Fine-Tuning Bidirectional Encoder Representations From Transformers (BERT)–Based Models on Large-Scale Electronic Health Record Notes: An Empirical Study

Fei Li1,2,3, PhD; Yonghao Jin1, MSc; Weisong Liu1,2,3, PhD; Bhanu Pratap Singh Rawat4, MSc; Pengshan Cai4, MSc; Hong Yu1,2,3,4, PhD

1Department of Computer Science, University of Massachusetts Lowell, Lowell, MA, United States
2Center for Healthcare Organization and Implementation Research, Bedford Veterans Affairs Medical Center, Bedford, MA, United States
3Department of Medicine, University of Massachusetts Medical School, Worcester, MA, United States
4School of Computer Science, University of Massachusetts, Amherst, MA, United States

Corresponding Author:
Hong Yu, PhD
Department of Computer Science
University of Massachusetts Lowell
1 University Avenue
Lowell, MA,
United States
Phone: 1 978 934 6132
Email: Hong_Yu@uml.edu

Abstract

Background: The bidirectional encoder representations from transformers (BERT) model has achieved great success in many natural language processing (NLP) tasks, such as named entity recognition and question answering. However, little prior work has explored this model to be used for an important task in the biomedical and clinical domains, namely entity normalization.

Objective: We aim to investigate the effectiveness of BERT-based models for biomedical or clinical entity normalization. In addition, our second objective is to investigate whether the domains of training data influence the performances of BERT-based models as well as the degree of influence.

Methods: Our data was comprised of 1.5 million unlabeled electronic health record (EHR) notes. We first fine-tuned BioBERT on this large collection of unlabeled EHR notes. This generated our BERT-based model trained using 1.5 million electronic health record notes (EhrBERT). We then further fine-tuned EhrBERT, BioBERT, and BERT on three annotated corpora for biomedical and clinical entity normalization: the Medication, Indication, and Adverse Drug Events (MADE) 1.0 corpus, the National Center for Biotechnology Information (NCBI) disease corpus, and the Chemical-Disease Relations (CDR) corpus. We compared our models with two state-of-the-art normalization systems, namely MetaMap and disease name normalization (DNorm).

Results: EhrBERT achieved 40.95% F1 in the MADE 1.0 corpus for mapping named entities to the Medical Dictionary for Regulatory Activities and the Systematized Nomenclature of Medicine—Clinical Terms (SNOMED-CT), which have about 380,000 terms. In this corpus, EhrBERT outperformed MetaMap by 2.36% in F1. For the NCBI disease corpus and CDR corpus, EhrBERT also outperformed DNorm by improving the F1 scores from 88.37% and 89.92% to 90.35% and 93.82%, respectively. Compared with BioBERT and BERT, EhrBERT outperformed them on the MADE 1.0 corpus and the CDR corpus.

Conclusions: Our work shows that BERT-based models have achieved state-of-the-art performance for biomedical and clinical entity normalization. BERT-based models can be readily fine-tuned to normalize any kind of named entities.

(JMIR Med Inform 2019;7(3):e14830) doi: 10.2196/14830

KEYWORDS
natural language processing; entity normalization; deep learning; electronic health record note; BERT
Introduction

Background

Entity normalization (EN) is the process of mapping a named entity mention (eg, dyspnea on exertion) to a term (eg, 60845006: Dyspnea on exertion) in a controlled vocabulary (eg, Systematized Nomenclature of Medicine—Clinical Terms [SNOMED-CT]) [1]. It is a significant task for natural language processing (NLP) [2]. It is also an important step for other NLP tasks such as knowledge base construction and information extraction [3-6].

EN has been extensively studied in the biomedical and clinical domains [7,8]. Supervised approaches usually perform better than unsupervised approaches. However, their performance depends highly on the quantity and quality of annotated data [1,8-10]. Recently, deep representation-learning models, such as bidirectional encoder representations from transformers (BERT) and embeddings from language models (ELMo), have been shown to improve many NLP tasks [11,12]. These studies usually employ unsupervised pretraining techniques to learn language representations from large-scale raw text.

Deep representation-learning models learn word representations from large-scale unannotated data, which are more generalizable than the models trained only from annotated data with limited sizes. Therefore, deep representation-learning models can be fine-tuned to improve downstream NLP tasks. For example, BERT [11] has achieved new state-of-the-art results on 11 NLP tasks, including question answering and natural language inference. BioBERT [13], which has a similar architecture but was pretrained using PubMed and PubMed Central (PMC) publications, achieved new state-of-the-art results on three biomedical NLP tasks: named entity recognition, relation extraction, and question answering. However, little work has explored such models in biomedical and clinical entity normalization tasks.

Related Work

Previous work has studied various language models. For instance, the n-gram language model [2] assumes that the current word can be predicted via previous \(n\) words. Bengio et al [14] utilized feed-forward neural networks to build a language model, but their approach was limited to a fixed-length context. Mikolov et al [15] employed recurrent neural networks to represent languages, which can theoretically utilize an arbitrary-length context.

Besides language models, researchers have also explored the problem of word representations. The bag-of-words model [16] assumes that a word can be represented by its neighbor words. Brown et al [17] proposed a clustering algorithm to group words into clusters that are semantically related. Their approach can be considered as a discrete version of distributed word representations. As deep learning develops, some researchers leveraged neural networks to generate word representations [16,18].

Recently, researchers have found that many downstream applications can benefit from the word representations generated by pretrained models [11,12]. ELMo utilized bidirectional recurrent neural networks to generate word representations [12]. Compared to word2vec [16], their word representations are contextualized and contain subword information. BERT [11] utilizes two pretraining objectives, mask language model and next sentence prediction, which can naturally benefit from large unlabeled data. The BERT input consists of three parts: word pieces, positions, and segments. BERT uses bidirectional transformers to generate word representations, which are jointly conditioned on both the left and right context in all layers. BERT and its derivatives such as BioBERT [13] achieved new state-of-the-art results on various NLP or biomedical NLP tasks (eg, question answering, named entity recognition, and relation extraction) through simple fine-tuning techniques.

In this paper, we investigated the effectiveness of such an approach in a new task, namely, biomedical or clinical entity normalization. In the biomedical or clinical domain, MetaMap [19] is the tool that is widely used to extract terms and link them to the Unified Medical Language System (UMLS) Metathesaurus [3]. Researchers utilized MetaMap in various scenarios, such as medical concept identification in electronic health record (EHR) notes [20], vocabulary construction for consumer health [21], and text mining from patent data [22]. In this paper, we employed MetaMap as one of our baselines. Previous work consisting of entity normalization can be roughly divided into three types: (1) rule-based approaches [7] depend on manually designed rules, but they are not able to cover all situations; (2) similarity-based approaches [23] compute similarities between entity mentions and terms, but the metrics of similarities highly influence the performances of such approaches; (3) machine learning-based approaches [1,8-10] can perform better, but they usually require enough annotated data to train models from scratch. In this paper, we fine-tuned pretrained representation-learning models on the entity normalization task to show that they are more effective than traditional supervised approaches.

Objective

In this study, we proposed the following objectives:

1. We aimed to explore the effectiveness of BERT-based models for the entity normalization task in the biomedical and clinical domains. The overview of this paper’s methods is shown in Figure 1.
2. We aimed to investigate whether the domains of training data influence the performances of BERT-based models as well as the degree of influence.
Contributions
The main contributions of this paper are as follows:

1. We proposed a BERT-based model that was trained using 1.5 million EHR notes (EhrBERT). To facilitate the research of clinical NLP, the EhrBERT is publicly available at GitHub [24].

2. We evaluated EhrBERT on three entity normalization corpora in the biomedical and clinical domain. EhrBERT improved the F1s in three corpora by 2.36%, 1.98%, and 3.9% compared with state-of-the-art models such as MetaMap and disease name normalization (DNorm). EhrBERT also performed better than BioBERT and BERT in two corpora.

3. By comparing BERT, BioBERT, and EhrBERT, we found that the domain influences the performances of BERT-based models. However, if the domains of models and tasks are close, such an effect is generally not statistically significant. However, if their domains are distant, such an effect becomes large.

Methods
Overview
In this section, we will first describe how to generate the clinical representation-learning model using BERT and EHR notes. Next, the details of the models used for entity normalization will be introduced. Lastly, we will introduce the corpora used in this paper. Throughout this paper, we leveraged the PyTorch implementation of BERT developed by Hugging Face [25] to implement our models.
<table>
<thead>
<tr>
<th>Hyper-parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epoch</td>
<td>15</td>
</tr>
<tr>
<td>Maximal sequence length</td>
<td>128</td>
</tr>
<tr>
<td>Batch size</td>
<td>64</td>
</tr>
<tr>
<td>Learning rate</td>
<td>0.00003</td>
</tr>
<tr>
<td>Embedding size</td>
<td>768</td>
</tr>
<tr>
<td>Dropout probability</td>
<td>0.1</td>
</tr>
<tr>
<td>Transformer blocks</td>
<td>12</td>
</tr>
<tr>
<td>Self-attention heads</td>
<td>12</td>
</tr>
</tbody>
</table>

We utilized 15 epochs to train EhrBERT, which were selected based on prior work [27] and our data size. Based on the average sentence length in our data, the maximal sequence length was set as 128, which is shorter than that used by BERT. The batch size and learning rate were set as 64 and 0.00003, respectively, based on the recommendation settings in BERT. The settings of the hyper-parameters related to the model architecture are identical to those of BERT BASE [11]. Other hyper-parameters, such as the probabilities of masked language model and next sentence prediction, were set as the default values (15% and 50%, respectively). For either EhrBERT500k or EhrBERT1M, we used four Tesla P40 graphics processing units to simultaneously fine-tune BioBERT on our EHR data. EhrBERT500k takes approximately 12 hours per epoch and EhrBERT1M takes approximately 23 hours per epoch.

Models for Entity Normalization

As shown in Figure 2, we treated entity normalization as a text classification task. Following BERT and BioBERT, we employed the word representations from the top layer of transformers as the features for the normalization task. Concretely, a classifier token, [CLS], is padded before the given sequence of word pieces [28]. Thus, our model takes a sequence \([CLS], w_1, ..., w_N]\) as input. Here, \(w_n\) is not necessarily a word; it can also be a subword (aka, a word piece). Each word piece is mapped to a \(d_{emb}\)-dimensional embedding, \(E_n\). In addition, the input also includes segment and position embeddings with the same dimension, \(d_{emb}\), which are mixed with the word piece embeddings.

After a few layers of bidirectional transformers, \(Trm\), each word piece, \(w_n\), corresponds to a \(d_{Trm}\)-dimensional vector, \(T_n\). The \(d_{Trm}\)-dimensional representation, \(C\), for the padding token, [CLS], is used as the representation of the whole sequence. Then \(C\) is input into the SoftMax layer to compute the probability distribution of all classes. The class with the maximal probability is selected as the prediction.

In terms of parameter initialization, the BERT part of the model was initialized with EhrBERT. Other parameters were randomly initialized with a uniform distribution. During training, the objective is to maximize the log-likelihood of gold annotations. We used the standard back-propagation to update all the parameters and the Adam algorithms [29] to control the update process. For hyper-parameter setting, \(d_{emb}\) and \(d_{Trm}\) are set as 768, the batch size is 32, the learning rate is 1e-5, and the dropout rate is 0.1. The training will stop early if the performance has not increased for 20 epochs.
Figure 2. Model architectures. An example of entity normalization is shown and the named entity “dyspnea on exertion” is normalized to the term “60845006” in the Systematized Nomenclature of Medicine—Clinical Terms (SNOMED-CT) vocabulary (SNOMED International, 2019). The size of classes depends on the vocabularies used in a corpus, which is about 380,000 (Medical Dictionary for Regulatory Activities [MedDRA] and SNOMED-CT) for the Medication, Indication, and Adverse Drug Events (MADE) 1.0 corpus and 11,000 (MErged Disease voCabulary [MEDIC]) for the National Center for Biotechnology Information (NCBI) Disease and Chemical-Disease Relations (CDR) corpora. BERT: bidirectional encoder representations from transformers; C: \(d_{Trm}\)-dimensional representation; [CLS]: classifier token; E: \(d_{emb}\)-dimensional embedding; T: \(d_{Trm}\)-dimensional vector; Trm: bidirectional transformer.

Corpora

We employed the Medication, Indication, and Adverse Drug Events (MADE) corpus [30], which derives from the MADE 1.0 challenge. The corpus includes 1089 EHR notes, which were divided into 876 notes for training and 213 notes for testing. This corpus contains the annotations of mapping adverse drug events to the Medical Dictionary for Regulatory Activities (MedDRA) [31] terms and of mapping indications, signs, and symptoms to the SNOMED-CT [32] terms. The MedDRA and SNOMED-CT vocabularies include about 380,000 terms in total, which are computed based on the MRCONSO.RRF file in the UMLS Metathesaurus, version 2016 AA. In the MADE corpus, there are about 35,000 and 8000 mentions in the training and test sets, respectively.

Moreover, we also employed two nonclinical corpora, namely the National Center for Biotechnology Information (NCBI) disease corpus [33] and the Chemical-Disease Relations (CDR) corpus [34], to evaluate EhrBERT in different domains. The NCBI disease corpus consists of 793 PubMed abstracts, 6892 disease mentions, and 790 unique disease concepts. The abstracts are split into 593, 100, and 100 for training, development, and testing, respectively. The CDR corpus is composed of 500, 500, and 500 PubMed abstracts for training, development, and testing, respectively. It includes 5818 disease mentions for normalization. The objectives of both corpora are to map each disease mention to a term in the MErged Disease voCablary (MEDIC) [35], which contains approximately 11,000 terms.

Experimental Settings

For the MADE corpus, we utilized mention-level precision, recall, and F1 as evaluation metrics. A prediction is counted as true-positive only if both the boundary and term ID of the mention are correct. Besides using the gold entity mentions, we also utilized the mentions recognized by MetaMap [19] as the input for our models as a comparison. Because the outputs of MetaMap are the UMLS IDs [3], we also utilized the UMLS Metathesaurus to map SNOMED-CT and MedDRA terms to UMLS terms. During preprocessing, we transformed all the tokens in a mention or a term into lowercase. We also removed the punctuations but kept the numbers.

For the NCBI disease and CDR corpora, we utilized document-level precision, recall, and F1, following DNorm [1]. There are two ID sets for a document, namely the predicted ID set and the gold ID set. If a predicted ID is equal to a gold ID, we counted it as true-positive. The performance of the corpus is the macro-averaged performance of all documents. All the abbreviations are replaced with their full names using the
dictionaries provided by DNorm. We employed gold mentions as input in order to compare with DNorm.

Besides precision, recall, and F1, we also analyzed statistical significances between different models. First, the MetaMap and DNorm were run once on test sets using their off-the-shelf models that were released by the authors. We believe that these models are elaborately tuned and can achieve the best performance as strong baselines. Second, the experiments for BERT, BioBERT, and EhrBERT were run thrice. During each run, we utilized a different random seed to initialize the model. After training, the model was run on the test set to obtain precision, recall, and F1. Lastly, the t-test was utilized to determine if the performances of two models were statistically different based on the results of these runs.

Results

Table 2 shows the F1s and the standard deviations of the models. The models are ranked from low to high based on F1s. Precisions and recalls are provided in Multimedia Appendix 1.

Tables 3-5 show the P values of the different models for the MADE (predicted entities), NCBI disease, and CDR corpora, respectively. The performance of the model along each row is lower than the performance of the model along each column, as shown in Table 2. We utilized .05 as the threshold to determine statistical significance.

The results for entity normalization are shown in Table 2. We ran our experiments thrice for all the models using different random seeds. The results in Table 2 are the mean F1 scores of these runs. We can see that no matter whether we used gold entities or MetaMap-predicted entities in the MADE corpus, EhrBERT performed better than BioBERT, and BioBERT performed better than BERT. In addition, BERT-based models obtained better results compared with MetaMap, improving the F1s by 2.22% for BERT, 2.28% for BioBERT, and 2.36% for both EhrBERT$\text{500k}$ and EhrBERT$\text{1M}$. From Tables 3-5, we can see that EhrBERT performed significantly better than MetaMap, BERT, and BioBERT. However, the performance differences between BERT, BioBERT, and EhrBERT are not always discernible.

In both the NCBI disease and CDR corpora, the F1s of BERT-based models were higher than the F1s of DNorm, as shown in Table 2. In the NCBI disease corpus, BioBERT achieved the highest F1 (90.71%). As shown in Tables 3-5, BioBERT is statistically discernible from BERT but not from EhrBERT. In the CDR corpus, BioBERT performed slightly worse than EhrBERT (93.42% vs 93.82%). In Tables 3-5, there are no statistical differences between BioBERT and EhrBERT$\text{500k}$, but a statistical difference exists between BioBERT and EhrBERT$\text{1M}$. The similar performances of EhrBERT and BioBERT may be because the domains of EhrBERT and BioBERT are close. Moreover, all models performed much better in the NCBI disease and CDR corpora than in the MADE corpus. One likely reason is that the class number of the MADE corpus is tens of times larger than those of the NCBI disease and CDR corpora.

Comparing EhrBERT$\text{500k}$ and EhrBERT$\text{1M}$, EhrBERT$\text{1M}$ consistently performed better in all the corpora as shown in Table 2. This implies that the size of the pretraining data may be a factor that influences the performance of BERT-based models. However, the significance analysis in Table 5 shows that the performance of EhrBERT$\text{1M}$ is only significantly different from that of EhrBERT$\text{500k}$ in the CDR corpus. There are no statistical differences between EhrBERT$\text{500k}$ and EhrBERT$\text{1M}$ in the other two corpora.
Table 2. F1s and standard deviations.

<table>
<thead>
<tr>
<th>Corpus and model</th>
<th>F1 (%), mean (SD)</th>
<th>Improvement compared with MetaMap or DNorm&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MADE&lt;sup&gt;b&lt;/sup&gt; (gold entities&lt;sup&gt;c&lt;/sup&gt;)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERT&lt;sup&gt;d&lt;/sup&gt;</td>
<td>67.87 (0.25)</td>
<td>N/A&lt;sup&gt;e&lt;/sup&gt;</td>
</tr>
<tr>
<td>BioBERT</td>
<td>68.22 (0.11)</td>
<td>N/A</td>
</tr>
<tr>
<td>EhrBERT&lt;sub&gt;500k&lt;/sub&gt;&lt;sup&gt;f&lt;/sup&gt;</td>
<td>68.74 (0.14)</td>
<td>N/A</td>
</tr>
<tr>
<td>EhrBERT&lt;sub&gt;1M&lt;/sub&gt;&lt;sup&gt;g&lt;/sup&gt;</td>
<td>68.82 (0.29)</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>MADE (predicted entities&lt;sup&gt;h&lt;/sup&gt;)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MetaMap [19]</td>
<td>38.59 (0)</td>
<td>N/A</td>
</tr>
<tr>
<td>BERT</td>
<td>40.81 (0.08)</td>
<td>+2.22</td>
</tr>
<tr>
<td>BioBERT</td>
<td>40.87 (0.06)</td>
<td>+2.28</td>
</tr>
<tr>
<td>EhrBERT&lt;sub&gt;500k&lt;/sub&gt;</td>
<td>40.95 (0.04)</td>
<td>+2.36</td>
</tr>
<tr>
<td>EhrBERT&lt;sub&gt;1M&lt;/sub&gt;</td>
<td>40.95 (0.07)</td>
<td>+2.36</td>
</tr>
<tr>
<td><strong>NCBI&lt;sup&gt;i&lt;/sup&gt;</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DNorm [1]</td>
<td>88.37 (0)</td>
<td>N/A</td>
</tr>
<tr>
<td>BERT</td>
<td>89.43 (0.99)</td>
<td>+1.06</td>
</tr>
<tr>
<td>EhrBERT&lt;sub&gt;500k&lt;/sub&gt;</td>
<td>90.00 (0.48)</td>
<td>+1.63</td>
</tr>
<tr>
<td>EhrBERT&lt;sub&gt;1M&lt;/sub&gt;</td>
<td>90.35 (1.12)</td>
<td>+1.98</td>
</tr>
<tr>
<td>BioBERT</td>
<td>90.71 (0.37)</td>
<td>+2.34</td>
</tr>
<tr>
<td><strong>CDR&lt;sup&gt;j&lt;/sup&gt;</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DNorm [1]</td>
<td>89.92 (0)</td>
<td>N/A</td>
</tr>
<tr>
<td>BERT</td>
<td>93.11 (0.54)</td>
<td>+3.19</td>
</tr>
<tr>
<td>BioBERT</td>
<td>93.42 (0.10)</td>
<td>+3.50</td>
</tr>
<tr>
<td>EhrBERT&lt;sub&gt;500k&lt;/sub&gt;</td>
<td>93.45 (0.09)</td>
<td>+3.53</td>
</tr>
<tr>
<td>EhrBERT&lt;sub&gt;1M&lt;/sub&gt;</td>
<td>93.82 (0.15)</td>
<td>+3.90</td>
</tr>
</tbody>
</table>

<sup>a</sup>DNorm: disease name normalization.

<sup>b</sup>MADE: Medication, Indication, and Adverse Drug Events.

<sup>c</sup>We used gold entity mentions as input.

<sup>d</sup>BERT: bidirectional encoder representations from transformers.

<sup>e</sup>N/A: not applicable.

<sup>f</sup>EhrBERT<sub>500k</sub>: BERT-based model that was trained using 500,000 electronic health record notes.

<sup>g</sup>EhrBERT<sub>1M</sub>: BERT-based model that was trained using 1 million electronic health record notes.

<sup>h</sup>We used MetaMap-predicted entity mentions as input.

<sup>i</sup>NCBI: National Center for Biotechnology Information.

<sup>j</sup>CDR: Chemical-Disease Relations.
Table 3. P values of the different models for the Medication, Indication, and Adverse Drug Events (predicted entities) corpus.

<table>
<thead>
<tr>
<th>Model</th>
<th>Model, P value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BERT&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>MetaMap</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>BERT</td>
<td>.17</td>
</tr>
<tr>
<td>BioBERT</td>
<td>.04</td>
</tr>
<tr>
<td>EhrBERT&lt;sub&gt;500k&lt;/sub&gt;</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>BERT: bidirectional encoder representations from transformers.

<sup>b</sup>EhrBERT<sub>500k</sub>: BERT-based model that was trained using 500,000 electronic health record notes.

<sup>c</sup>EhrBERT<sub>1M</sub>: BERT-based model that was trained using 1 million electronic health record notes.

Table 4. P values of the different models for the National Center for Biotechnology Information disease corpus.

<table>
<thead>
<tr>
<th>Model</th>
<th>Model, P value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BERT&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>DNorm&lt;sup&gt;d&lt;/sup&gt;</td>
<td>.10</td>
</tr>
<tr>
<td>BERT</td>
<td>.25</td>
</tr>
<tr>
<td>EhrBERT&lt;sub&gt;500k&lt;/sub&gt;</td>
<td></td>
</tr>
<tr>
<td>EhrBERT&lt;sub&gt;1M&lt;/sub&gt;</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>BERT: bidirectional encoder representations from transformers.

<sup>b</sup>EhrBERT<sub>500k</sub>: BERT-based model that was trained using 500,000 electronic health record notes.

<sup>c</sup>EhrBERT<sub>1M</sub>: BERT-based model that was trained using 1 million electronic health record notes.

<sup>d</sup>DNorm: disease name normalization.

Table 5. P values of the different models for the Chemical-Disease Relations corpus.

<table>
<thead>
<tr>
<th>Model</th>
<th>Model, P value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BERT&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>DNorm&lt;sup&gt;d&lt;/sup&gt;</td>
<td>.004</td>
</tr>
<tr>
<td>BERT</td>
<td>.18</td>
</tr>
<tr>
<td>BioBERT</td>
<td></td>
</tr>
<tr>
<td>EhrBERT&lt;sub&gt;500k&lt;/sub&gt;</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>BERT: bidirectional encoder representations from transformers.

<sup>b</sup>EhrBERT<sub>500k</sub>: BERT-based model that was trained using 500,000 electronic health record notes.

<sup>c</sup>EhrBERT<sub>1M</sub>: BERT-based model that was trained using 1 million electronic health record notes.

<sup>d</sup>DNorm: disease name normalization.

**Discussion**

**Principal Findings**

As shown in the results, BERT-based models outperformed MetaMap and DNorm. However, the performance differences between BERT-based models are not quite as large. Therefore, any kind of BERT-based models should be effective for entity normalization if one does not pursue 1%-2% performance improvements. Moreover, we also found that the domain of pretrained data has an effect on BERT-based models, but the effect is slight by further adding pretrained data. We will discuss these topics in the following sections.

**Effect of Domains**

In this section, we analyzed the effect of domains from two aspects. First, we investigated whether in-domain models performed better than out-domain models and whether the performance differences are statistically significant. For example, if the corpus belongs to the clinical domain (eg, MADE), the in-domain model (eg, EhrBERT) should theoretically perform better than out-domain models (eg, BERT or BioBERT). As shown in Multimedia Appendix 2 graph (a),
In-domain models performed better than out-domain models in two corpora (i.e., MADE and NCBI disease) out of three. In addition, statistical significance only emerged in the MADE corpus. By contrast, there are fewer corpora where out-domain models performed better than in-domain models. In the CDR corpus, the out-domain model (i.e., EhrBERT) performed better than the in-domain model (i.e., BioBERT); meanwhile, statistical significance exists. These results show that domains have an impact on the performances of models but the impact is not significantly visible between the biomedical and clinical domain.

Second, we analyzed whether clinical or biomedical domain models (e.g., BioBERT or EhrBERT) performed better than general domain models (e.g., BERT). As illustrated in Multimedia Appendix 2 graph (b), at least one model (i.e., BioBERT or EhrBERT) of biomedical and clinical domains performed better than the general domain model (i.e., BERT) in all corpora. More importantly, the performances of BioBERT or EhrBERT are significantly higher than that of BERT in all corpora. Therefore, the similarities of domains have a direct effect on the performances of models. Because biomedical and clinical domains are close to each other, the models trained using related data achieved similar results. By contrast, BERT achieved worse results in the biomedical or clinical corpora, since it was trained using the data from the general domain.

**Effect of the Data Size**

In this section, we discuss the effect of the data size on the performance of EhrBERT. To this end, we split up our EHR notes for pretraining models into a smaller part (500,000 notes) (i.e., EhrBERT\textsubscript{500k}) and a larger part (1 million notes) (i.e., EhrBERT\textsubscript{1M}). From Table 2, we observed that EhrBERT\textsubscript{1M} performed better than EhrBERT\textsubscript{500k} in all corpora, improving the F1s by 0.08%, 0.35%, and 0.37%. Thus, it may be helpful to enlarge the size of pretraining data to generate high-quality models. However, the significance analysis in Tables 3-5 indicates that the performance of EhrBERT\textsubscript{1M} is only statistically better than that of EhrBERT\textsubscript{500k} in the CDR corpus. In other corpora, they are not statistically discernable. Therefore, we cannot reach the conclusion that the larger the size of the pretraining data, the better the model becomes. One likely reason is that EhrBERT was not pretrained from scratch. It was fine-tuned from BioBERT, which was fine-tuned from BERT. Thus, EhrBERT may only need a certain amount of data to transfer from one domain to another domain. For most downstream tasks, we believe that using EhrBERT\textsubscript{500k} is effective enough. We leave further investigation of the data size for future work.

**Case Study**

To better understand EhrBERT, we manually analyzed about 100 cases in the MADE corpus and selected some typical cases that were predicted correctly or incorrectly. In addition, we also built a dot-attention [36] layer on top of EhrBERT to show the weight of each word. As illustrated in Figure 3, we learned the following points based on our observation.

First, short and simple entity mentions are easy to normalize. For example, the mention *fevers* was correctly normalized to the gold term *Fever* in the vocabulary. Moreover, complex words such as *osteoporosis* can be normalized correctly by our BERT-based models. In previous work, such words usually bring trouble, since they are out-of-vocabulary and cannot be well represented by models. However, our BERT-based models, which are built based on word pieces rather than words, can benefit from subword information and alleviate the out-of-vocabulary problem. Furthermore, long mentions, which consist of multiple words, are more difficult to be normalized. Through the visualization of attention weights, we found that EhrBERT can sometimes make valid predictions by concentrating on keywords and by neglecting noise at the same time. For instance, our model paid more attention to weight and gain in the mention *weight loss or gain*, it successfully linked the mention to the correct term, *Weight gain*.
Through the case study, we also learned some lessons. First, EhrBERT sometimes paid more attention to irrelevant words, leading to incorrect predictions. For example, since EhrBERT gave more attention to calculus in the mention ureteral calculus, it missed the important information from ureteral. Therefore, it linked the mention ureteral calculus to an invalid term, Kidney stone. Second, as the mention lengths became longer, it was more difficult for EhrBERT to focus on the correct words. For example, regarding the mention complications of his stone retrieval, since EhrBERT concentrated on the part near stone rather than complications, it linked the mention to Kidney stone rather than to the valid term Complication of procedure. Third, we found that even though EhrBERT sometimes paid more attention to proper words, it still failed to make correct predictions. For example, body and ache attained higher weights in the mention body aches, but the mention was not linked to the right term, Pain. One likely reason is that the model needs to truly understand the similarity between Pain and ache. Lastly, we observed some cases that are difficult even for us. For instance, the mention reactions to drugs is ambiguous. It is hard to know the true reason for reaction based on limited information. Therefore, such a situation may need more information to disambiguate mentions, such as context or background knowledge.

Limitations
One limitation of our work is that entity normalization is treated as a single-label classification problem; however, it is not possible to handle this type of problem when an entity can be linked to more than one term in the vocabulary. To address this limitation, one could leverage the multi-label classification approach [37] via the binary cross-entropy loss to train the model. Another limitation is that our model has not made full use of the information in vocabularies, such as synonyms and hierarchical relationships. In the future, this can be explored via other models such as graph convolutional neural networks [38]. Lastly, we have observed that there is a bias in our model as shown in Multimedia Appendix 3. Like most machine learning models, our model prefers highly frequent words in the dataset.

Conclusions
In this paper, we investigated the effectiveness of BERT-based models for the entity normalization task in the biomedical and clinical domain. We found that BERT-based normalization models outperformed some state-of-the-art systems. Moreover, the performance can be further improved by pretraining our models on large-scale EHR notes. Furthermore, we found that domains have an impact on the performance of BERT-based models. The impact depends on the similarities between the domains of models and tasks. In the future, our approach will be evaluated in more clinical NLP tasks.

Acknowledgments
This work was supported by two grants from the National Institutes of Health (grant numbers: 5R01HL125089 and 5R01HL135219) and an Investigator-Initiated Research grant from the Health Services Research and Development Program of the US Department of Veterans Affairs (grant number: I01HX001457-01).

Conflicts of Interest
None declared.

Multimedia Appendix 1
Full results on three corpora of entity normalization.
[PDF File (Adobe PDF File), 62KB-Multimedia Appendix 1]

Multimedia Appendix 2
Statistics of the domain effect.
[PDF File (Adobe PDF File), 60KB-Multimedia Appendix 2]

Multimedia Appendix 3
Analysis of the model bias.
[PDF File (Adobe PDF File), 60KB-Multimedia Appendix 3]

References


Abbreviations

BERT: bidirectional encoder representations from transformers
C: $d^\text{Trm}$-dimensional representation
CDR: Chemical-Disease Relations
[CLS]: classifier token
DNorm: disease name normalization
E: $d^{\text{emb}}$-dimensional embedding
EHR: electronic health record
EhrBERT: BERT-based model that was trained using 1.5 million electronic health record notes
EhrBERT$\text{1M}$: BERT-based model that was trained using 1 million EHR notes
EhrBERT$\text{500k}$: BERT-based model that was trained using 500,000 EHR notes
ELMo: embeddings from language models
EN: entity normalization
MADE: Medication, Indication, and Adverse Drug Events
MedDRA: Medical Dictionary for Regulatory Activities
MEDIC: MEFerged DIsease voCabulary
NCBI: National Center for Biotechnology Information
NLP: natural language processing
PMC: PubMed Central
SNOMED-CT: Systematized Nomenclature of Medicine—Clinical Terms
T: $d^\text{Trm}$-dimensional vector
Trm: bidirectional transformer
UMLS: Unified Medical Language System
w: word piece