EMS-BERT: A Pre-Trained Language Representation Model for the Emergency Medical Services (EMS) Domain

M Arif Rahman
Dept. of Computer Science
University of Virginia
mir6zw@virginia.edu

Sarah Masud Preum
Dept. of Computer Science
Dartmouth College
Sarah.Masud.Preum@dartmouth.edu

Ronald D. Williams
Dept. of Electrical and Computer
Engineering
University of Virginia
rdw@virginia.edu

Homa Alemzadeh
Dept. of Electrical and Computer
Engineering
University of Virginia
ha4d@virginia.edu

John Stankovic
Dept. of Computer Science
University of Virginia
jas9f@virginia.edu

ABSTRACT

Emergency Medical Services (EMS) is an important domain of healthcare. First responders save millions of lives per year. Machine learning and sensing technologies are actively being developed to support first responders in their EMS activities. However, there are significant challenges to overcome in developing these new solutions. One of the main challenges is the limitations of existing methods for EMS text mining, and developing a highly accurate language model for the EMS domain. Several important Bidirectional Encoder Representations from Transformer (BERT) models for medical domains, i.e., BioBERT and ClinicalBERT, have significantly influenced biomedical text mining tasks. But extracting information from the EMS domain is a separate challenge due to the uniqueness of the EMS domain, and the significant scarcity of a high-quality EMS corpus. In this research, we propose EMS-BERT - a BERT model specifically developed for EMS text-mining tasks. For data augmentation on our small, classified EMS corpus which consists of nearly 2.4M words, we use a simultaneous pre-training method for transfer-learning relevant information from medical, bio-medical, and clinical domains; and train a high-performance BERT model. Our thorough evaluation shows at least 2% to as much as 11% improvement of F-1 scores for EMS-BERT on different classification tasks, i.e., entity recognition, relation extraction, and inferring missing information when compared both with existing state-of-the-art clinical entity recognition tools, and with various medical BERT models.

CCS CONCEPTS

• **Information systems** → **Specialized information retrieval**; • **Applied computing** → **Health care information systems**; *Document management and text processing*.

KEYWORDS

Emergency Medical Services (EMS) Data Processing and Analysis, EMS Entity Recognition, Medicine and Health, Language Model

1 INTRODUCTION

Emergency Medical Services (EMS) provide emergency medical care to patients who are involved in an incident that causes serious illness or injury. EMS play an intricate role in healthcare, each component of EMS performs coordinated efforts for providing emergency medical care to the patient(s). An EMS system does not exist in isolation, rather it is integrated with other healthcare related services intended to maintain and enhance a community's health and safety. Emergency services often provide the most timely initial care to begin the recovery process for the patient. In the USA alone, EMS providers save thousands of lives everyday and initiate a primary phase of patient recovery through the healthcare system [1]. To improve healthcare and provide better services to the patients, the EMS domain can not be ignored. Sometimes, the whole recovery phase of the patient is conducted by EMS. EMS providers perform different interventions on the patient, and collect lots of data during an EMS scene for future treatment. The information regarding the recovery process and patient health are documented after each EMS episode. Using this data from the EMS scene, novel methods such as an EMS specific language model can be built to analyze patient information and predict patient outcome. State-of-the-art assistants for medical care heavily rely on correct detection of EMS related medical information. An EMS specific language model can also be utilized in EMS based applications such as [24, 26, 27] for developing a better, robust and more automated healthcare system.

EMS reports hold significant data related to different EMS protocols, interventions, and clinical conditions of the patient [27]. As EMS scenes occur frequently (especially during the COVID-19

pandemic) [1], and these reports are always generated afterwards, analysis of such information can also play an important role in optimizing the entire process, i.e., save money, time, and lives by better understanding of the EMS information and their correlation to improve performance in the future [11].

Different machine learning techniques exist in the literature to uncover patterns and improve predictions [17, 37, 39]. However, unstructured, high-dimensional, and sparse information such as EMS reports are difficult to use in traditional machine learning models. In recent years, advances in deep learning and transformers have led to great progress towards generic and personalized predictions in different medical domains. A key contributing factor to this success is the introduction of large multimodal health data such as electronic health records (EHR) [30]. Each individual's EHR can link data from many sources, i.e. doctor visits and hospital episodes. This data contains entities/concepts such as diagnoses, interventions, lab tests, clinical narratives, and more. The adoption of EHR systems has greatly impacted the frequency of hospitalization of patients [9, 14] and detection of severe illnesses [3, 23]. On the contrary, the EMS domain has seen almost no advancements in processing the EMS data for understanding the patient condition and personalized treatment generation. Just like the EHR data, an EMS dataset can be utilized to develop a domain specific language model for text-mining purposes in EMS based applications.

For developing a domain-specific language model, pre-training of the language model on large-scale raw textual corpus has already made a tremendous contribution for transfer learning in natural language processing (NLP). Introduction of transformer-based language models, such as Bidirectional Encoder Representations from Transformers (BERT) has significantly improved the performance of information extraction from free text in the general domain [6]. For domain-specific purposes, many studies showed that additional pre-training of the BERT model on a domain-specific corpus results in better performance in their specific text-mining tasks. Two of the most contributing factors for developing a domain adapted language model are the size of the training corpus, and the relevance of training dataset. For example, BERT models such as BioBERT and ClinicalBERT localize on biomedical and clinical text, respectively [2, 13]. However, these models are developed for the medical domain, and the EMS domain is significantly separate from both. Although the performance of the medical BERT models is good for entity and relation detection tasks, other barriers exist to relate the localization to the EMS domains. For example, information extraction and correlation detection for the EMS domain is unique from the previous two domains when compared to the lexicon of in-hospital medical and clinical corpora [11]. The EMS domain has its own uniqueness because of the specialized vocabulary which the first responders use. These are the main reasons which limit the applicability of current medical and clinical solutions. Compared to existing clinical and medical dataset, the EMS dataset is often specialized, unstructured and noisy. Our experiments suggest that concept detection as well as semantic inference from EMS data, i.e., negation detection, temporal expression detection, and value association for accurate information extraction requires different approaches compared to the clinical state-of-the-art methods and tools [24]. However, due to the lack of available datasets in the EMS domain, we devise a solution to utilize the overlapping portion of

clinical and medical datasets for augmenting our experimental EMS dataset. Since both domains are based on medical issues, there also exists overlap between EMS and the clinical domain. Some portions of EMS concepts are similar to clinical and medical concepts such as disease names and medication names.

As our EMS dataset is limited, we utilize data augmentation from related clinical and medical BERT models to develop **EMS-BERT: A Pre-Trained Language Representation Model for Emergency Medical Services (EMS) Domain** for text-mining purposes in EMS. For developing EMS-BERT, we implement simultaneous pre-training [34] method using two relevant types of corpora and combine them to create a sizable corpus of over 1.5B words. We augment our training corpus with amplified vocabulary from these related domains as well as from the general domain. First, we show the efficiency of our method for downstream text-mining tasks, i.e., entity/concept recognition, relation extraction, and inferring missing information using comparison with predefined EMS protocols on EMS documents. Then, we also demonstrate that when applied on the EMS domain, our approach provides a better pre-trained model that outperforms existing BERT models from the general, medical, clinical, and bio-medical domains, i.e., BERT-Base, BioBERT, and ClinicalBERT, and existing clinical concept recognition methods, i.e., MetaMap, CLAMP, and cTAKES. The main contributions of this research are:

- To the best of our knowledge, EMS-BERT is the first BERT based language model for clinical entity/concept related text-mining purposes on an EMS specific textual corpus. The novelty lies in the new EMS domain where we apply the BERT technique, and in creating the training instances for the model.
- Leveraging the state-of-the-art method of simultaneous pre-training, EMS-BERT is developed using an EMS corpus of real-scene transcripts and post-scene summary reports. This corpus is larger than any other EMS corpus available; and the corpus is augmented with medical, clinical, bio-medical and general corpora with the simultaneous pre-training method. Subsequently, we show that the localization of EMS entities with EMS-BERT is feasible using our method.
- We compared the performance of EMS-BERT with state-of-the-art techniques and without the simultaneous pre-training method. EMS-BERT with simultaneous pre-training outperforms existing BERT based models from relevant domains (i.e. BERT-Base, BioBERT, ClinicalBERT) and existing clinical concept recognition methods (i.e. MetaMap, CLAMP, cTAKES, EMSContExt) in F-1 scores for EMS entity recognition and relation extraction from EMS corpora by 5% to 11%, and by 2% to 6%, respectively. Also, using standard EMS protocols and guidelines, EMS-BERT infers missing EMS information from EMS documents with at least 7% to as much as 14% better accuracy compared to other methods.

The rest of the paper is organized as follows: section 2 highlights related works for this paper and presents related background knowledge on BERT. Section 3 provides technical challenges for the problem, and presents our solution. Experimental details and evaluation results are discussed in section 4. We conclude with discussion and future plans in section 5 and 6.

2 RELATED WORK

Different tools exist for extracting information from unstructured clinical texts, including, MetaMap [4], cTAKES [28], CLAMP [31] and EMSContExt [25]. MetaMap combines natural language processing (NLP) with knowledge-intensive approaches for clinical concept recognition and mapping for normalization. The Clinical Text Analysis and Knowledge Extraction System (cTAKES) combines rule-based and machine learning techniques to achieve this. CLAMP is a comprehensive clinical Natural Language Processing (NLP) software that enables recognition and automatic encoding of clinical information in narratives. EMSContExt uses a weakly supervised approach for recognition of EMS concepts from textual corpus leveraging lexical, medical and EMS domain knowledge integration. All of these tools and methods use either the Unified Medical Language System (UMLS), or lexicon expansion approach to extract medical concepts. Two of the main drawbacks of using these tools for EMS entity recognition is their inability for categorizing the contexts in finer granularity, and lack of correlation understanding. For example, MetaMap has Concept Unique Identifiers (CUI) and semantic type lists which signify whether a clinical concept is 'Disease' or 'Medication'. But there is no way to differentiate whether the disease or medication is the current condition of the patient or an occurrence from the past, i.e., recognition of context or relations between entities. Our proposed system EMS-BERT, on the other hand, uses a domain specific bidirectional transformer (BERT) based language model and a simultaneous pre-training technique to recognize entity, their relation and inferring missing information from an unstructured, relatively small-sized EMS corpora. There have been some other works on clinical document summarization and information extraction including, [5, 7, 16]. However, these works focus only on a subset of information and are not specialized for EMS domain.

The introduction of transformer based language models such as the Bidirectional Encoder Representations from Transformers (BERT) [6] has significantly increased the performance of information extraction from free text. Previously, authors in [20] proposed a vector representation for words called GloVe embeddings. GloVe does not explore the context while creating the word embeddings which means that the meaning of any specific word in different contexts will render the same embeddings. To address this limitation, authors in [21] came up with the idea of contextualized word-embeddings called ELMo, which created word embeddings using a bidirectional LSTM. ELMo is trained with a language modeling objective. ULMFiT [8] is another successful model for training a neural network with language modeling objective and fine-tuning for a specific task. However, all these models take into account the next occurring words and disregard the context from the previous words. BERT on the other hand, addresses the limitations in these prior works by taking the contexts of both the previous and next words into account instead of just looking at the next set of words for context. BERT [6] is a contextualized word-representation model which is based on masked language modeling (MLM). BERT model is pre-trained using bidirectional transformers [33]. There are two steps in the BERT framework: pre-training and fine-tuning. During pre-training, the model is trained on unlabeled, large-sized corpora. For fine-tuning, the BERT model is first initialized with pre-trained

weights, and all the weights are fine-tuned using labeled data from the downstream tasks. BERT pre-training is optimized for two unsupervised classification tasks - masked language modeling (MLM) and next sentence prediction (NSP). The training instance of MLM is a single modified sentence. Each token in the sentence has a 15% chance of being replaced by a special token [MASK]. The chosen token is replaced with 80% of the time, 10% with another random token, and the remaining 10% with the same token. The MLM objective is to find a cross-entropy loss on predicting the masked tokens. Next sentence prediction (NSP) is a binary classification task for predicting whether two segments follow each other in the original text. Positive instances are created by taking consecutive sentences from the text corpus. Negative instances are created by pairing segments from different documents. Positive and negative instances are sampled with equal probability. The NSP objective is designed to improve the performance of downstream tasks, such as natural language inference, which require reasoning regarding the relationships between pairs of sentences. Figure 1 [34] shows the basic architecture of a BERT model.
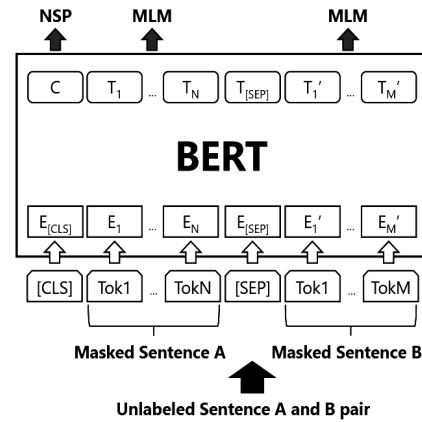

Figure 1: Basic architecture of BERT

Recent BERT models such as RoBERT and ToBERT [18] provide solutions for classification on long text, KnowBERT [22] incorporates different knowledge bases into BERT. In the masked language modeling approach of BERT, words in a sentence are randomly erased and replaced with a special token. A transformer is used to generate a prediction for the masked word based on the unmasked words surrounding it. With the masked language modeling objective, BERT has achieved improved results for many NLP tasks. Different research with BERT such as BioBERT [13], ClinicalBERT [2], and SciBERT [19] showed that additional pre-training of BERT models on a large domain-specific text corpus results in satisfactory performance in their specific text-mining tasks. For the clinical domain, models such as BioBERT and ClinicalBERT perform text-mining on structured clinical corpora. However, EMS or Emergency Medical Service domain is very different from traditional clinical corpora. The EMS dataset is mostly unstructured, the providers use different sets of semantics and lexicons during their communication and post-incident summary reports. For data mining on our EMS corpora

Pre-training for EMS-BERT
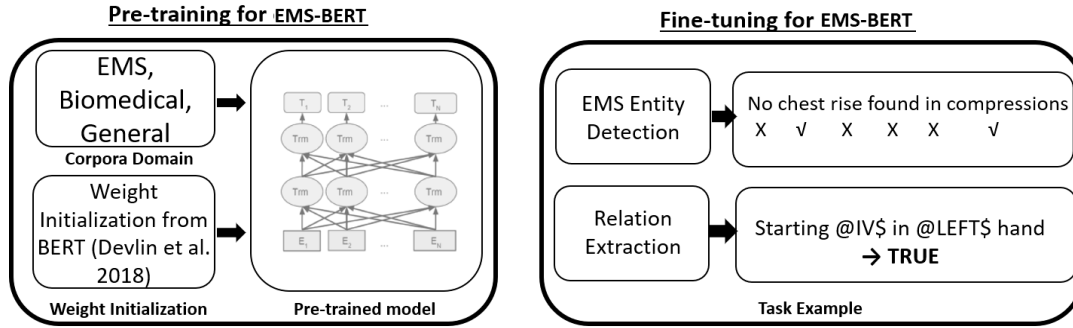
Fine-tuning for EMS-BERT

Figure 2: Overview of EMS-BERT model

which consists of live-transcriptions and post-scene narratives, existing medical and clinical BERT models do not perform as well as they perform on their respective domain. Hence we develop the EMS-BERT language model for entity recognition, relation extraction and inferring missing information from an EMS corpus.

3 METHODOLOGY AND SOLUTION

In the following subsections, we describe the underlying methods for EMS-BERT. Figure 2 [34] shows an overview of the overall approach.

3.1 Data augmentation

Through regional collaborators, we have access to 40,000 EMS narratives from real EMS scenes. These EMS narratives constitute over 2.4M words which contain different EMS concepts or entities, i.e., signs and symptoms, interventions, and medication information. Even though we have 40,000 narratives, this is relatively small compared to other dataset sizes for different domain specific BERT models (1). Besides the size, most of this narrative corpus are structured as they are created as post-scene summary reports. However, EMS narratives are also created with unstructured, on-scene communication based transcripts. EMS-BERT should be able to extract critical information from on-scene EMS narratives which is created by speech transcriptions collected at emergency scenes. So we need an EMS dataset which contains attributes such as noise, sparsity and of an unstructured nature for training EMS-BERT. For noisy and distorted EMS entities, we created a mapping of distorted entities to original EMS entities for the EMS-BERT model. The tokenization uses this mapping for suggesting the potential correct entity. Besides the noise, to train EMS-BERT with a sizable corpus for including broader medical and clinical entities, we devise a simultaneous pre-training method using a corpus from relevant domains. We have augmented the dataset by the following two methods to include both of these features.

3.1.1 Textual noise insertion. We utilized different noise insertion methods in textual corpora to emulate EMS narratives created from on-scene transcripts. Since the speech-to-text conversion sometimes yields distortion and inappropriate homophones in the presence of noise, we have used the state-of-the-art noise insertion methods to simulate similar kinds of errors. These noisy textual narratives are used for training EMS-BERT to mimic on-scene EMS transcripts.

Authors in [27] discuss the possible kinds of noise found in textual data. The authors in [32] highlight on text produced by processing signals and demonstrate that they are often noisy for automated processing. We have implemented a modified version of SpellMess [32] to introduce spelling errors in the EMS corpus. This modified version can change and/or substitute phonetically similar segments in a word, e.g., replacing a word with a homophone. We have created a list of possible homophones found in clinical context from [12]. Besides insertion, deletion and substitution of letters, homophone substitution is highly correlated with the kind of impact noise found in EMS transcriptions.

3.1.2 Simultaneous pre-training of EMS-BERT.. The standard BERT model does not perform well in specialized domains [13]. To overcome this limitation, possible techniques include additional pre-training on domain-specific corpora from an existing pre-trained BERT model, or pre-training from scratch on domain-specific corpora. A main benefit of the former is that the computational cost of pre-training is lower than the latter. The main advantage of the latter is the availability of its custom vocabulary, but the disadvantage is that the pre-trained neural language model may be less adaptable if the number of documents in a specific domain is small. Due to the scarcity of public EMS corpora, both approaches seem infeasible. So we argue that transfer-learning from relevant and general domain will create a more accurate language model for EMS domain.

For general corpora, state-of-the-art BERT-Base is pre-trained using English Wikipedia and the Books Corpus [6]. The vocabulary is quite different from EMS corpora, thus rendering this pre-training corpus only is quite inappropriate. BioBERT is the first BERT model released for the biomedical domain [13] which is initialized from BERT-Base and trained using PubMed abstracts. ClinicalBERT is also a clinically oriented BERT model [2] which is initialized from BioBERT v1.0 and trained with additional steps using MIMIC-III clinical notes. We use the BioBERT and ClinicalBERT vocabulary with BERT-base to augment our EMS corpora for simultaneous pre-training of EMS-BERT.

Table 1 summarizes the previous BERT-based dataset we use to augment our EMS corpora. Training a BERT model with a smaller corpus degrades the performance by introducing more false positives. As there is no public EMS corpus and collecting real-world EMS narrative is subject to different prohibitions, we adopted the method of simultaneous pre-training introduced by the authors in OuBioBERT
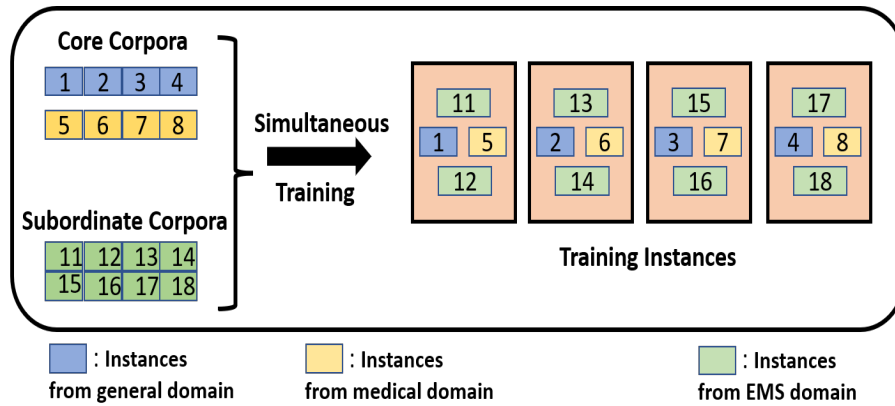
Figure 3: Simultaneous pre-training for EMS-BERT

[34] to increase the size and cover additional entities. Simultaneous pre-training of BERT with domain specific knowledge and generic corpora provides better results. This is achieved by increasing the frequency of pre-training instances for MLM (masked language modeling). Instead of only using the corpus from the medical domain, we use documents from the EMS domain too, with the general medical domain. Using the negative instances of NSP (next sentence prediction) where a sentence pair is constructed by pairing two random sentences from different documents, simultaneous pre-training method also increases the number of combinations of documents and enhances EMS word representations in the vocabulary.

The simultaneous pre-training technique is illustrated in Figure 3. This approach successfully creates an efficient pre-training corpora from multiple domains. While pre-training the EMS-BERT model, the core corpora is constituted from both the general and medical domains. The EMS corpora is considered as subordinate corpora here, which is used to create mixed training instances. During the implementation, the entire corpus was divided into smaller text files. This was particularly helpful to create simultaneous pre-training instances from different type of corpora. The combinations of NSP are determined within each split file, and the duplicate factor is set to define the number of times the sentences are used. There are two problems that arises in these cases. The first is that the duplicate factor is applied to the entire corpora of both core corpora and subordinate corpora. Thus, the smaller corpora remain relatively small. The second problem is that the combinations of NSP are limited to the file that was initially split. To solve these issues, both core corpora and subordinate EMS corpora are first divided into smaller documents with the same size for EMS-BERT. Later, we combine them to create pre-training instances. When we combined them, it was ensured that the documents in both of the corpora would be comparable in terms of their file sizes and diversity of the patterns. Using this technique, more instances from core corpora were used compared to those from subordinate corpora. With this homogeneously mixed dataset, the model achieved a higher increase in the frequency of pre-training for MLM. Using documents of core corpora for the process of pre-training creates larger training dataset than the original BERT method. It also generates an increased number of different combinations of documents compared to the original

method. Core corpora and subordinate corpora were combined so that their proportion were equal, thus a higher number of pre-training instances were created to train the EMS-BERT model. Comparing with the state-of-the-art BERT models and their pre-training dataset, our dataset volume is comprehensive and provides a better accuracy for the EMS domain.

3.2 Fine tuning EMS-BERT

For fine-tuning, a pre-trained language model generates a set of vectors with contextual representations. A task-specific prediction layer placed on top produces the final output for the application. Task-specific model parameters are trained from the task-specific training data. While training, BERT model parameters are fine-tuned by gradient descent using back-propagation. An input instance from EMS corpora goes through task-specific pre-processing and addition of special instance markers ([CLS], [SEP], etc.). The transformed input is then tokenized using the pre-training vocabulary of the neural language model. The sequence of vectors in contextual representations taken from the language model is then processed by a feature module and input into a prediction module to produce its final output of the given task. A sentence is transformed into an instance for BERT by replacing target entities with dummy tokens and adding special tokens. In the relation-extraction task, we use [CLS] BERT encoding as a featurizer and predict the relationship between the entities by multi-class classification. The relation extraction task predicts relations between two entities and their types mentioned in the sentence. We explored three entities from the EMS corpora - signs and symptoms, medications, and interventions. Our experiments predicted all the six pairs of relations among these three entities from the textual corpora.

Utilizing the approach discussed in the BLUE benchmark [19], this task is implemented as a sentence classification task by using anonymous entities within the sentences and predefined tags such as @SYMPTOM$ and @INTERVENTION$ [13]. Figure 4 shows a general architecture of fine-tuning a BERT model for downstream tasks [34]. We fine-tune EMS-BERT for the following two tasks: (i) EMS concept recognition, and (ii) relation extraction. Compared to regular clinical corpora, EMS concepts cover a wide range of clinical conditions, medications, and intervention. These entities may be

correlated depending on the recovery protocol. Relation extraction from EMS corpora signifies such dependencies and infers missing information from the narratives. Thus, accuracy of relation extraction task highlights potentially missing attributes from a given set of EMS interventions. For comprehensiveness of the evaluation, we compare EMS-BERT with state-of-the-art clinical concept detection tools, and with relevant BERT models for clinical and medical domain. Two evaluations are detailed in the following sections. First, we studied the EMS entity/concept recognition using state-of-the-art clinical concept recognition baseline tools such as MetaMap [4], cTAKES [28], and CLAMP [31]. Second, we showed the accuracy of relation extraction using EMS-BERT using the ground truth developed by EMS professionals.

4 EVALUATION OF DOWNSTREAM TASKS

In this section, we describe the experimental setup and dataset used to pre-train and test EMS-BERT. We also present the results of the experiments for recognition of EMS entities, relation extraction, and inferring missing information.

4.1 Experimental design and dataset

4.1.1 Setup. We use mixed precision training of FP16 computation for both pre-training and fine-tuning EMS-BERT. This method accelerates the computation significantly compared to other methods as it uses half-precision format. Two NVIDIA RTX-8000 of 32 GB size GPU are used for pre-training; a single GPU is used for fine-tuning. The configuration and weight initialization are almost same as the BERT-base. We modified the NVIDIA implementation to utilize FP16 computation, gradient accumulation, and a layer-wise adaptive based optimizer (LAMB) [38]. For pre-training, we set the maximum sequence length of 128 tokens and trained the model for 5,068 steps using the global batch size (GBS) of 65,536 and a LAMB optimizer with the learning rate (LR) of 6e–3. Subsequently, we continued to train the model allowing a sequence length up to 512 tokens for an additional 1,272 steps to learn positional embeddings. The size of the amplified vocabulary is 30,700.

For EMS entity recognition, EMS-BERT performs sequential labelling and detects the required entities in the given text. The BERT encoding of a given sequence of token predicts the label and recognizes the entity. The relation extraction task predicts relations between two entities and their types mentioned in the sentence. We explored three entities from the EMS corpora - signs and symptoms, medication, and intervention. Our experiments predicted all six relations among these three entities from the textual corpora. We also avoid overfitting by inserting dummy tags for entities, as depicted in Figures 2 and 4. Using the relation extraction task and a mapping of prerequisites of different entities developed by certified EMS professionals, we infer potentially missing information from the EMS test set. This information depicts how thoroughly each of the approaches cover the EMS entities in an EMS document.

4.1.2 Dataset and metrics. We utilized some of the datasets used in BERT-base [6], BioBERT [13] and ClinicalBERT [9] to pre-train EMS-BERT. BERT-base use English Wikipedia and Book Corpus as general domain corpora. BioBERT and ClinicalBERT use PubMed abstracts and PubMed Central Full-Text articles (PMC) [15], and Medical Information Mart for Intensive Care III dataset

(MIMIC-III) [10], respectively. These two datasets hold information specific to the medical and clinical domain. For EMS-BERT, we create our simultaneous training instances by combining the general, medical and clinical domain information with EMS corpora (depicted in Figure 3). Table 1 summarizes the datasets used to pre-train EMS-BERT. For EMS corpora, we used 36,000 EMS narratives for creating simultaneous training instances and 4,000 annotated EMS narratives for validation and testing EMS-BERT. The testing set includes both noisy, unstructured EMS transcripts and structured, post-scene EMS narratives. Certified EMS professionals supervised the annotation of the EMS dataset. Since our target is to measure how accurately EMS-BERT recognizes EMS entities, i.e., signs and symptoms, medications, and interventions, and extract relations between each of the entity pairs, we have selected *Precision, Recall and F-1 Score* as our accuracy metrics. We also compare EMS-BERT's simultaneous pre-training method with a knowledge integration based approach known as KnowBERT [22]. KnowBERT integrates knowledge bases into BERT using knowledge attention and a recontextualization component (KAR).

4.2 Experimental results

In this section, we detail the results obtained with EMS-BERT using the augmented dataset. We then compare these results with other state-of-the-art techniques and tools from the literature. For ablation studies of simultaneous pre-training, we also used a different pre-training of EMS-BERT which does not utilize a simultaneous pre-training method. The results obtained with this version is labelled under EMS-BERT-wsn to show the efficacy of simultaneous pre-training for our corpus.

4.2.1 EMS entity recognition. Table 2 shows the overall average scores of EMS entity recognition, i.e., of signs and symptoms, interventions, and medication. For the noisy and structured test dataset, the average F-1 score for EMS-BERT is 78.85 (72.91 and 81.68, respectively). EMS-BERT outperforms the other state-of-the-art tools by at least 5%. Comparison with EMS-BERT-wsn emphasizes the significant of simultaneous pre-training for the EMS corpora. The average F-1 score for EMS-BERT-wsn is only 52.91. We observe that for BERT-Base, which is pre-trained on only the general domain corpus, the result if very poor. The average F1-score is 51.89 for BERT-Base which is significantly lower than that of the other state-of-the-art models. On the other hand, BioBERT v1.1 achieves higher scores than ClinicalBERT for the EMS dataset. The better results of BioBERT v1.1 is due to the higher similarity of EMS corpus with PubMed abstracts and PubMed Central Full-Text articles (PMC), compared to the Medical Information Mart for Intensive Care III dataset (MIMIC-III). For KnowBERT, the knowledge integration approach shows good results with an average F-1 score of 66.75. However, our insight suggests that the non-overlapping entities of the EMS domain and other medical, clinical domain plays a significant role for this relatively lower score. All these low scores of the other BERT models on the EMS dataset can also be attributed to the following generic reasons: (i) the lack of a silver-standard dataset for training previous state-of-the-art models, and (ii) different training/test set splits used in previous work which were unavailable. For clinical concept recognition tools such as MetaMap, CLAMP, and cTAKES, these tools exhibit a high false positive rate. One possible
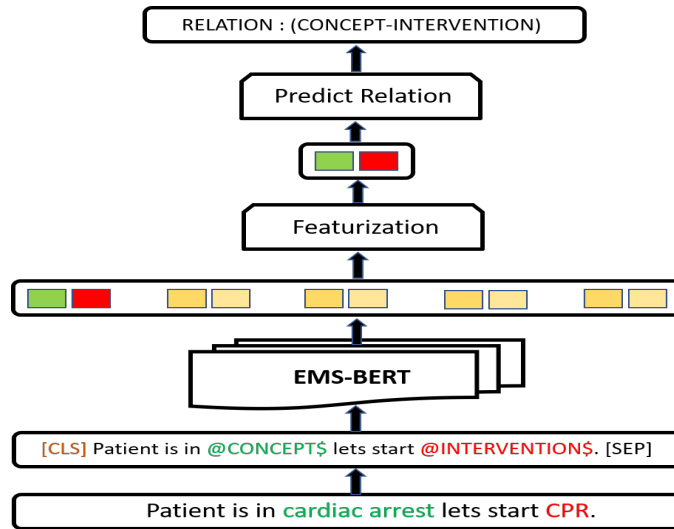
Figure 4: A fine-tuning example for EMS-BERT model

Table 1: Dataset for EMS-BERT

Model	Corpus	Number of Words	Domain
BERT-base	English Wikipedia, Book Corpus	2B	General
BioBERT	Wikipedia, Books, PubMed, PMC	3.5B	General, Medical
ClinicalBERT	MIMIC-III (Health Records)	1.5B	Medical
EMS-BERT	EMS Corpora	2.4M	EMS

Table 2: Entity/Concept recognition using EMS-BERT

Method/Metric	Precision	Recall	F-1 score
BERT-Base	52.58	51.21	51.89
EMS-BERT-wsn	64.87	61.09	62.91
CLAMP	69.29	62.58	63.42
cTAKES	62.94	65.81	64.34
KnowBERT	67.34	66.17	66.75
ClinicalBERT	69.59	68.24	68.91
MetaMap	71.23	68.47	69.82
BioBERT v1.1	73.84	70.81	72.29
EMSContExt	74.62	71.54	73.05
EMS-BERT	81.24	76.59	78.85

reason is the over-generalization of entities. A semi-supervised approach such as EMSConExt shows a better F-1 score compared to these three tools, but EMS-BERT also outperforms EMSContExt.

4.2.2 Entity relation extraction. The relation extraction results of different BERT models are shown in Table 3. We predict the relations between the following three EMS entities - signs and symptoms (S.&S.), intervention (Int.) and medication (Med.). EMS-BERT achieved better performance than the other state-of-the-art models.

On average, EMS-BERT obtained a higher F1 score (2%-5% higher) than original BERT-Base, KnowBERT and BioBERT v1.1 on EMS dataset. Table 3 shows that the accuracy of relation extraction for medication - medication is lower compared to the other pairs of relation. This is due to the inability of the model to detect other connecting contexts of the medications among themselves. Certain medications are labeled as prohibited with each other, however for the other pairs it is difficult to create a connection unless all of the underlying contexts are explicitly mentioned in the corpus and detected by the model. So any false positive or false negative in all other relation extraction may affect the relation extraction of multiple medications. For the other pairs of entities, the relation is much more straight forward, and EMS-BERT shows better results for those. Other models such as BioBERT v1.1 were not pre-trained using the augmented EMS corpora, so the lower performances are expected. KnowBERT shows good results compared to BERT-Base, this is because of the knowledge integration from the augmented EMS dataset. But compared to the results of EMS-BERT, the F-1 scores for relation extraction for each pair of entity are 5-12% lower on average. For BERT-Base, pre-training only with general domain corpus negatively impacts the outcome of relation extraction from the EMS corpus. The F-1 scores for relation extraction are 9-20% lower for each entity pair.

4.2.3 Inferring missing concept/entity. To measure the detection of missing information, we calculate the total coverage of concepts by EMS-BERT and other methods. Then, a concept dependency

Table 3: Relation extraction using EMS-BERT

Relation	Metric	BERT-Base	KnowBERT	BioBERT v1.1	EMS-BERT
Signs & Symptoms - Signs & Symptoms	P	60.51	73.51	77.29	79.58
	R	58.57	70.32	72.49	73.91
	F-1	59.52	71.88	74.81	76.64
Signs & Symptoms - Intervention	P	62.54	71.61	74.38	77.92
	R	59.78	70.87	71.94	76.51
	F-1	61.13	71.21	73.14	77.21
Signs & Symptoms - Medication	P	61.75	72.84	75.87	78.62
	R	57.39	69.74	74.94	75.54
	F-1	59.49	71.26	75.40	77.05
Intervention - Intervention	P	67.31	78.61	77.31	80.67
	R	64.32	74.62	73.88	77.92
	F-1	65.78	76.56	75.56	79.27
Intervention - Medication	P	64.38	66.94	71.62	73.94
	R	65.90	65.17	69.58	71.39
	F-1	65.13	66.04	70.59	72.64
Medication - Medication	P	51.84	60.25	63.57	68.67
	R	50.68	54.39	59.84	64.31
	F-1	51.25	57.17	61.65	66.42

Table 4: Total coverage of related EMS entities/concepts

Approach/Metric	Coverage (%)
ClinicalBERT	77.96%
BioBERT v1.1	84.47%
EMS-BERT	91.21%

model developed by certified EMS personnel predicts the missing information by comparing with the output of all the methods. For each of the entities found in an EMS narrative, there are some other entities which are correlated and expected to be preceded/followed in an EMS narrative. These are often prerequisites and post-requisites of various interventions and medications. Sometimes, they are not mentioned in the transcript or post-scene narrative. Using each of the entities in our test set, our EMT collaborators developed a document with dependencies among the EMS entities. When an entity is detected by EMS-BERT, it checks the list of all the correlated entities against the detected entity and infers the potentially missing entities in the original transcript. For example, a cardiac arrest protocol which exhibits the intervention CPR, must also have information regarding an IV intervention in the corpora. Table 4 shows the comparison of BERT, BioBERT v1.1 and EMS-BERT for recognition of all possible EMS entities which are correlated. Here, EMS-BERT shows highest accuracy for inferring potentially missing information by detecting the maximum number of entities and their potentially correlated missing entities correctly. EMS-BERT outperforms the other two models by at least 7% for overall data dependency capture.

This improvement is also significant for understanding the context of the situation and providing personalized patient care in latter stages of recovery. Inferring potentially missing information from live EMS transcripts and post-scene narratives lead to better EMS training and performance.

5 DISCUSSION

For the data mixing strategy and ablations, we do not have any ablation study at the moment to support the equal nature in core and subordinate corpora for augmenting the dataset. We have the data mixing research as a future goal for the project. Our future study will target finding what proportion of mixing both kinds of corpora yield best results, and whether there exist other approaches for data augmentation with similar or better results. Different methods for augmenting a dataset with amplified vocabulary exist in the literature, such as LSTM and transfer learning based approaches. In this study, we adopt the simultaneous pre-training method and compared it with multiple knowledge base integration by KAR methods [22]. As a future milestone of this research, we will investigate other data augmentation methods, run more comprehensive ablation studies for simultaneous pre-training, and compare their results with our current approach. Wolf et al. in [36] discussed the construction of the uncased vocabulary via byte-pair encoding (BPE) [29] using tokenizers. We implemented the uncased vocabulary as a custom vocabulary to suit a small corpus. A small corpus often shows biases towards subordinate corpora. To solve this problem, we amplified the core corpora and made the corpus size the same as that of the subordinate corpora. The authors in [35] presented Bidirectional

LSTM and BERT approaches to detect entity from EMS audits from Singapore Civil Defense Force. However, our EMS dataset is comparatively unstructured, noisy and a portion of it is created from live transcripts from real EMS scenes. The authors in [35] mentioned that one probable reason for their low scores with BiLSTM is the inability to handle misspelling in the dataset. Our hypothesis for developing EMS-BERT precisely highlights this condition of our dataset. The authors used basic BERT-Base and ClinicalBERT models instead of developing a custom BERT model. As we are focused to develop a generic model to detect EMS concepts and understand their correlations, we concentrated on developing a custom BERT for EMS domain. Our future goal also includes using EMS-BERT for other downstreaming tasks such as negation detection, vitals validation, etc. from EMS corpora.

In this paper, we present a study for augmenting the dataset compared to using only the EMS corpora without data augmentation. The results of EMS-BERT without simultaneous pre-training and data augmentation are documented and the experiment results show significant improvement when the simultaneous pre-training method is applied. We show that EMS-BERT outperforms ClinicalBERT and BioBERT for entity recognition, relation extraction, and inferring missing information for our EMS corpora. However, ClinicalBERT and BioBERT were developed for the medical and bio-medical domain. They are not pre-trained for the EMS domain. A more comprehensive comparison with a BERT model specifically pre-trained on EMS corpus will further strengthen the significance of EMS-BERT and it's simultaneous pre-training technique. For application of EMS-BERT, there are multiple potential scope. Cognitive assistants developed for the emergency response domain may leverage from deploying EMS-BERT in their backend. For example, different types of cognitive assistants for the emergency domain such as [24, 26], automated form filling [27] and other applications require clinical and medical entity detection from an EMS corpus. EMS-BERT can be very effective for such assistants and applications. These previous systems used different clinical concept detection tools, but our experiments clearly indicate better F-1 scores with EMS-BERT for the EMS domain. Currently, we are working on adapting the EMS-BERT model for real-time applications. We envision developing a cognitive application for introducing automation in EMS training. EMS-BERT will be deployed to detect different concepts in this application. The application will provide customized suggestions and feedback according to the severity level of the training and experience level of the first responder.

6 CONCLUSION

To the best of our knowledge, EMS-BERT is the first language model specialized for the EMS domain. For amplifying the existing EMS corpus which consists of post-scene EMS narratives and live-transcripts, EMS-BERT also utilizes general, clinical, and medical corpus from state-of-the-art BERT, BioBERT and ClinicalBERT models. Using simultaneous pre-training technique on the amplified vocabulary, we demonstrated that a practical BERT based model can be constructed for EMS downstream tasks. Our thorough experimentation also demonstrates that EMS-BERT outperforms the existing state-of-the-art medical and clinical models by at least 2% to as much as 11% for F-1 scores in downstream tasks such as entity

recognition, relation extraction, and inferring missing information on EMS domain. Even though there is room for improvement for the accuracy, the results suggest that EMS-BERT can successfully handle the complex challenges, i.e., unstructured, sparse, noisy, and high-dimensional dataset for text-mining related tasks in EMS domain. EMS-BERT also emphasizes the significance of a specialized BERT based language model for EMS specific corpus, and distinguishes the EMS domain from medical and clinical datasets.

REFERENCES

[1] Alaa Al Amiry and Brian J Maguire. 2021. Emergency medical services (Ems) calls during covid-19: Early lessons learned for systems planning (a narrative review). *Open Access Emergency Medicine: OAEM* 13 (2021), 407.

[2] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323* (2019).

[3] Diego Ardila, Atilla P Kiraly, Sujeeth Bharadwaj, Bokyung Choi, Joshua J Reicher, Lily Peng, Daniel Tse, Mozziyar Etemadi, Wenxing Ye, Greg Corrado, et al. 2019. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature medicine* 25, 6 (2019), 954–961.

[4] Alan R Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.. In *Proceedings of the AMIA Symposium*. 17.

[5] Murthy Devarakonda and Ching-Huei Tsou. 2015. Automated problem list generation from electronic medical records in IBM Watson. In *Twenty-Seventh IAAI Conference*.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[7] Yulan He. 2016. Extracting topical phrases from clinical documents. In *Thirtieth AAAI Conference on Artificial Intelligence*.

[8] Jeremy Howard and Sebastian Ruder. 2018. Fine-tuned language models for text classification. *arXiv preprint arXiv:1801.06146* 194 (2018).

[9] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342* (2019).

[10] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* 3, 1 (2016), 1–9.

[11] Sion Kim, Weishi Guo, Ronald Williams, John Stankovic, and Homa Alemzadeh. 2021. Information Extraction from Patient Care Reports for Intelligent Emergency Medical Services. In *2021 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*. IEEE, 58–69.

[12] Myrna LaFleur-Brooks and Danielle S LaFleur. 2005. *Exploring medical language: A student-directed approach*. Elsevier Mosby.

[13] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (2020), 1234–1240.

[14] Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. 2020. BEHRT: transformer for electronic health records. *Scientific reports* 10, 1 (2020), 1–12.

[15] Johanna McEntyre and David Lipman. 2001. PubMed: bridging the information gap. *Cmaj* 164, 9 (2001), 1317–1319.

[16] Srikanth Mujjiga, Vamsi Krishna, Kalyan Chakravarthi, and J Vijayananda. 2019. Identifying Semantics in Clinical Reports Using Neural Machine Translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 9552–9557.

[17] Thanh Nguyen-Duc, Natasha Mulligan, Gurdeep S Mannu, and Joao H Bettencourt-Silva. 2021. Deep EHR Spotlight: a Framework and Mechanism to Highlight Events in Electronic Health Records for Explainable Predictions. *AMIA Summits on Translational Science Proceedings* 2021 (2021), 475.

[18] Raghavendra Pappagari, Piotr Zelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. 2019. Hierarchical transformers for long document classification. In *2019*

IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 838–844.

[19] Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474* (2019).

[20] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[21] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *https://arxiv.org/abs/1802.05365v1* (2018).

[22] Matthew E Peters, Mark Neumann, Robert L Logan IV, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. 2019. Knowledge enhanced contextual word representations. *arXiv preprint arXiv:1909.04164* (2019).

[23] Ryan Poplin, Avinash V Varadarajan, Katy Blumer, Yun Liu, Michael V Mc-Connell, Greg S Corrado, Lily Peng, and Dale R Webster. 2018. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering* 2, 3 (2018), 158–164.

[24] Sarah Preum, Sile Shu, Mustafa Hotaki, Ronald Williams, John Stankovic, and Homa Alemzadeh. 2019. CognitiveEMS: A Cognitive Assistant System for Emergency Medical Services. *SIGBED Rev.* 16, 2 (aug 2019), 51–60. https://doi.org/10.1145/3357495.3357502

[25] Sarah Masud Preum, Sile Shu, Homa Alemzadeh, and John Stankovic. 2020. Emscontext: EMS protocol-driven concept extraction for cognitive assistance in emergency response. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 13350–13355.

[26] Sarah Masud Preum, Sile Shu, Jonathan Ting, Vincent Lin, Ronald Williams, John Stankovic, and Homa Alemzadeh. 2018. Towards a cognitive assistant system for emergency response. In *2018 ACM/IEEE 9th International Conference on Cyber-Physical Systems*.

[27] M Arif Rahman, Sarah M Preum, Ronald Williams, Homa Alemzadeh, and John Stankovic. 2020. GRACE: generating summary reports automatically for cognitive assistance in emergency response. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 13356–13362.

[28] Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association* 17, 5 (2010), 507–513.

[29] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909* (2015).

[30] Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. 2017. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE journal of biomedical and health informatics* 22, 5 (2017), 1589–1604.

[31] Ergin Soysal, Jingqi Wang, Min Jiang, Yonghui Wu, Serguei Pakhomov, Hongfang Liu, and Hua Xu. 2017. CLAMP–a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association* 25, 3 (2017), 331–336.

[32] L Venkata Subramaniam, Shourya Roy, Tanveer A Faruquie, and Sumit Negi. 2009. A survey of types of text noise and techniques to handle noisy text. In *Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data*. ACM, 115–122.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[34] Shoya Wada, Toshihiro Takeda, Shiro Manabe, Shozo Konishi, Jun Kamohara, and Yasushi Matsumura. 2020. Pre-training technique to localize medical bert and enhance biomedical bert. *arXiv preprint arXiv:2005.07202* (2020).

[35] Han Wang, Wesley Lok Kin Yeung, Qin Xiang Ng, Angeline Tung, Joey Ai Meng Tay, Davin Ryanputra, Marcus Eng Hock Ong, Mengling Feng, and Shalini Arulanandam. 2021. A Weakly-Supervised Named Entity Recognition Machine Learning Approach for Emergency Medical Services Clinical Audit. *International journal of environmental research and public health* 18, 15 (2021), 7776.

[36] Yanshan Wang, Naveed Afzal, Sunyang Fu, Liwei Wang, Feichen Shen, Majid Rastegar-Mojarad, and Hongfang Liu. 2020. MedSTS: a resource for clinical semantic textual similarity. *Language Resources and Evaluation* 54, 1 (2020), 57–72.

[37] Cao Xiao, Edward Choi, and Jimeng Sun. 2018. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association* 25, 10 (2018), 1419–1428.

[38] Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. 2019. Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv preprint arXiv:1904.00962* (2019).

[39] Kun-Hsing Yu, Andrew L Beam, and Isaac S Kohane. 2018. Artificial intelligence in healthcare. *Nature biomedical engineering* 2, 10 (2018), 719–731.

A AVAILABILITY OF MATERIALS

The PyTorch model for EMS-BERT is available at the following online repository: https://github.com/arifimtiaz216/EMS-BERT/. However, our EMS dataset is confidential and due to copyright reasons it is not made available.

B A SAMPLE DOCUMENT FROM THE EMS CORPUS

(De-identified) "*Dispatched for a sick person, that was changed to abdominal pain en route. Arrived to find the patient sitting outside of his house on porch steps leaning over holding his abdomen. Abdominal Pain found. The patient stated that he was discharged yesterday from local hospital and was diagnosed with pancreatitis. His pain onset was yesterday after dilaudid wore off, prescribed percocets did no effect with pain management per patient. Pain described as sharp pain, 10 out of 10, no radiation. He stated that his pain was in his entire abdomen, but worse in lower quadrants. The patient has had a heart attack, 2 stents put in. He stated that he has history of bradycardia. The patient was also given medication for nausea, but it did not work and he is still nauseous. PMH, medications, and medication allergies as noted. Alert and oriented person, place, event, and time are aware. Airway is patent, unobstructed, and self maintained. Breathing has equal bilateral chest rise and fall, clear lung sounds. Circulation has pms x 4, no bleeding noted, good turgor, no diaphoresis, skin warm pink and dry. The patient denies chest pain, shortness of breath, vomiting, back pain, syncope, lightheadedness, or blurry vision. Vitals obtained and documented. 12 lead: Sinus Bradycardia. 20 G IV in right hand connected to a 10gtts/ml drop set connected to a bag of 2000ml NS. 250ML Bolus of NS administered. Transported the patient to local hospital per patient's request. Patient condition unchanged. Report and care given to RN in room. Unable to obtain the patient's insurance due to his high level of pain.*"