Journal of Biomedical Informatics 84 (2018) 93–102

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Medical concept normalization in social media posts with recurrent neural networks

Elena Tutubalina[a,d,*], Zulfat Miftahutdinov[a], Sergey Nikolenko[a,b,c], Valentin Malykh[e,b]

[a] Kazan Federal University, 18 Kremlyovskaya street, Kazan 420008, Russian Federation
[b] St. Petersburg Department of the Steklov Mathematical Institute, 27 Fontanka, St. Petersburg 191023, Russian Federation
[c] Neuromation OU, Tallinn 10111, Estonia
[d] Insilico Medicine, Baltimore, MD 21218, United States
[e] Neural Systems and Deep Learning Laboratory, Moscow Institute of Physics and Technology, 9 bld. 7 Instituski per., Dolgoprudny 141700, Russian Federation

ARTICLE INFO

Keywords:
Natural language processing
Information extraction
Medical concept normalization
Recurrent neural networks
User reviews
Social media

ABSTRACT

Text mining of scientific libraries and social media has already proven itself as a reliable tool for drug repurposing and hypothesis generation. The task of mapping a disease mention to a concept in a controlled vocabulary, typically to the standard thesaurus in the Unified Medical Language System (UMLS), is known as medical concept normalization. This task is challenging due to the differences in the use of medical terminology between health care professionals and social media texts coming from the lay public. To bridge this gap, we use sequence learning with recurrent neural networks and semantic representation of one- or multi-word expressions: we develop end-to-end architectures directly tailored to the task, including bidirectional Long Short-Term Memory, Gated Recurrent Units with an attention mechanism, and additional semantic similarity features based on UMLS. Our evaluation against a standard benchmark shows that recurrent neural networks improve results over an effective baseline for classification based on convolutional neural networks. A qualitative examination of mentions discovered in a dataset of user reviews collected from popular online health information platforms as well as a quantitative evaluation both show improvements in the semantic representation of health-related expressions in social media.

1. Introduction

Recent years have seen many new applications of natural language processing (NLP) to biomedical information. Much of this work has been focused on the central task of information extraction, in particular *named entity recognition* (NER) from scientific literature and electronic health records. However, comparatively little work has been carried out to automatically process social media comments of individuals undergoing medical treatment.

Social media nowadays is a virtually inexhaustible source of people's opinions on a wide variety of topics. In this work, we are focusing on patients' opinions on drug effects, i.e., patient reports. In effect, social media provide huge datasets of people's opinions complete with demographic information and often much more detailed data regarding a specific user. We expect that continuous advancement and improvement in the accuracy of text mining approaches applied to patient reports in social media will have a significant impact in several areas including pharmacovigilance (especially for new drugs), drug repurposing, and understanding drug effects in the context of other factors such as concurrent use of other drugs, diet, and lifestyle.

In this work, we study the problem of discovering disease-related medical concepts from patients' comments on social media. In the context of this problem, we translate a text written in "social media language" (e.g., "I can't fall asleep all night" or "head spinning a little") to "formal medical language" (e.g., "insomnia" and "dizziness", respectively). This goes beyond simple matching of natural language expressions and vocabulary elements: string matching approaches are not able to link social media language to medical concepts since the words may not overlap at all. We call the task of mapping everyday life language to medical terminology *medical concept normalization*. The main benefit of solving this task is bridging the gap between the language of the lay public and medical professionals.

This task is difficult given that patients use social media to discuss different concepts of illness (ranging from well-defined conditions such as "major depressive disorder" to informal phrases describing specific symptoms such as "woke up too early" or "mucus building up in my

* Corresponding author at: Kazan Federal University, 18 Kremlyovskaya street, Kazan 420008, Russian Federation.
E-mail addresses: ElVTutubalina@kpfu.ru (E. Tutubalina), zulfatmi@gmail.com (Z. Miftahutdinov), sergey@logic.pdmi.ras.ru (S. Nikolenko), valentin.malykh@phystech.edu (V. Malykh).

https://doi.org/10.1016/j.jbi.2018.06.006
Received 30 October 2017; Received in revised form 24 April 2018; Accepted 10 June 2018
Available online 12 June 2018

lungs") and a wide diversity of drug reactions (e.g., "excessive sweating at night", "slept like a baby", or "clearing up an infection"). Moreover, social network data usually contains a lot of noise, such as misspelled words, incorrect grammar, hashtags, abbreviations, and different variations of the same word.

Formally, this task is related to several NLP challenges, including paraphrase detection, word sense disambiguation, and entity linking where an entity mention is mapped to a unique concept in an ontology after solving the disambiguation problem [1,2]. To address the challenges described above, recent studies treat the task of linking a one- or multi-word expression to a knowledge base as a supervised sequence labeling problem. Miftahutdinov and Tutubalina [3] proposed an encoder-decoder model based on bidirectional recurrent neural networks (RNNs) to translate a sequence of words from a death certificate into a sequence of medical codes. Two research groups [4,5] presented two systems with similar performances that utilize RNNs for normalization of tweets' phrases at the AMIA 2017 Social Media Mining for Health Applications workshop, while Limsopatham and Collier [6] experimented with convolutional neural networks (CNNs) on social media data. These works demonstrate the first attempts to use deep learning methods for medical concept normalization.

2. Background

Automatic extraction of health-related information from social media is a strong trend in related research nowadays. This task provides a challenging and rich context to explore computational models of natural language, motivating new research in computer science and computational linguistics. For an excellent overview of the work on social analytics for healthcare done up to 2015, see [7], which demonstrates how social media data can be used to mine health-related knowledge.

There exist many applications where a system needs to mediate between natural language expressions and elements of a vocabulary in an ontology. Huang and Lu [8] survey the work done in the organization of biomedical NLP (BioNLP) challenge evaluations up to 2014.

In this section, we give an overview of major findings in previous research on terminology association. In biology, a common task is to identify gene and protein names in text and link them to standard sources such as *Entrez Gene*. Biomedical researchers have addressed the needs to automatically detect diseases as well as corresponding acronyms and abbreviations in the scientific literature (e.g., *BioCreative V* lab). Recent open challenge evaluations have also focused on named entity recognition (NER) of disease names in clinical notes (e.g., *ShARe/CLEF eHealth*, *SemEval 2014*). Ontologies of medical concepts such as the Unified Medical Language System (UMLS) [9], the Systematized Nomenclature Of Medicine Clinical Terms (SNOMED-CT) [10], the Medical Dictionary for Regulatory Activities (MedDRA) [11], and the International Classification of Diseases (ICD-9, ICD-10) are widely used for this task. We note that there is no agreed definition of a disease in general, and diseases may be classified differently by different clinicians [12]. UMLS is undoubtedly the largest lexico-semantic resource for medicine which represented over more than 150 lexicons with terms from 25 languages. In particular, ICD and SNOMED-CT are subsets of UMLS. Every concept is represented by its Concept Unique Identifier (CUI). UMLS has integrated resources used worldwide in clinical care, public health, and epidemiology.

Automatic approaches to BioNLP tasks roughly fall into two categories: (i) linguistic approaches based on dictionaries, association measures, morphological and syntactic properties of texts [13–17]; (ii) machine learning approaches [18–21,6,3]. Recent studies have employed deep learning models such convolutional neural networks [6,5] or recurrent neural network architectures [3,5,4].

In the rest of the section, we describe methods that were trained on publicly available data of different health-related sources and related to the shared tasks such as Social Media Mining shared tasks, *CLEF eHealth* tasks, and *SemEval* tasks.

2.1. Bio- and medical natural language processing

A lot of work in bio- and medical NLP has been focused on language evaluation, information retrieval, and extraction from electronic medical records and biomedical academic literature. In their book, Cohen and Demner-Fushman [22] gave an overview of major challenges and the work done in biomedical NLP up to 2014.

The most popular knowledge-based system for mapping texts to UMLS concept identifiers (CUI) is MetaMap [13]. MetaMap was developed by the National Library of Medicine (NLM) in 2001 and has become a de facto baseline method for many recent studies. This system is based on a linguistic approach using lexical lookup and variants by associating a score with phrases in a sentence. General limitations of the linguistic method include low recall of information extraction from social media and unavailability for under-resourced natural languages.

The *ShARe/CLEF eHealth 2013* lab addressed the problem of identification and normalization of disorders from clinical reports in Task 1 [23]. The corpus consists of discharge summaries and electrocardiogram, echocardiogram, and radiology reports received from US intensive care. Each disorder mention is mapped to a UMLS code or a SNOMED-CT code. The best results were achieved with a DNorm system by NCBI team [18]. Leaman et al. introduced a DNorm system for assigning disease mentions from PubMed abstracts a unique identifier from a MEDIC vocabulary, which combines terminology from Medical Subject Headings (MeSH) and Online Mendelian Inheritance in Man (OMIM) [19]. DNorm consists of a text processing pipeline, including the named entity recognizer to locate diseases in the text, and a normalization method. The normalization method is based on a pairwise learning-to-rank technique using the tokens from all mentions as features. DNorm outperformed MetaMap as the baseline.

The *SemEval 2014* lab addressed the problem of analysis of clinical data in Task 7 [1]. This task was a follow-up to the ShARe/CLEF eHealth 2013 task 1 using a larger test set and a larger set of unlabeled MIMIC notes to inform models and generalize lexical features. The best results were obtained by UTH_CCB and UWM teams [17,16]. In [16], the UWM team present a pattern-based system that consists of several steps. First, the system looks for exact matches with disorder mentions in the training data and in the UMLS. Second, for every mention without exact matches, suitable variations were generated based on Levenshtein distances between the variations. Edit distance patterns were computed between all synonyms of the disorder concepts is UMLS as well as between their mentions in the training data. In [17], the UTH_CCB team used the cosine similarity scores between disorder entity and all UMLS terms to rank candidate terms.

The *CLEF Health* 2016 and 2017 labs addressed the problem of mapping death certificates to ICD codes. Death certificates are standardized documents filled by physicians to report the death of a patient [24]. Most submitted methods utilized dictionary-based semantic similarity and, to some extent, string matching. Mulligen et al. [14] obtained the best results in task 2 by combining a Solr tagger with ICD-10 terminologies. The terminologies were derived from the task training set and a manually curated ICD-10 dictionary. They achieved an F-measure of 84.8%. Mottin et al. [15] used pattern matching approach and achieved the F-measure of 55.4%. Dermouche et al. [20] applied two machine learning methods: (i) a supervised extension of Latent Dirichlet Allocation (LDA), i.e., Labeled-LDA and (ii) Support Vector Machine (SVM) based on bag-of-words features. For Labeled-LDA, they used ICD-10 codes from the training set as classes. The Labeled-LDA and SVM classifier achieved F-measures of 73.53% and 75.19%, respectively. This study did not focus on designing effective features to obtain better classification performance. Zweigenbaum and Lavergne [25] utilized a hybrid method combining simple dictionary projection and mono-label supervised classification. They used Linear SVM trained on the full training corpus and the 2012 dictionary provided for CLEF participants. This hybrid method obtained an F-measure of 85.86%. The participants of task 2 did not use word embeddings or deep neural

networks. Recently, Miftakhutdinov and Tutubalina [3] obtained the best results in the CLEF eHealth 2017 task 1, training an LSTM-based encoder-decoder architecture. As input, the network uses the certificates' text lines that contained terms that could be directly linked to a single ICD-10 code or several codes. As output, the network predicts a sequence of codes. The model obtained an F-measure of 85.01% on English texts. Word embeddings trained on social media texts improved the results significantly over embeddings on medical articles downloaded from BioMed Central.

2.2. Word sense disambiguation and entity linking

Capturing semantic similarity or relatedness between one- or multi-word expressions has many applications in NLP including paraphrase detection, machine translation, word sense disambiguation (WSD), and entity linking (EL). Over the last decade, NLP community has tackled EL (or entity disambiguation) which aims to link mentions of entities to a knowledge base (KB) such as, e.g., *Wikipedia*. Word sense disambiguation (WSD) is another closely related problem that addresses the lexical ambiguity of text by making explicit the meaning of words occurring in a context [26]. Herein, we briefly describe recent research on EL and WSD which is a closely related task to concept normalization studied in this work. A detailed overview of the research on these topics can be found in [26,27].

Recent research on EL has focused on training neural networks and word embeddings for capturing characteristics of an entity at various levels of information. He et al. proposed a deep learning approach that learns context-entity similarity measure in which representations are learned with Wikipedia annotations [28]. Huang et al. leveraged both structured and contextual information from large-scale semantic knowledge graphs and deep neural networks for EL [29]. Gupta et al. proposed a training objective to learn unified entity embeddings that encode text description in KB, contexts from documents, and sets of fine-grained types [30]. Babelfy [31] is the state-of-the-art unified approach to WSD and EL. This graph-based approach exploits the semantic network which encodes rich structural and lexical information of WordNet and various KBs. Raganato et al. [32] compared several neural architectures for WSD including bidirectional LSTM and a sequence-to-sequence model. They adopted the SemCor corpus which consists of a subset of the Brown Corpus (approximately 240,000 sense-annotated words). LSTM with an attention layer outperformed the encoder-decoder model and achieved state-of-the-art results over several supervised and knowledge-based systems in benchmarks.

Thus, there are several main challenges in adopting WSD and EL systems for medical concept normalization in social media texts related to the health domain. First, it is difficult to match the phrase to the correct concept in social media texts due to shorter local and document-level context of an entity. Second, there is a large language difference between medical terminology and patient vocabulary [33]. Third, UMLS includes a large number of non-noun concepts [34], while many existing knowledge-based systems such as Babelfy [31] take into account only nominal and named entity mentions occurring within a text. Finally, Wikipedia contains different kinds of information about entities such as textual description, linked mentions, while the content of UMLS is limited and complex. In [34], Nadkarni et al. explored the feasibility of using UMLS to identify concepts in discharge summaries and surgical notes. They concluded that concept indexing by UMLS can't be the only production-mode means of preprocessing medical narrative and indicated potential problems. In [35], Polepalli et al. evaluated a method for linking EHR notes to three knowledge resources: Wikipedia, UMLS, and MedlinePlus. In particular, their method utilized first three lines of a Wikipedia page and definitions of UMLS concepts. Experiments showed that linking to Wikipedia yielded the best performance, yet Wikipedia does not integrate classification and coding standards. Therefore, development of WSD and EL methods for health domain is challenging research topic, but will not be discussed further in this paper. We note this topic as a possible direction for further work.

2.3. Concept normalization in social media posts

While there has been a lot of work on named entity recognition from social media posts done over the past 7 years [36,37,2,38–43], relatively few researchers have looked at assigning social media phrases to medical identifiers. The first Social Media Mining shared task workshop (organized as part of the Pacific Symp. on Biocomputing 2016) was designed to mine pharmacological and medical information from social media, with a competition based on a published dataset [40]. Task 3 of this competition is devoted to medical concept normalization, where participants were required to identify the UMLS concept for a given ADR. The evaluation set consisted of 476 ADR instances. Sarker et al. [40] noted that there had been no prior work on normalization of concepts expressed in social media texts, and task 3 did not attract much attention from the academic community.

Recently, two teams, namely UKNLP [4] and gnTeam [5], participated in the Second Social Media Mining for Health (SMM4H) Shared Task and submitted their systems for automatic normalization of ADR mentions to MedDRA concepts. For the task 3, Sarker et al. [44] created a new dataset of tweets' phrases. The training set for this task contains 6,650 phrases mapped to 472 concepts, while the testing set consisted of 2,500 phrases mapped to 254 classes. We also note that organizers of this task did not describe the corpus creation in details as well as not providing corpus statistics, e.g., the overlap percentage between training and testing sets. Teams' systems showed similar results. The gnTeam's approach contained three components for preprocessing and classification. The first two components corrected spelling mistakes and converted sentences into vector-space representation, respectively. For the third step, GnTeam adopted multinomial logistic regression model which achieved the accuracy of 0.877, while the bidirectional GRU achieved the accuracy of 0.855. As input, the network adopted the GoogleNews embeddings trained on a Google News corpus due to higher results the highest performance over embeddings trained on tweets. The ensemble of both classifiers showed slightly better performance and achieved the accuracy of 0.885. The UKNLP's system adopted hierarchical LSTM in which a phrase is segmented into words and each word is segmented into characters. Word embeddings were trained on a Twitter corpus. Hierarchical Char-LSTM achieved the accuracy of 0.872, while hierarchical Char-CNN performed slightly better and achieved the accuracy of 0.877. We note this corpus of tweets for future work since the official test data is available for the shared task participants only by the time of publication.

Recently, Limsopatham and Collier [6] experimented with Convolutional Neural Networks (CNN) and pre-trained word embeddings for mapping social media texts to medical concepts. For evaluation, three different datasets were used. The authors created two datasets with 201 and 1,436 Twitter phrases which mapped to concepts from a SIDER database. The third dataset is the CSIRO Adverse Drug Event Corpus (CADEC) [2] which consists of user reviews from askapatient.com. The authors observed that training can be effectively achieved at 40–70 epochs. As input, the network concatenated embeddings of words. The GoogleNews embeddings improved results significantly over embeddings on medical articles. Experiments showed that CNN (accuracy 81%) outperformed DNorm (accuracy 73%), RNN (accuracy 80%) and a multi-class logistic regression (accuracy 77%) on the AskAPatient corpus (as well as corpora of tweets). This work is the closest to ours in the use of deep learning technology and semantic representation of words. However, we found that only approximately 40% of expressions in the test data are unique, while the rest of expressions occur in the training data. Therefore, the presented accuracy may be too optimistic. We believe that future research should focus on developing extrinsic test sets for medical concept normalization.

To sum up this section, we note that there has been little work on medical concept normalization in social media posts, and most methods in the biomedical domain have so far dealt with extracting information from the mention itself, ignoring the broader context of the text

document where it occurred.

3. Material and methods

In this section, we discuss major challenges in the medical concept normalization task and present recurrent neural networks and architectures such as LSTM and GRU.

3.1. Recognition of different word variants

The task of medical concept normalization is closely related to the problem of word sense disambiguation and terminological variation. We briefly describe major challenges faced by disease mention recognition methods as well as term extraction methods:

(a) lexical, morphological, and syntactic variants: failure weight gain – failure to gain weight (CUI C0231246), hypertension – high blood pressure disorder (CUI C0020538), auricular fibrillation – atrial fibrillation (CUI C0004238);
(b) paraphrases, synonyms: acute disorder – acute disease (CUI C0001314), a drop of cholesterol – lower total cholesterol (CUI C1868135), concentration lack – unable to concentrate (CUI C0235198);
(c) abbreviations: ADDH, ADHD – attention deficit hyperactivity disorder (CUI C1263846), AIDS – acquired immunodeficiency (CUI C0596032), AF – atrial fibrillation (CUI C0004238);
(d) ambiguity: aspiration – pulmonary aspiration (CUI C0700198) or aspiration pneumonia (CUI C0032290);
(e) misspellings, slang terms, and shortened forms of words: probs with sleeping, dont sleep to well – difficulty sleeping (CUI C0235162), not sure footed as I walked – unsteady when walking (CUI C0231686), afib, a fib – atrial fibrillation (CUI C0004238).

The examples described above are associated with Concept Unique Identifiers (CUI) from UMLS.

3.2. Neural networks

At present, neural networks represent a widely used machine learning framework applicable to a wide variety of tasks, especially data-intensive ones and tasks dealing with unstructured data such as images or natural language. A neural network (NN) itself is a computational graph of a specific kind. The simplest kind of NNs are the so-called Feed-Forward Networks (FFN). One feed-forward layer in a FFN is a transformation of an input vector to a output vector by multiplying it by some weight matrix and applying some non-linear function afterward. More information on basic NNs can be found in, e.g., [45,46].

3.2.1. Recurrent neural networks

While FFNs are simplest and in a way most general kind of NNs, there are other specific types of computational graphs that are especially useful for particular tasks. One of such specific types are Recurrent Neural Networks (RNN). The RNNs are applied for processing of sequential data such as time series or word sequences. The key feature of that architecture is information sharing between timesteps. More specifically, popular variants of RNNs that we discuss later in this chapter use the notion of a *state* for RNN, which at every time step is received by the RNN from the previous timestep. This mechanism is depicted on Fig. 1a.

3.2.2. Long short-term memory

The Long Short-Term Memory (LSTM) architecture is inspired by human brain short-term memory mechanism [47]. The idea behind it is to have some explicit memory cell for information storage and then an additional mechanism to operate with this memory. This mechanism consists of two [47] or three [48] non-linear functions applied to:

- input, i.e. what should be added to the memory cell,
- output, i.e. how strong we want the output signal to be,
- forget, i.e. how much of stored information we want to throw off.

These non-linear functions are called gates. These considerations are more formally spelled out in Eq. (1), and the graphical representation can be found on Fig. 2a.

Formally speaking, an LSTM has an *input gate*, *forget gate*, and *output gate*, together with the actual recurrent cell with a hidden state. We denote by $\boldsymbol{x}_t$ the input vector at time t; by $\boldsymbol{h}_t$, the hidden state vector at time t; by $W_{x.}$ (with different second subscripts), matrices of weights applied to the input; by $W_{h.}$, matrices of weights in recurrent connections; by $\boldsymbol{b}$, the bias vectors. In this notation we get the following formal definition: on step t, having received input $\boldsymbol{x}_t$, previous hidden state $\boldsymbol{h}_{t-1}$, and cell state $\boldsymbol{c}_{t-1}$, LSTM computes $\boldsymbol{h}_t$.

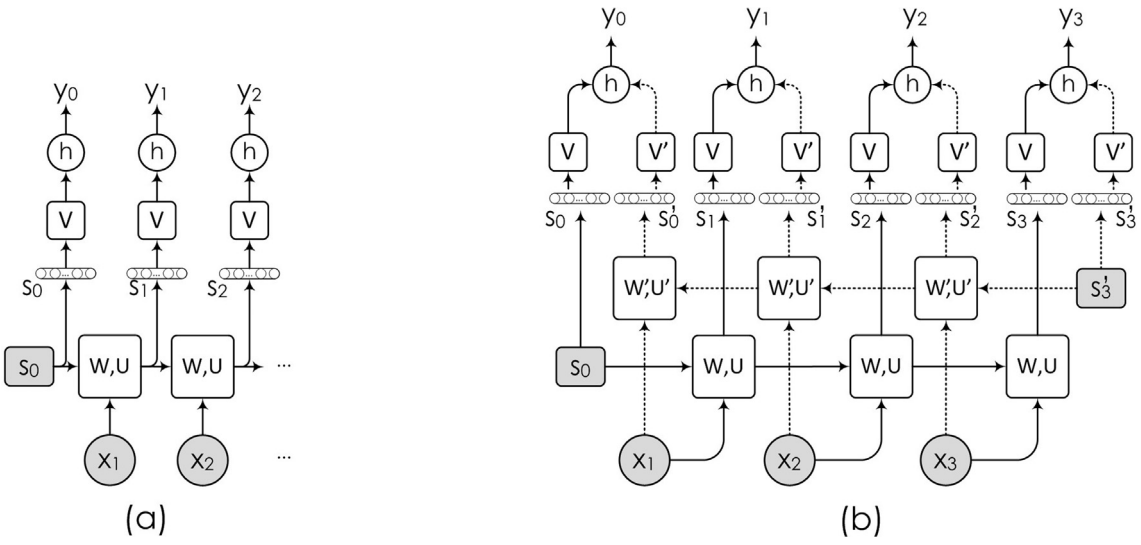

Fig. 1. Recurrent neural networks: (a) a regular RNN; (b) a bidirectional RNN.

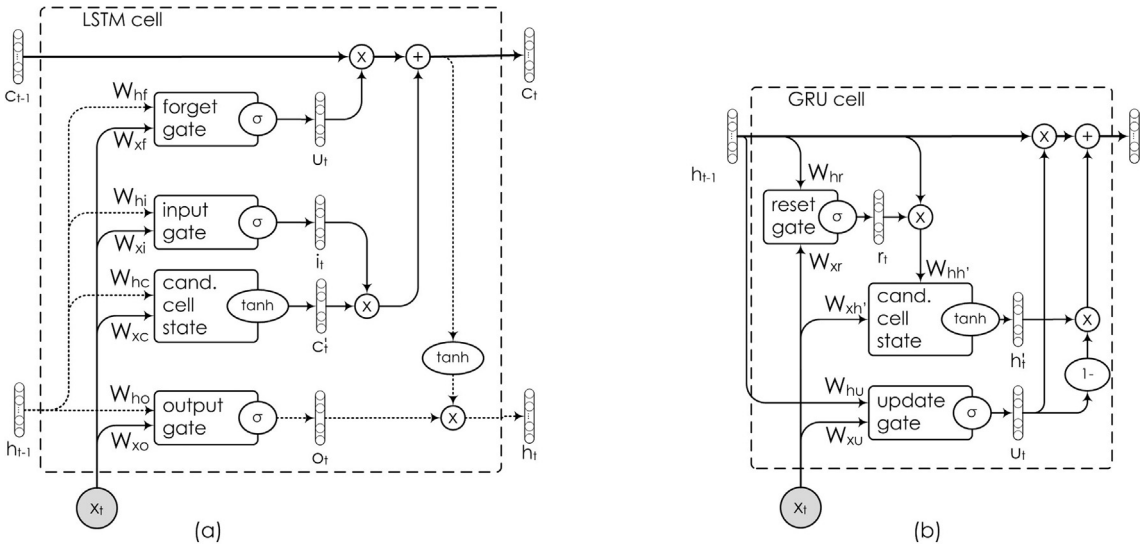

Fig. 2. Modern RNN units: (a) LSTM; (b) GRU.

$$
\begin{aligned}
c_t' &= \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_{c'}) && \textit{candidate cell state}\\
\imath_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) && \textit{input gate}\\
f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) && \textit{forget gate}\\
\text{ø}_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) && \textit{output gate}\\
c_t &= f_t \odot c_{t-1} + \imath_t \odot c_t', && \textit{cell state}\\
h_t &= \text{ø}_t \odot \tanh(c_t) && \textit{block output}
\end{aligned} \tag{1}
$$

3.2.3. Gated recurrent units

The Gated Recurrent Unit (GRU) is another common approach to memorize the state between timesteps. It was introduced in [49]. The GRU has only two gates, namely reset & update gates:

- reset gate is analogous to the forget gate in LSTM,
- update gate can be considered as combination of input and output gates.

The graphical representation can be found on Fig. 2b. A formal representation of GRU is shown in Eq. (2). Denoting by $\boldsymbol{x}_t$ the input vector at time t, by $\boldsymbol{h}_t$ the hidden state vector at time t, by $W_{x\cdot}$ (with different second subscripts) the matrices of weights applied to the input, by $W_{h\cdot}$ matrices of weights in recurrent connections, and by $\boldsymbol{b}$ the bias vectors, we get the following formal definition: on step t, having received input $\boldsymbol{x}_t$, and previous hidden state $\boldsymbol{h}_{t-1}$, GRU computes $\boldsymbol{h}_t$ as follows.

$$
\begin{aligned}
u_t &= \sigma(W_{xu}x_t + W_{hu}h_{t-1} + b_u),\\
r_t &= \sigma(W_{xr}x_t + W_{hr}h_{t-1} + b_r),\\
h_t' &= \tanh(W_{xh'}x_t + W_{hh'}(r_t \odot h_{t-1})),\\
h_t &= (1-u_t) \odot h_t' + u_t \odot h_{t-1}.
\end{aligned} \tag{2}
$$

Recent extensive practical comparisons indicate that GRUs achieve similar performance on sequence modeling problems [50], and due to fact that GRU has lesser number of trainable parameters, GRU becomes more and more popular.

3.2.4. Bidirectional RNN

Another option to alleviate the forgetfulness of RNNs is to use two RNNs in opposite directions on the input: one in standard direction from start to end, and another one from end to start [51]. RNN outputs for the corresponding sequence points are concatenated together. This approach, known as *bidirectional RNNs*, is presented on Fig. 1b.

3.3. Neural network architecture for concept normalization

We propose a deep learning approach for mapping entity mentions to medical codes. We first convert each mention into a semantic representative vector using bidirectional LSTM or GRU with an attention mechanism on top of the embedding layer. For activation, we use the hyperbolic tangent $\tanh(a) = (e^a - e^{-a})/(e^a + e^{-a})$. Then, a set of features are extracted using the cosine similarity between mentions and medical concepts from the UMLS Metathesaurus. For model training, we use the cross-entropy error between the data distribution and predicted distribution as the loss function. This model is depicted in Fig. 3.

3.3.1. Semantic similarity features

We extract a set of features to enhance the representation of the phrases. These features consist of the cosine similarity between the vectors of the input phrase and a concept in a medical terminology dictionary. This dictionary includes medical codes and synonyms from the UMLS Metathesaurus (version 2017 AA), where codes are presented in the CADEC corpus. We apply three strategies to create representations of a concept and a mention and compute the cosine similarity between the representations of each pair:

- TF-IDF (ALL): we represent a medical code as a single document by concatenating synonymous terms; then, we apply the TF-IDF transformation on the code and the entity mention and compute the cosine similarity;
- TF-IDF (MAX): we represent a medical code as a set of terms; for each term, we compute the cosine distance between its TF-IDF representation and the entity mention and then select the largest similarity;
- W2V (ALL): we represent a medical code as a single document by concatenating synonymous terms; then we embed a code and an entity mention as averaged sums of the embeddings of its words and compute the cosine similarity.

We use all these additional features for the experiments.

3.3.2. Pre-trained word embeddings

Neural networks require word representations as inputs. We investigate the use of several different pre-trained word embeddings. Recent advances have made *distributed word representations* into a method of choice for modern NLP [52]. In this model, each word from the dictionary is mapped to a Euclidean space $\mathbb{R}^d$ (i.e., to a vector of d

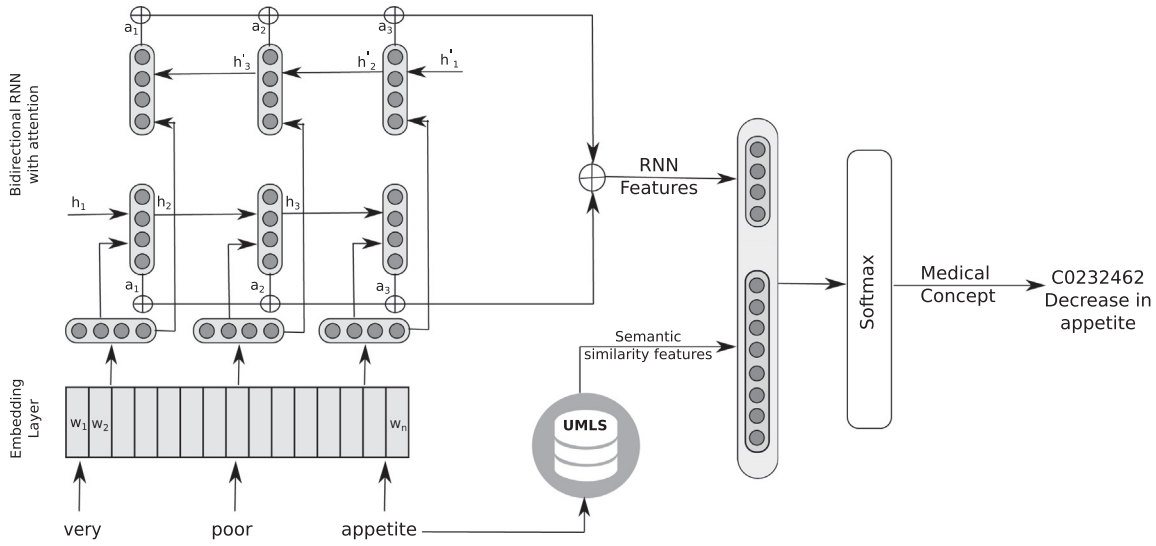

Fig. 3. Proposed architecture for medical concept normalization.

Table 1
Statistics of the dataset used in the experiments.

Entity type	Total	Unique phrases	Unique SNOMED codes
ADR	5838	3241	788
Disease	266	165	108
Drug	1657	290	124
Finding	399	270	180
Symptom	251	128	78

real numbers), in an attempt to capture the semantic relationships between the words as geometric relationships in $\mathbb{R}^d$. In a way, word embeddings rely upon what is long known in linguistics as the *distributional hypothesis*: words with similar meaning will occur in similar contexts [53]. The two most commonly used modern models for word embeddings, *Continuous Bag-of-Words* (CBOW) and *skip-gram*, were both introduced in [54].

We utilize word embeddings named *HealthVec*: publicly available 200-dimensional embeddings that were trained on 2,607,505 unlabeled user comments (93,526 terms) from health information websites using the CBOW model in [41]. We also experimented with another published 200-dimensional embeddings named *PubMedVec* (2,351,706 terms), which were trained on biomedical literature indexed in PubMed [55].

4. Experimental evaluation

In this section, we present the results of our evaluation study. The purpose of our evaluation is to determine how well recurrent neural networks can identify the corresponding medical concepts based on informal language from patients' texts.

4.1. Data set

We conducted experiments on a collection of user reviews obtained from the CADEC corpus [2]. This corpus contains 1,250 reviews and consists of four predefined disease-related types: ADR (6,318 entities), Disease (283 entities), Symptom (275 entities), and Clinical Finding (435 entities). Authors reported that only 39.4% of the annotations (including drugs) were unique; people generally discussed similar reactions. Disease and Symptom specify the reason for taking the drug. Patients may mention the name of a disease or the symptoms that led to them taking a drug. Findings are any adverse side effects, diseases, or symptoms that were not directly experienced by the reporting patient. We did not distinguish between these types and joined them into one class of annotations named *Disease*.

All entities in the CADEC corpus were mapped to SNOMED CT-AU (SCT-AU) by a clinical terminologist. SNOMED CT is a clinical terminology that provides codes, synonyms, and definitions of clinical terms, and can be accessed through the UMLS Metathesaurus. Additionally, concepts identified in the SNOMED CT were associated with MedDRA identifiers. In this work, we adopted only SNOMED CT identifiers and removed 'concept less' or ambiguous mentions for evaluation purpose. Table 1 shows final statistics for the CADEC corpus. The total number of unique codes was 1,029.

4.2. Preprocessing and experimental settings

Preprocessing includes spelling correction and lemmatization using the Natural Language Toolkit (NLTK). We performed 5-fold cross-validation to evaluate our methods. We found that a standard cross-validation method creates a high overlap of expressions in an exact match between training and testing parts. Therefore, the split procedure has a specific feature in our setup.

For preprocessing, we first removed all duplicates in each dataset. Second, we grouped medical records we are working with into sets that were each related to a specific medical code. Every such set was split independently into k folds, and all these folds were merged into final k folds. This procedure is formalized in Algorithm 1. We have made the resulting folds publicly available[1].

Algorithm 1. Split dataset into the folds for evaluation

procedure SPLIT (D, C, k)
  D - set of medical records
  C - medical codes
  k - number of folds
  F - final split
  for $i = 0$ to k **do**
    $F_i := \emptyset$
  end for
  for c in C **do**
    $g = \{ \forall\, d \in D$: d relates to $c\}$
    $E = kfoldsplit(g, k)$
    for $i = 0$ to k **do**

[1] https://yadi.sk/d/oLBTUpXg3RtCzd.

13: $F_i := F_i \bigcup E_i$
14: **end for**
15: **end for**
16: return F
17: **end procedure**

4.3. Baseline system

For comparison, we applied state-of-the-art baselines based on convolutional neural networks. In [6], experiments showed that CNN outperformed existing strong baselines such as DNorm and Logistic Regression, as well as unsupervised metrics such as TF-IDF, BM25, and cosine similarity between a phrase and a medical concept.

In order to obtain local features from a text with CNNs, we used multiple filters of different lengths [56]. The architecture is illustrated in Fig. 4. At the same time, each filter on a hidden layer was replicated across the entire input vector, learning the same localized features in every part of the input. Convolutional layers are usually interleaved with pooling or subsampling layers that combine the subsets of the input and output the maximum values of all features; here the idea is that a higher-level feature's exact location is less important than its interaction with other neighboring features. In one-dimensional CNNs, these are usually the max-over-time pooling layers, which output the maximal value of a feature map along with a window.

4.4. Model configuration and training

Since neural networks, especially deep neural networks, have a very large number of free parameters, problems with overfitting are inevitable, and some form of regularization is required. In the now-common dropout technique [57], units in a neural network are "switched off" at random during training, thus making each unit learn a useful feature "by itself" since it cannot rely upon other units to be present to form compositions with it. We used a dropout rate of 0.5 after the embedding layer (and before the networks' layers).

Another standard technique in modern deep learning, batch normalization [58], was designed to cope with a problem known as covariate shift. For all networks, we set the mini-batch size to 128 to minimize the negative log-likelihood of correct predictions.

The last important set of advances deal with actually training the model. We used a popular adaptive gradient descent variation, Adam [59]. Embedding layers are trainable for all networks. The number of outputs of the layer with the softmax activation equals to the number of unique concept codes. Additionally, we separated out 10% of the training set to form the validation set that was used to evaluate different model parameters. The number of epochs is determined by early stopping on the validation set. We employed early stopping after two epochs with no improvement on the validation set. The final number of epochs was 15.

Table 2
The accuracy performance of neural networks.

Model	Parameters	Accuracy
CNN	HealthVec, 100 feature maps	46.19
CNN	PubMedVec, 100 feature maps	45.79
LSTM	HealthVec, 200 hidden units	64.51
LSTM	PubMedVec, 200 hidden units	64.24
GRU	HealthVec, 200 hidden units	63.05
GRU	PubMedVec, 200 hidden units	62.73
LSTM + Attention	HealthVec, 200 hidden units	65.73
LSTM + Attention	PubMedVec, 200 hidden units	64.92
LSTM + Attention	HealthVec, 100 hidden units	64.83
GRU + Attention	HealthVec, 200 hidden units	**67.08**
GRU + Attention	PubMedVec, 200 hidden units	66.55
GRU + Attention	HealthVec, 100 hidden units	66.56
	With prior knowledge	
LSTM + Attention	HealthVec, 100, similarity: TF-IDF (ALL)	67.63
LSTM + Attention	HealthVec, 200, similarity: TF-IDF (ALL)	66.83
GRU + Attention	HealthVec, 100, similarity: TF-IDF (ALL)	69.92
GRU + Attention	HealthVec, 200, similarity: TF-IDF (ALL)	69.42
GRU + Attention	HealthVec, 100, similarity: W2V (ALL)	69.14
GRU + Attention	HealthVec, 100, similarity: TF-IDF (MAX)	**70.05**

For RNN, we utilized either a 100- or 200-dimensional hidden layer for each RNN chain. For CNN, we adopted effective parameters from [56,6]. We used the filter w with the window size h of [3, 4, 5], each of which had 100 feature maps. Pooled features were fed to a fully connected feed-forward neural network (with dimension 100) to make an inference, using rectified linear units as output activation.

We found 91% and 88% of words from the CADEC corpus vocabulary in the vocabularies of HealthVec and PubMedVec word embeddings respectively. For other words, their representations were uniformly sampled from the range of embedding weights [60].

4.5. Results

The standard technique for evaluating concept normalization is to compare correctly normalized disorder mentions against gold standard entities [23,1]. We used accuracy as the performance measure, defined as follows:

$$Accuracy = \frac{N_{correct}}{T_g}, \tag{3}$$

where $N_{correct}$ is the number of correctly normalized disorder mentions and T_g is the total number of disorder mentions in the gold standard.

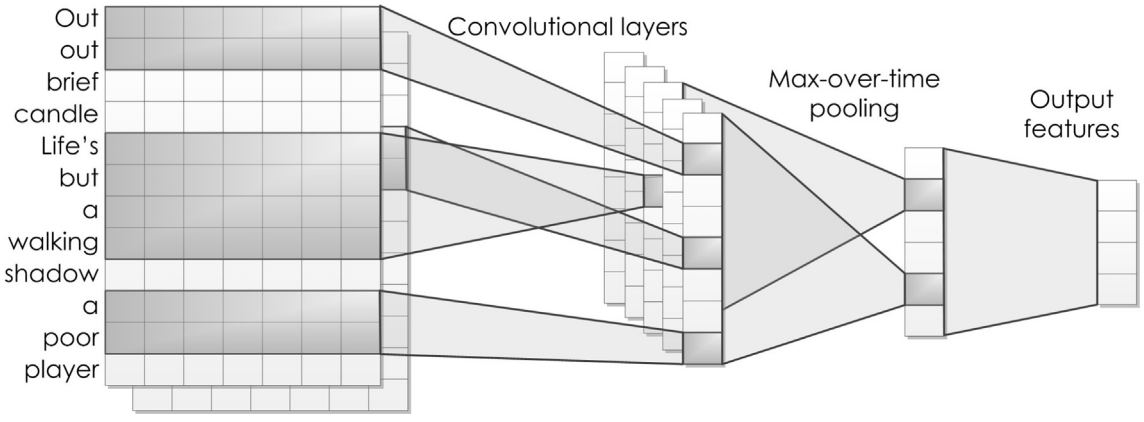

Fig. 4. A convolutional neural network with 1D convolutions over a text.

Table 3
Summary of statistics of entity mention phrases of different length.

Length of a mention	# Mentions
1	11
2	558
3	1064
4	739
5	531
6 or longer	662

Table 4
The performance of GRU + Attention with sim. features TF-IDF (MAX).

Length of a mention	Accuracy
1	67.16
2	71.43
3	72.91
4	72.15
5	65.49
6 or longer	50.12

Table 5
The accuracy performance of our proposed model and the state-of-the-art methods on AskAPatient folds.

Model	Accuracy
DNorm [6]	73.39
CNN [6]	81.41
RNN [6]	79.98
GRU + Attention (HealthVec, 100, TF-IDF (MAX))	85.71

We present experimental results of different neural architectures in Table 2. Attention-based GRU with prior knowledge achieved an accuracy of 70.05%. The best results were obtained while using vectors trained on social media posts. GRU consistently outperformed CNNs and LSTM in terms of accuracy. Attention mechanism and prior knowledge from the UMLS Metathesaurus indeed led to quality improvements for both GRU and LSTM.

We evaluated the performance of our best model on entity mention phrases of different lengths. Statistics of phrases is presented in Table 3. As can be seen from the results in Table 4, GRU with attention and similarity features TF-IDF (MAX) achieved the best results at 72.91% accuracy and 72.15% accuracy on three-word and four-word expressions, respectively.

The comparison between the proposed method and existing state-of-the-art methods on the CADEC corpus is provided in Table 5. We trained and evaluated GRU with attention and similarity features TF-IDF (MAX) on folds[2] (namely AskAPatient folds) from [6]. Experimental results shown in Table 5 show significant improvements in the concept normalization performance using GRU with attention and prior knowledge. Additionally, we evaluated our best model with different sizes of HealthVec embedding vectors. As it can be seen from Table 6, changing the size of the embedding vector from 200 to 400 did not significantly influence the results.

We explored the quality metrics for the number of training epochs ranging from 0 to 60. Fig. 5 presents the results achieved on the testing set by training with different numbers of epochs. It shows that training of neural networks can be effectively achieved at around 8–15 epochs before improvements in accuracy become small.

[2] DOI: doi:https://doi.org/10.5281/zenodo.55013.

Table 6
The accuracy performance of GRU + Attention with sim. features TF-IDF (MAX) and different embedding dimensions

Embeddings	Accuracy
100-dimensional HealthVec	65.82
200-dimensional HealthVec	70.05
300-dimensional HealthVec	69.48
400-dimensional HealthVec	69.37

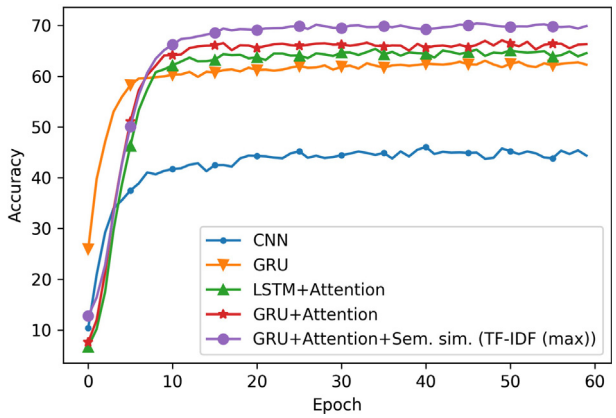

Fig. 5. Performance on the testing set achieved by training with different numbers of epochs.

5. Discussion

5.1. Failure analysis

The primary goal of our model was to map disease-related mentions into medical codes. We evaluated the capability of the attention-based GRU in fulfilling this goal and examined the output of our network. One limitation of deep learning technique is the need for sufficient training data; otherwise, RNNs do not perform well on rare, long, discontiguous, and overlapping expressions. For example, the mention “toes became so painful” is automatically associated with the concept “Pain” (SNOMED ID 22253000), while the ground truth label is the less frequent concept “Pain in toe” (SNOMED ID 285365001). Handling discontiguous expressions may involve word reordering: “toes became so painful” into “painful toes”. Another group of errors is related to expressions that overlap in meaning and share similar characteristics. For example, GRU associates the mention “could only walk less than 100 meters” with the concept “Walking disability” (SNOMED ID 228158008), while the ground truth concept is “Reduced mobility” (SNOMED ID 8510008). The mention “foggy thinking” is automatically associated with “Unable to think clearly” (SNOMED ID 247640008), while the ground truth concept is “Mentally dull” (SNOMED ID 419723007). The group of errors is related to the problem of ambiguity of one-word expressions. For example, “stiff” is automatically associated with “Stiff legs” (SNOMED ID 225609009) instead of “Stiffness” (SNOMED ID 271587009). We believe that these errors may be solved with advanced language modeling techniques and mark these models as future work.

5.2. Qualitative analysis

The primary goal of applying bidirectional RNNs to medical concept normalization is to capture “semantic representation” of a text based on not only the past but also the future context on every time step. Therefore, we selected several examples of disease-related entity mentions with corresponding medical concepts discovered in social media posts. First, many patients are regularly unsure of the spelling of medical terms such as *diarrhea*, which justifies the need for spelling

correction. Second, patients report disease mentions with multi-word expressions more frequently than with single words. Finally, patients tend to reformulate some medical terms into semantically similar words. For example, the sense of *pain* linked with the concept of "Pain in lower limb" is frequently replaced by words such as *killing*, *hurting*, and *aching*. The sense of the specific term *impairment*, associated with the concept of "Memory impairment", is frequently replaced by more general words such as *loss*, *poor*, *fog*, *weakened*, and *difficulty*. The sense of the word *raised* in the concept of "Serum cholesterol raised" is frequently replaced by words such as *high*, *went up*, *elevated*, and *jumped*, *climbed*. Therefore, the social media domain poses additional challenges that string matching techniques are not able to handle. The semantic information encoded in word embeddings trained on a large corpus of health-related consumer comments is critically important for model performance.

5.3. Limitations

We acknowledge three groups of limitations to this study. First, there are certain limitations associated with the data. One potential drawback of any supervised model is the lack of data for its development and evaluation. One of the main challenges is the lack of annotated corpora that cover various domains of texts across languages, categories of diseases, and concepts from biological thesauri. Existing corpora contain texts about a specific group of drugs or diseases. Hence, the coverage of rare adverse events or diseases is limited, and trained models might not accurately assign codes of rare terms to textual fragments. Second, most existing corpora contains disease mentions linked to their corresponding concepts in a particular resource, such as SNOMED, MeSH, ICD-10, or MedDRA. Therefore, models are limited by a controlled vocabulary and do not predict a variety of related medical concepts with different identifiers. In this paper, our model predicts medical codes in SNOMED only. We note that the linking of some SNOMED terms with UMLS concepts are ambiguous and one-to-one mapping between entries in SNOMED and another terminology dictionary (e.g., MedDRA) using UMLS is not always possible.

Second, the use of large-scale prior knowledge from UMLS brings new challenges. The number of unique concepts in a large knowledge base such as the UMLS Metathesaurus is often very large (more than 3 million concepts), which makes classification methods and integration of linguistic knowledge very computationally expensive. In this study, we did not integrate all medical concepts from UMLS. As future work, we plan to investigate the use of UMLS for learning concept embeddings.

Finally, another limitation is that the current model architecture does not allow for predicting a sequence of medical concepts associated with a particular disease mention. Our model is insufficient for sentences, overlapping annotations or multi-word expressions associated with more than one concept. For instance, the expression "severe pain in my left arm" is associated with both "Severe pain" (SNOMED ID 76948002) and "Pain in left arm" (SNOMED ID 287045000) in the CADEC corpus of user reviews, while the sentence "CAD/s/pCABG/Volume overload" is associated with "Acute coronary artery disease" (ICD code I251) and "Fluid overload" (ICD code E877) in the CDC corpus of death certificates used in the CLEF eHealth 2017 Evaluation Lab Task 1. This task has recently received research attention and some of the proposed methods, such as the encoder-decoder model, may be useful in this regard in the future.

6. Conclusion

In this work, we have applied various deep neural networks, in particular LSTM- and GRU-based architectures with attention, to the problem of normalizing medical concepts expressed in the free-form language of social networks. We obtained very promising results, both quantitatively and qualitatively. Finally, we also showed that adding hand-crafted features does further improve performance.

We foresee three directions for future work. First, the use of novel architectures based on the recurrent neural networks that are currently used for machine translation and similar problems (e.g., dialogue and conversational models) for biomedical text processing looks promising and remains to be explored. Future work might focus on paraphrase generation and an encoder-decoder architecture since RNNs can be naturally used to probabilistically model a sequence. Second, a promising research direction is the integration of linguistic knowledge into the models. Third, future research might focus on developing extrinsic test sets for medical concept normalization.

Conflict of Interest

The authors declared that there is no conflict of interest.

Funding

This work was supported by the Russian Science Foundation Grant No. 18-11-00284.

References

[1] S. Pradhan, N. Elhadad, W.W. Chapman, S. Manandhar, G. Savova, Semeval-2014 task 7: Analysis of clinical text., in: SemEval@ COLING, 2014, pp. 54–62.
[2] S. Karimi, A. Metke-Jimenez, M. Kemp, C. Wang, Cadec: A corpus of adverse drug event annotations, J. Biomed. Informat. 55 (2015) 73–81.
[3] Z. Miftakhutdinov, E. Tutubalina, Kfu at clef ehealth 2017 task 1: Icd-10 coding of english death certificates with recurrent neural networks, CLEF, 2017.
[4] S. Han, T. Tran, A. Rios, R. Kavuluru, Team uknlp: Detecting adrs, classifying medication intake messages, and normalizing adr mentions on twitter, CEUR Workshop Proceedings 1996, 2017, pp. 49–53.
[5] M. Belousov, W. Dixon, G. Nenadic, Using an ensemble of generalised linear and deep learning models in the smm4h 2017 medical concept normalisation task, CEUR Workshop Proceedings 1996, 2017, pp. 54–58.
[6] N. Limsopatham, N. Collier, Normalising Medical Concepts in Social Media Texts by Learning Semantic Representation, in: ACL, 2016.
[7] A. Kotov, Social Media Analytics for Healthcare (2015) 309–340 <http://www.crcnetbase.com/doi/abs/10.1201/b18588-11>.
[8] C.-C. Huang, Z. Lu, Community challenges in biomedical text mining over 10 years: success, failure and the future, Briefings Bioinformat. 17 (1) (2015) 132–144.
[9] P.L. Schuyler, W.T. Hole, M.S. Tuttle, D.D. Sherertz, The umls metathesaurus: representing different views of biomedical concepts. Bull. Med. Libr. Assoc. 81 (2) (1993) 217.
[10] K.A. Spackman, K.E. Campbell, R.A. Côté, SNOMED RT: a reference terminology for health care, Proceedings of the AMIA Annual Fall Symposium, American Medical Informatics Association, 1997, p. 640.
[11] E.G. Brown, L. Wood, S. Wood, The medical dictionary for regulatory activities (meddra), Drug Safety 20 (2) (1999) 109–117.
[12] L. Biesecker, Mapping phenotypes to language: a proposal to organize and standardize the clinical descriptions of malformations, Clin. Genet. 68 (4) (2005) 320–326.
[13] A.R. Aronson, Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program, Proceedings of the AMIA Symposium, American Medical Informatics Association, 2001, p. 17.
[14] E. Van Mulligen, Z. Afzal, S.A. Akhondi, D. Vo, J.A. Kors, Erasmus MC at CLEF eHealth 2016: Concept recognition and coding in French texts, CLEF, 2016.
[15] L. Mottin, J. Gobeill, A. Mottaz, E. Pasche, A. Gaudinat, P. Ruch, Bitem at clef ehealth evaluation lab 2016 task 2: Multilingual information extraction., in: CLEF (Working Notes), 2016, pp. 94–102.
[16] O. Ghiasvand, R.J. Kate, Uwm: Disorder mention extraction from clinical text using crfs and normalization using learned edit distance patterns., in: SemEval@ COLING, 2014, pp. 828–832.
[17] Y.Z.J.W.B. Tang, Y.W.M. Jiang, Y.C.H. Xu, Uth_ccb: a report for semeval 2014–task 7 analysis of clinical text, SemEval 2014 (2014) 802.
[18] R. Leaman, R. Khare, Z. Lu, Ncbi at 2013 share/clef ehealth shared task: disorder normalization in clinical notes with dnorm, Radiology 42 (21.1) (2011) 1–941.
[19] R. Leaman, R. Islamaj Doğan, Z. Lu, DNorm: disease name normalization with pairwise learning to rank, Bioinformatics 29 (22) (2013) 2909–2917.
[20] M. Dermouche, V. Looten, R. Flicoteaux, S. Chevret, J. Velcin, N. Taright, ECSTRA-INSERM@ CLEF eHealth2016-task 2: ICD10 code extraction from death certificates, CLEF, 2016.
[21] P. Zweigenbaum, T. Lavergne, LIMSI ICD10 coding experiments on CépiDC death certificate statements, CLEF, 2016.
[22] K.B. Cohen, D. Demner-Fushman, Biomedical natural language processing vol. 11, John Benjamins Publishing Company, 2014.
[23] H. Suominen, S. Salanterä, S. Velupillai, W.W. Chapman, G. Savova, N. Elhadad, S. Pradhan, B.R. South, D.L. Mowery, G.J. Jones, et al., Overview of the share/clef ehealth evaluation lab 2013, International Conference of the Cross-Language

Evaluation Forum for European Languages, Springer, 2013, pp. 212–231.
[24] A. Névéol, R.N. Anderson, K.B. Cohen, C. Grouin, T. Lavergne, G. Rey, A. Robert, C. Rondet, P. Zweigenbaum, Clef ehealth 2017 multilingual information extraction task overview: Icd10 coding of death certificates in english and french, in: CLEF 2017 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS, 2017.
[25] P. Zweigenbaum, T. Lavergne, Hybrid methods for icd-10 coding of death certificates, EMNLP 2016 (2016) 96.
[26] R. Navigli, Word sense disambiguation: A survey, ACM Comput. Surv. (CSUR) 41 (2) (2009) 10.
[27] W. Shen, J. Wang, J. Han, Entity linking with a knowledge base: Issues, techniques, and solutions, IEEE Trans. Knowl. Data Eng. 27 (2) (2015) 443–460.
[28] Z. He, S. Liu, M. Li, M. Zhou, L. Zhang, H. Wang, Learning entity representation for entity disambiguation., in: ACL, No. 2, 2013, pp. 30–34.
[29] H. Huang, L. Heck, H. Ji, Leveraging deep neural networks and knowledge graphs for entity disambiguation, arXiv preprint arXiv:1504.07678.
[30] N. Gupta, S. Singh, D. Roth, Entity linking via joint encoding of types, descriptions, and context, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 2671–2680.
[31] A. Moro, A. Raganato, R. Navigli, Entity linking meets word sense disambiguation: a unified approach, Trans. Assoc. Comput. Linguist. 2 (2014) 231–244.
[32] A. Raganato, C.D. Bovi, R. Navigli, Neural sequence learning models for word sense disambiguation, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 1167–1178.
[33] S. Karimi, C. Wang, A. Metke-Jimenez, R. Gaire, C. Paris, Text and data mining techniques in adverse drug reaction detection, ACM Comput. Surv. (CSUR) 47 (4) (2015) 56.
[34] P. Nadkarni, R. Chen, C. Brandt, Umls concept indexing for production databases: a feasibility study, J. Am. Med. Inform. Assoc. 8 (1) (2001) 80–91.
[35] R.B. Polepalli, T. Houston, C. Brandt, H. Fang, H. Yu, Improving patients' electronic health record comprehension with noteaid, Studies Health Technol. Informat. 192 (2013) 714.
[36] R. Leaman, L. Wojtulewicz, R. Sullivan, A. Skariah, J. Yang, G. Gonzalez, Towards internet-age pharmacovigilance: Extracting adverse drug reactions from user posts to health-related social networks, in: Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, BioNLP '10, Association for Computational Linguistics, Stroudsburg, PA, USA, 2010, pp. 117–125.
[37] A. Nikfarjam, A. Sarker, K. OConnor, R. Ginn, G. Gonzalez, Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features, J. Am. Med. Inform. Assoc. (2015) ocu041.
[38] M. Oronoz, K. Gojenola, A. Pérez, A.D. de Ilarraza, A. Casillas, On the creation of a clinical gold standard corpus in spanish: Mining adverse drug reactions, J. Biomed. Informat. 56 (2015) 318–332.
[39] 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, J. Biomed. Informat. 62 (2016) 148–158.
[40] A. Sarker, A. Nikfarjam, G. Gonzalez, Social media mining shared task workshop, in: Proc. Pacific Symposium on Biocomputing 2016, 2016, pp. 581–592.
[41] Z. Miftahutdinov, E. Tutubalina, A. Tropsha, Identifying disease-related expressions in reviews using conditional random fields, in: Proceedings of International Conference on Computational Linguistics and Intellectual Technologies Dialog, vol. 1, 2017, pp. 155–167.
[42] E. Tutubalina, S. Nikolenko, Combination of deep recurrent neural networks and conditional random fields for extracting adverse drug reactions from user reviews, J. Healthcare Eng. (2017), http://dx.doi.org/10.1155/2017/9451342.
[43] C. VanDam, S. Kanthawala, W. Pratt, J. Chai, J. Huh, Detecting clinically related content in online patient posts, J. Biomed. Informat. (2017).
[44] A. Sarker, G. Gonzalez-Hernandez, Overview of the second social media mining for health (smm4h) shared tasks at amia 2017, CEUR Workshop Proceedings 1996, 2017, pp. 43–48.
[45] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016 < http://www.deeplearningbook.org > ..
[46] Y. Bengio, A. Courville, P. Vincent, Representation learning: A review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1798–1828, http://dx.doi.org/10.1109/TPAMI.2013.50.
[47] S. Hochreiter, J. Schmidhuber, Long Short-Term Memory, Neural Comput. 9 (8) (1997) 1735–1780 based on TR FKI-207-95, TUM (1995).
[48] F.A. Gers, J. Schmidhuber, Recurrent nets that time and count, Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on, vol. 3, IEEE, 2000, pp. 189–194.
[49] K. Cho, B. van Merrienboer, Ç. Gülçehre, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, CoRR abs/1406.1078. URL http://arxiv.org/abs/1406.1078.
[50] J. Chung, Ç. Gülçehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, CoRR abs/1412.3555. URL http://arxiv.org/abs/1412.3555.
[51] M. Schuster, K.K. Paliwal, Bidirectional recurrent neural networks, IEEE Trans. Signal Process. 45 (11) (1997) 2673–2681.
[52] Y. Goldberg, A primer on neural network models for natural language processing, CoRR abs/1510.00726. URL http://arxiv.org/abs/1510.00726.
[53] H. Rubenstein, J.B. Goodenough, Contextual correlates of synonymy, Commun. ACM 8 (10) (1965) 627–633, http://dx.doi.org/10.1145/365628.365657 http://doi.acm.org/10.1145/365628.365657.
[54] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, CoRR abs/1301.3781. URL http://arxiv.org/abs/1301.3781.
[55] S. Moen, T.S.S. Ananiadou, Distributional semantics resources for biomedical text processing (2013).
[56] Y. Kim, Convolutional neural networks for sentence classification, arXiv preprint arXiv:1408.5882.
[57] N. Srivastava, G.E. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (2014) 1929–1958.
[58] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: International Conference on Machine Learning, 2015, pp. 448–456.
[59] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, CoRR abs/1412.6980. URL http://arxiv.org/abs/1412.6980.
[60] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1026–1034.