EasyChair Preprint

№ 2268

BertMCN: Mapping Colloquial Phrases to Standard Medical Concepts Using BERT and Highway Network

Katikapalli Subramanyam Kalyan and S. Sangeetha

January 5, 2021

BertMCN: Mapping colloquial phrases to standard medical concepts using BERT and Highway Network

Katikapalli Subramanyam Kalyan*, Sivanesan Sangeetha

Text Analytics and NLP Lab, Department of Computer Applications, NIT Trichy, India

Abstract

In the last few years, people started to share lots of information related to health in the form of tweets, reviews and blog posts. All these user generated clinical texts can be mined to generate useful insights. However, automatic analysis of clinical text requires identification of standard medical concepts. Most of the existing deep learning based medical concept normalization systems are based on CNN or RNN. Performance of these models is limited as they have to be trained from scratch (except embeddings). In this work, we propose a medical concept normalization system based on BERT and highway layer. BERT, a pre-trained context sensitive deep language representation model advanced state-of-the-art performance in many NLP tasks and gating mechanism in highway layer helps the model to choose only important information. Experimental results show that our model outperformed all existing methods on two standard datasets. Further, we conduct a series of experiments to study the impact of different learning rates and batch sizes, noise and freezing encoder layers on our model.

Keywords: Medical Concept Normalization, Clinical Natural Language Processing, BERT, Highway Network

1. Introduction

Social media with an increasing number of users in recent times, evolved as a rich source of data for many domains, including healthcare. People use twitter[1], facebook[2] and online health forums and often share many things including their treatment experiences, symptoms while consuming a drug etc. This rich clinical data is underutilized which can be leveraged in many applications to offer better services [1].

The task of medical concept normalization aims to map health related entity mentions identified in free-form text to formal medical concepts in standard vocabulary like Unified Medical Language System (UMLS), Medical Dictionary for Regulatory Activities (MEDRA) or Systematized Nomenclature of Medicine – Clinical Terms (SNOMED-CT) (see Figure 1). Here, entity mention refers to adverse drug reaction, symptom, finding, drug or disease. Such a mapping is required because of variation in the languages of general public and healthcare professionals. Most of the general public express their health conditions in layman terms rather than formal

*Corresponding author

Email address: kalyan.ks@yahoo.com (Katikapalli Subramanyam Kalyan)

[1]https://twitter.com

[2]https://www.facebook.com

medical terms i.e., in a descriptive way which reveals how they feel. For example, '*insomnia*' is expressed in layman terms as '*could not sleep much*'. Further, the same health condition can be expressed in multiple ways which makes the task more challenging. Medical concept normalization also called Entity Linking or Entity Encoding is one of the fundamental tasks in information extraction with applications in tasks like Question and Answering, Pharmacovigilance etc. However, it is less explored when compared to other information extraction tasks like named entity recognition and relation extraction.

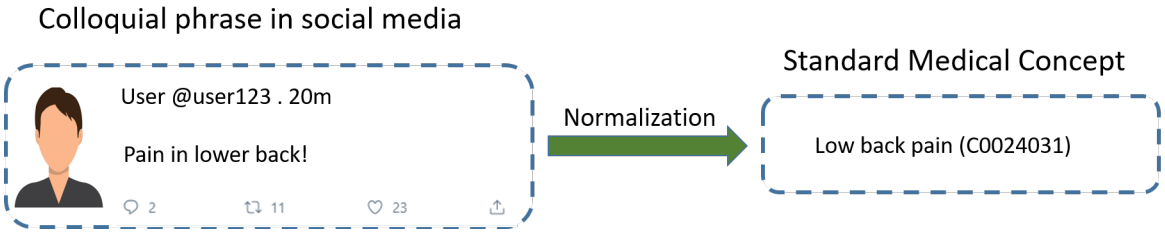

Figure 1: Example to illustrate medical concept normalization

Most of the traditional approaches for entity normalization applied string matching techniques [2, 3, 4]. For example, MetaMap tool maps biomedical text to UMLS concepts and it makes use of knowledge base and computational linguistic techniques [2]. Tsuruoka et al. [3] used character bigrams while McCallum et al. [4] used string edit distances. String matching techniques fail when there is no overlap between entity mention and the corresponding concept (e.g., '*could not sleep much*' → '*insomnia*', '*head spinning a little*' → '*dizziness*'). The application of machine learning techniques to entity normalization started with DNorm proposed by [5] followed by [6] and [7]. However, these methods failed to take semantics into consideration which significantly affected the performance.

Recent studies [8, 9, 10, 11] approached the task of concept normalization as a multi-class text classification problem. All these systems are deep learning based with embeddings as input features. The two drawbacks in these deep learning based systems are a) **Use of traditional embeddings** – Traditional word embeddings are learned using shallow neural network models like Word2Vec. Shallow neural networks are unable to encode more information in vector representations and hence quality of word vectors is limited. The context insensitive nature of traditional word embeddings further limits their quality. b)**Training downstream model from scratch** - With embeddings as input features, the downstream model based on CNN or RNN has to be trained from scratch. A model trained from scratch requires more training examples for better performance. With small size datasets, the performance of downstream models trained from scratch is limited.

In recent times, learning representations using deep language models achieved promising results in many NLP tasks. Some of the popular deep language representation models are ELMo [12], ULMFiT [13], GPT [14] and BERT [15]. ELMo and UMLFiT use recurrent neural network while GPT and BERT are transformer based. ELMo and ULMFiT use BiLSTM for language modeling which is sequential in nature. Further, the representations learned are shallow bidirectional. As GPT uses unidirectional language modeling objective, it is unable to encode information from both left and right contexts. BERT overcomes the drawback in ELMo,

ULMFiT and GPT by learning bidirectional representations using Masked Language Modeling objective and achieved state-of-the-art performance in eleven NLP tasks. In case of BERT a) representations learned are bidirectional and context sensitive b) model is pre-trained on large volumes of unlabeled text using stack of transformer encoders. This iterative approach of generating representations, helps the model to learn lots of language information. c) Task specific layers are added on the top of BERT and entire model is fine-tuned using task specific labeled dataset. As BERT model learns lots of language information during unsupervised pre-training itself, it can be fine-tuned even with small datasets and hence performs better compared to CNN or RNN based models which are to be trained from scratch.

We consider medical concept normalization as multi-class text classification problem and propose a system based on BERT and highway layer. Miftahutdinov and Tutubalina [16] achieved state-of-the-art performance in medical concept normalization using BERT based fine-tuned model. They experimented with only general BERT model pre-trained over text from Wikipedia and books. We believe that domain specific BERT models can better represent medical terms and there is a need for comprehensive evaluation of these models in the task of medical concept normalization. Recently, few research works evaluated the effectiveness of biomedical and clinical BERT models in the tasks of named entity recognition [17], hospital readmission prediction [18] and biomedical concept normalization [19]. However, there is no work which conducted evaluation of general as well as domain specific BERT models to normalize medical concepts in social media text. The work of Ji et al. [19] has also conducted evaluation of BERT based models for the normalization task on three different biomedical datasets. However, our work differs from the work of Ji et al. [19]. We treat medical concept normalization (MCN) as multi-classification task while Ji et al. [19] addresses it as information retrieval task. CADEC-MCN and TWADR-L datasets contain phrases which are written by online users in a colloquial language using descriptive words while concepts names are written by trained professionals in formal language. As the languages used in user generated phrases and concept names differ significantly in many aspects (colloquial vs formal, descriptive words vs standard single words, noisy vs clean), candidate concepts retrieved by BM25 don't include ground truth concepts in many of the cases which significantly affects the performance of the model (explained in detail in Section 6.2.1). As reported in Table 3, the model proposed by Ji et al. [19] is able to achieve only 46.98%, 61.38% and 32.45% on CADEC-Custom, CADEC-Random and TwADR-L datasets while our model achieves 82.62%, 89.95% and 48.32% on these datasets.

In this paper, we provide comprehensive evaluation of general as well as domain specific BERT models. Our key contributions can be summarized as

- Study the effectiveness of BERT based fine-tuned models to normalize medical concepts.
- As per our knowledge, it is the first work to provide comprehensive evaluation of general as well as domain specific BERT models to normalize medical concepts.
- We show that inclusion of highway layer before softmax layer improves the performance of model by filtering irrelevant information.
- Our best model based on BioBERT and highway layer outperforms all existing systems and achieves state-of-the-art accuracy on two standard datasets.

- Study the impact of different learning rates, batch sizes and freezing encoder layers on our best performing model.
- Study the robustness of our best performing model against different noises.

2. Related Work

2.1. Word2Vec to BERT

Machine learning or deep learning based models applied for NLP tasks require representation of text in numerical vectors. Traditional text representations which are based on various measures like word frequency, tf-idf suffer from high dimensionality, lack of language information and require more computation power for processing. The concept of learning distributed representations started with [20, 21, 22, 23, 24]. Bengio et al. [22] used shallow neural network architecture for language modeling. The neural network consists of *tanh* and *softmax* activations in hidden and output layers. Apart from predicting next word in the sequence, the model also learns distributed representations of words. Later, Collobert and Weston [23] learned distributed representation of words in an unsupervised manner using language modeling and then used these learned representations in various supervised downstream tasks. Models like Word2vec [25] and Glove [26] with simple and effective architectures made embeddings a default choice for text representation in NLP models. Word2vec is a prediction based model which learns vector representations using shallow neural network with three layers while glove being a counted based regression model learns vector representations using both local context information as well as global co-occurrence statistics from training corpus. Both Word2vec and Glove models are unable to a) leverage sub-word information and b) provide vectors for words which are missing in the training corpus. To overcome these two drawbacks, Bojanowski et al. [27] proposed FastText embedding model which modifies skipgram model with the introduction of character n-grams. In this model, word representation is based on vectors of its character n-grams.

The limitations of Word2vec, Glove and FastText models are a) *Use of shallow neural network to learn representations* - Word2vec and FastText models use a three layered neural network while glove is log-bilinear global regression model. These shallow models limits the amount of language information encoded in vector representations and hence the quality of vectors is limited. b) *Context insensitive representations* - All these models assign single representation to a word irrespective of its context.

To encode complex relations and make representations sensitive to context, models like ELMo [12], ULMFiT [13], GPT[14] and BERT [15] were proposed. The state-of-the-art performance of these models in many tasks illustrated the effectiveness of learning representations using deep language models over large volumes of text. Further these models except ELMo, changed the approach for NLP tasks from using a model trained from scratch to using a pre-trained model. Peters et al. [12] proposed ELMo which consists of two layers of BiLSTM with inputs generated by CNN and Highway network. Radford et al. [14] introduced GPT model based on Transformer decoder and Devlin et al. [15] proposed BERT based on Transformer encoder. The pre-trained language models can be used in two ways namely *feature based* and *fine-tuned*. In

feature based approach, embeddings learned by model are used as input features to downstream architectures and model has to be trained from scratch (except embeddings) using task specific labeled dataset. In *fine-tuning* approach, one or two task specific layers are added on the top of pre-trained model and entire model is fine-tuned using task specific labeled dataset. ELMo is feature based approach, GPT follows fine-tuning approach while BERT can be in used in both feature-based and fine-tuning approaches.

2.2. Social Media for Health care

With evolution of internet and various social media websites, common people started to share lots of data in the form of tweets, blog posts, questions and answers in discussion forums etc. The data shared by public includes information related to various domains including health. Mining publicly available health related social media data results in useful insights [1].

Traditional disease surveillance systems involves collection of data from health care centers and then processing of collected data. It is truly a time-consuming process and delay in data processing can have severe impacts. Modern disease surveillance systems [28, 29, 30, 31] based on real time social media data helps in early prediction of diseases and reduce the harm. Moreover, early prediction gives more time to handle the situation. Apart from disease surveillance, research studies utilized social media data for extraction of medical concepts [32, 33, 34, 35] like disease, symptoms, adverse drug reactions etc. Recently, there has been raising interest in research community in the form of shared tasks [36, 37, 38] related to identification of text containing drug mentions, medication intake, adverse drug reactions etc.

2.3. Normalizing concepts in social media text

O'Connor et al. [39] proposed a model based on Apache Lucene to normalize Adverse Drug Reaction (ADR) expressions in tweets to UMLS Concept Unique Identifiers (CUI). For a given ADR expression, Apache Lucene retrieves the relevant UMLS concepts. Limsopatham and Collier [7] proposed a model which involves phrase based machine translation and cosine similarity to normalize medical concepts. Medical concept is assigned to twitter phrase based on similarity score obtained as sum of cosine similarity between twitter phrase and concept and translation score calculated using phrase based translation model. The proposed model improved accuracy by upto 55% compared to baselines. Limsopatham and Collier [8] experimented with Google News embeddings as well as embeddings inferred from biomedical articles. They showed that CNN with Google News embeddings achieved better performance when compared to CNN with randomly initialized or biomedical embeddings on three datasets. Further they showed that updating GNews embeddings improved accuracy only on AskAPatient which is larger in size compared to other datasets (TwADR-L and TwADR-S).

Lee et al. [9] experimented with CNN and RNN based models. As the size of training corpus influence the quality of inferred embeddings, they generated embeddings using word2vec over clinical text collected from various sources. RNN with clinical embeddings inferred from combined corpora outperformed all others on two datasets created from tweets and online health forum reviews. Tutubalina et al. [10] proposed BiGRU+Attention model with embeddings inferred from Askapatient.com reviews and UMLS based semantic features as input. The proposed

model achieved an accuracy of 70.05% on custom folds and 85.71% on random folds of Aska-Patient dataset. Niu et al. [40] system is based on multi task char level attention network. With character embeddings matrix as input, auxiliary task with attention mechanism generates weights. CNN applies convolution and pooling operations on character embeddings matrix added with attention weights and predicts the concept.

Recently Miftahutdinov and Tutubalina [16] investigated context sensitive models like ELMo and BERT to normalize medical concepts. ELMo being a feature based embedding model, was used as input features to BiGRU+Attention model. BiGRU+Attention with ELMo+HealthVec as input features outperformed BiGRU+Attention model with only HealthVec embeddings. Further they showed that BERT based fine-tuned model achieved state-of-the-art performance.

Our work is closely related to [16] in applying BERT based fine-tuned model to medical concept normalization. However, Miftahutdinov and Tutubalina [16] experimented with only general BERT models while we do comprehensive evaluation of general as well as domain specific BERT models to normalize concepts. Further, we conduct series of experiments to study the a) impact of inclusion of highway network layer on the top of BERT before softmax layer b) impact of different learning rates, batch sizes and freezing encoder layers on our best model and c) robustness of our best model against different noises.

3. BERT Model

3.1. Description

BERT model consists of an embedding layer followed by a stack of bidirectional transformer encoders. Embedding layer maps sequence of tokens in input to list of vectors. Each transformer encoder [41] applies multi-head self attention and feed forward neural network to list of vectors and returns output to next encoder in the stack. Self-attention mechanism helps to encode bidirectional contextual information in token representations while feed forward network generates hierarchical features. ResNet [42] followed by layer normalization [43] is applied on each of the sub layers - multi-head self attention and feed forward network, to overcome the issue of vanishing and exploding gradients.

3.1.1. Embedding Layer

Input is added with special tokens [CLS] and [SEP] at the start and end respectively. Embedding layer maps sequence of tokens in input $\{[CLS], t_1, t_2, .., t_n, [SEP]\}$ to sequence of vectors $\mathbf{X} = \{x_{[CLS]}, x_1, x_2, .., x_n, x_{[SEP]}\}$ where each x_i is obtained as sum of three embeddings namely word embedding, position embedding and segment embedding.

$$X = W + P + S$$

where $X \in \mathbf{R}^{l \times d_{emb}}$ is input embedding matrix, $W \in \mathbf{R}^{l \times d_{emb}}$ is word embedding matrix, $S \in \mathbf{R}^{l \times d_{emb}}$ is segment embedding matrix,$P \in \mathbf{R}^{l \times d_{emb}}$ is position embedding matrix and each row of all these matrices correspond to a word. All these three embeddings are of equal dimension d_{emb} and have their own significance.

Word embeddings encode language information and BERT model uses WordPiece embeddings [44]. The advantage with WordPiece embeddings is a)Fixed and small size vocabulary

of 0.3M words b)Any word that is not available in vocabulary is represented in terms of sub-words available in vocabulary. Position embeddings encode information related to position of words in the sequence. It is required to include position embeddings because unlike RNN or CNN, self-attention is unable to capture order of words. Segment embedding differentiate words of different sequences. All these three embeddings are updated during pre-trainig as well as fine-tuning. Word embeddings are initialized with WordPiece embeddings while position and segment embeddings are initialized randomly.

3.1.2. Bidirectional Transformer Encoder

Each bidirectional transformer encoder consists of multi-head self attention and feed forward network layers. Self attention mechanism allows each token to attend to all tokens in the sequence and encode context information in vector representations. It is calculated using three weight matrices $W_Q \in \mathbf{R}^{d_{emb} \times d_k}$, $W_K \in \mathbf{R}^{d_{emb} \times d_k}$ and $W_V \in \mathbf{R}^{d_{emb} \times d_v}$

$$SA(X) = Softmax(\frac{QK^T}{\sqrt{d_k}})V \in \mathbf{R}^{l \times d_v}$$

where $Q \in \mathbf{R}^{l \times d_k}$, $K \in \mathbf{R}^{l \times d_k}$ and $V \in \mathbf{R}^{l \times d_v}$ are query, key and value matrices obtained by multiplying $X \in \mathbf{R}^{l \times d_{emb}}$ with the corresponding weight matrices.

$$Q = X \bullet W_Q, K = X \bullet W_K, V = X \bullet W_V$$

where $\bullet$ represents matrix multipication.
To obtain representations from different subspaces, self-attention is computed h times using different weight matrices. The outputs of all self-attention operations are concatenated to get $CONCAT = [SA_1(X), SA_2(X), ...SA_h(X)] \in \mathbf{R}^{l \times hd_v}$. Finally a linear transformation with weight matrix $W_O \in \mathbf{R}^{hd_v \times d_{emb}}$ is applied to get $MHSA(X) \in \mathbf{R}^{l \times d_{emb}}$.

$$MHSA(X) = CONCAT \bullet W_O$$

To avoid vanishing and exploding gradients, ResNet followed by layer normalization is applied.

$$\tilde{G} = LN(X + MHSA(X))$$

To generate non-linear hierarchical features, a position wise feed forward networks is applied separately for each position. Gelu [45] layer in between two linear layers constitutes position wise feed forward network i.e., $PwFFN(x) = Gelu(xW_1 + b_1)W_2 + b_2$. After applying ResNet followed by layer normalization, we get

$$G = LN(\tilde{G} + PwFFN(\tilde{G}))$$

BERT consists of a stack of such bidirectional transformer encoders and the depth of stack is 12 in case of $BERT_{Base}$ and 24 in case of $BERT_{Large}$. Each transformer encoder generates representation of input sequence by capturing bidirectional contextual information. This iterative process of generating sequence representation using a stack of encoders helps the model to learn

complex relationships.

$$\tilde{G}_m = LN(G_{m-1} + MHSA(G_{m-1}))$$

$$G_m = LN(\tilde{G}_{m-1} + PwFFN(\tilde{G}_{m-1}))$$

where $\tilde{G}$ is the intermediate result of m^{th} encoder , G_m is the output of m^{th} encoder and $G_0 = X$

3.2. Training Procedure

BERT framework consists of two steps: Unsupervised pre-training and Supervised fine-tuning. Unsupervised pre-training helps the model to learn parameters from scratch using tasks like Masked Language Modeling and Next Sentence Prediction. Training the model with these tasks helps it to learn semantics at both word and sentence levels. Once the model is pre-trained, it can be adapted to downstream tasks using supervised fine-tuning.

3.2.1. Self-Supervised Pre-training

Pre-trainig model involves two tasks namely Masked Language Modeling and Next Sentence Prediction. The authors selected these two tasks because Masked Language Modeling helps the model to encode bidirectional context features while Next Sentence Prediction helps to learn relationships between sentences.

Masked Language Modeling Language Modeling computes the probability of a word using previous or subsequent words. Forward language model predicts the word x_t using previous $t-1$ words $\{x_1, x_2, ...x_{t-1}\}$

$$P(x_t|x_1, x_2, ...x_{t-1})$$

Backward language model predicts the word x_t using the next $t-1$ words $\{x_{t+1}, x_{t+2}, .., x_{2t-1}\}$

$$P(x_t|x_{t+1}, x_{t+2}, ...x_{2t-1})$$

GPT is unidirectional as it is based on forward language model while ELMo is shallow bidirectional as ELMo representations are obtained from the concatenation of forward and backward language model representations. The main drawback in unidirectional language modeling objective is its inability to encode information from both left and right contexts simultaneously. BERT overcomes the drawback of unidirectional language model in GPT and ELMo with Masked Language Modeling. In Masked language modeling, a randomly masked word is predicted using words in both left and right contexts.

$$P(x_t|x_1, x_2, ..., x_{t-1}, \tilde{x}_t, x_{t+1}, x_{t+2}, .., x_n)$$

where $\tilde{x}_t$ is masked representation of x_t. The authors randomly masked 15% of tokens in each sequence. There will be masked tokens only during pre-training phase. To reduce mismatch between pre-training and fine-tuning, the authors introduced a special masking procedure. Each of the randomly sampled token a) is replaced with [MASK], 80% of time b) is replaced with random word, 10% of time and c) is left unchanged remaining times.

Next Sentence Prediction This pre-training task aims to help the model to learn semantics at sentence level. Learning relationships between sentences is useful for downstream tasks involving more than one sentence like question and answering, natural language inference etc. It

is basically, a binary classification task with two labels, '*IsNext*' and '*IsNotNext*'. For a given pair of sentences (x,y), the model has to predict whether y is next sentence of x or just a random sentence in the training corpus. Sentence pairs are generated from training corpus in a way that a) combined length of two sentences should not exceed 512 b) 50% of times, second sentence is actual next sentence and 50% of times, second sentence is a random sentence. The corpus used for pretraining BERT model includes text from BookCorpus having 800M words and English Wikipedia having 2500M words.

3.2.2. Supervised Fine-tuning

It helps the model to adjust to downstream task. Here task specific layers are added on the top of BERT. All the parameters of BERT and task specific layers are fine-tuned using task specific labeled data set. Different downstream tasks will have different fine-tuned models, though all of them are initialized with the same pre-trained BERT model.

4. Highway Networks

Highway Networks introducted by Srivastava et al. [46] filters out irrelevant information from input vector. It improves ResNet layer [42] with inclusion of gating mechanism. Kim et al. [47] showed the use of highway network layer as a potential filter of irrelevant information in character aware neural language model. Highway Network layer is defined as:

$$HN(x) = h(x) \odot t(x) + x \odot (1 - t(x)) \tag{1}$$

where h(x)= $ReLU(xW_h + b_h)$, t(x) = $Softmax(xW_t + b_t)$ is Transform gate, $1 - t(x)$ is Carry gate. Here $\odot$ represents element wise multiplication, W_h and W_t are weights, b_h and b_t are biases. Further $h(x)$ represents traditional non-linear path and x represents skip path. $t(x)$ and $1-t(x)$ act as gates and regulate the flow of information through non-linear and skip paths.

5. Methods

5.1. Datasets

In this work, we experiment with custom and random folds of CADEC-MCN and TwADR-L datasets. TWADR-L was generated from tweets while CADEC-MCN was generated from health related reviews on Askapatient.com which is an online health discussion forum.

CADEC-MCN Karimi et al. [32] developed a dataset called CADEC(CSIRO Adverse Drug Event Corpus) from AskAPatient[3] forum posts. This dataset consists of 1253 user posts having 7398 sentences and each identified entity is mapped to adverse effect, drug, symptom, disease or finding, using three vocabularies namely SNOMED-CT, MEDRA and AMT (The Australian Medicines Terminology). We represent this dataset as CADEC-MCN. Random and custom folds of CADEC-MCN are taken from [8] and [10] respectively. CADEC-MCN Custom consists of five folds with each fold having train and test sets (number of unique medical concepts is 181). CADEC-MCN Random consists of ten folds with each fold having train, validation and test sets

[3]https://www.askapatient.com

(number of unique medical concepts is 1036). For more details, refer Tables A.13 and A.14 in appendix section.

TwADR-L Limsopatham and Collier [8] created TwADR-L dataset which contains twitter ADR phrases mapped to medical concepts from Side Effect Resource (SIDER)[4]. The authors collected tweets generated over a span of three months related to fixed set of drugs, manually extracted and annotated ADR phrases with SIDER medical concepts. The datasets includes 1436 ADR phrases mapped to one of 2200 SIDER medical concepts. The authors divided dataset into ten folds with each fold having train, validation and test sets (number of unique medical concepts is 2200). For more details, refer Table A.15 in appendix section.

5.2. Problem Definition

Medical concept normalization is treated as multi class classification problem. Given, health related entity mention M and a label space $\{C_1, C_2, ..., C_K\}$, the normalization system maps M to one of the concepts in label space. Here K represents number of unique concepts in dataset.

- **Input**: Health related entity mention expressed as [CLS] M [SEP].
- **Output**: Probability vector $\mathbf{q} \in \mathbb{R}^{1xK}$ such that q_i represents probability that the entity mention belongs to concept C_i. The concept with maximum probability is assigned to the mention.

5.3. Model Configuration

In this work, we experiment with two BERT based fine-tuned models namely BertForSequenceClassification and BERT+Highway Network. The first model is pre-trained BERT added with Classifier on the top while second model is pre-trained BERT added with Highway Network+Classifier on the top (see Figure 2) .

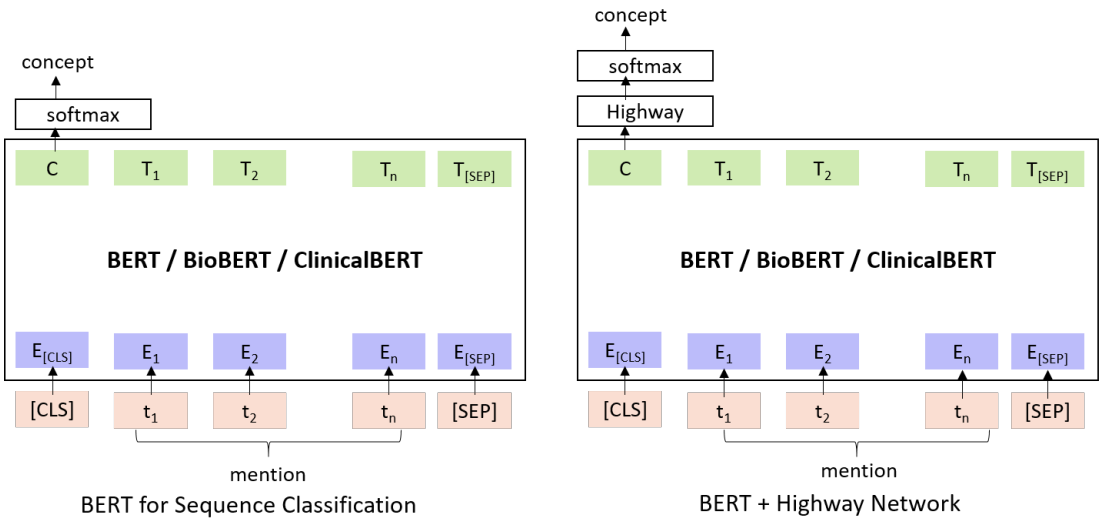

Figure 2: Architecture of BERT based fine-tuning models for medical concept normalization.

[4]http://sideeffects.embl.de/

5.3.1. BertForSequenceClassification

It is the default BERT model applied for text classification. In BERT model, the final hidden vector of the [CLS] token is considered to represent input text. So, this vector is given to softmax layer which outputs a vector containing label probabilities.

$$q = BERT(mention) \tag{2}$$

$$logits = qW^T + b \tag{3}$$

$$p = Softmax(logits) \tag{4}$$

Here $q \in \mathbb{R}^{1xH}$ is final hidden state vector of [CLS] token and H is BERT hidden vector dimension. $W \in \mathbb{R}^{KxH}$ and $b \in \mathbb{R}$ are weights and bias of classifier layer. $p \in \mathbb{R}^{1xK}$ is a vector with label probabilities where K is size of label space.
The model is trained by fine-tuning all the parameters of BERT model and classifier layer.

5.3.2. BERT + Highway Network

As show in Figure 2, this model is an improvement over default BERT model with addition of highway network layer before classifier layer. Gating mechanism in highway network layer filters out irrelevant information. So, we believe that by passing final hidden vector of [CLS] token through highway network and then through classifier layer, improves the performance of model.

$$q = BERT(mention) \tag{5}$$

$$r = h(q) \odot t(q) + q \odot (1 - t(q)) \tag{6}$$

$$logits = rW^T + b \tag{7}$$

$$p = Softmax(logits) \tag{8}$$

Here $q \in \mathbb{R}^{1xH}$ is final hidden state vector of [CLS] token and H is BERT hidden vector dimension. $r \in \mathbb{R}^{1xH}$ is output vector of highway network. $W \in \mathbb{R}^{KxH}$ and $b \in \mathbb{R}$ are weights and bias of classifier layer. $p \in \mathbb{R}^{1xK}$ is a vector with label probabilities where K is size of label space.
The model is trained by fine-tuning all the parameters of BERT model, highway network and classifier layer.

5.4. Evaluation Metric

Following the previous state-of-the-art methods [8, 10, 16], we considered accuracy as evaluation metric. Here accuracy refers to percentage of entity mentions that are assigned concepts correctly.

$$Accuracy = \frac{\#\text{EntityMentions}_{\text{Correctly Mapped}}}{\#\text{EntityMentions}_{\text{Total}}} \tag{9}$$

The accuracy values obtained over all the folds are averaged to get the final accuracy.

Model	Training Corpus	Initialized from
$\text{BERT}_{\text{base\_uncased}}$	Books Corpus and English Wikipedia	-
$\text{BERT}_{\text{base\_cased}}$	Books Corpus and English Wikipedia	-
$\text{BioBERT}_{\text{PubMed\_1M}}$	PubMed abstracts (1 Million)	$\text{BERT}_{\text{base\_cased}}$
$\text{BioBERT}_{\text{PubMed\_200K}}$	PubMed abstracts (200K)	$\text{BERT}_{\text{base\_cased}}$
$\text{BioBERT}_{\text{PMC\_270K}}$	PMC full text articles (270K)	$\text{BERT}_{\text{base\_cased}}$
$\text{BioBERT}_{\text{PubMed+PMC\_470K}}$	PubMed abstracts (200K) + PMC full text articles (270K)	$\text{BERT}_{\text{base\_cased}}$
$\text{ClinicalBERT}_{\text{scratch}}$	100K Clinical Notes from MIMIC-III	-
$\text{ClinicalBERT}_{\text{300K}}$	All Clinical Notes from MIMIC-III	$\text{BERT}_{\text{base\_cased}}$
$\text{ClinicalBERT}_{\text{clinical}}$	All Clinical Notes from MIMIC-III	$\text{BioBERT}_{\text{PubMed+PMC\_470K}}$
$\text{ClinicalBERT}_{\text{discharge}}$	All Discharge Notes from MIMIC-III	$\text{BioBERT}_{\text{PubMed+PMC\_470K}}$

Table 1: Summary of various BERT models. A model trained from scratch is indicated by '-'.

5.5. Pre-trained BERT Models

In this paper, we experiment with three different pre-trained BERT models namely, general BERT [15] models trained on Books and Wikipedia corpus, BioBERT [48] models trained on biomedical corpus and ClinicalBERT [49, 17, 18] models trained on medical corpus. Lee et al. [48] released four BioBERT models ($\text{BioBERT}_{\text{PubMed\_1M}}$, $\text{BioBERT}_{\text{PubMed\_200K}}$, $\text{BioBERT}_{\text{PMC\_270K}}$ and $\text{BioBERT}_{\text{PubMed+PMC\_470K}}$) trained on 1 million PubMed abstracts, 200K PubMed abstracts, 270K PubMed Central (PMC) full text articles and 200K PubMed abstracts + 270K PMC articles respectively. All these four models were initialized from $\text{BERT}_{\text{base\_cased}}$. Alsentzer et al. [49] released two ClinicalBERT models ($\text{ClinicalBERT}_{\text{clinical}}$ and $\text{ClinicalBERT}_{\text{discharge}}$) trained on clinical notes and discharge summaries from MIMIC-III [50]. Both these models were initialized from $\text{BioBERT}_{\text{PubMed+PMC\_470K}}$) model. Huang et al. [18] released $\text{ClinicalBERT}_{\text{scratch}}$ model trained from scratch with 100K clinical notes from MIMIC-III. Si et al. [17] released $\text{ClinicalBERT}_{\text{300K}}$ model initialized from $\text{BERT}_{\text{base\_cased}}$ and trained for 300K steps using MIMIC-III clinical notes. Table 1 shows a brief summary of different pre-trained BERT models.

6. Results and Discussions

We conduct experiments in two phases. In first phase, we evaluate general BERT, biomedical BERT and clinical BERT based fine-tuned models with and without including highway network layer on CADEC-MCN custom folds dataset. Then, we evaluate our best performing model on TwADR-L and CADEC-MCN random folds. Section 6.1 discuss the impact of including highway layer and Section 6.2 compare our best model with existing systems. In second phase, we conduct a series of experiments using CADEC-MCN custom folds to study the impact of different batch sizes, learning rates, noises and freezing encoder layers on our best model.

As there is significant overlap between train and test sets in TwADR-L and CADEC-MCN random, to study the impact of different batch sizes, learning rates, noises and freezing encoder layers on our best model, we use CADEC-MCN custom folds only. Section 6.3 discuss the impact of different batch sizes of 16, 32, 64 and 128 (learning rate fixed at 3e-5) on our best model. Section 6.4 disuss the impact of different learning rates of 2e-5, 3e-5, 4e-5 and 5e-5 (batch size fixed at 128) on our best model. Section 6.5 discuss the impact of freezing first 1, 2, 4, 6, 8, 10

Model	Accuracy	
	without HN*	with HN‡
$\text{BERT}_{\text{base\_uncased}}$	80.91	81.12
$\text{BERT}_{\text{base\_cased}}$	81.37	81.36
$\text{BioBERT}_{\text{PubMed\_1M}}$	**82.35**	**82.62**
$\text{BioBERT}_{\text{PubMed\_200K}}$	81.03	81.57
$\text{BioBERT}_{\text{PMC\_270K}}$	81.08	81.14
$\text{BioBERT}_{\text{PubMed+PMC\_470K}}$	81.77	81.46
$\text{ClinicalBERT}_{\text{scratch}}$	80.42	80.83
$\text{ClinicalBERT}_{\text{300K}}$	81.23	82.40
$\text{ClinicalBERT}_{\text{clinical}}$	81.20	81.27
$\text{ClinicalBERT}_{\text{discharge}}$	82.10	82.21

Table 2: Accuracy of various BERT based models on custom folds of CADEC-MCN dataset. **HN** stands for Highway Network, $\star$ represents BERT for Sequence Classification model and ‡ represents BERT+Highway Network Model.

and 11 encoder layers (batch size and learning rate are fixed at 128 and 3e-5 respectively) on our best model. Section 6.6 discuss the robustness of our best model against different noises.

Table 2 shows accuracy of different BERT based models evaluated on CADEC-MCN custom folds. From Table 2, it is clear that (1) In case of general BERT models, $\text{BERT}_{\text{base\_cased}}$ (without HN) with an accuracy of 81.37% outperformed other general models. (2) In case of BioBERT models, $\text{BioBERT}_{\text{PubMed\_1M}}$ which was initialized from $\text{BERT}_{\text{base\_cased}}$ and trained on 1 Million PubMed abstracts achieved an accuracy of 82.62% (with HN) and outperformed other biomedical models. (3) In case of ClinicalBERT models, $\text{ClinicalBERT}_{\text{300K}}$ which was initialized from $\text{BERT}_{\text{base\_cased}}$ and trained for 300K steps using all clinical notes from MIMIC-III achieved an accuracy of 82.40% (with HN) and outperformed other clinical models. (4) $\text{BioBERT}_{\text{PubMed\_1M}}$+HN achieved highest accuracy of 82.62% on CADEC-MCN custom folds. Further, we evaluated our best model $\text{BioBERT}_{\text{PubMed\_1M}}$+HN on CADEC-MCN random folds and TwADR-L and achieved accuracy of 89.95% and 48.32% respectively.

In case of general BERT models, $\text{BERT}_{\text{base\_cased}}$ outperformed $\text{BERT}_{\text{base\_uncased}}$. This shows that cased BERT models encode more information compared to uncased BERT models. This is the reason why all the domain specific BioBERT and ClinicalBERT (except $\text{ClinicalBERT}_{\text{scratch}}$ which is trained from scratch) models were initialized from BERT cased models rather than BERT uncased models.

In case of BioBERT models, $\text{BioBERT}_{\text{PubMed\_1M}}$ outperformed all other biomedical models with an accuracy of 82.62% (with HN). It is expected because $\text{BioBERT}_{\text{PubMed\_1M}}$ is trained on a large corpus of 1M PubMed abstracts compared to $\text{BioBERT}_{\text{PubMed\_200K}}$, $\text{BioBERT}_{\text{PMC\_270K}}$, $\text{BioBERT}_{\text{PubMed+PMC\_470K}}$ which were trained on relatively small corpus of 200K PubMed abstracts, 270K PubMed Central full text articles and (200K PubMed abstracts + 270K PubMed Central full text articles) respectively. Further, $\text{BioBERT}_{\text{PubMed\_1M}}$ and $\text{BioBERT}_{\text{PubMed+PMC\_470K}}$ outperformed $\text{BERT}_{\text{base\_cased}}$. Both $\text{BioBERT}_{\text{PubMed\_1M}}$ and $\text{BioBERT}_{\text{PubMed+PMC\_470K}}$ were initialized from $\text{BERT}_{\text{base\_cased}}$ and then further pre-trained on domain specific biomedical corpus. This shows that further pre-training general BERT models on domain specific corpus improves the performance. However, $\text{BioBERT}_{\text{PMC\_270K}}$ achieved lower performance than

$\text{BERT}_{\text{base\_cased}}$. This may be because it was further pre-trained using a relatively small corpus of 270K PubMed Central full text articles compared to $\text{BioBERT}_{\text{PubMed\_1M}}$ and $\text{BioBERT}_{\text{PubMed+PMC\_470K}}$ which were further pre-trained using relatively large corpus of 1M PubMed abstracts and (200K PubMed abstracts + 270K PubMed Central full text articles) respectively. In case of ClinicalBERT models, $\text{ClinicalBERT}_{\text{300K}}$ trained using all the clinical notes from MIMIC-III outperformed other clinical models with an accuracy of 82.40% (with HN).

$\text{BioBERT}_{\text{PubMed\_1M}}$+HN achieved the best performance on CADEC-MCN custom folds data set. We expected $\text{ClinicalBERT}_{\text{300K}}$+HN to achieve the best performance however it achieved 0.22% accuracy lower than $\text{BioBERT}_{\text{PubMed\_1M}}$+HN. We believe that further pre-training the model for more number of steps or further pre-training the model using medical related Wikipedia pages can improve the performance. We would like to explore these options in future. Further, $\text{ClinicalBERT}_{\text{scratch}}$ achieved the lowest performance compared to all the models including general BERT models. This is because it was trained from scratch using a relatively small corpus of 100K clinical notes. In future, we would like to investigate whether further pre-training this model using more clinical notes and medical related Wikipedia pages can improve the performance.

6.1. Impact of Highway Network

The performance of various BERT based fine-tuned models after including Highway network layer is reported in Table 2. From Table 2, it is clear that highway network has improved the performance in all the cases except $\text{BERT}_{\text{base\_cased}}$ and $\text{BioBERT}_{\text{PubMed+PMC\_470K}}$. The improvement is highest in case of $\text{ClinicalBERT}_{\text{300K}}$(1.17%) and lowest in case of $\text{BioBERT}_{\text{PMC\_270K}}$ (0.06%). Highway network layer consists of two gates namely $t(x)$ - transform and $1 - t(x)$ - carry gates. These two gates regulate the flow of data through non-linear and skip paths. This will help the model to choose only important information and hence the model performance increases.However, highway networks didn't improve the performance in case of $\text{BERT}_{\text{base\_cased}}$ and $\text{BioBERT}_{\text{PubMed+PMC\_470K}}$. This may be, because of inclusion of an additional layer, the model is over fitted. In these two cases, changing the dropout applied to Highway network layer or a better learning rate can improve the performance.

6.2. Comparison with previous systems

We compare our best performing model with previous systems which includes systems based on a) machine learning (DNorm and Logistic Regression [8]) b) deep learning with i) traditional embeddings (CNN [8] and Multi-task Char-CNN + Att [40]) ii) ELMo embeddings (GRU + Att, GRU + Att + tf-idf(Max) [16]) c) fine-tuned BERT (BERT, BERT + tf-idf(Max) [16]) and d) information retrieval based methods (BM25[51], BM25+BioBERT_1M[19])

- **DNorm** [8] - Applies pairwise rank learning technique to normalize medical concepts.
- **Logistic Regression** [8] - Multi-class logistic regression classifier with phrase vector as input and phrase vector is obtained by concatenating embeddings of words in phrase.
- **CNN** [8] - CNN with Google News embeddings as input.

Model	CADEC-MCN		TwADR-L
	Custom	Random	
DNorm [8]	-	73.39	30.99
Logistic Regression [8]	-	77.67	34.09
CNN [8]	-	81.41	45.90
Multi-task Char-CNN [40]	-	84.65	46.46
GRU+Att [16]	71.68	85.06	-
GRU+Att+tf-idf(max) [16]	74.70	85.71	-
BM25$^{\Upsilon}$	30.43	55.46	23.00
BM25 + BioBERT_1M$^{\Upsilon}$	46.98	61.38	32.45
BERT [16]	79.83	88.69	44.15$^{\perp}$
BERT+tf-idf(max) [16]	79.25	88.84	44.51$^{\perp}$
Our Best Model	**82.62**	**89.95**	**48.32**

Table 3: Performance comparison of our best model BioBERT$_{\text{PubMed_1M}}$+HN with existing methods on TwADR-L and CADEC-MCN datasets. $\perp$ - we evaluated BERT baseline models on TwADR-L and reported the accuracy. Υ - we evaluated BM25 and BM25 + BioBERT_1M on CADEC-MCN(Custom and Random) and TwADR-L datasets and reported the accuracy.

- **Multi-task Char-CNN + Att** [40] - CNN applies convolution and max-pooling operations on character embeddings matrix added with attention weights generated by auxiliary task and then predicts the concept.
- **GRU + Att, GRU + Att + tf-idf(Max)** [16] - GRU + Att with ELMo, HealthVec embeddings as input. UMLS based similarity features are calculated using tf-idf.
- **BM25** [51] - Ranking function which retrieve relevant concepts for the given colloquial phrase and the concept with maximum score is assigned to colloquial phrase.
- **BM25 + BioBERT_1M** [19] - BioBERT_1M re-ranks the relevant concepts retreived by BM25 and then the concept with maximum score is assigned to colloquial phrase.
- **BERT, BERT + tf-idf(max)** [16] - BERT based fine-tuned model without and with UMLS based similarity features calculated using tf-idf.

Table 3 shows comparison of our best performing model with existing systems on TwADR-L, custom and random folds of CADEC-MCN. Our best model based on BioBERT$_{\text{PubMed_1M}}$ and highway network outperformed all the existing systems. Based on the values reported in Table 3, it is clear that our best model based on BioBERT and highway layer outperformed existing deep learning systems based on traditional embeddings as well as systems based on ELMo embeddings. Traditional word embeddings which are learned using shallow neural networks are unable to encode more information in vector representations. Moreover, these representations are context insensitive which further limits the quality of vectors. Though ELMo is context sensitive, it is shallow bidirectional i.e., the representations are obtained as concatenation of representations from forward and backward LSTMs. Further, traditional word embeddings or ELMo embeddings are used as input features to downstream models which are then trained from scratch using task specific labeled data set. As downstream models are to be trained from scratch (except embeddings), they require more training instances to perform better. However in case of

BERT a) representations learned are bidirectional and context sensitive b) model is pre-trained on large volumes of unlabeled text using stack of transformer encoders. This iterative approach of generating representations, helps the model to learn lots of language information. c) Task specific layers are added on the top of BERT and entire model is fine-tuned using task specific labeled dataset. As BERT model learns lots of language information during unsupervised pre-training itself, it does not require large labeled data sets for fine-tuning. So, our best model achieved better performance compared to traditional embedding or ELMo based deep learning systems.

6.2.1. Why Information Retrieval based methods (BM25, BM25+BioBERT_1M) failed?

BM25 [51] also called as Okapi BM25 is a probabilistic ranking function used to retrieve relevant documents for a given query. BM25 ranks the documents based on a score which involves statistical measures like term frequency, document frequency, query length, average length of documents etc. In medical concept normalization, for a given colloquial phrase, BM25 retrieve top n (here n value is 10) candidate concept names and the concept with maximum score is assigned to the phrase. From Table 5, it is clear that it successfully maps concepts when a)colloquial phrase lexically matches with concept names. For example, colloquial phrase- *abdominal pain* and concept name – *abdominal pain*. b)colloquial phrase significantly overlaps with concept name. For example, colloquial phrase - *constant muscle tension in legs* and concept name - *muscle tension*. Here, colloquial phrase and concept name have *'muscle tension'* in common.

However, as the function ranks the concepts based on statistical measures, it ignores sub word information as well as semantic information which makes it to fail in many of the instances falling under three cases namely (from Table 4) a) No common words in phrase and concept name. For example, *heart attack* (phrase) and *myocardial infarction* (concept name) have no words in common. b) Lexical variants. For example, *diahorea* (phrase) is a lexical variant of *diarrhea* (concept name). c) One or two common words in phrase and the predicted concept. For example, *coronary disease* is mapped to *parkinson's disease* (instead of *heart disease* which is ground truth) as they have '*disease*' in common.

BM25 + BioBERT_1M approach involves two phases namely a)generation of candidate concepts – BM25 retrieve top 10 candidate concepts for the given colloquial phrase b) re-ranking of candidate concepts – BioBERT with colloquial phrase and candidate concept as input, predicts the similarity. The candidate concept with maximum similarity is assigned to the colloquial phrase. As shown in Figure 3, re-ranking using BioBERT brings the appropriate concept at the top and hence accuracy improves in comparison to using only BM25.

However, as BM25 scoring function does not consider sub word and semantic information, in many of the cases as shown in Figure 4, the ground truth concept is not in the retrieved top 10 concepts and so, even after re-ranking, the top candidate concept is not same as ground truth. Hence, the performance of BM25+BioBERT_1M is very low compared to our model. We strongly believe that as the languages used in user generated phrases and concept names differs significantly in many aspects (colloquial vs formal, descriptive words vs single words, noisy vs clean etc.), candidate concept generation should be semantic based rather than string matching based like BM25. We consider this as future work.

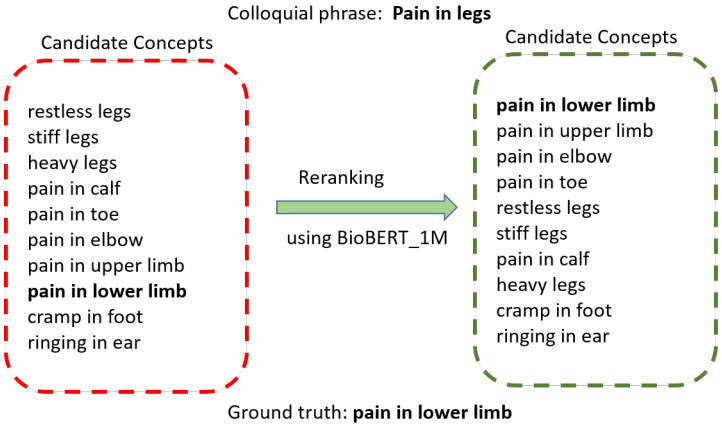

Figure 3: Re-ranking using BioBERT places appropriate concept at the top.

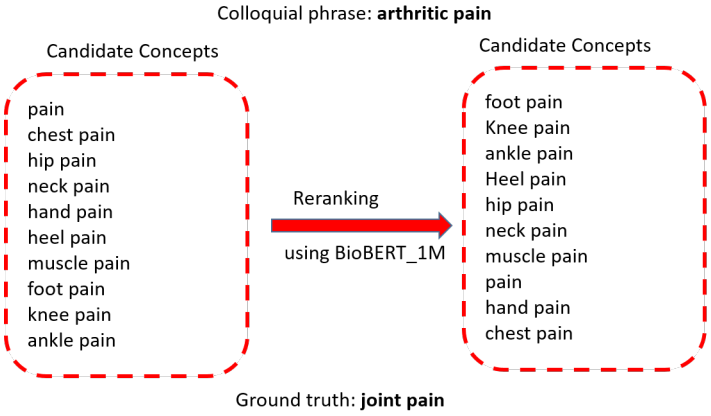

Figure 4: Ground truth concept is not in the concepts retrieved by BM25 and hence even after re-ranking by BioBERT, the top concept is not same as ground truth concept.

Colloquial Phrase	Prediction	Ground Truth
Case 1 (No common words)		
extremely sick	paresthesia of lower extremity	generally unwell
heart attack	heart disease	myocardial infarction
decreased sex drive	paresthesia of lower extremity	reduced libido
reduced mental capabiities	reduced libido	impaired cognition
Case 2 (Lexical variants)		
dizzieness	paresthesia of lower extremity	dizziness
excruciatig pain	pain	excruciating pain
ibruprofen	paresthesia of lower extremity	ibuprofen
diaharrea	paresthesia of lower extremity	diarrhea
diahorea	paresthesia of lower extremity	diarrhea
Case 3 (Overlapping)		
chest *ache*	stomach *ache*	chest pain
coronary *disease*	parkinson's *disease*	heart disease
lack of enthusiasm	*lack of* libido loss of motivation	loss of motivation
elevated levels of high cholesterol	*elevated* blood pressure	serum cholesterol raised

Table 4: Incorrectly classified phrases by BM25

Colloquial Phrase	Prediction	Ground Truth
Case 1 (Lexical Match)		
hip pain	hip pain	*hip pain*
fatigue	fatigue	*fatigue*
abdominal pain	abdominal pain	*abdominal pain*
excruciating pain	excruciating pain	*excruciating pain*
Case 2 (Significant overlap)		
constant *muscle tension* in legs	muscle tension	*muscle tension*
legs are *restless*	restless legs	*restless legs*
weakness in *muscles*	muscle weakness	*muscle weakness*
right *heel pain*	heel pain	*heel pain*

Table 5: Correctly classified phrases by BM25.

6.2.2. Baseline BERT vs Our best performing model

Tables 6,7 and 8 contain the instances illustrating the impact of BioBERT_1M+HN model in case of CADEC-MCN Custom, CADEC-MCN Random and TwADR-L datasets respectively. The base line BERT model is pre-trained on Books Corpus and Wikipedia i.e., general text corpus while BioBERT model is initialized from BERT and further pre-trained on PubMed abstracts i.e., domain related text corpus.

- In Case 1, the baseline BERT model assigns a related but wrong concept. For example, *menstrual cramps →Menorrhagia (386692008)* (Table 6), *clotting, and horrible periods →Menstrual cramp (431416001)* (Table 6), *Pain in left arm Pain in left arm →Pain in wrist (56608008)* (Table 7) and *Psychotic Disorders Mental illness →Mental disorders (C0004936)* (Table 8) Here, the words *menstrual*, *periods*, *arm* and *psychotic* are related to *Menorrhagia*[5], *menstrual*, *wrist* and *mental* respectively.

- In Case 2, the baseline BERT model assigns concepts which overlap with the colloquial phrase i.e., one to two words in common. For example, *fatigue in forearms →Fatigue (84229001)* (Table 6), *Severe dehydration extreme dehydration →Dehydration (34095006)* (Table 7) and *Anxiety aggravated incr anxiety →Anxiety (C0003467)* (Table 8).

- In Case 3, the baseline BERT model assigns concepts which are close in meaning with subtle difference. For example, *difficulty thinking →Poor concentration (26329005)* (Table 6), *agitated*[6] *→Depressive disorder (35489007)* (Table 6), *Impatient character less patience →Personality change (102943000)* (Table 7) and *Yawning yawns →Drowsiness (C0013144)* (Table 8).

In all these three cases, the baseline BERT model which is pre-trained on general text corpus assigns wrong concepts due to lack of enough domain specific information. Our model based on BioBERT_1M (initialized from BERT and further pre-trained on PubMed abstracts) with rich domain specific information generate better phrase representations and hence assigns

[5]Menorrhagia means menstrual periods with abnormally heavy or prolonged bleeding
[6]Agitated means feeling or appearing troubled or nervous

Colloquial Phrase	Prediction (BERT)$^{\Gamma}$	Prediction(BioBERT_1M+HN)$^{\Upsilon}$
Case 1 (related)		
menstrual cramps	386692008-Menorrhagia	431416001-Menstrual cramp
clotting, and horrible periods	431416001-Menstrual cramp	386692008-Menorrhagia
terrible pains in big toe	47933007-Foot pain	285365001-Pain in toe
numbness in left foot	309539008-Numbness of toe	309538000-Numbness of foot
sleeplesness	77692006-Hypersomnia	193462001-Insomnia
Case 2 (Overlapping)		
fatigue in forearms	84229001-Fatigue	80449002-Muscle fatigue
cramping in hamstrings	55300003-Cramp	449917004-Cramp in lower limb
aching joints shoulders	45326000-Shoulder pain	267949000-Shoulder joint pain
muscles in my chest started aching	29857009-Chest pain	68962001-Muscle pain
elbows burning	74323005-Pain in elbow	90673000-Burning sensation
Case 3 (Subtle variations)		
difficulty thinking	26329005-Poor concentration	247640008-Unable to think clearly
lack of sexual desire	8357008-Reduced libido	248096004-Lack of libido
agitated	35489007-Depressive disorder	24199005-Feeling agitated
no sleep	301345002-Difficulty sleeping	248255005-Cannot sleep at all
foggy thinking	247640008-Unable to think clearly	419723007-Mentally dull

Table 6: Baseline vs Our model in CADEC-MCN Custom. Here Γ - predicted and ground truth concepts are different, Υ - predicted and ground truth concepts are same.

Colloquial Phrase	Prediction (BERT)$^{\Gamma}$	Prediction(BioBERT_1M+HN)$^{\Upsilon}$
Case 1 (related)		
pale yellow complexion	16386004-Dry skin	398979000-Pale complexion
pain in left arm	56608008-Pain in wrist	287045000-Pain in left arm
thumb pain	56608008-Pain in wrist	300955002-Pain in thumb
aching joints wrists	202480001-Elbow joint pain	202482009-Wrist joint pain
Case 2 (Overlapping)		
over stressed	73595000-Stress	424582000-Stress overload
extreme dehydration	34095006-Dehydration	450316000-Severe dehydration
swollen joints	84445001-Joint stiffness	271771009-Joint swelling
aches and stiffness hips	49218002-Hip pain	249914008-Hip stiff
Case 3 (Subtle variations)		
less patience	102943000-Personality change	286755001-Impatient character
sleepy all or most sf the time	301345002-Difficulty sleeping	248262001-Always sleepy
yellow color to my skin	278528006-Facial swelling	225549006-Yellow skin
feeling depressed	420038007-Feeling unhappy	272022009- feeling depressed

Table 7: Baseline vs Our model in CADEC-MCN Random. Here Γ - predicted and ground truth concepts are different, Υ - predicted and ground truth concepts are same.

concepts correctly. For example, *yawning* and *drowsiness* are close in meaning with subtle difference. The baseline BERT model with insufficient domain information is unable to identify this subtle difference while our model with rich domain information maps *yawning yawns* to *yawning-C0043387* instead of *Drowsiness-C0013144*.

Colloquial Phrase	Prediction (BERT)$^{\Gamma}$	Prediction(BioBERT_1M+HN)$^{\Upsilon}$
Case 1 (related)		
mental illness	C0004936-Mental disorders	0033975-Psychotic Disorders
cramps	C0000737-Abdominal Pain	C0000729-Abdominal Cramps
get my shoulder right	C0000737-Abdominal Pain	0037011-Shoulder Pain
Case 2 (Overlapping)		
increased anxiety	C0003467-Anxiety	C0549259-Anxiety aggravated
suicide thoughts	C0086132-Depressive Symptoms	C1269683-Major Depressive Disorder
migraines	C0149931-Migraine Disorders	C0235890-Migraine aggravated
my knee has gotten all swollen	C0038999-Swelling	C0853619-Localized swelling
Case 3 (Subtle variations)		
wrecking my sleep	C0917801-Sleeplessness	C1262141-Poor quality sleep
knock you out	C0037317-Sleep disturbances	C0851578-Sleep Disorders
threatened to hurt me	C0002957-Anger	C0001807-Aggressive behavior
yawns	C0013144-Drowsiness	C0043387-Yawning

Table 8: Baseline vs Our model in TwADR-L. Here Γ - predicted and ground truth concepts are different, Υ - predicted and ground truth concepts are same.

Colloquial Phrase	Prediction (BioBERT_1M+HN)	Ground Truth
CADEC-MCN Custom		
damaging my muscles	68962001-Muscle pain	129565002-Disorder of muscle
itching of the skin	418290006-Itching	418363000-Itching of skin
constant sleepiness	193462001-Insomnia	77692006-Hypersomnia
CADEC-MCN Random		
runny nose	301202006-Nasal sinus problem	64531003-Nasal discharge
joints in my angles hurt	247373008-Ankle pain	202490009-Ankle joint pain
cold hands	309086004-Paresthesia of hand	271584002-Cold hands
TwADR-L		
need prozac	C0011570-Mental Depression	C0011581-Depressive disorder
cold sweat	C0038990-Sweating	C0232431-Cold sweat
accidentally double dosed	C1963951-Acute overdose	C0151821-Accidental overdose

Table 9: Failure analysis of our model (BioBERT_1M+HN)

6.2.3. Failure Analysis of our model (BioBERT_1M+HN)

Table 9 contains the instances for which our model assigned concepts wrongly. From Table 9, it is clear that predicted and ground truth concepts are close in meaning. For example, "*insomnia*" and "*hypersomnia*" are related as both represents disorders of sleep. One drawback of deep learning models is that they require sufficient training data i.e., good number of instances related to each class. When there is an imbalance, model prefer to assign the frequently occurring concept over less frequently occurring concept. Here, as ground truth concepts occur less frequently compared to predicted concepts, our model assigned concepts wrongly.

6.3. Impact of batch size

To study the impact of batch size on our best performing model, we evaluated it at different batch sizes of 16, 32, 64 and 128 using CADEC-MCN custom folds. In all these experiments, learning rate is fixed at 3e-5. The performance of our best model at different batch sizes is

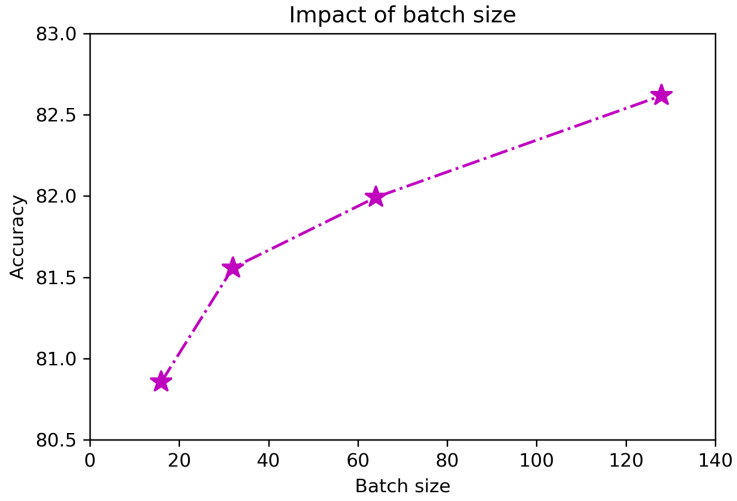

Figure 5: Our best model performance on CADEC-MCN custom folds at different batch sizes.

shown in Figure 5. Figure 5 shows that performance of model increases with increase in batch size and highest accuracy is achieved at batch size=128.

6.4. Impact of learning rate

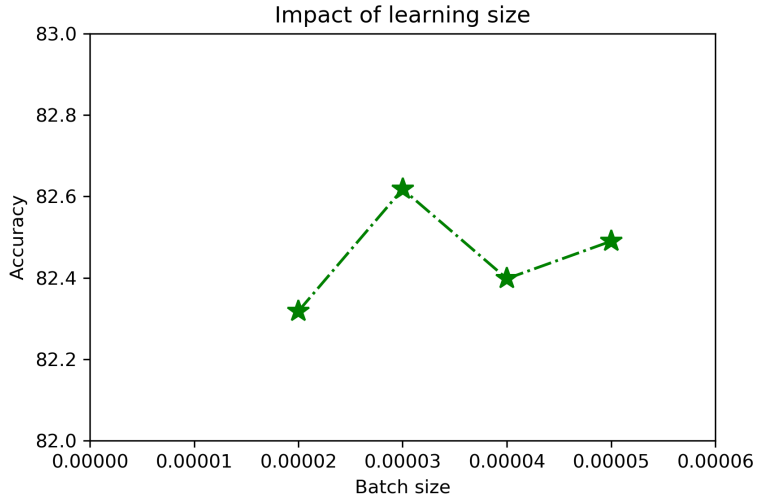

Figure 6: Our best model performance on CADEC-MCN custom folds at different learning rates.

To study the impact of learning on our best performing model, we evaluated it at different learning rates of 2e-5, 3e-5, 4e-5 and 5e-5 using CADEC-MCN custom folds. In all these experiments, batch size is fixed at 128. The performance of our best model at different learning rates is shown in Figure 6. Figure 6 shows that performance of model increases in the beginning and then decreases. Our best model achieved highest accuracy at learning rate=3e-5.

6.5. Impact of freezing encoder layers

Freezing a layer means, parameters of layer are not updated while fine-tuning the model. BERT consists of an embedding layer and stack of transformer encoder layers in which lower

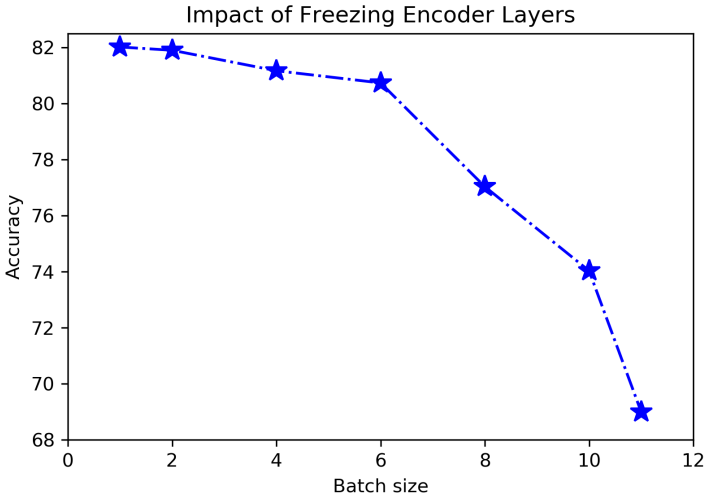

Figure 7: Our best model performance on CADEC-MCN custom folds at different learning rates.

layers capture syntactic information while upper layers capture semantic information. As syntactic information is common across domains and tasks, we believe that there is no need to further update the parameters of first few layers. Further freezing first few layers, allows the model to focus on learning more task specific information in upper layers which improves the performance of model. To study the impact of freezing encoder layers on performance of our best model, we conducted a series of experiments by freezing embedding layer along with first 1, 2, 4, 6, 8, 10 and 11 encoder layers while fine-tuning. In all these experiments, batch size and learning rate are fixed at 128 and 3e-5 respectively. From Figure 7, freezing encoder layers did not improve the performance of model. Freezing up to 6 encoder layers did not hurt the performance of model much and further, it increased speed of fine-tuning also. Freezing 8, 10 or 11 encoder layers reduced the performance considerably. The model achieved least accuracy when all the encoder layers were freezed.

BioBERT was initialized from general BERT and further pre-trained on biomedical text. Biomedical text authored by researchers is less noisy with standard terms while CADEC-MCN phrases authored by general public are more noisy with lots of colloquial and misspelled terms. Due to these variations, freezing first few layers while fine tuning didn't improve the performance of model, as expected.

6.6. Impact of Noise

To study robustness of our model against noise, we created four noisy datasets from custom folds of CADEC-MCN dataset. In each colloquial phrase, a word is chosen at random and one of the following types of noise is added (see Table 10).

- Addition of a special character at the start of word (Type 1)
- Deletion of a randomly chosen character (Type 2)
- Repetition (2 times) of a randomly chosen character (Type 3)
- Swapping a randomly chosen character with its adjacent character (Type 4)

	Original phrase	Noisy phrase
Type 1	**belly** weight gain	**#belly** weight gain
Type 2	**belly** weight gain	**bely** weight gain
Type 3	**belly** weight gain	**bellyy** weight gain
Type 4	**belly** weight gain	**blely** weight gain

Table 10: Example to illustrate different types of noise added.

	Noisy test set*	Noisy train and test sets$^{\Upsilon}$
Type 1	81.22	**82.31**
Type 2	60.39	**72.35**
Type 3	66.91	**75.69**
Type 4	60.65	**73.90**

Table 11: Performance (accuracy) of our best performing model on noisy datasets generated from CADEC-MCN custom folds. $\star$- accuracy obtained when our best model is trained on original train set and evaluated on noisy test set. Υ- accuracy obtained when our best model is trained and evaluated on noisy train and test sets respectively.

Our best performing model is evaluated in two ways namely a) Fine-tuned on original train set and evaluated on noisy test set and b) Fine-tuned on noisy train set and evaluated on noisy test set.
From Table 11, we observe

- Except in case of Type 1 noise, in all other cases the performance of model is significantly reduced in both evaluations. This is because, type 1 noise just adds a special character at the beginning of word, while other noises modify word by randomly deleting or swapping or repeating one of its characters.
- Our model performs much better when it is fine-tuned and evaluated on noisy instances compared to fine-tuned on original instances and evaluated on noisy instances. This is expected because the model performs better on noisy instances if it sees noisy instances during training itself.

6.7. Impact of BERT model

To show the impact of BERT model, we fine-tuned our best model using different sizes of training set from CADEC-MCN custom folds and then evaluated it. From Table 12, we observe that our best model outperforms ELMo based existing system [16] by 1.2% (75.90 vs 74.70) even when it is trained using 60% of training set. This is because ELMo based system has to be trained from scratch so it requires more training instances to perform better. Our best model is based on fine-tuned BioBERT and highway layer. As BERT model learns lots of language information during unsupervised pre-training itself, it can be fine-tuned even with small datasets and hence performs better compared to CNN or RNN downstream based models which are to be trained from scratch.

Our best model trained on	Accuracy
Full training set	**82.62**
95% of training set	81.93
90% of training set	81.16
85% of training set	80.82
80% of training set	79.25
70% of training set	78.38
70% of training set	77.82
65% of training set	76.59
60% of training set	75.90
Model	Accuracy
GRU+Att+tf-idf(max) * [16]	74.70

Table 12: Performance (accuracy) of our best model on training sets of different sizes from CADEC-MCN custom folds. $\star$ - model is trained on full training set.

7. Conclusion

In this study, we proposed a deep neural network based architecture to normalize medical concepts in social media text. Our deep neural network architecture consists of pre-trained BERT and task specific classifier which includes highway layer followed by softmax layer. We experimented with two general, four biomedical and four clinical BERT models to normalize concepts. As per our knowledge, it is the first work to do comprehensive evaluation of BERT based fine-tuned models in medical concept normalization. Our best model based on BioBERT pre-trained on 1M PubMed abstracts and highway layer outperformed other BERT models as well as existing systems and achieved best performance on TwADR-L, custom and random folds of CADEC-MCN. We also conducted series of experiments to study the impact of different batch sizes, learning rates and freezing encoder layers on the performance of our best model. Further we evaluated our best model on noisy datasets created from CADEC-MCN custom folds, to study its robustness against noise. In future, we would like to explore possible ways to a) make our model robust against noises and b) incorporate knowledge from sources like UMLS which can potentially improve the performance of model.

References

[1] K. S. Kalyan, S. Sangeetha, Secnlp: A survey of embeddings in clinical natural language processing, Journal of Biomedical Informatics 101 (2020) 103323.

[2] A. R. Aronson, Effective mapping of biomedical text to the umls metathesaurus: the metamap program., in: Proceedings of the AMIA Symposium, American Medical Informatics Association, 2001, p. 17.

[3] Y. Tsuruoka, J. McNaught, J. c. Tsujii, S. Ananiadou, Learning string similarity measures for gene/protein name dictionary look-up using logistic regression, Bioinformatics 23 (2007) 2768–2774.

[4] A. McCallum, K. Bellare, F. Pereira, A conditional random field for discriminatively-trained finite-state string edit distance, arXiv preprint arXiv:1207.1406 (2012).

[5] R. Leaman, R. Islamaj Doğan, Z. Lu, Dnorm: disease name normalization with pairwise learning to rank, Bioinformatics 29 (2013) 2909–2917.

[6] R. Leaman, Z. Lu, Automated disease normalization with low rank approximations, in: Proceedings of BioNLP 2014, 2014, pp. 24–28.

[7] N. Limsopatham, N. Collier, Adapting phrase-based machine translation to normalise medical terms in social media messages, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 1675–1680.

[8] N. Limsopatham, N. Collier, Normalising medical concepts in social media texts by learning semantic representation, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, pp. 1014–1023.

[9] K. Lee, S. A. Hasan, O. Farri, A. Choudhary, A. Agrawal, Medical concept normalization for online user-generated texts, in: 2017 IEEE International Conference on Healthcare Informatics (ICHI), IEEE, 2017, pp. 462–469.

[10] E. Tutubalina, Z. Miftahutdinov, S. Nikolenko, V. Malykh, Medical concept normalization in social media posts with recurrent neural networks, Journal of biomedical informatics 84 (2018) 93–102.

[11] M. Belousov, W. Dixon, G. Nenadic, Using an ensemble of generalised linear and deep learning models in the smm4h 2017 medical concept normalisation task, 2017.

[12] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018, pp. 2227–2237.

[13] J. Howard, S. Ruder, Universal language model fine-tuning for text classification, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 328–339.

[14] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training (2018).

[15] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186.

[16] Z. Miftahutdinov, E. Tutubalina, Deep neural models for medical concept normalization in user-generated texts, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, Association for Computational Linguistics, Florence, Italy, 2019, pp. 393–399. URL: https://www.aclweb.org/anthology/P19-2055.

[17] Y. Si, J. Wang, H. Xu, K. Roberts, Enhancing clinical concept extraction with contextual embedding, arXiv preprint arXiv:1902.08691 (2019).

[18] K. Huang, J. Altosaar, R. Ranganath, Clinicalbert: Modeling clinical notes and predicting hospital readmission, arXiv:1904.05342 (2019).

[19] Z. Ji, Q. Wei, H. Xu, Bert-based ranking for biomedical entity normalization, arXiv preprint arXiv:1908.03548 (2019).

[20] D. E. Rumelhart, G. E. Hinton, R. J. Williams, Learning representations by back-propagating errors, nature 323 (1986) 533.

[21] J. L. Elman, Distributed representations, simple recurrent networks, and grammatical structure, Machine learning 7 (1991) 195–225.

[22] Y. Bengio, R. Ducharme, P. Vincent, C. Jauvin, A neural probabilistic language model, Journal of machine learning research 3 (2003) 1137–1155.

[23] R. Collobert, J. Weston, A unified architecture for natural language processing: Deep neural networks with multitask learning, in: Proceedings of the 25th international conference on Machine learning, ACM, 2008, pp. 160–167.

[24] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, S. Khudanpur, Recurrent neural network based language model, in: Eleventh Annual Conference of the International Speech Communication Association, volume 2, 2010, p. 3.

[25] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in neural information processing systems, 2013, pp. 3111–3119.

[26] J. Pennington, R. Socher, C. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[27] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Transactions of the Association for Computational Linguistics 5 (2017) 135–146.

[28] K. Lee, A. Agrawal, A. Choudhary, Real-time disease surveillance using twitter data: demonstration on flu and cancer, in: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2013, pp. 1474–1477.

[29] K. Lee, A. Agrawal, A. Choudhary, Mining social media streams to improve public health allergy surveillance, in: 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2015, pp. 815–822.

[30] L. Chen, K. T. Hossain, P. Butler, N. Ramakrishnan, B. A. Prakash, Syndromic surveillance of flu on twitter using weakly supervised temporal topic models, Data mining and knowledge discovery 30 (2016) 681–710.

[31] S. Shan, Y. Jia, J. Zhao, Same influenza, different responses: Social media can sense a regional spectrum of symptoms, arXiv preprint arXiv:1905.01778 (2019).

[32] S. Karimi, A. Metke-Jimenez, M. Kemp, C. Wang, Cadec: A corpus of adverse drug event annotations, Journal of biomedical informatics 55 (2015) 73–81.

[33] A. Nikfarjam, A. Sarker, K. O'Connor, R. Ginn, G. Gonzalez, Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features, Journal of the American Medical Informatics Association 22 (2015) 671–681.

[34] I. Korkontzelos, A. Nikfarjam, M. Shardlow, A. Sarker, S. Ananiadou, G. H. Gonzalez, Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts, Journal of biomedical informatics 62 (2016) 148–158.

[35] C. VanDam, S. Kanthawala, W. Pratt, J. Chai, J. Huh, Detecting clinically related content in online patient posts, Journal of biomedical informatics 75 (2017) 96–106.

[36] A. Sarker, G. Gonzalez-Hernandez, Overview of the second social media mining for health (smm4h) shared tasks at amia 2017, Training 1 (2017) 1239.

[37] D. Weissenbacher, A. Sarker, M. J. Paul, G. Gonzalez-Hernandez, Overview of the third social media mining for health (SMM4H) shared tasks at EMNLP 2018, in: Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 13–16. doi:10.18653/v1/W18-5904.

[38] D. Weissenbacher, A. Sarker, A. Magge, A. Daughton, K. O'Connor, M. J. Paul, G. Gonzalez-Hernandez, Overview of the fourth social media mining for health (SMM4H) shared tasks at ACL 2019, in: Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task, Association for Computational Linguistics, Florence, Italy, 2019, pp. 21–30. doi:10.18653/v1/W19-3203.

[39] K. O'Connor, P. Pimpalkhute, A. Nikfarjam, R. Ginn, K. L. Smith, G. Gonzalez, Pharmacovigilance on twitter? mining tweets for adverse drug reactions, in: AMIA annual symposium proceedings, volume 2014, American Medical Informatics Association, 2014, p. 924.

[40] J. Niu, Y. Yang, S. Zhang, Z. Sun, W. Zhang, Multi-task character-level attentional networks for medical concept normalization, Neural Processing Letters (2018) 1–18.

[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in neural information processing systems, 2017, pp. 5998–6008.

[42] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[43] J. L. Ba, J. R. Kiros, G. E. Hinton, Layer normalization, arXiv preprint arXiv:1607.06450 (2016).

[44] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al., Google's neural machine translation system: Bridging the gap between human and machine translation, arXiv preprint arXiv:1609.08144 (2016).

[45] D. Hendrycks, K. Gimpel, Bridging nonlinearities and stochastic regularizers with gaussian error linear units, arXiv preprint arXiv:1606.08415 (2016).

[46] R. K. Srivastava, K. Greff, J. Schmidhuber, Highway networks, arXiv preprint arXiv:1505.00387 (2015).

[47] Y. Kim, Y. Jernite, D. Sontag, A. M. Rush, Character-aware neural language models, in: Thirtieth AAAI Conference on Artificial Intelligence, 2016.

[48] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: pre-trained biomedical language representation model for biomedical text mining, arXiv preprint arXiv:1901.08746 (2019).

[49] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, M. McDermott, Publicly available clinical BERT embeddings, in: Proceedings of the 2nd Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 72–78. doi:10.18653/v1/W19-1909.

[50] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, R. G. Mark, Mimic-iii, a freely accessible critical care database, Scientific data 3 (2016) 160035.

[51] S. Robertson, H. Zaragoza, et al., The probabilistic relevance framework: Bm25 and beyond, Foundations and Trends® in Information Retrieval 3 (2009) 333–389.

Appendix A. Datasets

Appendix A.1. CADEC-MCN Custom

Table A.13 contains detailed statistics of CADEC-MCN Custom dataset.

Fold	#train samples	#test samples	#unique medical concepts
0	1655	504	181
1	1700	459	181
2	1735	424	181
3	1762	397	181
4	1784	375	181

Table A.13: Statistics of CADE-MCN Custom dataset

Appendix A.2. CADEC-MCN Random

Table A.14 contains detailed statistics of CADEC-MCN Random dataset.

Fold	#train samples	#valid samples	#test samples	# unique medical concepts
0	15612	845	867	1036
1	15631	826	867	1036
2	15700	758	866	1036
3	15672	786	866	1036
4	15630	828	866	1036
5	15675	783	866	1036
6	15710	748	866	1036
7	15659	799	866	1036
8	15647	811	866	1036
9	15716	742	866	1036

Table A.14: Statistics of CADEC-MCN Random dataset

Appendix A.3. TwADR-L

Table A.15 contains detailed statistics of TwADR-L dataset.

Fold	#train samples	#valid samples	#test samples	# unique medical concepts
0	4816	115	143	2200
1	4817	114	143	2200
2	4791	140	143	2200
3	4812	119	143	2200
4	4811	120	143	2200
5	4801	130	143	2200
6	4819	112	143	2200
7	4790	142	142	2200
8	4788	144	142	2200
9	4812	120	142	2200

Table A.15: Statistics of TwADR-L dataset