Medical Concept Normalization for Online User-Generated Texts

Kathy Lee[†,§], Sadid A Hasan[§], Oladimeji Farri[§], Alok Choudhary[†], Ankit Agrawal[†]
[§]AI Lab, Philips Research North America, Cambridge, MA
{kathy.lee_1, sadid.hasan, dimeji.farri}@philips.com
[†]EECS Department, Northwestern University, Evanston, IL
{kathy.lee, choudhar, ankitag}@eecs.northwestern.edu

***Abstract*—Social media has become an important tool for sharing content in the last decade. People often talk about their experiences and opinions on different health-related issues e.g. they write reviews on medications, describe symptoms and ask informal questions about various health concerns. Due to the colloquial nature of the languages used in the social media, it is often difficult for an automated system to accurately interpret them for appropriate clinical understanding. To address this challenge, this paper proposes a novel approach for medical concept normalization of user-generated texts to map a health condition described in the colloquial language to a medical concept defined in standard clinical terminologies. We use multiple deep learning architectures such as convolutional neural networks (CNN) and recurrent neural networks (RNN) with input word embeddings trained on various clinical domain-specific knowledge sources. Extensive experiments on two benchmark datasets demonstrate that the proposed models can achieve up to 21.28% accuracy improvements over the existing models when we use the combination of all knowledge sources to learn neural embeddings.**

I. Introduction

On social media and online health communities, people often share their experiences and opinions on various health topics including personal health issues and symptoms. Especially, on medical forums, consumers ask health related questions, write reviews on medications and describe negative side effects they experience while taking a drug. Moreover, patients and their families can get emotional support by sharing their stories of overcoming illnesses.

Medical concept normalization for user-generated texts aims at mapping a health condition described in colloquial language to a medical concept in standard ontologies such as Unified Medical Language System (UMLS) [18] via concept unique identifiers (CUIs). This task has many applications for improving patient care such as: 1) understanding questions and providing answers to patients/families seeking medical knowledge, 2) early detection of patients who need immediate attention and medical support (e.g., people with suicidal ideation), 3) digital disease surveillance (e.g., monitoring of pandemics), and 4) clinical paraphrasing to improve patient engagement by helping patients understand their clinical reports.

While consumers describe their health conditions in colloquial language, clinical knowledge sources such as biomedical literature present medical terms in scientific language. This gap in the use of languages between patients/consumers and clinicians requires mapping of one to the other. In order to generate solutions to a given medical problem (e.g. to answer questions posted on an online health community), health conditions in user-generated texts need to be normalized to medical concepts in standard ontologies. Once the solution is generated, it needs to be translated back to colloquial language for users to easily understand.

Table I shows examples of user-generated texts from social media that describe medical concepts. The labels in the top row are medical concepts from the standard medical ontologies and the phrases in the same column denote example phrases from social media that describe the concept. The examples very well illustrate the characteristics of colloquial language or non-standard terms used to describe medical conditions on social media. As can be seen in the table, the challenges for medical concept normalization include: 1) alternative descriptions for health conditions in colloquial language (e.g., 'sore and stiff ankles', 'terrible pain in my ankles', 'ankles ache so bad' $\rightarrow$ *ankle pain*; 'trouble sleeping', 'cannot sleep', 'hard time sleeping' $\rightarrow$ *difficulty sleeping*, and 2) no overlaps of terms between colloquial language and scientific/medical terms describing the same health condition (e.g., 'couldn't remember' $\rightarrow$ *memory impairment*, 'sight loss' $\rightarrow$ *visual impairment*, 'trouble remembering', 'foggy brain' $\rightarrow$ *memory impairment*). In the latter case, basic string matching approaches without understanding semantics of the text will result in a poor performance in a medical concept normalization task. Other challenges include misspellings or typos as shown for the concept 'diarrhoea'.

In this work, we aim to address the aforementioned challenges using deep learning-based architectures and study the impact of different types of input data used to build neural embeddings on the medical concept normalization performance.

Our key contributions are:

- We investigate the use of various domain-specific text data to build neural embeddings to learn semantic features of medical concepts for normalization.
- We demonstrate that two deep learning models (CNN and RNN) can better predict the medical concepts when we use neural embeddings trained on domain-specific clinical texts compared to those trained on a larger general domain text corpus.

978-1-5090-4881-6/17 $31.00 © 2017 IEEE
DOI 10.1109/ICHI.2017.59

TABLE I
MEDICAL CONCEPTS AND EXAMPLE SOCIAL MEDIA PHRASES

loss of hair	memory impairment	ankle pain	diarrhoea	difficulty sleeping
hair falling out	memory problem	ankle hurt	direar	can not sleep
hair loss	memory failure	ankles started aching	diaharrea	difficult to sleep
hair losss	memory deficits	pain in ankles	diahhrea	hard time sleeping
losing my hair	poor memory	sore and stiff ankles	diahrea	inability to sleep well
thinning hair	trouble remembering	ankles seized up	diarrehea	lousy sleeping at night
hair has started falling out	memory weakened	sore ankles	dioreah	poor sleep
hair is getting very thin	couldn't remember	terrible pain in my ankles	dioreaha	problems sleeping
hair was falling out	foggy brain	ankles ache so bad	bathroom with the runs	trouble sleeping

- Our best results present the new state-of-the-art for two benchmark datasets, outperforming the accuracy of a strong normalization model by up to +21.17% on the Twitter data set and up to +21.28% on the AskAPatient data set.

This paper is organized as follows. In section II, we present related work on deep neural network models, social media for healthcare, and medical concept normalization. In section III, we describe CNN and RNN models we use for concept normalization. In section IV, we describe how we re-created the social media datasets and present details of text data from various clinical knowledge sources used to build neural embeddings. In section V, we present our experimental results, followed by conclusion in section VI.

II. RELATED WORK

A. Social Media for Healthcare

Social media has been widely used as a new medium for real-time information transmission in various domains including health to track volume of mentions of disease, drugs, and symptoms [14], [15], predict influenza activities, and detect adverse drug events (ADE) earlier than the traditional influenza or ADE surveillance systems that have significant time delays in data processing [6], [16]. For automatic extraction of medical concepts from social media, researchers have used machine learning approaches such as CRF (Conditional Random Fields) and HMM (Hidden Markov Model) to extract phrases that describe medical concepts (e.g., disease, drugs, symptoms) [22], [26], identify relationships between two medical concepts (e.g., duration, frequency, dosage, route for a drug, indication, side effects, etc.), and to classify texts into different categories (e.g., health vs. non-health, ADE vs. non-ADE) [29], [27], [16].

B. Deep Neural Network Models

Recurrent neural network (RNN) models have shown to be very effective in many natural language processing (NLP) tasks. Unlike traditional neural network models, RNNs use sequential information. Hence they are well-suited for tasks such as machine translation, speech recognition, language modeling and image caption generation. Traditionally, convolutional neural network (CNN) models have been widely used in image processing tasks (e.g., automatic recognition of handwritten numbers, object detection) because of their ability to learn task-relevant features. However, with the recently proposed word embedding models (word2vec) by Mikolov et al. [20], [21], deep neural network models for NLP tasks have gained popularity. Kim [11] showed that a simple one layer CNN model trained on top of pre-trained word vectors outperform several state-of-the-art models for text classification such as sentiment analysis and question classification. Lee et al. [16] explored semi-supervised CNN models to detect adverse drug events in tweets and demonstrated that neural word embeddings trained on a smaller domain-specific dataset helps more than the one trained on a larger random dataset for ADE classification. Deep learning models have also shown to be highly effective in other healthcare tasks such as clinical diagnostic inferencing [24] and clinical neural phraphrase generation [9], [23].

C. Concept Normalization

Traditional approaches used for medical concept normalization include lexicon-based string matching, heuristic string matching, and rule-based text mapping to a set of pre-defined variants of terms [25], [1], [19]. DNorm [13] is a state-of-the-art concept (disease name) normalization system that is based on pairwise learning to rank that learns similarities between mentions and concept names. Limsopatham et al. used a machine translation approach in which a social media phrase is translated into a formal medical concept. More recently, Limsopatham et al. [17] showed that simple deep learning models, convolutional neural network (CNN) and recurrent neural network (RNN), with pre-trained word embeddings induced from a large collection of Google News (GNews) and BioMed Central (BMC) articles improved the performance over previous state-of-the-art concept normalization models and reported that GNews was more effective than BMC for both CNN and RNN across all datasets.

Our work significantly improves on the results from Limsopatham et al. [17] by refining their original datasets and leveraging neural embeddings of various health-related text to better learn the semantic characteristics of medical concepts and provide a new state-of-the-art accuracy for medical concept normalization.

III. MODEL DESCRIPTION

In this section, we describe two deep learning models, convolutional neural network (CNN) and recurrent neural network (RNN), we use for medical concept normalization.

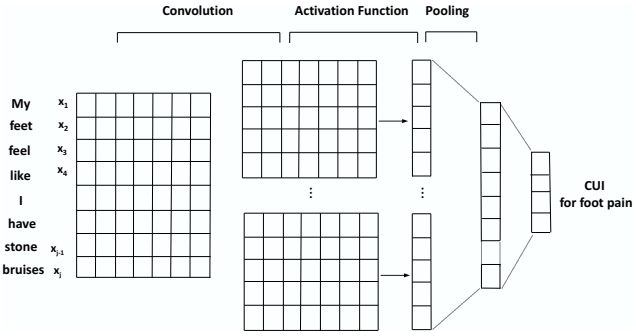

Fig. 1. Generic convolutional neural network architecture.

A. Convolutional Neural Network (CNN)

CNN is a feed-forward neural network model that learns task-relevant semantic features for text classification. Figure 1 depicts a simple CNN with an input layer, followed by a convolutional layer with multiple filters, a pooling layer, and a final softmax classifier. The input layer of CNN are phrases or sentences represented as a matrix. Each row of the matrix is a low-dimensional vector (word embeddings) representing a token or a word. Formally, given an input phrase x of length j, where $x = x_i, x_{i+1}, \ldots, x_{i+j}$ denotes a sequence of words, and x_i denotes a k-dimensional word vector, a filter $w \in R^{hk}$ is applied to a window of h words to produce a new feature in a convolution layer. For example, a feature c_i is generated as follows:

$$c_i = f(w \cdot x_{i:i+h-1} + b) \tag{1}$$

from a window of words $x_{i:i+h-1}$ where b is a bias and f is a nonlinear activation function. Each feature is applied to the input matrix to produce a feature map. Then the features are passed to a fully connected softmax layer to output the most probable label [11]. For example, for the eight word phrase 'my feet feel like I have stone bruises' using 300-dimensional embedding, the input to the CNN would be a 8 x 300 matrix and the output would be a CUI representing the medical concept 'foot pain'.

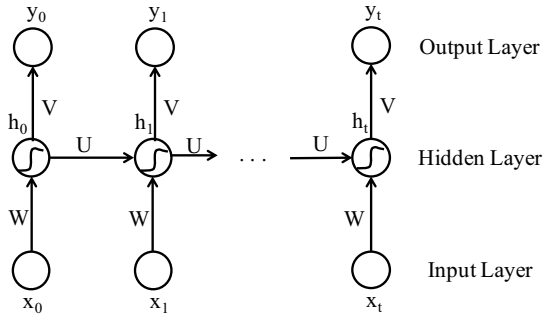

Fig. 2. Generic recurrent neural network architecture.

B. Recurrent Neural Network (RNN)

RNN is similar to the standard feedforward neural network, but the hidden unit activation at time t is dependent on that of time $t-1$, which allows the model to deal with variable-length input and output [28] and make it suitable for modeling sequences. Figure 2 shows an unrolled RNN architecture, where x_t, y_t, h_t are the input, output, and hidden state at time step t, and W, U, V are the parameters of the model corresponding to *input*, *hidden*, and *output* layer weights (shared across all time steps). The hidden state h_t can be formulated as follows:

$$h_t = f(Wx_t + Uh_{t-1}), \tag{2}$$

where the h_{t-1} is previous hidden state, x_t is the the current input, and f is an element-wise nonlinear activation function.

Although RNN is a powerful model to encode sequences, it suffers from the vanishing gradient problem while it tries to efficiently learn long-range dependencies [2]. We use gated recurrent unit (GRU) [7], which is known to be a successful remedy to the vanishing gradient problem.

The hidden state of GRU h_t can be formulated as follows:

$$\begin{aligned} z_t &= \sigma(W^z x_t + U^z h_{t-1}) \\ r_t &= \sigma(W^r x_t + U^r h_{t-1}) \\ k_t &= \tanh(W^k x_t + U^k (r_t \odot h_{t-1})) \\ h_t &= (1 - z_t) \odot k_t + z_t \odot h_{t-1}, \end{aligned} \tag{3}$$

where z_t, r_t are the update gate and the reset gate, and k_t is the candidate hidden state. z_t, r_t are computed using different weight parameters where z_t determines how much of the old memory to keep while r_t denotes how much new information is needed to be combined with the old memory. Finally, k_t is computed by exploiting r_t, and h_t is calculated to denote the amount of information needed to be transmitted to the following layers.

IV. EXPERIMENTAL SETUP

A. Data

We use two data sets, TwADR-L (from Twitter) and AskA-Patient, used by Limsopatham et al. [17] for medical concept normalization[1]. TwADR-L was created by the authors, and AskAPatient dataset was created by Karimi et al. [10] for ADR (adverse drug reaction), from which the authors extracted the gold-standard mappings of phrases to medical concepts.

In the original dataset, the TwADR-L had 48,057 training, 1,256 validation and 1,427 test examples. The test set (all test samples from 10 folds combined) consists of 765 unique phrases and 273 unique classes (medical concepts). The AskA-Patient dataset contained 156,652 training, 7,926 validation, and 8,662 test examples. The entire test set (all test samples from 10 folds combined) consists of 3,749 unique phrases and 1,035 unique classes (medical concepts). The authors randomly split each dataset into ten equal folds, ran 10-fold

[1]Available at https://zenodo.org/record/55013#.WKXwdxIrLde

TABLE II
DATA STATISTICS AFTER REMOVING DUPLICATES FROM COMBINED TRAINING, VALIDATION, AND TEST DATA

	TwADR-L	AskAPatient
# unique phrases	2,944	4,469
# unique labels	2,220	1,036
# unique phrase-label pairs	3,157	4,496
# phrases with multiple labels	173	26
Min # examples per label	1	1
Max # examples per label	36	141
Avg # examples per label	1.42	4.35

TABLE III
EXAMPLES OF PHRASES WITH MULTIPLE LABELS

Social Media Phrase	**Multi-Labels (Medical Concepts)**
shaking	shivering, trembling, tremor
mad	anger, rage
have no emotion	emotional disorder, indifferent mood
mood swings	bipolar disorder , disturbance in mood
sore	pain, myalgia
high blood pressure	increased venous pressure, hypertension, findings of increased blood pressure

cross validation and reported the accuracy averaged across the ten folds.

We found that, in the original data set, many phrase-label pairs appeared multiple times within the same training data file and also across the training and test data sets in the same fold. In the AskAPatient data set, on average 35.82% of the test data overlapped with training data in the same fold. In the Twitter (TwADR-L) dataset, on average 8.62% of the test set had an overlap with the training data in the same fold. Having a large overlap between the training and the test data can potentially introduce bias in the model and contribute to high accuracy. It is not unlikely that the high model performance reported in the original paper may be triggered by the the large overlap between the training and test sets.

Therefore to remove the bias, we further cleaned and recreated the training, validation, and test sets such that each phrase-label pair appears only once in the entire dataset (either in training, validation or test set).

First, we combined all examples in training, validation and test data from the original data set and then removed all duplicate phrase-label pairs (examples that have the same phrase and label pair and appear more than once in training/validation/test datasets). Table II shows the statistics of the new dataset (after removing duplicates). The Twitter data set had 3,157 unique phrase-label pairs and 2,220 unique labels (medical concepts) while 173 phrases had multiple labels (i.e., they were assigned to more than one label). Many concepts had only one example, and the concept that had the most number of examples had 36 phrases. On average, each concept had 1.42 examples. The AskAPatient data set had 4,496 unique phrase-label pairs, 1,036 unique labels while 26 phrases had multiple labels. Table III shows examples of phrases that had multiple labels. For example, 'mad' can be mapped to 'anger' or 'rage' and 'sore' can be mapped to 'pain' or 'myalgia'.

Second, we remove all concepts that had less than five examples. The statistics of the final data are shown in Table IV.

TABLE IV
DATA STATISTICS AFTER REMOVING CONCEPTS THAT HAVE LESS THAN FIVE EXAMPLES

	TwADR-L	AskAPatient
# unique phrases	543	2,494
# unique labels	65	228
# unique phrase-label pairs	617	1,427
# phrases with multiple labels	173	26
Min # examples per label	5	5
Max # examples per label	36	78
Avg # examples per label	9.5	11

Third, we divide all examples without multiple labels into random 10 folds such that each unique phrase-label pair appears once in one of the 10 test sets. We add the pairs with multiple labels into the training data. This final 10-folds dataset is used in all our experiments.

B. Data Sources for Word Embedding

In this section, we describe different types of unlabeled text data we use for building neural embeddings.

Fig. 3. Definition, example sentence, synonyms, related words, near antonyms and antonyms for the word 'sore' obtained from Merriam-Webster Thesaurus.

1) Thesaurus (TH): For each word in TwADR-L and AskAPatient dataset (both phrases and labels), we obtained the following six types of information from the Merriam-Webster thesaurus[2]: definition, example sentence, synonyms, related words, near antonyms, and antonyms. Figure 3 illustrates the information that was obtained for the word 'sore', the last example shown in Table III. The definition of 'sore' include the label 'pain' and also the list of synonyms include 'painful' (an adjective form of the label 'pain'). Therefore, the word embeddings built with the thesaurus will help the model learn the semantics and predict the label 'pain'.

Fig. 4. Medical definition of the term 'myalgia' obtained from Merriam-Webster Medical Dictionary.

2) Medical Dictionary (MD): We collected definitions from the Merriam-Webster Medical Dictionary[3], which contains 60,000 words and phrases used by healthcare professionals. It is also used in the National Library of Medicine's consumer

[2]https://www.merriam-webster.com/thesaurus
[3]https://www.merriam-webster.com/medical

TABLE V
MEDICAL CONCEPTS AND SIMILAR WORDS BASED ON COSINE SIMILARITY OBTAINED FROM WORD EMBEDDINGS BUILT WITH DIFFERENT TEXT CORPORA.

Medical Concept	Clinical Text (CT)	Medical Dictionary (MD)	Thesaurus (TH)	Health-related Tweets (HT)
depression	dysthymia	arthritic	recession	boredom
	anxiety	mood	disorder	weightgain
	schizophrenia-like	diminution	collapse	obesityWHO
	benzodiazepine-induced	exertion	lassitude	irritability
	hopelessness	fatigue	lethargy	anxiety
insomnia	apnea		sleeplessness	depressionchronic
	derealization		wakefulness	migraines
	sleep	–	restlessness	weightgain
	dysthymic			hyperexcitability
	awakening			stressrelated
dizzy	lightheaded	verge	woozy	lightheaded
	faint	restless	fainting	nauseous
	nauseated	light-headed	whirling	headache
	swaying	lamely	faint	lethargic
	shaky	paranoia	feeble	sleepfeeling
myalgia	backache			arthralgia
	arthralgia			athralgia
	asthenia	–	–	muscleampjoint
	aches			odynophagia
	fatigability			bodymuscle
hypertension	dyslipidemia	arterial		diseaseheart
	renovascular	hypotension		diabetes
	nephrosclerosis	narrowing	–	dyslipidemia
	beta-antagonists	weakness		pressurehigh
	Gestosis	diallation		arteriosclerosis

health website to help consumers with spelling of medical words and understanding of medical notes written by physicians[4]. For each unique word in TwADR-L and AskAPatient dataset, we obtained a medical definition (if present) using the Merriam-Webster medical dictionary API[5]. The dictionary contains clinical terms that may not be found in the thesaurus. We found that while definitions for some terms are same in both the thesaurus and the medical dictionary, for other terms, either they use slightly different words/phrases, or one or both do(es) not have a definition at all. For example, the word 'myalgia' was in the medical dictionary, but not in the thesaurus. As shown in Figure 4, we were able to collect the definition for the word 'myalgia', a medical term that was not found in the thesaurus.

3) Clinical Texts (CT): is a collection of sentences from the following sources in the medical domain.

Adverse Drug Reaction Classification System (ADReCS)[6] : is a comprehensive ADR ontology database that provides both standardization and hierarchical classification of ADR terms [5]. The database integrates ADR and drug information collected from various public medical repositories like DailyMed[7], MedDRA [4], SIDER2 [12], DrugBank[8], PubChem[9], UMLS. It contains 6.7K unique ADR terms and 1,698 drug names, and 154K drug-ADR pairs. For each term in the ADReCS database, we collected its definition and synonyms. For example, the definition of the word 'myalgia' is 'painful sensation in the muscles' and its synonyms are myalga, myaigia, soreness, muscle pain, muscle ache, etc.

Biomedical Literature: We collect 301,790 sentences from all wikipedia pages that are under the category of clinical medicine[10]. We also collect 4,271 sentences from PubMed articles from the adverse drug events benchmark corpus [8].

Medical Concept to Lay Term Dictionaries: We use two medical to lay terms dictionaries to create a collection of sentences[11,12]. These dictionaries contain professional medical terms and their definitions described in lay language. For example, the medical term 'anesthesia' is defined in lay language as 'loss of sensation or feeling', the term 'cephalalgia' as 'headache', and the term 'dyspnea' as 'hard to breathe' or 'short of breath'. From these dictionaries, we generate sentences (e.g., 'Anesthesia refers to loss of sensation or feeling', 'cephalalgia means headache') by combining a term and its definition with a connecting phrase randomly chosen from a small preselected set (e.g., stands for, refers to, indicates, means, etc.). We create a total of 1,556 sentences from these sources.

UMLS Medical Concept Definitions: We extract a total of 167,550 sentences that define medical terms in the UMLS Metathesaurus [3], a large biomedical thesaurus consisting of millions of medical concepts and used by professionals for patient care and public health.

We combine all of the above mentioned sentences.

[4]https://www.nlm.nih.gov/news/mplusdictionary03.html
[5]https://www.dictionaryapi.com/products/api-medical-dictionary.htm
[6]http://bioinf.xmu.edu.cn/ADReCS/
[7]https://dailymed.nlm.nih.gov/dailymed/
[8]https://www.drugbank.ca/
[9]https://pubchem.ncbi.nlm.nih.gov/
[10]https://en.wikipedia.org/wiki/Category:Clinical_medicine
[11]http://gsr.lau.edu.lb/irb/forms/medical_lay_terms.pdf
[12]https://depts.washington.edu/respcare/public/info/Plain_Language_Thesaurus_for_Health_Communications.pdf

TABLE VI
CLASSIFICATION ACCURACY (%) USING 10-FOLD CROSS VALIDATION (TH = THESAURUS, MD = MEDICAL DICTIONARY, CT = CLINICAL TEXTS, HT = HEALTH-RELATED TWEETS, BATCH SIZE = 50, NUMBER OF EPOCH = 100, VECTOR DIMENSION = 300)

Word Embeddings	TwADR-L CNN	TwADR-L RNN	AskAPatient CNN	AskAPatient RNN
Rand	16.06	22.05	40.95	58.54
GNews	15.57	23.17	45.73	64.41
TH	14.43	20.43	32.66	57.17
MD	15.73	19.62	41.90	58.26
CT	14.77	22.21	45.49	61.81
HT	16.69	24.63	45.46	64.08
TH + MD + CT + HT	**19.46**	**25.30**	**55.46**	**65.04**

TABLE VII
ABLATION STUDY. COMPARISON OF MODELS' ACCURACY (%) WHEN A FEATURE IS REMOVED FROM ALL POSSIBLE FEATURE SETS (TH = THESAURUS, MD = MEDICAL DICTIONARY, CT = CLINICAL TEXTS, HT = HEALTH-RELATED TWEETS). THE NUMBERS IN PARENTHESIS INDICATE THE PERFORMANCE DROP WHEN THE FEATURE IS REMOVED.

Word Embeddings	TwADR-L CNN	TwADR-L RNN	AskAPatient CNN	AskAPatient RNN
All - HT	18.80 (-0.66)	22.54 **(-2.76)**	46.37 (-9.09)	62.97 (-2.07)
All - TH	16.38 (-3.08)	25.44 (+0.14)	45.29 (-10.17)	62.96 **(-2.08)**
All - CT	15.58 **(-3.88)**	24.96 (-0.34)	45.61 (-9.85)	64.09 (-0.95)
All - MD	17.69 (-1.77)	26.60 (+1.3)	44.50 **(-10.96)**	63.93 (-1.11)
All	19.46	25.30	55.46	65.04

4) Health-related Tweets (HT): We collected 100 million publicly available health-related tweets that mention 116 common diseases and symptoms (e.g., flu, depression, insomnia, diabetes, obesity, heart disease, anxiety disorder, etc.) using the Twitter streaming API[13], which provides approximately 1% of all publicly available tweets. As preprocessing steps, we remove non-English tweets, tokenize the text, normalize to lowercase, and replace hyperlinks, numerics and Twitter screen names with special tokens: 'URL', 'NUMBER' and 'USER'.

Table V shows medical concepts and examples of top 20 similar words by cosine similarity based on the word embeddings built with individual data source.

V. RESULTS

Table VI shows the accuracy of classification models using 10-fold cross validation, averaged over ten folds. The first two rows are our baseline models[14] [17] where CNN and RNN models use a randomly generated embeddings (Rand) and a publicly available pre-trained word embeddings generated from 100 billion words from Google News (GNews) using word2vec [21] as inputs. The next four rows (rows 3-6) present the performance of the same CNN and RNN as the baseline models but using word embeddings we built on top of various clinical texts described in section IV-B. The last row presents the performance when the models use word embeddings built using combination of all four data sources as an input. All experiments including the baseline models are trained and evaluated on the cleaned and newly-created datasets (described in section IV-A).

[13]https://dev.twitter.com/streaming/public

[14]Code available at https://github.com/nutli/concept_normalisation

Among the individual datasets (TH, MD, CT, HT), the health-related tweets (HT) had the most significant impact on the classification performance. Both the CNN and the RNN models performed comparable to (for AskAPatient dataset) or better (for TwADR-L dataset) than the best baseline models. When we combined all individual datasets, it largely improved the classification accuracy over all baseline models and all our models. Compared to the best baseline accuracy, the improvement was +21.17% on TwADR-L CNN, +9.19% on TwADR-L RNN, +21.28% on AskAPatient CNN, and +0.98% on AskAPatient RNN. The improvement was substantial for CNN. For all models, we used the following hyperparameters: batch size = 50, number of epochs = 100, vector dimension = 300, number of neurons in hidden layer = 100, dropout rate = 0.5, non-linear activation function = rectifier, and max-pooling for CNN.

A. Ablation Study

Next we conduct experiments to study the effects of removing a dataset from training. Table VII presents the performance loss when each dataset is removed from the set of all possible resources (TH + MD + CT + HT). Interestingly, each of the four data sources appeared to be the most important for different deep learning models and datasets. The performance dropped by 3.88% (from 19.46% to 15.58%) when clinical texts (CT) was removed, indicating that CT is the most important feature for TwADR-L CNN among the four individual features. For TwADR-L RNN, helath-related tweets (HT) was the most helpful feature, indicated by the performance drop of 2.76% when removed.

While the definitions from the medical dictionary (MD) contributed the most for AskAPatient CNN model (with 10.96% performance drop when removed), the definitions, synonyms, and antonyms from the thesaurus (TH) was the

TABLE VIII
TWADR-L EXAMPLES THAT SHOULD HAVE MULTIPLE LABELS

Social Media Phrase	CUI-Ground Truth	Concept - Ground Truth	CUI - Predicted	Concept - Predicted
feel like crap	C0011570	mental depression	C0344315	depressed mood
not being able to eat	C1971624	loss of appetite	C0232462	decrease in appetite
feeling weird	C1443060	feeling abnormal	C0278061	abnormal mental state
depressive emotions and thoughts	C0011570	mental depression	C0086132	depressive symptoms
wide awake	C0455769	energy increased	C0043012	wakefulness

most significant feature for the AskAPatient RNN model (with 2.08% performance drop when removed). These results indicate that each text data from different healthcare domain is very helpful for the deep learning models learn clinical semantics for normalization. Word embeddings built with the larger dataset that combines texts from multiple healthcare domains significantly contributed to improving models performance across both Twitter and AskAPatient datasets when compared to that built from a larger general domain corpus like google news.

B. Qualitative Analysis

Table VIII shows examples that our best model incorrectly predicted. The first column shows example phrases of the social media posts that describe medical conditions, the second and the third columns show the annotated CUIs (unique concept identifier) and corresponding medical concept descriptions, and the fourth and fifth columns show the predicted CUIs and corresponding concept descriptions by our best model (TH + MD + CT + HT). These examples are false positives based on the ground truth labels (i.e., the predicted CUIs do not match the labeled CUIs). However, we can easily observe that, although the CUIs are different, the social media phrases can actually be mapped to both predicted and labeled concepts. For example, the predicted concept 'decrease in appetite' and the label 'loss of appetite' have similar meanings, therefore predicting the phrase 'not being able to eat' to the concept 'decrease in appetite' should be considered correct. While some phrases in the dataset have multiple labels, there are still many more that should have multiple labels (such as those shown in Table VIII).

This suggests several future directions for designing a normalization system. First, it is necessary to have a list of CUIs that represent similar medical concepts so that, when a normalization system predicts a CUI, the mapping can automatically be associated with other CUIs in the same set. Second, the normalization task should be cast as a multi-class multi-label classification problem since each phrase can be mapped to multiple concepts (as shown in Tables III and VIII) and each concept can have many social media phrases (as shown in Table I).

VI. CONCLUSION AND FUTURE WORK

In this work, we explored building neural word embeddings using unlabeled text data from various clinical knowledge sources for medical concept normalization from user-generated social media texts. We have shown that two deep learning models (CNN and RNN) can better predict the medical concepts when we use various clinical domain-specific neural embeddings compared to embeddings trained on a larger general domain text corpus. Our experiments showed that the proposed models with neural embeddings trained on the combined clinical data sources can improve the accuracy up to +21.17% on the Twitter data set and up to +21.28% on the AskAPatient data set. In the future, we would like to explore multi-label normalization approach that integrates hierarchical structure of medical concepts from the standard ontology databases such that the normalization system can automatically associate multiple concepts with similar meanings.

ACKNOWLEDGMENT

This work is supported in part by the following grants: NSF award CCF-1409601; DOE awards DE-SC0007456, DE-SC0014330, and Northwestern Data Science Initiative.

REFERENCES

[1] A. R. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, pages 17–21, 2001.

[2] Y. Bengio, P. Simard, and P. Frasconi. Learning Long-Term Dependencies with Gradient Descent is Difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994.

[3] O. Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32:D267–D270, 2004.

[4] E. G. Brown, L. Wood, and S. Wood. The medical dictionary for regulatory activities (meddra). *Drug Safety*, 20(2):109–117, 2012.

[5] M. Cai, Q. Xu, Y. Pan, W. Pan, N. Ji, Y. Li, H. Jin, K. Liu, and Z. Ji. Adrecs: an ontology database for aiding standardization and hierarchical classification of adverse drug reaction terms. *Nucleic Acids Research*, 43(Database-Issue):907–913, 2015.

[6] L. Chen, K. S. M. T. Hossain, P. Butler, N. Ramakrishnan, and B. A. Prakash. Flu gone viral: Syndromic surveillance of flu on twitter using temporal topic models. In *2014 IEEE International Conference on Data Mining, ICDM 2014, Shenzhen, China, December 14-17, 2014*, pages 755–760, 2014.

[7] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio. On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, 2014.

[8] H. Gurulingappa, A. M. Rajput, A. Roberts, J. Fluck, M. Hofmann-Apitius, and L. Toldo. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics*, pages 885 – 892, 2012.

[9] S. A. Hasan, B. Liu, J. Liu, A. Qadir, K. Lee, V. Datla, A. Prakash, and O. Farri. Neural clinical paraphrase generation with attention. *ClinicalNLP 2016*, page 42, 2016.

[10] S. Karimi, A. Metke-Jimenez, M. Kemp, and C. Wang. Cadec: A corpus of adverse drug event annotations. *Journal of Biomedical Informatics*, 55:73 – 81, 2015.

[11] Y. Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014.

[12] M. Kuhn, I. Letunic, L. J. Jensen, and P. Bork. The SIDER database of drugs and side effects. *Nucleic Acids Research*, 44(Database-Issue):1075–1079, 2016.

[13] R. Leaman, R. I. Dogan, and Z. Lu. Dnorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, 29(22):2909–2917, 2013.

[14] K. Lee, A. Agrawal, and 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*, KDD '13, pages 1474–1477, New York, NY, USA, 2013. ACM.

[15] K. Lee, A. Agrawal, and 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)*, pages 815–822, Aug 2015.

[16] K. Lee, A. Qadir, S. A. Hasan, V. Datla, a. prakash, J. Liu, and O. Farri. Adverse drug event detection in tweets with semi-supervised convolutional neural networks. In *Proceedings of the Twenty-Sixth International World Wide Web conference (WWW 2017)*, Perth, Australia, 2017.

[17] N. Limsopatham and 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, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*, 2016.

[18] D. Lindberg, B. Humphreys, and A. McCray. The Unified Medical Language System. *Methods of Information in Medicine*, 32(4):281–291, 1993.

[19] A. McCallum, K. Bellare, and F. C. N. Pereira. A conditional random field for discriminatively-trained finite-state string edit distance. *CoRR*, abs/1207.1406, 2012.

[20] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[21] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 27th Annual Conference on Neural Information Processing Systems NIPS 2013*, 2013.

[22] A. Nikfarjam, A. Sarker, K. O'Connor, R. Ginn, and 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*, 22:671–681, 2015.

[23] A. Prakash, S. A. Hasan, K. Lee, V. V. Datla, A. Qadir, J. Liu, and O. Farri. Neural paraphrase generation with stacked residual LSTM networks. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 2923–2934, 2016.

[24] A. Prakash, S. Zhao, S. A. Hasan, V. Datla, K. Lee, A. Qadir, and O. F. Joey Liu. Condensed memory networks for clinical diagnostic inferencing. In *The 31st AAAI Conference on Artificial Intelligence (AAAI 2017)*, 2017.

[25] E. S. Ristad and P. N. Yianilos. Learning string-edit distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(5):522–532, May 1998.

[26] H. Sampathkumar, X. Chen, and B. Luo. Mining adverse drug reactions from online healthcare forums using hidden markov model. *BMC Medical Informatics and Decision Making*, 14(1):1–18, 2014.

[27] A. Sarker and G. Gonzalez. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of Biomedical Informatics*, 53:196 – 207, 2015.

[28] I. Sutskever, J. Martens, and G. E. Hinton. Generating Text with Recurrent Neural Networks. In *Proceedings of ICML*, pages 1017–1024, 2011.

[29] S. Tuarob, C. S. Tucker, M. Salathe, and N. Ram. Discovering health-related knowledge in social media using ensembles of heterogeneous features. In *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management*, CIKM '13, pages 1685–1690, New York, NY, USA, 2013. ACM.