MedLit data is not significantly better. The bottom rows of the table show the scenario where no in-domain (ClvC) data exists, so we train on out-of-domain (i2b2) data. In this case, training on i2b2 alone is better than pre-training first on MedLit. When there is enough EHR data (even from a different source), the MedLit data does not help.

Table 3b shows results using the i2b2 test set as target. With the large amount of in-domain i2b2 training data available, using MedLit does not help. The bottom rows of Table 3b shows performance if we did not have i2b2 training data. Using the out-of-domain (ClvC) training data performs poorly. Transferring MedLit with a very small amount of out-of-domain (ClvC) data significantly outperforms using just the ClvC data for both the RNN and BERT, improving average F-score from 53% to 81% for the RNN and from 84% to 92% for BERT.

4.2 Sentence Classification

Prior work on section-level classification assumes that the section segmentation and header are known. In practice, this is not always the case. It is common for headers to be missing or unclear. In our experiments, therefore, we don't assume to know where the header is and instead of trying to classify a section we classify all sentences in the EHR. For training data for this task we take the class label annotated on each section in our datasets and attach it to each sentence in the section. So each sentence in the section (including the header) is considered an instance of the class. The sentence predictions could be combined to provide section level boundaries. We discuss this further as future work in Section 6. In the MedLit training data, we exclude sentences that are too small (fewer than 15 characters) or too large (more than 400 characters). As with our section-level experiments (Section 4.1), we evaluate our performance with the ClvC and i2b2 test sets. We first tuned the MedLit model using the MedLit development set as described in Section 4. We found that with BERT, only one epoch was necessary, with more epochs resulting in over-fitting.

[6]There is some change in the performance when using Wikipedia only, but not in the overall trends

Experiment	RNN F1	BERT F1
MedLit	0.30	0.37
ClvC (UB)	0.61^{+}	0.73^{+}
MedLit + TR ClvC	**0.62***	**0.74***
i2b2	0.56	0.62
MedLit + TR i2b2	0.53	0.60

(a) Testing on ClvC

Experiment	RNN F1	BERT F1
MedLit	0.30	0.41
i2b2 (UB)	0.64^{+}	**0.71**
MedLit + TR i2b2	**0.65***	**0.71**
ClvC	0.48^{+}	0.59^{+}
MedLit + TR ClvC	0.52*	0.60*

(b) Testing on i2b2

Table 4: Sentence-Level results for testing on the (a) ClvC and (b) i2b2 EHRs in Micro Avg F1-score. The best results in each column are highlighted in bold. The results are grouped by MedLit, in-domain ((a) ClvC, (b) i2b2) and out-of-domain ((a) i2b2, (b) ClvC) training. * is significantly higher than + at $p <= .01$.

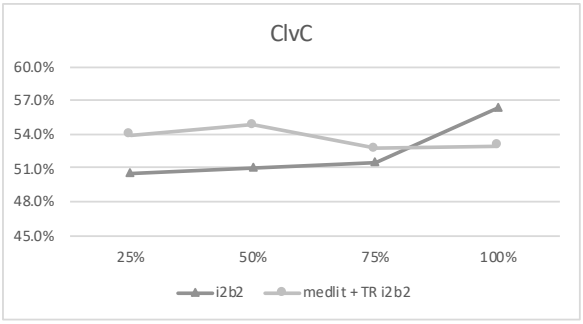

Figure 3: F-score trends of our RNN model training on smaller portions of the i2b2 data for the i2b2 and MedLit + TR i2b2 experiments and testing on ClvC

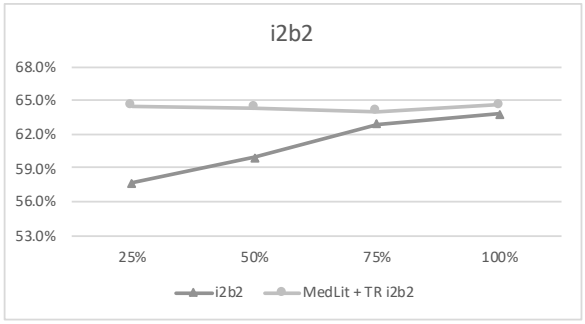

Figure 4: F-score trends of our RNN model training on smaller portions of the i2b2 data for the i2b2 and MedLit + TR i2b2 experiments and testing on i2b2

The results for testing on the ClvC data are shown in Table 4a. Although we call the configuration where we train and test on ClvC data the "upper bound", the MedLit + TR ClvC performs significantly ($p <= .01$) better for the RNN (62% vs 61%) and BERT models (74% vs 73%). We believe that this is because the ClvC dataset is small so the upper bound is not achieved. On the other hand, because the i2b2 training dataset is large, training on the MedLit data first does not provide a significant improvement with either model. We analyzed this result further by experimenting on training with subsets (25%, 50%, and 75%) of the i2b2 training data and testing on ClvC for the i2b2 and MedLit + TR i2b2 experiments. Figure 3 shows that with less out-of-domain EHR training data, transfer learning does help significantly (p-value $<= .01$) in all cases for the RNN model. In fact, there is no significant difference between using 25% or 100% of the i2b2 training data in the MedLit + TR i2b2 experiment.

The sentence-level results for testing with i2b2 as target are shown in Table 4b. Using the MedLit data in the RNN provides a significant improvement (65% vs 64%). However, there is no difference in the BERT models, likely because of the a large amount of i2b2 training data. Here again, we explored how having a smaller subset of the in-domain training data would impact the results. We found that running the MedLit + TR i2b2 experiment with just 25% of the i2b2 dataset with the RNN model performs as well (no significant difference) as training on the full i2b2 dataset (Figure 4), and 12% better than the model using 25% of the i2b2 data alone. In contrast, when using the BERT model, MedLit + TR i2b2 only performs better than i2b2 alone if less than 25% of the training data is available. This indicates that BERT can compensate for the lack of data while the RNN cannot, and MedLit is needed to improve the model. In the scenario where no in-domain training data is available we show that MedLit + TR ClvC does significantly better than just ClvC when testing on i2b2 for the RNN model (52% vs 48%) and BERT (60% vs 59%). We discuss these results further in the next section

Experiment	F1@1	F1@2	F1@3
RNN ClvC	0.48	0.66	0.75
RNN MedLit + TR ClvC	0.52	0.69	0.79
BERT ClvC	0.59	0.75	0.82
BERT MedLit + TR ClvC	0.60	0.75	0.83

Table 5: Average F1 score for training on ClvC and testing on i2b2 when examining the top 3 predictions.

5 Discussion

5.1 How much do out-of-domain EHRs help?

One of our most interesting results using the full datasets is the improvement we are able to achieve training with a small amount of ClvC (out-of-domain) EHR data and testing on i2b2 as the target domain (the last two rows of Table 3b and Table 4b). All of these results on the sentence level are quite low due to similar text that may be found in multiple sections. For example, the text found for *Chief Complaint* and *Past Medical History* can be quite similar with the only difference being *when* the problems occurred. To determine if this kind of misclassification was common we looked at our performance based on whether the correct class was in the top 1, 2, or 3 predictions (F1@3). Both ClvC and MedLit+TR ClvC showed large improvements for the RNN and BERT models as shown in Table 5 with roughly a 27 point improvement for the RNN models and a 22 point improvement for the BERT models. We analyzed the confusion matrix and the most common misclassifications were to the majority MedLit and ClvC classes (*Assessment and Plan* and *Chief Complaint*).

5.2 Do models impact sections differently?

We also compared performance on the individual classes when transferring from MedLit. To do this, we explored the training on i2b2 and testing on BERT i2b2, and BERT MedLit + TR i2b2 experiments (See the middle rows in Table 4b) where both models performed the same. We found the BERT i2b2 performed better for *Allergies* (6.8%), but BERT MedLit + TR i2b2 performed better for *Family History* (5.8%), and *Procedures* (7.9%). This suggests that it may be worth exploring an ensemble approach.

5.3 What is the quality of the MedLit data?

How well our approach performs on EHR section classification relies on the MedLit data. The MedLit dataset is cheaply labeled data of a different genre than the target data, so we would not expect the MedLit model to perform particularly well on its own on EHR data. To judge the quality of the MedLit-trained model in *its own genre*, we analyzed its performance on the dev set made up solely of medical literature. The best average F-score was 65.2% using the RNN and 71.2% for BERT on the sentence classification task. In particular, we found that *Assessment and Plan* performs quite poorly. We examined the confusion matrix and the predictions are distributed among the other classes indicating that it is more vague. Looking to see if the correct label is at the second- or third-highest scoring predictions, we find the F1@3 is 88.9% for the RNN model and 92.6% for the BERT model. This is significantly better, but still shows that it is weakly labeled data. We also analyzed the performance from our different medical literature sources. This analysis caused us to drop one of our initial guideline sources because it had few useful labels. The two relevant labels it did have were easy to predict correctly causing a bias toward the source. Finally, we inspected the text for the section headers. There were 66 unique headers in the medical literature dev set ranging over the 11 classes. "Physical", which maps to the *Examination* class was the most common section header and was predicted correctly 74% and 79% of the time for the RNN and BERT models respectively. Some other good strings include "History of Present Illness" and "signs and symptoms" for the *Chief Complaint* class and "medication" for the *Medications* class. Our *Review of Systems* headers included common subsections found in this area of the EHR such as "eye", "neurology", and "genitourinary". Some of these did not perform as well, probably because they are likely to appear in many different sections. We also analyzed the confusion matrix to determine the common misclassifications. In most cases the common misclassifications were to *Chief Complaint*, the majority class. We also found that *Past Medical History* was often incorrectly labeled as *Medications*. We found this analysis to be consistent across both models indicating that the data is not model dependent.

6 Conclusion

In this paper we describe a novel approach to classifying sections in Electronic Health Records

when a limited amount of in-domain training data is available. We present a new dataset for EHR section prediction from Medical Literature, of which the Wikipedia part is available to the public for research purposes. We show that combining a very small amount of in-domain EHR data with a large amount of automatically labeled, out-of-domain, out-of-genre Medical Literature data can perform as well as using a large amount of in-domain EHR data at the section and sentence level. We also show that combining out-of-domain, out-of-genre Medical Literature data with out-of-domain EHRs can provide significant improvement over using just out-of-domain EHRs at the section and sentence level, depending on training data size. These results indicate that even though the data in Medical Literature is very different in style, the content can bridge between the domain-specific vocabularies of different EHR systems. We show that our approach can be used to achieve good results on *new unseen EHR datasets* with minimal or even no training data. In the future we would also like to explore using both i2b2 and ClvC together to see if a multi-task learning approach would provide additional improvements. In addition to BERT, we also briefly explored BioBERT (Lee et al., 2019), a BERT model pre-trained on a medical corpus. In our initial experiments BioBERT performed worse than BERT, but we would like to explore this further.

Finally, in this work we focus on individual sentence and section classification and show that we can achieve improvements in this regard. In addition, we could also exploit the structure of the document to provide additional improvements. The structure often follows a pattern in EHRs (for example, *Chief Complaint* tends to be the first section). Prior work has looked at CRFs and HMMs (Li et al., 2010; Dai et al., 2015) to exploit this property. We would like to explore whether we can improve our model by combining it with a model that takes into account trends at the document level. Using LSTM-CRFs (Lample et al., 2016; Huang et al., 2015) as a second level with BERT as pre-training on our model may provide such an improvement.

7 Acknowledgments

We thank the physicians and IT staff at Cleveland Clinic who provided de-identified EMRs under an IRB protocol and the students from New York Medical College and SUNY Downstate Medical Center for their enthusiastic and dedicated annotation work.

References

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

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder – decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

Hongjie Dai, Syed Abdul Shabbir, 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.

Joshua C. Denny, Anderson Spickard, III, Kevin B. Johnson, Neeraja B. Peterson, Josh F. Peterson, and Randolph A. Miller. 2009. Evaluation of a method to identify and categorize section headers in clinical documents. *Journal of the American Medical Informatics Association*, 16(6):806–815.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Peter J. Haug, Xinzi Wu, Jeffrey P. Ferraro, Guergana Savova, Stanley M. Huff, and Christopher G. Chute. 2014. Developing a section labeler for clinical documents. In *AMIA 2014, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 15-19, 2014*.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270. Association for Computational Linguistics.

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. *CoRR*, abs/1901.08746.

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*, IHI '10, pages 744–750, New York, NY, USA. ACM.

Benjamin Shickel, Patrick Tighe, Azra Bihorac, and Parisa Rashidi. 2017. Deep EHR: A survey of recent advances on deep learning techniques for electronic health record (EHR) analysis. *CoRR*, abs/1706.03446.

Amber Stubbs, Christopher Kotfila, Hua Xu, and Özlem Uzuner. 2015. Identifying risk factors for heart disease over time: Overview of 2014 i2b2/uthealth shared task track 2. *Journal of Biomedical Informatics*, 58:S67 – S77. Proceedings of the 2014 i2b2/UTHealth Shared-Tasks and Workshop on Challenges in Natural Language Processing for Clinical Data.

Amber Stubbs and Özlem Uzuner. 2015. Annotating risk factors for heart disease in clinical narratives for diabetic patients. *Journal of Biomedical Informatics*, 58:S78 – S91. Proceedings of the 2014 i2b2/UTHealth Shared-Tasks and Workshop on Challenges in Natural Language Processing for Clinical Data.

Michael Tepper, Daniel Capurro, Fei Xia, Lucy Vanderwende, and Meliha Yetisgen-Yildiz. 2012. Statistical section segmentation in free-text clinical records. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*. European Language Resources Association (ELRA).